Mixture composite regression models with multi-type feature selection

Tsz Chai (Samson) Fung

RiskLab, Department of Mathematics
ETH Zürich

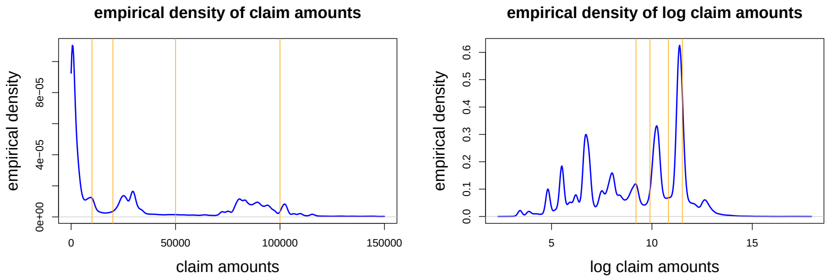
Joint work with George Tzougas and Mario V. Wüthrich

Insurance Data Science Conference, 16 Jun 2021

Outline

Motivating dataset – Greek automobile dataset

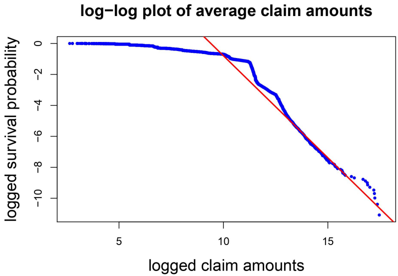
- Motor third-party liability (MTPL) insurance policies
- 64,923 non-zero property claim severities for underwriting years 2013 to 2017.
- Empirical density of claim amounts:

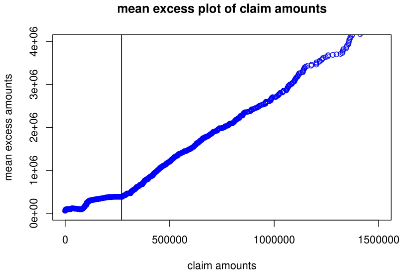


  - Multimodality of distribution
  - Not meaningful to capture all distributional nodes for small claim sizes

Motivating dataset – Greek automobile dataset

- Log-log plot and mean excess plot of claim amounts:





  - Heavy-tailedness of distribution (tail index $\eta \approx 1.3$)
  - Mismatch between body and tail behavior

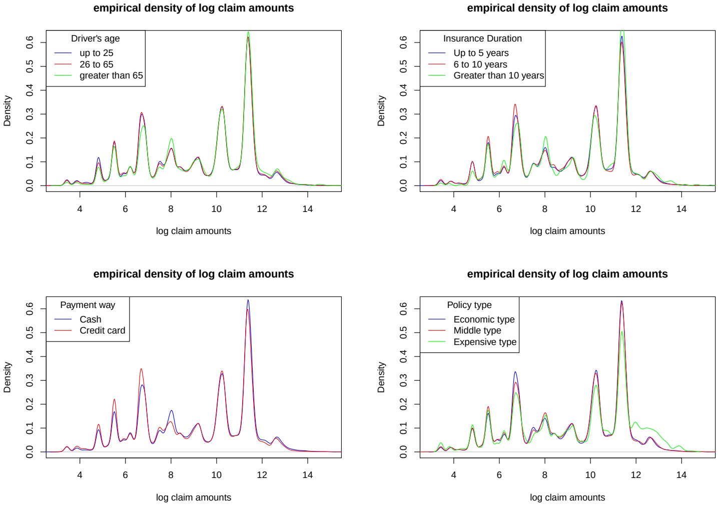
Motivating dataset – Greek automobile dataset

- Explanatory variables under consideration:

Name	**Short Description**	**Categories**	**Type**	**Categories Description**
DriverAge	Driver's age	18-74	Continuous	From 18 to 74+ years old
VechicleBrand	Automobile Brand	B1 - B31	Unordered	31 different brands
CC	Car cubism	0 - 18	Ordered	19 different categories
PolicyType	Policy Type	C1 C2 C3	Ordered	Economic type Middle type Expensive type
FHP	Automobile horsepower	1 - 13	Ordered	13 categories of horsepower
InsuranceDuration	Insurance duration	ID1 ID2 ID3	Ordered	Up to 5 years From 6 to 10 years Greater than 10 years
PaymentWay	Payment way	C1 C2	Unordered	Cash Credit card
Region	City population	1-2; 4-14	Unordered	13 Administrative Regions
VehicleAge	Vehicle age	C1 C2 C3	Ordered	New car Middle Old
SumInsured.1	Sum Insured	C1 C2 C3	Ordered	Up to 5,000 Euros 5,001 to 10,000 Euros Greater than 10,000 Euros

Motivating dataset – Greek automobile dataset

- How do the covariates impact the claim severity distribution?



