Mixture Composite Regression Models with Multi-type Feature Selection

Tsz Chai Fung,1 George Tzougas,2 and Mario V. Wüthrich3

1Department of Risk Management and Insurance, Georgia State University, Atlanta, Georgia
2Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, United Kingdom
3RiskLab, ETH Zurich, Zurich, Switzerland

The aim of this article is to present a mixture composite regression model for claim severity modeling. Claim severity modeling poses several challenges such as multimodality, tail-heaviness, and systematic effects in data. We tackle this modeling problem by studying a mixture composite regression model for simultaneous modeling of attritional and large claims and for considering systematic effects in both the mixture components as well as the mixing probabilities. For model fitting, we present a group-fused regularization approach that allows us to select the explanatory variables that significantly impact the mixing probabilities and the different mixture components, respectively. We develop an asymptotic theory for this regularized estimation approach, and fitting is performed using a novel generalized expectation-maximization algorithm. We exemplify our approach on a real motor insurance dataset.

1. INTRODUCTION

Insurance claim severity modeling is a very challenging problem in actuarial science. Motivated by a Greek motor third party liability (MTPL) insurance dataset that is further described in Section 5.1, we observe that insurance claim severity data sets often exhibit several peculiar characteristics: Firstly, claim severity distributions are often multimodal, coming from the fact that there are systematic effects in the data due to unobserved heterogeneity and latent factors such as different claim types. Secondly, a claim severity distribution ranges over several magnitudes, from small attritional claims to large claim events, and often exhibits a heavy-tailed nature and a mismatch between body and tail behavior. Thirdly, insurance data are often accompanied by multiple types of policyholder attributes, including several continuous variables (e.g., driver’s age), ordered categorical variables (e.g., sum insured categories), and nominal categorical variables (e.g., car brand). These variables may have different explanatory powers to different parts of the severity distribution.

For insurance pricing, reserving, and risk management, it is crucial to have accurate descriptions of claim severity distributions and to understand clearly the influence of policy attributes to the claim distribution. Therefore, it is essential to devise a claim severity modeling framework that possesses all of the following features to address the aforementioned modeling challenges:

1. **Distributional multimodality**: The model must enable sufficient flexibility to capture distributional multimodality.
2. **Tail-heaviness**: The severity distribution has to be heavy-tailed in nature and it should allow for robust estimation of tail-heaviness.
3. **Covariates’ influence**: The model needs to capture the covariates’ influence on various parts of the severity distribution, including (i) the probabilities that assign observations into various clusters or nodes (clustering probabilities), (ii) systematic effects in claim severity distributions conditioned on each cluster (body part), and (iii) the tail-heaviness of the distribution (tail part).

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

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10920277.2022.2099426.
4. **Variable selection**: Realizing that not all variables are important, a variable selection strategy must also be employed to determine which variables are influential to which of the aforementioned three parts of the severity distribution. Moreover, the strategy has to be adapted to a multi-type variable setting.

There are several actuarial research works in addressing each of the above modeling requirements. To capture distributional multimodality and tail-heaviness (points 1 and 2), there are three main streams of claim severity modeling approaches as follows:

- **Finite mixture models** constitute an easily extendable model class for approximating general distribution functions in a semiparametric way and accounting for unobserved heterogeneity. Notable actuarial contributions include, for example, Lee and Lin (2010), Tzougas, Vrontos, and Frangos (2014), Miljkovic and Grün (2016), and Fung, Badescu, and Lin (2021). Recent work such as Tzougas, Vrontos, and Frangos (2018) and Blostein and Miljkovic (2019) combines both light-tailed and heavy-tailed distributions to cater to the mismatch between body and tail behavior. In this case, however, we will show empirically in Section 5.2 that the tail estimation is unrobust due to the overlapping density region between the body and the tail where the small claims can severely impact the estimation of the tail index.

- **Composite or splicing models** comprise of light-tailed distributions up to a threshold for modeling moderate losses and heavy-tailed distributions beyond the threshold for large losses to potentially capture the mismatch between the body and the tail behavior. Composite models also address the tail unrobustness problem inherited from finite mixture models because there is no density overlap between body and tail parts. The relevant actuarial literature includes, for instance, Cooray and Ananda (2005), Scollnik (2007), Pigeon and Denuit (2011), Scollnik and Sun (2012), Nadarajah and Bakar (2014), Bakar et al. (2015), Calderín-Ojeda and Kwok (2016), Grün and Miljkovic (2019) and Parodi (2020).

- **Combinations finite mixtures and composite models** were proposed by Reynkens et al. (2017), who developed a global fitting strategy for a splicing model with an Erlang mixture distribution for the body and a Pareto distribution for the tail.

To understand how the claim severity distribution is influenced by certain risk factors, covariates’ influence (point 3) has been extensively explored in actuarial literature using various types of severity regression models (Nelder and Wedderburn 1972). We refer readers to Frees (2009) for a comprehensive summary.

Considering variable selection techniques (point 4), a popular approach is the use of penalty functions, such as least absolute shrinkage and selection operator (LASSO; see Tibshirani 1996) or smoothly clipped absolute deviation (SCAD; see Fan and Li 2001), to shrink unimportant regression coefficients to zero. In actuarial literature, Jeong, Chang, and Valdez (2021) used nonconvex regularization methods in order to obtain stable estimation of loss development factors in insurance claims reserving. In a multi-type variable setting, Devriendt et al. (2021) is currently the only paper that considers multi-type feature selection under a Poisson regression framework for claim frequency modeling.

Though the existing literature addresses some of the aforementioned claim severity modeling needs, a universal modeling framework that not only provides versatility to fit a multimodal heavy-tailed severity distribution but also explains the covariates’ influence on multiple distributional parts with variable selection is still lacking. As a result, the goal of this article is to integrate, adapt, and extend the existing modeling techniques and devise a universal insurance claim severity modeling framework that simultaneously addresses all of the four modeling needs mentioned above. To this end, we make the following contributions.

Firstly, we introduce a mixture composite regression model for claim severity modeling based on the use of the available covariate information. This extends the setup of Reynkens et al. (2017), who used a finite mixture distribution for the body and a Pareto-type distribution for the tail of the distribution without using covariates, by incorporating covariates’ impacts on all three parts of the severity distribution: clustering probabilities, body part, and tail part.

Secondly, we propose a group-fused regularization approach for variable selection. This approach allows us to select three different sets of variables that significantly impact the previously mentioned three parts of the claim severity distribution, respectively. The set of variables chosen is homogeneous across all mixture components to preserve model interpretability. Furthermore, this approach enables regularization under multi-type variable settings.

Thirdly, we develop an asymptotic theory to theoretically justify the proposed regularization approach. The proposed method is able to accurately merge and shrink regression coefficients across the various modeling parts provided that the sample size is large enough. Furthermore, it is still feasible to quantify model uncertainty under our proposed method. We can
construct several quantities of interest, including Wald-type confidence intervals and Efron percentile bootstrap confidence intervals of model parameters.

Finally, we present a novel generalized expectation-maximization (GEM) algorithm for estimating the parameters of the proposed model with parameter regularization. The GEM algorithm is demonstrated to perform satisfactorily when a mixture-Gamma Lomax composite regression model is fitted to a Greek MTPL dataset that inherits all of the previously described features.

The remainder of this article proceeds as follows. In Section 2, we introduce the framework of the mixture composite regression model. Section 3 presents the feature selection approach, which can be used for selecting important variables for explaining the claim severity distribution in the presence multi-type covariates. The maximum likelihood estimation (MLE) procedure for our proposed model via the GEM algorithm is presented in Section 4. In Section 5, we describe the MTPL dataset that we use for our empirical analysis and provide estimation and model comparison for various benchmark distributions. We also fit the proposed mixture composite distribution and, subsequently, the mixture composite regression model with feature regularization. Concluding remarks are given in Section 6, and other theoretical and computational details are included in the Appendixes.

2. MODELING FRAMEWORK

This section summarizes the features that are incorporated in a regression modeling framework to address the challenges encountered in claim severity datasets in general insurance. In particular, motivated by the characteristics of the multimodal and heavy-tailed Greek MTPL insurance dataset studied below, we propose the following mixture composite regression model.

Let $Y \in \mathbb{R}^+$ be the claim severity random variable, and let $x \in \mathbb{R}^D$ be the vector of covariate information.\(^1\) The density of the mixture composite regression model is given by

$$
    h_Y(y; \mathbf{x}, \mathbf{\beta}, \phi, \theta, v, x) = \sum_{j=1}^{g+1} \pi_j(y; \mathbf{x}) \frac{f(y; \exp \{\beta_j^T x\}, \phi_j)}{F(\tau; \exp \{\beta_j^T x\}, \phi_j)} \{y \leq \tau\} + \pi_{g+1}(x; \mathbf{x}) \frac{h(y; \theta, \exp \{v^T x\})}{1 - H(\tau; \theta, \exp \{v^T x\})} \{y > \tau\},
$$

where $\pi_j(x; \mathbf{x}), 1 \leq j \leq g+1$, are covariate-dependent component weights given by

$$
    \pi_j(x; \mathbf{x}) = \frac{\exp \{a_j^T x\}}{\sum_{j=1}^{g+1} \exp \{a_j^T x\}},
$$

with $a_{g+1} = 0$ for model identifiability, and $\mathbf{a} = (\mathbf{a}_1, ..., \mathbf{a}_{g+1}) \in \mathbb{R}^{D \times (g+1)}$. Hence, the weight function $\pi_j(x; \mathbf{x})$ forms a logistic GLM with a canonical link. $f$ and $h$ are the body and tail density functions, respectively, such that the first $g$ mixture components are specialized in capturing small to moderate claim amounts and the last component focuses on extreme claims. $F$ and $H$ are the corresponding cumulative distribution functions.

In this article, we specify $f$ and $h$ as Gamma (body) and Lomax (tail, also called Pareto type II) density functions given by, respectively,

$$
    f(y; \mu, \phi) = \frac{(\phi \mu)^{-1/\phi} \Gamma(1/\phi)}{\Gamma(1/\phi)} y^{1/\phi - 1} e^{-y/(\phi \mu)},
$$

and

$$
    h(y; \theta, \zeta) = \frac{\zeta \theta^\zeta}{(y + \theta)^{\zeta+1}}.
$$

\(^1\)Note also that all vectors are assumed to be column vectors.
The choice of Gamma density is motivated by its light-tailed and uni-modal characteristics to capture small to moderate claims. Also, mixtures of Gammas provide sufficient flexibility to capture complex distributional structures like multimodality, thanks to the denseness property of Gamma mixtures. The Gamma components also form a GLM with exponential links. The choice of the Lomax density for the tail is motivated by its polynomial tail characteristics with the tail index $\zeta$ describing the tail-heaviness of the distribution. The analytical form of the truncated Lomax distribution also makes the model estimation procedures computationally desirable.

**Remark 1.** One may choose other plausible model specifications as long as $f$ is a unimodal lighter-tailed density and $h$ is a heavier-tailed density. Alternative possible choices of $f$ include the Weibull and log-normal densities. Note, however, that mixtures of any unimodal distribution can serve as a flexible density approximation tool to capture the body part of the data flexibly. Hence, the proposed mixture composite model in Equation (2.1) can resemble well a wide range of complex distributional characteristics, regardless of the choice of the unimodal finite mixture components. As a result, we do not expect that replacing the Gamma components with other unimodal densities (such as Weibull) will substantially impact either model interpretation or fitting results. To avoid distorting the focus of this article and given that the fitting results of the real Greek automobile dataset (Section 5.3) are satisfactory, we focus on this model.

**Remark 2.** Note that there is a discontinuity of the proposed density function (Eq. [2.1]) at $\tau$. In actuarial literature, it is common to impose constraints on the parameter space so that the resulting density is continuous and differentiable in $\tau$. However, we refrain from doing so because the resulting parameter estimation procedures will require a constrained optimization with respect to (w.r.t.) high-dimensional parameters (including multiple sets of regression coefficients for the finite mixture distribution), which is computationally prohibitive. On the other hand, the impact of the discontinuity issue of the density function is minimal in practice: Finite mixtures are well known to be very flexible in capturing complex distributional phenomena. Fitting our proposed mixture-based model to the data, we foresee that the resulting fitted model will closely agree with the empirical distribution (which is approximately continuous). Therefore, the distributional discontinuity at $\tau$ is expected to be small for the fitted model in most practical cases, even if no constraints are imposed on the composite density function.

Furthermore, $\beta = (\beta_1, \ldots, \beta_g) \in \mathbb{R}^{D \times g}$ and $\nu \in \mathbb{R}^D$ are the regression coefficients for the body and tail distributions, respectively. The proposed distribution is characterized by a splicing threshold $\tau > 0$, which is predetermined using expert opinion via performing, for example extreme value analysis instead of treating it as a parameter estimated by a likelihood approach; this is mainly motivated by estimation stability and is adopted by, for example Reijnkens et al. (2017).

The mean of $Y|x$ is given by

$$
E[Y|x] = \sum_{j=1}^{g} \pi_j(x; \mathbf{a}) \frac{\theta + \tau}{\exp \{\nu^T x\} - 1 + \tau}.
$$

**Remark 3.** Equation (2.5) is finite if and only if the tail index $\zeta := \exp \{\nu^T x\}$ is greater than 1. To ensure that the mean resulting from the fitted model is always finite for every $x \in \mathbb{R}^D$, one may need to consider an alternative regression link to the tail index, such as the inverse of sigmoid function.

The composite model in Equation (2.1) can alternatively be regarded as a mixture of $g$ right-truncated Gamma distributions for the body and a left-truncated Lomax distribution for the tail. Each claim is classified to one of the $g+1$ subgroups ($g$ subgroups for body and one subgroup for tail) with probabilities $(\pi_j(x; \mathbf{a}))_{j=1, \ldots, g+1}$, where each subgroup may correspond to a different claim subtype. Regression coefficient $\mathbf{a}$ explains the heterogeneity of the assignment probabilities across different claims, and $\beta$ explains the systematic effects of the claims within the given subgroups. The regression coefficients $\nu$ for the tail distribution, on the other hand, capture the effect of covariates to the tail-heaviness of claims.

The motivation of introducing a composite model in Equation (2.1) instead of using a mixture-Gamma Lomax model is that there are no overlapping density regions between the body and tail distributions under the proposed framework. We will show in our motivating application in Section 5 that this results in a robust and stable estimation procedure of the tail index, because it is not distorted by attritional claims from the body of the distribution. One should, however, note that the mixture probabilities connect tail and body regression parameter estimation; that is the proposed composite regression model does not decouple into independent estimation parts.
3. FEATURE SELECTION METHOD

In this section, we propose the group-fused penalty approach to select the variables to describe the systematic effects in claim severities under a multi-type covariates setting. We will select three potentially different sets of variables that may influence, respectively, the subgroup probabilities, body, and tail of the distribution. For the sake of model interpretability, we select the same set of variables for all mixture components of the body of the data and the mixing probabilities.

