Investigating High-Dimensional Problems in Actuarial Science, Dependence Modelling, and Quantitative Risk Management

Christopher Blier-Wong

Laboratoire ACT&RISK
École d'actuariat
Université Laval, Québec, Canada

chblw@ulaval.ca

18 septembre 2023

Université Laval — Faculté des sciences et de génie, École d'actuariat · Institut intelligence et données · crdm.ul — Centre de recherche en données massives de l'Université Laval · Quantact

Introduction

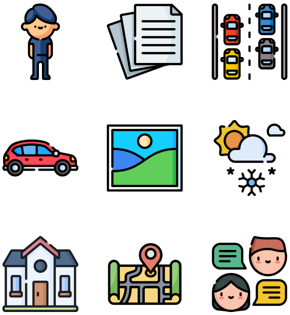
Introduction

- Topic of this thesis: high-dimensional actuarial science
- Motivation: solve problems in actuarial science where *dimension*, *size* or *computation scaling* is an issue
- Objective: leverage new machine learning techniques to simplify modelling
- Objective: develop new mathematical (probabilistic, number theoretic) tools to avoid tedious computations

From classical to unstructured ratemaking

Low-dimensional, vectorial data

High-dimensional, unstructured data

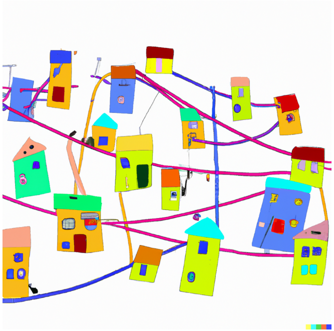


Images from Freepik and fjstudio on `flaticon.com`

High-dimensional risk aggregation with dependence

Small group of similar risks

Large group of heterogeneous, dependent risks



Images generated by DALL-E 2

Chapter	Issues due to high dimensions	Solution proposed	Results
1 and 2	- High-dimensional variables - Complicated format	Representation learning	- Low-dimensional features - Simple vectorial format
3 and 4	- Many parameters with elaborate parameter space - Uninterpretable dependence structure	Stochastic representation of FGM copulas with multivariate symmetric Bernoulli random vectors	- Few dependence parameters - Interpretable dependence parameters
5	- Tedious expressions - Uninterpretable dependence structure	Stochastic representation of FGM copulas with order statistics	- Simple and intuitive expressions - Interpretable dependence parameters
6	- No solution for arbitrary discrete rvs - Tedious convolution operations	Transformation into the Fourier space	- Efficient numerical methods for discrete rvs - Simple and intuitive expressions

Part 1: High-dimensional data in ratemaking models

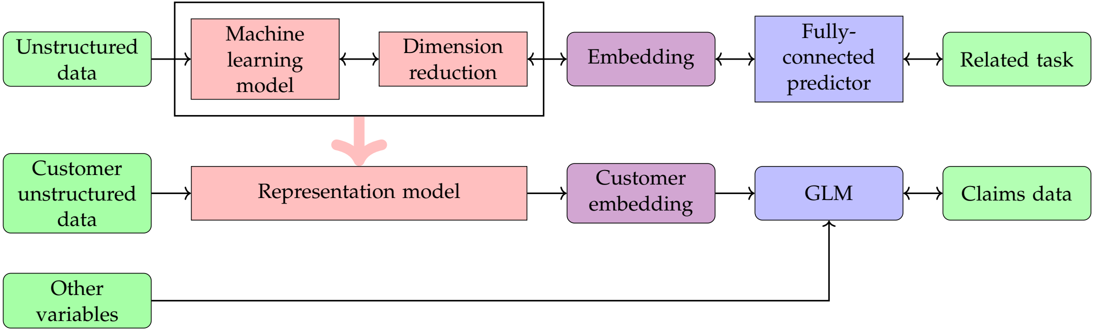
Big data and actuarial science

To improve pricing, an insurance company may

1. Gather more data
   - Gather more observations (costly)
   - Ask more questions upon quoting
   - Collect data online
2. Use more flexible predictive models to capture non-linear transformations and interactions
   - E.g. going from GLMs to GBMs or neural networks
   - Need more data (10x more data than degrees of freedom)
3. Use better representations
   - Find useful non-linear transformations and interactions
   - Without using the response variable (no impact on degrees of freedom)
   - Lose interpretability

