

Mesures de dépendance entre variables : corrélation, copules, et autres coefficients de dépendance

Renata Alcoforado : Universidade Federal de Pernambuco, Brésil & I2M

Denys Pommeret : Institut de Mathématiques de Marseille (I2M)

Samuel Stocksieker : Actunéo & I2M



Objectifs de la présentation

- Une corrélation inattendue, ou comment remettre en cause certaines hypothèses d'indépendance
- Au-delà de la corrélation : d'autres coefficients de dépendance
- Les copules : une approche efficace, appuyée par un package simple *COPULA*
- Des mesures de dépendance complexe, notamment l'HGR
- Illustrations dans le cas de deux variables ou plus

Corrélation entre deux variables

Coefficient de Pearson

$$\rho_{X,Y} = Cov(X, Y) / (\sigma_X \sigma_Y)$$

- Interprétation : mesure de dépendance linéaire.
- Points forts :
 - Capture les relations linéaires
- Faiblesses :
 - Ne capte pas les relations non linéaires
 - Sensible aux outliers

Tau de Kendall

$$\tau = \frac{C - D}{\frac{1}{2}n(n - 1)}$$

Avec C = paires concordantes et D = paires discordantes

- Points forts :
 - Robuste aux valeurs extrêmes
 - Capte les relations monotones
- Faiblesses :
 - Moins efficace sur petits échantillons

Corrélation de Spearman

$$\rho_s = \text{Cov}(\text{rg}(X), \text{rg}(Y)) / (\sigma_{\text{rg}(X)} \sigma_{\text{rg}(Y)})$$

où rg désigne le rang des observations

- Points forts :
 - Robuste aux valeurs extrêmes
 - Capte les relations monotones non linéaires
- Faiblesses :
 - Moins efficace si la relation n'est pas monotone

Real Data

The data consist of two different branches that we work with as three groups:

- Total portfolio – with all data of the company
- Group 1 – liability insurance
- Group 2 – housing insurance

Real Data

Period: Every claim that happened in the period from 01/01/2015 until 31/12/2019.

Frequency of claims: For each group, it represents the number of claims per day during this five-year period (1826 days).

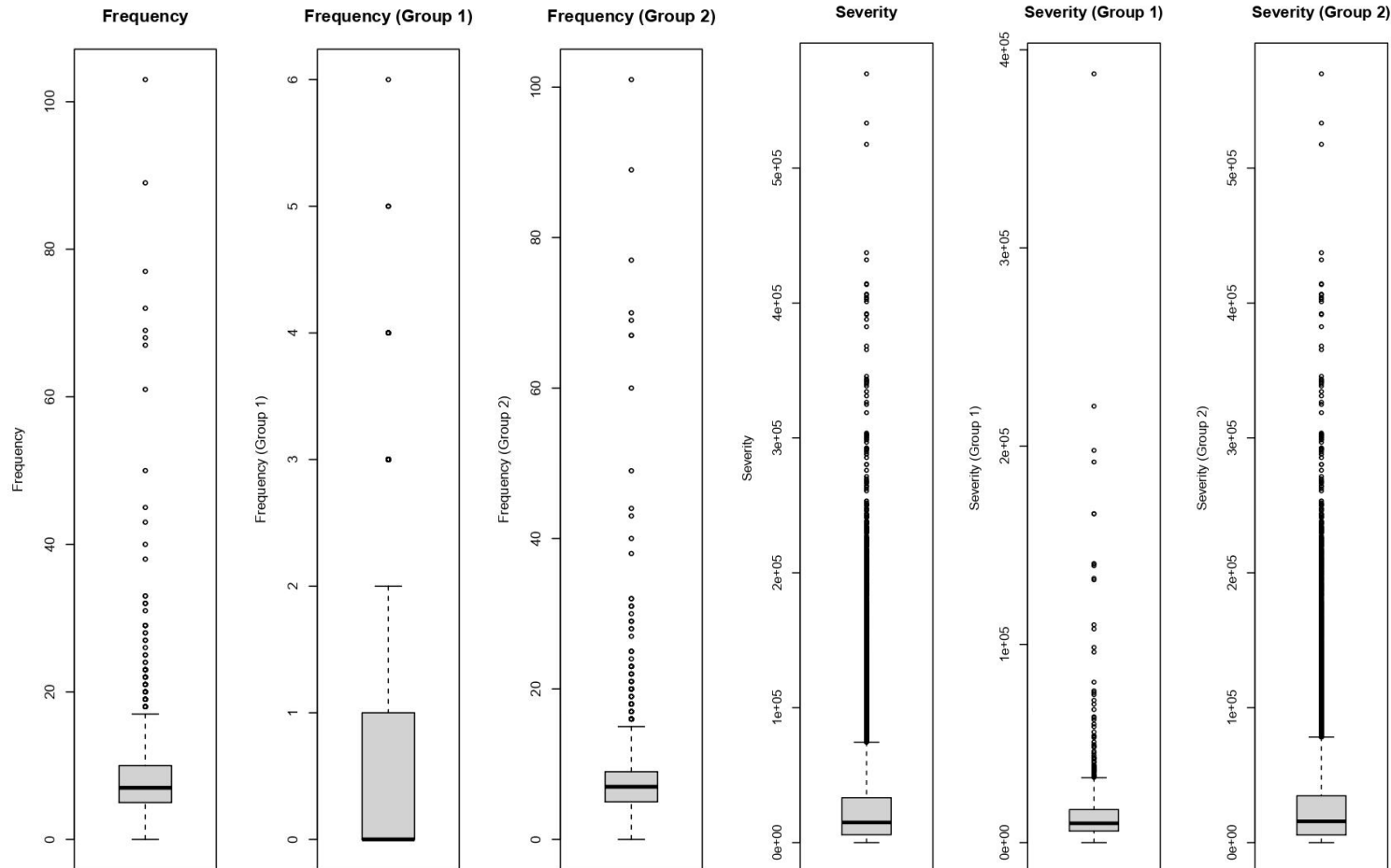
Claims: 15,665 (966 G1 and 14,699 G2)

Deflation: The severity of claims was deflated using inflation data from the World Bank, and all amounts were deflated to the first day of the period under study (01/01/2015).

