Modeling the dependence in compound model using copula representation

Jae Youn Ahn[1] (Joint work with Woojoo Lee[2] and Rosy Oh[1])

[1] **Department of Statistics, Ewha Womans University, Korea**
[2] Department of Statistics, Inha University, Korea

The 3RD Pacific RIM Statistical Conference for Production Engineering
National Tsing Hua University,
Taiwan, 2018

Table of Contents

- Introduction I: Background Knowledge
- Introduction II: Literature Review
- Our Copula Model
- Data Analysis and Conclusion

- Classical Compound Model

$$S = \begin{cases} \sum_{i=1}^{N} Y_i, & N > 0 \\ 0, & N = 0 \end{cases} \quad \text{and} \quad M = \begin{cases} \frac{S}{N}, & N > 0 \\ \emptyset, & N = 0 \end{cases}$$

with $N, Y_1, Y_2, \cdots$ are independent and Y_i are identically distributed.

- Premium:

$$\mathbb{E}[S] = \mathbb{E}[N]\,\mathbb{E}[Y]$$

- Reserve:

$$\mathbb{V}ar(S) = \mathbb{E}[N]\,\mathbb{V}ar(Y) + (\mathbb{E}[X])^2\mathbb{V}ar(N)$$

- Question: What if N and $(Y_1, Y_2, \cdots,)$ are not independent?

Two classical approach modelling dependence

- Two(three) popular way to model the dependence.
- Copula
  - Conceptually simple
  - Hard to model the systematic dependence (Vine copula)
    - **Hard to choose right copula.**
  - No posterior distribution available
  - Easy to determine goodness of fit
- Random Effect
  - Can carefully organize the systematic dependence
  - Marginal and dependence effect can be mixed
  - Posterior distribution available: posterior risk classification
  - Hard to determine goodness of fit
- Graphical Model

Dependence in Frequency-Severity

- Copula: Czado et al. (2012), **Cossette et al. (2018)**
- Random effect: Boudreault et al. (2006), Hernández-Bastida et al. (2009), Baumgartner et al. (2015), **Oh et al. (2018)**
- Graphical model: Frees et al. (2014); Shi et al. (2015); Garrido et al. (2016)

Table of Contents

- Introduction I: Background Knowledge
- Introduction II: Literature Review
- Our Copula Model
- Data Analysis and Conclusion

Copula: Modelling the dependence

Summarization of Czado et al. (2012)'s model.

- Frequency: $N \sim \mathrm{Pois}(\lambda)$
- Mean severity: $M \sim \mathrm{Gamma}(\mu, v)$ with mean μ and variance μ^2/v
- $N \sim F_1$ and $M \sim F_2$
- Joint Modelling:

$$(N, M) \sim C_\rho(F_1, F_2)$$

  where C_ρ is Gaussian copula with correlation coefficient ρ.
- Estimation result: $\rho = 0.1366$ with Full comprehensive car insurance in Germany in the year 2000.
  - Contradiction to the literature: Strong positive correlation between frequency and severity ?
  - Literature (Frees et al., 2014; Shi et al., 2015; Garrido et al., 2016)
    - Comprehensive insurance: strong negative dependence
    - Liability insurance: weak dependence

Joint Modelling:

$$(N, M) \sim C(F_1, F_2)$$

with $N \sim \mathrm{Pois}(\lambda)$ and $M \sim \mathrm{Gamma}(\mu, v)$.

Issues in Copula Modelling

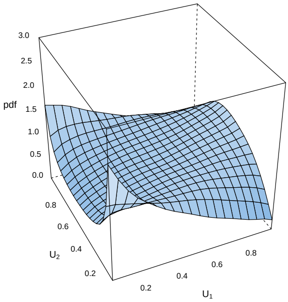
- Is $M \sim \mathrm{Gamma}(\mu, v)$ proper?
- Is the choice of Copula family C_Θ proper?
  - What is the corresponding copula $C^\perp$ if frequency and severities are independent?
  - Good candidate for copula family C_Θ should include $C^\perp$