Representation learning framework

Use a representation learning framework proposed in [Blier-Wong et al., 2021]



GEOGRAPHIC RATEMAKING WITH SPATIAL EMBEDDINGS

BY

CHRISTOPHER BLIER-WONG, HÉLÈNE COSSETTE,
LUC LAMONTAGNE AND ETIENNE MARCEAU

ASTIN Bulletin
The Journal of the International Actuarial Association

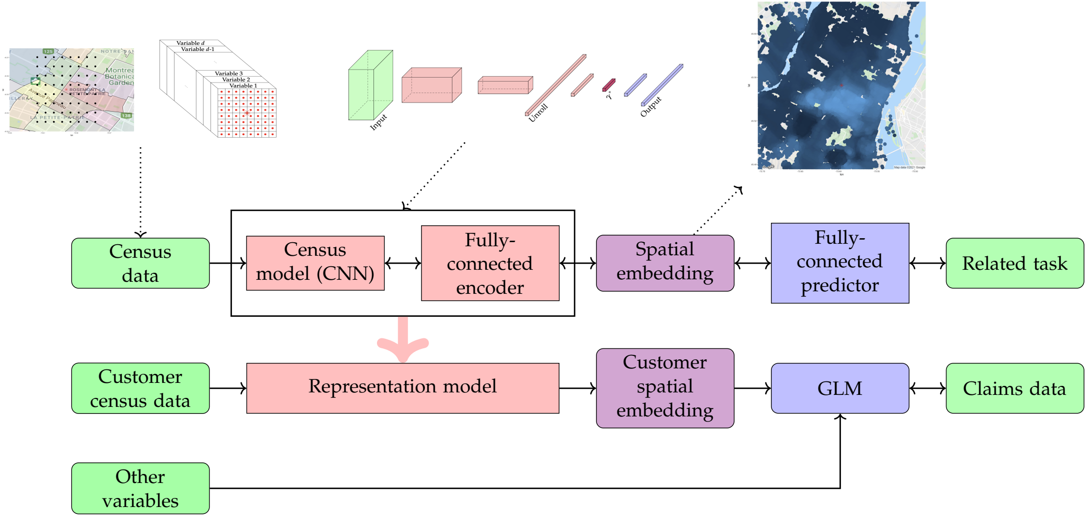
- Geographic ratemaking captures the spatial effects that parametric components fail to model
- Spatial embeddings = Captures spatial effects in a vector
- Focus instead on what actually generates spatial risk
  - Landform
  - Weather
  - People

Spatial embeddings should have desirable attributes

1. Spatial embeddings are coordinate-based
2. Spatial embeddings encode relevant external information
3. Spatial embeddings must follow Tobler's first law of geography

Chapter 1: spatial ratemaking

Geographic ratemaking with spatial embeddings



Application setup:

- Accident frequency prediction
- Home insurance in Québec
- Over 2 000 000 contracts

Poisson GAM:

$$\ln(E[Y_i]) = \beta_0 + \ln \omega + \underbrace{\sum_{j=1}^{p} x_{ij}\alpha_j,}_{\text{traditional component}} + \underbrace{f_k(lon_i,\, lat_i)}_{\text{spline component}} + \underbrace{\sum_{j=1}^{\ell} \gamma^*_{ij}\beta_j,}_{\text{embedding component}}$$

