A CLASS OF MIXTURE OF EXPERTS MODELS FOR GENERAL INSURANCE: APPLICATION TO CORRELATED CLAIM FREQUENCIES

BY

TSZ CHAI FUNG, ANDREI L. BADESCU
AND X. SHELDON LIN

ABSTRACT

This paper focuses on the estimation and application aspects of the Erlang count logit-weighted reduced mixture of experts model (EC-LRMoE), which is a fully flexible multivariate insurance claim frequency regression model. We first prove the identifiability property of the proposed model to ensure that it is a suitable candidate for statistical inference. An expectation conditional maximization (ECM) algorithm is developed for efficient model calibrations. Three simulation studies are performed to examine the effectiveness of the proposed ECM algorithm and the versatility of the proposed model. The applicability of the EC-LRMoE is shown through fitting an European automobile insurance data set. Since the data set contains several complex features, we find it necessary to adopt such a flexible model. Apart from showing excellent fitting results, we are able to interpret the fitted model in an insurance perspective and to visualize the relationship between policyholders' information and their risk level. Finally, we demonstrate how the fitted model may be useful for insurance ratemaking.

KEYWORDS

Erlang count models, expectation conditional maximization algorithm, logit-weighted gating functions, mixture of experts models, multivariate count regression

1. INTRODUCTION

Claim frequency is one of the two major components under the frequency/severity framework used in Property & Casualty (P&C) insurance for ratemaking purposes. While frequency modeling helps insurers gain insight

Astin Bulletin 49(3), 647-688. doi:10.1017/asb.2019.25

on the claim characteristics, it is also useful for the risk management and regulatory requirements (Frees *et al*., 2016). Modeling claim frequencies has its own challenges that occur due to the fact that one needs to consider not only the effect of policyholders' risk profile to the claim frequency distribution, but also the dependence among various types of claims. In order to capture the aforementioned features, it is very important to develop a flexible multivariate frequency regression model that is well justified to be used in practice.

In Fung *et al*. (2019a), we proposed the logit-weighted reduced mixture of experts model (LRMoE) for multivariate claim frequency/severity regression. The model contains two components: a gating function that governs the probability of each policyholder being classified into different latent homogeneous subgroups; and an expert function that determines the frequency/severity distributions given that a policyholder belongs to a particular subgroup. By letting the gating functions (but not the expert functions) depend on the covariates, the LRMoE is a reduced version of the generalized mixture of experts model (GMoE), which is first introduced by Jacobs *et al*. (1991). The LRMoE enjoys several desirable properties. It is dense in the space of any frequency/severity regression distributions under mild restrictions and with suitably chosen expert functions (called the "denseness condition"), meaning that it can be fully flexible in capturing the underlying distribution, dependence, and regression patterns. As a result, the input data and the output model will share similar characteristics. Also, the LRMoE is closed under response and covariate marginalization, and has reduced-form expressions for the moments and measures of associations, making it mathematically tractable in terms of premium and risk measure calculations.

After identifying a class of multivariate regression models that is theoretically justified, the remaining problem is to choose an appropriate frequency expert function, which is crucial because it determines the flexibility, interpretability, mathematical tractability, and the computational efficiency of the LRMoE. In this paper, we propose the use of the Erlang count (EC) distribution as a suitable expert function that makes the resulting LRMoE (called the EC-LRMoE) possess all of the above-mentioned desirable properties. Further, since the distribution function of the EC model exhibits an analytical form, the computational cost for model calibration is controllable.

In this paper, we first focus on a statistically very important property of the class of EC-LRMoE: identifiability. We prove that the EC-LRMoE is identifiable up to translation and permutation. Identifiability ensures that model fitting is unique and avoids multiple interpretations on a fitted model. Besides these, unidentifiability will cause problems in statistical inference, such as to meaningfully determine the standard errors of the fitted parameters. Identifiability problem for finite mixture models was first formulated in Teicher (1963). Cappé *et al*. (2005) and Fung *et al*. (2019b) extended the identifiability concept to the hidden Markov model. For example, the phase-type distribution (Asmussen *et al*., 1996), which is widely considered as a flexible model, is unidentifiable. Such a problem becomes more challenging for the LRMoE.

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

Jiang and Tanner (1999) demonstrated that the class of mixture of experts models (MoE) not only inherits the permutational invariance issue from finite mixture models, but also is subjected to the translational invariance issue, so that the LRMoE is generally not fully identifiable. Yet, they show that under certain choices of exponential class expert functions, the MoE is still identifiable up to translation and permutation. While EC distribution is not in the exponential class, this paper shows that such an identifiability property will hold for the EC-LRMoE, making the model appealing for the purpose of statistical inference.

Another very important feature of our proposed EC-LRMoE class is the computational tractability. To this end, in this paper we develop an efficient and easily implementable algorithm with reasonable computational costs for the model calibration. For finite mixture models, a popular and efficient approach for parameter estimation is the expectation-maximization (EM) algorithm (McLachlan and Peel, 2000). In a regression setting, Wedel and DeSarbo (1995) formulated the EM algorithm for the finite mixture of generalized linear models (GLMs), and Badescu *et al*. (2015) integrated the usage of built-in statistical computing functions (e.g., GLM function in R) to such an algorithm, making the procedures easy to be implemented. Parameter estimation for the EC-LRMoE through the EM algorithm is much more difficult, as the standard EM algorithm requires a computationally undesirable high-dimensional optimization in the M-step. Therefore, in this paper we propose to use the expectation conditional maximization (ECM) algorithm by Meng and Rubin (1993). This separates the M-step into several substeps so that the problem is reduced to several computationally manageable lower-dimensional convex optimizations. In addition, we need to estimate the integer-valued shape parameters in the EC distributions. Motivated by Gui *et al*. (2018), we propose integrating a similar local search strategy into the M-step to minimize the computational burden. All the above challenges are solved in this paper, as all the steps in the proposed ECM algorithm either involve analytical formulas or only require low-dimensional convex/concave optimizations. Hence, we can easily and efficiently estimate the model parameters under controllable computational costs.

By being interpretable, by possessing all the desirable properties mentioned above and by having an efficient and easy-to-implement algorithm for model calibration, the EC-LRMoE is deemed to be an appropriate multivariate claim count regression model. In the last part of the paper, we demonstrate the practical necessity and applicability of using such a complex class of models. The insurance claim count data set we obtained from an European major automobile insurer exhibits a few unusual features. The mean, dispersion ratio, and tail behavior of the claim counts differ greatly between the two types of coverage. Also, there exist some nonlinear relationships between the response variables and the covariates. We first examine the use of some classic actuarial models, such as the negative binomial (NB) GLM and its zero-inflated version, to fit the marginals of the data set. However, the fitting results are relatively poor,

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

showcasing the need of adopting a highly flexible model. On the contrary, the EC-LRMoE captures well a very wide range of complicated structures implied by the data set.

The rest of the paper is organized as follows. Section 2 reviews the LRMoE and the EC expert function proposed in Fung *et al.* (2019a). The identifiability problem for the proposed model is discussed in Section 3. In Section 4, we present the ECM algorithm to estimate the parameters of the proposed model. Through three simulation studies, in Section 5 we confirm the adequacy of the proposed ECM algorithm, gain insights on the identifiability problem in practice, and verify the full versatility of the proposed model. The application of the proposed model to a real automobile insurance data set is discussed in Section 6. Not only do we analyze the fitted model with interpretations and visualizations, but we also demonstrate the predictive ratemaking power of the fitted model. The paper is concluded in Section 7 with a brief review of our findings, identifying some practical concerns and providing some future research directions.

2. The EC-LRMoE regression model

In this section, we provide a review of the Erlang count logit-weighted reduced mixture of experts model (EC-LRMoE) proposed in Fung *et al.* (2019a). The proofs and derivations related to the model and the motivations of using such a model are discussed in detail in that paper.

Suppose that the insurer issues n bundled insurance contracts, each of which consists of K types of coverage (or called "claim types"). For policyholder $i \in \{1, \ldots, n\}$, denote $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iK})^T$ and $\boldsymbol{y} = (y_{i1}, \ldots, y_{iK})^T$, respectively, as the number of claims random vector (response count variable) and the corresponding realization. Also, define $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{iP})^T$ as the policyholder's risk profile (covariates), where $x_{i0} = 1$. We assume that the policyholders are independent of each other.

2.1. Logit-weighted reduced mixture of experts model (LRMoE)

Under the LRMoE, the probability mass function (pmf) of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is

$$h(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{\Theta}, g) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) \prod_{k=1}^{K} f(y_{ik}; \boldsymbol{\theta}_{jk}), \tag{2.1}$$

where g is the number of latent classes, $\pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha})$ (the gating function) is the mixing weight for the jth class, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$ and $\boldsymbol{\alpha}_j = (\alpha_{j0}, \ldots, \alpha_{jP})^T \in \mathbb{R}^{P+1}$ are the parameters for the regressions of the mixing weights, $f(y_{ik}; \boldsymbol{\theta}_{jk})$ (the expert function) is the pmf of a count distribution with parameters $\boldsymbol{\theta}_{jk}$, and $\boldsymbol{\Theta} =$

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

$\{\boldsymbol{\theta}_{jk};\ j=1,\ldots,g,\ \ k=1,\ldots,K\}$. Moreover, we choose the logit-type gating function for $\pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha})$:

$$\pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha})=\frac{exp\{\boldsymbol{\alpha}_j^T\boldsymbol{x}_i\}}{\sum_{j'=1}^{g}exp\{\boldsymbol{\alpha}_{j'}^T\boldsymbol{x}_i\}},\qquad j=1,\ldots,g. \tag{2.2}$$

Several features and interpretations of the LRMoE in an insurance context are now in place.

- The entire population of policyholders is classified into g unobservable subgroups.
- Policyholder's risk level varies among different subgroups, but is homogeneous within a subgroup.
- The probability that a policyholder belongs to particular subgroup depends on the covariates, so that policyholders with more undesirable characteristics are more likely to be classified into a more risky subgroup, that is, a subgroup that results to more claims on average.
- The regression coefficients $\boldsymbol{\alpha}$ for the gating functions represent how the covariates impact subgroup assignments. Large positive regression coefficient α_{jp} represents a higher chance for an individual to be classified as subgroup j when x_{ip} is large.
- Conditioned on the unobservable subgroups, the number of claims for a given policyholder is independent among the K types of coverages. On the other hand, the number of claims is unconditionally dependent.

Remark 2.1. *The LRMoE includes covariates only for the mixing weights (gating functions). Alternatively, one may consider the GMoE, which also allows regression relationships for the count distributions (gating functions). However,* Fung *et al.* (2019a) *has shown several benefits of using the LRMoE over the GMoE. Firstly, the model flexibility is not impeded when the GMoE is simplified to the LRMoE. Secondly, the LRMoE involves less parameters and contains a simpler mathematical form compared to the GMoE. Thirdly, several desirable mathematical and statistical properties of the LRMoE cannot be satisfied by the GMoE. Overall, the LRMoE is deemed to be a more parsimonious model than the GMoE.*

Remark 2.2. *The LRMoE captures the dependence structure among claim types through introducing a mixture on the conditionally independent distributions, which is very different from the classical dependence models (such as copulas) commonly adopted in actuarial practice. We would regard the LRMoE as a better alternative to copula modeling in terms of the flexibility of capturing dependence structures. The denseness property presented in* Section 2.3 *guarantees that the proposed model can capture any dependence structures. In contrast, there are specified functional forms for the parametric copulas, limiting its model flexibility.* Lee and Lin (2012) *also discussed the superiority of the multivariate mixture models over the copula models.*

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

2.2. Erlang count expert function

Following the discussions on the frequency expert functions in Section 5.1 of Fung *et al.* (2019a), we propose an EC distribution as the expert function f. The frequency distribution is modeled through the waiting times $\{\tau_s; s \in \mathbb{N}\}$ between the $(s-1)$th and the sth event. Also, let $\nu_s = \sum_{s'=1}^{s} \tau_{s'}$ be the time of occurrence of the sth event. Then, the number of events occurring up to time T is given by $N_T = \sup_{s\in\mathbb{N}}\{s; \nu_s \le T\}$. EC model assumes τ_s iid following Erlang distribution with $E[\tau_s] = m/\beta$ and $\mathrm{Var}[\tau_s] = m/\beta^2$. Without much loss of generality and for simplicity, we assume $T = 1$. Then, the expert function is written as

$$f(y;\boldsymbol{\theta}) := P(N_1 = y; m, \beta) = e^{-\beta} \sum_{b=0}^{m-1} \frac{\beta^{my+b}}{(my+b)!}, \qquad y = 0, 1, 2, \ldots, \tag{2.3}$$

where $\boldsymbol{\theta} = (m, \beta)$. In general insurance, conditioned on the subgroup of the policyholder, the inter-arrival time of the claims is independent Erlang distributed. Also, the cumulative distribution has an analytical formula:

$$F(y;\boldsymbol{\theta}) := P(N_1 \le y; m, \beta) = 1 - e^{-\beta} \sum_{b=0}^{my-1} \frac{\beta^b}{b!}, \qquad y = 0, 1, 2, \ldots \tag{2.4}$$

Winkelmann (1995) showed that a renewal model can cater for under-dispersed and equi-dispersed discrete data. From Proposition 4.5 of Fung *et al.* (2019a), mixture modeling can increase the dispersion ratio of a distribution. Therefore, the resulting EC-LRMoE can also capture over-dispersed distributions.

Remark 2.3. *If $m = 1$, the resulting distribution becomes the Poisson distribution, and the corresponding claim arrival process is a homogeneous Poisson process with claim rate β. As a result, the proposed EC-LRMoE contains the class of the Poisson-LRMoE.*

Remark 2.4. *Under the EC model, m is assumed to be a positive integer. If the integer assumption is removed, it will result to Gamma Count model. However, there are no closed-form representations of the probability functions for Gamma Count model. This will increase the computational burden for model fitting. Further, since the EC expert function already fulfills the denseness condition, removing the integer assumption will not improve the flexibility of the corresponding LRMoE. Therefore, it is unnecessary to adopt the Gamma Count expert function.*

Remark 2.5. *Since the EC model can be interpreted as a renewal process, we can view T as the policyholder's exposure. The assumption of $T = 1$ implies that every policyholder has the same exposure period. Unlike the real automobile insurance data set, we will analyze in* Section 6*, which consists of 1-year contracts only, it is possible in practice that the contract periods are different among policyholders. In this case, our model must be adjusted to cater for the exposure information.*

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

One method is to treat it as a covariate for the gating function, but the resulting model will lose some interpretability. Alternatively, one can remove the assumption of $T = 1$. *Then, the density of the EC in Equation* (2.3) *can be easily modified as*

$$f(y;\boldsymbol{\theta},T) := P(N_1 = y; m, \beta, T) = e^{-\beta T}\sum_{b=0}^{m-1}\frac{(\beta T)^{my+b}}{(my+b)!}, \qquad y = 0, 1, 2, \ldots \tag{2.5}$$

The computational aspect of introducing exposures is discussed in Remark 4.2. *We find that there are no computational implications by removing such an assumption.*

2.3. Desirable properties

In the model fitting perspective, it is crucial that the proposed model has a full flexibility to capture any distribution, dependence, and regression patterns, such that the data generated from the fitted model will be highly synchronous to the input data, even if the characteristics of the data set are highly complicated. In other words, the class of candidate models is "dense" in the space of all possible distributions. We first provide definitions of some relevant terms. Note that we have dropped the subscript i for $\boldsymbol{x}_i$ (only) in this subsection to allow for a cleaner presentation and make the notations coherent to that from Fung *et al.* (2019a).