Real Data

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Severity	0.0194	5820	14910	28820	33270	569918
Frequency	0	5	7	8.579	10	103
Severity 1	1.35	4347.08	7897.25	11972.21	12711.83	315022.6
Frequency 1	0	0	0	0.529	1	6
Severity 2	0.0194	4381	12125	22884	26582	462884
Frequency 2	0	5	7	8.05	9	101

Boxplot for the Frequency and Severity of Claims



Correlation

In this case we used the following variables:

- Frequency
- Severity per day
- Mean of the Severity per day
- Median of the Severity per day

Since none of the data samples were drawn from a Normal universe, we use Kendall and Spearman coefficients.

Spearman's correlation Table

	Frequency 1	Severity_Day 1	Mean_Sev 1	Median_Sev 1	Frequency 2	Severity_Day 2	Mean_Sev 2	Median_Sev 2
Frequency 1	-	0.5865	0.1575	0.0888	0.1129	0.0692	-	-
Severity_Day 1	0.5865	-	0.8737	0.7254	0.1243	0.0765	-	-
Mean_Severity_Day 1	0.1575	0.8737	-	0.9576	0.1253	0.0726	-	-
Median_Severity_Day 1	0.0888	0.7254	0.9576	-	0.0964	0.053	-	-
Frequency 2	0.1129	0.1243	0.1253	0.0964	-	0.6551	-	-0.0843
Severity_Day 2	0.0692	0.0765	0.0726	0.053	0.6551	-	0.7167	0.2735
Mean_Severity_Day 2	-	-	-	-	-	0.7167	-	0.5077
Median_Severity_Day 2	-	-	-	-	-0.0843	0.2735	0.5077	-

Spearman's correlation Table

	Frequency 1	Severity_Day 1	Mean_Sev 1	Median_Sev 1	Frequency 2	Severity_Day 2	Mean_Sev 2	Median_Sev 2
Frequency 1	-	0.5865	0.1575	0.0888	0.1129	0.0692	-	-
Severity_Day 1	0.5865	-	0.8737	0.7254	0.1243	0.0765	-	-
Mean_Severity_Day 1	0.1575	0.8737	-	0.9576	0.1253	0.0726	-	-
Median_Severity_Day 1	0.0888	0.7254	0.9576	-	0.0964	0.053	-	-
Frequency 2	0.1129	0.1243	0.1253	0.0964	-	0.6551	-	-0.0843
Severity_Day 2	0.0692	0.0765	0.0726	0.053	0.6551	-	0.7167	0.2735
Mean_Severity_Day 2	-	-	-	-	-	0.7167	-	0.5077
Median_Severity_Day 2	-	-	-	-	-0.0843	0.2735	0.5077	-

Some additional results

- Statistically significant correlation between Frequency and aggregate severity of claims per day in both groups.
- The correlation between Severity_Day and both Median and Mean shows that when the aggregate amount of severity per day increases, it is **not only due to the increase in Frequency** but also to the severity per claim on that day. This holds for both groups.

Models propose particular dependence

- Linear models describe dependence through a linear relationship between the mean of Y and the covariates X .
 - Generalized Linear Models (GLM) and Generalized Additive Models (GAM) extend this framework by introducing a link (or non parametric) function to relate $E(Y)$ to X .
 - GAMLSS further generalizes this idea by allowing dependence to arise through affine transformations of the covariates.
- ➔ For illustration, we apply GAMLSS to our dataset.
- ➔ It is important to keep in mind that **if a covariate is not significant in this model, it only means that it is not significant *within this specific modelling framework*. Other types of dependence may still exist, and we will investigate them in the following sections.**

GAMLSS Modelling

- We use Generalized Additive Models for Location, Scale and Shape (GAMLSS) because they provide a general framework for univariate regression-type statistical problems.
- GAMLSS allows flexibility in specifying the distribution of the response variable and makes it easy to include additional explanatory variables.
- As Denuit et al. (2019) state, in the GAMLSS approach the exponential dispersion distribution assumption for the response is relaxed, which removes the restriction that actuarial analyses had regarding the distributions used in the classical GLM/GAM setting.
- In our case, we have one main explanatory variable, and we also include dummy variables representing the year of the data.

GAMLSS Modelling: Fitted Distribution Families

We fitted each variable to identify which GAMLSS family provides the best fit according to the AIC.

Variable	Family	AIC
Frequency	Delaporte	10275.7
Severity	Box-Cox Power Exponential	342791
Median_Severity_Day	Exponential Gaussian	38092.5
Mean_Severity_Day	Box-Cox-t	39281.1
Frequency 1	Double Poisson	3592.03
Severity 1	Box-Cox Power Exponential	19506.7
Median_Severity_Day 1	Pareto Type 2	-58173.4
Mean_Severity_Day 1	Box-Cox Power Exponential	13430.9
Frequency 2	Delaporte	10079.5
Severity 2	Generalized beta 2 (second kind)	322628
Median_Severity_Day 2	Exponential Gaussian	38397.7
Mean_Severity_Day 2	Box-Cox-t	39429.9

