North American Actuarial Journal

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

Fitting Censored and Truncated Regression Data Using the Mixture of Experts Models

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

To cite this article: Tsz Chai Fung, Andrei L. Badescu & X. Sheldon Lin (2022): Fitting Censored and Truncated Regression Data Using the Mixture of Experts Models, North American Actuarial Journal, DOI: 10.1080/10920277.2021.2013896

To link to this article: https://doi.org/10.1080/10920277.2021.2013896

Published online: 20 Jan 2022.

North American Actuarial Journal, 0(0), 1–25, 2022

ISSN: 1092-0277 print / 2325-0453 online
DOI: 10.1080/10920277.2021.2013896

Fitting Censored and Truncated Regression Data Using the Mixture of Experts Models

Tsz Chai Fung,[1] Andrei L. Badescu,[2] and X. Sheldon Lin[2]
[1]*Department of Risk Management and Insurance, Georgia State University, Atlanta, Georgia*
[2]*Department of Statistical Sciences, University of Toronto, Ontario Power Building, Toronto, Ontario, Canada*

The logit-weighted reduced mixture of experts model (LRMoE) is a flexible yet analytically tractable non-linear regression model. Though it has shown usefulness in modeling insurance loss frequencies and severities, model calibration becomes challenging when censored and truncated data are involved, which is common in actuarial practice. In this article, we present an extended expectation–conditional maximization (ECM) algorithm that efficiently fits the LRMoE to random censored and random truncated regression data. The effectiveness of the proposed algorithm is empirically examined through a simulation study. Using real automobile insurance data sets, the usefulness and importance of the proposed algorithm are demonstrated through two actuarial applications: individual claim reserving and deductible ratemaking.

1. INTRODUCTION

In general insurance ratemaking and reserving applications, it is often of interest to model the frequencies, severities, and reporting delays of insurance claims, and examine how the distributions of these quantities are influenced by policyholders' attributes and risk profiles. Hence, it is important to build a suitable model that not only enables capturing the distributional complexity (such as multimodality) but also flexibly captures the relationships between policyholders' characteristics and the claim distributions (including nonlinear patterns and interactions among explanatory variables). A suitable regression modeling framework can help insurance companies to appropriately set different premiums among policyholders with varying risk profiles (see, e.g., Fung, Badescu, and Lin 2019a)), and accurately determine the claim reserves (see, e.g., Fung, Badescu, and Lin 2020; Wang, Wu, and Qiu 2021).

To address the aforementioned modelling challenges, the logit-weighted reduced mixture of experts models (LRMoE), a class of flexible nonlinear regression models, was theoretically formulated by Fung, Badescu, and Lin (2019b) and has recently been useful in both ratemaking and reserving applications (Fung, Badescu, and Lin 2019a; 2020). LRMoE is regarded as a regression extension of the finite mixture model, where the effects of the policyholder's risk profile (covariates) are incorporated in the component weights of the mixture model. The LRMoE inherits merits from both parametric models and machine learning models, the two mainstream modeling frameworks. For example, being closed under marginalization and having simplified-form expressions for moments and measures of associations, the LRMoE is mathematically and statistically tractable like traditional parametric models. Also, satisfying several denseness properties, the LRMoE is potentially a universal approximator to any distribution and regression structures, and hence its flexibility is comparable to machine learning models such as neural networks.

Apart from the aforementioned desirable properties, the expectation–conditional maximization (ECM) algorithm presented by Fung, Badescu, and Lin (2019a) for efficient parameter estimations ensures that the LRMoE is computationally tractable. The ECM algorithm originated from the expectation–maximization (EM) algorithm introduced by Dempster, Laird, and Rubin (1977), which is a widely adopted approach that fits finite mixture models to insurance loss frequency or severity data; see, for example, Lee and Lin (2010), Badescu et al. (2015) and Miljkovic and Grün (2016). The ECM algorithm further divides the M-step into several computational feasible substeps such that the algorithm only requires low-dimensional convex optimizations that are easy to evaluate. Recently, Fung, Badescu, and Lin (2020) extended the ECM algorithm to incorporate parameter

Address correspondence to Tsz Chai Fung, Department of Risk Management and Insurance, Georgia State University, 35 Broad St NW, Atlanta, GA 30303. E-mail: tfung@gsu.edu

penalization. This addresses some problems inherited from finite mixture models such as unbounded likelihood and spurious fitted model.

Though the standard ECM algorithm requires that the observed data are complete, it is often not the case in practice when censored and truncated data are involved across various actuarial areas. From an insurance ratemaking perspective, insurance losses are often left truncated and right censored due to the presence of deductibles and policy limits (Frees and Valdez 2008). Reinsurers also encounter left-truncated loss data because insurance companies often report only the losses greater than a predetermined threshold to the reinsurers; see, for example, the Secura Re dataset discussed in Beirlant et al. (2006). Similarly, in risk management, operational risk datasets are left truncated because immaterial operational losses are not recorded (Badescu et al. 2015). Claim reserving is another actuarial area where the issue of incomplete data must be considered. Under individual reserving framework (see, e.g., Antonio and Plat 2014, Badescu, Lin, and Tang 2016; Verrall and Wüthrich 2016; Badescu et al. 2019), modeling the reporting delay of claims is essential for an adequate prediction of the incurred but not reported claims (IBNR). Reporting delay data can be interval censored because it is often recorded as the number of days instead of the exact time value. Claims are not observed if they are reported after an evaluation date, so the reporting delay data is also right truncated.

To address the above practical concerns, a few articles in actuarial science developed fitting algorithms for censored and truncated data. Verbelen et al. (2015) and Verbelen, Antonio, and Claeskens (2016) developed an EM algorithm that efficiently fits the mixture of Erlang distribution to censored and truncated loss data, using the approach of Lee and Scott (2012). Under this approach, censoring is handled by including uncensored observations in the complete data, and truncation is handled by reweighting the component weights of the finite mixture model. See also Reynkens et al. (2017) and Blostein and Miljkovic (2019) for other extensions using the same approach. Though this approach allows a convenient and efficient model calibration to censored and truncated data, two major limitations arise. Firstly, the algorithms assume a constant fixed truncation interval across all observations; however, as will soon be illustrated, random truncation is a common issue in actuarial practice. Secondly and more important, the algorithms focus on fitting loss distributions without considering regression, but some actuarial applications (such as ratemaking) involve policyholder information as covariates. To the best of our knowledge, these two shortcomings cannot be trivially addressed through extending the existing approach directly.

In addition to the actuarial literature, censoring and truncation problems for mixture-type models are widely studied in statistics and engineering literature. For example, Jaspers et al. (2014) studied the penalized mixture approach as a semi-parametric model for interval-censored data, and Bordes and Chauveau (2016) derived a stochastic EM algorithm for right-censored data under parametric and semi parametric mixture models. Ducros and Pamphile (2018) further developed a Bayesian restoration maximization algorithm for censored data under a Weibull mixture. These studies, however, did not incorporate regression that is necessary in insurance modeling at a granular level in the mixture models. In the context of mixture-based regression models, Mirfarah, Naderi, and Chen (2021) studied the EM estimation problem of mixture of linear experts model for censored data, with a special specification of scale mixture normal class as the mixture components. Nonetheless, the random truncation issue is yet to be investigated.

Motivated by the practical importance of modeling incomplete regression data and the limitations of the existing fitting algorithms, the main contribution in this article is to develop an extended ECM algorithm for fitting the LRMoE to random censored and truncated data. Handling random truncated data that appear in many practical problems, one cannot follow the same approach as in the current actuarial literature, which manipulates the component weights. Inspired by (yet not identical to) the (random) missing data construction technique proposed by Dempster, Laird, and Rubin (1977), in this article we construct "hypothetical" complete data where each observation itself "generates" some random missing data beyond the truncation intervals. This serves as an auxiliary tool that makes the complete data likelihood function computationally desirable and hence facilitates the implementation of the ECM algorithm under data censoring and truncation. Using this technique, our proposed ECM algorithm is capable of fitting random censored and random truncated regression data, addressing the two main limitations of the existing algorithms.

After deriving an efficient fitting algorithm, we illustrate its usefulness through two important non-life actuarial applications that involve random censored and truncated regression data. The first application is on individual claim reserving, where random truncation exists in the reporting delay data. In addition to showing that our proposed algorithm fits the reporting delay data very well, we introduce a new semiparametric method as a convenient alternative tool to predict the number of IBNR claims. With this method, the ECM fitting procedures (for reporting delay) automatically produce a good IBNR frequency prediction without the need to parametrically model the claim arrival process, as considered by most actuarial papers. Due to its simplicity and due to the inclusion of covariate information, our semi parametric method may produce very accurate reserve estimates and can replace the classical macrolevel models (chain ladder, Bornhuetter-Ferguson, etc.) that actuaries use in practice.

The second application is on deductible ratemaking, where the loss severity data involve random truncation because different policyholders may choose different deductible levels. Though treating the deductible level as a left truncation point (this article's approach) is the theoretically correct approach to modeling insurance losses, practitioners may also be interested in the regression approach, a more convenient and computationally less expensive approach where the deductible is treated as a covariate for ratemaking purposes (Lee 2017). Using a real automobile insurance dataset, we compare and contrast these two approaches from modeling and ratemaking perspectives, focusing on the shortcomings the later method may produce.

In a recent paper, Tseung et al. (2021) introduced a new Julia package, LRMoE.jl, statistical software tailor-made for actuarial applications that allows actuarial researchers and practitioners to model and analyze insurance loss frequencies and severities using the LRMoE model. LRMoE.jl offers several new actuarially motivated features. Key features include a wider coverage of actuarial distributions, and the flexibility to vary classes of distributions across components. The parameter estimations under data censoring and truncation based on the ECM algorithm proposed in the current article are under current development and continuous updating in our LRMoE.jl package. The source code and package documentation are available at https://github.com/sparktseung/LRMoE.jl and https://sparktseung.github.io/LRMoE.jl/dev/. Furthermore, we have developed an R package with similar functionalities for users interested in running the package in R instead. We refer such readers to Tseung et al. (2020) for the vignette, and https://github.com/sparktseung/LRMoE for the code and documentation.

This article is organized as follows. We first revisit the class of LRMoE introduced by Fung, Badescu, and Lin (2019b) in Section 2. Then, the censoring and truncation framework is introduced in Section 3. In Section 4, we develop the ECM algorithm that efficiently fits the LRMoE to random censored and truncated regression data. To examine the effectiveness of our proposed fitting algorithm and to illustrate the necessity of using our proposed algorithm to fit censoring and truncated regression data, a simulation study is presented in Section 5. Using real automobile insurance datasets, Sections 6 and 7 demonstrate the practical applicability of the proposed fitting algorithm to two important actuarial areas: individual claim reserving and deductible ratemaking. The main findings of this article and future research directions are summarized in Section 8.

2. THE MIXTURE OF EXPERTS MODEL

This section briefly revisits the class of logit-weighted reduced mixture of experts models (LRMoE) proposed by Fung, Badescu, and Lin (2019b) as a flexible regression model. Suppose that there are a total of n observations. Denote $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ and $\boldsymbol{y} = (y_1, ..., y_n)^T$ respectively the response variable column vector and the corresponding realization. For each sample $i = 1, ..., n,$ we also define $\boldsymbol{x}_i = (x_{i0}, ..., x_{iP})^T$ (with $x_{i0} = 1$) as the corresponding covariates. Assuming that $Y_1, ..., Y_n$ are mutually independent, the probability density function (pdf) of LRMoE is given by

$$h(y_i; \boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{\Psi}, g) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) f(y_i; \boldsymbol{\psi}_j), \qquad y_i > 0, \tag{2.1}$$

where g is the number of latent classes, $f(y_i; \boldsymbol{\psi}_j)$ is an expert function that governs the distributional property of Y_i, $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_g)$ 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 a gating function that governs the mixing weight for the *jth* class, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_g)$ with $\boldsymbol{\alpha}_j = (\alpha_{j0}, ..., \alpha_{jP})^T \in \mathbb{R}^{P+1}$ are the regression parameters of the mixing weights. To ensure model identifiability, we fix $\boldsymbol{\alpha}_g = \boldsymbol{0}$. Similarly, the cumulative density function (cdf) of the LRMoE is given by

$$H(y_i; \boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{\Psi}, g) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) F(y_i; \boldsymbol{\psi}_j), \qquad y_i > 0, \tag{2.2}$$

where $F(y_i; \boldsymbol{\psi}_j) = \int_0^{y_i} f(y; \boldsymbol{\psi}_j) dy$. We choose gamma expert function throughout this article with $\boldsymbol{\Psi} = (\boldsymbol{m}, \boldsymbol{\theta})$, $\boldsymbol{\psi}_j = (m_j, \theta_j)$ and

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

where $\boldsymbol{m} = (m_1, ..., m_g)$ and $\boldsymbol{\theta} = (\theta_1, ..., \theta_g)$ are respectively the shape and scale parameters of the gamma distribution.

The key motivation of using the Gamma-LRMoE is its model flexibility, where Fung, Badescu, and Lin (2019b) proved that the Gamma-LRMoE is dense in the space of any univariate severity regression distributions. This means that the Gamma-LRMoE is fully flexible in capturing any distributional and regression structures, including distributional multimodality, nonlinear regression links, and interactions among covariates, even if the LRMoE contains only linear regressions on the gating functions, suggesting that the LRMoE is a parsimonious model class. As a result, regardless of the complexities of the model generating the input data, the characteristics of the fitted LRMoE will be highly synchronous to those of the input data. This theoretical result was also empirically justified through several simulation studies and real insurance data analyses by Fung, Badescu, and Lin (2019a; 2020). The Gamma-LRMoE also possesses several desirable properties, including mathematical tractability and model identifiability, thoroughly examined by Fung, Badescu, and Lin (2019a,b).

Remark 1. A competing model class to the LRMoE is the finite mixture of regression (FMR) model (McLachlan and Peel 2000). Instead of incorporating regressions in the gating function, the regression links in FMR are incorporated in the expert functions through the parameters $\boldsymbol{\Psi}$. Though both LRMoE and FMR are flexible in capturing complex distributional characteristics including multimodality, the denseness property generally does not hold for the class of FMR in regression setting (Fung, Badescu, and Lin 2019b), meaning that the FMR is not guaranteed to well resemble the characteristics of all dataset, especially for regression links and more complex features that can hardly be visualized in practice (e.g., interactions between regression links and distributional structures; see, for example, figure 5 of Fung, Badescu, and Lin 2020). This motivates us to adopt the LRMoE instead of the FMR as the modeling framework.

Needless to say, the main focus of this article is to introduce a novel computational strategy to deal with censored and truncated data (Sections 3 and 4) and demonstrate the empirical and practical importance of appropriately treating data censoring and truncation (Sections 5 to 7). As will be discussed in Remarks 3 and 6, our proposed strategy is necessary and useful for any mixture-based regression models (including not only the LRMoE but also the FMR) under censoring and truncation mechanisms. Because model selections between the LRMoE and FMR are not our main research interest, we present our results only under the LRMoE modelling framework for the purpose of conciseness.

The LRMoE is also interpretable in the general insurance context. Each observation (e.g., policyholder or claim) is classified into one of the g latent risk subgroups through the gating functions. The probabilities of subgroup assignments can be affected by the covariates (e.g., policyholder or claim characteristics) $\boldsymbol{x}_i$. Loss distributions, which are governed by the expert functions, vary among subgroups but are homogeneous within a subgroup, so different subgroups have different risk levels.