Definition 2.1 (Regression distribution). *A class of regression distributions* $\mathcal{C}(\mathcal{A})$ *(where* $\mathcal{A}$ *is the support of the covariates* $\boldsymbol{x}$*) is a set, where each element* $F(\mathcal{A}) := \{F(\cdot;\boldsymbol{x}); \boldsymbol{x} \in \mathcal{A}\}$ *in* $\mathcal{C}(\mathcal{A})$ *is itself a set of probability distributions.*

Definition 2.2 (Denseness property in the context of multivariate regression distributions). *Let* $\mathcal{A}$ *be the support of the covariates* $\boldsymbol{x}$. *Also, denote* $\mathcal{C}_1(\mathcal{A})$ *and* $\mathcal{C}_2(\mathcal{A})$ *as two classes of regression distributions.* $\mathcal{C}_1(\mathcal{A})$ *is dense in* $\mathcal{C}_2(\mathcal{A})$ *if and only if for all* $F(\mathcal{A}) \in \mathcal{C}_2(\mathcal{A})$, *there exists a sequence of regression distributions* $\{G_n(\mathcal{A})\}_{n=1,2,\ldots}$ *with* $G_n(\mathcal{A}) \in \mathcal{C}_1(\mathcal{A})$ *for* $n = 1, 2, \ldots$ *such that for all* $\boldsymbol{x} \in \mathcal{A}$, $G_n(\boldsymbol{y};\boldsymbol{x}) \xrightarrow{\mathcal{D}} F(\boldsymbol{y};\boldsymbol{x})$ *as* $n \to \infty$. *If the convergence* $G_n(\boldsymbol{y};\boldsymbol{x}) \to F(\boldsymbol{y};\boldsymbol{x})$ *is uniform on* $\boldsymbol{x} \in \mathcal{A}_{\boldsymbol{y}}$ *for any* $\boldsymbol{y}$, *where* $\mathcal{A}_{\boldsymbol{y}}$ *is the set of* $\boldsymbol{x}$ *such that* $\boldsymbol{y}$ *is a continuity point of* $F(\boldsymbol{y};\boldsymbol{x})$, *then* $\mathcal{C}_1(\mathcal{A})$ *is uniformly dense in* $\mathcal{C}_2(\mathcal{A})$.

The denseness property of the proposed EC-LRMoE below is proved in Fung *et al.* (2019a):

Theorem 2.1. *Let* $\mathcal{G}_1(\mathcal{A})$ *be a class of multivariate frequency regression distributions. For each element* $G^*(\mathcal{A}) \in \mathcal{G}_1(\mathcal{A})$ *where* $G^*(\mathcal{A}) := \{G^*(\cdot;\boldsymbol{x}); \boldsymbol{x} \in \mathcal{A}\}$, $\{G^*(\cdot;\boldsymbol{x})\}_{\boldsymbol{x}\in\mathcal{A}}$ *is tight and* $G^*(\boldsymbol{y};\boldsymbol{x})$ *is Lipschitz continuous on* $\boldsymbol{x} \in \mathcal{A}$ *for all* $\boldsymbol{y}$. *Assume that* $\mathcal{A} = \{1\} \times [m_{\min}, m_{\max}]^P$, *where* $m_{\min}$ *and* $m_{\max}$ *are finite. Then, the*

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

class of EC-LRMoE defined in Equations (2.1)–(2.3) *with covariates* $\boldsymbol{x} \in \mathcal{A}$ *is uniformly dense in* $\mathcal{G}_1(\mathcal{A})$.

Note that any assumptions/restrictions imposed in the theorem are very mild and are not of any concern in practice. Apart from the denseness property, the EC-LRMoE also enjoys several desirable distributional and moment properties: it is closed under response and covariates marginalization, and various moments/measures of association corresponding to the proposed model can be expressed in a simple and easily computable form. For example, the mean number of claims for the kth type of coverage is given by

$$E[Y_k \mid \boldsymbol{x}] = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) \sum_{l=0}^{\infty} \left(1 - e^{-\beta_{jk}} \sum_{b=0}^{m_{jk}l-1} \frac{\beta_{jk}^b}{b!}\right), \tag{2.6}$$

where $\boldsymbol{\theta}_{jk} := (m_{jk}, \beta_{jk})$ are the EC parameters for the jth subgroup and the kth type of coverage. Based on the covariate marginalization property, we also obtain

$$E[Y_k \mid \boldsymbol{x}^c] = \sum_{j=1}^{g} \tilde{\pi}_j(\boldsymbol{x}^c; \boldsymbol{\alpha}) \sum_{l=0}^{\infty} \left(1 - e^{-\beta_{jk}} \sum_{b=0}^{m_{jk}l-1} \frac{\beta_{jk}^b}{b!}\right). \tag{2.7}$$

Here, $\boldsymbol{x}^c$ is a subset of the complete covariates $\boldsymbol{x}$, $\tilde{\pi}_j(\boldsymbol{x}^c; \boldsymbol{\alpha}) := \int_{\boldsymbol{x}^u \in D^u} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) dW(\boldsymbol{x}^u; \boldsymbol{x}^c)$ is the covariate-marginalized weight, where $\boldsymbol{x}^u := \boldsymbol{x} \backslash \boldsymbol{x}^c$ represents the missing covariates, D^u is the support of $\boldsymbol{x}^u$, and $W(\boldsymbol{x}^u; \boldsymbol{x}^c)$ is the distribution of $\boldsymbol{x}^u$ conditioned on $\boldsymbol{x}^c$. These properties are crucial in insurance applications in terms of premium and risk measure calculations. Detailed descriptions and the derivations of these properties can be found in Fung *et al.* (2019a).

3. Model identifiability

In the modeling perspective, it is desirable that the model is identifiable, that is, there is a one-to-one mapping between regression distributions and model parameters. Otherwise, there may be issues for statistical inference and there may exist multiple interpretations for the same model. This section discusses the identifiability issues of the proposed EC-LRMoE, following the logic of Jiang and Tanner (1999).

As discussed by Jiang and Tanner (1999), it is impossible for the general class of the LRMoE to be identifiable because of the following two invariance properties:

1. Permutational invariance: From Equation (2.1), it is easy to see that $h_Y(\boldsymbol{y}; \boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{\Theta}, g) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) \prod_{k=1}^{K} f(y_{ik}; \boldsymbol{\theta}_{jk}) = \sum_{j=1}^{g} \pi_{c(j)}(\boldsymbol{x}_i; \boldsymbol{\alpha}) \prod_{k=1}^{K} f(y_{ik}; \boldsymbol{\theta}_{c(j)k})$, where $\{c(1), \ldots, c(g)\}$ is a permutation of $\{1, \ldots, g\}$. Therefore, the model is invariant under the transformation $\boldsymbol{\alpha}_j \mapsto \boldsymbol{\alpha}_{c(j)}$ and

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

$\boldsymbol{\theta}_{jk} \mapsto \boldsymbol{\theta}_{c(j)k}$. Such invariance is also common in many models such as finite mixture models and hidden Markov models.

2. Translational invariance: From Equation (2.2), we can see that

$$\pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}) = \frac{exp\{\boldsymbol{\alpha}_j^T \boldsymbol{x}_i\}}{\sum_{j'=1}^{g} exp\{\boldsymbol{\alpha}_{j'}^T \boldsymbol{x}_i\}} = \frac{exp\{(\boldsymbol{\alpha}_j+\boldsymbol{\delta})^T \boldsymbol{x}_i\}}{\sum_{j'=1}^{g} exp\{(\boldsymbol{\alpha}_{j'}+\boldsymbol{\delta})^T \boldsymbol{x}_i\}},$$

where $\boldsymbol{\delta}$ is a column vector with length $P+1$. Hence, the model is invariant under the transformation $\boldsymbol{\alpha}_j \mapsto \boldsymbol{\alpha}_j + \boldsymbol{\delta}$.

The above issues can be addressed by introducing an ordering on $\boldsymbol{\Phi}_j = (\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j)$ with $\boldsymbol{\theta}_j = (\boldsymbol{\theta}_{j1}, \ldots, \boldsymbol{\theta}_{jK})$, assigning subgroup indexes based on the order of $\boldsymbol{\Phi}_j$ and fixing $\boldsymbol{\alpha}_{j'} = \mathbf{0}$ for one of the $j' \in \{1, \ldots, g\}$. Without addressing these issues, however, it is still interesting to see if any model unidentifiability (two different sets of parameters lead to the same regression distributions) of the LRMoE can only be resulted from the translational and permutational invariance properties, that is, the LRMoE is identifiable up to translation and permutation. Such a nontechnical statement is defined rigorously below in accordance with Jiang and Tanner (1999).

Definition 3.1. *Let $\mathcal{G}$ be the class of LRMoE with the pmf in the form of Equation (2.1). Each element $G_{\boldsymbol{\Phi},g} \in \mathcal{G}$ is a regression distribution with covariates $\boldsymbol{x}_i \in \Omega$, parameter setting $\boldsymbol{\Phi} = (\boldsymbol{\alpha}, \boldsymbol{\Theta})$, and the number of latent classes g, where $\Omega \subseteq \mathbb{R}^{P+1}$ is the support of $\boldsymbol{x}_i$. A subclass $\bar{\mathcal{G}} \subseteq \mathcal{G}$ is identifiable up to translation and permutation whenever $G_{\boldsymbol{\Phi}^*,g^*}, G_{\boldsymbol{\Phi},g} \in \bar{\mathcal{G}}$, $(\boldsymbol{\alpha}_{j_1}^*, \boldsymbol{\theta}_{j_1}^*) \neq (\boldsymbol{\alpha}_{j_2}^*, \boldsymbol{\theta}_{j_2}^*)$ for all $j_1 \neq j_2 \in \{1, \ldots, g^*\}$ and $(\boldsymbol{\alpha}_{j_1}, \boldsymbol{\theta}_{j_1}) \neq (\boldsymbol{\alpha}_{j_2}, \boldsymbol{\theta}_{j_2})$ for all $j_1 \neq j_2 \in \{1, \ldots, g\}$, if*

$$\sum_{j=1}^{g^*} \pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}^*) \prod_{k=1}^{K} f(y_{ik};\boldsymbol{\theta}_{jk}^*) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}) \prod_{k=1}^{K} f(y_{ik};\boldsymbol{\theta}_{jk}), \tag{3.1}$$

for all $\boldsymbol{x}_i \in \Omega$ and $\boldsymbol{y}_i \in \{0, 1, \ldots\}^K$, it implies that $g^* = g$ and $(\boldsymbol{\alpha}_j^*, \boldsymbol{\theta}_j^*) = (\boldsymbol{\alpha}_{c(j)} + \boldsymbol{\delta}, \boldsymbol{\theta}_{c(j)})$ for $j = 1, \ldots, g$, where $\{c(1), \ldots, c(g)\}$ is a permutation of $\{1, \ldots, g\}$ and $\boldsymbol{\delta}$ is a vector that is constant across all $j = 1, \ldots, g$.

Remark 3.1. *In Definition 3.1, the condition that $(\boldsymbol{\alpha}_{j_1}, \boldsymbol{\theta}_{j_1}) \neq (\boldsymbol{\alpha}_{j_2}, \boldsymbol{\theta}_{j_2})$ (and similar for $(\boldsymbol{\alpha}_j^*, \boldsymbol{\theta}_j^*)$) is necessary, or otherwise $G_{\boldsymbol{\Phi},g}$ can be easily reduced to the same regression distribution with a smaller number of components. Also, Equation (3.1) means that the pmf of $G_{\boldsymbol{\Phi},g}$ matches with that of $G_{\boldsymbol{\Phi}^*,g^*}$.*

After presenting what is meant by "identifiable" for the general class of LRMoE, we now demonstrate the identifiability property of the proposed EC-LRMoE.

Theorem 3.1. *The EC-LRMoE is identifiable up to translation and permutation, subject to the restriction that $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g$ are distinct and Ω spans $\mathbb{R}^P$.*

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

Proof. We follow the three-step approach proposed by Fung *et al.* (2019b) to prove the identifiability property.

First, it can be proved that the class of finite mixture of univariate EC is identifiable. Applying Theorem 2 of Teicher (1963), it suffices to show that

1. There exists a transform $\phi_\xi(t)$ defined for $t \in \mathcal{S}_{\phi_\xi}$ such that the mapping $M : F_\xi \to \phi_\xi$ is linear and one-to-one:
   In this case, $\xi = (m, \beta)$ and the probability generating function is taken as the transform, so that we have

$$\phi_\xi(t) = \sum_{y=0}^{\infty} e^{-t^{1/m}\beta} \sum_{i=0}^{m-1} \frac{(t^{1/m}\beta)^{my+i}}{(my+i)!} t^{-i/m} e^{(t^{1/m}-1)\beta}.$$

   The support of t is $\mathcal{S}_{\phi_\xi} = (0, \infty)$. Note that when $t > 1$, we have the following inequalities:

$$e^{(t^{1/m}-1)\beta} t^{-1} \leq \phi_\xi(t) \leq e^{(t^{1/m}-1)\beta}.$$

2. There exists a total ordering of $\mathcal{F}$ such that $F_{\xi_1} \prec F_{\xi_2}$ implies: (i) $\mathcal{S}_{\phi_{\xi_1}} \subseteq \mathcal{S}_{\phi_{\xi_2}}$; (ii) There exists $t^* \in \bar{\mathcal{S}}_{\phi_{\xi_1}}$ (t^* being independent of ϕ_{ξ_2}) such that $\lim_{t \to t^*} \phi_{\xi_2}(t)/\phi_{\xi_1}(t) = 0$:
   To demonstrate this, we order the EC distribution by $f(y; m_1, \beta_1) \prec f(y; m_2, \beta_2)$ when $(m_1 < m_2)$ or $(m_1 = m_2$ and $\beta_1 > \beta_2)$. Choosing $t^* = +\infty \in \bar{\mathcal{S}}_{\phi_{\xi_1}}$, we have

$$\lim_{t \to t^*} \log \left[\frac{\phi_{\xi_2}(t)}{\phi_{\xi_1}(t)} \right] \leq \lim_{t \to t^*} (\beta_2 t^{1/m_2} - \beta_1 t^{1/m_1} + \log t) + const. = -\infty.$$

Second, from the identifiability result of the univariate EC finite mixture and Theorem 2 of Teicher (1967), the class of finite mixtures of multivariate EC is identifiable.