Train and test deviance in Montréal

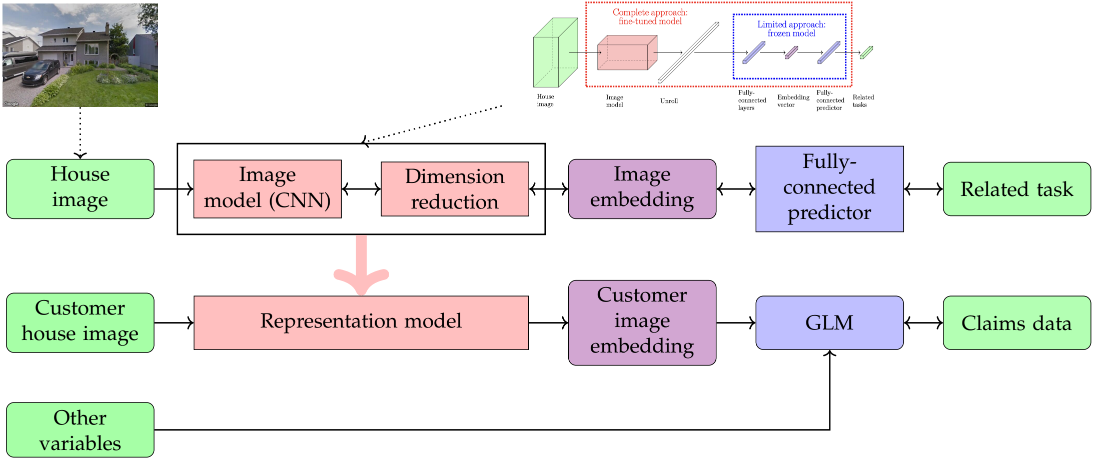
	Without embeddings				With embeddings γ^*			
k	Training	Test	DoF	Time (s)	Training	Test	DoF	Time (s)
0	–	–	–	–	66013	**14383**	19	58
3	66149	14390	6.69	242	65952	14399	23.70	483
5	65991	14400	16.49	108	65886	14402	33.61	1553
8	65838	14390	34.00	1306	65778	14388	48.65	1024
10	65766	14389	46.03	2201	65733	14388	58.21	1540
15	65691	14389	64.44	2533	65683	14391	73.42	3617
20	65652	14386	75.37	7733	65651	14387	82.29	12368
25	65644	14386	80.36	50902	65642	14387	86.39	50763

- A representation-learning approach for insurance pricing with images
- With Luc Lamontagne and Etienne Marceau
- Under revision in the ASTIN Bulletin

Chapter 2: image ratemaking

Image preparation

Representation learning framework



Application setup:

- Accident frequency prediction
- Home insurance in Québec City
- Over 50 000 contracts
- Different models for perils fire, theft, water, wind, sewer backup, other

Poisson GLM:

$$\ln(E[Y_i]) = \beta_0 + \ln \omega + \underbrace{\sum_{j=1}^{p} x_{ij}\alpha_j,}_{\text{traditional component}} + \underbrace{\sum_{j=1}^{\ell} \gamma^*_{ij}\beta_j,}_{\text{embedding component}}$$