3. CENSORING AND TRUNCATION

In general insurance practice, the true values of response variables $\boldsymbol{y}$ may be observed inexact (censoring) and may not be fully observed (truncation). For example, the reporting delay of a claim can only be observed if the claim is reported before a valuation date, and it is usually recorded as the number of days instead of an exact time (continuous value). This section formally constructs the mechanisms of censoring and truncation, and defines the relevant notations. We will also state the underlying assumptions and provides visualizations to illustrate the practical meanings of the mechanism.

3.1. Formalism

Define $\boldsymbol{T}_i := (T_i^l, T_i^u)$ as the lower and upper random truncation points of the ith observation with the corresponding realizations $\boldsymbol{t}_i := (t_i^l, t_i^u)$, where we have $0 \leq T_i^l < T_i^u \leq \infty$. $\boldsymbol{T}_i$ is said to be the *truncation mechanism* of observation i. By truncation we mean that the data point is observed conditioned on the situation that the response variable Y_i falls into the truncation range $[Y_i^l, Y_i^u]$.

Define $\mathcal{R}_i^U \subseteq [T_i^l, T_i^u]$ as the random uncensoring region of observation i and similarly $\mathcal{R}_i^C = [T_i^l, T_i^u] \backslash \mathcal{R}_i^U$ as the random censoring region. Denote $\{I_{i1}, ..., I_{iS_i}\}$ as the S_i disjoint random censoring intervals of observation i with $\cup_{s=1}^{S_i} I_{is} = \mathcal{R}_i^C$. Note that S_i is assumed to be random. We then call $\boldsymbol{C}_i := (\mathcal{R}_i^U, \mathcal{R}_i^C, S_i, \{I_{i1}, ..., I_{iS_i}\})$ the *censoring mechanism* of observation i.

It is assumed that the censoring and truncation mechanisms $(\boldsymbol{C}_i, \boldsymbol{T}_i)$ are independent of the true response variable Y_i conditioned on the covariates $\boldsymbol{x}_i$. We will explain in Sections 5 to 7 why this assumption is reasonable in a general insurance context. Define the lower and upper censoring points of observation i as

$$Y_i^l = \begin{cases} Y_i, & Y_i \in \mathcal{R}_i^U \\ \sum_{s=1}^{S_i} 1\{Y_i \in I_{is}\} \inf\{I_{is}\}, & Y_i \in \mathcal{R}_i^C \end{cases}, \quad \text{and} \quad Y_i^u = \begin{cases} Y_i, & Y_i \in \mathcal{R}_i^U \\ \sum_{s=1}^{S_i} 1\{Y_i \in I_{is}\} \sup\{I_{is}\}, & Y_i \in \mathcal{R}_i^C \end{cases}, \tag{3.1}$$

where both Y_i^l and Y_i^u are undefined if Y_i is outside of the truncation interval $[T_i^l, T_i^u]$. The above censoring mechanism is interpreted as follows. If the true response Y_i falls into the uncensoring range $\mathcal{R}_i^U$, we will have $Y_i^l = Y_i^u = Y_i$ such that observation i is said to be observed in exact or *uncensored*. Otherwise, if Y_i falls into the censoring range $\mathcal{R}_i^C$, the lower and upper censoring points (Y_i^l, Y_i^u) will be the lower and upper end points of the censoring interval I_{is} that the true response Y_i belongs to. In this case, we only know the range $[Y_i^l, Y_i^u]$ that the true response variable Y_i belongs to. In the special case where $T_i^l = Y_i^l < Y_i^u < T_i^u$, we call the observation *left censored*. The observation is *right censored* if $T_i^l < Y_i^l < Y_i^u = T_i^u$. In the general case where $T_i^l < Y_i^l < Y_i^u < T_i^u$, the observation is *interval censored*.

Overall, under the censoring and truncation mechanisms, instead of observing the true response Y_i, we observe only the corresponding lower and upper censoring points, as well as the censoring and truncation mechanisms, given that Y_i is inside the truncation interval, denoted as

$$\{(Y_i^l, Y_i^u, \boldsymbol{T}_i, \boldsymbol{C}_i) | Y_i \in \left[T_i^l, T_i^u\right]\},$$

and the corresponding realized ith observation is denoted by $(y_i^l, y_i^u, t_i^l, t_i^u, \boldsymbol{c}_i)$. The observations are assumed to be independent and identically distributed (iid) across $i = 1, ..., n$.

3.2. Visualization

We hereby present Figure 1 to help readers understand the censoring and truncation mechanisms through three examples. In the figure, the shaded areas are the uncensored intervals, the semi-open brackets represent the censored intervals, the two long vertical bars in each examples are the truncation ranges, and the stars are the true response Y_i, which is not fully observed in practice (as described by the previous subsection). Example 1 showcases a typical interval-censored observation where Y_i falls inside the censoring interval I_{i1}, so in practice we observe $(Y_i^l, Y_i^u]$ as the range of possible values of (unknown) response variable Y_i, with $Y_i^l < Y_i^u$ displayed in the appropriate places in the figure. Example 2 is the case where the observation is uncensored. Because the response Y_i falls inside an uncensored region $\mathcal{R}_i^U$, the response Y_i is observed in exact and we have $Y_i^l = Y_i^u = Y_i$. Example 3 is the case where the true response is outside the truncation interval $[T_i^l, T_i^u]$. This case is impossible and hence is not considered because the observations are conditioned on $Y_i \in [T_i^l, T_i^u]$.

4. PARAMETER ESTIMATION – AN ECM ALGORITHM

Model calibration of the LRMoE can be efficiently achieved through the ECM algorithm presented by Fung, Badescu, and Lin (2019a), which assumes that the response variables $\boldsymbol{y}$ are exact and fully observed. Therefore, in this section we develop an efficient fitting algorithm for the LRMoE that caters for censored and truncated data, which is the main contribution of this article. In Subsection 4.1, we identify the challenges for parameter estimation and present a novel complete data construction

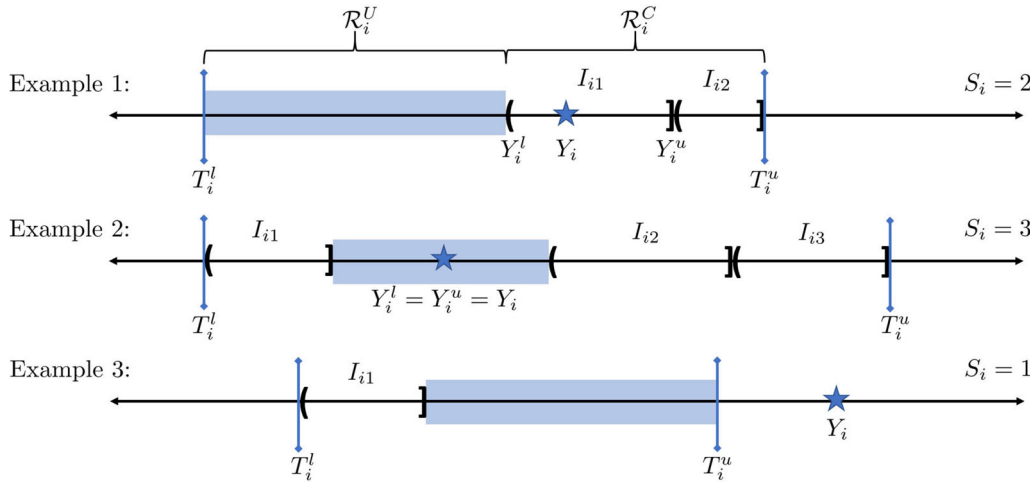

FIGURE 1. Visualizations of the Censoring and Truncation Mechanisms through Three Examples.

technique that makes an efficient ECM algorithm possible for the LRMoE even if the data are censored and truncated. Then, the procedures involved in the adjusted ECM algorithm will be presented in Subsections 4.2.1 to 4.2.3.

4.1. Likelihood Function and Complete Data Construction

Denote the observed data as $\mathcal{D}^{\text{obs}} = \{(y_i^l, y_i^u, t_i^l, t_i^u, \boldsymbol{c}_i)\}_{i=1,\ldots,n}$; the observed data likelihood is proportional to

$$\begin{aligned}\mathcal{L}^{\text{obs}}(\boldsymbol{\Phi}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}) &= \prod_{i\in\mathcal{C}} \frac{H(y_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}) - H(y_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi})}{H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}) - H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi})} \prod_{i\in\mathcal{U}} \frac{h(y_i; \boldsymbol{x}_i, \boldsymbol{\Phi})}{H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}) - H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi})} \\ &= \prod_{i\in\mathcal{C}} \frac{\sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) \left[F(y_i^u; \boldsymbol{\psi}_j) - F(y_i^l; \boldsymbol{\psi}_j)\right]}{\sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) \left[F(t_i^u; \boldsymbol{\psi}_j) - F(t_i^l; \boldsymbol{\psi}_j)\right]} \\ &\quad \times \prod_{i\in\mathcal{U}} \frac{\sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) f(y_i; \boldsymbol{\psi}_j)}{\sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) \left[F(t_i^u; \boldsymbol{\psi}_j) - F(t_i^l; \boldsymbol{\psi}_j)\right]},\end{aligned} \tag{4.1}$$

where $\boldsymbol{\Phi} = (\boldsymbol{\alpha}, \boldsymbol{\Psi}, g)$, $\mathcal{C} \subseteq \{1, \ldots, n\}$ is the subset of observations that are censored (i.e., $y_i^l \neq y_i^u$) and $\mathcal{U} \subseteq \{1, \ldots, n\}$ is the subset of observations that are uncensored (i.e., $y_i^l = y_i^u = y_i$). Note above that the censoring mechanism $\boldsymbol{c}_i$ is not involved in the log-likelihood function. It is difficult to optimize the likelihood directly.

Remark 2. Verbelen et al. (2015) proposed a re-weighting scheme that is effective in simplifying the likelihood function of finite mixture distributions under censoring and truncation and hence makes an EM algorithm implementable. However, such a special technique cannot be extended to simplify Equation (4.1) when regression is incorporated and different truncation intervals among observations are assumed. The arguments are as follows. Consider an uncensored observation $i \in \mathcal{U}$. Following the procedures of Verbelen et al. (2015), the individual observed data likelihood is written as

$$\begin{aligned}\mathcal{L}_i^{\text{obs}}(\boldsymbol{\Phi}) &:= \frac{\sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) f(y_i; \boldsymbol{\psi}_j)}{H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}) - H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi})} \\ &= \sum_{j=1}^{g} \frac{\pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) \left[F(t_i^u; \boldsymbol{\psi}_j) - F(t_i^l; \boldsymbol{\psi}_j)\right]}{H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}) - H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi})} \frac{f(y_i; \boldsymbol{\psi}_j)}{F(t_i^u; \boldsymbol{\psi}_j) - F(t_i^l; \boldsymbol{\psi}_j)} \\ &:= \sum_{j=1}^{g} \tilde{\pi}_j(\boldsymbol{x}_i, t_i^l, t_i^u; \boldsymbol{\alpha}) \tilde{f}(y_i; \boldsymbol{\psi}_j),\end{aligned}$$

with transformed weight $\tilde{\pi}_j(\boldsymbol{x}_i, t_i^l, t_i^u; \boldsymbol{\alpha})$ and transformed density $\tilde{f}(y_i; \boldsymbol{\psi}_j)$. The successful key by Verbelen et al. (2015) is that the transformed weight parameters $\tilde{\pi}_j$ are estimated instead of the original weight parameters. This requires that the transformed weights $\tilde{\pi}_j$ are unified across all observations, which is, however, not the case if either regression is incorporated or truncation intervals are random. Similar arguments hold for censored observation $i \in \mathcal{C}$.

Remark 3. Applying the same arguments as Remark 2, the transformed weight $\tilde{\pi}_j$ would still differ across observations even if the model class is FMR instead of LRMoE. Therefore, the reweighting scheme proposed by Verbelen et al. (2015) would also fail for censored and truncated data under the FMR. This highlights the importance of devising a new computational strategy suitable for fitting censored and truncated data using finite mixture-based regression models.

Motivated by the construction technique proposed in Subsection 4.2 of Dempster, Laird, and Rubin (1977), we introduce "hypothetical" complete data that lead to a much simpler likelihood function compared to Equation (4.1), enabling an extension of the ECM algorithm to censored and truncated data. The complete data are given by

$$\mathcal{D}^{\text{com}} = (\boldsymbol{y}, \boldsymbol{k}, \{\boldsymbol{y}_i'\}_{i=1,\ldots,n}, \{\boldsymbol{z}_i\}_{i=1,\ldots,n}, \{\boldsymbol{z}_{is}'\}_{i=1,\ldots,n; s=1,\ldots,k_i}), \tag{4.2}$$

which consists of the following five elements:

- $\boldsymbol{y} = (y_1, \ldots, y_n)^T$: a vector of true (uncensored) values within the truncation intervals.
- $\boldsymbol{k} = (k_1, \ldots, k_n)^T$: the realization of a multivariate random vector $\boldsymbol{K} = (K_1, \ldots, K_n)^T$, where K_i represents the number of missing sample points (i.e., points that falls outside the truncation interval) corresponding to "generated by" the ith observation, with covariates $\boldsymbol{x}_i$ and truncation interval $[t_i^l, t_i^u]$.
- $\boldsymbol{y}'_i = (y'_{i1}, \ldots, y'_{ik_i})$: the realization of a multivariate vector $\boldsymbol{Y}'_i = (Y'_{i1}, \ldots, Y'_{ik_i})$, which represents the missing sample points corresponding to the ith observation.
- $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ig})^T$: the realization of a multinomial latent random vector $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ig})$ such that $Z_{ij} = 1$ if the ith observation comes from the jth component of the LRMoE and $Z_{ij} = 0$ otherwise.
- $\boldsymbol{z}'_{is} = (z'_{is1}, \ldots, z'_{isg})^T$: the realization of a multinomial latent random vector $\boldsymbol{Z}'_{is} = (Z'_{is1}, \ldots, Z'_{isg})$ such that $Z'_{isj} = 1$ if the sth missing sample point corresponding to the ith observation comes from the jth component and $Z'_{isj} = 0$ otherwise.