Median Severity of Claims per Day

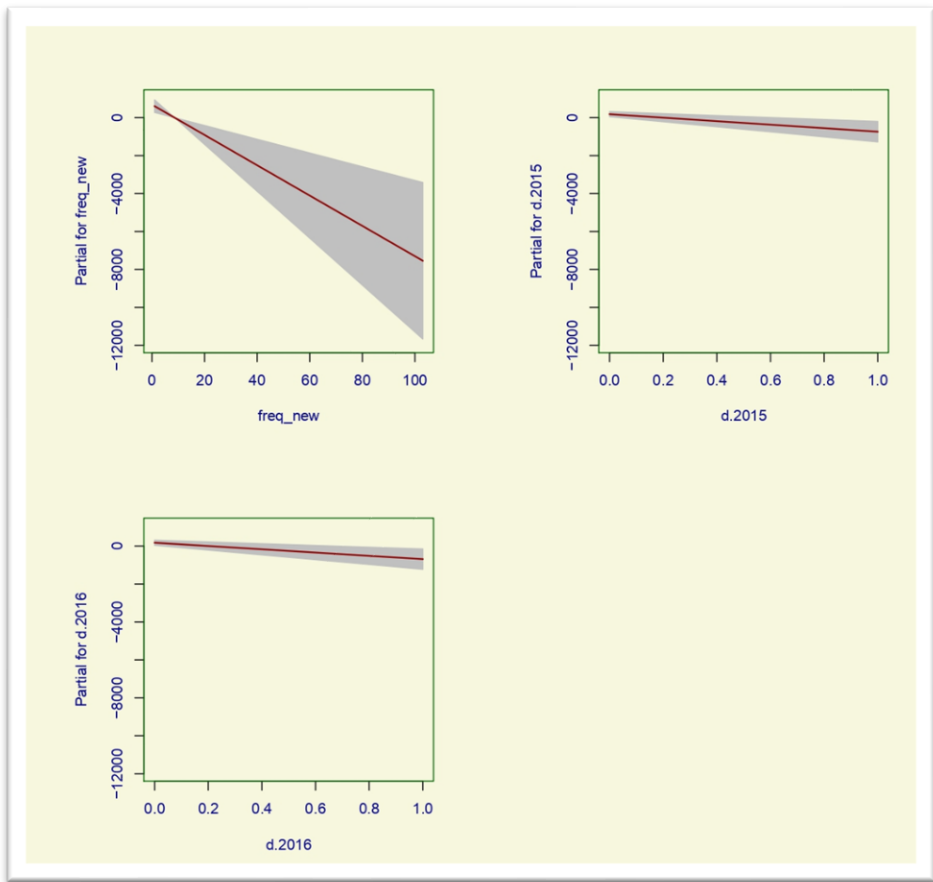
Using the Exponential Gaussian family of distributions, we specify three GAMLSS models for:

- Location (μ),
- Scale (σ),
- Shape (ν).

Link functions:

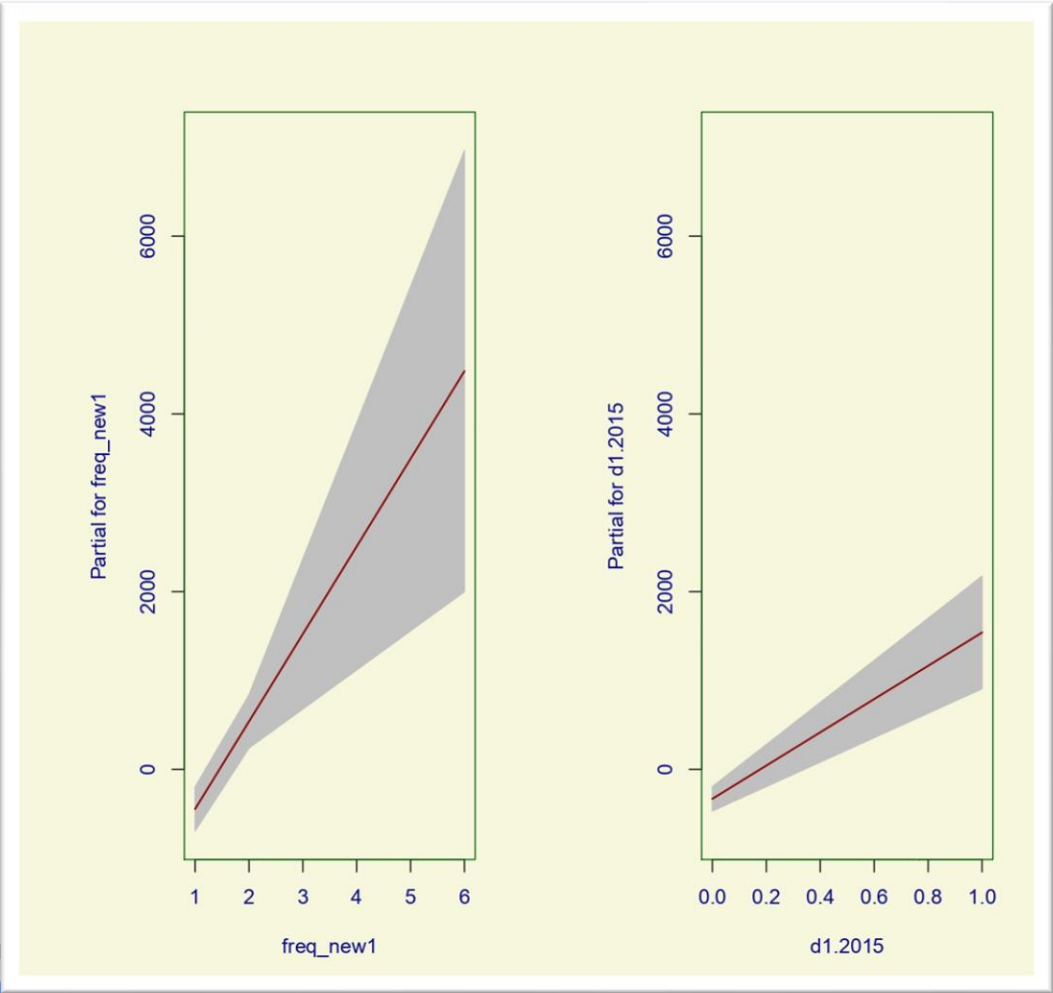
- For location (μ): identity
- For scale (σ): log
- For shape (ν): log

GAMLSS Modelling – Median Severity (Total Portfolio)