Choice of right copula

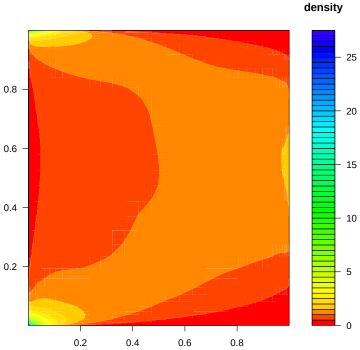
Question: What is copula $C^{\perp}$ if frequency and severities are independent?
Simulation study under independence assumption:

- $N \sim \mathrm{Pois}(\lambda)$
- $Y_1, \cdots, Y_N \overset{\text{i.i.d.}}{\sim} \mathrm{Gamma}(\mu, v)$, Note $(N \perp Y_i)$
- Define $M := \begin{cases} \dfrac{\sum_{i=1}^{N} Y_i}{N}, & N > 0 \\ \emptyset, & N = 0 \end{cases}$
- Simulation of $(N, M) \Rightarrow$ Empirical copula

Choice of right copula



(a) 3D Plot: Density of Copula



(b) Contour: Density of Copula

Figure: Density Estimation of Copula of (N, M) using Kernel Density Estimation

Choice of right copula?

- No known copula family includes the such copula!
- Construction of such copula family is not trivial.

Other option?

- Modelling with mirco-level data:

$$(N, Y_1, \cdots, Y_N) \sim C_\theta(F_1, F_2, \cdots, F_2)$$

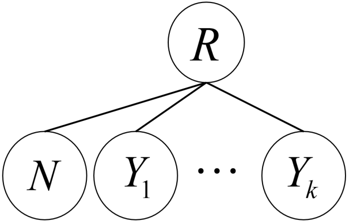
  where, for example, F_1: Poisson and F_2:Gamma
- Difficulty:
  - Choice of multi dimensional copula family C_θ
  - Not a classical statistical method: Dimension of $(N, Y_1, \cdots, Y_N)$ changes with N.

Table of Contents

- Introduction I: Background Knowledge
- Introduction II: Literature Review
- Our Copula Model
- Data Analysis and Conclusion

Review: Shared Random Effect Model

- Shared random effect model (Baumgartner et al., 2015)



- Frequency

$$N \mid R \sim \mathrm{Pois}(\lambda)$$

where $\lambda = \exp(\alpha_0 + R)$

- Severity

$$Y_j \mid R \overset{\text{i.i.d.}}{\sim} \mathrm{Gamma}(\mu, v)$$

where $\mu = \exp(\alpha_1 + \alpha_2 R)$

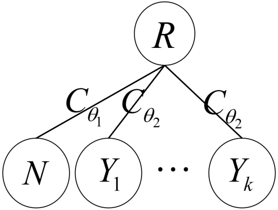
- Random effect $R \sim \pi$

- Handy joint distribution expression $(N, Y_1, \cdots, Y_k)$

$$\int f(r, n, y_1, \cdots, y_k)\mathrm{d}r = \int \pi(r) f(n \mid r) \prod_{j=1}^{k} f(y_j \mid r)\mathrm{d}r$$

Construction of Factor Copula Model

- Factor Copula Model



- Marginal Distributions:

$$N \sim F_1 \quad \text{and} \quad Y_j \sim F_2$$

- C_θ: a bivariate Gaussian copula with corr$= \theta$
- Random effect $R \sim N(0, 1)$

- Handy joint distribution expression of $(N, Y_1, \cdots, Y_k)$

$$\int f(r, n, y_1, \cdots, y_k)\mathrm{d}r = \int \phi(r) f(n|r) \prod_{j=1}^{k} f(y_j|r)\mathrm{d}r$$

- After some math: $(N, Y_1, \cdots, Y_k) \sim C_{\boldsymbol{\Sigma}}(F_1, F_2, \cdots, F_2)$ where $C_{\boldsymbol{\Sigma}}$ is a Gaussian copula with correlation matrix $\boldsymbol{\Sigma}$

- Model:

$$(N, Y_1, \cdots, Y_k) \sim C_{\mathbf{\Sigma}}(F_1, F_2, \cdots, F_2), \quad \text{for} \quad k \in \mathcal{N}$$