We assume that $k_1, ..., k_n$ are mutually independent and are independent of the remaining elements of the complete data. Also, k_i is artificially constructed to follow a geometric distribution with probability mass function

$$p(k_i; \boldsymbol{x}_i, \boldsymbol{\Phi}) = \left[1 - H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}) + H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi})\right]^{k_i} \left[H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}) - H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi})\right], \quad k_i = 0, 1, ... \tag{4.3}$$

In short, the complete data consist of uncensored observed data within the truncation interval and missing data outside the truncation interval. We call the complete data "hypothetical" because the distribution made above is unlikely to be realistic. Missing sample points are "generated by" the observed sample points and the covariates of the missing sample points are assumed to be identical to that of the observed sample points, but these obviously make no sense in reality. The geometric distribution in Equation (4.3) also may not adequately represent the true distribution on the number of missing sample points; for example, when the insurance claim arrival follows a Poisson process. Indeed, we do not claim that in reality the number of missing samples must follow the geometric distribution as mentioned. Instead, we will show that the complete data likelihood function is much more computationally desirable than the observed data likelihood function in Equation (4.1). In other words, the aforementioned artificial construction of hypothetical complete data provides a distribution extension of the missing samples that facilitates the implementation of the ECM algorithm under data censoring and truncation. The complete data likelihood function is given by

$$\begin{aligned}
&\mathcal{L}^{\text{com}}(\boldsymbol{\Phi}; \mathcal{D}^{\text{com}}, \boldsymbol{x}) \\
&= \prod_{i=1}^{n} \{P(\boldsymbol{Z}_i = \boldsymbol{z}_i, Y_i = y_i | Y_i \in \left[t_i^l, t_i^u\right], \boldsymbol{x}_i, \boldsymbol{\Phi}) \times p(k_i; \boldsymbol{x}_i, \boldsymbol{\Phi}) \\
&\quad \times \prod_{s=1}^{k_i} P(\boldsymbol{Z}'_{is} = \boldsymbol{z}'_{is}, Y'_{is} = y'_{is} | Y'_{is} \notin \left[t_i^l, t_i^u\right], \boldsymbol{x}_i, \boldsymbol{\Phi})\} \\
&= \prod_{i=1}^{n} \{P(Y_i = y_i | Y_i \in \left[t_i^l, t_i^u\right], \boldsymbol{x}_i, \boldsymbol{\Phi}) P(\boldsymbol{Z}_i = \boldsymbol{z}_i | Y_i = y_i, \boldsymbol{x}_i, \boldsymbol{\Phi}) p(k_i; \boldsymbol{x}_i, \boldsymbol{\Phi}) \\
&\quad \times \prod_{s=1}^{k_i} P(Y'_{is} = y'_{is} | Y'_{is} \notin \left[t_i^l, t_i^u\right], \boldsymbol{x}_i, \boldsymbol{\Phi}) P(\boldsymbol{Z}'_{is} = \boldsymbol{z}'_{is} | Y'_{is} = y'_{is}, \boldsymbol{x}_i, \boldsymbol{\Phi})\} \\
&= \prod_{i=1}^{n} \Bigg\{ \left[\prod_{j=1}^{g} \left(\frac{h(y_i; \boldsymbol{x}_i, \boldsymbol{\Phi})}{H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}) - H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi})} \times \frac{\pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) f(y_i; \boldsymbol{\psi}_j)}{h(y_i; \boldsymbol{x}_i, \boldsymbol{\Phi})} \right)^{z_{ij}} \right] \\
&\qquad \times \left[1 - H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}) + H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi})\right]^{k_i} \left[H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}) - H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi})\right] \\
&\qquad \times \prod_{s=1}^{k_i} \left[\prod_{j=1}^{g} \left(\frac{h(y'_{is}; \boldsymbol{x}_i, \boldsymbol{\Phi})}{1 - H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}) + H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi})} \times \frac{\pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) f(y'_{is}; \boldsymbol{\psi}_j)}{h(y'_{is}; \boldsymbol{x}_i, \boldsymbol{\Phi})} \right)^{z'_{isj}} \right] \Bigg\} \\
&= \prod_{i=1}^{n} \prod_{j=1}^{g} \left(\pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) f(y_i; \boldsymbol{\psi}_j)\right)^{z_{ij}} \times \prod_{i=1}^{n} \prod_{s=1}^{k_i} \prod_{j=1}^{g} \left(\pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) f(y'_{is}; \boldsymbol{\psi}_j)\right)^{z'_{isj}}.
\end{aligned} \tag{4.4}$$

Note that the geometric distribution on $k_1, ..., k_n$ acts as a key to cancel out $H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}) - H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi})$ and $1 - H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}) + H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi})$ in Equation (4.4) that consist of summation of terms. Therefore, optimizing the complete data likelihood function (involving the product of terms only) is much easier than optimizing the observed data likelihood function (Eq. [4.1], involving the product of sums).

A natural approach for model calibration is to find the parameters $(\boldsymbol{\alpha}, \boldsymbol{\Psi})$ that maximize the likelihood function. However, as discussed by McLachlan and Peel (2000) and Fung, Badescu, and Lin (2020), it is possible that the likelihood function is unbounded for severity distributions when $m_j \to \infty$ and $\theta_j \to 0$ for some j such that some mixture components have very small variances specially fitting only one observation, leading to a spurious fitted model. To address such an issue, we adopt the same approach as Fung, Badescu, and Lin (2020) that penalizes parameters taking extreme values through finding the maximum a posteriori estimates of the parameters. This would prevent the fitted parameters from inflating indefinitely and/or shrinking to zero, which leads to a spurious fitted model. By assuming that all parameters are a priori independent, the observed data posterior log-likelihood is given by

$$\begin{aligned}\tilde{l}^{\text{obs}}(\boldsymbol{\Phi}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}) &= \log\left[\frac{\mathcal{L}^{\text{obs}}(\boldsymbol{\Phi}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}) p(\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\theta})}{p(\boldsymbol{y}; \boldsymbol{x})}\right] \\ &= l^{\text{obs}}(\boldsymbol{\Phi}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}) + \log p(\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\theta}) + \text{const}, \end{aligned} \tag{4.5}$$

where

$$\log p(\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\theta}) = \sum_{j=1}^{g-1} \sum_{p=0}^{P} \log p_1(\alpha_{jp}) + \sum_{j=1}^{g} \log p_2(m_j) + \sum_{j=1}^{g} \log p_3(\theta_j), \tag{4.6}$$

$p(\cdot)$ represents the joint prior distribution of the parameters, and $p_1(\cdot)$, $p_2(\cdot)$ and $p_3(\cdot)$ are the marginal prior density functions. We also follow Fung, Badescu, and Lin (2020) in choosing the following prior distributions of parameters: $\alpha_{jp} \sim N(0, \sigma_{jp}^2)$ for $j = 1, ..., g-1$ and $p = 0, ..., P$, $m_j \sim Gamma(\nu_j^{(1)}, \lambda_j^{(1)})$ for $j = 1, ..., g$, and $\theta_j \sim Gamma(\nu_j^{(2)}, \lambda_j^{(2)})$ for $j = 1, ..., g$, where σ_{jp}, $\nu_j^{(1)}$, $\lambda_j^{(1)}$, $\nu_j^{(2)}$, $\lambda_j^{(2)}$ are all fixed numbers.

Remark 4. Following subsection 5.3.1 of Fung, Badescus, and Lin (2020), for each $j = 1, ..., g$ we choose the hyperparameters $\sigma_{j0} = 3$, $\sigma_{jp} = 2/(\max_{i=1,...,n}\{x_{ip}\} - \min_{i=1,...,n}\{x_{ip}\})$ $(p > 0)$, $\nu_j^{(1)} = \nu_j^{(2)} = 1$, and $\lambda_j^{(1)} = \lambda_j^{(2)} = 500$. The resulting prior distributions are weak priors, allowing for minimal distortions to the fitted model yet effectively addressing the unbounded likelihood problem.

Similarly, the complete data posterior log-likelihood is given by

$$\begin{aligned}&\tilde{l}^{\text{com}}(\boldsymbol{\Phi}; \mathcal{D}^{\text{com}}, \boldsymbol{x}) \\ &= \log \mathcal{L}^{\text{com}}(\boldsymbol{\Phi}; \mathcal{D}^{\text{com}}, \boldsymbol{x}) + \log p(\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\theta}) + \text{const.} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} \left[\log \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) + (m_j - 1) \log y_i - \frac{y_i}{\theta_j} - m_j \log \theta_j - \log \Gamma(m_j)\right] \\ &\quad + \sum_{i=1}^{n} \sum_{s=1}^{k_i} \sum_{j=1}^{g} z'_{isj} \left[\log \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) + (m_j - 1) \log y'_{is} - \frac{y'_{is}}{\theta_j} - m_j \log \theta_j - \log \Gamma(m_j)\right] \\ &\quad - \sum_{j=1}^{g} \sum_{p=0}^{P} \frac{\alpha_{jp}^2}{2\sigma_{jp}^2} + \sum_{j=1}^{g} \left((\nu_j^{(1)} - 1) \log m_j - \frac{m_j}{\lambda_j^{(1)}}\right) + \sum_{j=1}^{g} \left((\nu_j^{(2)} - 1) \log \theta_j - \frac{\theta_j}{\lambda_j^{(2)}}\right) + \text{const.} \end{aligned} \tag{4.7}$$

4.2. The ECM Algorithm

After formulating the complete data posterior log-likelihood function, which takes a simple analytical form, we are able to develop the ECM algorithm of the LRMoE under censoring and truncation.

4.2.1. E-Step

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

$$
\begin{aligned}
&Q(\boldsymbol{\Phi}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}) \\
&= E\left[\tilde{l}^{\text{com}}(\boldsymbol{\Phi}; \boldsymbol{Y}, \boldsymbol{K}, \{\boldsymbol{Y'}_i\}_{i=1,\ldots,n}, \{\boldsymbol{Z}_i\}_{i=1,\ldots,n}, \{\boldsymbol{Z'}_{is}\}_{i=1,\ldots,n;s=1,\ldots,k_i}, \boldsymbol{x}) | \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}^{(l)} \left[\log \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) + (m_j - 1)\, \widehat{\log y}_{ij}^{(l)} - \frac{\hat{y}_{ij}^{(l)}}{\theta_j} - m_j \log \theta_j - \log \Gamma(m_j)\right] \\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{g} k_i^{(l)} z_{ij}^{'(l)} \left[\log \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) + (m_j - 1)\, \widehat{\log y'}_{ij}^{(l)} - \frac{\hat{y'}_{ij}^{(l)}}{\theta_j} - m_j \log \theta_j - \log \Gamma(m_j)\right] \\
&\quad - \sum_{j=1}^{g}\sum_{p=0}^{P} \frac{\alpha_{jp}^2}{2\sigma_{jp}^2} + \sum_{j=1}^{g}\left((\nu_j^{(1)} - 1)\log m_j - \frac{m_j}{\lambda_j^{(1)}}\right) + \sum_{j=1}^{g}\left((\nu_j^{(2)} - 1)\log \theta_j - \frac{\theta_j}{\lambda_j^{(2)}}\right) + \text{const},
\end{aligned} \tag{4.8}
$$

where

$$
z_{ij}^{(l)} = P(Z_{ij} = 1 | \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}) = \begin{cases} \dfrac{\pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}^{(l-1)})\left[F(y_i^u; \boldsymbol{\psi}_j^{(l-1)}) - F(y_i^l; \boldsymbol{\psi}_j^{(l-1)})\right]}{H(y_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}^{(l-1)}) - H(y_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi}^{(l-1)})}, & i \in \mathcal{C} \\ \dfrac{\pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}^{(l-1)}) f(y_i; \boldsymbol{\psi}_j^{(l-1)})}{h(y_i; \boldsymbol{x}_i, \boldsymbol{\Phi}^{(l-1)})}, & i \in \mathcal{U} \end{cases}, \tag{4.9}
$$

$$
\begin{aligned}
\hat{y}_{ij}^{(l)} &= E(Y_i | \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}, Z_{ij} = 1) \\
&= \begin{cases} \dfrac{\left[F(y_i^u; m_j^{(l-1)} + 1, \theta_j^{(l-1)}) - F(y_i^l; m_j^{(l-1)} + 1, \theta_j^{(l-1)})\right] m_j^{(l-1)} \theta_j^{(l-1)}}{F(y_i^u; \boldsymbol{\psi}_j^{(l-1)}) - F(y_i^l; \boldsymbol{\psi}_j^{(l-1)})}, & i \in \mathcal{C} \\ y_i, & i \in \mathcal{U} \end{cases},
\end{aligned} \tag{4.10}
$$

$$
\widehat{\log y}_{ij}^{(l)} = E(\log Y_i | \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}, Z_{ij} = 1) = \begin{cases} \dfrac{\int_{y_i^l}^{y_i^u} \log y\, f(y; \boldsymbol{\psi}_j^{(l-1)}) dy}{F(y_i^u; \boldsymbol{\psi}_j^{(l-1)}) - F(y_i^l; \boldsymbol{\psi}_j^{(l-1)})}, & i \in \mathcal{C} \\ \log y_i, & i \in \mathcal{U} \end{cases}, \tag{4.11}
$$

$$
k_i^{(l)} = E(K_i | \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}) = \frac{1 - H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}^{(l-1)}) + H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi}^{(l-1)})}{H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}^{(l-1)}) - H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi}^{(l-1)})}, \tag{4.12}
$$

$$
z_{ij}^{'(l)} = P(Z'_{isj} = 1 | \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}) = \frac{\pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}^{(l-1)})\left[1 - F(t_i^u; \boldsymbol{\psi}_j^{(l-1)}) + F(t_i^l; \boldsymbol{\psi}_j^{(l-1)})\right]}{1 - H(t_i^u; \boldsymbol{x}_i, \boldsymbol{\Phi}^{(l-1)}) + H(t_i^l; \boldsymbol{x}_i, \boldsymbol{\Phi}^{(l-1)})}, \tag{4.13}
$$

$$
\begin{aligned}
\hat{y'}_{ij}^{(l)} &= E(Y'_{is} | \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}, Z'_{isj} = 1) \\
&= \frac{\left[1 - F(t_i^u; m_j^{(l-1)} + 1, \theta_j^{(l-1)}) + F(t_i^l; m_j^{(l-1)} + 1, \theta_j^{(l-1)})\right] m_j^{(l-1)} \theta_j^{(l-1)}}{1 - F(t_i^u; \boldsymbol{\psi}_j^{(l-1)}) + F(t_i^l; \boldsymbol{\psi}_j^{(l-1)})},
\end{aligned} \tag{4.14}
$$

$$\widehat{\log y'}_{ij}^{(l)} = E(\log Y'_{is} | \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}, Z'_{isj} = 1) = \frac{\int_0^{t_i^l} \log y \, f(y; \boldsymbol{\psi}_j^{(l-1)}) dy + \int_{t_i^u}^{\infty} \log y \, f(y; \boldsymbol{\psi}_j^{(l-1)}) dy}{1 - F(t_i^u; \boldsymbol{\psi}_j^{(l-1)}) + F(t_i^l; \boldsymbol{\psi}_j^{(l-1)})}. \tag{4.15}$$

4.2.2. CM-Step

The CM-step involves updating the parameters $\boldsymbol{\Phi}^{(l-1)}$ to $\boldsymbol{\Phi}^{(l)}$ such that $Q(\boldsymbol{\Phi}^{(l)}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}) \geq Q(\boldsymbol{\Phi}^{(l-1)}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)})$. One appealing feature of $Q(\boldsymbol{\Phi}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)})$ (Eq. [4.8]) is that many parameters can be linearly separated through the following decomposition:

$$Q(\boldsymbol{\Phi}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}) = S^{(l)}(\boldsymbol{\alpha}) + \sum_{j=1}^{g} T_j^{(l)}(\boldsymbol{\psi}_j) + \text{const}, \tag{4.16}$$

where

$$S^{(l)}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{j=1}^{g} \left(z_{ij}^{(l)} + k_i^{(l)} z_{ij}^{\prime(l)} \right) \log \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) - \sum_{j=1}^{g-1} \sum_{p=0}^{P} \frac{\alpha_{jp}^2}{2\sigma_{jp}^2}, \tag{4.17}$$

$$\begin{aligned} T_j^{(l)}(\boldsymbol{\psi}_j) = {} & (m_j - 1) \sum_{i=1}^{n} \left(z_{ij}^{(l)} \widehat{\log y}_{ij}^{(l)} + k_i^{(l)} z_{ij}^{\prime(l)} \widehat{\log y'}_{ij}^{(l)} \right) + \frac{1}{\theta_j} \sum_{i=1}^{n} \left(z_{ij}^{(l)} \hat{y}_{ij}^{(l)} + k_i^{(l)} z_{ij}^{\prime(l)} \hat{y'}_{ij}^{(l)} \right) \\ & - \left(m_j \log \theta_j + \log \Gamma(m_j) \right) \sum_{i=1}^{n} \left(z_{ij}^{(l)} + k_i^{(l)} z_{ij}^{\prime(l)} \right) \\ & + \left((\nu_j^{(1)} - 1) \log m_j - \frac{m_j}{\lambda_j^{(1)}} \right) + \left((\nu_j^{(2)} - 1) \log \theta_j - \frac{\theta_j}{\lambda_j^{(2)}} \right). \end{aligned} \tag{4.18}$$