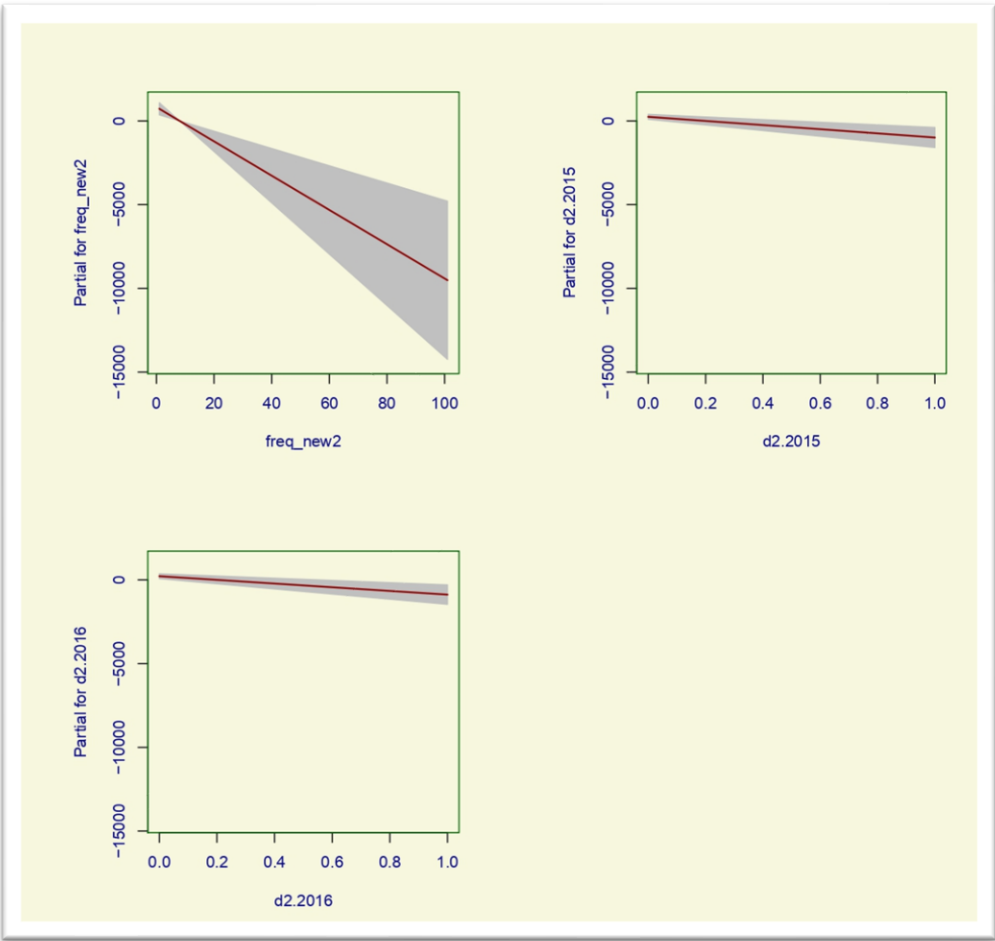
Coefficient	Estimate	Std Error	t value	p-value
mu link function: identity				
(Intercept)	5758.22	279.55	20.598	< 2e-16
freq	-79.91	25.58	-3.124	0.00181
d.2015	-924.26	385.96	-2.395	0.01674
d.2016	-863.89	403.05	-2.143	0.03221
sigma link function: log				
(Intercept)	7.80849	0.06537	119.4	< 2e-16
nu link function: log				
(Intercept)	9.20654	0.02985	308.4	< 2e-16

GAMLSS Modelling – Median Severity (Group 1)



Coefficient	Estimate	Std Error	t value	p-value
mu link function: identity				
(Intercept)	940.5	395.9	2.376	0.017799
freq	985.2	207.3	4.753	2.46E-06
d.2015	1873.7	482.1	3.887	0.000112
sigma link function: log				
(Intercept)	7.54157	0.07961	94.73	< 2e-16
nu link function: log				
(Intercept)	9.0607	0.04486	202	< 2e-16

GAMLSS Modelling – Median Severity (Group 2)



Coefficient	Estimate	Std Error	t value	p-value
mu link function: identity				
(Intercept)	5931.11	299.07	19.832	< 2e-16
freq	-102.39	28.25	-3.625	0.000297
d.2015	-1236.88	378.18	-3.271	0.001093
d.2016	-1100.27	373.64	-2.945	0.003273
sigma link function: log				
(Intercept)	7.86251	0.07243	108.6	< 2e-16
nu link function: log				
(Intercept)	9.2921	0.03035	306.1	< 2e-16

Final Remarks

Regarding the dependence calculation

- Dependence was assessed using Kendall's and Spearman's correlation coefficients.
- Significant dependence appears both within each group and between groups (cross-group effects).

Regarding the GAMLSS Modelling

- For the median of severity, we used the Exponential Gaussian family.
- For the Total Portfolio and for the Housing group, the frequency of claims has a negative impact on the median/mean of severity.
- Out of the 15,665 claims in our data, 14,699 are from Group 2, representing 93.83%.
- This pattern differs from the results reported in Garrido et al. (2016).
- For the Liability group, we detected the opposite effect.
- We discuss and conjecture possible reasons for this behaviour.

Pour aller au-delà des coefficients de dépendance

- On vient donc d'illustrer une situation où les variables Fréquence et Sinistre ne sont pas indépendantes.
- Pourtant cette hypothèse d'indépendance est souvent faite.
- Parfois cette dépendance est plus fine que la corrélation ou le coefficient de Spearman.
- Elle peut même échapper à des modèles plus ou moins classiques (GLM, GAM, GAMLSS, GMM, GLLM).

Pour aller au-delà des coefficients de dépendance : les copules

- Une copule est une fonction qui relie une loi jointe à ses lois marginales. Si C est la copule associée à un couple de variables aléatoires (X,Y) on a

$$F_{X,Y}(x, y) = C(F_X(x), F_Y(y))$$

où les F désignent les fonctions de répartition.

Pourquoi utiliser des copules ?