Finally, we will prove the identifiability for the EC-LRMoE. The second step implies that if Equation (3.1) holds, then $\boldsymbol{\theta}_j^* = \boldsymbol{\theta}_{c(j)}$ and $\pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}^*) = \pi_{c(j)}(\boldsymbol{x}_i; \boldsymbol{\alpha})$ for $j = 1, \ldots, g$, since $\pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}^*) > 0$. We have

$$\frac{exp\{\boldsymbol{\alpha}_j^{*T} \boldsymbol{x}_i\}}{\sum_{j'=1}^{g} exp\{\boldsymbol{\alpha}_{j'}^{*T} \boldsymbol{x}_i\}} = \frac{exp\{\boldsymbol{\alpha}_{c(j)}^{T} \boldsymbol{x}_i\}}{\sum_{j'=1}^{g} exp\{\boldsymbol{\alpha}_{c(j')}^{T} \boldsymbol{x}_i\}}, \qquad for\ j = 1, \ldots, g. \tag{3.2}$$

Choosing $j_1, j_2 \in \{1, \ldots, g\}$, plugging them into Equation (3.2) and taking division across the two equations obtained and taking logarithm, we have

$$(\boldsymbol{\alpha}_{j_1}^* - \boldsymbol{\alpha}_{j_2}^*)^T \boldsymbol{x}_i = (\boldsymbol{\alpha}_{c(j_1)} - \boldsymbol{\alpha}_{c(j_2)})^T \boldsymbol{x}_i,$$

for all $j_1, j_2 = 1, \ldots, g$ and $\boldsymbol{x} \in \Omega$. Since Ω spans $\mathbb{R}^{P+1}$, we have $\boldsymbol{\alpha}_{j_1}^* - \boldsymbol{\alpha}_{c(j_1)} = \boldsymbol{\alpha}_{j_2}^* - \boldsymbol{\alpha}_{c(j_2)}$ for all $j_1, j_2 = 1, \ldots, g$. Therefore, $\boldsymbol{\alpha}_j^* - \boldsymbol{\alpha}_{c(j)} := \boldsymbol{\delta} = const.$ for all $j = 1, \ldots, g$. ■

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

4. Parameter estimation: An ECM algorithm

In finite mixture-related models, a common approach for parameter estimation is to apply an EM algorithm (see e.g., Dempster *et al*., 1977; McLachlan and Peel, 2000). In MoE, however, the M-step requires optimization of a non-concave function over all regression coefficients $\boldsymbol{\alpha}$, which is computationally undesirable. A possible solution is to divide the M-step into several substeps and optimize a more mathematically tractable function over a lower dimensional space in each substep, which is possible using the ECM algorithm proposed in Meng and Rubin (1993).

We now present an ECM algorithm to fit the proposed model (Equation 2.1) to data. Assume that there are n independent observations $\{(\boldsymbol{Y}_i, \boldsymbol{x}_i); i = 1, \ldots, n\}$. Hereafter, denote $\boldsymbol{y} := \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ as all the response variables and $\boldsymbol{x} := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ as all the covariates. g is fixed at each ECM run. Its adjustments will be addressed later in this section. The parameters to be estimated are $\boldsymbol{\Phi} = (\boldsymbol{\alpha}, \boldsymbol{\Theta})$. Because of the concern of translational invariance for regression coefficients $\boldsymbol{\alpha}$, we fix $\boldsymbol{\alpha}_g = \boldsymbol{0}$. The log-likelihood of observed data is given by

$$l(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) \prod_{k=1}^{K} f(y_{ik}; \boldsymbol{\theta}_{jk}) \right]. \tag{4.1}$$

To formulate the ECM algorithm, as usual we introduce a latent random vector $\boldsymbol{Z_i} = (Z_{i1}, \ldots, Z_{ig})^T$ such that $Z_{ij} = 1$ if the observation $\boldsymbol{y}_i$ comes from the jth component and $Z_{ij} = 0$ otherwise for $i = 1, \ldots, n$. As a result, $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are independently following a multinomial distribution $Multi_g(1, \{\pi_1(\boldsymbol{x}_i; \boldsymbol{\alpha}), \ldots, \pi_g(\boldsymbol{x}_i; \boldsymbol{\alpha})\})$. The complete data log-likelihood is given by

$$l(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{Z}) = \sum_{i=1}^{n} \sum_{j=1}^{g} Z_{ij} \left(\log \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) + \sum_{k=1}^{K} \log f(y_{ik}; \boldsymbol{\theta}_{jk}) \right). \tag{4.2}$$

4.1. E-step

At the tth iteration, the expectation of the complete data log-likelihood given the observed data is

$$\begin{aligned} Q(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Phi}^{(t-1)}) &= E[l(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{Z}) | \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Phi}^{(t-1)}] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij}^{(l)} \left(\log \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) + \sum_{k=1}^{K} \log f(y_{ik}; \boldsymbol{\theta}_{jk}) \right), \end{aligned} \tag{4.3}$$

where, for $i = 1, \ldots, n$ and $j = 1, \ldots, g$, $z_{ij}^{(t)} = E[Z_{ij} | \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Phi}^{(t-1)}]$ under the EC-LRMoE is expressed as

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

$$z_{ij}^{(t)} = \frac{\pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}^{(t-1)}) \prod_{k=1}^{K} f(y_{ik}; \boldsymbol{\theta}_{jk}^{(t-1)})}{\sum_{j'=1}^{g} \pi_{j'}(\boldsymbol{x}_i; \boldsymbol{\alpha}^{(t-1)}) \prod_{k=1}^{K} f(y_{ik}; \boldsymbol{\theta}_{j'k}^{(t-1)})}. \tag{4.4}$$

4.2. CM-step

The goal of the CM-step is to maximize $Q(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Phi}^{(t-1)})$ with respect to $\boldsymbol{\Phi}$. However, it is too computational costly to directly find the global maximum. Instead, we aim to use a computational effective algorithm to find a near-maximum to update the parameters $\boldsymbol{\Phi}^{(t)}$ such that $Q(\boldsymbol{\Phi}^{(t)}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Phi}^{(t-1)}) \geq Q(\boldsymbol{\Phi}^{(t-1)}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Phi}^{(t-1)})$. It follows from Equation (4.3) that $Q(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Phi}^{(t-1)})$ can be decomposed into two parts, $Q_{\boldsymbol{\alpha}}^{(t)}$ and $Q_{\boldsymbol{\Theta}}^{(t)}$, such that

$$Q(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Phi}^{(t-1)}) = Q_{\boldsymbol{\alpha}}^{(t)} + Q_{\boldsymbol{\Theta}}^{(t)}, \tag{4.5}$$

where

$$Q_{\boldsymbol{\alpha}}^{(t)} = \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij}^{(t)} \log \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij}^{(t)} \left[\boldsymbol{\alpha}_j^T \boldsymbol{x}_i - \log \left(\sum_{j'=1}^{g} \exp\{\boldsymbol{\alpha}_{j'}^T \boldsymbol{x}_i\} \right) \right], \tag{4.6}$$

and

$$Q_{\boldsymbol{\Theta}}^{(t)} = \sum_{j=1}^{g} \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ij}^{(t)} \log f(y_{ik}; \boldsymbol{\theta}_{jk}) = \sum_{j=1}^{g} \sum_{k=1}^{K} Q_{\boldsymbol{\theta}_{jk}}^{(t)} \tag{4.7}$$

with

$$Q_{\boldsymbol{\theta}_{jk}}^{(t)} = \sum_{i=1}^{n} z_{ij}^{(t)} \left[-\beta_{jk} + \log \left(\sum_{b=0}^{m_{jk}-1} \frac{\beta_{jk}^{m_{jk} y_{ik} + b}}{(m_{jk} y_{ik} + b)!} \right) \right]. \tag{4.8}$$

Clearly, $Q_{\boldsymbol{\alpha}}^{(t)}$ depends only on $\boldsymbol{\alpha}$, $Q_{\boldsymbol{\Theta}}^{(t)}$ depends only on $\boldsymbol{\Theta}$, and $Q_{\boldsymbol{\theta}_{jk}}^{(t)}$ depends only on $\boldsymbol{\theta}_{jk}$. As a result, the problem is reduced to separately maximize the objective functions $Q_{\boldsymbol{\alpha}}^{(t)}$ (with respect to $\boldsymbol{\alpha}$) and $Q_{\boldsymbol{\theta}_{jk}}^{(t)}$ (with respect to $\boldsymbol{\theta}_{jk}$).

We first attempt to maximize $Q_{\boldsymbol{\alpha}}^{(t)}$ with respect to $\boldsymbol{\alpha}$. However, direct maximization is difficult because the dimension of $\boldsymbol{\alpha}$ is large. On the other hand, $Q_{\boldsymbol{\alpha}}^{(t)}$ is a concave function of $\boldsymbol{\alpha}_j$ if other parameters $\{\boldsymbol{\alpha}_{j'}; j' \neq j\}$ are fixed. Therefore, it is computationally simpler to implement the CM-steps that optimize the parameters sequentially for $j = 1, \ldots, g-1$. The CM-steps are as follows:

- Step 1: Obtain $\boldsymbol{\alpha}_1^{(t)}$ through maximizing $Q_{\boldsymbol{\alpha}}^{(t)}$ with $\{\boldsymbol{\alpha}_j; j = 2, \ldots, g-1\}$ fixed at $\boldsymbol{\alpha}_j^{(t-1)}$.
- Step 2: Obtain $\boldsymbol{\alpha}_2^{(t)}$ through maximizing $Q_{\boldsymbol{\alpha}}^{(t)}$ with $\boldsymbol{\alpha}_1$ fixed at $\boldsymbol{\alpha}_1^{(t)}$ and $\{\boldsymbol{\alpha}_j; j = 3, \ldots, g-1\}$ fixed at $\boldsymbol{\alpha}_j^{(t-1)}$.

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

- ...
- Step $g-1$: Obtain $\boldsymbol{\alpha}_{g-1}^{(t)}$ through maximizing $Q_{\boldsymbol{\alpha}}^{(t)}$ with $\{\boldsymbol{\alpha}_j; j=1,\ldots,g-2\}$ fixed at $\boldsymbol{\alpha}_j^{(t)}$.

For each CM-step, $\boldsymbol{\alpha}_j^{(t)}$ can be obtained through the iteratively reweighted least squares (IRLS) approach (Jordan and Jacobs, 1994), that is, perform the following iterations until convergence, using $\boldsymbol{\alpha}_j^{(t-1)}$ as the initialization:

$$\boldsymbol{\alpha}_j \leftarrow \boldsymbol{\alpha}_j - \left(\frac{\partial^2 Q_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\alpha}_j^T}\right)^{-1} \frac{\partial Q_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}_j}, \qquad j=1,\ldots,g-1, \tag{4.9}$$

where the derivatives are given by

$$\frac{\partial Q_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}_j} = \sum_{i=1}^{n}\left[z_{ij}^{(t)} - \frac{\exp\{\boldsymbol{\alpha}_j^T \boldsymbol{x}_i\}}{\sum_{j'=1}^{g}\exp\{\boldsymbol{\alpha}_{j'}^T \boldsymbol{x}_i\}}\right]\boldsymbol{x}_i, \tag{4.10}$$

$$\frac{\partial^2 Q_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\alpha}_j^T} = \sum_{i=1}^{n}\frac{(\exp\{\boldsymbol{\alpha}_j^T \boldsymbol{x}_i\} - \sum_{j'=1}^{g}\exp\{\boldsymbol{\alpha}_{j'}^T \boldsymbol{x}_i\})\exp\{\boldsymbol{\alpha}_j^T \boldsymbol{x}_i\}}{(\sum_{j'=1}^{g}\exp\{\boldsymbol{\alpha}_{j'}^T \boldsymbol{x}_i\})^2}\boldsymbol{x}_i \boldsymbol{x}_i^T. \tag{4.11}$$

Note that in the CM-steps above, $Q_{\boldsymbol{\alpha}}^{(t)}$ is not maximized globally. Instead, we have $Q_{\boldsymbol{\alpha}^{(t)}}^{(t)} \geq Q_{\boldsymbol{\alpha}^{(t-1)}}^{(t)}$, meaning that $Q_{\boldsymbol{\alpha}}^{(t)}$ is increased.

The remaining task is to maximize $Q_{\boldsymbol{\theta}_{jk}}^{(t)}$ with respect to $\boldsymbol{\theta}_{jk}$. The first step is to fix the shape parameter m_{jk} and globally maximize the objective function with respect to the rate parameter β_{jk}. We introduce the following Proposition to show that such maximization is easy to implement.

Proposition 4.1. *For a fixed $m_{jk} > 0$, $Q_{\boldsymbol{\theta}_{jk}}^{(t)}$ is a concave function on $\beta_{jk} \in (0,\infty)$.*

Proof. By Equation (4.8), it suffices to show that $f(x) := \log(g(x))$ is concave on $x \in (0,\infty)$, where $g(x) = \sum_{b=m}^{M} x^b/b!$ and M, m are positive integers with $m \leq M$. Note that $f''(x) \leq 0$ if and only if $h(x) := g(x)g''(x) - [g'(x)]^2 \leq 0$. By some algebraic manipulations, we have

$$\begin{aligned} h(x) &= \sum_{b=m}^{M}\frac{1}{b!}x^b \sum_{b=m}^{M}\frac{1}{(b-2)!}x^{b-2} - \left(\sum_{b=m}^{M}\frac{1}{(b-1)!}x^{b-1}\right)^2 \\ &= \sum_{b=M+m-1}^{M=m-2}\left(\frac{1}{M!(b-M)!}\right)\left(1-\frac{M}{b-M+1}\right)x^b \\ &\quad + \sum_{b=2m-3}^{M+m-2}\left(\frac{1}{(m-2)!(b-m+2)!}\right)\left(1-\frac{b-m+2}{m-1}\right)x^b \leq 0. \end{aligned}$$

■

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

Since $\lim_{\beta_{jk}\to 0} Q^{(t)}_{\boldsymbol{\theta}_{jk}} = \lim_{\beta_{jk}\to\infty} Q^{(t)}_{\boldsymbol{\theta}_{jk}} = -\infty$, it follows from Proposition 4.1 that $Q^{(t)}_{\boldsymbol{\theta}_{jk}}$ only has one (global) maximum point on $x \in (0,\infty)$ given that m_{jk} is fixed. Therefore, β_{jk} in Equation (4.8) can be optimized numerically with little computational cost. The second step is to find $m^{(t)}_{jk}$ that maximizes $Q^{(t)}_{m_{jk}} := \sup_{\beta_{jk}>0} Q^{(t)}_{\boldsymbol{\theta}_{jk}}$. This is not a trivial task since m_{jk} is discrete. One possible approach is to fix m_{jk} in each complete ECM run and adopt the element-wise +1/−1 variation strategy for each m_{jk} proposed in Lee and Lin (2010). However, this method requires a large number of ECM runs and prohibits parallel computing, so it is not computationally desirable. Motivated by the approach of the generalized EM (GEM), we do not aim to globally maximize $Q^{(t)}_{m_{jk}}$. Instead, we find $m^{(t)}_{jk}$ that can potentially increase $Q^{(t)}_{m^{(t-1)}_{jk}}$. Motivated also by Gui *et al*. (2018) and the +1/−1 strategy, we propose a local search strategy on m_{jk} within the CM-step. Denote $D := \{m^{(t-1)}_{jk} - 1, m^{(t-1)}_{jk}, m^{(t-1)}_{jk} + 1\}$. The updates of the shape and the scale parameters are then given by

$$m^{(t)}_{jk} = \underset{m_{jk}\in D}{\operatorname{argmax}}\, Q^{(t)}_{m_{jk}}; \qquad \beta^{(t)}_{jk} = \underset{\beta_{jk}>0}{\operatorname{argmax}}\, Q^{(t)}_{(m^{(t)}_{jk},\beta_{jk})}. \tag{4.12}$$

It is obvious that $Q^{(t)}_{\boldsymbol{\theta}^{(t)}_{jk}} \geq Q^{(t)}_{\boldsymbol{\theta}^{(t-1)}_{jk}}$, so $Q^{(t)}_{\boldsymbol{\theta}_{jk}}$ is also increased.