Suppose there are \( n \) independent claims \( Y = (Y_1, \ldots, Y_n)^T \), and denote their realizations by \( y = (y_1, \ldots, y_n)^T \). For each claim \( i = 1, \ldots, n \), we have a covariate vector \( x_i = (x_{i1}, \ldots, x_{id})^T \in \mathbb{R}^D \) with \( x_{i1} = 1 \) (for the intercept component). Define \( X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n \times D} \) as a full-rank design matrix containing the covariate information of all \( n \) observations. The observed data log-likelihood is given by

\[
\mathcal{L}_n(\Phi) := \mathcal{L}_n(\Phi; y, X) = \sum_{i=1}^{n} \log h_Y(y_i; \beta, \phi, \theta, v, x_i),
\]

where \( h_Y(\cdot) \) is the density function of \( Y_i \) given by Equation (2.1), and \( \Phi := (\beta, \phi, \theta, v) \) contains all model parameters. To incorporate variable selection, we propose a group-fused regularization approach, where the penalty function for the regression parameters is as follows:

\[
P_n(\Phi) = P_{\lambda_1, n}(\beta) + P_{\lambda_2, n}(\beta) + P_{\lambda_3, n}(v),
\]

with \( P_{\lambda_1, n}(\beta) \), \( P_{\lambda_2, n}(\beta) \), and \( P_{\lambda_3, n}(v) \) being the penalty functions on the regression parameters \( \beta \in \mathbb{R}^{D \times (g+1)} \), \( v \in \mathbb{R}^D \). These are given by

\[
P_{\lambda_1, n}(\beta) = \sum_{k=1}^{K_1} p_{1n}(\|c_{1k}^T \beta\|_2; \lambda_{1kn}), \quad P_{\lambda_2, n}(\beta) = \sum_{k=1}^{K_2} p_{2n}(\|c_{2k}^T \beta\|_2; \lambda_{2kn}),
\]

\[
P_{\lambda_3, n}(v) = \sum_{k=1}^{K_3} p_{3n}(\|c_{3k}^T v\|; \lambda_{3kn}),
\]

where \( \| \cdot \|_2 \) is the \( L^2 \)-norm, \( \lambda_{1kn}, \lambda_{2kn}, \) and \( \lambda_{3kn} \) are penalty tuning parameters; \( p_{1n}, p_{2n}, \) and \( p_{3n} \) are concave non-decreasing penalty functions (which will be chosen proportional to the sample size \( n \)); and \( K_1, K_2, \) and \( K_3 \) correspond to the numbers of penalization terms that govern how the regression coefficients are shrunk. Finally, \( \{c_{ik}\}_{k=1}^{3} \) are predetermined vectors of penalty coefficients that allow for different types of penalties, including standard LASSO to shrink continuous variables, fused LASSO to merge regression coefficients of various ordinal categorical variables, and generalized fused LASSO to merge regression coefficients for nominal categorical variables. For a full description on constructing predetermined vectors for each type of variables (continuous, ordinal categorical, and nominal categorical), we refer the reader to Oelker and Tutz (2017), in the statistics literature, and to Devriendt et al. (2021) in the actuarial literature. We hereby provide an example to give readers a better understanding on the notations:

**Example 1.** We consider a case with three explanatory variables: one continuous, one ordered categorical (with three levels), and one nominal (unordered) categorical (with four levels). Denote \( x_{i1} = 1 \) as the intercept, \( x_{i2} \) as the continuous covariate, \( x_{i3} - x_{i4} \) as the ordered categorical covariates (note that the dummy coding provides two covariates and one reference level for a three-level categorical variable), and \( x_{i5} - x_{i7} \) as the nominal categorical covariates (also using dummy coding). For simplicity, we only consider \( g = 1 \) in this example. A multi-type feature regularization approach requires the following setup:

- For a continuous variable, we penalize the absolute value of the regression coefficient to shrink it.
- For an ordered categorical variable, we penalize the absolute difference of regression coefficients between any two adjacent levels to fuse adjacent levels.
- For a nominal categorical variable, we penalize the differences of coefficients between any two levels to merge levels without the knowledge of their orderings.

Denote \( C_l := (c_{l1}, \ldots, c_{lK_l}) \) as a design matrix of penalty coefficients for \( l = 1, 2, 3 \). In this case, we have \( K_l = 1 + (3 - 1) + \binom{4}{2} = 9 \) and
The proposed feature selection method:

\[ C_j = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix} \]

for \( l = 1, 2, 3 \). The resulting penalty function for \( x \) in Equation (3.2) is then given by

\[
P_{\lambda,\eta}(x) = p_{\text{ln}}(|x_1|; \lambda_{11}) + p_{\text{ln}}(|x_1|; \lambda_{12}) + p_{\text{ln}}(|x_2 - x_1|; \lambda_{13}) \\
+ p_{\text{ln}}(|x_3|; \lambda_{14}) + p_{\text{ln}}(|x_3|; \lambda_{15}) + p_{\text{ln}}(|x_4|; \lambda_{16}) \\
+ p_{\text{ln}}(|x_6 - x_5|; \lambda_{17}) + p_{\text{ln}}(|x_7 - x_6|; \lambda_{18}) + p_{\text{ln}}(|x_8 - x_7|; \lambda_{19}).
\]

Analogous expressions can be obtained for \( P_{\lambda,\eta}(\beta) \) and \( P_{\lambda,\eta}(\gamma) \). The first term on the right-hand side of Equation (3.5) shrinks the regression coefficients of the continuous variable. The next two terms merge the ordered categorical levels, while the last six terms merge the nominal categorical levels.

The aim is to maximize the following objective function (penalized log-likelihood)

\[
\mathcal{F}_n(\Phi) = \mathcal{L}_n(\Phi) - P_n(\Phi).
\]

We will use the following two commonly used penalty functions (for \( l \in \{1, 2, 3\} \) and \( \psi \geq 0 \)) to illustrate the usefulness of the proposed feature selection method:

- **L\(^1\)**-norm (LASSO) penalty: \( p_{\text{ln}}(\psi; \eta) = n_\eta |\psi| \);

- (SCAD) penalty introduced by Fan and Li (2001): \( p_{\text{ln}}'(\psi; \eta) := p_{\text{ln}}'(\psi; \eta, a) = n_\eta \left[ 1 \{ |\psi| \leq \eta \} + \frac{(a|\psi|-\psi)}{(a-1)|\psi|} 1 \{ |\psi| > \eta \} \right] \), with \( p_{\text{ln}}(0; \eta) = 0 \) and \( a > 2 \) being a hyperparameter affecting the shape of the penalty function. A smaller \( a \) makes the penalty function more nonconcave, reducing the modeling bias introduced by regularization (Fan and Li 2001). On the other hand, the penalty function converges to the LASSO (which is concave) as \( a \to \infty \). Note that \( p_{\text{ln}}'(\psi; \eta) \) is denoted as the first derivative of \( p_{\text{ln}}(\psi; \eta) \) w.r.t. \( \psi \). We follow the guidance of Fan and Li (2001)\(^2\) by choosing the hyperparameter \( a \) as 3.7.

Here, we set \( n_1 = n \) for the total number of observations, \( n_2 := n_\eta = \sum_{i=1}^{n} 1 \{ y_i \leq \tau \} \) for the number of observations in the body; and \( n_3 := n_\tau = \sum_{i=1}^{n} 1 \{ y_i > \tau \} \) for the number of observations in the tail.

Each of the parameters \( x \) and \( \beta \) contains \( g \) sets of regressors (one for each mixture component of the body; we initialize \( x_{g+1} = 0 \)). For the sake of interpretability, the proposed fused-group regularization method shrinks and merges regression coefficients of any variable uniformly across all mixture components, and the sets of variables do not vary across mixture components. Therefore, the proposed method allows us to choose three sets of variables that significantly impact each of the three modeling parts: the subgroup probabilities, the body part, and the tail part of the severity distribution.

**Remark 4.** Alternatively, one can adopt a fused penalization (ungrouped) instead of a grouped one in Equation (3.3) for variable selection, resulting in different sets of variables being selected across mixture components. Though this may provide more modeling flexibility, model interpretation will become more difficult when the number of mixture components \( g \) becomes large, because we will expect slightly different shrinkage and mergrence of variable levels across multiple mixture components.

\(^2\)Fan and Li (2001) justified the choice of \( a = 3.7 \) by showing that it minimizes the Bayes risk function. Alternatively, Fan and Li (2001) mentioned that the optimal \( a \) can be searched through cross-validation, but such an implementation can be computationally prohibitive.
After presenting the above feature selection strategy, it is desirable to develop an asymptotic theory to justify the appropriateness of the proposed regularization approach. The technical details, mathematical theorems, and proofs are provided in Appendix A, and here we summarize our main theoretical results as follows:

1. The proposed method is consistent in terms of feature selection, in particular, as sample size goes to infinity, the proposed method will correctly merge and shrink regression coefficients across the various modeling parts.
2. The parameters of the reduced model, after merging and shrinking the regression coefficients, are asymptotically normal with zero mean, and their variances are the same as the parameter uncertainties obtained by fitting the same mixture composite regression model to the reduced model. As a result, we can construct Wald-type confidence intervals, Efron percentile bootstrap confidence intervals of model parameters, and other quantities of interest to quantify model uncertainty.

4. MODEL ESTIMATION

Direct optimization of the penalized log-likelihood in Equation (3.6) is difficult. Firstly, the log-likelihood \( \log h(y; x, \beta, \phi, \theta, v, x) \) is the logarithm of a sum of \((g + 1)\) mixture terms. Secondly, observe that \( \log h(y; x, \beta, \phi, \theta, v, x) \) contains a gamma distribution function \( F(z; \exp \{ b^T x_i \}, \phi) \), which is not available in closed form\(^3\); this is not the case for the Lomax distribution because \( H(z; \theta, \exp \{ b^T x_i \}) \) has an analytical form. Thirdly, the penalty functions are not continuously differentiable.

Model estimation of an ordinary splicing model is typically simple, thanks to the non-overlapping density parts between the body and tail, so that one may factor out the likelihood function, and separately calibrate the three parts of distribution: subgroup probability, body, and tail. Nonetheless, under the regression framework outlined in Equation (2.1) with variable selection techniques embedded, the weight regression parameters \( x \) share and interact across all \( g \) body, components and one tail component. Therefore, there is no straightforward way to segregate the likelihood function and simplify the estimation procedure.

Motivated by the computational challenges, this section presents the strategy to estimate the parameters and select important variables under the proposed modeling framework.

4.1. Construction of Complete Data

We first construct a hypothetical complete dataset that is a modified version of the method introduced by Fung, Badescu, and Lin (2022). The rationale is that the complete data would result in a much simpler penalized log-likelihood function compared to Equation (3.6). The simplified objective function enables us to apply the GEM algorithm for efficient estimation of parameters with manageable computational costs; see Section 4.2 below. Define the complete data

\[
D^{\text{com}} = \{ (y_i, z_i, m_i, y'_i) \}_{i=1,...,n},
\]

with three extra elements defined as follows:

- \( z_i = (z_{i1}, ..., z_{i(g+1)}) \) is the realization of a categorical latent random vector \( Z_i = (Z_{i1}, ..., Z_{i(g+1)}) \) such that \( Z_{ij} = 1 \) if the \( i \)th observation comes from the \( j \)th component of the proposed mixture distribution and \( Z_{ij} = 0 \) otherwise; in fact, \( Z_i \) is one-hot encoding of the selected mixture component that \( y_i \) is allocated to.
- \( m_i \) is the realization of \( M_i \), the number of missing sample points outside the truncation interval \((0, \tau)\) generated by the \( i \)th observation.
- \( y'_i = (y'_{i1}, ..., y'_{imu}) \) is the realization of \( Y'_i = (Y'_{i1}, ..., Y'_{imu}) \), the missing sample points from the \( i \)th observation.

Thus, we artificially expand beyond the truncation point, and this expansion is done such that it is convenient for parameter estimation. We assume that the cases \( (y_i, Z_i, M_i, Y'_i) \) are independent in \( 1 \leq i \leq n \). Moreover, we assume independence among \( M_i, Y_i \) and \( Y'_i \) conditioned on \( Z_i \), and that \( Y'_{i1}, ..., Y'_{imu} \) are independent and identically distributed (i.i.d.) given \( x_i \) and \( Z_i \). The categorical latent random vector \( Z_i \) independently follows a (single-trial) multinomial distribution with \( P(Z_{ij} = 1) = \pi_j(x_i; \phi) \) for \( j = 1, ..., g + 1 \). The number of missing sample points \( M_i \) given the latent subgroup \( Z_i \) is assumed to follow the geometric distribution.

\(^3\)Note that if \( F(\cdot) \) is a Weibull distribution (see Remark 1), then \( F(\cdot) \) can be expressed analytically. In this case, the statement “Secondly, observe that...” does not apply as a justification of implementing the proposed GEM algorithm. Still, there are two other reasons why a GEM algorithm should be implemented to mixture-Weibull composite models.
\[ P(M_i = m_i | Z_{ij} = 1, x_i, \Phi) = \left[ 1 - F(\tau; \exp \{ \beta^T_j x_i \}, \phi_j) \right]^{m_i} F(\tau; \exp \{ \beta^T_j x_i \}, \phi_j), \quad m_i \in \mathbb{N}_0, \]  

for \( j = 1, \ldots, g \). For completeness, we assume \( M_i = 0 \) almost surely if \( Z_{i(g+1)} = 1 \). The missing sample points \( Y'_{im} \) of \( Y_i \) are i.i.d. conditioned on \( Z_i \) with the following density function for \( j = 1, \ldots, g \) :

\[
f_{y_{im} | z_{01} = 1}(y; \beta_j, \phi_j, x_i) = \frac{f(y; \exp \{ \beta^T_j x_i \}, \phi_j)}{1 - F(\tau; \exp \{ \beta^T_j x_i \}, \phi_j)} 1 \{ y > \tau \}.
\]

The complete data log-likelihood function is given by (the derivation is shown in Appendix C.1)

\[
\mathcal{L}^{\text{com}}_n(\Phi; D^{\text{com}}, X) = \sum_{i=1}^n \sum_{j=1}^{g+1} \left( \log \pi_j(x_i; \alpha) + \log f(y_i; \exp \{ \beta^T_j x_i \}, \phi) 1 \{ y_i \leq \tau \} + \log \frac{h(y_i; \theta; \exp \{ v^T x_i \})}{1 - H(\tau; \theta; \exp \{ v^T x_i \})} 1 \{ y_i > \tau \} \right) + \sum_{i=1}^n \sum_{j=1}^{g+1} \sum_{m=1}^{m_i} z_{ij} \log f(y'_{im}; \exp \{ \beta^T_j x_i \}, \phi). 
\]

This is easier to evaluate and optimize compared to Equation (3.1) given that \( H \) has an analytical form, which is the case for the Lomax distribution. The complete data penalized log-likelihood is given by

\[
\mathcal{F}^{\text{com}}_n(\Phi) = \mathcal{L}^{\text{com}}_n(\Phi; D^{\text{com}}, X) - P_n(\Phi). 
\]

**Remark 5.** The choice of geometric distributions for \( M_i \) is motivated by the fact that it will allow for an efficient fitting algorithm, because it will lead to nice cancellations in the complete data likelihood function that cannot be achieved by other discrete distributions like Poisson and binomial. That is, this is an auxiliary tool that is computationally attractive. See Appendix C.1 for more details.

### 4.2. The GEM Algorithm

This subsection provides the technical details for the GEM algorithm to efficiently estimate the model parameters. A summary outline is also provided in Appendix C.3 to help readers implement the algorithm.

In the GEM algorithm, the M-step is a modified version of the penalized iteratively reweighted least squares (PIRLS) method proposed by Oelker and Tutz (2017), where this method is also technically justified. To perform the PIRLS algorithm, one needs to apply a Newton-based approach that requires computing the gradient and Hessian of the objective function. However, the penalty terms \( p_{1n}(\| c_k^T x \|_2^2; \lambda^{1kn}) \), \( p_{2n}(\| c_k^T \beta \|_2^2; \lambda^{2kn}) \), and \( p_{3n}(\| c_k^T v \|_2^2; \lambda^{3kn}) \) are not differentiable for both LASSO and SCAD penalties when \( c_k^T x = 0 \), \( c_k^T \beta = 0 \), or \( c_k^T v = 0 \). This is because of the nondifferentiability of the \( L^2 \)-norm \( \| \cdot \| \) and the absolute function \( | \cdot | \) at 0. Hence, the corresponding gradient and Hessian may be undefined. Motivated by this computational challenge, we follow Oelker and Tutz (2017) to perturb the penalty function in Equation (3.2) as follows:

\[
P_\epsilon(\Phi) = P_{k_1,n,\epsilon}(\alpha) + P_{k_2,n,\epsilon}(\beta) + P_{k_3,n,\epsilon}(v), 
\]

where the \( \epsilon \)-perturbed penalty functions \( P_{k_1,n,\epsilon}(\alpha) \), \( P_{k_2,n,\epsilon}(\beta) \), and \( P_{k_3,n,\epsilon}(v) \) are given by

\[
P_{k_1,n,\epsilon}(\alpha) = \sum_{k=1}^{K_1} p_{1n}(\| c_k^T x \|_2; \lambda_{1kn}), \quad P_{k_2,n,\epsilon}(\beta) = \sum_{k=1}^{K_2} p_{2n}(\| c_k^T \beta \|_2; \lambda_{2kn}), 
\]

\[
P_{k_3,n,\epsilon}(v) = \sum_{k=1}^{K_3} p_{3n}(\| c_k^T v \|_2; \lambda_{3kn}), 
\]

where \( K_1, K_2, K_3 \) are finite sets that are chosen based on the user's discretion. The details of these sets are provided in Appendix C.4.
with \( \|w\|_{2,\varepsilon} = (ww^T + \varepsilon)^{1/2} \) and \( |w|_{\varepsilon} = (w^2 + \varepsilon)^{1/2} \) for any vector \( w \) and scalar \( \varepsilon \). Because \( \|w\|_{2,\varepsilon} \) is bounded below by \( \varepsilon^{1/2} > 0 \), this ensures differentiability of the penalty functions \( p_{\varepsilon}(\cdot) \) \((l = 1, 2, 3)\). In the following, instead of maximizing Equation (4.5), we maximize the \( \varepsilon \)-perturbed complete data penalized log-likelihood given by

\[
F_{n,\varepsilon}(\Phi) = F_n(\Phi; D_{\text{com}}, X) - P_{n,\varepsilon}(\Phi).
\]

\( F_{n,\varepsilon}(\Phi) \) is continuously differentiable w.r.t. any parameter and, hence, it is computationally tractable. Furthermore, note that \( P_{n,\varepsilon}(\Phi) \rightarrow P_n(\Phi) \) and hence \( F_{n,\varepsilon}(\Phi) \rightarrow F_n(\Phi) \) as \( \varepsilon \rightarrow 0 \), so choosing a very small \( \varepsilon > 0 \), the perturbation of the estimated parameters \( \Phi \) will be negligible. In simulation studies and real data analysis, we find that the choice of \( \varepsilon = 10^{-10} \) works well. Note, however, that \( \varepsilon > 0 \) will not shrink components precisely to zero, and hence this needs to be done manually. The automated adjustment algorithm discussed in Section 4.5 will address this issue.