We first update $\boldsymbol{\alpha}^{(l-1)}$ to $\boldsymbol{\alpha}^{(l)}$ so that $S^{(l)}(\boldsymbol{\alpha}^{(l)}) \geq S^{(l)}(\boldsymbol{\alpha}^{(l-1)})$. We adopt a conditional maximization approach that sequentially maximizes $S^{(l)}(\boldsymbol{\alpha}_1^{(l)}, ..., \boldsymbol{\alpha}_{j-1}^{(l)}, \boldsymbol{\alpha}_j, \boldsymbol{\alpha}_{j+1}^{(l-1)}, ..., \boldsymbol{\alpha}_g^{(l-1)})$ with respect to $\boldsymbol{\alpha}_j$ to obtain $\boldsymbol{\alpha}_j^{(l)}$. This can be done by the iteratively reweighted least squares approach (Jordan and Jacobs 1994), which conducts the following iterations until convergence:

$$\boldsymbol{\alpha}_j \leftarrow \boldsymbol{\alpha}_j - \left(\frac{\partial^2 S^{(l)}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\alpha}_j^T} \right)^{-1} \frac{\partial S^{(l)}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_j}. \tag{4.19}$$

Because $S^{(l)}(\boldsymbol{\alpha})$ is a concave function, convergence to the global maximum is guaranteed. We then maximize $T_j^{(l)}(\boldsymbol{\psi}_j)$ with respect to $\boldsymbol{\psi}_j = (m_j, \theta_j)$ separately for $j = 1, ..., g$ to obtain an update $(\boldsymbol{m}^{(l)}, \boldsymbol{\theta}^{(l)})$:

$$m_j^{(l)} = \underset{m_j > 0}{\text{argmax}} \; T_j^{(l)}(m_j, \tilde{\theta}_j^{(l)}(m_j)); \qquad \theta_j^{(l)} = \tilde{\theta}_j^{(l)}(m_j^{(l)}), \tag{4.20}$$

where $\tilde{\theta}_j^{(l)}(m_j^{(l)})$ is given by

$$\begin{aligned} \tilde{\theta}_j^{(l)}(m_j) = {} & \frac{\lambda_j^{(2)}}{2} \Bigg\{ (\nu_j^{(2)} - 1) - m_j \sum_{i=1}^{n} \left(z_{ij}^{(l)} + k_i^{(l)} z_{ij}^{\prime(l)} \right) \\ & + \sqrt{\left(m_j \sum_{i=1}^{n} \left(z_{ij}^{(l)} + k_i^{(l)} z_{ij}^{\prime(l)} \right) - (\nu_j^{(2)} - 1) \right)^2 + \frac{4}{\lambda_j^{(2)}} \sum_{i=1}^{n} \left(z_{ij}^{(l)} \hat{y}_{ij}^{(l)} + k_i^{(l)} z_{ij}^{\prime(l)} \hat{y'}_{ij}^{(l)} \right)} \Bigg\}. \end{aligned} \tag{4.21}$$

Overall, the CM-step ensures that $Q(\boldsymbol{\Phi}^{(l)}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}) \geq Q(\boldsymbol{\Phi}^{(l-1)}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)})$, and hence the observed data posterior log-likelihood is guaranteed to converge to a local maximum. We terminate the ECM algorithm when the change of the

observed data posterior log-likelihood is smaller than a tolerance threshold of 10^{-4} or the maximum number of iterations of 500 is reached.

Remark 5. Involving numerical integration or numerical optimization, Equations (4.11), (4.15), and (4.20) are the relatively more computationally intensive substeps of the ECM iteration. The main challenge of computing Equations (4.11) and (4.15) is that a numerical integration is required for each observation $i = 1, ..., n$. Naive numerical integration may be computationally intensive if n is large, but we notice that for two different observations with the same censoring/truncation ranges, the corresponding $\widehat{\log y}_{ij}^{(l)}$ (or $\widehat{\log y'}_{ij}^{(l)}$) are identical. Therefore, performing a numerical integration for each unique range suffices and significantly reduces the computational cost if the number of unique ranges is much smaller than n, which usually holds in a general insurance context. For example, when modeling losses under deductibles (left truncation points), an insurance company usually offers only a limited choices of deductibles to customers. For Equation (4.20), the function to be optimized (i.e., $T_j^{(l)}$) seems complicated because n-term summations appear multiple times. Yet, these summations do not involve any parameters $\mathrm{m_j}$ or θ_j and hence they can be computed before that the function is optimized. As a result, the computational burden of Equation (4.20) is indeed minimal.

Remark 6. For efficient parameter estimations, our proposed hypothetical complete data approach (Eq. [4.2]) for censored and truncated data is useful not only to the LRMoE but also to finite mixture-based regression models in general including the FMR. We do not bog down the mathematical details for conciseness purpose, but we briefly discuss the logic as follows. After applying the proposed hypothetical complete data approach in the context of FMR, the resulting complete data posterior likelihood function would be the same as Equation (4.7), except that the parameters in the expert functions $(\boldsymbol{m}, \boldsymbol{\theta})$ would depend on the covariates $\boldsymbol{x}_i$ and some regression parameters. Then, the E-step can be computed using the same formulas presented above, and the CM-step can be computed by optimizing the complete data posterior log-likelihood with respect to the regression parameters in the expert functions instead of $(\boldsymbol{m}, \boldsymbol{\theta})$.

4.2.3. Initialization and Parameter Adjustments

We use a clusterized method of moments approach similar to Gui, Huang, and Lin (2018) and Fung, Badescu, and Lin (2020) to set the initial parameters $\boldsymbol{\Phi}^{(0)}$. This comes with the following steps:

1. Perform K-means clustering on $(y_i^l + y_i^u)/2$ (or just on y_i^l if $y_i^u = \infty$) with g clusters and obtain the clustering mean $\{\mu_j^{\text{cluster}}\}_{j=1,...,g}$, variance $\{(\sigma_j^{\text{cluster}})^2\}_{j=1,...,g}$, and weights $\{\pi_j^{\text{cluster}}\}_{j=1,...,g}$.
2. Set $m_j^{(0)} = (\mu_j^{\text{cluster}}/\sigma_j^{\text{cluster}})^2$ and $\theta_j^{(0)} = (\sigma_j^{\text{cluster}})^2/\mu_j^{\text{cluster}}$.
3. Set $\alpha_{j0}^{(0)} = \log(\pi_j^{\text{cluster}}/\pi_g^{\text{cluster}})$ and $\alpha_{jp}^{(0)} = 0$ for $p > 0$.

Following the above initialization steps, the information on the truncation range is disregarded and the true observed value is approximated by the mid-point between the two censoring points. Therefore, the moments of the initial model only very roughly match that of the data. However, the above initialization strategy is already found stable and robust for both simulated and real datasets presented in the preceding sections. Finally, the optimal number of subgroups (hyperparameter g) is chosen using some standard statistical criteria, such as the Akaike information criterion (AIC) and the Bayesian information criterion.

5. A SIMULATION STUDY

To examine the effectiveness of the proposed fitting algorithm in recovering the true model and to compare our proposed method to some existing methods that deal with censored and truncated data, this section presents a simulation study that is designed in the context of modeling insurance claim reporting delay.

Suppose that 20,000 claims occur to an insurance company within 5 years, where the choice of the number of samples is motivated by the typical portfolio size of insurance companies. We set the time unit as "day" so the (current) evaluation date is given by $\tau = 5 \times 365 = 1825$. Assume that the arrival date of claim $i \in \{1, ..., 20,000\}$ is $W_i := \lfloor \tilde{W}_i \rfloor$ where $\tilde{W}_i$ follows a pdf $f_{\tilde{W}_i}(w) = 2w/\tau^2 \cdot 1\{0 \le w \le \tau\}$. The density function represents that the business size or exposure of the insurance company is growing constantly over time. The reporting delay Y_i of claim i is generated from the Gamma-LRMoE with $g = 2$ components with pdf given by Equation (2.1). Corresponding to each claim, we also have $P = 2$ covariates that explain the policyholder/claim characteristics. The first covariate x_{i1} is a time-independent variable simulated from $N(0, 1)$. Meanwhile, the second covariate x_{i2} is a time-dependent variable simulated from $N(\tilde{W}_i/\tau, 1)$, meaning that policyholders' characteristics are

generally changing over time. In reality, reporting delay is usually observed as the number of days (instead of being observed in exact) and, further, each claim i is observed only if its reporting date $R_i = W_i + Y_i^u$ is no later than the evaluation date (i.e., $R_i \leq \tau$ or $0 \leq Y_i^u \leq \tau - W_i$). As a result, we have the following censoring and truncation mechanisms adopting the framework and notations from Section 3: For truncation mechanism $\boldsymbol{T}_i$, the lower and upper truncation points of claim i are respectively $T_i^l = 0$ and $T_i^u = \tau - W_i$. For censoring mechanism $\boldsymbol{C}_i$, we have $\mathcal{R}_i^C = (0, \tau - W_i] = (T_i^l, T_i^u]$ and $\mathcal{R}_i^U = \phi$ such that there is no uncensoring region. To reflect the one-day imprecision of measuring the reporting delay, we construct $S_i = \tau - W_i$ one-day intervals such that $I_{is} = (s-1, s]$ for $s = 1, ..., S_i$. Using Equation (3.1), the lower and upper censoring points that we are observing are $Y_i^l = \lfloor Y_i \rfloor$ and $Y_i^u = \lceil Y_i \rceil$. Overall in this simulation study, the observed data are given by $\mathcal{D}^{\text{obs}} = \{(y_i^l, y_i^u, t_i^l, t_i^u, \boldsymbol{c}_i)\}_{i; y_i^u \leq \tau - w_i}$, where $(y_i^l, y_i^u, t_i^l, t_i^u, w_i)$ are the realizations of $(Y_i^l, Y_i^u, T_i^l, T_i^u, W_i)$. We can then directly apply the proposed ECM algorithm in Section 4 for parameter estimations. Note that the number of claims observed can be significantly fewer than 20,000.

The remaining quantities we need to set are the parameters of the Gamma-LRMoE that simulate the reporting delay Y_i. Here we construct two cases: in the first one (Case A), we set the shape parameters $\boldsymbol{m} = (2, 1)$, scale parameters $\boldsymbol{\theta} = (2, 100)$, and regression parameters $\boldsymbol{\alpha}_1 = (2, 0.5, -0.5)^T$ so that the average reporting delay is around 21 days; in the second one (Case B), we adopt the same shape and regression parameters as Case A but the scale parameters are changed to $\boldsymbol{\theta} = (5, 250)$, so the average reporting delay is much longer (around 53 days). In both cases, the larger time-independent variable x_{i1} leads to a smaller average reporting delay, whereas the larger time-dependent variable x_{i2} leads to a greater average reporting delay.

Under each of the two cases described above, we fit the Gamma-LRMoE to the observed data using the proposed ECM algorithm for censored and truncated data. To ensure a thorough examination on the proposed algorithm, the whole simulation and fitting process is replicated 200 times. Using the AIC, the proposed algorithm identifies the correct number of components (i.e., $g = 2$) in 192 and 184 out of 200 replications respectively under Case A and Case B. Using the Bayesian information criterion, the algorithm even correctly detects the number of components in all replications under both cases. To examine whether the true model is recovered by the proposed algorithm, we plot the density function of each fitted parameter and compare it to

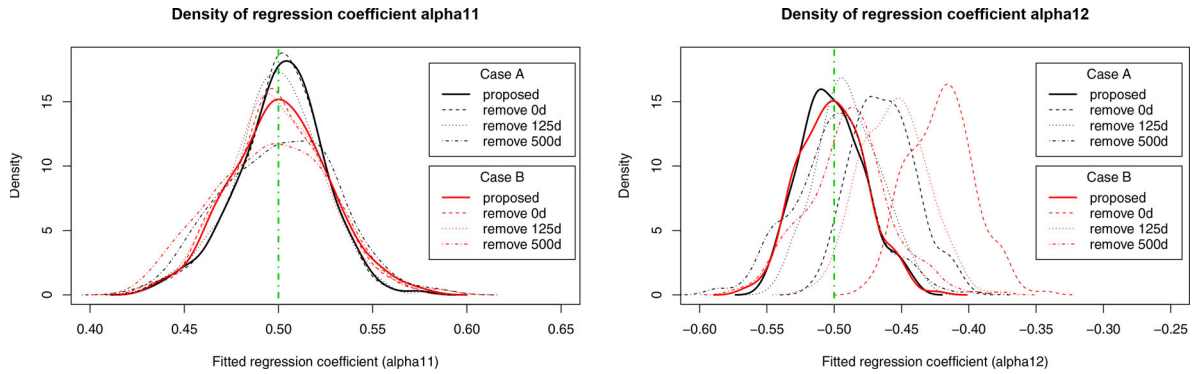

FIGURE 2. Density Plots of Fitted Regression Coefficients $\hat{\alpha}_{11}$ and $\hat{\alpha}_{12}$. *Note:* In the legends, "proposed" represents the use of our proposed ECM algorithm for censored and truncated data; "remove 0d," "remove 125d," and "remove 500d" represent removing claims observed with $t_i^u < t^*$ ($t^* = 0, 125, 500$) and using a standard ECM algorithm that ignores data truncation. The vertical dotted line is the true model parameter.

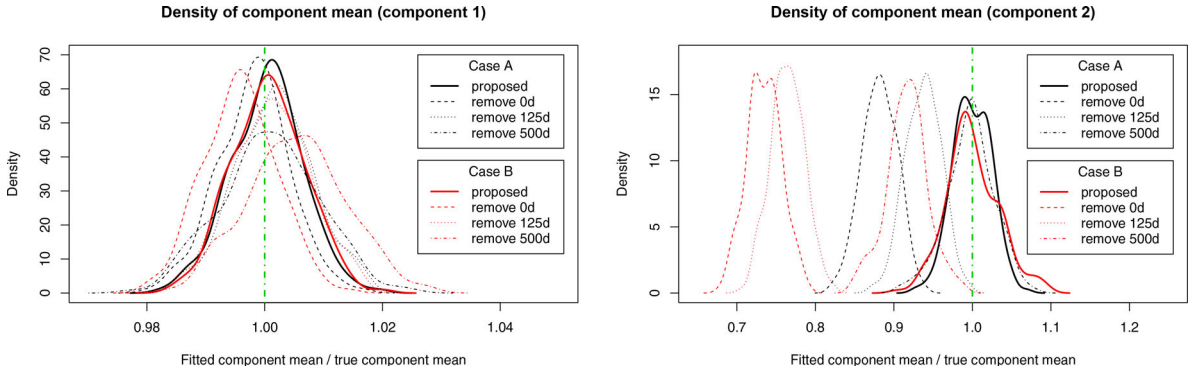

FIGURE 3. Density Plots of Fitted Component Mean Ratios $\hat{\mu}_j/\mu_j := \hat{m}_j\hat{\theta}_j/m_j\theta_j$ ($j = 1, 2$).