To sum up, the full procedures described in this subsection guarantee that $Q(\boldsymbol{\Phi}^{(t)}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Phi}^{(t-1)}) \geq Q(\boldsymbol{\Phi}^{(t-1)}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Phi}^{(t-1)})$. Therefore, the observed log-likelihood is non-decreasing for each iteration. The E-step and CM-step are iterated until the observed data log-likelihood is smaller than a tolerance threshold of 10^{-2}.

Remark 4.1. *The costs of computing the E-step and updating the regression parameters* $\boldsymbol{\alpha}$ *in the CM-step are much lower than that of maximizing* $Q^{(t)}_{\boldsymbol{\theta}_{jk}}$. *It is because the maximization of* $Q^{(t)}_{\boldsymbol{\alpha}}$ *involves only the IRLS algorithm similar to Newton's method, which is a fast convergence algorithm, while we have used numerical optimization functions in R to maximize* $Q^{(t)}_{\boldsymbol{\theta}_{jk}}$. *To reduce the run time, one may extend the ECM algorithm above to a multicycle version, which repeats computing the E-step and increasing* $Q_{\boldsymbol{\alpha}^{(t)}}$ *several times before increasing* $Q_{\boldsymbol{\theta}^{(t)}_{jk}}$ *once.*

Remark 4.2. *Considering policyholders' exposures as discussed in* Remark 2.5, *a similar ECM algorithm can be developed accordingly with Equation* (4.8) *modified to*

$$Q^{(t)}_{\boldsymbol{\theta}_{jk}} = \sum_{i=1}^{n} z^{(t)}_{ij} \left[-\beta_{jk} T_i + \log\left(\sum_{b=0}^{m_{jk}-1} \frac{(\beta_{jk} T_i)^{m_{jk} y_{ik} + b}}{(m_{jk} y_{ik} + b)!} \right) \right], \tag{4.13}$$

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

where T_i is the contract period of the ith policyholder. Since the mathematical form of Equation (4.13) is very similar to that of Equation (4.8), one should not expect that the introduction of exposure will introduce any extra computational burden.

4.3. Initialization and parameter adjustments

The performance of the proposed ECM algorithm depends on the initialization. Therefore, it is suggested to try multiple initializations and choose the one that yields the best fitting performance. Here, we propose a simple random initialization method that is already robust based on our experiments.

- For $j = 1, \ldots, g$ and $k = 1, \ldots, K$, sample $m_{jk}^{(0)}$ uniformly on $\{1, \ldots, C\}$, where C is a constant. Based on our experiments, the fitting results are stable unless C is too small ($C = 1$) or too large $C > 10$.
- For $j = 1, \ldots, g$ and $k = 1, \ldots, K$, set $\beta_{jk}^{(0)} = U_{jk} \times \{\beta : E[W_{jk}] = \sum_{i=1}^{n} y_{ik}/n\}$, where W_{jk} is a EC random variable (Equation 2.3) with shape parameter $m_{jk}^{(0)}$ and rate parameter β and U_{jk} is a positive random variable with mean 1 and a small standard deviation to perturb the initialization. Note that $E[W_{jk}] = \sum_{q=1}^{\infty} (1 - e^{-\beta} \sum_{b=0}^{m_{jk}^{(0)} q-1} \beta^b/b!)$.
- Set $\boldsymbol{\alpha}^{(0)} = \mathbf{0}$.

Remark 4.3. *Alternatively, one may apply the K-means clustering method for initializations (see e.g., Gui et al., 2018). First, perform K-means clustering on $\boldsymbol{y}$ with g clusters, which yields the clustering mean $\{\mu_{jk}^{cluster}\}_{j=1,\ldots,g;\, k=1,\ldots,K}$ and the clustering weights $\{\pi_j^{cluster}\}_{j=1,\ldots,g}$ (the proportion of observations classified in cluster j). Second, set $m_{jk}^{(0)}$ at the same way as simple random initialization method and set $\beta_{jk}^{(0)} = U_{jk} \times \{\beta : E[W_{jk}] = \mu_{jk}^{cluster}\}$. Third, set $\alpha_{j1}^{(0)} = \log(\pi_j^{cluster}/\pi_g^{cluster})$ and $\alpha_{jp}^{(0)} = 0$ for $p > 1$.*

Finally, we are to find g that optimizes the fitting result. We determine the optimal g based on the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). With the possibility of parallel computing, it is computationally feasible to try different g and find the one that optimizes the criterion.

Remark 4.4. *One major concern of the proposed algorithm is the overall run time, because for each of the simulation studies* (Section 5) *and for the real data analysis (Section 6), we need to try multiple (≥ 10) initializations, try a wide range of g, perform local search strategies on m_{jk}, and (for real data analysis) conduct bootstrapping for parameter uncertainties. For each data set, the whole*

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

fitting process involves no more than 50 CPUs (∼2GHz) running for less than a day. We realize that the computational burden is intensive, but is still feasible in practice because we only need to calibrate the parameters once for each data set. We will discuss some possible methods to significantly reduce the run time in Section 7.

5. Simulation studies

This section illustrates the flexibility of the proposed model and the effectiveness of the proposed algorithm via three simulation studies. The first study is a simple example involving fitting of simulated data from an EC-LRMoE model. Understanding the similarities and differences between the fitted model and the target model, we can examine the adequacy of the proposed ECM algorithm and also gain insight in the relationships between the identifiability in theory and in practice. The second and third study are based on highly different classes of models. The models are intentionally chosen to be complex and highly heterogeneous, involving various rather extreme correlation structures among response marginals, under/over-dispersed marginal distributions, linear/nonlinear regression patterns, and interactions among covariates. The main purposes of these studies are to demonstrate the versatility of the proposed model and verify that the denseness theory is applicable to empirical model fitting.

5.1. Two-components bivariate MoE with two covariates

This study examines the ability of the proposed algorithm to recover an EC-LRMoE model. In each simulation, we generate 20,000 sets of observations $\{(y_{i1}, y_{i2}), i = 1, \ldots, 20,000\}$ from the proposed multivariate EC MoE model with $g = 2$, $K = 2$, $P = 3$ and the following parameters:

$$\boldsymbol{\alpha} = \begin{pmatrix} -2.5 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}; \qquad \boldsymbol{m} = \begin{pmatrix} 4 & 1 \\ 3 & 5 \end{pmatrix}; \qquad \boldsymbol{\beta} = \begin{pmatrix} 4 & 1.5 \\ 1.5 & 1 \end{pmatrix}.$$

The covariates x_{i2} and x_{i3} are simulated from $U[0, 1]$ and Bernoulli $(p = 0.5)$, respectively. Table 1 shows the summary statistics, where the subgroup conditional mean and variance are, respectively, computed as

$$\begin{aligned} E[Y_{ik}|Z_{ij} = 1] &= \sum_{y=0}^{\infty} y f(y; \boldsymbol{\theta}_{jk}); \\ Var[Y_{ik}|Z_{ij} = 1] &= \sum_{y=0}^{\infty} y^2 f(y; \boldsymbol{\theta}_{jk}) - \left(E[Y_{ik}|Z_{ij} = 1]\right)^2, \end{aligned} \tag{5.1}$$

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

TABLE 1

SUMMARY STATISTICS OF THE TARGET MODEL.

$E[Y_{ik} \mid Z_{ij}=1]$	$k=1$	$k=2$	$Var[Y_{ik} \mid Z_{ij}=1]$	$k=1$	$k=2$	$P(Z_{ij}=1)$
$j=1$	0.619	1.500	$j=1$	0.342	1.500	0.198
$j=2$	0.196	0.004	$j=2$	0.166	0.004	0.802
$E[Y_{ik}]$	0.279	0.299	$Var[Y_{ik}]$	0.229	0.654	

TABLE 2

SUMMARY FOR FITTED PARAMETERS: MEDIAN AND 95% CONFIDENCE INTERVAL.

$\boldsymbol{\alpha}$	$p=1$	$p=2$	$p=3$		
$j=1$	−2.497 (−2.617, −2.384)	0.989 (0.862, 1.135)	1.002 (0.913, 1.085)		
$\boldsymbol{\beta}$	$k=1$	$k=2$	$\boldsymbol{m}$	$k=1$	$k=2$
$j=1$	4.004 (3.921, 4.083)	1.502 (1.446, 1.558)	$j=1$	4 (4, 4)	1 (1, 1)
$j=2$	1.500 (1.474, 1.549)	0.004 (0.000, 0.148)	$j=2$	3 (3, 3.025)	1 (1, 2)

where $\boldsymbol{\theta}_{jk}=(m_{jk}, \beta_{jk})$ and the function f comes from Equation (2.3). We also estimate

$$P(Z_{ij}=1) \simeq \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) \tag{5.2}$$

as the average probability of an individual being classified in the jth subgroup over the whole population of policyholders, and we choose $n_{sim}=5,000,000$ for accurate estimations. Further, the unconditional mean and variance are calculated by $E[Y_{ik}]=\sum_{j=1}^{g} E[Y_{ik}|Z_{ij}=1]P(Z_{ij}=1)$ and $\text{Var}[Y_{ik}]=\sum_{j=1}^{g} E[Y_{ik}^2|Z_{ij}=1]P(Z_{ij}=1)-(E[Y_{ik}])^2$.

It can be seen that Y_{ik} tends to be smaller when sample i belongs to the second component. Especially, $E[Y_{i2}|Z_{i2}=1]$ is very close to zero, so the marginal distribution of Y_{i2} is almost zero-inflated, which is a common characteristic of insurance claims data. Also, marginal 1 is under-dispersed while marginal 2 is over-dispersed.

We use the proposed ECM algorithm to fit the simulated data. The whole process is replicated by 200 times to ensure a complete examination of the proposed algorithm. Using both AIC and BIC, the algorithm correctly identifies that there are two components in 187 and 200 out of 200 replications, respectively. This aligns to the result by Kuha (2004) that BIC is superior in identifying the true model if the sample size is sufficiently large. The summary of parameter estimates is listed in Table 2. Most parameters can be almost recovered with high precision, except for (m_{22}, β_{22}) and in very rare cases (m_{21}, β_{21}). The differences can be explained by the empirical unidentifiability of the true model when the marginal component mean is too small. Denote

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

W an EC random variable with shape parameter m, shape parameter β, and mean $\mu := E[W]$. $\mu \approx 0$ implies that W follows approximately a degenerate(0) distribution regardless of the choice of m. Also, consider a limiting property of the dispersion ratio of the EC model $\lim_{\mu \to 0} Var[W]/\mu = 1$ with m fixed. It shows that if $\mu \approx 0$, W is still Poisson-like ($m = 1$) even if the true model has $m > 1$. Theoretically, EC models are identifiable as shown in Section 3. Nonetheless, EC distributions under different parameter settings can still be arbitrarily close to each other as long as $\mu \to 0$. Together with the sampling error of data generation, it is impossible to recover m_{22} and β_{22}.

To further verify the proposed initialization strategy and evaluate the impact of incorrectly identifying (m_{22}, β_{22}), we also initialize using the true model parameters in each simulation replication and see the difference of the fitting performance. Although this can yield correct (m_{22}, β_{22}), the resulting observed log-likelihood is negligibly different from that obtained by our proposed initialization strategy. Therefore, the impact of misspecifying (m_{22}, β_{22}) is minimal and the proposed fitting algorithm is overall deemed to be appropriate. Yet, this simulation study shows that one should not overinterpret the shape parameter m_{jk} and the shape parameter β_{jk} of the proposed model in real data analysis if the corresponding marginal component mean is close to zero.

5.2. Trivariate nonlinear regression model with one covariate

This study aims to evaluate the flexibility of the proposed model to simultaneously cater for various marginal distributional properties, dependence structures, and regression patterns. We use the proposed model to fit 10,000 observations $\{(y_{i1}, y_{i2}, y_{i3}), i = 1, \ldots, 10,000\}$ generated from a complex hypothetical model. Marginally, y_{ik} has the following pmf $f_k(y)$ for $k = 1, 2, 3$

$$f_1(y) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)}, \quad Z(\lambda, \nu) = \sum_{b=0}^{\infty} \frac{\lambda^b}{(b!)^\nu}; \tag{5.3}$$

$$f_2(y) = \binom{y+m-1}{y} \left(\frac{m}{\mu+m}\right)^m \left(\frac{\mu}{\mu+m}\right)^y; \qquad f_3(y) = \frac{\lambda^{*y} e^{-\lambda^*}}{y!}. \tag{5.4}$$

One covariate x_{i2}, which is simulated from $U[0, 1]$, is introduced. Equation (5.3) is the pmf of Conway–Maxwell–Poisson (CMP) distribution introduced by Conway and Maxwell (1962), which allows for both over-dispersion ($\nu < 1$) and under-dispersion ($\nu > 1$). We choose the parameters $\lambda = 15$ and $\nu = \exp\{0.5 + 0.25x_{i2}\}$ so that the resulting distribution is under-dispersed. CMP distribution has already been applied to motor vehicle crashes data (Lord *et al.*, 2008), which can be under-dispersed and can directly relate to insurance claims. The second marginal follows an NB distribution, which only caters for over-dispersion. We choose $m = 2$ and $\mu = \exp\{0.5 + 0.25x_{i2}\}$ such that it

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

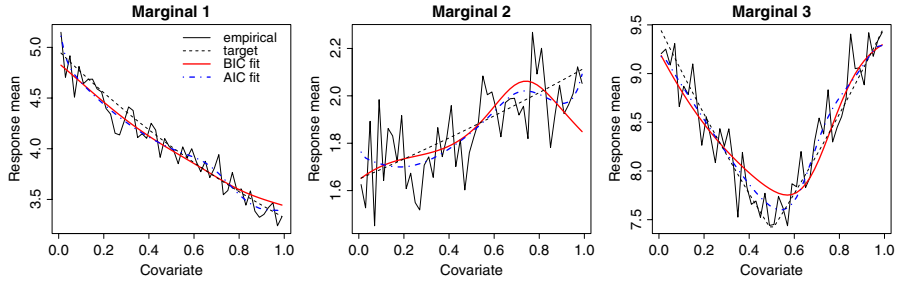

FIGURE 1: Regression patterns – Response mean versus the covariate.

is a standard NB regression. The third marginal follows the Poisson distribution that is undispersed. To induce challenges for model fitting, we choose $\lambda^* = \exp\{2 + 0.5|x_{i2} - 0.5|\}$ so that the regression pattern is nonlinear.