4.2.1. E-step

In the \( b \)th iteration, the expectation of the complete data \( \varepsilon \)-perturbed penalized log-likelihood is computed as follows:

\[
Q_b(\Phi; y, X, \Phi^{(l-1)}) = E \left[ F_{n,\varepsilon}(\Phi) \mid y, X, \Phi^{(l-1)} \right] = \sum_{i=1}^{n} \sum_{j=1}^{g} \left( \log \pi_j(x_i; \alpha) + \log f(y_i; \exp \{ \beta_j^T x_i \}, \phi_j) \mathbb{1}\{y_i \leq \tau \} + \log \frac{h(y_i; \theta, \exp \{ \mathbf{v}^T x_i \})}{1 - H(\tau; \theta, \exp \{ \mathbf{v}^T x_i \})} \mathbb{1}\{y_i > \tau \} \right) + \frac{1}{\phi_j} \left( \log \gamma_j^{(l)} - \log \phi_j - \beta_j^T x_i - \gamma_j^{(l)} \exp \{-\beta_j^T x_i\} \right) - \log \Gamma \left( \frac{1}{\phi_j} \right)
\]

where the updated quantities \( z_j^{(l)}, m_j^{(l)}, \gamma_j^{(l)}, \) and \( \log \gamma_j^{(l)} \) are displayed in Equations (C.2) to (C.5) of Appendix C. Note that \( \log \gamma_j^{(l)} \) does not have an analytical solution (i.e., \( \log \gamma_j^{(l)} \) is expressed as a [numerical] integral in Eq. (C.5)), so the computation burden is high by directly computing \( \log \gamma_j^{(l)} \). As a result, we adopt a stochastic EM approach: For each \( i = 1, \ldots, n \) and \( j = 1, \ldots, g \), we estimate \( \log \gamma_j^{(l)} \) by simulating (log-transformed) \( Y_j^{(i)} \) from the density of \( Y_j^{(i)} | Z_{ij} = 1 \) given by Equation (4.3) evaluated at parameters \( \beta_j^{(l-1)} \) and \( \phi_j^{(l-1)} \).

4.2.2. M-step

In this step, we attempt to find a parameter update \( \Phi^{(l)} \) in such that we will receive monotonicity \( Q_b(\Phi^{(l)}; y, X, \Phi^{(l-1)}) \geq Q_b(\Phi^{(l-1)}; y, X, \Phi^{(l-1)}) \). Though \( Q_b(\Phi; y, X, \Phi^{(l-1)}) \) is now differentiable w.r.t. any parameter, direct implementation of an iteratively re-weighted least squares (IRLS) algorithm is challenging due to concavity of the penalty functions. In this section, we propose the use of convex quadratic approximation to the penalty functions, analogous to Fan and Li (2001) and Oelker and Tutz (2017), such that the implementation of an IRLS algorithm is feasible. We approximate \( P_{n,\varepsilon}(\Phi) \) by

\[
\tilde{P}_{n,\varepsilon}(\Phi) = \sum_{k=1}^{K_1} \tilde{P}_{ln}(\|c_{1k}^T x\|_{2,\varepsilon}; \lambda_{1kn}) + \sum_{k=1}^{K_2} \tilde{P}_{2n}(\|c_{2k}^T \beta\|_{2,\varepsilon}; \lambda_{2kn}) + \sum_{k=1}^{K_3} \tilde{P}_{3n}(\|c_{3k}^T y\|_{2,\varepsilon}; \lambda_{3kn}),
\]

where

\[
\tilde{P}_{ln}(\|c_{1k}^T x\|_{2,\varepsilon}; \lambda_{1kn}) = p_{ln}(\|c_{1k}^T x^{(l-1)}\|_{2,\varepsilon}; \lambda_{1kn}) + \frac{1}{2} \frac{c_{1k}^T x^T c_{1k} - c_{1k}^T x^{(l-1)} x^{(l-1)^T} c_{1k}}{\|c_{1k}^T x^{(l-1)}\|_{2,\varepsilon}} p_{ln}'(\|c_{1k}^T x^{(l-1)}\|_{2,\varepsilon}; \lambda_{1kn}).
\]
\[
\tilde{p}_{2n}\left(\|e_{Tn}^T\beta\|_{2,\varepsilon}^2; \lambda_{2kn}\right) = p_{2n}\left(\|e_{Tn}^T\beta^{(l-1)}\|_{2,\varepsilon}^2; \lambda_{2kn}\right) \\
+ \frac{1}{2} \frac{c_{Tn}^T \beta T c_{2k} - c_{Tn}^T \beta^{(l-1)} T c_{2k}}{\|e_{Tn}^T\beta^{(l-1)}\|_{2,\varepsilon}^2} p_{2n}'\left(\|e_{Tn}^T\beta^{(l-1)}\|_{2,\varepsilon}^2; \lambda_{2kn}\right).
\] (4.12)

\[
\tilde{p}_{3n}\left(\|e_{Tn}^T v\|_{2,\varepsilon}^2; \lambda_{3kn}\right) = p_{3n}\left(\|e_{Tn}^T v^{(l-1)}\|_{2,\varepsilon}^2; \lambda_{3kn}\right) + \frac{1}{2} \frac{(c_{Tn}^T v^2 - (c_{Tn}^T v^{(l-1)})^2)}{\|e_{Tn}^T v^{(l-1)}\|_{2,\varepsilon}^2} p_{3n}'\left(\|e_{Tn}^T v^{(l-1)}\|_{2,\varepsilon}^2; \lambda_{3kn}\right).
\] (4.13)

The properties below justify the use of such approximations:

**Theorem 1.** As \(\epsilon \to 0\), \(\widehat{P}_{n,\epsilon}(\Phi)\) majorizes \(\mathcal{P}_{n,\epsilon}(\Phi)\), i.e., \(\widehat{P}_{n,\epsilon}(\Phi^{(l-1)}) = \mathcal{P}_{n,\epsilon}(\Phi^{(l-1)})\) and \(\widehat{P}_{n,\epsilon}(\Phi) \geq \mathcal{P}_{n,\epsilon}(\Phi)\), as \(\epsilon \to 0\), for any \(\Phi \neq \Phi^{(l-1)}\).

**Proof.** Without loss of generality it suffices to prove that \(\tilde{p}_{1n}(\|e_{Tn}^T x\|_{2,\varepsilon}; \lambda_{1kn})\) majorizes \(p_{1n}(\|e_{Tn}^T x\|_{2,\varepsilon}; \lambda_{1kn})\). Firstly, it is trivial that \(\tilde{p}_{1n}(\|e_{Tn}^T x^{(l-1)}\|_{2,\varepsilon}; \lambda_{1kn}) = p_{1n}(\|e_{Tn}^T x^{(l-1)}\|_{2,\varepsilon}; \lambda_{1kn})\). Secondly, we notice that we have \(\tilde{p}_{1n}(\|e_{Tn}^T x\|_{2,\varepsilon}; \lambda_{1kn}) \geq p_{1n}(\|e_{Tn}^T x\|_{2,\varepsilon}; \lambda_{1kn})\) for any \(x \neq x^{(l-1)}\) if the following condition is met: For any \(u \in \mathbb{R}\) satisfying \(u^2 \neq e_{Tn}^T x^{(l-1)} T x^{(l-1)} e_{Tn} =: u^2\), we have

\[
\tilde{p}_{1n}\left((u^2 + \epsilon)^{1/2}; \lambda_{1kn}\right) := p_{1n}\left((u^2 + \epsilon)^{1/2}; \lambda_{1kn}\right) \\
+ \frac{1}{2} \frac{u^2 - u^2}{(u^2 + \epsilon)^{1/2}} p_{1n}'\left((u^2 + \epsilon)^{1/2}; \lambda_{1kn}\right) \geq p_{1n}\left((u^2 + \epsilon)^{1/2}; \lambda_{1kn}\right).
\] (4.14)

As \(\epsilon \to 0\), the above equation obviously holds for \(u^* \neq 0\), because \(\tilde{p}_{1n}(u; \lambda_{1kn})\) is a convex function as opposed to \(p_{1n}(u; \lambda_{1kn})\) is a concave function with \(\tilde{p}_{1n}(u; \lambda_{1kn}) = p_{1n}(u; \lambda_{1kn})\). If \(u^* = 0\), then \(\tilde{p}_{1n}((u^2 + \epsilon)^{1/2}; \lambda_{1kn}) \to \infty\) as \(\epsilon \to 0\) and \(p_{1n}(u; \lambda_{1kn}) \to \infty\), so the result follows.

**Corollary 1.** As \(\epsilon \to 0\), if \(\tilde{Q}_\epsilon(\Phi^{(l)}; y, X, \Phi^{(l-1)}) \geq \tilde{Q}_\epsilon(\Phi^{(l)}; y, X, \Phi^{(l-1)})\), then we have \(Q_\epsilon(\Phi^{(l)}; y, X, \Phi^{(l-1)}) \geq \tilde{Q}_\epsilon(\Phi^{(l)}; y, X, \Phi^{(l-1)})\), where \(\tilde{Q}_\epsilon\) is the same as \(Q_\epsilon\) except that the term \(\mathcal{P}_{n,\epsilon}(\Phi)\) inside Equation (4.9) is replaced by \(\mathcal{P}_{n,\epsilon}(\Phi)\).

**Proof.** It follows by Theorem 1 that \(\tilde{Q}_\epsilon\) is a minorizer of \(Q_\epsilon\). Then the result follows by the ascending property of the minorization-majorization algorithm.

Next, we decompose \(\tilde{Q}_\epsilon(\Phi; y, X, \Phi^{(l-1)})\) into the following terms:

\[
\tilde{Q}_\epsilon\left(\Phi; y, X, \Phi^{(l-1)}\right) = S^{(l)}(x) + T^{(l)}(\beta, \phi) + V^{(l)}(\theta, v),
\] (4.15)

where

\[
S^{(l)}(x) = \sum_{i=1}^{n} \sum_{j=1}^{g} c_{ij}^{(l)} \log \pi_j(x; z) - \sum_{k=1}^{K_n} \tilde{p}_{2n}\left(\|e_{Tn}^T x\|_{2,\varepsilon}^2; \lambda_{2kn}\right),
\] (4.16)

\[
T^{(l)}(\beta, \phi) = \sum_{i=1}^{n} \sum_{j=1}^{g} \left( \log f(y_i; \exp \{v_{ij}^T x_i\}, \phi) + m_{ij}^{(l)} \log \tilde{f}\left(\gamma_{ij}^{(l)}, \log \tilde{y}_{ij}^{(l)}; \exp \{v_{ij}^T x_i\}, \phi\right) \right)
\] (4.17)

\[
- \sum_{k=1}^{K_n} \tilde{p}_{3n}\left(\|e_{Tn}^T x\|_{2,\varepsilon}^2; \lambda_{3kn}\right),
\]

\[
V^{(l)}(\theta, v) = \sum_{i=1}^{n} \log \frac{h(y_i; \theta, \exp \{v^T x_i\})}{1 - H(\tau; \theta, \exp \{v^T x_i\})} 1\{y_i > \tau\} - \sum_{k=1}^{K_n} \tilde{p}_{3n}\left(\|e_{Tn}^T v\|_{2,\varepsilon}^2; \lambda_{3kn}\right).
\] (4.18)

Update of \(x^{(l-1)}\) to \(x^{(l)}\) can be done by sequentially adopting the IRLS approach for \(j = 1, ..., g\).
the optimize function or the Newton-Raphson method, aiming to achieve

\[ \theta^{(l)} = \arg \max_{\theta > 0} V^{(l)}(\theta, v^{(l)}) . \]  

(4.23)

Because of Corollary 1, the M-step ensures \( Q_e(\Phi^{(l)}; y, X, \Phi^{(l-1)}) \geq Q_e(\Phi^{(l-1)}; y, X, \Phi^{(l-1)}) \) given a very small \( \epsilon > 0 \). The GEM algorithm is iterated until the observed data \( \epsilon \)-perturbed penalized log-likelihood is improved by less than a threshold of \( 10^{-2} \) or if the maximum number of iterations of 200 is reached.

4.3. Initialization Procedures

Initialization of parameters \( \Phi^{(0)} \) can be done using the clusterized method of moments approach proposed by Gui, Huang, and Lin (2018). It requires a \( K \)-means clustering method to assign observations \( y_i \) with \( y_i \leq \tau \) to one of the \( g \) subgroups for the body and observations \( y_i \) with \( y_i > \tau \) to the tail component. Then, we set initial parameters that match the first two moments for each mixture component, after initially fixing all of the regression parameters to be zero except for the intercepts. Refer to, for example, section 3.3.3 of Fung, Badescu, and Lin (2022) for more details.

4.4. Choice of the Number of Mixture Components

Usually, the choice of the number of mixture components \( g \) of the body can be determined based on standard specification criteria, including Akaike’s information criterion (AIC) and the Bayesian information criterion (BIC). However, for our motivating dataset, which will be described in Section 6, the AIC and BIC criteria would both lead to an excessively large number of components that would significantly impede the model interpretability. The main reason for obtaining a large \( g \) is that the claim severity distribution has many small nodes for small claim amounts (i.e., less than 10,000), as evidenced by Figure 1 in Section 5.1. Excessive fitting and modeling of such small claim amounts does not bring much insight from an insurance rate-making perspective because such small claims could even be modeled by an empirical distribution. As a result, for this particular dataset, we adopt a qualitative method that chooses \( g \) as the minimum number of components required for the proposed model to capture all nodes above a claim severity threshold of 10,000. The moderate and large claims exceeding the severity threshold dominate the total claim amount; that is, they constitute a large part of the total claim amount; see also Section 5.
4.5. Selection of Variables

The proposed GEM algorithm with group-fused penalty functions shrinks some regression coefficients to zero and merges some coefficients across different levels of a categorical variable. As a variant of the PIRLS approach, the proposed algorithm also caters for a wide range of concave penalty functions. However, as pointed out by Devriendt et al. (2021), the parameters obtained by the proposed algorithm are not exact. Therefore, in order to select the variables and reduce model complexity, after every model fit we need to perform an automated adjustment algorithm to remove parameters very close to zero and merge the parameters when their values are very close to each other.

Denote the fitted model parameter as \( \hat{\Phi} = (\hat{x}, \hat{\beta}, \hat{\phi}, \hat{\theta}, \hat{v}) \). Also, let \( \hat{z}_{ij}, \hat{m}_{ij}, \hat{y}_{ij}, \) and \( \hat{\log y}_{ij} \) be \( \hat{z}_{ij}^{(l)}, \hat{m}_{ij}^{(l)}, \hat{y}_{ij}^{(l)}, \) and \( \hat{\log y}_{ij}^{(l)} \) obtained by the E-step using the fitted parameters. Define the partial log-likelihood functions \( S(x) \), \( T(\beta, \phi) \), and \( V(\theta, v) \), respectively, for the mixing probabilities, body distributions, and tail distribution analogous to Equations (4.16) to (4.18) as follows:

\[
S(x) = \sum_{i=1}^{n} \sum_{j=1}^{g+1} \hat{z}_{ij} \log p_{ij}(x_{i}; x) - \sum_{k=1}^{K_{x}} p_{1n} \left( \| e_{ik}^{T} x_{i} \|_{2, \varepsilon} ; \lambda_{1n} \right) =: S_{0}(x) - P_{k_{1}, n, \varepsilon}(x), \tag{4.24}
\]

\[
T(\beta, \phi) = \sum_{i=1}^{n} \sum_{j=1}^{g} \hat{z}_{ij} \left[ \log f(y_{i}; \exp \{ \beta_{j}^{T} x_{i} \}, \phi_{j} \} + \hat{m}_{ij} \log \hat{f}(\hat{y}_{ij}, \hat{\log y}_{ij}^{(l)}; \exp \{ \beta_{j}^{T} x_{i} \}, \phi_{j} \} \right] \tag{4.25}
\]

\[
- \sum_{k=1}^{K_{x}} p_{2n} \left( \| e_{ik}^{T} \beta \|_{2, \varepsilon} ; \lambda_{2n} \right) =: T_{0}(\beta, \phi) - P_{k_{2}, n, \varepsilon}(\beta),
\]

\[
V(\theta, v) = \sum_{i=1}^{n} \log \frac{h(y_{i}; \theta, \exp \{ v^{T} x_{i} \})}{1 - H(\tau; \theta, \exp \{ v^{T} x_{i} \})} 1 \left\{ y_{i} > \tau \right\} - \sum_{k=1}^{K_{x}} p_{3n} \left( \| e_{ik}^{T} v_{e} \|_{e} ; \lambda_{3n} \right) =: V_{0}(\theta, v) - P_{k_{3}, n, \varepsilon}(v). \tag{4.26}
\]

The general principle of the automated adjustment algorithm is to fine-tune the regression parameters, so that the regression parameters (which are close to zero or very close to each other) are shrunked or merged in exact. Because fine-tuning of parameters would lead to another source of error, the automated adjustment algorithm needs to ensure that the likelihood-based quantities displayed above would not change significantly due to fine-tuning. We provide the step-by-step algorithm to Appendix C.4.