- Modéliser la dépendance indépendamment des marginales
- Capturer des dépendances non linéaires
- Prendre en compte la dépendance extrême (risques catastrophiques)
- Améliorer la tarification et la segmentation
- Simuler des sinistres multidimensionnels

La transformation copule

- Souvent au lieu de travailler sur X et Y , on se ramène à

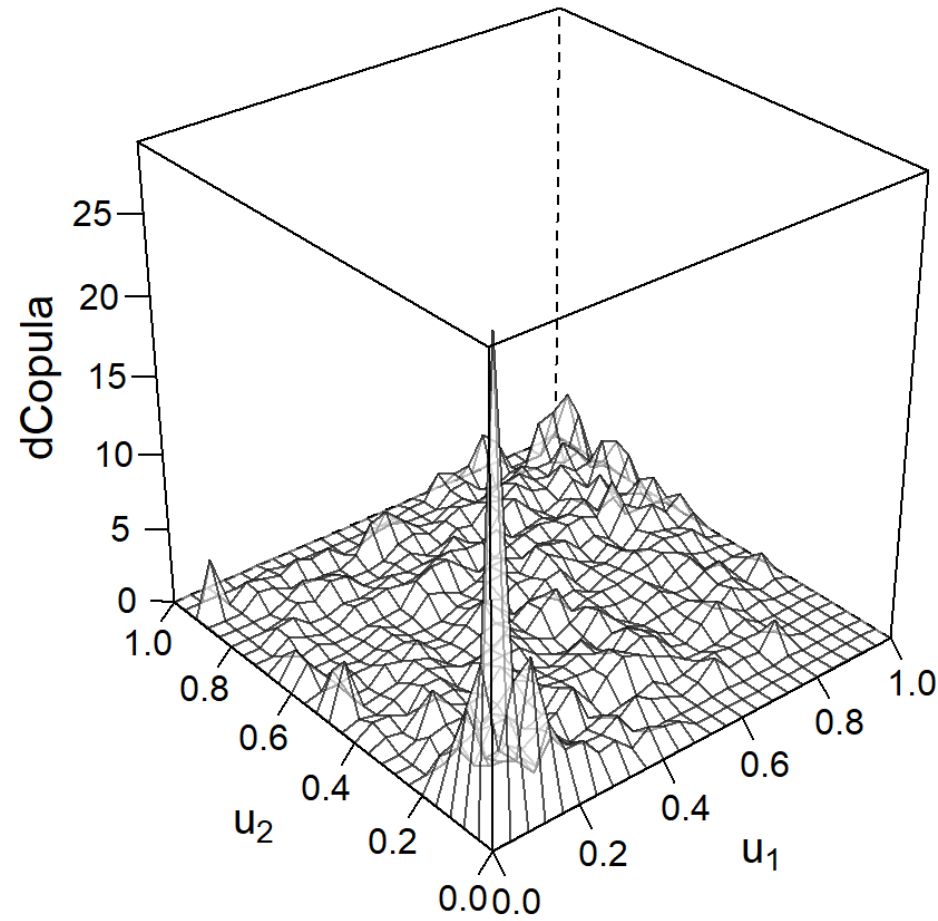
$$U = F_X(X) \text{ et } V = F_Y(Y)$$

Les couples (X, Y) et (U, V) ont même copule.

On va l'illustrer sur notre jeu de données.

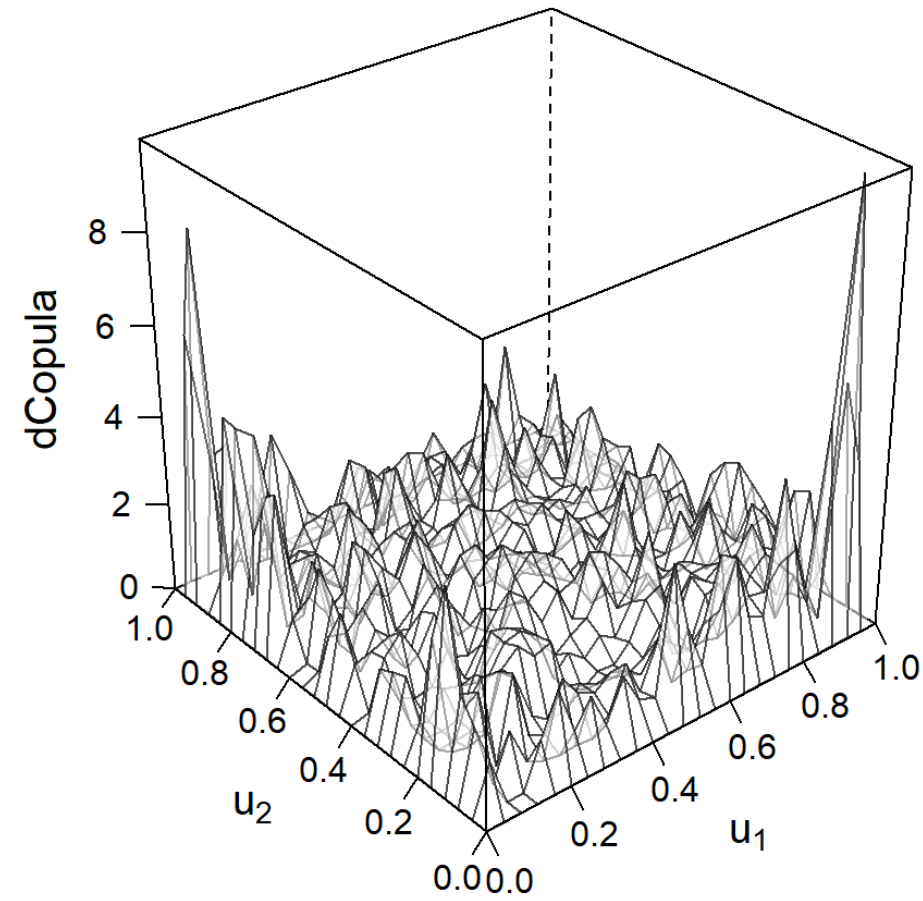
Analyse de la copule associée l'assurance responsabilité