Overall, there is a large heterogeneity among response marginals. In terms of the marginal distributions, we need to model under-dispersed, undispersed, and over-dispersed distributions at the same time. In terms of regression patterns, Figure 1 shows the existence of decreasing, increasing, and "V-shaped" patterns as a function of the covariate. Also, marginal 1 has a stronger regression pattern than marginal 2. The solid rough patterns in Figure 1 show the empirical marginal response mean under various covariate values, based on a uniform kernel with covariate bandwidth 0.01. Higher pattern fluctuations relative to the regression trend for marginal 2 reveal that its regression pattern is weaker.

The dependence among marginals is modeled by a Gaussian copula. Because of its mathematical tractability, it is commonly used in modeling dependence structure of insurance claim frequencies among business lines (see e.g., Shi and Valdez, 2014). The copula function is

$$C(u_1, u_2, u_3) = \Phi_{\mathbf{\Sigma}}\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3)\right), \tag{5.5}$$

where Φ is a standard normal cdf, Φ_{Σ} is the joint cdf of multivariate normal distribution with mean $\mathbf{0}$, covariance matrix $\mathbf{\Sigma}$, and $\text{diag}(\mathbf{\Sigma}) = \mathbf{1}$. We choose $(\mathbf{\Sigma})_{12} = 0.4$, $(\mathbf{\Sigma})_{13} = 0$, and $(\mathbf{\Sigma})_{23} = -0.6$ to allow for both positive and negative correlations.

Based on the proposed algorithm, 15 and 36 components are detected in the fitted model using BIC and AIC, respectively. We first compare the marginal distributions of the fitted model and the simulated observations through histograms. Figure 2 shows that our proposed model is versatile to capture different marginal distributional properties. Then, the dependence structures of fitted models and target model are compared through Kendall's tau. We apply the method proposed by Badescu *et al.* (2015) that considers two cases for the computations of Kendall's tau: with and without covariates' influence.

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

TABLE 3

KENDALL'S τ FOR THE FITTED MODEL VERSUS THE TARGET MODEL.

with covariates (BIC)				without covariates (BIC)			
	y_1	y_2	y_3		y_1	y_2	y_3
y_1		0.233	−0.008	y_1		0.203	−0.005
y_2	0.255		−0.416	y_2	0.218		−0.365
y_3	0.011	−0.426		y_3	0.014	−0.382	
with covariates (AIC)				**without covariates (AIC)**			
	y_1	y_2	y_3		y_1	y_2	y_3
y_1		0.248	0.006	y_1		0.218	0.010
y_2	0.255		−0.422	y_2	0.218		−0.375
y_3	0.011	−0.426		y_3	0.014	−0.382	

Upper and lower triangles represent fitted and target models, respectively.

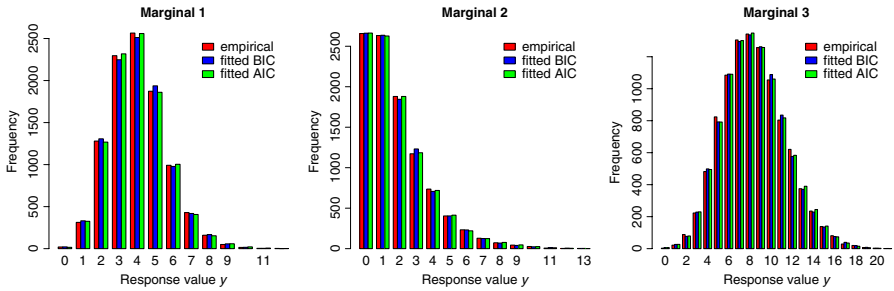

FIGURE 2: Barplots for the fitted model versus the target model.

From Table 3, it is concluded that the fitted models reflect well the dependence structure of the target model. Finally, regression patterns between fitted and target models are displayed in Figure 1. Solid and dotted thick curves are the fitted regression patterns using BIC and AIC models, respectively, while the thin dotted curve is calculated analytically from the target model. The fitted curves can generally capture various regression trends of the empirical data simulated from the target model. AIC fitted model, which contains more components, better fits the regression patterns, especially when the patterns obtained by empirical data are rather weak (e.g., $x_{i2} > 0.6$ for marginal 2) or rather abnormal (e.g., the V-shaped pattern for $0.4 < x_{i2} < 0.6$ in marginal 3). Under the model selection through AIC, which penalizes extra parameters less heavily, there can be a concern of over-fitting because the number of components obtained is rather large. However, the regression curves obtained are still smooth under AIC, addressing the over-fitting concern. With a good fitting performance to such a complicated model, this study confirms the denseness property of our proposed model.

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

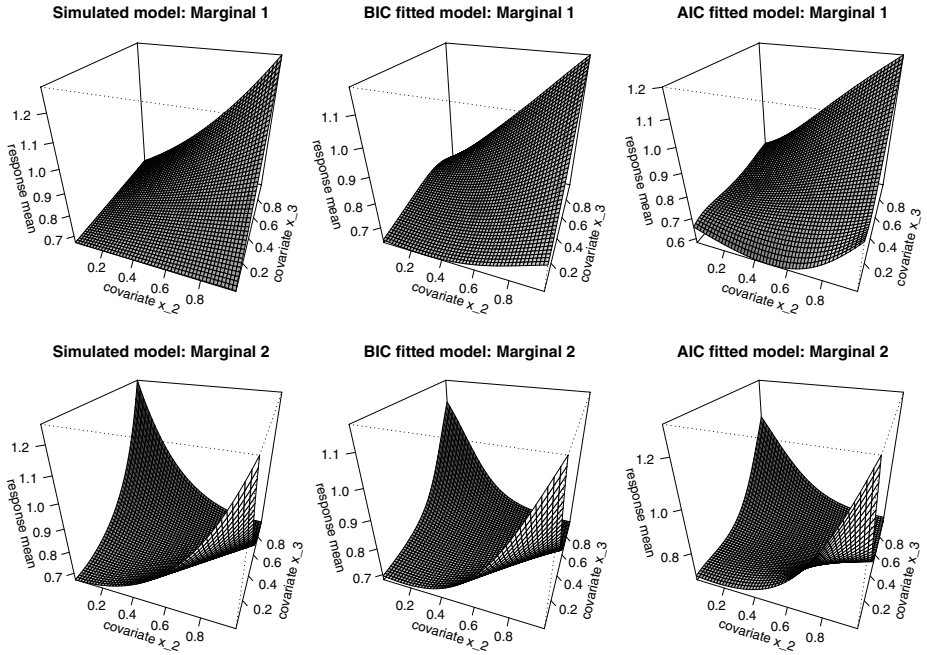

FIGURE 3: Regression patterns – Response mean versus the two covariates.

5.3. Bivariate regression model with covariates interactions

To better understand the flexibility of the proposed model, it is desirable to perform multiple simulation studies that are based on different classes of models. In this study, we simulate 10,000 observations $\{(y_{i1}, y_{i2}), i = 1, \ldots, 10,000\}$ from the Poisson common shock model with two covariates x_{i2} and x_{i3} generated independently from $U[0, 1]$. The common shock model is a popular model for multivariate insurance claim counts since it has a simple mathematical representation and has a physical interpretation that some incidents can lead to multiple types of claims simultaneously. See Bermúdez and Karlis (2011) for a relevant application. This model is given by $Y_{i1} = Y^*_{i1} + Z_i$ and $Y_{i2} = Y^*_{i2} + Z_i$. We choose $Y^*_{i1} \sim \text{Poi}(-1 + x_{i2}x_{i3})$, $Y^*_{i2} \sim \text{Poi}(-1 + (x_{i2} - x_{i3})^2)$, and the common shock $Z_i \sim \text{Poi}(0.3)$. We aim to investigate how well the proposed model can capture such dependence structure and covariates interactions behavior. The covariates interactions behavior can be visualized in the left panels of Figure 3.

Based on BIC and AIC, the resulting fitted model contains 6 and 12 components, respectively. The fitting performances for both criteria are evaluated as follows. Firstly, Table 4 shows that the first four moments of the fitted models match well to that of the target model. Secondly, Kendall's τ with or without covariates' influence is also displayed in Table 5. It shows that our proposed model can capture well the dependence structure of the target model. Thirdly,

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

TABLE 4

THE FIRST FOUR MOMENTS FOR THE FITTED MODEL VERSUS THE TARGET MODEL.

		Empirical	BIC fitted	% error (BIC)	AIC fitted	% error (AIC)
Marginal 1	Mean	0.772	0.772	0.037	0.771	−0.002
	Variance	0.795	0.790	−0.697	0.798	0.382
	Skewness	1.183	1.158	−2.123	1.199	1.399
	Kurtosis	4.466	4.326	−3.144	4.601	3.016
Marginal 2	Mean	0.739	0.739	0.019	0.739	0.004
	Variance	0.741	0.740	−0.154	0.740	−0.049
	Skewness	1.178	1.139	−3.313	1.173	−0.422
	Kurtosis	4.447	4.171	−6.213	4.452	0.104

TABLE 5

KENDALL'S τ FOR THE FITTED MODEL VERSUS THE TARGET MODEL.

	BIC		AIC	
	Target model	Fitted model	Target model	Fitted model
With covariates	0.322	0.316	0.322	0.320
Without covariates	0.212	0.214	0.212	0.200

we examine the performance of the fitted models in capturing the influence of covariates through three-dimensional plots displayed in Figure 3. The target model (left panels) is compared to the fitted models (middle and right panels). The z-axis corresponds to the mean of the response y_{ik} (e.g., For marginal 1, the mean is $\exp\{-1 + x_{i2}x_{i3}\} + 0.3$). For each marginal, the shapes of the resulting surfaces are similar to each other, confirming the ability of the proposed model in capturing various covariates interactions behavior. The smoothness of the surfaces produced by the fitted models also addresses the potential concerns of over-fitting. Overall, the performances of the fitted models are good regardless of the criterion used.

6. APPLICATION TO INSURANCE COUNT DATA REGRESSION

6.1. Data overview

The insurance claim counts data set comes from an European major automobile insurer. It contains the information of 18,019 policyholders who started or renewed their insurance contracts during the year of 2015. Any contracts involved are of 1-year term. When contracts are expired, policyholders may renew the contract so they can keep insured for another year.

The structure of this European insurance data set is similar to that of Bermúdez (2009) and Shi and Valdez (2014). The data set records the number

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

TABLE 6

SUMMARY OF THE COVARIATES.

Discrete valued covariates					
Variable	Description	Mean	SD	Minimum	Maximum
x_{i1}	Age of policyholder	51.000	11.702	20	88
x_{i2}	Car age	6.248	3.335	0	26

Categorical covariates			
Variable	Description	Levels	Proportions
x_{i3}	Car fuel	Diesel: $x_{i3} = 1$	0.383
		Gasoline: $x_{i3} = 0$	0.617
x_{i4}–x_{i5}	Policyholder's history	Renewal with claims last year: $x_{i4} = 1$	0.148
		New contract: $x_{i5} = 1$	0.235
		Renewal, no claims last year: $x_{i4}, x_{i5} = 0$	0.618
x_{i6}–x_{i9}	Geographical location	Region I: $x_{i6} = 1$	0.187
		Region II: $x_{i7} = 1$	0.146
		Region III: $x_{i8} = 1$	0.111
		Capital: $x_{i9} = 1$	0.420
		Region IV: $x_{i6}, x_{i7}, x_{i8}, x_{i9} = 0$	0.136
x_{i10}–x_{i11}	Car brand class	Class A: $x_{i10} = 1$	0.193
		Class B: $x_{i11} = 1$	0.513
		Class C: $x_{i10}, x_{i11} = 0$	0.284

of claims of both third-party liabilities (Y_{i1}) and car damages (Y_{i2}), which are the two types of coverage for each policyholder. Apart from the claim counts, we have access to various policyholder's characteristics (age, claim history, and location) and vehicle's characteristics (age, fuel type, and brand) that are useful for us to understand the risk profiles of each policyholder.

The summary statistics of the response variables $\boldsymbol{Y}_i$ and the covariates $\boldsymbol{x}_i$ are displayed in Tables 6 and 7. Since the covariates lead to heterogeneities among policyholders, we compute the fitted counts displayed in Table 7 as the total of all individual policyholders' marginal probabilities. For the response variables, there exist several heterogeneities between the two claim types. The average number of claims associated with car damages is much higher than that associated with third-party liabilities. Although both claim types exhibit over-dispersions, the dispersion ratio of Y_{i2} (1.907) is much higher than that of Y_{i1} (1.111). On the other hand, Y_{i1} has much higher skewness and kurtosis than Y_{i2}, indicating that Y_{i1} potentially has a heavier tail. For the covariates, the integer ages (in years) of the policyholder and the vehicle are captured by discrete variables x_{i1} and x_{i2}, respectively. x_{i3} is a binary variable indicating the energy source of the vehicle (diesel or gasoline). The power and size of a diesel vehicle is usually larger than that of a gasoline vehicle. The three possible contract statuses (new contract and renewal contract with or without claim

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

TABLE 7

SUMMARY OF RESPONSE AND GOODNESS-OF-FIT OF MARGINAL MODELS.

		Fitted					Fitted		
Y_{i1}	Empirical	NB GLM	ZINB GLM	EC-LRMoE	Y_{i2}	Empirical	NB GLM	ZINB GLM	EC-LRMoE
0	16,971	16,975.06	16,976.66	16,965.19	0	14182	14,177.32	14,205.60	14,188.88
1	991	972.64	969.88	1001.73	1	2499	2498.57	2386.71	2484.87
2	48	65.90	66.81	40.75	2	752	810.45	883.92	777.23
3	3	4.95	5.14	7.31	3	359	307.02	333.24	317.83
4	5	0.41	0.45	2.82	4	129	125.77	127.56	155.43
5+	1	0.04	0.05	1.20	5	66	54.19	49.51	64.01
					6	22	24.22	19.46	22.04
					7	7	11.15	7.74	6.52
					8+	3	10.30	5.26	2.19
χ^2		**81.31**	**70.61**	**5.66**	χ^2		**22.59**	**33.88**	**11.13**
loglik		−4224.94	−4213.99	−4208.77	loglik		−13,279.18	−13,204.95	−13,178.68
mean	0.062	0.062	0.062	0.062	mean	0.340	0.340	0.340	0.340
% diff		0.011%	−0.001%	0.161%	% diff		−0.011%	−0.037%	−0.007%
variance	0.069	0.068	0.068	0.069	variance	0.649	0.669	0.644	0.650
% diff		−2.148%	−1.857%	−0.426%	% diff		3.077%	−0.805%	0.221%
skewness	5.084	4.522	4.544	5.096	skewness	3.265	3.672	3.305	3.261
% diff		−11.049%	−10.615%	0.244%	% diff		12.451%	1.215%	−0.132%
kurtosis	40.248	26.386	26.755	41.938	kurtosis	16.509	23.063	17.988	16.399
% diff		−34.441%	−33.526%	4.199%	% diff		39.702%	8.962%	−0.668%

records last year) are captured in variates x_{i4} and x_{i5}. x_{i6} to x_{i9} classify which major geographical region (Regions I to IV and the capital) each policyholder belongs to. Based on the insurer's historical pricing information, the car brands are classified into Classes A, B, and C. Class A corresponds to "good" brands that are expected to have less claims, and vice versa for Class C.