Model	ℓ	Theft	Other	SBU	Water	Hail	Wind	Fire	Total
Baseline	0	1468.91	2028.72	2697.17	3831.07	163.47	884.08	479.96	8661.44
ResNet18	8	1468.06 (**1**)	2029.27 (0)	2670.60 (**3**)	3825.91 (**1**)	166.78 (0)	881.25 (**1**)	481.42 (**3**)	8665.47 (**2**)
	16	1468.12 (**8**)	2028.05 (**3**)	2667.97 (**7**)	3817.73 (**2**)	165.56 (0)	884.93 (0)	477.01 (0)	8657.23 (**7**)
	32	1460.68 (**2**)	2035.17 (**3**)	2662.17 (**9**)	3816.33 (**5**)	182.68 (**4**)	880.27 (0)	477.92 (**2**)	8656.09 (**2**)
ResNet50	8	1459.36 (**3**)	2030.22 (0)	2692.66 (**3**)	3825.50 (**1**)	163.87 (0)	885.85 (**1**)	477.11 (**3**)	8670.48 (**3**)
	16	1463.73 (**2**)	2028.17 (**2**)	2661.89 (**9**)	3820.33 (**7**)	170.42 (0)	884.61 (0)	475.02 (0)	8658.43 (**4**)
	32	1462.73 (**7**)	2029.40 (**3**)	2661.36 (**8**)	3816.36 (**2**)	177.95 (0)	878.15 (**2**)	481.98 (1)	8657.40 (**2**)
ResNet101	8	1460.39 (**4**)	2026.99 (**5**)	2671.30 (**5**)	3821.99 (0)	168.80 (0)	878.75 (**3**)	477.50 (**1**)	8659.38 (**3**)
	16	1464.28 (**7**)	2028.09 (0)	2663.49 (**6**)	3820.84 (**3**)	169.21 (0)	883.01 (**1**)	474.28 (0)	8659.49 (**6**)
	32	1468.98 (**3**)	2030.89 (**2**)	2665.01 (**8**)	3819.22 (**4**)	173.32 (**2**)	893.89 (1)	478.95 (1)	8664.47 (**10**)
DenseNet121	8	1473.56 (**5**)	2028.77 (**2**)	2672.94 (**5**)	3823.37 (**1**)	166.79 (**1**)	879.85 (**1**)	477.91 (0)	8664.57 (**2**)
	16	1467.33 (**2**)	2031.46 (**4**)	2665.68 (**6**)	3826.43 (0)	174.70 (0)	878.79 (**1**)	477.26 (0)	8661.09 (**4**)
	32	1477.16 (**6**)	2024.60 (**3**)	2668.62 (**7**)	3823.74 (**7**)	173.28 (0)	885.86 (0)	479.82 (0)	8660.95 (**5**)

Table: Testing deviance for frequency prediction with fine-tuned models.

Part 2: Stochastic representation of FGM copulas

- We consider a random vector $\boldsymbol{X}$
- Described by the joint cdf

$$F_{\mathbf{X}}(x_1, \ldots, x_n) = \Pr(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

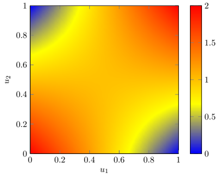
- This function has two components:
  1. The marginal cdfs

$$F_1(x_1) = \Pr(X_1 \leq x_1), \ldots, F_n(x_n) = \Pr(X_n \leq x_n)$$

  2. A dependence structure joining the marginal
- Copulas are mathematical objects that let us model/study dependence in probability models

UNIVERSITÉ LAVAL

- Farlie-Gumbel-Morgenstern (FGM) copula
- Bivariate:



$$C(u_1, u_2) = u_1 u_2 (1 + \theta_{12} \overline{u}_1 \overline{u}_2)$$

where $\overline{u} = 1 - u$

- Trivariate:

$$C(\boldsymbol{u}) = u_1 u_2 u_3 (1 + \theta_{12}\overline{u}_1\overline{u}_2 + \theta_{13}\overline{u}_1\overline{u}_3 + \theta_{23}\overline{u}_2\overline{u}_3 + \theta_{123}\overline{u}_1\overline{u}_2\overline{u}_3)$$

- Expression of the copula

$$C(\boldsymbol{u}) = \prod_{m=1}^{d} u_m \left(1 + \sum_{k=2}^{d} \sum_{1 \leq j_1 < \cdots < j_k \leq d} \theta_{j_1 \ldots j_k} \overline{u}_{j_1} \overline{u}_{j_2} \ldots \overline{u}_{j_k}\right) \quad \boldsymbol{u} \in [0,1]^d$$

- Number of parameters: $d^\star = 2^d - d - 1$

FGM copulas

Advantages of FGM copulas

- Capture multiple shapes of dependence ($2^d - d - 1$ parameters)
- Quadratic marginals: easy to integrate
- Admits analytic expressions

Limitations of the FGM copula

- Admits weak dependence, no tail dependence
- Dependence parameters difficult to interpret
- Tedious high-dimensional sampling
- Tedious stochastic comparison
- Tedious parameter space $\mathcal{T}_d$