- Avec le package COPULA, on peut déterminer la meilleure copule (au sens du critère AIC).
- Ici c'est une copule de Clayton (paramètre 2)
- Elle est caractéristique d'une dépendance plus forte dans la queue inférieure, ce qui signifie que quand la fréquence est faible, les montants ont tendance à être faibles aussi.
- Le taux de Kendall associé vaut 0.508



Analyse de la copule associée l'assurance habitation

- Ici c'est une copule Gaussienne.
- Elle est caractéristique d'une dépendance symétrique, plus faible dans les queues. Son paramètre est négatif (-0.3) ce qui signifie que quand la fréquence est élevée, le montant tend à être plus faible. Et quand la fréquence est faible, les montants tendent à être plus élevés.
- Ainsi les assurés qui déclarent **souvent** des sinistres sont typiquement des “petits sinistres”. Et les assurés qui déclarent **rarement** des sinistres présentent un risque de **gros sinistres**.
- Mais ceci devrait concerner les sinistres « raisonnable » (pas de liens extrêmes)



Clustering de copules

- Avec le package KCOP on peut faire du clustering de copules, c'ad regrouper automatiquement les portefeuilles qui ont les mêmes structures de dépendances.
- Intérêt : on identifie les portefeuilles qui peuvent réagir de la même manière) face un risque combiné.
- En effet, deux portefeuilles peuvent avoir :
- la **même fréquence**,
- la **même sévérité**,
- les **mêmes lois marginales...**
- mais **une dépendance totalement différente** entre fréquence et sévérité

Application du clustering

- On applique le clustering sur les 5 années (2015-2019) d'observations. Le résultat donne deux clusters :
 - ➔ 2016 : premier cluster
 - ➔ 2015-2017-2018-2019 : deuxième cluster
- Les deux copules associées sont les copules de Student suivantes
 - ➔ Pour 2016 : Student $(-0.25, 6.6)$
 - ➔ Pour 2015-2017-2018-2019 : Student $(-0.09, 5.3)$

Conclusion

On peut interpréter ces copules (notamment les coefficients de dépendance de queues)

- ➔ 2016 : forte dépendance (négative) globale, mais plus faible dans les queues
- ➔ 2015-2017-2018-2019 : faible dépendance qui, au contraire, s'accroît dans les queues (une fréquence plus rare va davantage correspondre à des montants plus élevés)

Pour aller plus loin

On peut considérer les copules associées à des vecteurs multidimensionnels.

On peut aussi définir des coefficients de copules.

Mais il existe d'autres mesures de dépendance...

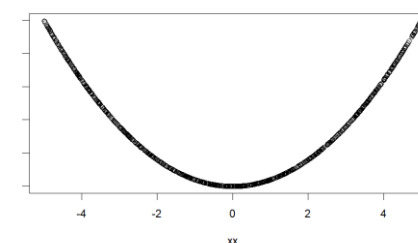
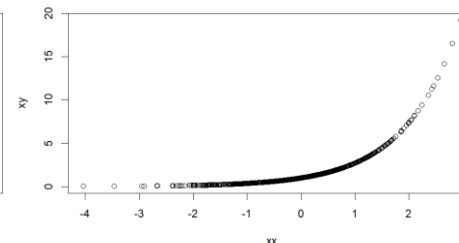
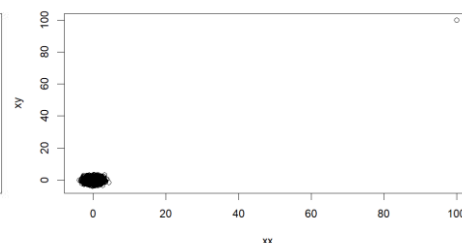
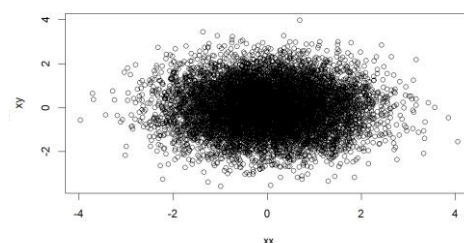
Mesures de corrélation complexe

Non-linéarité, liens multivariés, etc.

Mesures de corrélation complexe

- Mesures de corrélation traditionnelles (e.g. Pearson, Spearman ou Kendall) :
 - ➔ Echec dans la capture de lien non-linéaire
 - ➔ Impossibilité de traiter les corrélations d'ordre supérieur à 2
- Coefficient de Pearson, utilisé souvent par défaut, souffrent de plusieurs inconvénients : paramétriques, relation linéaire uniquement, sensibilité aux valeurs extrêmes
- Coefficients de Spearman ou Kendall : se limitent à retranscrire des liens monotones, entre deux variables univariées

Mesures de corrélation complexe



Corr. \ Lien	Indép.	Val. extr.	Lien exp.	Lien quad.
Pearson	0,0	0,5	0,8	0,0
<i>pvalue - Test de S</i>	0,2	0,0	0,0	0,9
Spearman	0,0	0,0	1,0	0,0
<i>pvalue - Test de S</i>	0,2	1,0	0,0	0,8
Kendall	0,0	0,0	1,0	0,0
<i>pvalue - Test de S</i>	0,2	1,0	0,0	0,6

Mesures de corrélation complexe

- La corrélation maximale de Hirschfeld-Gebelein-Rényi (HGR) résout les limites précédentes en recherchant la relation non linéaire la plus forte possible entre deux variables
- Corrélation maximale HGR $\Leftrightarrow$ maximisation du coefficient de Pearson (r) entre 2 transformations non linéaires f_U, f_V de deux variables U, V :

$$HGR(U, V) := \sup_{f_U, f_V} r(f_U(U), f_V(V))$$

Mesures de corrélation complexe

- Estimation de l'HGR reste difficile en raison de la complexité de son optimisation non linéaire
- Estimation neuronale (Grari et al. (2021)) : transformations HGR à l'aide de réseaux neuronaux afin de saisir les relations non linéaires. Minimisation de la fonction de perte suivante :

$$- \sup r \left(f_{\theta_u}(u), f_{\theta_v}(v) \right)$$

- Estimation neuronale de l' $HGR(u, v)$:

$$\widehat{HGR}_{\Theta}(u, v) = r \left(f_{\theta_u}^*(u), f_{\theta_v}^*(v) \right)$$

$$\text{avec } (f_{\theta_u}^*, f_{\theta_v}^*) = \operatorname{argmax}_{f_{\theta_u}, f_{\theta_v}} r \left(f_{\theta_u}(u), f_{\theta_v}(v) \right)$$