Remark 6.1. *In practice, it is crucial to consider covariate selections for regressions because we may collect a large amount of policyholders' information where not all is useful. In this real data analysis, we have included all of the above-mentioned variables for regressions, and we will show in Section 6.2 that all variables are significantly impactful to the risk levels of policyholders. Also, since the main focus of this paper is to showcase a novel EC-LRMoE as a fully flexible regression model useful in general insurance, we do not investigate on covariate selection issues. Instead, we put it as an important future research direction, see Section 7 for more discussions.*

The challenges of modeling this insurance data set cannot be undermined. We try to fit the marginal of the data using the NB GLM model proposed by Shi and Valdez (2014), which provides a good fit to their insurance count data set. Such model is in the NB-II form with density

$$f^{NB}(y; \boldsymbol{x}_i, \boldsymbol{\beta}, m) = \binom{y+m-1}{y} \left(\frac{m}{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}+m}\right)^m \left(\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}+m}\right)^y, \qquad y = 0, 1, 2, \ldots \tag{6.1}$$

The goodness-of-fit results exhibited in Table 7 show that the NB GLM model provides poor fittings to both marginals. The tail heaviness is significantly underestimated for Y_{i1} and overestimated for Y_{i2}, leading to severe mismatches of the higher moments between the fitted models and the empirical distributions. The χ^2 statistics are also very high for both marginals. For the main reason of poor fits, one may suggest that NB distributions may not be sufficiently flexible to capture the excessive zeros appearing in our data set. Thus, we further fit a zero-inflated NB (ZINB) GLM with density

$$f^{ZINB}(y; \boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, m) = (1 - p(\boldsymbol{x}_i; \boldsymbol{\alpha}))\, 1\{y = 0\} + p(\boldsymbol{x}_i; \boldsymbol{\alpha}) f^{NB}(y; \boldsymbol{x}_i, \boldsymbol{\beta}, m), \quad y = 0, 1, 2, \ldots, \tag{6.2}$$

where $p(\boldsymbol{x}_i; \boldsymbol{\alpha}) = 1/(1 + \exp\{\boldsymbol{x}_i^T\boldsymbol{\alpha}\})$ represents a logistic regression on the probability of extra zeros. ZINB GLM better captures the tail of Y_{i2}, having improvements in matching the moments. However, the overall fitting is still unsatisfactory because of the high χ^2 statistics and little improvements in fitting the distribution of Y_{i1}. Therefore, the problems are much more complicated than modeling excessive zeros, providing motivations to adopt a highly flexible, data-driven model such as ours for data fitting.

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

TABLE 8

MODEL SELECTION TABLE.

Number of components	2	3	4	5	6
Log-likelihood	−17,123.55	−17,007.63	−16,966.05	−16,939.34	−16,925.21
AIC	34,287.10	34,087.26	34,036.10	**34,014.68**	34,018.42
BIC	34,443.08	**34,368.03**	34,441.66	34,545.02	34,673.55

6.2. Estimation results

We fit the bivariate EC-LRMoE to the insurance claim count data set described in the previous subsection. Using the proposed ECM algorithm, three and five components are detected in the fitted model using the BIC and the AIC criteria, respectively (Table 8). BIC is better in identifying the true model (Kuha, 2004), but this is not the aim in practice since the true model is never known for any real data sets. However, AIC models better predict future data. Hence, we choose only to present the fitting results of the AIC model for conciseness purpose. The fitted model parameters are displayed in Table 9 and the left panel of Table 10. Because of the response marginalization property for the class of LRMoE (Proposition 4.1 of Fung *et al.* (2019a)), we are able to display the marginal goodness-of-fit of the proposed model in Table 7. With much lower χ^2 statistics and tighter matches of the higher moments to the empirical data, the proposed model adequately caters for the complicated distributional characteristics of the data set. Also, the Kendall's tau for the fitted model is very close to that of the empirical data (Table 11), showing that the dependence structure of the data is well captured by the proposed model.

Remark 6.2. *One may be concerned about the number of parameters for the fitted EC-LRMoE, because it has a much more complicated structural form compared to the ZINB GLM. Based on the AIC, the EC-LRMoE involves 68 parameters in total, capturing not only both marginal distributions, but also for the dependence structure between two business lines. The ZINB GLM, on the other hand, involves 25 parameters per marginal without capturing the dependence. Hence, the number of parameters involved in the EC-LRMoE is not that large and one should not expect there is an over-fitting problem.*

The fitted model may be interpreted as follows. Each policyholder is classified in one of the five possible homogeneous subgroups. From the subgroup conditional mean $E[Y_{ik}|Z_{ij} = 1]$ displayed in Table 9, we can see that subgroups (components) 1 and 3 represent the safest and the most dangerous driver group, respectively. Some subgroups (e.g., subgroup 5) are relatively more prone to third-party liabilities while some (e.g., subgroup 2) are relatively more prone to car damages. The last column of Table 9 suggests that most policyholders are safe drivers. The regression coefficients in the left panel of Table 10

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

TABLE 9

ESTIMATES OF THE SCALE PARAMETERS, THE SHAPE PARAMETERS, THE SUBGROUP CONDITIONAL MEAN, AND THE COMPONENT WEIGHTS.

	$\hat{\boldsymbol{m}}$		$\hat{\boldsymbol{\beta}}$		$E[Y_{ik} \mid Z_{ij} = 1]$		
	$k=1$	$k=2$	$k=1$	$k=2$	$k=1$	$k=2$	$P(Z_{ij}=1)$
$j=1$	15	3	3.299	0.732	0.000	0.038	0.743
	(12.975,15)	(3,3)	(2.550,7.123)	(0.374,0.799)	(0.000,0.008)	(0.007,0.048)	(0.689,0.756)
$j=2$	3	1	1.470	2.042	0.188	2.042	0.087
	(1,3)	(1,1)	(0.207,1.587)	(1.839,2.229)	(0.147,0.236)	(1.839,2.229)	(0.073,0.107)
$j=3$	1	1	1.585	2.487	1.585	2.487	0.003
	(1,3)	(1,15)	(0.807,7.661)	(1.899,27.036)	(0.752,2.742)	(1.449,3.128)	(0.001,0.010)
$j=4$	2	4	0.650	5.232	0.143	0.933	0.119
	(2,3)	(4,4)	(0.506,1.239)	(4.653,5.570)	(0.100,0.183)	(0.786,1.018)	(0.098,0.166)
$j=5$	6	3	5.665	1.902	0.513	0.310	0.047
	(4,8)	(3,3)	(2.621,7.648)	(1.415,2.465)	(0.191,0.578)	(0.172,0.420)	(0.037,0.076)

The brackets are the 95% confidence intervals. The component weights $P(Z_{ij}=1)$ are estimated using Equation (5.2) with $n_{sim} = n = 18,019$.

TABLE 10

ESTIMATES OF THE REGRESSION PARAMETERS.

$\hat{\boldsymbol{\alpha}}$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$\hat{\boldsymbol{\alpha}}^*$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
x_{i0}	**4.381** (3.556, 6.718)	**2.266** (1.241, 4.624)	−0.923 (−77.382, 1.240)	**3.844** (2.798, 6.284)	0.000	x^*_{i0}	0.000	**−1.440** (−2.035, −0.836)	−2.347 (−35.244, 0.077)	−0.238 (−0.705, 0.217)	**−3.529** (−4.876, −2.653)
x_{i1}	−0.001 (−0.019, 0.013)	−0.011 (−0.030, 0.006)	−0.021 (−0.081, 0.042)	−0.011 (−0.032, 0.006)	0.000	x^*_{i1}	0.000	**−0.009** (−0.019, −0.001)	−0.019 (−0.082, 0.045)	**−0.010** (−0.017, −0.002)	0.001 (−0.013, 0.019)
x_{i2}	**−0.134** (−0.294, −0.086)	**−0.120** (−0.276, −0.065)	**−0.359** (−0.595, −0.231)	**−0.270** (−0.417, −0.205)	0.000	x^*_{i2}	0.000	0.014 (−0.016, 0.050)	**−0.226** (−0.405, −0.091)	**−0.136** (−0.198, −0.086)	**0.134** (0.086, 0.294)
x_{i3}	−0.186 (−0.557, 0.136)	0.066 (−0.330, 0.422)	0.618 (−0.537, 42.641)	−0.015 (−0.444, 0.370)	0.000	x^*_{i3}	0.000	**−0.252** (−0.420, −0.058)	−0.804 (−42.778, 0.221)	−0.171 (−0.351, 0.004)	−0.186 (−0.557, 0.136)
x_{i4}	**−0.996** (−2.143, −0.691)	**−1.027** (−2.022, −0.635)	0.973 (−0.619, 43.352)	−0.172 (−1.408, 0.237)	0.000	x^*_{i4}	0.000	0.082 (−0.174, 0.332)	−0.839 (−43.774, 0.532)	−0.090 (−0.337, 0.224)	**−0.842** (−1.955, −0.530)
x_{i5}	**−0.842** (−1.955, −0.530)	**−0.924** (−1.992, −0.562)	−0.003 (−1.911, 42.979)	**−0.752** (−2.148, −0.344)	0.000	x^*_{i5}	0.000	0.051 (−0.340, 0.400)	1.130 (−0.176, 3.195)	**0.734** (0.456, 1.078)	0.154 (−0.261, 0.534)

x_{i6}	−0.014	0.222	1.262	−0.103	0.000	x^*_{i6}	0.000	−0.270	−0.039	−0.126	0.190
	(−0.590,	(−0.377,	(−1.053,	(−0.784,				(−0.632,	(−1.278,	(−0.409,	(−0.584,
	0.677)	1.131)	45.810)	0.782)				0.111)	2.182)	0.266)	0.940)
x_{i7}	−0.274	0.423	0.956	−0.530	0.000	x^*_{i7}	0.000	0.190	−0.086	−0.294	0.449
	(−0.903,	(−0.248,	(−37.977,	(−1.358,				(−0.109,	(−46.557,	(−0.727,	(−0.172,
	0.210)	1.162)	44.650)	0.158)				0.586)	2.739)	0.104)	1.266)
x_{i8}	0.176	**0.682**	1.491	0.213	0.000	x^*_{i8}	0.000	**−0.506**	−1.315	−0.037	0.176
	(−0.472,	(0.022,	(−1.377,	(−0.620,				(−0.988,	(−46.724,	(−0.403,	(−0.472,
	0.989)	1.523)	47.107)	1.181)				−0.102)	1.389)	0.347)	0.989)
x_{i9}	−0.334	0.437	−0.283	−0.362	0.000	x^*_{i9}	0.000	0.265	−1.264	−0.065	**0.509**
	(−0.868,	(−0.134,	(−43.564,	(−0.974,				(−0.012,	(−46.908,	(−0.355,	(0.043,
	0.055)	1.075)	43.856)	0.204)				0.592)	2.109)	0.285)	1.174)
x_{i10}	0.086	**−0.594**	0.444	−0.195	0.000	x^*_{i10}	0.000	**−0.680**	0.358	**−0.281**	−0.086
	(−0.332,	(−1.048,	(−1.174,	(−0.722,				(−0.962,	(−1.224,	(−0.557,	(−0.509,
	0.509)	−0.021)	2.303)	0.347)				−0.409)	1.934)	−0.033)	0.332)
x_{i11}	−0.005	−0.339	−0.390	−0.333	0.000	x^*_{i11}	0.000	**−0.334**	−0.385	**−0.328**	0.005
	(−0.360,	(−0.717,	(−30.225,	(−0.712,				(−0.532,	(−30.219,	(−0.501,	(−0.367,
	0.367)	0.061)	1.392)	0.163)				−0.139)	1.326)	−0.105)	0.360)

Left panel and right panel are the parameters before and after transformations, respectively. The brackets are the 95% confidence intervals.

TABLE 11

KENDALL'S τ FOR THE FITTED MODEL VERSUS THE EMPIRICAL DATA.

τ	empirical data	fitted model
with covariates influence	0.241	0.240
without covariates influence	0.315	0.315

determine how the covariates affect policyholder's subgroup assignment. For example, for each 1-year increase in car age (x_{i2}), the probability ratio for a policyholder to be classified into subgroup 1 relative to subgroup 5 (we call it the "baseline subgroup") will be decreased by $1 - e^{-0.134} = 12.5\%$. Also, new contracts ($x_{i5} = 1$) or past claim records ($x_{i4} = 1$) result to a lower chance for being classified into subgroup 1 relative to subgroup 5.

To understand how precise the estimated model parameters are, we take parameter uncertainties into account and construct a confidence interval (CI) on each of the estimated parameters $\hat{\boldsymbol{\Phi}} := (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{m}})$. This can be done through bootstrapping. For each run b, we resample $n = 18,019$ independent observations from the original data set $\{(\boldsymbol{Y}_i, \boldsymbol{x}_i); i = 1, \ldots, n\}$ with replacement. Then, we refit the proposed model to the resampled data using the $\hat{\boldsymbol{\Phi}}$ as the initialization to obtain the refitted model parameters $\tilde{\boldsymbol{\Phi}}^b$. Perform and repeat the whole procedures described above for $b = 1, 2, \ldots, B$, where we choose $B = 400$. Using the refitted parameters $\{\tilde{\boldsymbol{\Phi}}^b; b = 1, \ldots, B\}$, not only we can construct the CIs of the estimated parameters (Table 9 and 10), but also we can calculate the CIs of any other quantities that can be derived from the fitted model, such as the subgroup conditional mean and the component weights.

In Table 9, it is observed that the CIs for $\hat{\boldsymbol{\beta}}$ are very wide and unstable for some subgroups and marginals (e.g., $j = 3, k = 2$), because β_{jk} is greatly affected by the changes of m_{jk} for a given subgroup conditional mean. On the other hand, the CIs for the subgroup conditional mean and the component weights are much more stable. The instabilities of the CIs are also found for $\hat{\boldsymbol{\alpha}}$. The left panel of Table 10 shows that several very positive and negative values appear in the CIs for subgroup 3, because only very few (around $0.003 \times 18,019 = 52.3$) policyholders are classified into subgroup 3. This makes us unconfident on some of the estimated regression coefficients. Another issue we observed is that the regression coefficients for some covariates (e.g., x_{i1} and x_{i3}) are not significantly different from zero for any subgroups ($j = 1, 2, 3, 4$). This may lead to a misconception that these covariates would not have significant impacts on the risk level of the policyholders. We should note that these regression coefficients govern the probability of subgroup assignments relative to the baseline subgroup. The subgroup assignments relative to other subgroups, however, can still be strongly influenced by these covariates.