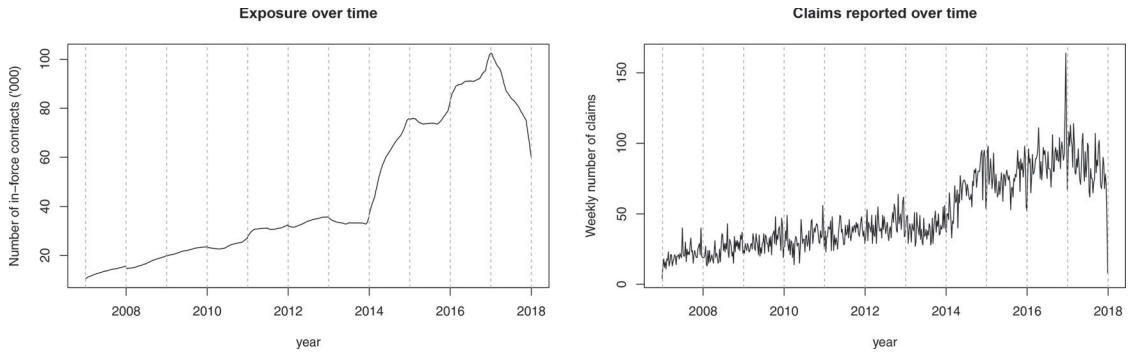

FIGURE 4. Exposure and Weekly Number of Claims Reported versus Time.

the true model parameter. For the sake of preciseness, we only present the density plots for the fitted regression coefficient $\hat{\alpha}_{1p}$ ($p = 1, 2$) in Figure 2 and the fitted component mean (ratio) defined by $\hat{\mu}_j/\mu_j := \hat{m}_j\hat{\theta}_j/m_j\theta_j$ ($j = 1, 2$) in Figure 3. As observed from these two figures, the peaks and medians of the thick, solid density curves (generated using our proposed ECM algorithm) are very close to the true values (vertical dotted lines), showcasing empirically the unbiasedness of estimated parameters and revealing the effectiveness of the proposed algorithm in recovering the true model.

To demonstrate the usefulness and importance of our proposed fitting algorithm, we compare it with some commonly used methods that fits truncated reporting delay data. The first method (named "remove 0d"), which is adopted by, for example, Badescu et al. (2019), simply fits the untruncated reporting delay distribution to the observed data. Without a proper fitting algorithm for truncated data, ignoring the truncation points makes model calibration computationally feasible, but the true reporting delay distribution will be underestimated. To mitigate such a bias, one can discard the data points with small upper truncation points (i.e., $t_i^u < t^*$ for some t^*). By selecting a large enough t^*, the biases induced from data truncation may be significantly reduced at the expense of having fewer data points for model fitting procedure; see remark 5.1 of Badescu et al. (2019) for a more detailed discussion. This results in the second and third methods (called "remove 125d" and "remove 500d"), where $t^* = 125$ (roughly the 95% quantile of reporting delay in Case A) and $t^* = 500$ (roughly the 97.5% quantile in Case B) are chosen, respectively.

For each method and each case, the simulation and fitting procedures are replicated by 200 times. The resulting distributions of the fitted parameters are displayed as dotted curves in Figures 2 and 3. From the right panels of Figures 2 and 3, the density curves of $\hat{\alpha}_{12}$ and $\hat{\mu}_2/\mu_2$ by the method "remove 0d" substantially deviate from the true values under both cases, revealing a significant bias of the fitted model if the data truncation issue is ignored. Such a bias is larger in Case B (where the average reporting delay is longer) than in Case A. Though the bias is reduced using methods "remove 125d" or "remove 500d," it cannot be completely removed in Case B even if 500 days of reporting delay data are discarded. From the left panels of Figures 2 and 3, we observe no substantial deviations on the density curves of $\hat{\alpha}_{11}$ and $\hat{\mu}_1/\mu_1$ from the corresponding true values under any methods. However, the distributions of these fitted parameters under method "remove 500d" are more dispersed (see the lower peaks) than those under other methods, implying that removal of reporting delay data would induce a significant increase in the uncertainties of fitted parameters.

Overall, without an efficient fitting algorithm catering for data truncation, there exists a bias–variance trade-off in fitting truncated reporting delay data. If the truncation issue is neglected, the fitted reporting delay distribution will be biased. Though the bias can be mitigated through discarding some data points, this would reduce the number of observations and thus result in greater parameter uncertainties. In contrast, using our proposed ECM algorithm, one does not need to concern about the bias–variance trade-off problem.

6. APPLICATION: MODELING REPORTING DELAY

In this section, we apply our proposed ECM algorithm to fit reporting delay data from a real insurance dataset. We will also demonstrate how the proposed algorithm facilitates convenient yet accurate prediction for the number of IBNR.

6.1. Data Overview

The dataset is from a major European insurer that was also analyzed by Fung, Badescu, and Lin (2020). It contains two files: "policies" and "claims." The policies file contains 594,908 third-party liability insurance contracts from January 1, 2007,

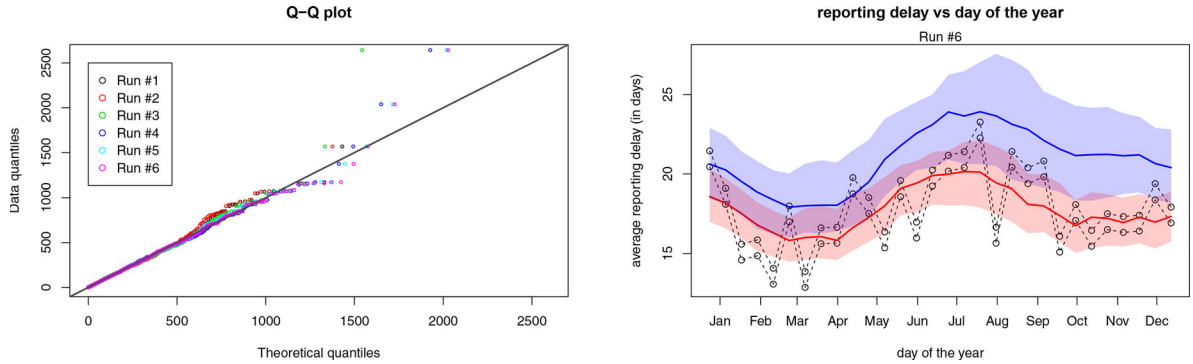

FIGURE 5. (a) – Q-Q Plots for Each Run. *Note:* (b) Average Reporting Delay versus Day of the Year (in Run 6) Using the Nonparametric Approach. *Note:* Two dotted curves represent the empirical patterns of upper and lower censoring points of reporting delay. Two solid curves represent the patterns generated by the fitted model, where the lower (red) one caters to data truncation but the upper (blue) one does not. The shaded areas are the 95% confidence intervals.

TABLE 1
Summary of the Covariates for the ith Claim, Where w_i Is the Accident Date and τ_0 = January 1, 2017

Variable	Description	Type	Notes
x_{i1}	Policyholder age	Discrete	
x_{i2}	Car age	Discrete	
x_{i3}	Car fuel	Categorical	Diesel: $x_{i3} = 1$ Gasoline: $x_{i3} = 0$
x_{i4}–x_{i7}	Geographical location	Categorical	Region I: $x_{i4} = 1$ Region II: $x_{i5} = 1$ Region III: $x_{i6} = 1$ Region IV: $x_{i7} = 1$ Capital: $x_{i4} = x_{i5} = x_{i6} = x_{i7} = 0$
x_{i8}–x_{i9}	Car brand class	Categorical	Class A: $x_{i8} = 1$ Class B: $x_{i9} = 1$ Class C: $x_{i8} = x_{i9} = 0$
x_{i10}	Contract type	Categorical	Renewal contract: $x_{i10} = 1$ New contract: $x_{i10} = 0$
x_{i11}–x_{i14}	Day of the year	Continuous	$x_{i11} = \sin(2\pi(w_i - \tau_0)/365.25)$ $x_{i12} = \cos(2\pi(w_i - \tau_0)/365.25)$ $x_{i13} = \sin(4\pi(w_i - \tau_0)/365.25)$ $x_{i14} = \cos(4\pi(w_i - \tau_0)/365.25)$

TABLE 2
Summary of Data Segmentation Process, Fitted Model and p Values of In-Sample Residual Tests under Six Different Runs

	Data segmentation				Fitted model	Residual test p values		
Run	IS training	OS test	τ	IS data no.	Component no.	Kolmogorov-Smirnov test	χ^2 test	AD test
1	2007–2012	2013–2017	1/1/2013	9,608	5	0.9728	0.4086	0.9300
2	2007–2013	2014–2017	1/1/2014	11,699	7	0.9504	0.3301	0.9245
3	2007–2014	2015–2017	1/1/2015	15,061	7	0.9771	0.5048	0.9620
4	2007–2015	2016–2017	1/1/2016	18,994	13	0.9662	0.7178	0.9770
5	2007–2016	2017–2017	1/1/2017	23,668	13	0.9588	0.8017	0.9571
6	2007–2017	Nil	1/1/2018	28,256	12	0.9440	0.6165	0.9528

to December 31, 2017. For each of the contracts we have the contract number, start date, end date, and several policyholders' information. The claims file contains $n^o = 28,256$ claims incurred and reported until December 31, 2017. For each claim, we have the corresponding contract number, accident date w_i, and reporting date r_i. The number of in-force contracts (total exposure) over time, as well as the weekly number of claims reported, are displayed in Figure 4.

The contract numbers provide a link between two files so that we can append the policyholder information in each claim. Then, in this application we mainly deal with the claims file because our aim is to model the truncated reporting delay data. The censoring and truncation mechanisms in this application are exactly the same as in the simulation study (Section 5), because the precision unit of the accident and reporting dates is still one day (interval censoring), and the data observed are already given that the reporting date is no later than the validation date. The lower and upper censoring points for each observed claim i are respectively given by $y_i^l = r_i - w_i$ and $y_i^u = y_i^l + 1$. The truncation points are given by $t_i^l = 0$ and $t_i^u = \tau - w_i$, where τ is the validation date, which will be described in the following paragraph. Similar to the simulation study, the observed data are given by $\{(y_i^l, y_i^u, t_i^l, t_i^u, \boldsymbol{c}_i\}_{i=1,\ldots,n^o}$, where we recall $\boldsymbol{c}_i$ as the censoring mechanism, which is irrelevant to the log-likelihood function (see Subsection 4.1). The summary of the covariates $\boldsymbol{x}_i$ is presented in Table 1. Note that apart from the policyholder attributes $x_{i1} - x_{i10}$, we added four linearly independent sinusoidal functions $x_{i11} - x_{i14}$ to flexibly capture the effect of the the accident's day of the year to its reporting delay, which will be found to be important in the right panel of Figure 5. With all variables properly defined, the proposed ECM algorithm is then directly applied to estimate the parameters.

To examine the predictive power of the proposed modeling framework, which will be discussed in Subsection 6.3, we segment the claim dataset into two parts: an in-sample (IS) training set containing all claims reported before the validation date τ and an out-of-sample (OS) test set containing the remaining claims that are not yet observed until τ but are reported before January 1, 2018. To make our analysis comprehensive, we perform model fitting and out-of-sample testing under six different validation dates summarized in the "Data segmentation" column of Table 2. For example, in run 3, 15,061 claims reported before the validation date $\tau = 1/1/2015$ are fitted to our proposed model and the resulting IBNR prediction as at January 1, 2015, is compared to the realized IBNR from the test set claims, which are reported between January 2, 2015, and December 31, 2017.

6.2. Estimation Results and In-Sample Validation Tests

In each run, we fit the Gamma-LRMoE to the reporting delay training set using the proposed ECM algorithm for censored and truncated data. Using the AIC, the resulting fitted models contain 5 to 13 components, displayed in Table 2. In general, a larger number of observed data points results in more subgroups for the optimal fitted model.

After model fitting, it is important to assess the goodness-of-fit. We first construct a Q-Q plot as follows: Using the parameters from the fitted model, simulate the reporting delay of each claim i given that it is below the upper truncation point (i.e., simulate $\{Y_i|\boldsymbol{x}_i, Y_i \le t_i^u, \hat{\boldsymbol{\Phi}}\}_{i=1,\ldots,n^o}$, where $\hat{\boldsymbol{\Phi}}$ represents the fitted Gamma-LRMoE parameters) from the truncated density function $h(y_i;\boldsymbol{x}_i,\hat{\boldsymbol{\Phi}})/H(t_i^u;\boldsymbol{x}_i,\hat{\boldsymbol{\Phi}})$. Denote $\hat{y}_i$ as the realization from the simulation and $\hat{y}_i^l := \lfloor \hat{y}_i \rfloor$ the lower censoring point of the simulated reporting delay. Then, compare the quantiles of $\{\hat{y}_i^l\}_{i=1,\ldots,n^o}$ to that of the empirical left censoring points $\{y_i^l\}_{i=1,\ldots,n^o}$. The resulting Q-Q plot for each run is displayed in the left panel of Figure 5, showing that the Gamma-LRMoE fits the reporting delay data very well in all runs.

Apart from the Q-Q plot, we evaluate the goodness of fit through a residual test that tests the null hypothesis (H_0) that the reporting delay data are generated from the fitted model against the alternative hypothesis (H_1) that H_0 is false. To do so, we first compute the fitted cdfs of truncated data $\hat{H}_i^l := H(y_i^l;\boldsymbol{x}_i,\hat{\boldsymbol{\Phi}})/H(t_i^u;\boldsymbol{x}_i,\hat{\boldsymbol{\Phi}})$ and $\hat{H}_i^u := H(y_i^u;\boldsymbol{x}_i,\hat{\boldsymbol{\Phi}})/H(t_i^u;\boldsymbol{x}_i,\hat{\boldsymbol{\Phi}})$ for each claim i. We then simulate $\hat{H}_i \sim U[\hat{H}_i^l, \hat{H}_i^u]$. If H_0 is true, then $\hat{H}_i$ will follow $U[0,1]$ for every $i = 1,\ldots,n^o$. Finally, three common goodness-of-fit statistics (from the Kolmogorov-Smirnov test, the chi-square test [using 200 equiprobable intervals] and the Anderson-Darling test) are computed to examine the difference between the empirical distribution of the simulated data $\{\hat{H}_i\}_{i=1,\ldots,n^o}$ and the $U[0,1]$ distribution. The resulting p values are displayed in Table 2. All tests are passed with significant margins in all runs, so H_0 cannot be rejected and our model fitting performances are robust.