Mesures de corrélation complexe

- Estimation neuronale de l'HGR reste sensible aux valeurs extrêmes, limités à l'estimation de 2 variables (univariées ou multivariées), retranscrit un lien linéaire
- Extension du concept de l'HGR au Spearman – HGR (SHGR) en travaillant sur les transformations copules (F.d.R)

$$SHGR(U, V) := \sup_{f_U, f_V} r \left(F_{f_U(U)}(f_U(U)), F_{f_V(V)}(f_V(V)) \right)$$

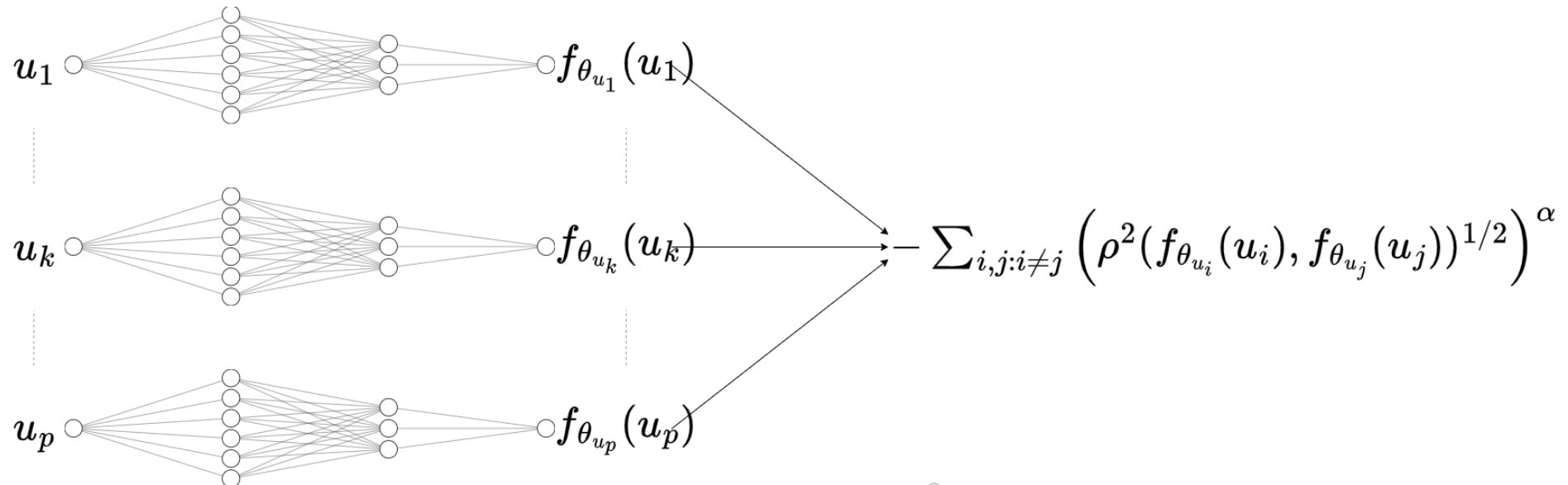
Mesures de corrélation complexe

- Estimation neuronale du SHGR, basée sur les rangs:

$$\widehat{SHGR}_{\Theta}(u, v) = r \left(\widehat{F}_{f_{\theta_u}^*} \left(f_{\theta_u}^*(u) \right), \widehat{F}_{f_{\theta_v}^*} \left(f_{\theta_v}^*(v) \right) \right)$$

avec $(f_{\theta_u}^*, f_{\theta_v}^*) = \operatorname{argmax}_{f_{\theta_u}, f_{\theta_v}} r \left(\widehat{F}_{f_{\theta_u}} \left(f_{\theta_u}(u) \right), \widehat{F}_{f_{\theta_v}} \left(f_{\theta_v}(v) \right) \right)$