4.6. Tuning of Hyperparameters

The remaining problem is to select appropriate tuning parameters \( \lambda := (\lambda_{1}, \lambda_{2}, \lambda_{3}) \) that control the model complexity and select variables useful for explaining different parts of the claim severity distribution. The current theory in Appendix A only provides guidance on the order of \( \lambda \), but in application it is obvious that grid search on \( \lambda \) is computationally prohibitive.
because of the curse of dimensionality. As a result, we adopt an adaptive standardization approach similar to Devriendt et al. (2021), where we restrict $\lambda_{1,m} = w_{1k}\lambda_1$, $\lambda_{2,0} = w_{2k}\lambda_2$, and $\lambda_{3,0} = w_{3k}\lambda_3$. Here,

$$w_{1k} = w_k^{(ad)} w_k^{(st)}, \quad w_{2k} = w_k^{(ad)} w_k^{(st)}, \quad w_{3k} = w_k^{(ad)} w_k^{(st)}, \quad \text{with} \quad w_k^{(st)} = \frac{p_G - 1}{r_G} \sqrt{\frac{n_{d_1} + n_{d_2}}{n}},$$

(4.27)

where $w_k^{(ad)} = \|e_{1k} \hat{\beta}\|_2^{-1}$, $w_k^{(ad)} = \|e_{2k} \hat{\beta}\|_2^{-1}$, and $w_k^{(ad)} = \|e_{3k} \hat{\phi}\|_2^{-1}$ are the adaptive terms, and $w_k^{(st)}$ is the standardization term. Note here that the estimated parameters $\hat{x}$, $\hat{\beta}$, and $\hat{\phi}$ are obtained on the fitting procedures obtained in Section 4.2 starting with very small tuning parameters $\lambda$ (or even $\lambda = 0$). $d_1$ and $d_2$ are the two categories that the 4th penalty term is attempting to merge for categorical variables, and $(n_{d_1}, n_{d_2})$ is the number of observations being classified to those respective categories. $p_G$ is the number of categories for the respective explanatory variable, and $r_G$ is the number of penalty terms for the respective explanatory variable. Note that $r_G = p_G - 1$ for ordinal variables and $r_G = p_G(p_G - 1)/2$ for nominal variables. For continuous variables, we set $w_k^{(st)} = 1$ instead. The motivation of adopting the adaptive standardization approach is as follows:

- The adaptive weights $w_k^{(ad)} (l = 1, 2, 3)$ mitigate the scaling effects of regression coefficients, where the order of magnitude of some regression coefficients may be larger than the others. Also, Zou (2006) showed that adaptive weights facilitate more efficient shrinkage or merger of regression coefficients, achieving the oracle property.
- The standardization weight $w_k^{(st)}$ addresses the issue of level imbalances (with the term $\sqrt{(n_{d_1} + n_{d_2})/n}$, where some levels may contain more observations than others. Furthermore, the term $(p_G - 1)/r_G$ in $w_k^{(st)}$ addresses the imbalances of the number of penalty terms involved in an explanatory variable.

After specifying the weights, we can perform a grid search on $(\lambda_1, \lambda_2, \lambda_3)$ to find optimal tuning parameters. Motivated by the likelihood-based deviance approach by Khalili (2010), we propose the following method, which allows us to perform a separate grid search for $\lambda_1$, $\lambda_2$, and $\lambda_3$.

After obtaining the fitted model parameters $\Phi$ starting with a small $\lambda$, we compute the estimated latent variables $\hat{z}_{ij}$, $\hat{m}_{ij}$, $\hat{y}_{ij}$, and $\log \hat{y}_{ij}$, outlined in Appendix C.2.1, and we assume that they are fixed during the whole process of grid search. Then, for each $\lambda_1$, $\lambda_2$, and $\lambda_3$ within separately specified (one-dimensional) grids, we refit the models by maximizing the penalized partial log-likelihood functions $S(\mathbf{z})$ in Equation (4.24) (which only requires iterating Eqs. [4.19]), $T(\mathbf{\beta}, \mathbf{\phi})$ in Equation (4.25) (iterating Eqs. [4.20] and [4.21]) and $V(\theta, \nu)$ in Equation (4.26) (iterating Eqs. [4.22] and [4.23]). We denote the resulting fitted parameters as $\hat{x}(\lambda_1)$, $\hat{\beta}(\lambda_2)$, $\hat{\phi}(\lambda_2)$, and $\hat{\theta}(\lambda_3)$, $\hat{\nu}(\lambda_3)$. Because the parameters $\mathbf{z}$, $\mathbf{\beta}$, and $\mathbf{\nu}$ are estimated separately, our proposed algorithm avoids repeating the whole GEM procedure over a multidimensional grid of $(\lambda_1, \lambda_2, \lambda_3)$ which is computationally prohibitive. Optimal $\lambda_1$, $\lambda_2$, $\lambda_3$, can be determined by various selection criteria, where in this article we will present partial AIC (pAIC), partial BIC (pBIC), and K-fold cross-validation (CV). For the pAIC or pBIC approach, we define

$$pAIC_1(\lambda_1) = -2S_0(\hat{x}(\lambda_1)) + 2N_1(\lambda_1) \log n,$$

(4.28)

$$pAIC_2(\lambda_2) = -2T_0(\hat{\beta}(\lambda_2), \hat{\phi}(\lambda_2)) + 2N_2(\lambda_2) \log n,$$

(4.29)

$$pAIC_3(\lambda_3) = -2V_0(\hat{\theta}(\lambda_3), \hat{\nu}(\lambda_3)) + 2N_3(\lambda_3) \log n,$$

(4.30)

where $N_1(\lambda_1)$, $N_2(\lambda_2)$, and $N_3(\lambda_3)$ are the effective number of parameters (i.e., the number excluding zeroes and redundant parameter values) for $\hat{x}(\lambda_1)$, $\hat{\beta}(\lambda_2)$, $\hat{\phi}(\lambda_2)$, and $\hat{\theta}(\lambda_3)$, $\hat{\nu}(\lambda_3)$, respectively. Recall that $n_b$ and $n_t$ are the number of observations allocated to body and tail components, respectively. Now, $\lambda_1$, $\lambda_2$, and $\lambda_3$ can be chosen one at a time through minimizing the pAICs or pBICs.

For K-fold CV, we partition the data into K disjoint folds and measure the performance on each fold after training the remaining K−1 folds. The performance metric we use in this article is called “partial deviance”, given by

$$pD_1(\lambda_1) = -2S_0(\hat{x}(\lambda_1)); \quad pD_2(\lambda_2) = -2T_0(\hat{\beta}(\lambda_2), \hat{\phi}(\lambda_2)); \quad pD_3(\lambda_3) = -2V_0(\hat{\theta}(\lambda_3), \hat{\nu}(\lambda_3)).$$

(4.31)

For $l = 1, 2, 3$, the optimal $\lambda_l$ is the largest value such that the corresponding partial deviance is within one standard deviation of its minimum.
Because the regularization functions make the estimated parameters of the fitted model biased toward zero, it is important to collapse regression parameters \(\hat{a}^{(k_1)}, \hat{b}^{(k_2)}, \hat{m}^{(k_3)}\) and covariate matrix \(X\) and re-estimate the model with penalties excluded (i.e., \(k = 0\)), using the full GEM algorithm outlined in Section 4.2. This will also update the estimated latent variables for better accuracy. A related approach can be found by Devriendt et al. (2021).

5. MOTIVATING DATASET: GREECE MTPL CLAIM AMOUNTS

In this section, we present a dataset that motivates the proposed modeling and feature selection framework described above. The characteristics of the dataset are first described in Section 5.1. Then, we fit some state-of-the-art models to the dataset in Section 5.2 to show the necessity of adopting the proposed modeling framework. The complete fitting results under the proposed modeling framework will be discussed in Section 5.3.

5.1. Data Description

The dataset for our study was kindly provided by a major insurance company operating in Greece. It consists of 64,923 MTPL insurance policies with non-zero property claims for underwriting years 2013 to 2017. The sample comprised policy-holders with complete records; that is with the availability of all explanatory variables under consideration, and with at least one reported accident over the five underwriting years. These explanatory variables are summarized in Table 1.

An exploratory analysis was carried out in order to identify the challenges that need to be surmounted to efficiently model these property damage claim costs based on the subset of explanatory variables with the highest predictive power. Firstly, as we observe from Figures 1 and 2, the empirical claim severity distribution is multimodal and heavy-tailed. In particular, the empirical density plot of the claim amounts in the left panel of Figure 1 shows that there are at least three major nodes or clusters in the empirical density function: one for small claim severities of \(<10,000\), one for claim severities of about 30,000, and one for claim severities of about 80,000–100,000. Note that 54.6% of the claims are at least moderate (\(>10,000\)), but these claims constitute 98.4% of the total claim amount. Additionally, the density for the log claim amounts in the right panel of Figure 1 reveals even more complex distribution characteristics, especially for small claim sizes. Furthermore, regarding the

| Name          | Short description | Categories | Type       | Category Descriptions                      |
|---------------|-------------------|------------|------------|--------------------------------------------|
| DriverAge     | Driver’s age      | 18–74      | Continuous | From 18 to 74+ years old                   |
| VehicleBrand  | Automobile brand  | B1–B31     | Unordered  | 31 different brands                        |
| CC            | Car cubism        | 0–18       | Ordered    | 19 different categories                    |
| PolicyType    | Policy Type       | C1         | Ordered    | Economic type–only MTPL coverage           |
|               |                   | C2         |            | Middle type–includes other types           |
|               |                   | C3         |            | Expensive type–own coverage                |
| FHP           | Automobile horsepower | 1–13     | Ordered    | 13 categories of horsepower                |
| InsuranceDuration | Insurance duration | ID1    | Ordered    | Up to 5 years                             |
|               |                   | ID2        |            | From 6 to 10 years                         |
|               |                   | ID3        |            | Greater than 10 years                      |
| PaymentWay    | Payment way       | C1         | Unordered  | Cash                                      |
|               |                   | C2         |            | Credit card                               |
| Region        | City population   | 1–2; 4–14  | Unordered  | 13 Administrative Regions of Greece        |
| VehicleAge    | Vehicle age       | C1         | Ordered    | New car (up to 7 years old)                |
|               |                   | C2         |            | Middle (from 8 to 15 years old)            |
|               |                   | C3         |            | Old (greater than 15 years old)            |
| SumInsured    | Sum insured       | C1         | Ordered    | Up to 5000 euros                          |
|               |                   | C2         |            | Between 5001 and 10,000 euros              |
|               |                   | C3         |            | Greater than 10,000 euros                  |

Note: *Though driver’s age is by nature a continuous variable, in data analysis of this article we treat it as an ordered categorical variable with 57 levels instead.*
heavy-tailed nature of the data, the log-log plot in the left panel of Figure 2 seems asymptotically linear (an asymptotic red straight line is fitted) with a slope of roughly \(-1.3\) (this represents the tail index \(\xi\) of the empirical distribution; see, e.g., chapter 4.6.5 of Resnick 2007). Furthermore, the mean excess plot (see, e.g., chapter 6.2.2 of Embrechts, Klüppelberg, and Mikosch 1997) which is depicted in the right panel of Figure 2 appears linear when the claim size exceeds a threshold of around 270,000 (black vertical line), with asymptotic slope of 3.35, which also suggests \(\xi \approx 1.3\). Secondly, as far as the explanatory power of the variables is concerned, we studied the influence of the explanatory variables on the claim amounts through plotting the empirical density plots across each level of each variable. The results for the variables driver’s age,
insurance duration, payment method, and policy type are displayed in Figure 3. For example, we select only the losses with the driver’s age greater than 65 years, and the green curve in the top left panel gives the resulting empirical density (of the selected observations only). From Figure 3, we observe that the density functions are generally quite closely agreed across different levels of each variable, reflecting that the covariates are not decisive in determining the overall shape of the claim severity distribution. However, we still see that some variables have some apparent effects on the peak (or probability) of each cluster instead of the position of each cluster. For example, from the bottom right panel, the policy with “expensive type” has a higher probability assigned to the tail cluster and lower probabilities assigned to the remaining body clusters. Finally, it should be noted that considering all explanatory variables to be categorical, the 10 explanatory variables lead to 137 covariates in total. Therefore, because, previously mentioned, the impact of covariates on the claim severity distribution could be multifold (e.g., covariates may affect cluster assignment probabilities, average claim severity given a particular cluster and/or tail-heaviness), an appropriate regression model for this Greek MTPL dataset should contain multiple regressors. Obviously, this will lead to a large number of parameters without parameter regularizations, which can potentially result in an over fitting problem and impede model interpretations. Therefore, these issues outline the importance of variable selection.

5.2. Preliminary Model Fitting

In this subsection we first explore probability distributions that may appropriately fit the distribution of claim amounts, ignoring the effects of covariates. Because we observed multiple nodes and heavy-tailed characteristics of the claim amount distribution, it is natural to consider a finite mixture model with both light- and heavy-tailed mixture components to capture such characteristics. Motivated by Blostein and Miljkovic (2019), who proposed a finite mixture of various classes of distributions, a plausible benchmark is a mixture-Gamma Lomax model where the claim severity \( Y \) is modelled by a density function

\[
h_Y(y; \pi, \mu, \phi, \theta, \eta) = \sum_{j=1}^{g} \pi_j f(y; \mu_j, \phi_j) + \pi_{g+1} h(y; \theta, \zeta),
\]

where \( \pi = (\pi_1, ..., \pi_{g+1}) \) are the mixture probabilities and \( \mu = (\mu_1, ..., \mu_g) \) and \( \phi = (\phi_1, ..., \phi_g) \) are the mean and dispersion parameters. \( f \) is the Gamma density function for modeling the body, and \( h \) is the Lomax density function for modeling the tail given by Equations (2.3) and (2.4), respectively.

Observing three major nodes in the density shown in the left panel of Figure 1, we first start with the above model with \( g = 3 \) components for the body. Summary statistics are shown in Table 2, which compares the above model to several other models existing in the literature, including:

- Classical unimodal models, such as the Gamma (GA), Weibull (WEI), Weibull type three (WEI3), generalized Gamma (GG), and generalized Pareto (GP) distributions.
- Finite mixture of actuarial distributions (Miljkovic and Grün 2016) with four components, including Weibull mixture (Mix-Wei), Lognormal mixture (Mix-LN), and Burr mixture (Mix-Burr). We apply the flexmix package in \( R \) (Leisch 2004) to fit the models.

| Model               | DF | Log-likelihood | AIC    | BIC    |
|---------------------|----|----------------|--------|--------|
| GA                  | 2  | −742,516       | 1,485,036 | 1,485,054 |
| WEI                 | 2  | −740,360       | 1,480,725 | 1,480,743 |
| GG                  | 3  | −740,248       | 1,480,503 | 1,480,530 |
| GP                  | 2  | −748,596       | 1,497,197 | 1,497,215 |
| Mix-Wei (4 components) | 11 | −740,275       | 1,480,573 | 1,480,673 |
| Mix-LN (4 components) | 11 | −741,940       | 1,483,902 | 1,484,002 |
| Mix-Burr (4 components) | 15 | −725,620       | 1,451,270 | 1,451,406 |
| NPMLE               | 9  | −732,695       | 1,465,409 | 1,465,491 |
| 3-Gamma Lomax       | 11 | −723,447       | 1,446,917 | 1,447,017 |
The corresponding density functions are given by Equations (1.1) to (1.8) of the Supplementary Material.

The results show that classical unimodal models provide unsatisfactory fits to the data observing that the log-likelihoods are very low. Despite the distributional multimodality, Weibull and lognormal mixture models still do not show substantial fitting performance compared to the classical models because they fail to extrapolate the tail-heaviness of the claim severities. Though the Burr mixture and NPMLE fit much better than all other preliminary models, the three-component mixture-Gamma Lomax model performs the best. Therefore, it is necessary and important to adopt a mixture-based model that contains both light- and heavy-tailed components to capture distributional multimodality and the mismatch between body and tail behavior.