Using the fitted model, we can also construct a visualization tool to understand how the covariates impact the reporting delay of a claim. The methodology is similar to Fung, Badescu, and Lin (2019a; non-parametric approach) and is described as follows. We define $\mathcal{X}$ as the covariates space and partition it into Q disjoint subspaces $\mathcal{X}_1,\ldots,\mathcal{X}_Q$ such that $\mathcal{X}_1 \cup \cdots \cup \mathcal{X}_Q = \mathcal{X}$. The average reporting delay of a claim given that its covariates $\boldsymbol{x}_i$ belong to $\mathcal{X}_q$ can be estimated by

$$\bar{Y}^{\text{npar}}(q) = \frac{1}{N(q)} \sum_{i \le n^o : \boldsymbol{x}_i \in \mathcal{X}_q} E\big[Y_i|\boldsymbol{x}_i,\hat{\boldsymbol{\Phi}}\big] = \frac{1}{N(q)} \sum_{i \le n^o : \boldsymbol{x}_i \in \mathcal{X}_q} \sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}}) \hat{m}_j \hat{\theta}_j, \tag{6.1}$$

where $N(q) = \sum_{i=1}^{n} 1\{\boldsymbol{x}_i \in \mathcal{X}_q\}$ is the number of claims with covariates $\boldsymbol{x}_i \in \mathcal{X}_q$. By plotting $\bar{Y}^{\text{npar}}(q)$ across q, we can examine the influence of the covariates to the reporting delay. The right panel of Figure 5 demonstrates an example of the visualization plot, where the average reporting delay is plotted against the day of the year of the claim's accident date. In this example, we choose $Q = 30$ and construct $\mathcal{X}_q = \{\boldsymbol{x} : (q-1.5)/29 \le (\text{day of the year})/365.25 \le (q-0.5)/29\}$ (note that the day of the year of each claim i can be retrieved by the covariates x_{i11}–x_{i14}). The resulting $\bar{Y}^{\text{npar}}(q) \sim q$ plot is represented by the upper (blue) solid curve. One may observe that this curve is consistently higher than the average reporting delay from the empirical data (two dotted curves). This is because the empirical (observed) reporting delay data points are right truncated, but Equation (6.1) is evaluated based on untruncated distributions. To make the solid curve directly comparable to the empirical data, we also calculate the average reporting delay of a claim given that it is reported on or before τ with covariates $\boldsymbol{x}_i$ belong to $\mathcal{X}_q$:

$$\bar{Y}^{\text{npar},t}(q) = \frac{1}{N(q)} \sum_{i \le n^o : \boldsymbol{x}_i \in \mathcal{X}_q} E\left[Y_i | Y_i \le t_i^u, \boldsymbol{x}_i, \hat{\boldsymbol{\Phi}}\right] = \frac{1}{N(q)} \sum_{i \le n^o : \boldsymbol{x}_i \in \mathcal{X}_q} \frac{\sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) F(t_i^u; \hat{m}_j + 1, \hat{\theta}_j) \hat{m}_j \hat{\theta}_j}{H(t_i^u; \boldsymbol{x}_i, \hat{\boldsymbol{\Phi}})}. \tag{6.2}$$

Plotting $\bar{Y}^{\text{npar},t}(q)$ against q results in the lower (red) solid curve in the right panel of Figure 5. This curve aligns well with the empirical average reporting delay. Both solid curves suggest that June–August (February–April) are the months where claims are expected to experience longer (shorter) reporting delays. Apart from the day-of-the-year effect, a similar visualization tool can be deployed to analyze the impact of other variables. Though we refrain from showing them one by one for conciseness, several variables that significantly affect the reporting delay are summarized as follows:

- Claims from younger drivers ($\le$40 years old) experience longer reporting delays (by 2.5 days on average) than those from older drivers (>40 years).
- The average reporting delay of claims from the capital region (24.7 days) is much longer than that from other regions (19.1 days).
- Claims from car class A (the best class) have shorter reporting delays than those from other car classes (17.4 versus 22.2 days on average).

6.3. IBNR Prediction and Out-of-Sample Tests

In the microlevel reserving framework, a topic that received plenty of interest in recent years, prediction of the number of IBNR requires modeling both claim frequencies and reporting delays either separately (Antonio and Plat 2014; Badescu et al. 2019; Crevecoeur, Antonio, and Verbelen 2019) or jointly (Verbelen et al. 2018). Therefore, the claim arrival process and its regression link need to be specified before the IBNR prediction can be obtained. In this section, we propose an alternative semiparametric approach for the number of IBNR prediction, where the fitted reporting delay distribution itself suffices to produce an adequate IBNR count prediction without the need to fit a claim frequency distribution. This is a very simple microlevel procedure that produces very accurate estimates and can potentially be applied by practitioners with minimal computational costs. A paper that will focus solely on the estimation of the IBNR and the reported but not settled reserve using the class of LRMoE and the comparisons with macro- and micro- level methods including covariates is part of our current research objectives, as described in more details in the last section.

For simplicity, two assumptions are made as follows. Note that relaxation of the assumptions may be possible, but it is beyond the scope of this article, which mainly focuses on the calibration of censored and truncated data.

- The claim arrival process of each contract i follows an inhomogeneous Poisson process with intensity $\lambda(\tilde{\boldsymbol{x}}_i, t)1\{t \in \mathcal{T}_i\}$, where $\mathcal{T}_i$ is the contract period (from contract start date to end date) and $\tilde{\boldsymbol{x}}_i$ is the policyholder information (including only x_{i1}–x_{i10} in Table 1).
- The claim arrival process is independent of the reporting delay distribution $P(\cdot; \boldsymbol{x}_i)$.

Suppose that there are a total of N^* contracts. Denote N, N^o, and N^{IBNR} as the total number of claims that occurred, observed (reported) claims, and IBNR at time τ. The standard thinning property of the Poisson process results in the following:

$$N \sim \text{Poi}\left(\sum_{i=1}^{N^*} \int_{\tau_0}^{\tau} \lambda(\tilde{\boldsymbol{x}}_i, t)1\{t \in \mathcal{T}_i\} dt\right) := \text{Poi}(\bar{\lambda}), \tag{6.3}$$

$$N^o \sim \text{Poi}\left(\sum_{i=1}^{N^*}\int_{\tau_0}^{\tau}\lambda(\tilde{\boldsymbol{x}}_i,t)P(\tau\ -\ t;\boldsymbol{x}_i)1\{t\in\mathcal{T}_i\}dt\right) := \text{Poi}(\bar{\lambda}^o), \tag{6.4}$$

$$N^{\text{IBNR}} \sim \text{Poi}\left(\sum_{i=1}^{N^*}\int_{\tau_0}^{\tau}\lambda(\tilde{\boldsymbol{x}}_i,t)(1\ -\ P(\tau\ -\ t;\boldsymbol{x}_i))1\{t\in\mathcal{T}_i\}dt\right) := \text{Poi}(\bar{\lambda}^{\text{IBNR}}), \tag{6.5}$$

where $\tau_0 =$ January 1, 2007, is the business start date. To predict the number of IBNR, it suffices to find a reasonable predictor of $\bar{\lambda}^{\text{IBNR}}$, which results in the following proposition.

Proposition 1. *Suppose that the IBNR process is given by Equation (6.5). Then*

$$\hat{k} := \sum_{i=1}^{N^o}\hat{k}_i := \sum_{i=1}^{N^o}\frac{1-P(\tau-W_i;\boldsymbol{x}_i)}{P(\tau\ -\ W_i;\boldsymbol{x}_i)} \tag{6.6}$$

is an unbiased estimator of $\bar{\lambda}^{\text{IBNR}}$, where W_i is the accident (occurrence) time of the ith claim.

Proof. We take an expectation on $\hat{k}$.

$$\begin{aligned}
E\left[\sum_{i=1}^{N^o}\frac{1-P(\tau-W_i;\boldsymbol{x}_i)}{P(\tau\ -\ W_i;\boldsymbol{x}_i)}\right] &= E[N^o]E\left[\frac{1-P(\tau-W;\boldsymbol{x})}{P(\tau\ -\ W;\boldsymbol{x})}\right] \\
&= \bar{\lambda}^o \times \int_{\tau_0}^{\tau}\sum_{i=1}^{N^*}\frac{1-P(\tau-t;\boldsymbol{x}_i)}{P(\tau\ -\ t;\boldsymbol{x}_i)} \times \frac{\lambda(\tilde{\boldsymbol{x}}_i,t)1\{t\in\mathcal{T}_i\}P(\tau-t;\boldsymbol{x}_i)}{\bar{\lambda}^o}dt \\
&= \sum_{i=1}^{N^*}\int_{\tau_0}^{\tau}\lambda(\tilde{\boldsymbol{x}}_i,t)1\{t\in\mathcal{T}_i\}(1\ -\ P(\tau\ -\ t;\boldsymbol{x}_i))dt = \bar{\lambda}^{\text{IBNR}}.
\end{aligned}$$

□

Now, assume that the fitted reporting delay distribution $H(\cdot;\boldsymbol{x}_i,\hat{\boldsymbol{\Phi}})$ synchronizes well with the true reporting delay distribution $P(\cdot;\boldsymbol{x}_i)$. Then $\hat{k}_i$ in Equation (6.6) can be accurately estimated by

$$\hat{k}_i = \frac{1-P(\tau-W_i;\boldsymbol{x}_i)}{P(\tau\ -\ W_i;\boldsymbol{x}_i)} \approx \frac{1-H(t_i^u;\boldsymbol{x}_i,\hat{\boldsymbol{\Phi}})}{H(t_i^u;\boldsymbol{x}_i,\hat{\boldsymbol{\Phi}})}, \tag{6.7}$$

which is equivalent to Equation (4.12) involved in the E-step of the proposed fitting algorithm. Note that the above assumption is reasonable under the class of LRMoE as reporting delay distribution, because of its denseness property (Fung, Badescu, and Lin 2019b), which guarantees its flexibility to synchronize well any true distributions. The semi-parametric IBNR prediction is now given by

$$\hat{k} = \sum_{i=1}^{n^o}\frac{1-H(t_i^u;\boldsymbol{x}_i,\hat{\boldsymbol{\Phi}})}{H(t_i^u;\boldsymbol{x}_i,\hat{\boldsymbol{\Phi}})}. \tag{6.8}$$

Overall, under our proposed semi parametric approach, the parameter calibration procedure automatically produces the IBNR estimate. Also, this approach makes no assumptions on the impact of covariates and time to the claim arrival intensity (regression link), enabling convenient yet realistic prediction of the IBNR.

To obtain a full predictive distribution of the IBNR, it is also important to take parameter uncertainties into account. To do so, we apply a bootstrapping technique as follows for each run b: Firstly, simulate the number of observed claims $N^{o(b)} \sim \text{Poi}(\bar{\lambda}^o)$. Then, for $i = 1,...,N^{o(b)}$, bootstrap $\{(y_i^{l(b)}, y_i^{u(b)}, t_i^{l(b)}, t_i^{u(b)}, \boldsymbol{x}_i^{(b)})\}$ iid with replacement from the observed claim dataset. After that, we re-fit the proposed model to the re sampled data using the proposed ECM algorithm, with the fitted model parameters $\hat{\boldsymbol{\Phi}}$ being the initialization. The re-fitted model parameters $\hat{\boldsymbol{\Phi}}^{(b)}$ is finally obtained. The whole process above is repeated for $b = 1,...,B$, where we choose $B = 200$.

TABLE 3
Summary of the OS Predictions on the Number of IBNR

	IBNR prediction period		IBNR prediction results					
Run	Start date	End date	Total IBNR	Truncated IBNR	Lower 95% CI	Upper 95% CI	Realized	p value
1	1/1/2013	1/1/2018	133.2958	133.0842	106	165	144	0.4606
2	1/1/2014	1/1/2018	139.7749	139.5201	107	175	120	0.2775
3	1/1/2015	1/1/2018	195.0955	193.6148	159	233	212	0.3400
4	1/1/2016	1/1/2018	213.0416	200.1773	167	236	223	0.2067
5	1/1/2017	1/1/2018	268.2051	222.5120	189	259	217	0.7982
6	1/1/2018	Infinite	230.4764	230.4764	194	271	Nil	Nil

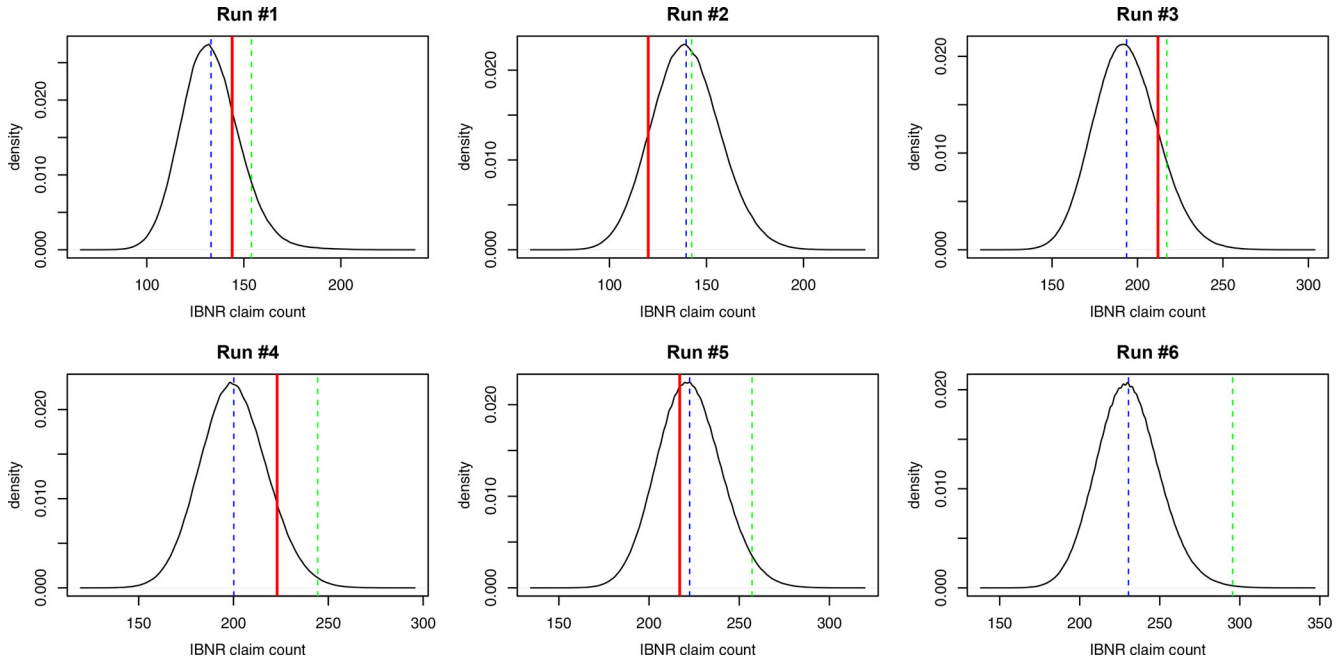

FIGURE 6. The Predictive Distributions of the Truncated IBNR. *Note:* Red vertical line: Realized value from the test set. Blue dotted lines: Predictions from the proposed method. Green dotted lines: Predictions from chain ladder approach.

Now, the IBNR predictive distribution that caters to parameter uncertainties can be obtained through simulations. We compute $\hat{k}^{(b)} = \sum_{i=1}^{n^{o(b)}} (1 - H(t_i^{u(b)}; \boldsymbol{x}_i^{(b)}, \hat{\boldsymbol{\Phi}}^{(b)}))/H(t_i^{u(b)}; \boldsymbol{x}_i^{(b)}, \hat{\boldsymbol{\Phi}}^{(b)})$ and generate 10,000 random samples from $\text{Poi}(\hat{k}^{(b)})$ for each $b = 1, \ldots B$. This will generate 2 million simulated points in total.

The out-of-sample IBNR predictions for each of the six runs are summarized in Table 3 and Figure 6. In Table 3, "Total IBNR" represents the predicted IBNR computed by Equation (6.8). However, the total IBNR prediction cannot be directly compared to the realized value from the test set because our dataset only contains information up to December 31, 2017. Any claims reported after December 31, 2017, are not observed. To properly examine the predictive power of our proposed framework, we also compute the estimates of "Truncated IBNR" that considers only the claims that occurred before τ, reported on or after the "Start date" (τ_{start}) and reported before the "End date" (τ_{end}). Following arguments similar to Proposition 1, the truncated IBNR estimation is given by $\hat{k}^t(\tau_{\text{end}}) = \sum_{i=1}^{n^o} (H(t_i^u + \tau_{\text{end}} - \tau; \boldsymbol{x}_i, \hat{\boldsymbol{\Phi}}) - H(t_i^u; \boldsymbol{x}_i, \hat{\boldsymbol{\Phi}}))/H(t_i^u; \boldsymbol{x}_i, \hat{\boldsymbol{\Phi}})$. The lower and upper 95% confidence intervals (CIs) in Table 3 are based on the truncated IBNR. They are computed using the afore-mentioned bootstrap technique. Because the realized (truncated) IBNR falls inside the 95% CI of the truncated IBNR predictive distribution for all runs 1–5, we conclude that our proposed modeling framework provides adequate predictions on the IBNR.