A d-variate FGM copula exists if $\boldsymbol{\theta} \in \mathcal{T}_d$, where

$$\mathcal{T}_d = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{d^\star} : 1 + \sum_{k=2}^{d} \sum_{1 \leq j_1 < \cdots < j_k \leq d} \theta_{j_1 \ldots j_k} \varepsilon_{j_1} \varepsilon_{j_2} \ldots \varepsilon_{j_k} \geq 0 \right\}, \quad \{\varepsilon_{j_1}, \varepsilon_{j_2}, \ldots, \varepsilon_{j_k}\} \in \{-1, 1\}^d$$

Chapter 3: Stochastic representation of FGM copulas

Motivation

Computational Statistics and Data Analysis 173 (2022) 107506

ELSEVIER

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Stochastic representation of FGM copulas using multivariate Bernoulli random variables

Check for updates

Christopher Blier-Wong, Hélène Cossette, Etienne Marceau [*,1]

École d'actuariat, Université Laval, Canada

Theorem

Let $\boldsymbol{U}$ be a random vector such that $F_{\boldsymbol{U}}(\boldsymbol{u})$ is a FGM copula, then there exists a $\boldsymbol{I}$ such that $\boldsymbol{U}$ admits the representation

$$\boldsymbol{U} = (\boldsymbol{1} - \boldsymbol{I})\boldsymbol{V}_{[1]} + \boldsymbol{I}\boldsymbol{V}_{[2]},$$

where

- $\boldsymbol{V}_{[1]}$ is a vector of iid rvs distributed as the minimum order statistic of a uniform rvs out of a sample of two
- $\boldsymbol{V}_{[2]}$ is a vector of iid rvs distributed as the maximum order statistic of a uniform rvs out of a sample of two
- $\boldsymbol{I}$ is a vector of multivariate symmetric Bernoulli rvs
- $\boldsymbol{V}_{[1]}, \boldsymbol{V}_{[2]}, \boldsymbol{I}$ are independent

Limitations of the FGM copula

- ~~Dependence parameters difficult to interpret~~
- ~~Tedious high-dimensional sampling~~
- ~~Tedious stochastic comparison~~
- ~~Tedious parameter space $\mathcal{T}_d$~~
- Admits weak dependence, no tail dependence

New advantages of FGM copulas

- Interpretable dependence structures
- High-dimensional sampling
- Stochastic comparison
- Construction of subfamilies
- Reveal properties of FGM copulas
- Stochastic representation useful for applications

Adv. Appl. Probab. 1–30 (2023)
*doi:*10.1017/apr.2023.19

EXCHANGEABLE FGM COPULAS

CHRISTOPHER BLIER-WONG,* **
HÉLÈNE COSSETTE,* AND
ETIENNE MARCEAU,* *Université Laval*

Definition (Exchangeability)

A vector of rvs $\boldsymbol{U}$ is said exchangeable if

$$(U_1, \ldots, U_d) \overset{d}{=} (U_{\pi(1)}, \ldots, U_{\pi(d)})$$

for all permutation $(\pi(1), \ldots, \pi(d))$ of $(1, \ldots, d)$.

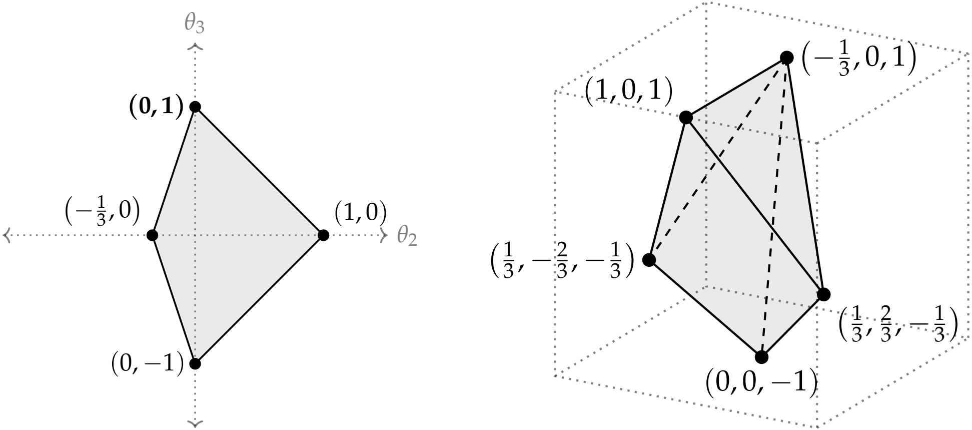
eFGM copula:

$$C_d(u_1, \ldots, u_d) = \prod_{j=1}^{d} u_j \left(1 + \sum_{k=2}^{d} \sum_{1 \leq j_1 < \cdots < j_k \leq d} \theta_k \overline{u}_{j_1} \ldots \overline{u}_{j_k}\right), \quad (u_1, \ldots, u_d) \in [0,1]^d$$

Parameter space

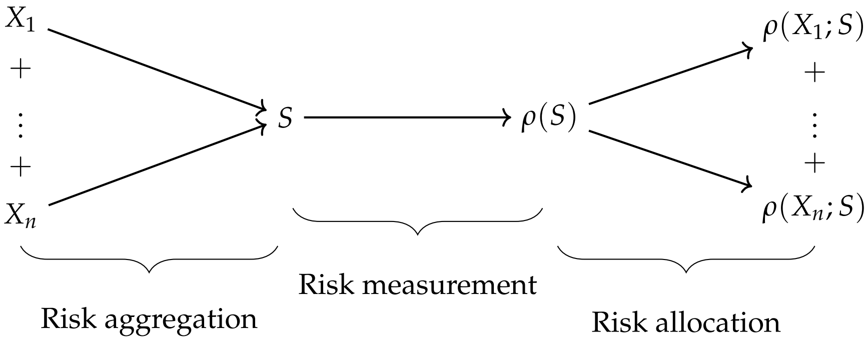
$$\mathcal{T}_d^* = \left\{(\theta_2, \ldots, \theta_d) \in \mathbb{R}^{d-1} : 1 + \sum_{k=2}^{d} \sum_{1 \leq j_1 < \cdots < j_k \leq d} \theta_k \varepsilon_{j_1} \ldots \varepsilon_{j_k} \geq 0\right\}$$

Efficient numerical method to obtain extreme points of $\mathcal{T}_d^*$



Part 3: High-dimensional risk aggregation

Risk aggregation and risk allocation



Direct approach: $d-1$ multiple integral. Risk aggregation: for $y > 0$,

$$f_S(y) = \int_0^y \int_0^{y-x_1} \cdots \int_0^{y-\sum_{j=1}^{d-2} x_i} f_{X_1, X_2, \ldots, X_d}\left(x_1, x_2, \ldots, y - \sum_{j=1}^{d-1} x_i\right) \mathrm{d}x_{d-1} \ldots \mathrm{d}x_2 \, \mathrm{d}x_1$$

Insurance: Mathematics and Economics 111 (2023) 102–120

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Risk aggregation with FGM copulas

Christopher Blier-Wong *, Hélène Cossette, Etienne Marceau

École d'actuariat, Université Laval, Québec, Canada

- We consider the agregate rv $S = X_1 + \cdots + X_n$
- We study the effect of dependence when C is a FGM copula and

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = C(F_{X_1}(x_1), \ldots, F_{X_n}(x_n))$$

- Research questions:
  - What is the distribution of S?
  - How can we compare S under different dependence structures?
  - How can we allocate risk measures or share risks?
- Already studied in [Cossette et al., 2013], using the *direct approach*

Theorem

Let $\boldsymbol{X}$ be a random vector such that $F_{\boldsymbol{X}}(\boldsymbol{x}) = C(F_1(x_1), \ldots, X_d(x_d))$ where C is a FGM copula, then there exists a random vector $\boldsymbol{I}$ such that $\boldsymbol{X}$ admits the representation

$$\boldsymbol{X} = (\boldsymbol{1} - \boldsymbol{I})\boldsymbol{X}'_{[1]} + \boldsymbol{I}\boldsymbol{X}'_{[2]}.$$