However, one drawback of the mixture-Gamma Lomax model is the instability of the estimation of the implied tail index \( \xi_0 := \exp \{ \nu_0 \} \), which corresponds to the parameter \( \xi \) in the Lomax distribution (Eq. [2.4]). Here, \( \nu_0 \) is the log-transformed implied tail index. Model fitting has been tested across various numbers of Gamma components \( g \) and we examine how robust the estimates of tail-heaviness across different \( g \)s are. The results are shown in Table 3. We see that the implied tail index \( \xi_0 \) fluctuates greatly from smaller than 1.5 to greater than 1.8 across \( g \), which does not make sense in practice because \( g \) should control the body part of severity distribution only and bring very little impact on the estimated tail index. The main reason of seeing such an undesirable phenomenon is that the Lomax distribution, which is designed to capture the tail distribution, also calibrates to the body of the distribution and the MLE approach is not very stable in estimating the tail parameter. This motivates the use of the composite model proposed in Section 2, where the tail component only interacts with the body via the mixture probability. Also, note that though both AIC and BIC suggest a bigger number of components for the body (the optimal \( g \) goes way beyond 15), this mainly reflects improvements of fitting small claims below 3000. This should not be over-weighted because exact prediction of these small claims is less relevant in pricing, and excessive model complexity may impede interpretability. As a result, AIC and BIC may be less appropriate in determining the number of mixture components under this dataset.

### 5.3. Fitting Results

We hereby analyze the performance under the proposed mixture composite model with multi-type feature regularization for the covariates.

#### 5.3.1. Distributional Fitting

As in the preliminary analysis, we first fit the distribution of claim amounts under the proposed modeling framework, without considering covariates. Notice that in the mean excess plot of claim amounts (right panel of Figure 2) under the preliminary analysis, the plot becomes linear beyond claim severity of 270,000 indicated by the vertical line of the plot. As a result, a reasonable choice of the splicing threshold is \( \tau = 270,000 \). After fitting the proposed model across various choices of the number of body components \( g \), we find that \( g = 5 \) is the minimum number of components required to capture all of the density nodes above a claim severity of 10,000. The summary statistics of the fitted model is presented in Table 4, the fitted versus empirical density plots are shown in Figure 4, and the Q-Q and log-log plots of claim sizes are illustrated in Figure 5. The model estimated tail index of 1.3817 roughly resembles that estimated by the asymptotic slope of the log-log plot, which is 1.3 (left panel of Figure 2). Also, as expected, we find that the model-estimated tail index is robust across various choices of \( g \). The empirical and fitted densities in Figure 4 agree quite closely, reflecting a satisfactory goodness-of-fit of the flexible
mixture-Gamma Lomax model. In particular, the fitted distribution captures all nodes representing a larger claim amounts, with multiple small nodes for smaller claims explained smoothly by one single component (to be precise, by the subgroup $j = 1$ indicated by Figure 4). From the Q-Q plot, we see that the fitting performance is satisfactory except for immaterial claims ($y < 100$). The fitted versus empirical log-log plot also indicates satisfactory fitting performance for the tail part. The fitted log-likelihood is $-719,309.1$, with $\text{AIC} = 1,438,652$ and $\text{BIC} = 1,438,640$, which is slightly superior compared to the 5-Gamma Lomax distribution illustrated in Table 3; that is, we slightly prefer this splicing model over the plain-vanilla mixture case.

5.3.2. Effects of the Covariates

We now include all variables described in Table 1 and fit our proposed mixture composite regression model with LASSO and SCAD regularizations. Because all variables are included as categorical covariates, there is a total of $D = 138$ parameters for each set of regressors. The grid search is performed on a very wide range of tuning parameters (we chose $\lambda_1 \in \{0.1 \cdot 2^0/n, 0.1 \cdot 2^1/n, ..., 0.1 \cdot 2^{16}/n\} = \{0.1/n, 0.2/n, ..., 6553.6/n\}$, $\lambda_2 \in \{0.1/n_b, 0.2/n_b, ..., 6553.6/n_b\}$, and

| Subgroup $j$ | 1    | 2    | 3    | 4    | 5    | 6    |
|-------------|------|------|------|------|------|------|
| Classification probability | 0.3790 | 0.0450 | 0.1184 | 0.2111 | 0.2099 | 0.0365 |
| Subgroup mean | 1,339  | 9,184 | 27,527 | 71,439 | 88,818 | 616,666 |
| Subgroup dispersion $\phi_j$ | 0.0331 | 0.9985 | 0.0119 | 0.0169 | 1.3734 | 1.3817 |

FIGURE 4. Empirical vs. Fitted Density of Claim Amounts (Left Panel) and Log Claim Amounts (Right Panel).

FIGURE 5. Left Panel: Q-Q Plot; Right Panel: Empirical (Blue Dots) vs. Fitted (Red Dots) Log-Log Plot of Claim Amounts.
\[ \lambda_3 \in \{0.1/n_1, 0.2/n_1, \ldots, 6553.6/n_1\} \] to find optimal tuning parameters. The fitting performances for different model settings (without regression versus with regression), penalty settings (without regularization versus with regularization), and model selection criteria (pAIC, pBIC, or CV with one standard deviation rule) are summarized in Table 5. With a large number of covariates, we first note from the table that regularization of regression coefficients is a must; otherwise, some parameters diverge to very large values (due to overfitting), causing the algorithm to collapse eventually because of numerical instability (spurious solutions). As a result, for a full model as a benchmark for comparison, we need to apply a weak LASSO penalty, which sets relatively small \[ \hat{\lambda}_l > 0 \ (l = 1, 2, 3) \] such that no covariates are removed or merged. For this dataset, we find that the choices of \[ \hat{\lambda}_1 = 10/n, \hat{\lambda}_2 = 10/n_0, \text{ and } \hat{\lambda}_3 = 10/n \] achieve such a purpose. We next investigate the effect of the model selection criteria to the resulting fitted model. For both LASSO and SCAD penalties chosen as regularization function, pAIC results to very large models with a total of \( N = 809 \) parameters for LASSO and \( N = 613 \) for SCAD, indicating that many variables have predictive power on explaining all parts (body, tail, and subgroup probabilities) of the claim severity distribution. The large number of predictors, however, makes the fitted models very difficult to interpret. Also, the selected model severities can vary greatly across various choices of initializations or grids for tuning parameters, because we find that model sizes within a range of about 150 to 1000 parameters all have very similar AICs. In contrast, pBIC heavily penalizes the regression parameters and leads to a very small fitted model that chooses very few or even no covariates.

On the other hand, using CV with a one standard deviation rule provides fitted models with more reasonable complexity (\( N = 112 \) under LASSO or \( N = 197 \) under SCAD). Both LASSO and SCAD penalties suggest that there is not any systematic effect in the tail that is explained by the available variables. On the other hand, both penalty functions reveal similar sets of variables important to explain the body and subgroup probability parts. The higher model complexity under SCAD is mainly due to more granular mergers among different levels of some variables (such as driver’s age). Under LASSO, the resulting AIC under the CV approach is close to that under the corresponding pAIC approach, whereas the BIC is just slightly inferior to the pBIC approach.

Table 5 also shows the performance of the LASSO and SCAD CV-selected models after the model refit procedure. Recall from Section 4.6 that the refitting procedure involves re-estimation of parameters for the shrinked model with regularization terms excluded to reduce biasedness in the estimated parameters. For the LASSO penalty, the improvements of the log-likelihood, AIC, and BIC are all expected after refitting. For the SCAD penalty, because the concavity of SCAD penalty function already mitigates the biasedness of estimated parameters (Fan and Li 2001), there is no apparent improvement of the fitting performance after performing the refitting procedure. After refitting, the LASSO penalty approach results to superior fitting performance compared to the SCAD approach, as evidenced by lower AICs and BICs. As a result for conciseness concern, we focus solely on the CV approach with LASSO penalty as our model selection criterion in the following analysis.

The final refitted model suggests that the subgroup probabilities \( \pi_j(x; \theta), 1 \leq j \leq g + 1 \), are influenced by the variables as follows.

- Driver’s age: The model merges this variable into six categories: \{18 – 30, 31 – 34, 35 – 41, 42 – 51, 52 – 72, 73+\}.
- Car cubism: three categories: \{0 – 12, 13 – 14, 15 – 18\}.

**TABLE 5**

| Model selection criteria            | No. of parameters | Log-likelihood | AIC       | BIC       |
|-------------------------------------|-------------------|----------------|-----------|-----------|
| \( \mathcal{L}_n \) with without regression | 17                | -719,309       | 1,438,652 | 1,438,807 |
| \( \mathcal{L}_n \) with without penalty | NA               | NA             | NA        | NA        |
| \( \mathcal{L}_n \) + weak penalty only | 1,524             | -719,969       | 1,438,987 | 1,452,826 |
| \( \mathcal{L}_n \) + LASSO penalty w/ pAIC before refit | 809               | -718,312       | 1,438,242 | 1,445,589 |
| \( \mathcal{L}_n \) + LASSO penalty w/ pBIC before refit | 42                | -719,139       | 1,438,362 | 1,438,743 |
| \( \mathcal{L}_n \) + LASSO penalty w/ CV before refit | 112               | -719,029       | 1,438,282 | 1,439,299 |
| \( \mathcal{L}_n \) + LASSO penalty w/ CV after refit | 112               | -718,779       | 1,437,781 | 1,438,798 |
| \( \mathcal{L}_n \) + SCAD penalty w/ pAIC before refit | 613               | -718,324       | 1,437,873 | 1,443,439 |
| \( \mathcal{L}_n \) + SCAD penalty w/ pBIC before refit | 17                | -719,309       | 1,438,652 | 1,438,807 |
| \( \mathcal{L}_n \) + SCAD penalty w/ CV before refit | 197               | -718,925       | 1,438,244 | 1,440,033 |
| \( \mathcal{L}_n \) + SCAD penalty w/ CV after refit | 197               | -718,925       | 1,438,244 | 1,440,033 |

\[ n_1 = n_0 + n_6 = 6553.6, \quad n_2 = 10, \quad n_3 = 10, \quad n_4 = 10, \quad n_5 = 10. \]
Policy type: Expensive type causes higher probability of a claim falling into the tail component.

Horsepower: two categories: \( f_{13}, 4 \). Larger tail probability for higher horsepower.

Payment method: Cash payment results in higher tail probability.

Region: four regions (Regions 4, 8, 9, 12) differs the subgroup probabilities from other regions.

Car brand, insurance duration, vehicle age, and sum insured: No significant impacts.

These results are also presented by plots in Figure 6, which display the probability being classified to tail component versus various variables. The points indicated as triangle (Δ) and square ([ ]) correspond to the fitted and empirical probabilities, respectively. As we can see from the figure, conditioned on any categories/levels of any explanatory variables, the fitted and empirical probabilities match very well, reflecting the ability of the proposed regression model to capture well the covariates, influence. The green dotted line is the overall empirical tail probability across all observations. The blue and red intervals are,
respectively, the 95% Wald-type and Efron bootstrap confidence intervals (CIs) derived in Appendix A. The CIs generated by the two approaches reconcile well.

The model chooses a smaller set of variables which are important in explaining the body distributions $f(y_i; \exp \{ \beta_j^T x_i, \phi_j \})$, reflecting more heterogeneity among subgroup probabilities than within-subgroup average claim sizes:

- Driver’s age: three categories: $\{ 18 - 28, 29 - 69, 70+ \}$.
- Car cubism: two categories: $\{ 0 - 15, 16 - 18 \}$.
- Payment way: Cash payment results in a generally higher within-subgroup mean claim severity.
- Region: two categories: $\{ \text{Region 4, Others} \}$.
- Other variables are excluded.

Finally, the fitted model suggests that none of the explanatory variables are significantly influential to the tail distribution $h(y_i; \theta, \exp \{ v^T x_i \})$.

Overall, the effects on various variables to the average claim severity (Eq. [2.5]) are demonstrated in Figure 7. Note that the average claim severity depends not only on the body distributions $f(y_i; \exp \{ \beta_j^T x_i, \phi_j \})$ but also on the subgroup probabilities $\pi_j(x; \alpha)$. As a result, though the policy type and horsepower bring no influence on the body distributions, they still substantially impact the expected claim amount because of their influence on the subgroup probabilities.

5.3.3. Comparison Studies

To further justify the advantages of using the proposed mixture composite modeling framework with multi-type variable selection approach, we perform comparison studies with the following state-of-the-art modeling frameworks:

- Generalized linear model (GLM) with multi-type variable selection (Oelker and Tutz 2017): Choosing the Gamma distribution, it is a special case of the proposed model with $g = 1$ and $\tau \to \infty$.
- Finite mixture of regression (FMR) model (Frees, Derrig, and Meyers 2014): Choosing the same (mixture composite) component functions as the proposed model, it is a special case of the proposed model with $\pi_j(x; \alpha) = \pi_j$ independent of $x$; that is no regression on the weight functions.
- Logit-weighted reduced mixture of experts (LRMoE) model (Fung, Badescu, and Lin 2019, 2021): Choosing the same component functions as the proposed model, it is a special case of the proposed model with $\exp \{ \beta_j^T x \} = \mu_j$ and $\exp \{ v^T x \} = \zeta$ independent of $x$; that is, no regression on the parameters of the component functions.

The fitting performances for different models (state-of-the-art and proposed models) and regularization approaches (no regression, no penalty, weak penalty, and LASSO penalty with CV) are summarized in Table 6. As expected, the GLM fits poorly to the dataset, as indicated by the high AIC and BIC. Following the approach of Oelker and Tutz (2017) to perform variable selection using LASSO (with CV approach), all variables are eliminated and the resulting fitted model recovers to a “no regression” model. It is because the estimated GLM parameters are heavily distorted by a few outliers, and hence the regression parameters tend to overfit the extreme claims, providing little or no predictive power. Therefore, the variable selection approach of Oelker and Tutz (2017) does not work if the distributional model class is severely misspecified.

Among the three mixture-based models (FMR, LRMoE, and the proposed model), the proposed mixture-composite model apparently performs the best. After conducting variable selections, the proposed model results in the lowest AIC (with a significant margin compared to other models), sacrificing the BIC only a bit compared to the FMR. Therefore, incorporating the covariates’ influences (with variable selections) on both subgroup probabilities and the body part of the severity distribution can substantially improve the predictive power of the fitted model.

5.3.4. Summary Findings

In this real data analysis, we get a deeper understanding of the influence of policyholder attributes on the claim severity distribution with a highly complex structure including multimodality and tail-heaviness. Using the proposed mixture composite modeling framework embedded with a variable selection approach, we find that the explanatory variables most prominently impact the subgroup probabilities of the severity distribution, explaining the unobserved heterogeneity of policyholder risk.

4Following the results presented in Table 5, the weak penalty incorporated in Table 6 is a LASSO penalty with $\lambda_1 = 10/n$, $\lambda_2 = 10/n_0$, and/or $\lambda_3 = 10/n_0$. 

---

416 T. C. FUNG ET AL.
profiles and/or claim types. Fewer variables explain well the body part of the distribution, reflecting relatively homogeneous claim severity distributions conditioned on the subgroups to which each claim belongs. This finding is in contrast to many traditional regression models widely adopted in actuarial practice, including GLM and GAM, where regression links are set to capture the systematic effects in distributions instead of the subgroup heterogeneity.

Further, we do not find any variables significantly influencing the tail-heaviness of the claim severity distribution, which may be the result of the scarcity of large claims (only around 2400 claims exceed the splicing threshold $\tau$) to allow for statistically significant covariates influencing the tail part. This empirically verifies the legitimacy of actuarial practice where covariates, influence is often excluded in modeling large claims. In actuarial literature, we refer to Laudagé, Desmettre, and Wenzel (2019), who also refrained from incorporating a regression in the tail part of their severity distribution.

FIGURE 7. Average Claim Severities vs. Several Variables.
Finally, the significance of the proposed modeling framework is justified by showing its superior fitting performance to other advanced modeling tools, including the FMR and LRMoE.

6. DISCUSSION

In this article, we considered a mixture composite regression model for addressing several challenges when modeling claim severities such as multimodality and tail-heaviness of claims, extending the framework of Reynkens et al. (2017), who considered the case without covariates. For variable selection, we proposed a group-fused regularization approach. Our covariates may influence the mixture probabilities and the body and the tail of the claim size distribution in such way that model interpretability is preserved. This approach enables regularization under multi-type variable settings. For this setup, we developed an asymptotic estimation theory that justified the efficiency of the proposed method. In particular, we showed that the method we presented is (i) consistent in terms of covariate selection because when the sample size goes to infinity, it will merge and shrink correctly regression coefficients across all modeling parts and (ii) the parameters of the reduced model are asymptotically normal. The implementation was illustrated by a real data application that involved fitting claim size data from a Greek automobile insurance company. Maximum likelihood estimation of the model parameters was achieved through a novel GEM algorithm that was demonstrated to perform well.