Research problem

To address the challenges of modelling the motivating dataset, we need a claim severity model with the following characteristics:

- Sufficient flexibility to model distributional multimodality
- Heavy tail in nature and robustness for estimating the tail-heaviness
- Capture covariates influence on various parts of the distribution:
  - Probabilities assigning observations into various clusters
  - Systematic effects in distributions conditioned on each clusters
  - Tail-heaviness of the distribution
- Enable variable selection:
  - Not all variables are important
  - Different variables impact different parts of the distribution
  - Multi-type variable settings – Continuous, ordered and unordered categorical

Mixture-Gamma Lomax composite regression model

- Probability Density:

$$h_Y(y;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\phi},\theta,\boldsymbol{\nu},\boldsymbol{x}) = \sum_{j=1}^{g}\pi_j(\boldsymbol{x};\boldsymbol{\alpha})\frac{f(y;\exp\{\boldsymbol{\beta}_j^T\boldsymbol{x}\},\phi_j)}{F(\tau;\exp\{\boldsymbol{\beta}_j^T\boldsymbol{x}\},\phi_j)}1\{y\le\tau\} + \pi_{g+1}(\boldsymbol{x};\boldsymbol{\alpha})\frac{h(y;\theta,\exp\{\boldsymbol{\nu}^T\boldsymbol{x}\})}{1-H(\tau;\theta,\exp\{\boldsymbol{\nu}^T\boldsymbol{x}\})}1\{y>\tau\}$$

  - $Y\in\mathbb{R}^+$: claim severity random variable
  - $\boldsymbol{x}\in\mathbb{R}^D$: vector of covariates
  - $\pi_j(\boldsymbol{x};\boldsymbol{\alpha})$: Clustering probabilities
  - $f(y;\exp\{\boldsymbol{\beta}_j^T\boldsymbol{x}\},\phi_j)$: Density capturing moderate claim amounts
  - $h(y;\theta,\exp\{\boldsymbol{\nu}^T\boldsymbol{x}\})$: Density capturing extreme claim amounts
  - τ: Splicing threshold of composite model

Model specifications

Mixture-Gamma Lomax composite regression model

$$h_Y(y;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\phi},\theta,\boldsymbol{\nu},\boldsymbol{x}) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x};\boldsymbol{\alpha}) \frac{f(y;\exp\{\boldsymbol{\beta}_j^T\boldsymbol{x}\},\phi_j)}{F(\tau;\exp\{\boldsymbol{\beta}_j^T\boldsymbol{x}\},\phi_j)} 1\{y \le \tau\} + \pi_{g+1}(\boldsymbol{x};\boldsymbol{\alpha}) \frac{h(y;\theta,\exp\{\boldsymbol{\nu}^T\boldsymbol{x}\})}{1 - H(\tau;\theta,\exp\{\boldsymbol{\nu}^T\boldsymbol{x}\})} 1\{y > \tau\}$$

- Clustering probabilities: logit-linear function

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

  - Simple formulation to incorporate covariates effects on clustering probabilities
  - Denseness theory (Fung et al. (2019)): Flexible in capture a wide range of regression structures
  - Favorable for likelihood-based inference: $\log \pi_j(\boldsymbol{x};\boldsymbol{\alpha})$ is concave w.r.t. $\boldsymbol{\alpha}_j$.