On the other hand, Figure 6 shows that the truncated IBNRs determined under the traditional distribution-free chain ladder (CL) approach overestimate the true value in all five runs, suggesting a potential biasedness of the CL approach in estimating the IBNR under the dataset we used.

Careful readers may further notice in Table 3 that the total IBNR prediction in run 6 decreases suddenly, as opposed to an increasing trend from run 1 to run 5. To see why this prediction looks abnormal, we compare it to the predicted IBNR using the CL approach under run 6. As also displayed in Figure 6, the total IBNR prediction using our proposed framework (230.5) is 22% fewer than that using the CL approach (295.5). Such a large discrepancy can be explained qualitatively as follows. Controlling for the average business exposure over the previous year (i.e., year 2017 in run 6), an increasing trend in the exposure during the year will lead to greater expected IBNR than a decreasing trend. This is because in the former case a larger proportion of claims arrive toward the end of the year (close to the validation date τ), which are less likely to be reported before τ. On the other hand, though the CL approach captures the change of the yearly average exposures across years, it does not cater to any within-year exposure change effects because the claim development data are aggregated by each calendar year. From the left panel of Figure 4, the business exposure decreases drastically during the year 2017, compared with a prolonged increasing trend from year 2007 to 2016. This explains why our proposed framework, which uses granular claim data, produces a significantly smaller IBNR estimate than the CL approach in run 6. Though the OS data are not available under run 6, we expect that the CL approach will continue to overestimate the IBNR because it fails to capture the micro-effects of an IBNR process.

7. APPLICATION: DEDUCTIBLE RATEMAKING

Apart from reporting delay, insurance loss data involve censoring and truncation if insurance contracts are subject to deductibles and policy limits. In this section we investigate the usefulness of the proposed ECM algorithm in the context of deductible ratemaking.

7.1. Data Overview

In this study, we analyze $n = 10,032$ car damages claim that occurred during 2016. For each claim $i = 1, ..., n,$ we have the policyholder information the same as x_{i1}–x_{i10} described in Table 1. We also have the claim amount y_i^*, deductible d_i, and policy limit u_i corresponding to each claim. There are six choices of deductibles: 0, 100, 200, 300, 500, and 1000 euros, where over 25% of claims have nonzero deductibles. The policy limits u_i are continuously distributed, ranging from 900 to 183,610 euros with an average of 12,126 euros. The claim amounts y_i^* range widely from 0.9 to 45,449 euros, with mean and median of 837.3 and 443.8 euros, respectively. Also, note that the observed loss amount of claim i is given by $\tilde{y}_i := y_i^* + d_i$ and it is right censored at u_i. Therefore, if the observed loss amount is exactly u_i, the actual loss amount of claim i will not be fully observed (i.e., can be greater than the observed one). Only 6 out of 10,032 observations hit the policy limit.

To understand the body and tail behavior of claim severities, a preliminary analysis is performed by fitting Gamma, Log-normal, and Pareto (Lomax) distributions to $\{y_i^*\}_{i=1,...,n}$ without considering covariates. To assess the goodness of fit, three normalized residuals Q-Q plots are displayed in Figure 7. It is obvious that the Gamma distribution undercaptures the heaviness of the right tail implied by the data, whereas the Pareto distribution misfits the body of the data. The Log-normal distribution provides a better fit, but all three goodness-of-fit statistics (Kolmogorov-Smirnov test, χ^2 test and Anderson-Darling (AD)

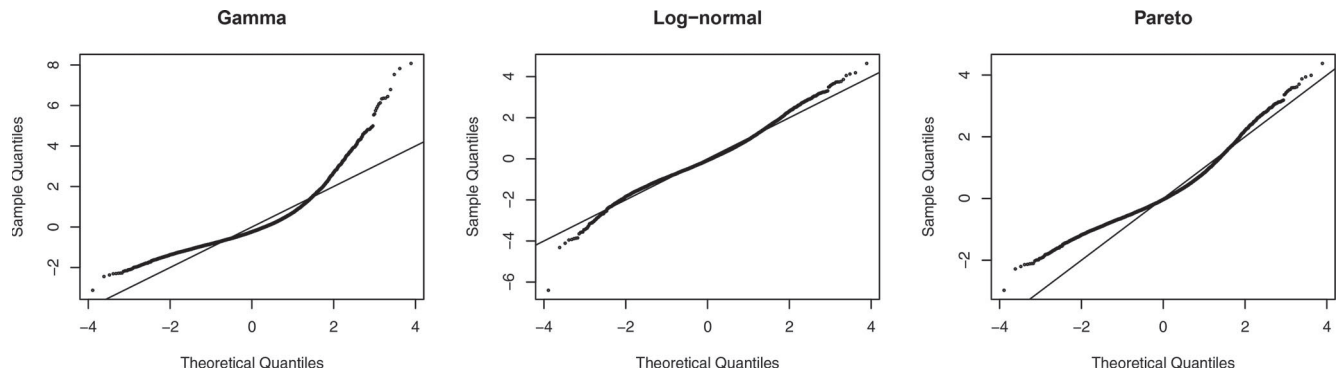

FIGURE 7. The Q-Q Plot of the Normalized Residuals Based on Three Preliminary Distributions Fitted to the Claim Amounts y_i^*.

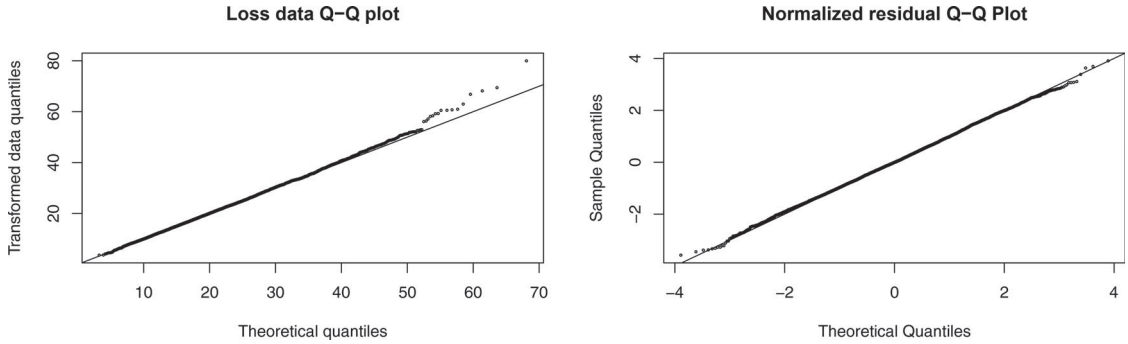

FIGURE 8. Loss Data and Normalized Residual Q-Q Plots (Deductible Excluded as Covariate).

test) report extremely small p values ($< 10^{-7}$), so the fitting performance can still be substantially improved through a flexible modeling framework.

Our main goal is to model loss severities subject to deductibles and policy limits. The censoring and truncation mechanisms are as follows. The (observed realizations of) the truncation points are apparently $t_i^l = d_i$ and $t_i^u = \infty$, showcasing that actual loss amounts of smaller than the deductible level d_i are not observed. For the censoring mechanism, we have $S_i = 1$, $\mathcal{R}_i^U = (d_i, u_i]$ and $\mathcal{R}_i^C = I_{i1} = (u_i, \infty]$, reflecting right censoring of loss amounts. Using Equation (3.1), the (observed realizations of) censoring points are $y_i^l = y_i^u = \tilde{y}_i$ if $\tilde{y}_i < u_i$ and $(y_i^l, y_i^u) = (\tilde{y}_i, \infty)$ if $\tilde{y}_i = u_i$. Because the preliminary analysis reveals that the Gamma distribution under-captures the tail-heaviness of the empirical loss severity, fitting the proposed Gamma-LRMoE directly to the dataset may not help extrapolate the tail-heaviness effectively. Instead, we follow the transformation technique proposed by Fung, Badescu, and Lin (2020), which transform the heavy-tailed data into lighter-tailed data first before fitting the Gamma-LRMoE. Define a Box-Cox transformation function

$$\xi_\gamma(y) = \begin{cases} \dfrac{(1+y)^\gamma - 1}{\gamma}, & \gamma > 0 \\ \log(1+y), & \gamma = 0 \end{cases}, \tag{7.1}$$

where γ is the transformation hyperparameter. If $\gamma < 1$, applying ξ_γ to the dataset will reduce its tail-heaviness and vice versa. We transform the dataset $(\tilde{y}_i^l, \tilde{y}_i^u, \tilde{t}_i^l, \tilde{t}_i^u) = (\xi_\gamma(y_i^l), \xi_\gamma(y_i^u), \xi_\gamma(t_i^l), \xi_\gamma(t_i^u))$ and now the input of the ECM algorithm (observed data) is given by $\mathcal{D}^{\text{obs}} = \{(\tilde{y}_i^l, \tilde{y}_i^u, \tilde{t}_i^l, \tilde{t}_i^u)\}_{i=1,\ldots,n}$.

7.2. Estimation Results

We fit the (transformed) observed data via the proposed calibration algorithm using the policyholder information x_{i1}–x_{i10} in Table 1 as well as the (transformed) policy limit $x_{i11} := \log(1+u_i)$ as the covariates $\boldsymbol{x}_i$. For each number of subgroups g, we try various values of transformation parameter $\gamma \in \{0, 0.1, 0.2, ..., 1\}$ and choose the γ that maximizes the observed data log-likelihood. Then, the optimal g is determined based on the AIC. Note that because the observed data are transformed, the likelihood function obtained by Equation (4.1) is distorted. In order to make a fair comparison of the log-likelihood among different values of γ, we should use the untransformed observed data likelihood $\mathcal{L}^{\text{obs}*}$ obtained by simple probabilistic arguments:

$$\log \mathcal{L}^{\text{obs}*}(\boldsymbol{\Phi}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \gamma) = \log \mathcal{L}^{\text{obs}}(\boldsymbol{\Phi}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \gamma) + (\gamma - 1) \sum_{i=1}^{n} \log(1+y_i) 1\{y_i < u_i\}, \tag{7.2}$$

where $\mathcal{L}^{\text{obs}}(\boldsymbol{\Phi}; \mathcal{D}^{\text{obs}}, \boldsymbol{x}, \gamma)$ is simply computed by Equation (4.1).

The optimal fitted model contains seven components with $\gamma = 0.3$. Two goodness-of-fit tests very similar to those presented in Subsection 6.2 are performed to assess the in-sample performance of the fitted model. The first one, the loss data Q-Q plot, simulates the lower censoring points $\{\hat{\tilde{y}}_i^l\}_{i=1,\ldots,n}$ from the fitted model and compares them to the empirical left censoring points $\{\tilde{y}_i^l\}_{i=1,\ldots,n}$. The resulting Q-Q plot displayed in the left panel of Figure 8 demonstrates an excellent fit. The second test, the residual test, computes the fitted truncated cdfs $\hat{H}_i^l := H(\tilde{y}_i^l; \boldsymbol{x}_i, \hat{\boldsymbol{\Phi}}) / (1 - H(\tilde{t}_i^l; \boldsymbol{x}_i, \hat{\boldsymbol{\Phi}}))$, and sets $\hat{H}_i = \hat{H}_i^l$ if $y_i < u_i$ or simulates $\hat{H}_i \sim U[\hat{H}_i^l, 1]$ if $y_i = u_i$. The three goodness-of-fit statistics, which compare $\{\hat{H}_i\}_{i=1,\ldots,n}$ to $U[0,1]$, report p values all greater than the 5% significance threshold (0.2505, 0.1921, and 0.1033, respectively). We also display the normal Q-Q plot of the

normalized residuals $\{\mathcal{N}^{-1}(\hat{H}_i)\}_{i=1,\ldots,n}$ ($\mathcal{N}^{-1}(\cdot)$ is the inverse normal cdf) in the right panel of Figure 8. The fitting is greatly improved compared to traditional parametric models (Fig. 7).

Using the same visualization technique presented by Equation (6.1) and the right panel of Figure 5 (not fully displayed in this section for conciseness), it is also possible to identify several risk factors that are related to larger loss severities. In summary, the model reveals that younger drivers, car age of about 5 years, diesel cars, driving in Region I or III, car class C, new contract, and larger policy limit are the higher risk drivers.

In the above analysis, the deductible is treated only as a truncation point. Many actuarial studies (see, e.g., Lee 2017; Lee and Shi 2019), in contrast, consider the deductible as an explanatory variable under a regression framework, providing a convenient way to model insurance losses and price insurance contracts in the presence of deductibles. Therefore, here we include (transformed) deductible $x_{i12} := \log(1+d_i)$ as a covariate, refit the proposed model, and contrast the result to that when the deductible is excluded as a covariate. We will show that through (from a modeling perspective) the deductible level has some predictive power to the claim severity (from a ratemaking perspective), inclusion of the deductible as a covariate would lead to unreasonably high premiums charged to policyholders choosing high deductible levels. Hence, this may suggest that deductible levels should be treated as a truncation point instead of an explanatory variable.

The optimal new refitted model contains nine components with $\gamma = 0.3$. Though the standard visualization tools show that the influence of covariates x_{i1}–x_{i11} produced by the new model (deductible included as covariate) is very similar to that produced by the original model (deductible excluded as a covariate), the overall fitting performance of the new model is significantly better than the original model (AIC improved from 150,899 to 150,480), revealing an intrinsic relationship between the choice of deductible and the loss distribution that cannot be fully explained by the truncation effect.

To quantify the effect of deductible choice to the expected loss after controlling for other variables, we calculate the expected loss severity across each deductible level using the partial dependence approach (a similar approach also studied by eq. [6.4] of Fung, Badescu, and Lin 2019a). The expected severity per accident/loss with deductible $d \in \{0; 100; 200; 300; 500; 1000\}$ is

$$\bar{Y}^{\text{prtl}}(d) = \frac{1}{n}\sum_{i=1}^{n} E\left[Y_i|\boldsymbol{x}_i^*(d), \hat{\boldsymbol{\Phi}}\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i^*(d);\hat{\boldsymbol{\alpha}}) \int_0^{\infty} \xi_\gamma^{-1}(y) f(y;\hat{\boldsymbol{\psi}}_j)dy, \tag{7.3}$$

where $\boldsymbol{x}_i^*(d) = (x_{i0}, x_{i1}, \ldots, x_{i11}, \log(1+d))^T$ and ξ_γ^{-1} is the inverse function of ξ_γ. The result computed in Table 4 shows that policyholders choosing a higher deductible contract are more prone to more severe losses. This is theoretically unreasonable and counter-intuitive because one should not expect that an alteration of the deductible level itself would directly change a policyholder's driving behavior and risk characteristics. In other words, the deductible level and loss severity are linked by confounding variables and their relationship is not causal. A possible interpretation of the result is that dangerous drivers may have a preference for in choosing high-deductible contracts, and at the same time they are more likely to incur severe losses. This interpretation is related to deductible selection behavior, which was studied by, for example, Sydnor (2010). Though such a behavioral insurance problem is beyond the scope of this paper, practicing actuaries should be especially vigilant to interpret the true relationship between deductible level and loss severities very carefully under a regression framework.