Furthermore, it is worth noting that instead of following a data driven approach for selecting the number of mixture components in the body area based on specification criteria, as is done herein, an interesting direction of further research would be to extend the framework to a non-parametric maximum likelihood estimation approach that can be utilized for automated selection of the number of mixture components.

Finally, it is worth noting that though the proposed composite model mitigates instabilities of tail index estimations inherited by finite mixture models, selection of the splicing threshold is often subjective. Therefore, it would be worth to exploring alternative approaches for robust estimation of the tail index. One possible way is to modify the maximum likelihood approach for parameter estimation such that an observation with a larger claim severity has a higher relative importance in determining the model parameters. Another possible way is to explore models that bridge the gap between finite mixture models and composite models and share the advantages of both model classes.

ACKNOWLEDGMENTS

The authors thank the editor and two anonymous referees for their useful comments, which helped us to improve the article.
Bakar, S. A., N. A. Hamzah, M. Maghsoudi, and S. Nadarajah. 2015. Modeling loss data using composite models. *Insurance: Mathematics and Economics* 61:146–54. doi:10.1016/j.insmatheco.2014.08.008

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

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

Cooray, K., and M. M. Ananda. 2005. Modeling actuarial data with a composite lognormal-Pareto model. *Scandinavian Actuarial Journal* 2005 (5):321–34. doi:10.1080/0346123050099763

Devriendt, S., K. Antonio, T. Reynkens, and R. Verbelten. 2021. Sparse regression with multi-type regularized feature modeling. *Insurance: Mathematics and Economics* 96:248–61. doi:10.1016/j.insmatheco.2020.11.010

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

Fan, J., and R. Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96 (456):1348–60. doi:10.1198/016214501753382273

Frees, E. W. 2009. *Regression modeling with actuarial and financial applications*. New York: Cambridge University Press.

Frees, E. W., R. A. Derrig, and G. Meyers. 2014. *Predictive modeling applications in actuarial science*. Vol. 1. New York: Cambridge University Press.

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

Fung, T. C., A. L. Badescu, and X. S. Lin. 2021. A new class of severity regression models with an application to IBNR prediction. *North American Actuarial Journal* 25 (2):206–31. doi:10.1080/10920277.2020.1729813

Fung, T. C., A. L. Badescu, and X. S. Lin. 2022. Fitting censored and truncated regression data using the mixture of experts models. *North American Actuarial Journal*. doi:10.1080/10920277.2021.2031896

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

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

Jeong, H., H. Chang, and E. A. Valdez. 2021. A non-convex regularization approach for stable estimation of loss development factors. *Scandinavian Actuarial Journal* 2021 (9):779–803. doi:10.1080/03461238.2021.1882550

Khalili, A. 2010. New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics* 38 (4):519–39. doi:10.1002/cjs.10083

Khalili, A., and J. Chen. 2007. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* 102 (479):1025–38. doi:10.1198/016214507000000590

Laudagé, C., S. Desmettre, and J. Wenzel. 2019. Severity modeling of extreme insurance claims for tariﬁﬁcation. *Insurance: Mathematics and Economics* 88:77–92. doi:10.1016/j.insmatheco.2019.06.002

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

Leisch, F. 2004. Flexmix: A general framework for ﬁnite mixture models and latent class regression in R. *Journal of Statistical Software* 343:189–205. doi:10.1523/jos.343.2004.11.010

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

Nadarajah, S., and S. A. Bakar. 2014. New composite models for the Danish fire insurance data. *Scandinavian Actuarial Journal* 2014 (2):180–87. doi:10.1080/03461238.2012.695746

Nelder, J. A., and R. W. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* 135 (3):370–84. doi:10.2307/2344614

Oelker, M.-R., and G. Tutz. 2017. A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classiﬁcation* 11 (1):97–120. doi:10.1007/s11634-015-0205-y

Parodi, P. 2020. A generalised property exposure rating framework that incorporates scale-independent losses and maximum possible loss uncertainty. *ASTIN Bulletin* 50 (2):513–53. doi:10.1080/05335136.2020.110

Pigeon, M., and M. Denuit. 2011. Composite lognormal-Pareto model with random threshold. *Scandinavian Actuarial Journal* 2011 (3):177–92. doi:10.1080/03461231103690754

Resnick, S. I. 2007. *Heavy-tail phenomena: Probabilistic and statistical modeling*. New York: Springer Science & Business Media.

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

Scollnik, D. P. 2007. On composite lognormal-Pareto models. *Scandinavian Actuarial Journal* 2007 (1):20–33. doi:10.1080/03461230601110447

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

Seidel, W., K. Sever, and H. Sevečková. 2006. Efﬁcient calculation of the NPMLE of a mixing distribution for mixtures of exponentials. *Computational Statistics & Data Analysis* 50 (5):1248–71. doi:10.1016/j.csda.2004.11.016

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1):67–88.

Zou, H. 2006. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* 101 (467):1418–29. doi:10.1198/016214506000000735

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1):67–88.
APPENDIX A. ASYMPTOTIC PROPERTIES

This section presents two asymptotic theorems regarding the proposed mixture composite model with the feature selection method. We will only present the key results and discuss their implications in this Appendix. All construction details, including the assumptions and proof details, are postponed to Appendix B. Suppose \( Y_i \), given \( x_i \), is generated by the model of Equation (2.1) with true model parameter \( \Phi_0 = (\mathbf{x}_0, \mathbf{b}_0, \phi_0, \theta_0, \upsilon_0) \). We first have the following theorem.

Theorem A.1. Assume H1–H4 outlined in Appendix B.1 hold for the penalty functions \( p_{\ell n}(\psi; \lambda_{\ell n}) \). Let \( \mathbf{V}_i = (Y_i, x_i) \), \( i = 1, \ldots, n \), be a random sample from a density function \( h(\mathbf{v}; \Phi) \) that satisfies regularity conditions RI–RS5 outlined in Appendix B.2. Then, there exists a local maximizer \( \hat{\Phi}_n \) of the penalized log-likelihood function \( \mathcal{F}_n(\Phi) \) for which \( \|\hat{\Phi}_n - \Phi_0\|_2 = O_p(n^{-1/2}) \) as \( n \to \infty \).

As discussed in Appendix B.1, both LASSO and SCAD penalty functions can be constructed to satisfy all assumptions H1 to H4. Therefore, the above theorem says that the estimated parameters \( \hat{\Phi}_n \) under the proposed model setup will converge to the true model parameters as \( n \to \infty \).

In addition to consistency, it is important to show sparsity of the proposed feature selection method, enabling consistent variable selection. To do so, we need to linearly transform the parameter space and formulate the asymptotic properties in the transformed space. We first define \( C_l = (c_{l1}, \ldots, c_{lk}) \) as a design matrix of penalty coefficients. Further, denote \( Z_1 = \{k : \|c_{l1}\|_2 = 0\} \), \( Z_2 = \{k : \|c_{l2}\|_2 = 0\} \) and \( Z_3 = \{k : c_{l3} = 0\} \), representing the regression coefficients to be merged or shrunk. Without loss of generality we hereafter assume that under the true model, \( Z_l = \{1, 2, \ldots, s_l\} \) for \( l = 1, 2, 3 \), and we construct a reduced matrix \( \hat{C}_{\text{red}, l} := (c_{l1}, \ldots, c_{lD}) \) of the linearly independent vectors that span the space of the vectors \( \{c_{l1}, \ldots, c_{ln}\} \). Note that we always have \( J_l = \min(s_l, P) \). Further, we construct linearly independent vectors \( \hat{C}_{\text{ind}, l} := (c_{l1}, \ldots, c_{lD}) \) that are also linearly independent of all vectors in \( \hat{C}_{\text{red}, l} \). Then, define \( C_l = (\hat{C}_{\text{red}, l}, \hat{C}_{\text{ind}, l}) \), which is a \( D \times D \) full rank matrix, and define the transformed parameters \( \mathbf{x}^* = \hat{C}_T^l \mathbf{x}, \mathbf{b}^* = \hat{C}_T^l \mathbf{b} \) and \( \mathbf{v}^* = \hat{C}_T^l \mathbf{v} \). Note that the transformed parameters can also be decomposed as \( \mathbf{x}^* = (\mathbf{x}_{\text{red}}^T \mathbf{z}_{\text{ind}}^T)^T, \mathbf{b}^* = (\mathbf{b}_{\text{red}}^T \mathbf{z}_{\text{ind}}^T)^T, \) and \( \mathbf{v}^* = (\mathbf{v}_{\text{red}}^T \mathbf{z}_{\text{ind}}^T)^T, \) where \( \mathbf{x}_{\text{red}}^* = \hat{C}_{\text{red}, l}^T \mathbf{x}, \mathbf{b}_{\text{red}}^* = \hat{C}_{\text{red}, l}^T \mathbf{b}, \mathbf{b}_{\text{ind}}^* = \hat{C}_{\text{ind}, l}^T \mathbf{b}, \mathbf{v}_{\text{red}}^* = \hat{C}_{\text{red}, l}^T \mathbf{v}, \text{ and } \mathbf{v}_{\text{ind}}^* = \hat{C}_{\text{ind}, l}^T \mathbf{v}. \) The above mathematical construction allows us to rewrite the penalized log-likelihood as a function of the transformed parameters \( \Phi^* := (\mathbf{x}^*, \mathbf{b}^*, \phi, \theta, \upsilon)^T \) as follows:

\[
\mathcal{F}'_n(\Phi^*) = \mathcal{F}_n(\Phi) = \mathcal{L}'_n(\Phi^*) - \mathcal{P}'_n(\Phi^*),
\]

with log-likelihood \( \mathcal{L}'_n(\Phi^*) = \mathcal{L}_n(\Phi) \) and penalty term

\[
\mathcal{P}_n(\Phi^*) = \sum_{k=1}^{K_1} p_{\ell n}(\|\tilde{c}_{1k}^T \mathbf{x}^*\|_2; \lambda_{1kn}) + \sum_{k=1}^{K_2} p_{2n}(\|\tilde{c}_{2k}^T \mathbf{b}^*\|_2; \lambda_{2kn}) + P_{k, n}(\|\tilde{c}_{3k}^T \mathbf{v}^*\|_1; \lambda_{3kn}),
\]

where \( \tilde{c}_{ik} = \hat{C}_{\text{red}, l}^{-1} c_{ik} \) for \( l = 1, 2, 3 \). Suppose that the true model parameters are given by \( \Phi_0^* := (\mathbf{x}_0^*, \mathbf{b}_0^*, \phi_0, \theta_0, \upsilon_0^*) \). This can be decomposed as \( \Phi_0^* := (\Phi_{\text{red}, 0}^*, \Phi_{\text{ind}, 0}^*) \), with \( \Phi_{\text{red}, 0}^* := (\mathbf{x}_{\text{red}, 0}^*, \mathbf{b}_{\text{red}, 0}^*, \mathbf{v}_{\text{red}, 0}^*) \) and \( \Phi_{\text{ind}, 0}^* := (\mathbf{b}_{\text{ind}, 0}^*, \mathbf{b}_{\text{ind}, 0}^*, \mathbf{v}_{\text{ind}, 0}^*) \). Note that by construction \( \Phi_{\text{red}, 0}^* = 0 \). Finally, denote \( \Phi_n^* := (\Phi_{\text{red}, n}^*, \Phi_{\text{ind}, n}^*) \) as the corresponding estimator of model parameters. We have the following theorem, which is an extension of the oracle property given by Fan and Li (2001):
Theorem A.2. Assume that the conditions of Theorem A.1 are fulfilled and the conditions H1–H4 hold for $p_{ln}(\psi; \lambda_{ln})$. Then, for any $\sqrt{n}$-consistent local maximizer $\Phi_{n}^{*}$ of the regularized log-likelihood function $F_{n}^{*}(\Phi^{*})$ as $n \to \infty$, we have

(a) Consistency of feature selection: $P(\Phi_{\text{red}, n}^{*} = 0) \to 1$ as $n \to \infty$.

(b) Asymptotic normality:

$$\sqrt{n} \left[ \left[ I_{\text{ind}}^{*}(\Phi_{\text{ind}, 0}^{*}) - \mathcal{P}_{n}^{*}(\Phi_{\text{ind}, 0}^{*})/n \right] \left( \Phi_{\text{ind}, n}^{*} - \Phi_{\text{ind}, 0}^{*} \right) + \mathcal{P}_{n}^{*}(\Phi_{\text{ind}, 0}^{*})/n \right] \overset{d}{\to} N(0, I_{\text{ind}}^{*}(\Phi_{\text{ind}, 0}^{*}))$$

as $n \to \infty$, where $I_{\text{ind}}^{*}(\Phi_{\text{ind}, 0}^{*})$ is the Fisher information matrix and $\mathcal{P}_{n}^{*}(\Phi_{\text{ind}, 0}^{*})$ (respectively $\mathcal{P}_{n}^{*}(\Phi_{\text{ind}, 0}^{*})$) is the first (second) derivative of the penalty functions under the true (reduced) model after fixing $\Phi_{\text{red}, 0} = 0$.

The above theorem shows that when the sample size is large, the proposed feature selection method merges and shrinks parameters correctly with high probability. Moreover, the estimated parameters of the reduced model are asymptotically normal. Though detailed discussions are leveraged to Remark B.1 of Appendix B.4, the adjustment and bias terms $\mathcal{A}_{n}(\Phi_{\text{ind}, 0}^{*}, \Phi_{\text{ind}, 0}^{*})$ and $\mathcal{B}_{n}(\Phi_{\text{ind}, 0}^{*}, \Phi_{\text{ind}, 0}^{*})$ are both asymptotically negligible when the penalty function is LASSO with an adaptive approach (which was discussed in Section 4.6) or SCAD. For large sample sizes, the estimated parameters are approximately unbiased and we may approximate the variance of the estimated transformed parameters as

$$\text{Var}(\Phi_{\text{ind}, n}^{*}) \approx \frac{1}{n} \left[ \hat{I}_{\text{ind}}^{*}(\Phi_{\text{ind}, n}^{*}) \right]^{-1}, \quad (A.3)$$

where $\hat{I}_{\text{ind}}^{*}(\Phi_{\text{ind}, n}^{*})$ is the sample Fisher information of the reduced model. In other words, parameter uncertainty under the proposed feature selection method is equivalent to that under the reduced model after selecting the variables. With this regards, the construction of CIs of parameters is straightforward:

- Wald-type CIs: Denote $\psi_{0,q}^{*}$ as the $q$-th element of $\Phi_{\text{ind}, 0}^{*}$. A two-sided Wald-type CI for $\psi_{0,q}^{*}$ is

$$\left[ \psi_{0,q}^{*} - \frac{z_{1/2} \kappa_{q}}{\sqrt{n}} \sqrt{\left[ \hat{I}_{\text{ind}}^{*}(\Phi_{\text{ind}, n}^{*}) \right]_{q,q}^{-1}}, \psi_{0,q}^{*} + \frac{z_{1/2} \kappa_{q}}{\sqrt{n}} \sqrt{\left[ \hat{I}_{\text{ind}}^{*}(\Phi_{\text{ind}, n}^{*}) \right]_{q,q}^{-1}} \right], \quad (A.4)$$

where $z_{1/2}$ is the $\kappa$-quantile of the standard normal distribution and $\left[ \hat{I}_{\text{ind}}^{*}(\Phi_{\text{ind}, n}^{*}) \right]_{q,q}^{-1}$ is the $q$th diagonal element of $\left[ \hat{I}_{\text{ind}}^{*}(\Phi_{\text{ind}, n}^{*}) \right]^{-1}$. For other quantities of interest (e.g., mean claim amounts), one may apply a delta method or simulate parameters from $N(\hat{\Phi}_{\text{ind}, n}^{*}, \text{Var}(\hat{\Phi}_{\text{ind}, n}^{*}))$ to analytically or empirically approximate their CIs.

- Efron percentile bootstrap CIs: Consider a parametric bootstrap procedure that generates the bootstrap samples $\{(y^{(b)}, X)\}_{b=1,...,B}$, where $y^{(b)}$ is simulated from the reduced model with parameters $\Phi_{\text{ind}, 0}^{*}$. For each $b = 1,...,B$, refit the bootstrap sample $(y^{(b)}, X)$ to the reduced model (the procedure was presented in Section 4) to obtain bootstrap fitted parameters $\Phi_{\text{ind}, n}^{(b)}$. The Efron percentile bootstrap CI of a quantity of interest is then represented by its empirical quantiles based on the $B$ sets of bootstrap fitted parameters $\{\Phi_{\text{ind}, n}^{(b)}\}_{b=1,...,B}$.

APPENDIX B. SUPPLEMENTARY DETAILS IN APPENDIX A

The proof techniques are in general followed by the arguments of Fan and Li (2001), Khalili and Chen (2007) and Khalili (2010), and we here extend to the setting of the proposed group-fused regularization method for multi-type feature selection.