- $\mathbf{\Sigma}$ is a $(k+1) \times (k+1)$ correlation matrix

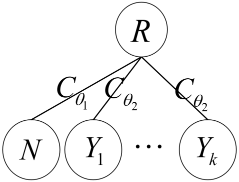
$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho_1 & \rho_1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & 1 & \rho_2 & \rho_2 & \cdots & \rho_2 \\ \rho_1 & \rho_2 & 1 & \rho_2 & \cdots & \rho_2 \\ \rho_1 & \rho_2 & \rho_2 & 1 & \cdots & \rho_2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_1 & \rho_2 & \rho_2 & \rho_2 & \cdots & 1 \end{pmatrix}$$

where $\rho_1 = \theta_1\theta_2$ and $\rho_2 = \theta_2^2$ which is positive definite if and only if

$$|\theta_1| < 1 \quad \text{and} \quad |\theta_2| < 1 \tag{1}$$

Factor Copula Model $\Rightarrow$ Copula Model?

- Factor Copula (Random effect) Model



$\Downarrow$ Integrating R

- Copula Model:

$$(N, Y_1, \cdots, Y_k) \sim C_{\boldsymbol{\Sigma}}(F_1, F_2, \cdots, F_2)$$

- Question:
  - Why starting from factor copula model?
  - Define copula model first?

Factor Copula Model $\Rightarrow$ Copula Model?

- Defining a copula model

$$(N, Y_1, \cdots, Y_k) \sim C_{\mathbf{\Sigma}}(F_1, F_2, \cdots, F_2)$$

is possible without the help of the factor copula model.

- but with the help of factor copula representation...
  - Easy extension into multi-year compound model (Valdez, Oh, Jung, and Ahn, 2019)

$$N_1, \cdots, N_t, (Y_{1,1}, \cdots, Y_{1,k}), \cdots, (Y_{t,1}, \cdots, Y_{t,k})$$

    - Complex covariance matrix in multiple year model
    - Convenient posterior ratemaking
  - Generation of various copula models is possible
    - t-copula (Valdez, Oh, Jung, and Ahn, 2019)
    - Archimedean copula (Cossette et al., 2018)

Table of Contents

- Introduction I: Background Knowledge
- Introduction II: Literature Review
- Our Copula Model
- Data Analysis and Conclusion

Data Analysis: Massachusetts Automobile Claims

- Insurance automobile claims in 2006 (Ferreira Jr and Minikel, 2012):
  - 3.25 million policies from liability insurances

Number of Claims	0	1	2	3	4	5
Count	3,164,447	128,311	5,463	310	31	3

- Explanatory variables
  - Primary driver charasteristics: Adult, Business, Youth, Senior
  - Territory group: 6 sectors
- Previous study (Frees et al., 2014; Shi et al., 2015) using the following model:

$$Y_i \big| N \sim \mathrm{Gamma}(\mu_\mathrm{i}, \mathrm{v})$$

with

$$\log \mu_i = \boldsymbol{X_i}^{\mathrm{T}} \boldsymbol{\beta} + \beta_0 N,$$

conclude weak (positive) dependence between frequency and severity.

			95%CI	
parameters	Est	S.E.	lower	upper
α_0	8.1105	0.0087	8.0934	8.1275*
α_{12}	0.0725	0.0115	0.0500	0.0950*
α_{13}	0.0657	0.0130	0.0401	0.0913*
α_{14}	0.1035	0.0110	0.0820	0.1250*
α_{15}	0.1498	0.0107	0.1288	0.1709*
α_{22}	0.2875	0.0112	0.2655	0.3095*
α_{23}	0.0785	0.0260	0.0274	0.1295*
α_{24}	0.0954	0.0117	0.0725	0.1184*
α_{25}	0.0927	0.0128	0.0675	0.1178*
α_{26}	-0.0881	0.0092	-0.1062	-0.0701*
ν	0.7272	0.0024	0.7226	0.7318*
β_0	-3.5432	0.0074	-3.5577	-3.5287*
β_{12}	0.1350	0.0098	0.1158	0.1542*
β_{13}	0.1805	0.0111	0.1587	0.2022*
β_{14}	0.3016	0.0093	0.2832	0.3199*
β_{15}	0.4808	0.0091	0.4629	0.4987*
β_{22}	0.7108	0.0095	0.6920	0.7295*
β_{23}	0.2088	0.0222	0.1653	0.2522*
β_{24}	0.8135	0.0099	0.7940	0.8330*
β_{25}	0.4955	0.0109	0.4742	0.5169*
β_{26}	0.0157	0.0078	0.0004	0.0311*
ρ_0	-0.0139	0.0039	-0.0215	-0.0063*
$\rho 1$	0.0351	0.0128	0.0100	0.0603*