To gain a better understanding of the above issues, we perform some transformations on the covariates and the regression coefficients to see how they will

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

affect the CIs and the significance of the estimated regression coefficients. The transformations are as follows:

- Redefining some categorical covariates: We set high/above-average risk categories (diesel for car fuel, new contract for policyholder's history, Region III for geographical location, and Class C for car brand class) as reference/control categories. Let $\boldsymbol{x}_i^*$ be the transformed covariates. Changing the reference categories, we can define: $x_{ip}^* = x_{ip}$ for $p = 1, 2, 10, 11$; $x_{i3}^* = 1 - x_{i3}$ represents diesel vehicles; $x_{i4}^* = 1$ and $x_{i5}^* = 1$ correspond to renewal without and with claims last year, respectively; $x_{i6}^* = 1$, $x_{i7}^* = 1$, $x_{i8}^* = 1$; and $x_{i9}^* = 1$ are Regions I, II, IV, and capital, respectively.
- Changing the baseline subgroup as subgroup 1: Subgroup 1 is the lowest risk group that contains a majority of policyholders. Making it as a baseline subgroup can enhance the model interpretability, because the regression coefficients show directional relationships between the covariates and the expected claim counts. A positive regression coefficient means a higher chance for a policyholder to be classified into a non-baseline higher risk subgroup if the value of the corresponding covariate is large, and vice versa.

To find the estimated regression coefficients after both transformations $\hat{\boldsymbol{\alpha}}^*$, model refitting is not needed. Instead, only linear transformations of $\hat{\boldsymbol{\alpha}}$ are required. Denote $\hat{\boldsymbol{\alpha}}^{(t)}$ as the regression coefficients just after categorical covariates transformations. Since $\boldsymbol{x}_i^*$ can be easily expressed in terms of linear combinations of $\boldsymbol{x}_i$, the coefficients of $\boldsymbol{x}_i^*$ (i.e., $\hat{\boldsymbol{\alpha}}^{(t)}$) can also be written as linear transformations of the coefficients of $\boldsymbol{x}$ (i.e., $\hat{\boldsymbol{\alpha}}$). For baseline subgroup transformation, the resulting regression coefficients can be easily obtained as $\hat{\boldsymbol{\alpha}}^* = \hat{\boldsymbol{\alpha}}^{(t)} - \mathbf{1}\hat{\boldsymbol{\alpha}}_5^{(t)T}$ because of the translational invariance property of the LRMoE, where $\mathbf{1}$ is a five-element column vector.

The estimated regression parameters after the transformations are displayed in the right panel of Table 10. The instability of the CIs for subgroup 3 is reduced as no more very positive values appear. It indicates that no policyholders are classified into such a highest risk subgroup with (almost) certainty. However, several very negative values still exist, so we still should not rule out a chance that policyholders with certain risk characteristics (almost) never belong to subgroup 3. Contrary to the original regression coefficients, the transformed regression coefficients for covariates x_{i1}^* and x_{i3}^* are significantly different from zero in some subgroups. Such negative coefficients suggest that older drivers and gasoline vehicle drivers are more likely to be safer drivers who file less claims on average.

6.3. Model visualization

6.3.1. *Influences of the covariates*

Interpreting the regression coefficients $\hat{\boldsymbol{\alpha}}$ directly may help understand how the covariates affect the probabilities of classifying a policyholder into various

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

subgroups. However, to gain more insights on how the policyholder's characteristics affect their risk level, it is desirable to visualize the impacts of each individual covariate on the expected claim counts for each claim type.

To do so, we first plot the mean number of claims against the values of a particular covariate (e.g., policyholder's age). Note that multiple covariates are involved in the fitted model so this task is not trivial and we need to aggregate the effects of the other covariates. We make use of Equation (2.7) for the calculation of the mean and the remaining problem is to estimate the covariate-marginalized weights $\tilde{\pi}_j(\boldsymbol{x}^c;\boldsymbol{\alpha})$. Two simple methods are introduced to approximate $\tilde{\pi}_j(\boldsymbol{x}^c;\boldsymbol{\alpha})$ and hence to estimate the mean number of claims conditioned only on covariate p. The first one is a pure nonparametric approach:

$$\bar{Y}^{\text{npar}}_{k;p}(s) = \sum_{j=1}^{g} \frac{1}{N^{\text{npar}}_p(s)} \sum_{i\le n:\ x_{ip}=s} \pi(\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}})\mu_k^{(j)}, \tag{6.3}$$

where $N^{\text{npar}}_p(s)$ is the number of $i \le n$ such that $x_{ip} = s$ and $\mu_k^{(j)} = E[Y_{ik}|Z_{ij} = 1]$ is independent of i. This method yields an unbiased estimate of the mean claim count. However, $N^{\text{npar}}_p(s)$ can be small for certain p and s (e.g., since only nine policyholders are 85 years old, we have $N^{\text{npar}}_1(85) = 9$), causing the estimate sometimes rather unstable. Alternatively, we demonstrate the partial dependence plot that is commonly used in machine learning (Friedman, 2001). It assumes independence between covariate p and the other covariates:

$$\bar{Y}^{\text{indep}}_{k;p}(s) = \sum_{j=1}^{g} \frac{1}{n} \sum_{i\le n} \pi(\tilde{\boldsymbol{x}}_i;\hat{\boldsymbol{\alpha}})\mu_k^{(j)}, \tag{6.4}$$

where $\tilde{x}_{ip'} = x_{ip'}$ for $p' \ne p$ and $\tilde{x}_{ip} = s$. This method uses all policyholders' features, making the estimation smooth and stable. If the independence assumption does not hold, then this approach will introduce some biases to the estimates.

The impacts of policyholder's age on the expected claim counts using the pure nonparametric plots and the partial dependence plots are displayed in Figures 4 and 5, respectively. The gray regions are the 95% CIs considering parameter uncertainties. Both kinds of plots suggest that older drivers are expected to have less claims for both claim types, while the effects of policyholder's age to car damages are more significant than to third-party liabilities. As expected, the fitted curves from the partial dependence plots are smoother than that from the nonparametric plots. Empirical studies show that the correlations between policyholder's age and the other covariates are (very) weak, so the trends obtained by the two methods are similar and the biases of the partial dependence plots are minimal.

Next, we analyze the interactive effects between car age and past claim record. Similar to Equations (6.3) and (6.4), we compute the mean number

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

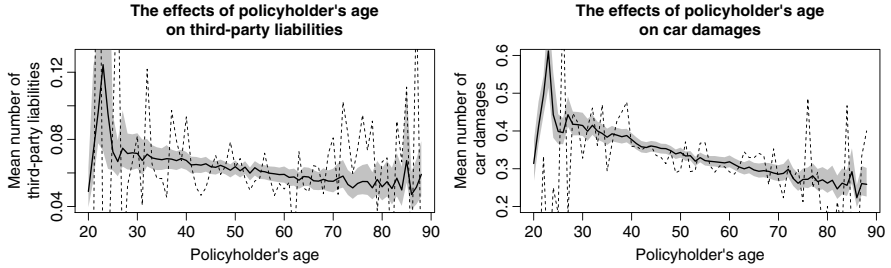

FIGURE 4: Expected number of claims versus policyholder's age (x_1) evaluated using the nonparametric approach. Dotted curve: Empirical mean pattern; Solid curve: Fitted/estimated pattern using the proposed model; Colored region: 95% confidence intervals of the estimations.

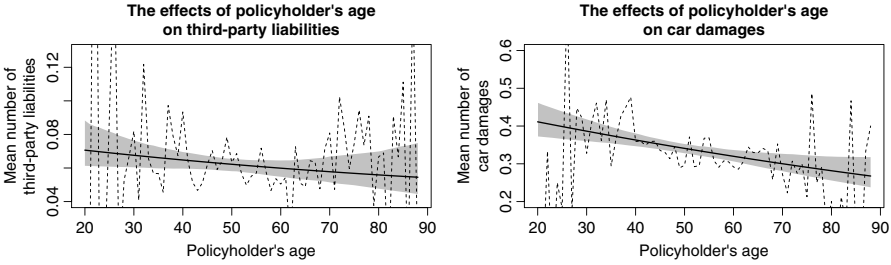

FIGURE 5: Expected number of claims versus policyholder's age assuming the independence between x_1 and other covariates.

of claims conditioned on covariates p_1 and p_2 for the nonparametric plots and the partial dependence plots:

$$\bar{Y}^{\text{npar}}_{k;\,p_1,p_2}(s_1,s_2)=\sum_{j=1}^{g}\frac{1}{N^{\text{npar}}_{p_1,p_2}(s_1,s_2)}\sum_{\substack{i\le n:\\ x_{ip_1}=s_1,x_{ip_2}=s_2}}\pi(\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}})\mu_k^{(j)} \tag{6.5}$$

$$\bar{Y}^{\text{indep}}_{k;\,p_1,p_2}(s_1,s_2)=\sum_{j=1}^{g}\frac{1}{N^{\text{indep}}_{p_2}(s_2)}\sum_{i\le n:\ x_{ip_2}=s_2}\pi(\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}})\mu_k^{(j)} \tag{6.6}$$

where $N^{\text{npar}}_{p_1,p_2}(s_1,s_2)$ is the number of $i\le n$ such that $x_{ip_1}=s_1$ and $x_{ip_2}=s_2$, while $N^{\text{indep}}_{p_2}(s_2)$ is the number of $i\le n$ satisfying $x_{ip_2}=s_2$. Note that under the partial dependence plots, we assume that covariate p_1 is independent of other covariates except covariate p_2. In this analysis, we have $p_1=2$ (car age) and $p_2=4$ (past claim record). Empirical analysis shows that the correlations between car age and the other covariates (except past claim record) are (very) weak.

The visualizations are displayed in Figures 6 and 7. For third-party liabilities, whether or not the policyholder has past claim records, the expected claim counts versus the car age exhibit "U"-shapes. It means that medium-aged (4–7 years) vehicles are less involved in claims associated with third-party liabilities. For car damages, the expected claim counts decrease significantly when the car

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

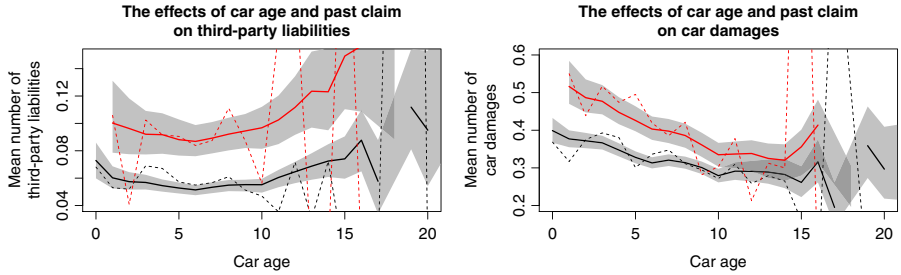

FIGURE 6: Expected number of claims versus car age (x_2) and past claim (x_4) using the nonparametric approach. Red and black dotted curves are, respectively, the empirical mean patterns with and without claim records last year (i.e., $x_4 = 1$ or $x_4 = 0$). Red and black solid curves are the corresponding fitted patterns.

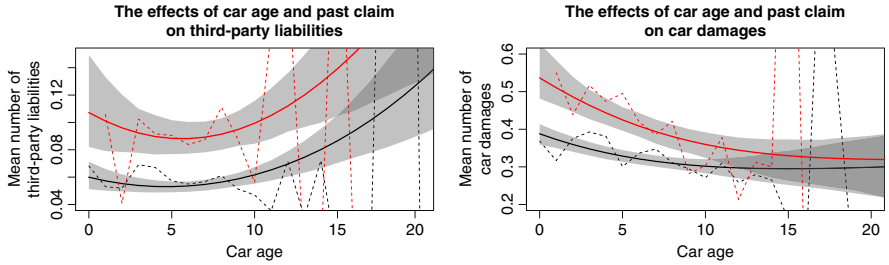

FIGURE 7: Expected number of claims versus car age and past claim assuming that x_2 is independent of all other covariates except x_4.

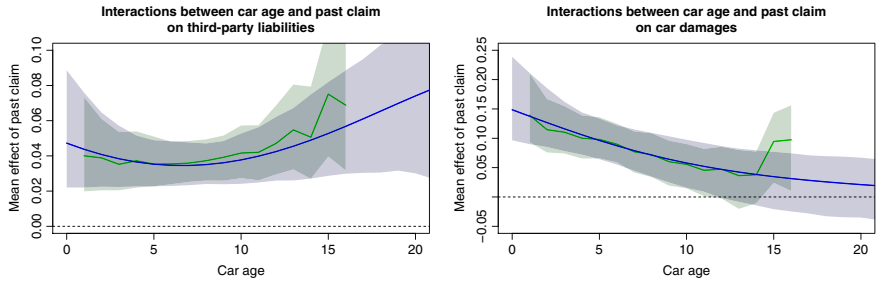

FIGURE 8: Interactions between car age and past claim. The effect of past claim (difference between the two curves in Figures 6 or 7) is plotted against the car age. The rough and smooth curves are from the nonparametric plots and the partial dependence plots, respectively.

age increases from 0 to 10, and then become flat after 10 years. Figure 8 shows the effect of past claim record to the number of claims, which is plotted against the car age. We can see that for newer cars, past claim record will lead to at least 0.1 more claims associated with car damages on average per policyholder. Such an effect, however, is small and insignificant for older cars (>12 years). Hence, interactions exist between these two covariates.

Using the nonparametric approach in Equation (6.3), we may also evaluate the effects of the categorical covariates on the claim counts. Note that some

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

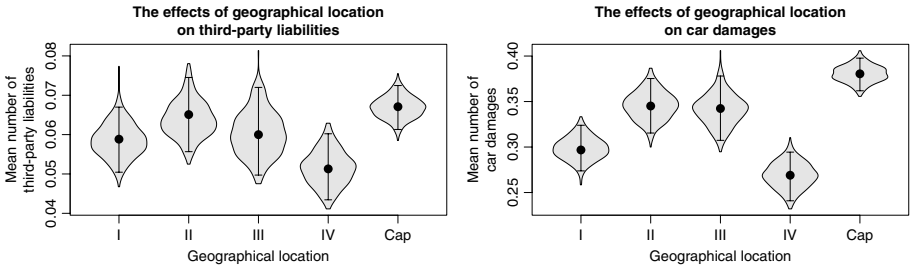

FIGURE 9: Violin plot with 95% confidence interval bands: Expected number of claims under various geographical locations.

categorical variables (e.g., x_{i6} to x_{i9} for geographical locations) are obviously (negatively) related to each other, so the use of partial dependence plots is inappropriate. Figure 9 contains the violin plots that show the expected number of claims under various geographical locations. We find that policyholders in the capital region are more risky and policyholders in Region IV are less risky for both claim types. Also, Table 12 confirms the statistical significance that geographical locations affect the risk level of the policyholders. Similar analyses can be performed for other categorical variables but these are not fully presented in this paper for preciseness purpose. Instead, the main results that are highly statistically significant are summarized as follows:

- Compared to gasoline vehicles, diesel vehicles result to 0.014 more third-party liabilities claims and 0.063 more car damages claims per policyholder.
- Renewal contracts with past claim records are expected to file the largest number of claims per policyholder for both claim types, followed by new contracts, and then renewal contracts without any claims last year.
- Compared to Class C for car brands, policyholders driving with cars classified as Class A or B are expected to file less claims for both claim types. The difference between Class A and Class B is significant only for the number of claims associated with car damages.