Model specifications

Mixture-Gamma Lomax composite regression model

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

- Density for moderate claims: Gamma distribution

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

  - Light-tailed: Model the body of distribution
  - Distributional multimodality achieved under mixture of $g>1$ components.
  - Linear regression on the mean parameter for each mixture components

Model specifications

Mixture-Gamma Lomax composite regression model

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

- Density for extreme claims: Lomax distribution

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

  - Heavy-tailed: Model the (polynomial) tail of distribution
  - Tail index η governs tail-heaviness
  - Linear regression on tail index

Model specifications

Mixture-Gamma Lomax composite regression model

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

- Composite modelling framework with splicing threshold τ
  - Remove overlapping of density functions between body and tail parts
  - More robust estimation of tail index
  - The denominators $F(\tau;\exp\{\boldsymbol{\beta}_j^T\boldsymbol{x}\},\phi_j)$ and $1-H(\tau;\theta,\exp\{\boldsymbol{\nu}^T\boldsymbol{x}\})$ ensure that the density functions f_Y is proper.

Statistical inference with feature selection

- Suppose we observe n independent claims. Notations:
  - $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$: Claim size random vector
  - $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times D}$: Design matrix containing the covariates information
- For parameter estimation, we maximize the penalized log-likelihood

$$\mathcal{F}_n(\boldsymbol{\Phi}) = \mathcal{L}_n(\boldsymbol{\Phi}) - \mathcal{P}_n(\boldsymbol{\Phi})$$

- $\mathcal{L}_n(\boldsymbol{\Phi})$: Log-likelihood function

$$\mathcal{L}_n(\boldsymbol{\Phi}) := \mathcal{L}_n(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} \log h_Y(y_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi}, \theta, \boldsymbol{\nu}, \boldsymbol{x}_i)$$

- $\mathcal{P}_n(\boldsymbol{\Phi})$: Group-fused regularization term to penalize regression parameters for variable selection

$$\mathcal{P}_n(\boldsymbol{\Phi}) = P_{\boldsymbol{\lambda}_1,n}(\boldsymbol{\alpha}) + P_{\boldsymbol{\lambda}_2,n}(\boldsymbol{\beta}) + P_{\boldsymbol{\lambda}_3,n}(\boldsymbol{\nu})$$

$$P_{\boldsymbol{\lambda}_1,n}(\boldsymbol{\alpha}) = \sum_{k=1}^{K_1} p_{1n}\left(\|\boldsymbol{c}_{1k}^T\boldsymbol{\alpha}\|_2; \lambda_{1kn}\right); \quad P_{\boldsymbol{\lambda}_2,n}(\boldsymbol{\beta}) = \sum_{k=1}^{K_2} p_{2n}\left(\|\boldsymbol{c}_{2k}^T\boldsymbol{\beta}\|_2; \lambda_{2kn}\right); \quad P_{\boldsymbol{\lambda}_3,n}(\boldsymbol{\nu}) = \sum_{k=1}^{K_3} p_{3n}\left(|\boldsymbol{c}_{3k}^T\boldsymbol{\nu}|; \lambda_{3kn}\right)$$

Statistical inference with feature selection

Penalty function: $\mathcal{P}_n(\boldsymbol{\Phi}) = P_{\boldsymbol{\lambda}_1,n}(\boldsymbol{\alpha}) + P_{\boldsymbol{\lambda}_2,n}(\boldsymbol{\beta}) + P_{\boldsymbol{\lambda}_3,n}(\boldsymbol{\nu})$

$$P_{\boldsymbol{\lambda}_1,n}(\boldsymbol{\alpha}) = \sum_{k=1}^{K_1} p_{1n}\left(\|\boldsymbol{c}_{1k}^T\boldsymbol{\alpha}\|_2; \lambda_{1kn}\right); \quad P_{\boldsymbol{\lambda}_2,n}(\boldsymbol{\beta}) = \sum_{k=1}^{K_2} p_{2n}\left(\|\boldsymbol{c}_{2k}^T\boldsymbol{\beta}\|_2; \lambda_{2kn}\right); \quad P_{\boldsymbol{\lambda}_3,n}(\boldsymbol{\nu}) = \sum_{k=1}^{K_3} p_{3n}\left(|\boldsymbol{c}_{3k}^T\boldsymbol{\nu}|; \lambda_{3kn}\right)$$

- p_{1n}, p_{2n}, p_{3n}: concave non-decreasing penalty terms (e.g. LASSO, SCAD) – shrink unimportant regression parameters to zero
- $\{\boldsymbol{c}_{lk}\}_{l=1,2,3}$: predetermined penalty coefficients – flexible to deal with multi-type features
  - Shrink regression coefficients for continuous variables
  - Merge regression coefficients for ordinal/unordered categorical variables
- λ_{1kn}, λ_{2kn}, λ_{3kn}: penalty tuning parameters – Larger values mean more regression parameters are shrunk/merged.