Then, S admits the representation

$$S = \sum_{k=1}^{d} \left\{ (1 - I_k) X_{k,[1]} + I_k X_{k,[2]} \right\}$$

Proposition

If the marginals of $\mathbf{X}$ *have cdfs belonging to the same family of distributions that are closed under*

1. *order statistics*
2. *convolution*
3. *mixtures*

then the cdf of $S = X_1 + \cdots + X_n$ *will belong to the same family of distributions.*

- Mixed Erlang (exponential, gamma, generalized Erlang, phase-type)
- Matrix Exponential
- We compute risk measures for S (mean, variance, VaR, TVaR) and risk allocations based on Euler's rule

Generating function method for the efficient computation of expected allocations*

Christopher Blier-Wong[†], Hélène Cossette, and Etienne Marceau

École d'actuariat, Université Laval, Québec, Canada

- Peer-to-peer insurance pricing schemes: compute the contribution of participants according to risk sharing rule [Denuit, 2020]
- Conditional mean risk sharing rule [Denuit and Dhaene, 2012]: popular choice
- Satisfies desirable properties [Denuit et al., 2022], axiomatic characterization [Jiao et al., 2022]
- If $S = k$, price for the ith participant is

$$E[X_i|S = k] = \frac{E\left[X_i \times 1_{\{S=k\}}\right]}{\Pr(S = k)}, \quad i \in \{1, \ldots, n\}$$

- Similar requirement for contributions to (T)VaR(S) based on Euler's principle

Support of each rv: $h\mathbb{N}_0 = \{0, h, 2h, \dots\}$, with some fixed $h > 0$ (suppose $h = 1$ for notational simplicity)

Objectives

1. Provide convenient representations for the values of $E\left[X_i \times 1_{\{S=k\}}\right]$ for $i \in \{1, \dots, n\}$ and $k \in \mathbb{N}_0$
2. Provide efficient computation methods for $E\left[X_i \times 1_{\{S=k\}}\right]$

- Large pools/portfolios
- Heterogeneous risks
- Dependent risks

Main result: generating function for the (cumulative) expected allocations of risk X_1

Theorem

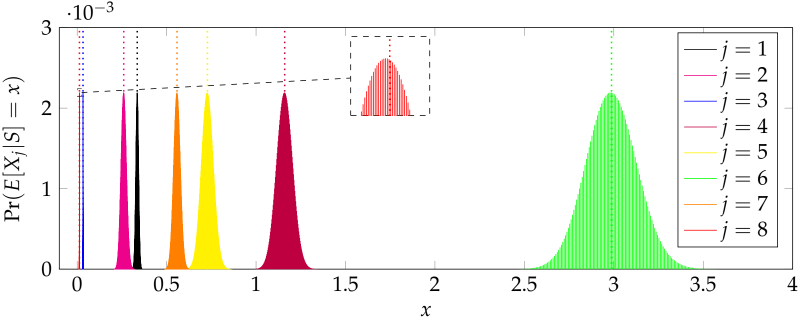
We have

$$\mathcal{P}_S^{[1]}(t) := \left[t_1 \times \frac{\partial}{\partial t_1}\mathcal{P}_{\mathbf{X}}(t_1, t_2, \ldots, t_n)\right]_{t_1=\cdots=t_n=t} = \sum_{k=0}^{\infty} t^k E\left[X_1 \times 1_{\{S=k\}}\right].$$

Then, $\mathcal{P}_S^{[1]}(t)$ *is the generating function for the sequence of expected allocations.*

- Consider a portfolio of 10 000 independent risks
- Compound Poisson distributions with rate λ_j
- Severity rv $B_j \sim NBinom(r_j, q_j)$
- Distinct and arbitrarily fixed parameters
- Computes 1 600 $\times$ 10 000 conditional means at once
- Takes approximately 16 seconds on a personal computer

Distribution of conditional means for the first eight contracts



Parameters for first eight risks

j	1	2	3	4	5	6	7	8
λ_j	0.16	0.03	0.03	0.24	0.12	0.47	0.15	0.01
q_j	0.49	0.42	0.46	0.45	0.49	0.44	0.44	0.48
r_j	2	6	1	4	6	5	3	1
$E[X_j]$	0.34	0.26	0.03	1.16	0.73	2.99	0.56	0.01