6.3.2. *Subgroup probabilities for individual policyholders*

One interesting insight of the proposed model is that it can generate the prior and posterior probabilities that a policyholder belongs to a certain latent subgroup. Similar to Section 4, we denote $Z_{ij} = 1$ if the ith policyholder belongs to the jth subgroup and $Z_{ij} = 0$ otherwise. Then, the prior and posterior probabilities are, respectively, given by $P(Z_{ij} = 1; \boldsymbol{x}_i, \hat{\boldsymbol{\Phi}}) = \pi_j(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}})$ and $P(Z_{ij} = 1; \boldsymbol{y}_i, \boldsymbol{x}_i, \hat{\boldsymbol{\Phi}}) = \pi_j(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) \prod_{k=1}^{K} f(y_{ik}; \hat{\boldsymbol{\theta}}_{jk}) / \sum_{j'=1}^{g} \pi_{j'}(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) \prod_{k=1}^{K} f(y_{ik}; \hat{\boldsymbol{\theta}}_{j'k})$.

We select three representative policyholders (exhibited in Table 13) from the real data set and compare the prior and posterior probabilities. Policyholder A has a lot of undesirable risk characteristics (e.g., young driver, new diesel vehicle under a poor car class, claim record last year) but no claims

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

TABLE 12

THE EFFECTS OF GEOGRAPHICAL LOCATION ON THE EXPECTED CLAIM COUNTS PRESENTED IN MATRICES.

Y_{i1}	I	II	III	IV	Cap	Y_{i2}	I	II	III	IV	Cap
I		0.006	0.001	−0.008	0.008	I		0.048	0.046	−0.028	0.084
II	0.395		−0.005	−0.014	0.002	II	**0.015**		−0.003	−0.076	0.035
III	0.785	0.665		−0.009	0.007	III	**0.035**	0.860		−0.073	0.038
IV	0.265	**0.040**	0.215		0.016	IV	0.100	**0.000**	**0.000**		0.111
Cap	0.100	0.580	0.290	**0.000**		Cap	**0.000**	0.055	**0.040**	**0.000**	

Upper triangle: The differences of the mean claim counts between two locations. Lower triangle: The two-sided p-values testing whether such differences significantly differ from zero.

TABLE 13

THREE SELECTED POLICYHOLDERS TO BE CONSIDERED FOR THE CALCULATIONS OF SUBGROUP PROBABILITIES.

Policyholder	y_1	y_2	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
A	0	0	36	1	1	1	0	0	0	1	0	0	0
B	0	1	59	7	1	0	1	0	0	1	0	0	1
C	1	2	72	7	0	0	0	1	0	0	0	1	0

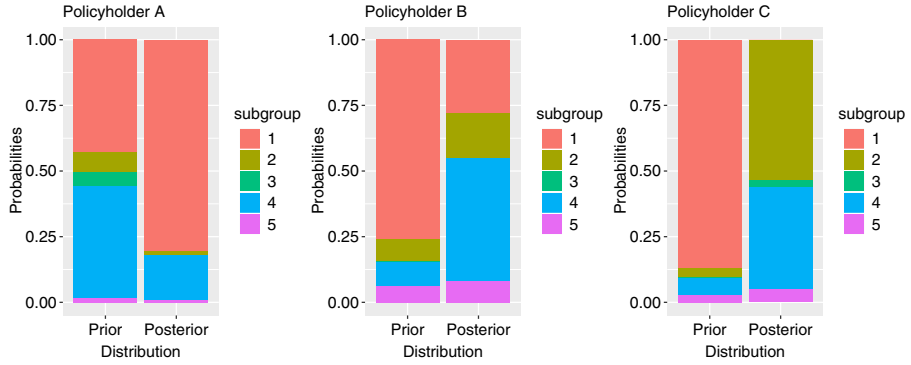

FIGURE 10: The prior and posterior subgroup probabilities for the three selected policyholders.

are observed during the contract period. Policyholder B has an average risk profile. The risk characteristics of Policyholder C is relatively desirable but eventually three claims occurred during the contract period.

The subgroup probabilities are visualized in Figure 10. As expected, the prior probability that Policyholder A (C) belongs to subgroup 1, a very low risk subgroup, is relatively low (high). Given that no claims are observed for Policyholder A, the posterior probability that the policyholder belongs to subgroup 1 is much higher than the prior probability. On the other hand, since a large number of claims are observed for policyholder C, the posterior subgroup 1 probability reduces to almost zero, and there is now a substantial probability that policyholder C belongs to the most dangerous driver group (subgroup 3).

6.4. Predictive applications

Insurance ratemaking for bundled contracts is one of the important applications of multivariate regression models, where the insurer provides multiple types of coverage in a single contract. In this subsection, we aim to study the total claim count variable $L_i = Y_{i1} + Y_{i2}$, which is the basis for premium calculations, for hypothetical risk profile i. Similar studies have been performed in Bermúdez (2009), Bermúdez and Karlis (2011), and Shi and Valdez (2014) using the multivariate regression models they proposed.

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

TABLE 14

SIX DIFFERENT HYPOTHETICAL RISK PROFILES TO BE CONSIDERED.

Profile	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
Best	60	8	0	0	0	0	0	0	0	1	0
Good	40	5	0	0	0	1	0	0	0	1	0
Average 1	40	6	0	0	0	0	0	0	1	0	1
Average 2	80	10	1	0	0	0	0	1	0	0	0
Bad	30	0	1	0	1	0	1	0	0	0	1
Worst	30	1	1	1	0	0	0	0	1	0	0

TABLE 15

MEAN AND VARIANCE OF THE NUMBER OF CLAIMS BY RISK PROFILE.

	Mean and variance of claims					
Profile	$E[Y_{i1}]$	$\mathrm{Var}[Y_{i1}]$	$E[Y_{i2}]$	$\mathrm{Var}[Y_{i2}]$	$E[L_i]$	$\mathrm{Var}[L_i]$
Best	0.033	0.035	0.179	0.316	0.212	0.396
Good	0.040	0.048	0.246	0.440	0.286	0.565
Average 1	0.050	0.051	0.347	0.710	0.397	0.847
Average 2	0.053	0.054	0.331	0.710	0.383	0.850
Bad	0.098	0.147	0.482	0.873	0.580	1.237
Worst	0.121	0.152	0.677	0.909	0.798	1.227

For demonstration purposes, we construct six hypothetical risk profiles exhibited in Table 14. Depending on the risk levels, they are named as Best, Good, Average 1, Average 2, Bad, and Worst. For example, a policyholder with the "Best" profile is 60 years old, drives a 8-year-old gasoline vehicle that is classified as Class A, signs a renewal contract without claims last year and lives in Region IV. These are all favorable characteristics that lead to a lower risk level. Also, we have constructed two average profiles, each of them consist of a mixture of desirable and undesirable risk characteristics. The main difference between the two types of profiles is that the features of an Average 1 profile are more common than that of an Average 2 profile. In other words, very few policyholders share similar characteristics as an Average 2 profile (e.g., 80-year-old policyholders are very rare; relatively few policyholders live in Region III). We will show that the fair premium charged to policyholders can be affected by such a difference.

The summary statistics for L_i, Y_{i1}, and Y_{i2} are displayed in Table 15. As expected, lower risk profiles make less claims on average, and vice versa. The expected claim counts for the worst risk profile is almost four times as that for the best risk profile. Also, we observe that the variance of the number of claims generally increases with the mean. However, the proposed model shows that

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

TABLE 16

QUANTILE PREMIUM, STANDARD DEVIATION PREMIUM, AND SD PREMIUM (ASSUMING INDEPENDENCE BETWEEN TWO TYPES OF COVERAGE) BY RISK PROFILE.

	Quantile Premium		SD Premium		SD Prem (Indep)	
Profile	75%	95%	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 0.1$	$\gamma = 0.5$
Best	0.219	0.232	0.275	0.526	0.271	0.508
Good	0.302	0.321	0.361	0.662	0.356	0.635
Average 1	0.411	0.427	0.489	0.857	0.484	0.833
Average 2	0.407	0.452	0.476	0.844	0.471	0.820
Bad	0.595	0.680	0.691	1.136	0.681	1.085
Worst	0.835	0.921	0.908	1.352	0.901	1.313

such a relationship is nonlinear. The dispersion ratio of the number of third-party liabilities claims is the greatest for profile ["Bad,"], while the dispersion ratio of the claim counts on car damages is the largest for profile ["Average 2."]. This phenomenon is in contrast to the assumption of the NB GLM framework that the variance of marginal claim counts is linear to the mean.

The expected total number of claims $E[L_i]$ is the basis for premium calculations. However, under parameter uncertainties, $E[L_i]$ can still be random. The quantiles of the distribution of $E[L_i]$ may be viewed as the percentile premiums. The distribution of $E[L_i]$ can be easily and directly computed from the $B = 400$ sets of refitted parameters $\{\tilde{\boldsymbol{\Phi}}^b; b = 1, \ldots, B\}$, which have already been obtained by bootstrapping. The 75th and the 95th percentile premiums for each of the six risk profiles are exhibited in Table 16. One interesting observation is that although the risk levels of the two average profiles are similar (the expected total claim counts of profile "Average 2" is slightly lower than that of profile "Average 1"), the 95th percentile premium for profile "Average 2" is about 6% higher than that for profile "Average 1". As mentioned before, very few policyholders share similar risk characteristics as an Average 2 profile. Since the insurer has less relevant claim information, the proposed model tells us that the effects of parameter uncertainties are greater for such a profile, leading to greater-than-usual percentile premiums.

Table 16 also displays the premiums based on the standard deviation premium principle, which is $E[L_i] + \gamma\sqrt{Var[L_i]}$. We further compare them to the SD premiums assuming independence between two types of coverage, which are calculated as $E[L_i] + \gamma\sqrt{\text{Var}[Y_{i1}] + \text{Var}[Y_{i2}]}$. It is obvious that the SD premiums will be underestimated assuming independence between two claim types, especially if we choose a larger γ, because our fitted model identifies a positive correlation between Y_{i1} and Y_{i2}. Note that the calculations of the quantile premium and the SD premiums are based on completely different rules and concepts, so the premiums calculated by the two approaches can be quite different.

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

7. Concluding remarks

In this paper, we consider the estimation and application aspects of the EC-LRMoE model, which is regarded as a fully flexible multivariate count regression model. We first proved the identifiability property that makes the proposed model an excellent candidate for statistical inference. Then, an ECM algorithm is presented to estimate the model parameters. The steps involve either analytical formulas or low-dimensional convex/concave optimizations, so they are easily computable. The effectiveness of the proposed ECM algorithm and the flexibility of the EC-LRMoE model are verified through three simulation studies. In applications, we fit the EC-LRMoE model to a real automobile insurance data set, which possesses complicated characteristics. The EC-LRMoE model captures well the distribution, dependence, and regression structures implied by the data set. Also, the effect of the policyholder's characteristics to his/her risk level, as well as the prior and posterior probabilities that a specific policyholder belongs to a certain subgroup, can be visualized. Finally, we demonstrate the use of the model to insurance ratemaking.

The current work opens up some possible future research directions. Firstly, while this paper considers frequency expert functions for the LRMoE, it is also worthwhile to study the applications of the severity LRMoE in general insurance context, especially when the insurance claim size distributions are very heavy tailed. Secondly, as discussed in Remark 4.1, the proposed algorithm is practically feasible but is still computationally intensive. One may easily apply many existing tools, such as the stochastic EM algorithm, mini-batch computing, more advanced local searching strategies (for m_{ij} and g), and more efficient built-in functions from various statistical software, to reduce the run time significantly. Thirdly, one could perform variable selection of the proposed model, which shrinks some regression parameters in the gating functions. This could be an important issue, especially when we have a large amount of policyholders' information where not all is useful. While Fan and Li (2001) proposed the SCAD penalty function for variable selections under the linear regression model and Yin and Lin (2016) proposed the iSCAD penalty function to effectively choose the number of components in finite mixture models, it is worthwhile to extend their results in order to apply them to our LRMoE model.

References

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

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

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press

BERMÚDEZ, L. (2009). A priori ratemaking using bivariate Poisson regression models. *Insurance: Mathematics and Economics* **44**(1), 135–141.
BERMÚDEZ, L. and KARLIS, D. (2011). Bayesian multivariate Poisson models for insurance ratemaking. *Insurance: Mathematics and Economics* **48**(2), 226–236.
CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in Hidden Markov Models*. Heidelberg: Springer.
CONWAY, R. W. and MAXWELL, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering* **12**(2), 132–136.
DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.
FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**(456), 1348–1360.
FREES, E. W., LEE, G. and YANG, L. (2016). Multivariate frequency/severity regression models in insurance. *Risks* **4**(1), 4.
FRIEDMAN, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* **29**(5), 1189–1232.
FUNG, T. C., BADESCU, A. L. and LIN, X. S. (2019a). A class of mixture of experts models for general insurance: Theoretical developments. Submitted.
FUNG, T. C., BADESCU, A. L. and LIN, X. S. (2019b). Multivariate Cox hidden Markov models with an application to operational risk. *Scandinavian Actuarial Journal*, accepted.
GUI, W., HUANG, R. and LIN, X. S. (2018). Fitting the Erlang mixture model to data via a GEM-CMM algorithm. *Journal of Computational and Applied Mathematics* **343**, 189–205.
JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. and HINTON, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation* **3**(1), 79–87.
JIANG, W. and TANNER, M. A. (1999). On the identifiability of mixtures-of-experts. *Neural Networks* **12**(9), 1253–1258.
JORDAN, M. I. and JACOBS, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* **6**(2), 181–214.
KUHA, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research* **33**(2), 188–229.
LEE, S. C. K. AND LIN, X. S. (2010). Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal* **14**(1), 107–130.
LEE, S. C. K. and LIN, X. S. (2012). Modeling dependent risks with multivariate Erlang mixtures. *ASTIN Bulletin: The Journal of the IAA* **42**(1), 153–180.
LORD, D., GUIKEMA, S. D. and GEEDIPALLY, S. R. (2008). Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention* **40**(3), 1123–1134.
MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics.
MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**(2), 267–278.
SHI, P. and VALDEZ, E. A. (2014). Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics* **55**, 18–29.
TEICHER, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, **34**(4), 1265–1269.
TEICHER, H. (1967). Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics* **38**(4), 1300–1302.
WEDEL, M. AND DESARBO, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification* **12**(1), 21–55.
WINKELMANN, R. (1995). Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics* **13**(4), 467–474.
YIN, C. and LIN, X. S. (2016). Efficient estimation of Erlang mixtures using iSCAD penalty with insurance application. *ASTIN Bulletin: The Journal of the IAA* **46**(3), 779–799.

TSZ CHAI FUNG
Department of Statistical Sciences
University of Toronto
100 St George Street
Toronto, ON M5S 3G3, Canada
E-Mail: tszchai.fung@mail.utoronto.ca

ANDREI L. BADESCU (Corresponding author)
Department of Statistical Sciences
University of Toronto
100 St George Street
Toronto, ON M5S 3G3, Canada
E-Mail: badescu@utstat.toronto.edu

X. SHELDON LIN
Department of Statistical Sciences
University of Toronto
100 St George Street
Toronto, ON M5S 3G3, Canada
E-Mail: sheldon@utstat.toronto.edu

https://doi.org/10.1017/asb.2019.25 Published online by Cambridge University Press