Statistical properties and parameter estimations

Theoretical properties to justify the feature selection method

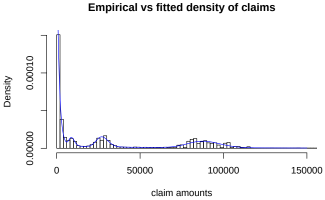
- The proposed method is consistent in terms of feature selection
  - Correctly merge and shrink regression coefficients as $n \to \infty$
- The reduced model parameters are asymptotically normal with zero mean and easily computable covariance matrix
  - Easy to construct Wald-type and Efron percentile bootstrap confidence intervals of model parameters

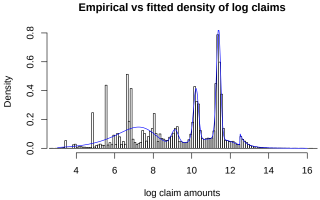
Model calibration techniques

- Novel augmented Generalized Expectation-Maximization (GEM) algorithm to estimate the parameters
- Adaptive standardization approach for more efficient tuning of hyperparameters $\boldsymbol{\lambda} := (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3)$
- Selection of hyperparameters using AIC, BIC or K-fold CV-based approaches.

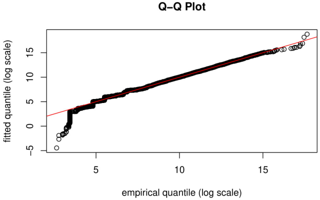
Let's return to the Greek dataset – Distributional fitting

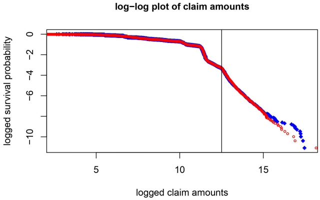
- We need at least $g = 5$ components for the body part to adequately capture all density nodes except for very small claims.
- Empirical vs fitted density of claim amounts:





- Q-Q plot and log-log plot of claim amounts:





Effects of the covariates

- Summary of regression model selection and performance across various model selection criteria

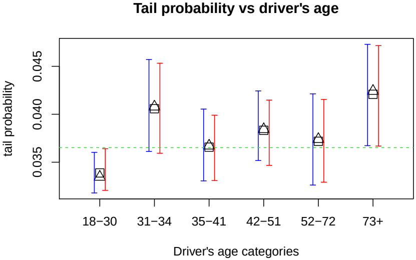
Model selection criteria	# parameters	log-likelihood	AIC	BIC
$\mathcal{L}_n$ with without regression	17	-719,309	1,438,652	1,438,807
$\mathcal{L}_n$ with without penalty	NA	NA	NA	NA
$\mathcal{L}_n$ + weak penalty only	1,524	-717,969	1,438,987	1,452,826
$\mathcal{L}_n$ + LASSO penalty w/ AIC before refit	809	-718,312	1,438,242	1,445,589
$\mathcal{L}_n$ + LASSO penalty w/ BIC 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/ AIC before refit	613	-718,324	1,437,873	1,443,439
$\mathcal{L}_n$ + SCAD penalty w/ BIC 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

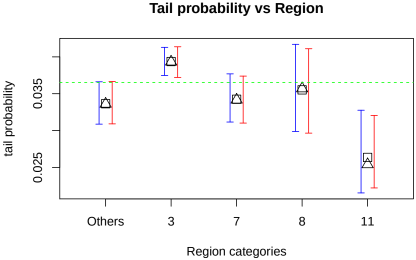
- The best model uses LASSO penalty with K-fold CV as variable selection criterion

Effects of the covariates

Brief summary

- Many variables well explain the clustering probabilities $\pi_j(\boldsymbol{x};\boldsymbol{\alpha})$: Driver's age, car cubism, policy type, payment way, region etc.
- Fewer variables well explain the body distributions: Driver's age, car cubism, payment way and region.
- No variables well explain the tail distribution.





Concluding remarks

- **Mixture composite regression model** to address several challenges when modelling claim severities such as multimodality and heavy-tailedness of claims
- **Group-fused regularization approach** for multi-type variable selection
- **Covariates** may influence the **mixture probabilities**, **body** and **tail** of the distribution, so model interpretability is preserved