7.3. Deductible Ratemaking

A natural application of censored and truncated regression models is deductible ratemaking. In this subsection, we illustrate how policyholder information, deductible level, and policy limit affect several quantities (e.g., distribution and moment) related to the loss severity random variable (of the lth loss) Y_{il}, claim amount of a loss $Y_{il}^p(d_i) := (Y_{il} - d_i)_+ \wedge (u_i - d_i)$, aggregated claim amount $S_i^p(d_i) := \sum_{l=1}^{N_i} Y_{il}^p(d_i)$, and claim frequency $N_i^*(d_i) = \sum_{l=1}^{N_i} 1\{Y_{il} > d_i\}$, where N_i is the number of losses/accidents that occurred. Such quantities serve as a basis to calculate premiums for a policyholder under various deductible choices. For illustrative purposes, three hypothetical risk profiles, namely, "Good," "Average," and "Bad," are constructed in

TABLE 4
Expected Severity per Loss versus Deductible Choice under Partial Dependence Approach
(Deductible Included as Covariate)

	Deductible					
	0	100	200	300	500	1000
Expected severity per loss	750.08	1126.63	1185.69	1220.86	1265.84	1328.05

TABLE 5
Three Different Hypothetical Risk Profiles to Be Considered

Profile	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	x_{i7}	x_{i8}	x_{i9}	x_{i10}
Good	60	10	0	0	1	0	0	1	0	1
Average	45	2	1	0	0	1	0	0	1	0
Bad	25	5	1	1	0	0	0	0	0	0

TABLE 6
Relativities, Expected Loss, and Expected Payments (Deductible Excluded as Covariate)

		Relativity with deductible of						
Profile	Policy limit	100	200	300	500	1000	Expected loss	Expected payment
Good	6,000	0.828	0.681	0.569	0.414	0.226	583.63	574.14
Average	6,000	0.853	0.719	0.613	0.476	0.303	711.99	670.19
Bad	6,000	0.919	0.842	0.777	0.676	0.509	1487.21	1213.45
Good	15,000	0.867	0.747	0.645	0.487	0.27	745.79	744.99
Average	15,000	0.902	0.811	0.733	0.619	0.456	1036.10	1008.15
Bad	15,000	0.956	0.913	0.874	0.804	0.669	2410.96	2260.14

Table 5. For example, a policyholder with a "Bad" profile has many high-risk characteristics, such as young driver, diesel vehicle, and poor car class (Class C).

To avoid excessive complications, in the following analysis we assume a classical frequency–severity framework for aggregated loss, where loss frequency is independent of severity. From a pricing perspective, deductible relativity is a simple yet widely adopted indicator that measures the proportion of covered loss retained by introducing a deductible d_i. The deductible relativity for aggregate loss is defined by $\mathrm{REL}_i(d_i) = E[S_i^p(d_i)|\boldsymbol{x}_i]/E[S_i^p(0)|\boldsymbol{x}_i]$ and satisfies the following relationships:

$$\begin{aligned}\mathrm{REL}_i(d_i) &= \frac{E\left[S_i^p(d_i)|\boldsymbol{x}_i\right]}{E\left[S_i^p(0)|\boldsymbol{x}_i\right]} = \frac{E[N_i|\boldsymbol{x}_i]E\left[(Y_i-d_i)_+\wedge(u_i-d_i)|\boldsymbol{x}_i\right]}{E[N_i|\boldsymbol{x}_i]E[Y_i\wedge u_i|\boldsymbol{x}_i]} = \frac{E\left[(Y_i-d_i)_+\wedge(u_i-d_i)|\boldsymbol{x}_i\right]}{E[Y_i\wedge u_i|\boldsymbol{x}_i]} \\ &= \frac{\sum_{j=1}^{g}\pi_j(\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}})\left[\int_{\xi_\gamma(d_i)}^{\xi_\gamma(u_i)}(\xi_\gamma^{-1}(y)-d_i)f(y;\hat{\boldsymbol{\psi}}_j)dy+(u_i-d_i)(1-F(\xi_\gamma(u_i);\hat{\boldsymbol{\psi}}_j))\right]}{\sum_{j=1}^{g}\pi_j(\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}})\left[\int_{0}^{\xi_\gamma(u_i)}\xi_\gamma^{-1}(y)f(y;\hat{\boldsymbol{\psi}}_j)dy+u_i(1-F(\xi_\gamma(u_i);\hat{\boldsymbol{\psi}}_j))\right]}.\end{aligned} \tag{7.4}$$

As a result, under the classical frequency–severity framework, the deductible relativity for aggregate loss is equal to the deductible relativity for each loss, which can be computed directly from the fitted model. We first consider the original model where the deductible is excluded as a covariate. The relativities are computed in Table 6 across five possible choices of (nonzero) deductible levels $d_i \in \{100, 200, 300, 500, 1000\}$ and two different policy limits $u_i \in \{6000, 15,000\}$. Note that the expected loss and payment are $E[Y_i|\boldsymbol{x}_i]$ and $E[(Y_i-d_i)_+\wedge(u_i-d_i)|\boldsymbol{x}_i]$, and they are, as expected, greater for policyholders with poorer risk profiles and larger policy limits. Given a fixed deductible level, the relativity is also greater with poorer risk profiles and larger policy limits. This is rather intuitive, because such policyholders file larger claims in general, so a fixed amount of deductible should lead to a relatively less impact on the expected claims. This result is in contrast to the GLM regression approach for deductible ratemaking proposed by Lee (2017), which results in a unified value of relativity (for each fixed deductible level) across all policyholders with different risk profiles.

Another quantity of interest in deductible ratemaking is related to coverage modification. With deductibles, the expected number of claims $E[N_i^*|\boldsymbol{x}_i]$ is smaller than the expected number of losses $E[N_i|\boldsymbol{x}_i]$ because any loss amounts below the

TABLE 7
Claim Probability Given a Loss $P(Y_i > d_i|\boldsymbol{x}_i)$ (Deductible Excluded as Covariate)

		$P(Y_i > d_i|\boldsymbol{x}_i)$ with deductible of				
Profile	Policy limit	100	200	300	500	1000
Good	6,000	0.989	0.944	0.857	0.664	0.330
Average	6,000	0.990	0.939	0.818	0.618	0.392
Bad	6,000	0.993	0.956	0.876	0.730	0.513
Good	15,000	0.991	0.959	0.885	0.692	0.341
Average	15,000	0.993	0.961	0.880	0.726	0.518
Bad	15,000	0.996	0.977	0.932	0.832	0.644

TABLE 8
Relativities, Expected Loss, and Expected Payments (Deductible Included as Covariate)

		Relativity with deductible of						
Profile	Policy limit	100	200	300	500	1000	Expected loss	Expected payment
Good	6,000	1.390	1.273	1.139	0.893	0.482	550.63	544.77
Average	6,000	1.640	1.620	1.557	1.425	1.146	637.33	604.16
Bad	6,000	1.730	1.743	1.707	1.603	1.338	869.29	789.75
Good	15,000	1.232	1.128	1.013	0.804	0.462	681.44	681.11
Average	15,000	1.868	1.952	1.966	1.937	1.794	905.57	883.43
Bad	15,000	1.831	1.910	1.927	1.902	1.759	1300.38	1253.25

deductible threshold are not reported as claims. These two quantities satisfy $E[N_i^*|\boldsymbol{x}_i] = E[N_i|\boldsymbol{x}_i]P(Y_i > d_i|\boldsymbol{x}_i)$, where $P(Y_i > d_i|\boldsymbol{x}_i) = 1 - H(\xi_\gamma(d_i); \boldsymbol{x}_i, \hat{\boldsymbol{\Psi}})$, the probability that a loss results to a payment, is usually treated as an offset (same treatment as policyholder exposure) under the claim frequency regression model. The probabilities across different risk profiles, deductible levels, and policy limits are displayed in Table 7. It is expected that a poorer risk profile results in a larger claim probability (given the same deductible level and policy limit), but, surprisingly it is not always true. For example, at a deductible of 300 and policy limit of 6000, the claim probability for an "Average" profile (0.818) is slightly lower than that of a "Good" profile (0.857). In other words, our proposed flexible modeling framework does not assume a (first-order) stochastic dominance relationship of loss distributions between two different risk profiles. In contrast, this is usually implicitly assumed under the GLM framework.

As a comparison, we also investigate the implication of including the deductible level as a covariate from an insurance ratemaking perspective. Table 8 shows the deductible relativities under the new refitted model (deductible included as covariate). As discussed in the previous section, the new refitted model reveals a positive correlation between deductible level and loss severity. Hence, the expected loss and payment obtained from the new refitted model (Table 8), which are calculated assuming a zero deductible, are generally smaller than those obtained from the original model (Table 6). More important, the key observation is that the relativities obtained by the new refitted model are sometimes greater than 1 and are not monotonically decreasing as a function of deductible level. This is very unreasonable, because higher-deductible contracts pay strictly less than lower-deductible contracts. If the pricing policy is set according to Table 8, most (if not all) policyholders will simply switch to zero-deductible contracts, which are now underpriced compared to prices obtained by the original model (Table 6). In this case, the deductible preference of policyholders is distorted, so the link between deductible levels and loss severities implied by the new refitted model will no longer hold true.

Overall, including the deductible as an explanatory variable for ratemaking purposes is a tricky problem. Unlike other covariates, deductible level can be easily altered or manipulated subject to policyholder selection behavior. Also, as mentioned in the previous subsection, the positive correlation between deductible levels and loss severities is not a casual relationship and can be affected by the ratemaking policies. Assuming that a change of deductible level itself does not directly change the driver's risk characteristics, we therefore believe that more reasonable and fair premiums for different policyholders may be

produced if the deductible is excluded as a covariate, even though it is statistically more preferable (based on the AIC) to include the deductible as a covariate.

8. CONCLUDING REMARKS

In this article, we extend the ECM algorithms presented by Fung, Badescu, and Lin (2019a; 2020) such that a non-linear flexible regression model (called LRMoE) can be fitted to random censored and random truncated regression data. The key feature of the extended ECM algorithm is that the complete data are constructed in a way where each observed data point itself randomly "generates" some unobserved data points beyond the truncation interval. The complete data are a pure hypothetical construction because they not resemble reality. Nonetheless, it serves as a very convenient tool that makes the ECM fitting procedures efficiently implementable in the context of censored and truncated regression.

We then demonstrate the usefulness and importance of proposed fitting algorithm through two real data case studies.

The first one fits the proposed algorithm to insurance claim reporting delay data, which are interval censored and random right truncated, with an application to IBNR prediction. Several goodness-of-fit tests reveal that our proposed algorithm fits the reporting delay data very well. To predict the number of IBNR claims, we propose a new semiparametric approach such that by assuming a Poisson claim arrival process, an adequate IBNR predictive distribution is automatically produced by the proposed ECM fitting procedures. The proposed approach is convenient because there is no need to model the claim arrival process explicitly in order to obtain the number of IBNR claims. An appealing extension to our proposed IBNR count prediction approach is to relax the Poisson assumption. Our future research investigations will focus on the possibility of relaxing such assumptions to more complex point arrival processes for which the unbiasedness property of the estimator in Equation (6.1) may still hold. Moreover, it will be interesting to investigate whether this very simple and convenient nonparametric method can be applied directly to raw data without the use of any special point process structure. Assuming success in the above mentioned steps, this microlevel method may represent a viable alternative to the triangular methods currently used in practice and can be grouped with claim amount estimation in order to produce very accurate IBNR reserves.

The second application considers an insurance loss dataset that is random left truncated due to deductibles and right censored due to policy limits, with an application to deductible ratemaking. The proposed model not only fits the loss data excellently but also produces reasonable ratemaking related quantities (e.g., deductible relativity) across various risk profiles and deductible levels. We further study the effects when the deductible level is included as a covariate (instead of only as a truncation point), which is a common actuarial practice. The resulting new refitted model, however, recommends unreasonably high premiums charged to insurance contracts with higher deductible levels. This may be attributed to the possibility that deductible selection preference differs among policyholders with different risk attitudes. This investigation raises big doubts on the current actuarial approach of considering deductible as a covariate in premium and ratemaking calculation. Our future investigations will therefore focus on studying the implications of policyholders' risk appetite and insurance pricing policies to the deductible selection behavior. All of the above issues can only be addressed by a proper estimation procedure for censored and truncated data such as the one proposed in this article.

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.

REFERENCES

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., L. Gong, X. S. Lin, and D. Tang. 2015. Modeling correlated frequencies with application in operational risk management. *Journal of Operational Risk* 10 (1):1–43. doi:10.21314/JOP.2015.157

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

Beirlant, J., Y. Goegebeur, J. Segers, and J. L. Teugels. 2006. *Statistics of extremes: Theory and applications*. Hoboken, NJ: John Wiley & Sons.

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

Bordes, L., and D. Chauveau. 2016. Stochastic EM algorithms for parametric and semiparametric mixture models for right-censored lifetime data. *Computational Statistics* 31 (4):1513–538. doi:10.1007/s00180-016-0661-7

Crevecoeur, J., K. Antonio, and R. Verbelen. 2019. Modeling the number of hidden events subject to observation delay. *European Journal of Operational Research* 277 (3):930–44. doi:10.1016/j.ejor.2019.02.044

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

Ducros, F., and P. Pamphile. 2018. Bayesian estimation of Weibull mixture in heavily censored data setting. *Reliability Engineering & System Safety* 180: 453–62. doi:10.1016/j.ress.2018.08.008

Frees, E. W., and E. A. Valdez. 2008. Hierarchical insurance claims modeling. *Journal of the American Statistical Association* 103 (484):1457–469. 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* 49:647–88. 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

Fung, T. C., A. L. Badescu, and X. S. Lin. 2020. A new class of severity regression models with an application to IBNR prediction. *North American Actuarial Journal* 25 (2): 206–31. doi:10.1080/10920277.2020.1729813

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

Jaspers, S., M. Aerts, G. Verbeke, and P.-A. Beloeil. 2014. A new semi-parametric mixture model for interval censored data, with applications in the field of antimicrobial resistance. *Computational Statistics & Data Analysis* 71:30–42.

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

Lee, G., and C. Scott. 2012. EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis* 56 (9):2816–829. doi:10.1016/j.csda.2012.03.003

Lee, G. Y. 2017. General insurance deductible ratemaking. *North American Actuarial Journal* 21 (4):620–38. doi:10.1080/10920277.2017.1353430

Lee, G. Y., and P. Shi. 2019. A dependent frequency–severity approach to modeling longitudinal insurance claims. *Insurance: Mathematics and Economics* 87:115–29. doi:10.1016/j.insmatheco.2019.04.004

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.

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

Mirfarah, E., M. Naderi, and D.-G. Chen. 2021. Mixture of linear experts model for censored data: A novel approach with scale-mixture of normal distributions. *Computational Statistics & Data Analysis* 158:107182. doi:10.1016/j.csda.2021.107182

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

Sydnor, J. 2010. (Over) insuring modest risks. *American Economic Journal: Applied Economics* 2 (4):177–99. doi:10.1257/app.2.4.177

Tseung, S., A. L. Badescu, T. C. Fung, and X. S. Lin. 2020. LRMoE: An R package for flexible actuarial loss modelling using mixture of experts regression model. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3740215.

Tseung, S., A. L. Badescu, T. C. Fung, and X. S. Lin. 2021. LRMoE.jl: Flexible actuarial loss modelling using mixture of experts regression model. *Annals of Actuarial Science* 15 (2):419–40. doi:10.1017/S1748499521000087

Verbelen, R., K. Antonio, and G. Claeskens. 2016. Multivariate mixtures of Erlangs for density estimation under censoring. *Lifetime Data Analysis* 22 (3): 429–55. doi:10.1007/s10985-015-9343-y

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 Papers Department of Accountancy, Finance and Insurance (AFI), Leuven 623951, KU Leuven, Faculty of Economics and Business (FEB), Department of Accountancy, Finance and Insurance (AFI), Leuven.

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, 1–36. doi:10.3390/risks4030025

Wang, Z., X. Wu, and C. Qiu. 2021. The impacts of individual information on loss reserving. *ASTIN Bulletin* 51 (1):303–47. doi:10.1017/asb.2020.42

Discussions on this article can be submitted until October 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.