B.1. Assumptions on the Penalty Functions

Denote $Z_{1} = \{ k : \|c_{k}^{x}x_{0}\|_{2} = 0 \}$, $Z_{2} = \{ k : \|c_{k}^{x}b_{0}\|_{2} = 0 \}$, and $Z_{3} = \{ k : \|c_{k}^{x}v_{0}\| = 0 \}$. We define the following quantities that are helpful for presenting the asymptotic results:
where \( p'_n(\psi; \eta_n) \) and \( p''_n(\psi; \eta_n) \) are the first and second derivatives of the penalty functions \( p_n(\psi; \eta_n) \) w.r.t. \( \psi \), for \( l = 1, 2, 3 \).

We require the following conditions on the penalty functions \( p_n(\psi; \eta_n) \), \( l = 1, 2, 3 \):

**H1.** For all \( n, \lambda_{\delta kn} \), and \( k = 1, \ldots, K_l \), we have \( p_n(0; \lambda_{\delta kn}) = 0 \), and \( p_n(\psi; \lambda_{\delta kn}) \) is non-decreasing and twice differentiable in \( \psi \in (0, \infty) \) except in a finite set.

**H2.** \( b'_n = o_P(1) \) as \( n \rightarrow \infty \).

**H3.** For \( T_n = \{ \psi : 0 < \psi \leq n^{-1/2} \log n \} \), we have \( \lim_{n \rightarrow \infty} \inf_{\psi \in T_n} p'_n(\psi; \lambda_{\delta kn}) / \sqrt{n} = \infty \) for every \( k \in Z_l \).

**H4.** \( b_n = O_P(1) \) as \( n \rightarrow \infty \).

It is easy to check that LASSO and SCAD penalties with \( \lambda_{\delta kn} = O_P(n^{-1/2}) \) for any \( k \in Z_l (l = 1, 2, 3) \) both satisfy the aforementioned assumptions.

### B.2. Regularity Conditions

Let \( h(v; \Phi) \) be the density function of \( V = (Y, x) \) with parameters \( \Phi \in \Omega \). For a more concise presentation on the regularity conditions, we write \( \Phi = (\psi_1, \ldots, \psi_Q) \) where \( Q \) is the total number of parameters in the model. The regularity conditions are equivalent to Khalili (2010) given by the following:

**R1.** \( h(v; \Phi) \) has common support in \( v \) for all \( \Phi \in \Omega \), and \( h(v; \Phi) \) is identifiable in \( \Phi \) up to a permutation of mixture components.

**R2.** \( h(v; \Phi) \) admits third partial derivatives with respect to \( \Phi \) for each \( \Phi \in \Omega \) and for almost all \( v \).

**R3.** For all \( j_1, j_2 = 1, \ldots, Q \), the first two derivatives of \( h(v; \Phi) \) satisfy

\[
E \left[ \frac{\partial}{\partial \psi_{j_1}} \log h(v; \Phi) \right] = 0; \quad (B.1)
\]

\[
E \left[ \frac{\partial}{\partial \psi_{j_1}} \log h(v; \Phi) \frac{\partial}{\partial \psi_{j_2}} \log h(v; \Phi) \right] = -E \left[ \frac{\partial^2}{\partial \psi_{j_1} \partial \psi_{j_2}} \log h(v; \Phi) \right]. \quad (B.2)
\]

**R4.** The Fisher information matrix is finite and positive definite at \( \Phi = \Phi_0 \):

\[
I(\Phi) = E \left[ \left( \frac{\partial}{\partial \Phi} \log h(v; \Phi) \right) \left( \frac{\partial}{\partial \Phi} \log h(v; \Phi) \right)^T \right]. \quad (B.3)
\]

**R5.** There exists a function \( M(v) \) such that

\[
\left| \frac{\partial}{\partial \psi_{j_1}} \log h(v; \Phi) \right| \leq M(v), \quad \left| \frac{\partial^2}{\partial \psi_{j_1} \partial \psi_{j_2}} \log h(v; \Phi) \right| \leq M(v), \quad \left| \frac{\partial^3}{\partial \psi_{j_1} \partial \psi_{j_2} \partial \psi_{j_3}} \log h(v; \Phi) \right| \leq M(v). \quad (B.4)
\]

### B.3. Proof of Theorem A.1

Let \( r_n = n^{-1/2}(1 + b_{1n} + b_{2n} + b_{3n}) \). It suffices to show that for any \( \epsilon > 0 \), there exists a large constant \( M_\epsilon \) such that

\[
\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{|u| = M_\epsilon} \mathcal{F}_n(\Phi_0 + r_n u) < \mathcal{F}_n(\Phi_0) \right\} \geq 1 - \epsilon, \quad (B.5)
\]
where \( u = (u_x, u_\beta, u_\phi, u_\theta, u_r) \) represents a vector of (the change of) all parameters. This corresponds to the existence of local maximizer \( \Phi_n \) with \( \| \Phi_n - \Phi_0 \|_2 = O_P(n^{-1/2}(1 + b_{1n} + b_{2n} + b_{3n})) \).

Denote \( D_n(u) := F_n(\Phi_0 + r_n u) - F_n(\Phi_0) \), which satisfies the following inequality

\[
D_n(u) \leq \left[ \mathcal{L}_n(\Phi_0 + r_n u) - \mathcal{L}_n(\Phi_0) \right] \\
- \sum_{k=1}^{K_1} \left[ p_{1n}(\| e_{1k}^T (x_0 + r_n u_x) \|_2; \lambda_{1kn}) - p_{1n}(\| e_{1k}^T x_0 \|_2; \lambda_{1kn}) \right] \\
- \sum_{k=1}^{K_2} \left[ p_{2n}(\| e_{2k}^T (\beta_0 + r_n u_\beta) \|_2; \lambda_{2kn}) - p_{2n}(\| e_{2k}^T \beta_0 \|_2; \lambda_{2kn}) \right] \\
- \sum_{k=1}^{K_3} \left[ p_{3n}(\| e_{3k}^T (v_0 + r_n u_v) \|_2; \lambda_{3kn}) - p_{3n}(\| e_{3k}^T v_0 \|_2; \lambda_{3kn}) \right] \\
:= D_{1n}(u) - D_{2n}(u) - D_{3n}(u) - D_{4n}(u),
\]

where \( D_{1n}(u), D_{2n}(u), D_{3n}(u), D_{4n}(u) \) are the four corresponding terms expressed in \( D_n(u) \) above, and recall that \( J_i \) is defined in Appendix A as the number of linearly independent vectors in a reduced designed matrix. Taylor’s expansion and triangular inequality yield

\[
D_{1n}(u) = n^{-1/2}(1 + b_{1n} + b_{2n} + b_{3n})\mathcal{L}_n'(\Phi_0)^T u - \frac{1}{2} \left( 1 + b_{1n} + b_{2n} + b_{3n} \right)^2 u^T \mathcal{I}(\Phi_0) u \left( 1 + O_P(1) \right),
\]

and

\[
|D_{2n}(u)| = \left| \sum_{k=1}^{K_1} \left[ p_{1n}(\| e_{1k}^T x_0 \|_2; \lambda_{1kn}) \right] - \frac{1}{2} \left( 1 + b_{1n} + b_{2n} + b_{3n} \right)^2 \max_{k=1}^{K_1} \| e_{1k}^T u_x \|_2 \right|
\]

\[
+ \sum_{k=1}^{K_2} \left[ p_{2n}(\| e_{2k}^T x_0 \|_2; \lambda_{2kn}) \right] - \frac{1}{2} \left( 1 + b_{1n} + b_{2n} + b_{3n} \right)^2 \max_{k=1}^{K_2} \| e_{2k}^T u \|_2 \right|
\]

\[
\leq (K_1 - J_1) b_{1n}(1 + b_{1n} + b_{2n} + b_{3n}) \max_{k=1}^{K_1} \| e_{1k}^T u_x \|_2
\]

\[
+ (K_1 - J_1) \frac{1}{2} b_{1n}(1 + b_{1n} + b_{2n} + b_{3n})^2 \max_{k=1}^{K_1} \| e_{1k}^T u_x \|_2
\]

\[
\leq (K_1 - J_1) \left[ b_{1n}(1 + b_{1n} + b_{2n} + b_{3n}) \| u_x \|_2 + \frac{1}{2} b_{1n}^2 (1 + b_{1n} + b_{2n} + b_{3n})^2 \| u_x \|_2 \right],
\]

where \( C_{\text{max}} \) is a fixed constant determined by the design matrix \( C_1 \). By similar arguments,

\[
|D_{3n}(u)| \leq (K_2 - J_2) \left[ b_{2n}(1 + b_{1n} + b_{2n} + b_{3n}) \| u_{\beta} \|_2 + \frac{1}{2} b_{2n}^2 (1 + b_{1n} + b_{2n} + b_{3n})^2 \| u_{\beta} \|_2 \right],
\]

and

\[
|D_{4n}(u)| \leq (K_3 - J_3) \left[ b_{3n}(1 + b_{1n} + b_{2n} + b_{3n}) \| u_{\phi} \|_2 + \frac{1}{2} b_{3n}^2 (1 + b_{1n} + b_{2n} + b_{3n})^2 \| u_{\phi} \|_2 \right].
\]

Performing an order analysis while keeping in mind that penalty function conditions H1–H2 and regularity conditions R1–R5 hold, we know that \( \mathcal{L}_n'(\Phi_0) = O_P(n^{1/2}) \) and that

\[
-\frac{1}{2} \left( 1 + b_{1n} + b_{2n} + b_{3n} \right)^2 u^T \mathcal{I}(\Phi_0) u \left( 1 + o_P(1) \right) < 0
\]

is the sole leading term in \( D_n(u) \) after choosing large enough \( M_r \). This shows that Equation (B.5) holds and hence the result follows.
B.4. Proof of Theorem A.2

We first start with the following lemma.

Lemma B.1. Under the conditions of Theorem A.2, for any \( \Phi^* \) satisfying \( \| \Phi^* - \Phi^0 \|_2 = O(n^{-1/2}) \), we have
\[
P\{ F_n^*((\Phi^*_{\text{red}}, \Phi^*_{\text{ind}})) < F_n^*((0, \Phi^*_{\text{ind}})) \} \rightarrow 1 \quad \text{as} \quad n \rightarrow \infty.
\]

Proof. We first notice that
\[
F_n^*((\Phi^*_{\text{red}}, \Phi^*_{\text{ind}})) - F_n^*((0, \Phi^*_{\text{ind}})) = [L_n^*((\Phi^*_{\text{red}}, \Phi^*_{\text{ind}})) - L_n^*((0, \Phi^*_{\text{ind}}))] - [P_n^*((\Phi^*_{\text{red}}, \Phi^*_{\text{ind}})) - P_n^*((0, \Phi^*_{\text{ind}}))].
\]

Following the proof techniques by theorem 2 of Khalili and Chen (2007) and lemma 2 of Khalili (2010), by the mean value theorem and using condition R5, we have
\[
L_n^*((\Phi^*_{\text{red}}, \Phi^*_{\text{ind}})) - L_n^*((0, \Phi^*_{\text{ind}})) = \left[ \frac{\partial L_n^*((\xi_n, \Phi^*_{\text{ind}}))}{\partial \Phi^*_{\text{red}}} \right]^T \Phi^*_{\text{red}},
\]
for some \( \xi_n \) satisfying \( \| \xi_n \|_2 \leq \| \Phi^*_{\text{red}} \|_2 = O(n^{-1/2}) \) and
\[
\left\| \frac{\partial L_n^*((\xi_n, \Phi^*_{\text{ind}}))}{\partial \Phi^*_{\text{red}}} - \frac{\partial L_n^*((0, \Phi^*_{\text{ind}}))}{\partial \Phi^*_{\text{red}}} \right\|_2 = O_P(n^{1/2}).
\]

Combining the above two equations, we have
\[
L_n^*((\Phi^*_{\text{red}}, \Phi^*_{\text{ind}})) - L_n^*((0, \Phi^*_{\text{ind}})) = O_P(n^{1/2}) \times \| \Phi^*_{\text{red}} \|_2 = O_P(1).
\]

On the other hand, for the penalty terms we have
\[
P_n^*((\Phi^*_{\text{red}}, \Phi^*_{\text{ind}})) - P_n^*((0, \Phi^*_{\text{ind}}))
= \sum_{k=1}^{K_t} [p_{1n}(|\hat{c}_{1k}^T (\vec{z}^T_{\text{red}} \vec{z}^T_{\text{ind}})|^2; \hat{\lambda}_{1k}) - p_{1n}(|\hat{c}_{1n}^T (0^T \vec{z}^T_{\text{ind}})|^2; \hat{\lambda}_{1n})]
+ \sum_{k=1}^{K_t} [p_{2n}(|\hat{c}_{2k}^T (\vec{b}^T_{\text{red}} \vec{b}^T_{\text{ind}})|^2; \hat{\lambda}_{2k}) - p_{2n}(|\hat{c}_{2n}^T (0^T \vec{b}^T_{\text{ind}})|^2; \hat{\lambda}_{2n})]
+ \sum_{k=1}^{K_t} [p_{3n}(|\hat{c}_{3k}^T (\vec{v}^T_{\text{red}} \vec{v}^T_{\text{ind}})|^2; \hat{\lambda}_{3k}) - p_{3n}(|\hat{c}_{3n}^T (0^T \vec{v}^T_{\text{ind}})|^2; \hat{\lambda}_{3n})].
\]

We now perform an order analysis on the first term of the right-hand side of the above equation as follows:
\[
P_{1n}(|\hat{c}_{1k}^T (\vec{z}^T_{\text{red}} \vec{z}^T_{\text{ind}})|^2; \hat{\lambda}_{1k}) - P_{1n}(|\hat{c}_{1n}^T (0^T \vec{z}^T_{\text{ind}})|^2; \hat{\lambda}_{1n})
= p'_{1n}(|\vec{z}^T_{\text{ind}} \vec{z}^T_{\text{ind}} + \xi_n|_2^2; \hat{\lambda}_{1n}) \times \frac{\vec{c}_{1n}^T (\vec{z}^T_{\text{ind}} \vec{z}^T_{\text{ind}} + \xi_n|_2^2; \hat{\lambda}_{1n})}{\vec{c}_{1n}^T (\vec{z}^T_{\text{ind}} \vec{z}^T_{\text{ind}} + \xi_n|_2^2; \hat{\lambda}_{1n})}
= p'_{1n}(|\vec{z}^T_{\text{ind}} \vec{z}^T_{\text{ind}} + \xi_n|_2^2; \hat{\lambda}_{1n}) \times n^{1/2} \times O_P(1) \times O_P(n^{-1/2})
= p'_{1n}(|\psi_{n,k}; \hat{\lambda}_{1n}|) \times O_P(1),
\]
where we have decomposed \( \hat{c}_{1k} = (\vec{c}_{\text{red},1k}, \vec{c}_{\text{ind},1k}) \) such that \( \hat{c}_{1k}^T (\vec{z}^T_{\text{red}} \vec{z}^T_{\text{ind}}) = \hat{c}_{1k}^T \vec{z}^T_{\text{red}} + \hat{c}_{1k}^T \vec{z}^T_{\text{ind}}, \) for some \( \| \xi_n \|_2 \leq \| \vec{c}_{\text{red},1k}^T \vec{c}_{\text{ind},1k} \|_2 = O_P(n^{-1/2}). \) Note from Equation (B.19) that \( \psi_{n,k} = O_P(n^{-1/2}) \) for \( k = 1, ..., J_1 \) and \( \psi_{n,k} = O_P(1) \) for \( k = J_1 + 1, ..., K_1. \) As a result, under condition H3, Equation (B.19) has an order greater than \( O_P(1) \) for \( k = 1, ..., J_1, \) whereas, under condition H4, Equation (B.19) has an order equal to \( O_P(1) \) for \( k = J_1 + 1, ..., K_1. \) Applying similar arguments as above for the second and third terms on the right-hand side of Equation (B.18), by comparing the orders it is clear that
\[ -\sum_{j=1}^{J_1} p_{1n}(\|\mathbf{c}_1^T(\mathbf{z}_{\text{red}}; \mathbf{z}_\text{ind})\|_2; \lambda_{1kn}) - p_{1n}(\|\mathbf{c}_1^T(0^T; \mathbf{z}_\text{ind})\|_2; \lambda_{1kn}) \\
- \sum_{j=1}^{J_2} p_{2n}(\|\mathbf{c}_2^T(\mathbf{p}_{\text{red}}^T; \mathbf{p}_{\text{ind}})\|_2; \lambda_{2kn}) - p_{2n}(\|\mathbf{c}_2^T(0^T; \mathbf{p}_{\text{ind}})\|_2; \lambda_{2kn}) \\
- \sum_{j=1}^{J_3} p_{3n}(\|\mathbf{c}_3^T(\mathbf{v}_{\text{red}}^T; \mathbf{v}_{\text{ind}}^T)\|_2; \lambda_{3kn}) - p_{3n}(\|\mathbf{c}_3^T(0^T; \mathbf{v}_{\text{ind}}^T)\|_2; \lambda_{3kn}) < 0 \] (B.20)

is the dominant term of \( F_n^*(\{\Phi_{\text{red}}, \Phi_{\text{ind}}^*\}) - F_n^*(\{0, \Phi_{\text{ind}}^*\}) \) in Equation (B.14). Note that this term must be negative because, for example, \( p_{1n}(\|\mathbf{c}_1^T(0^T; \mathbf{z}_\text{ind})\|_2; \lambda_{1kn}) = 0 \) for \( k = 1, ..., J_1 \) by construction. Hence, the result follows.