- Spearman's rho between N and Y_i (Within a year):

$$\rho_S(N, Y_i) = \frac{6}{\pi} \cdot \arcsin(\rho_1 \cdot 2) = -0.013$$

  - Similar with Frees et al. (2014); Shi et al. (2015); Garrido et al. (2016)

- Spearman's rho between Y_j and Y_i (Within a year):

$$\rho_S(Y_j, Y_i) = \frac{6}{\pi} \cdot \arcsin(\rho_2 \cdot 2) = 0.034$$

  - Dependence between Y_i's are stronger than dependence between frequency-severity

Data Analysis: Massachusetts Automobile Claims

Estimation under Czado et al. (2012)'s model using Massachusetts Automobile Claims.

- Frequency: $N \sim F_1 = \mathrm{Pois}(\lambda)$
- Mean severity: $M \sim F_2 = \mathrm{Gamma}(\mu, v)$ with mean μ and variance μ^2/v
- Joint Modelling:

$$(N, M) \sim C_\rho(F_1, F_2)$$

  where C_ρ is Gaussian copula with correlation coefficient ρ.
- Estimation result: $\rho = 0.1413$
- Comparable to our result: $\mathrm{corr}\,(N, Y) = -0.013$

Summary and Conclusion

- Modelling the dependence in frequency and severity:
  - Use of N in explanatory variable of severity
  - Random effect model
  - Copula model
- In copula modelling, caution is required.
  - Impossibility of modelling N and M
  - Required to define joint model $(N, Y_1, \cdots, Y_N)$ (via factor copula model)
- Actual data analysis:
  - Dependence between frequency and severities were minor compared with dependence between severities.
- Our model can be extended into multiple years (Valdez, Oh, Jung, and Ahn, 2019)

Thank You.

References I

Baumgartner, C., Gruber, L. F., and Czado, C. (2015). Bayesian total loss estimation using shared random effects. *Insurance: Mathematics and Economics*, 62:194–201.

Boudreault, M., Cossette, H., Landriault, D., and Marceau, E. (2006). On a risk model with dependence between interclaim arrivals and claim sizes. *Scand. Actuar. J.*, (5):265–285.

Cossette, H., Marceau, E., and Mtalai, I. (2018). Collective risk models with hierarchical archimedean copulas.

Czado, C., Kastenmeier, R., Brechmann, E. C., and Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scand. Actuar. J.*, (4):278–305.

Ferreira Jr, J. and Minikel, E. (2012). Measuring per mile risk for pay-as-you-drive automobile insurance. *Transportation Research Record: Journal of the Transportation Research Board*, (2297):97–103.

Frees, E. W., Derrig, R. A., and Meyers, G. (2014). *Predictive Modeling Applications in Actuarial Science*, volume 1. Cambridge University Press.

Garrido, J., Genest, C., and Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, 70:205 – 215.

Hernández-Bastida, A., Fernández-Sánchez, M. P., and Gómez-Déniz, E. (2009). The net Bayes premium with dependence between the risk profiles. *Insurance Math. Econom.*, 45(2):247–254.

References II

Oh, R., Shi, P., and Ahn, J. Y. (2018). Implementation of bonus-malus system under the dependence between frequency and severity. *Working Paper*.

Shi, P., Feng, X., and Ivantsova, A. (2015). Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics*, 64:417–428.

Copula model with micro-level data: Dimensions

- Modelling of $(N, Y_1, \cdots, Y_N) \sim C_{\boldsymbol{\Sigma}^{(N)}_{\rho_1,\rho_2}}(F_1, F_2, \cdots, F_2)$ can be nonsense.
  - Dimension changes with N
- Solution:
  - Use the following model:

$$(N, Y_1, \cdots, Y_\infty) \sim C_{\boldsymbol{\Sigma}^{(\infty)}_{\rho_1,\rho_2}}(F_1, F_2, \cdots, F_2)$$

  - Assumption: among full data

$$(N, Y_1, \cdots, Y_N, Y_{N+1}, \cdots, Y_\infty)$$

    we only observe

$$(N, Y_1, \cdots, Y_N)^1$$

[1] This setting is common in missing data analysis.