Conclusion

Conclusion

Summary and future research

Chapter	Issues due to high dimensions	Solution proposed	Results	Future research
1 and 2	- High-dimensional variables - Complicated format	Representation learning	- Low-dimensional features - Simple vectorial format	- Telematics, textual data - Interpretability - Update through time
3 and 4	- Many parameters with elaborate parameter space - Uninterpretable dependence structure	Stochastic representation of FGM copulas with multivariate symmetric Bernoulli random vectors	- Few dependence parameters - Interpretable dependence parameters	- Generalized FGM copulas - High-dimensional estimation
5	- Tedious expressions - Uninterpretable dependence structure	Stochastic representation of FGM copulas with order statistics	- Simple and intuitive expressions - Interpretable dependence parameters	- Collective risk models - Bernstein copulas
6	- No solution for arbitrary discrete rvs - Tedious convolution operations	Transformation into the Fourier space	- Efficient numerical methods for discrete rvs - Simple and intuitive expressions	- Resolve numerical instability (tilting) - Continuous version with LST

Conclusion

Summary of contributions

1. Blier-Wong, C., Cossette, H., Lamontagne, L., & Marceau, E. (2022). Geographic ratemaking with spatial embeddings. ASTIN Bulletin: The Journal of the IAA, 52(1), 1-31.
2. Blier-Wong, C., Lamontagne, L., & Marceau, E. (2022). A representation-learning approach for insurance pricing with images. ASTIN Bulletin: The Journal of the IAA. Under revision.
3. Blier-Wong, C., Cossette, H., & Marceau, E. (2022). Stochastic representation of FGM copulas using multivariate Bernoulli random variables. Computational Statistics & Data Analysis, 173, 107506.
4. Blier-Wong, C., Cossette, H., & Marceau, E. (2024). Exchangeable FGM copulas. Advances in Applied Probability.
5. Blier-Wong, C., Cossette, H., & Marceau, E. (2023). Risk aggregation with FGM copulas. Insurance: Mathematics and Economics, 111, 102-120.
6. Blier-Wong, C., Cossette, H., & Marceau, E. (2022). Generating function method for the efficient computation of expected allocations. arXiv preprint

Conclusion

Thank you for your attention.

References I

Blier-Wong, C., Baillargeon, J.-T., Cossette, H., Lamontagne, L., and Marceau, E. (2021).
Rethinking representations in P&C actuarial science with deep neural networks.
arXiv:2102.05784 [stat].

Cossette, H., Côté, M.-P., Marceau, E., and Moutanabbir, K. (2013).
Multivariate distribution defined with Farlie–Gumbel–Morgenstern copula and mixed Erlang marginals: Aggregation and capital allocation.
Insurance: Mathematics and Economics, 52(3):560–572.

Denuit, M. (2020).
Investing in your own and peers' risks: The simple analytics of P2P insurance.
European Actuarial Journal, 10(2):335–359.

References II

Denuit, M. and Dhaene, J. (2012).
Convex order and comonotonic conditional mean risk sharing.
Insurance: Mathematics and Economics, 51(2):265–270.

Denuit, M., Dhaene, J., and Robert, C. Y. (2022).
Risk-sharing rules and their properties, with applications to peer-to-peer insurance.
Journal of Risk and Insurance.

Fontana, R. and Semeraro, P. (2018).
Representation of multivariate Bernoulli distributions with a given set of specified moments.
Journal of Multivariate Analysis, 168:290–303.

References III

Jiao, Z., Liu, Y., and Wang, R. (2022).
An axiomatic theory for anonymized risk sharing.
arXiv preprint arXiv:2208.07533.

Sharakhmetov, S. and Ibragimov, R. (2002).
A Characterization of Joint Distribution of Two-Valued Random Variables and Its Applications.
Journal of Multivariate Analysis, 83(2):389–408.