For part (a) of Theorem 3, it suffices to show that \( F_n^*(\{\Phi_{\text{red}}, \Phi_{\text{ind}}^*\}) - F_n^*(\{0, \Phi_{\text{ind}}^*\}) < 0 \) in probability for any \( \Phi^* := (\Phi_{\text{red}}^*, \Phi_{\text{ind}}^*) \) such that \( ||\Phi^* - \Phi_0||_2 = O_P(n^{-1/2}) \). Following the arguments of Theorem 2 of Khalili and Chen (2007) and Theorem 3 of Khalili (2010), note that we have \( F_n^*(\{\Phi_{\text{red}}, \Phi_{\text{ind}}^*\}) - F_n^*(\{0, \Phi_{\text{ind}}^*\}) \leq F_n^*(\{\Phi_{\text{red}}, \Phi_{\text{ind}}^*\}) - F_n^*(\{0, \Phi_{\text{ind}}^*\}) < 0 \) in probability using Lemma 1.

For part (b) of Theorem A.2, the arguments are completely identical to theorem 2 of Khalili and Chen (2007) and theorem 3 of Khalili (2010) after establishing the results of Lemma B.1, so the proof is omitted.

**Remark B.1.** Note that the second derivative \( \mathcal{P}''(\Phi, \mathcal{D}, \alpha) \) in part (b) of Theorem A.2 is asymptotically negligible due to condition H2. On the other hand, we need \( b_{kn} = o_P(1) \), which is stronger than condition H4, in order for the bias term \( \mathcal{P}''(\Phi, \mathcal{D}, \alpha) \) to be negligible. For the SCAD penalty, this is not a problem by choosing \( \lambda_{kn} = O_P(n^{-1/2}) \). For the LASSO penalty, we may use an adaptive approach by choosing \( \lambda_{kn} = o_P(n^{-1/2})/\|\mathbf{c}_1^T \mathbf{z}_0 \|_2 \) (analogously for \( \lambda_{2kn} \) and \( \lambda_{3kn} \)), where \( \mathbf{z}_0 \) is the MLE without penalty. Note that \( \|\mathbf{c}_1^T \mathbf{z}_0 \|_2 = O_P(n^{-1/2}) \) for \( k \in \mathcal{Z}_1 \) and \( \|\mathbf{c}_1^T \mathbf{z}_0 \|_2 = O_P(1) \) for \( k \notin \mathcal{Z}_1 \). This allows that the conditions of \( b_{kn} = o_P(1) \) and H1–H4 still hold simultaneously to preserve both consistency and asymptotic normality.

**APPENDIX C. SUPPLEMENTARY INFORMATION IN SECTION 4**

**C.1. Derivation of the Complete Data Log-Likelihood**

The complete data likelihood is given by

\[ \exp \{ L_n^{\text{com}}(\Phi, \mathcal{D}^{\text{com}}, X) \} = \prod_{i=1}^{n} P(Z_i = z_i, Y_i = y_i, M_i = m_i, Y'_i = y'_i | x_i, \Phi) \\
= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{g} \left( \prod_{i=1}^{m} \left( \frac{f(y_i; \exp \{ \beta_j x_i \}, \phi_j \}}{\exp \{ \beta_j x_i \}, \phi_j \}} \right)^{z_{ij}} \right) \right\} \\
\times \left( \prod_{i=1}^{n} \left( \prod_{j=1}^{g} \left( \frac{h(y_i; \theta, \exp \{ v_j x_i \}}{1 - H(\theta, \exp \{ v_j x_i \})} \right)^{z_{ij}} \right) \right) \right\} \\
\times \left( \prod_{i=1}^{n} \left( \prod_{j=1}^{g} \left( \frac{f(y'_i; \exp \{ \beta_j x_i \}, \phi_j \}}{1 - F(\theta, \exp \{ \beta_j x_i \}, \phi_j \}} \right)^{z_{ij}} \right) \right) \}

(C.1)
C.2. Supplementary Details of the GEM Algorithm

In this section, we provide the missing details of the GEM algorithm presented in the article.

C.2.1. E-step

The updated quantities for the E-step are given by the following formulas:

\[
\zeta_{ij}^{(l)} = P(Z_{ij} = 1 | y, x, \Phi^{(l-1)}) = \pi_i(x_i; \theta^{(l-1)}) \frac{f(y_i; \exp \{ \beta_j^{(l-1)} x_i \}, \phi_j^{(l-1)})}{h_y(y_i; x_i, \Phi^{(l-1)})} \left[ \frac{1}{1 - H(\tau; \theta^{(l-1)}, \exp \{ v^{(l-1)} x_i \})} I\{y_i \leq \tau\} + \frac{h(y_i; \theta^{(l-1)}, \exp \{ v^{(l-1)} x_i \})}{1 - H(\tau; \theta^{(l-1)}, \exp \{ v^{(l-1)} x_i \})} I\{y_i > \tau\} \right],
\]

(C.2)

\[
m_{ij}^{(l)} = E[M_j | y, x, \Phi^{(l-1)}, Z_{ij} = 1] = \frac{1 - F(\tau; \exp \{ \beta_j^{(l-1)} x_i \}, \phi_j^{(l-1)})}{F(\tau; \exp \{ \beta_j^{(l-1)} x_i \}, \phi_j^{(l-1)})},
\]

(C.3)

\[
\tilde{\gamma}_{ij}^{(l)} = E[Y_{im} | y, x, \Phi^{(l-1)}, Z_{ij} = 1] = \int_{\tau}^{\infty} f(y_i; \exp \{ \beta_j^{(l-1)} x_i \}, \phi_j^{(l-1)}) dy \frac{1 - F(\tau; \exp \{ \beta_j^{(l-1)} x_i \}, \phi_j^{(l-1)})}{1 - F(\tau; \exp \{ \beta_j^{(l-1)} x_i \}, \phi_j^{(l-1)})}
\]

\[
= \frac{\phi_j^{(l-1)} \exp \{ \beta_j^{(l-1)} x_i \} \Gamma(1/\phi_j^{(l-1)})}{\Gamma(1/\phi_j^{(l-1)})(1 - F(\tau; \exp \{ \beta_j^{(l-1)} x_i \}, \phi_j^{(l-1)}))} \times \int_{\tau}^{\infty} (\phi_j^{(l-1)} \exp \{ \beta_j^{(l-1)} x_i \})(1 + \phi_j^{(l-1)})^{-\phi_j^{(l-1)}}
\]

\[
\times y^{1/\phi_j^{(l-1)} - 1} \exp \{-y/(\phi_j^{(l-1)} \exp \{ \beta_j^{(l-1)} x_i \})(1 + \phi_j^{(l-1)})\} dy
\]

\[
= \frac{1 - F(\tau; \exp \{ \beta_j^{(l-1)} x_i \}, \phi_j^{(l-1)})(1 + \phi_j^{(l-1)})}{1 - F(\tau; \exp \{ \beta_j^{(l-1)} x_i \}, \phi_j^{(l-1)})} \exp \{ \beta_j^{(l-1)} x_i \},
\]

(C.4)

where \( \tilde{\gamma}_{ij}^{(l-1)} = \phi_j^{(l-1)}/(1 + \phi_j^{(l-1)}) \). Moreover,

\[
\log \tilde{y}_{ij}^{(l)} = E[\log Y_{im} | y, x, \Phi^{(l-1)}, Z_{ij} = 1] = \int_{\tau}^{\infty} \log(y_i) \frac{f(y_i; \exp \{ \beta_j^{(l-1)} x_i \}, \phi_j^{(l-1)})}{1 - F(\tau; \exp \{ \beta_j^{(l-1)} x_i \}, \phi_j^{(l-1)})} dy.
\]

(C.5)

Note that \( \log \tilde{y}_{ij}^{(l)} \) can be evaluated numerically using the \texttt{integrate()} function in R. Alternatively, \( \log \tilde{y}_{ij}^{(l)} \) can be analytically expressed using an incomplete digamma function (details are omitted).

C.2.2. M-step

The equations with regards to the derivatives under the IRLS approach in Equations (4.19), (4.20), and (4.22) are showcased as follows:

Taking the logarithm in the above equation would lead to the complete data log-likelihood in Equation (4.4). The complete data log-likelihood (Eq. [4.4]) is analytically tractable because the geometric distribution (the third line of the second equality in Eq. [C.1]) cancels exactly with the denominators in the first and fourth lines of the second equality in Equation (C.1). Such a nice cancellation cannot be achieved by any other discrete distributions for \( M_i \) such as Poisson.
\[
\frac{\partial S^{(t)}(\mathbf{x})}{\partial x_j} = \sum_{i=1}^{n} \left[ \frac{\exp \{ x_j^T x_i \}}{\sum_{j'=1}^{p} \exp \{ x_{j'}^T x_i \}} x_i - \sum_{k=1}^{K} \rho'_{ik} \left( \| c_{ik}^T x^{(t-1)} \|_2, z_{2k} \right) \frac{c_{ik}^T x_i}{\| c_{ik}^T x^{(t-1)} \|_2} \right] x_i,
\]
\[
\frac{\partial^2 S^{(t)}(\mathbf{x})}{\partial x_j \partial x_j'} = \sum_{i=1}^{n} \left[ \frac{\exp \{ x_j^T x_i \}}{\sum_{j'=1}^{p} \exp \{ x_{j'}^T x_i \}} x_i x_i' - \sum_{k=1}^{K} \rho'_{ik} \left( \| c_{ik}^T x^{(t-1)} \|_2, z_{2k} \right) \frac{c_{ik}^T x_i x_i'}{\| c_{ik}^T x^{(t-1)} \|_2} \right] \frac{c_{ik}^T x_i x_i'}{\| c_{ik}^T x^{(t-1)} \|_2},
\]
\[
\frac{\partial T^{(t)}(\beta, \phi_j^{(t-1)})}{\partial \beta_j} = \frac{1}{\phi_j^{(t-1)}} \sum_{i=1}^{n} \frac{z_{ij}^{(t)}}{y_i} \left[ \left( y_i \exp \{ -\beta_j^T x_i \} - 1 \right) + m_j^{(t)} \left( y_j^{(t)} \exp \{ -\beta_j^T x_i \} - 1 \right) \right] x_i - \sum_{k=1}^{K} \rho'_{2k} \left( \| c_{2k}^T \beta^{(t-1)} \|_2, z_{2k} \right) \frac{c_{2k}^T x_i}{\| c_{2k}^T \beta^{(t-1)} \|_2} \beta_j,
\]
\[
\frac{\partial^2 T^{(t)}(\beta, \phi_j^{(t-1)})}{\partial \beta_j \partial \beta_j'} = -\frac{1}{\phi_j^{(t-1)}} \sum_{i=1}^{n} \frac{z_{ij}^{(t)}}{y_i} \left[ y_i + m_j^{(t)} y_j^{(t)} \right] \exp \{ -\beta_j^T x_i \} x_i x_i' - \sum_{k=1}^{K} \rho'_{2k} \left( \| c_{2k}^T \beta^{(t-1)} \|_2, z_{2k} \right) \frac{c_{2k}^T x_i x_i'}{\| c_{2k}^T \beta^{(t-1)} \|_2},
\]
\[
\frac{\partial V^{(t)}(\theta^{(t-1)}, \nu)}{\partial \nu} = \sum_{i=1}^{n} \left[ 1 - \log \left( \frac{y_i - \tau + \theta^{(t-1)}}{\theta^{(t-1)}} \right) \right] \exp \{ \nu^T x_i \} 1 \{ y_i > \tau \} x_i - \sum_{k=1}^{K} \rho'_{3k} \left( \| c_{3k}^T \nu^{(t-1)} \|_2, z_{3k} \right) \frac{c_{3k}^T x_i}{\| c_{3k}^T \nu^{(t-1)} \|_2} \nu,
\]
\[
\frac{\partial^2 V^{(t)}(\theta^{(t-1)}, \nu)}{\partial \nu \partial \nu'} = -\sum_{i=1}^{n} \log \left( \frac{y_i - \tau + \theta^{(t-1)}}{\theta^{(t-1)}} \right) \exp \{ \nu^T x_i \} 1 \{ y_i > \tau \} x_i x_i' - \sum_{k=1}^{K} \rho'_{3k} \left( \| c_{3k}^T \nu^{(t-1)} \|_2, z_{3k} \right) \frac{c_{3k}^T x_i x_i'}{\| c_{3k}^T \nu^{(t-1)} \|_2}.
\]

C.3. Outline of the GEM Algorithm
Algorithm 1 presents a summary outline to implement the GEM algorithm presented in Section 4.2.

Algorithm 1 GEM algorithm

Require: Estimate the fitted model parameters \( \Phi = (\mathbf{x}, \beta, \phi, \theta, \nu) \).

Input: Initialized model parameters \( \Phi^{(0)} = (\mathbf{x}^{(0)}, \beta^{(0)}, \phi^{(0)}, \theta^{(0)}, \nu^{(0)}) \), observed data \( (y, X) \), hyperparameters \( (\lambda_1, \lambda_2, \lambda_3) \), penalty perturbation \( \epsilon > 0 \).

Output: The fitted model parameters \( \Phi \) and the fitted latent variables \( \tilde{z}_{ij}, \tilde{m}_{ij}, \tilde{y}_{ij}, \text{ and } \tilde{\log}{y}_{ij} \).

1: Set \( l \leftarrow 0 \) and define \( F^\text{com}_{n, \epsilon}(\Phi^{(1)}) \leftarrow -\infty \) \( \left\{ \text{I/E-step} \right\} \)
2: while \( F^\text{com}_{n, \epsilon}(\Phi^{(l)}) - F^\text{com}_{n, \epsilon}(\Phi^{(l-1)}) > 10^{-2} \) and \( l \leq 200 \) do
3: \( l \leftarrow l + 1 \)
4: Compute \( z_{ij}^{(l)}, m_{ij}^{(l)}, \tilde{y}_{ij}^{(l)} \) and \( \log{y}_{ij}^{(l)} \) using Equations (C.2) to (C.5). \( \left\{ \text{I/M-step} \right\} \)
5: for \( 1 \leq j \leq g \) do
6: \( \left\{ \text{I/M-step} \right\} \)
7: end for
8: end while

The automated adjustment algorithm is outlined in Algorithm 2, where in practice we find that setting the hyperparameters by Eqs. [C.6] and [C.7].

Automated adjustment algorithm

Require: Reduce complexity of fitted model to achieve variable selection.

Output: The adjusted fitted model parameters \( \hat{\Phi} \).

1: \( \text{run}_s \leftarrow \text{TRUE} \)
2: \( \text{while} \ \text{run}_s = \text{TRUE} \) do
3: \( \hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}}' \)
4: \( \text{for} \ 2 \leq d_1 < d_2 \leq D \) do
5: \( \quad \text{if} \ |\hat{\mathbf{z}}_{d_1} - \hat{\mathbf{z}}'_{d_1}|_2 < \delta \) then \( \hat{\mathbf{z}}_{d_1} \leftarrow \hat{\mathbf{z}}'_{d_1} \) \( \quad \) //Merge levels when the fitted parameters are close
6: \( \text{end if} \)
7: \( \text{end for} \)
8: \( \text{for} \ 2 \leq d \leq D \) do
9: \( \quad \text{if} \ |\hat{\mathbf{z}}_{d}'|_2 < \delta \) then \( \hat{\mathbf{z}}_d \leftarrow 0 \) \( \quad \) //Force parameters to become zero if they are close to zero
10: \( \text{end if} \)
11: \( \text{end for} \)
12: \( \text{if} \ S(\hat{\mathbf{x}}') - S(\hat{\mathbf{x}}) < -\xi \) then \( \text{run}_s \leftarrow \text{FALSE} \) \( \quad \) //Stop when parameter adjustment decreases log-likelihood by a lot
13: \( \text{end if} \)
14: \( \hat{\delta} \leftarrow \hat{\delta}(1 + \eta) \)
15: end while
16: \( \text{return} \ \hat{\mathbf{x}} \) \( \quad \) //Return the adjusted parameters
17: Similarly, repeat the above lines to update the adjusted values for \( \hat{\beta} \) and \( \hat{v} \), with \( S \) replaced by \( T \) and \( V \), respectively, and \( \mathbf{x} \) replaced by \( \hat{\beta} \) and \( \hat{v} \), respectively.