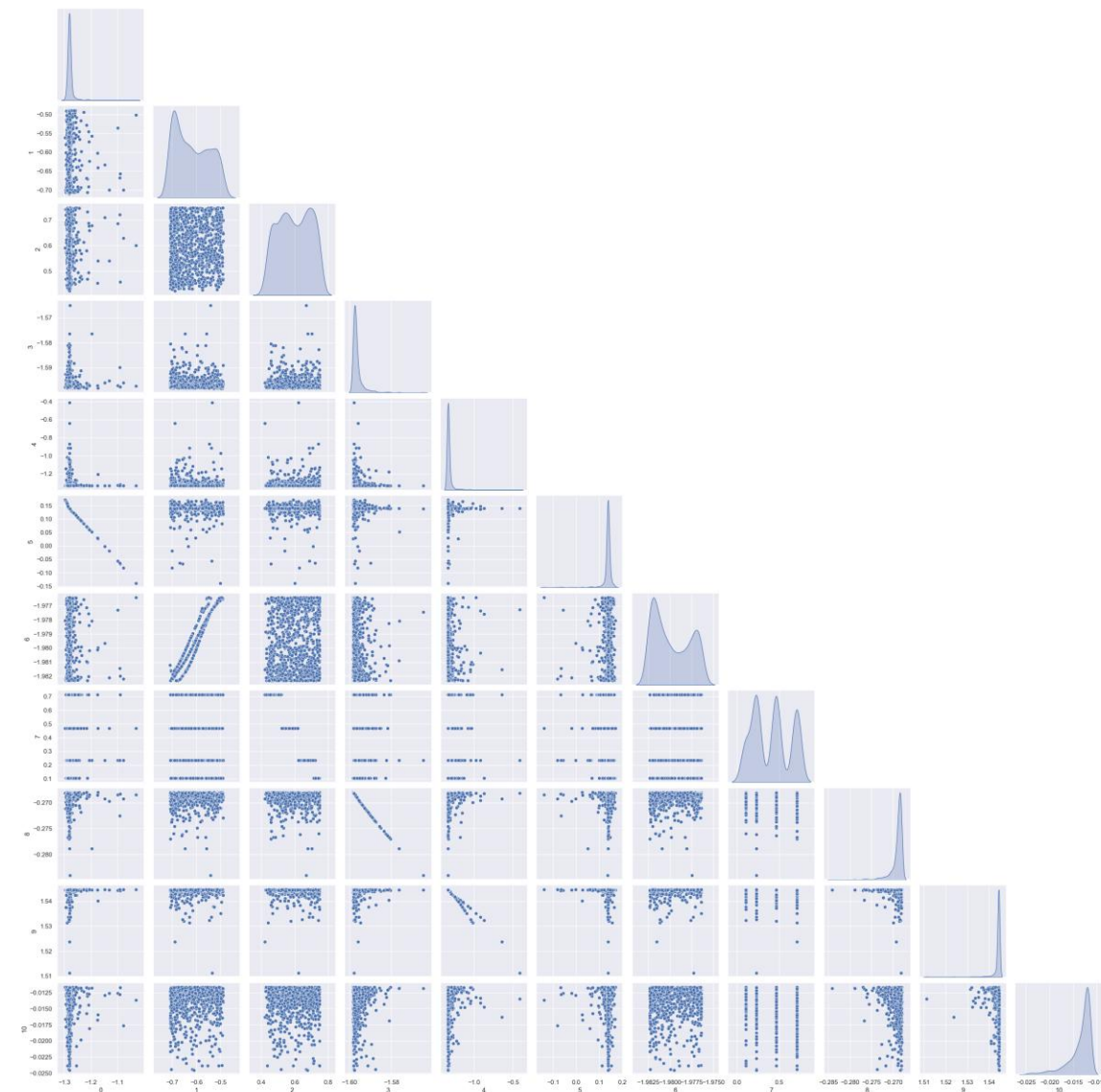
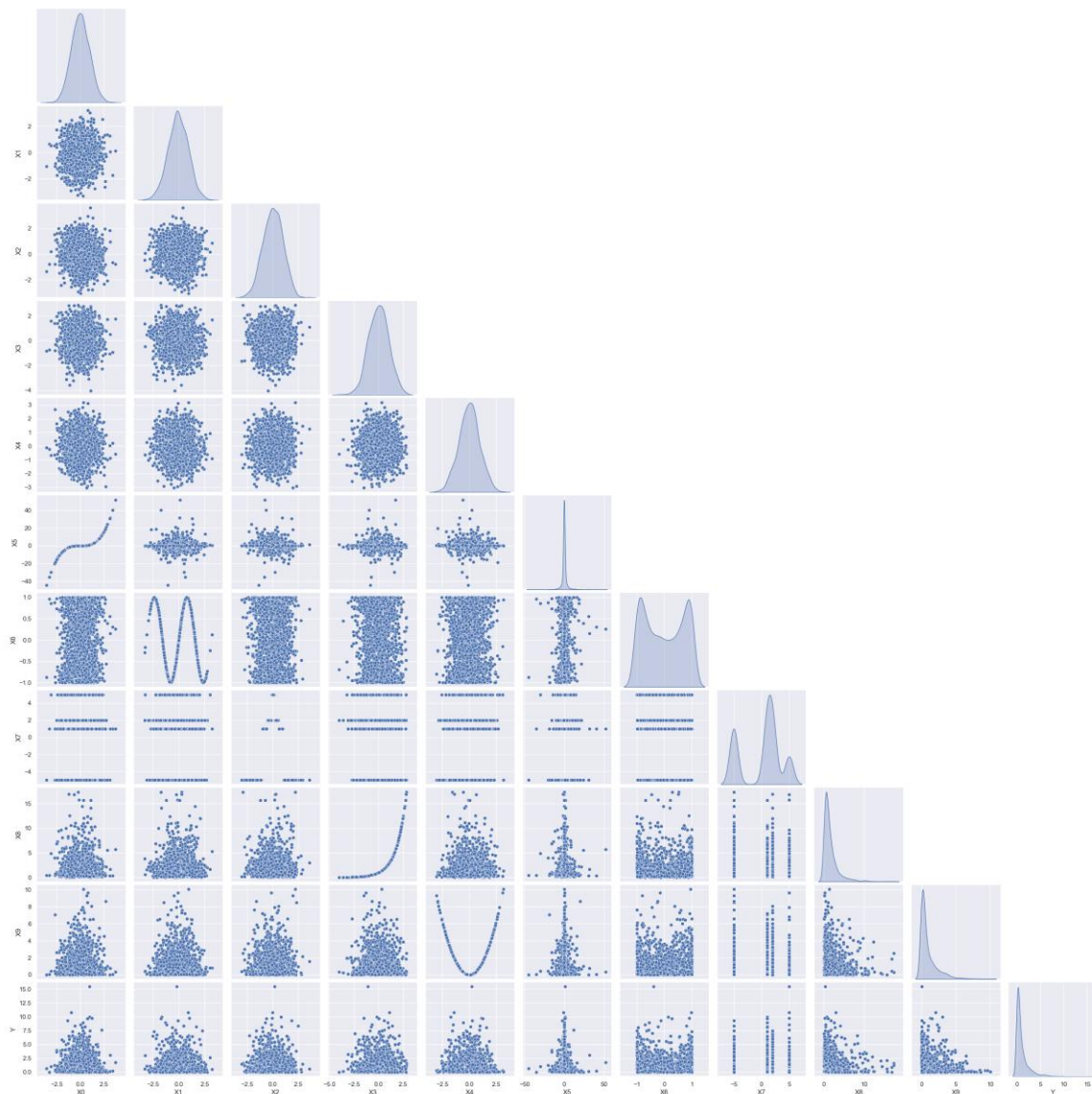
- Stacked Cross-Encoder Architecture : estimation simultanée



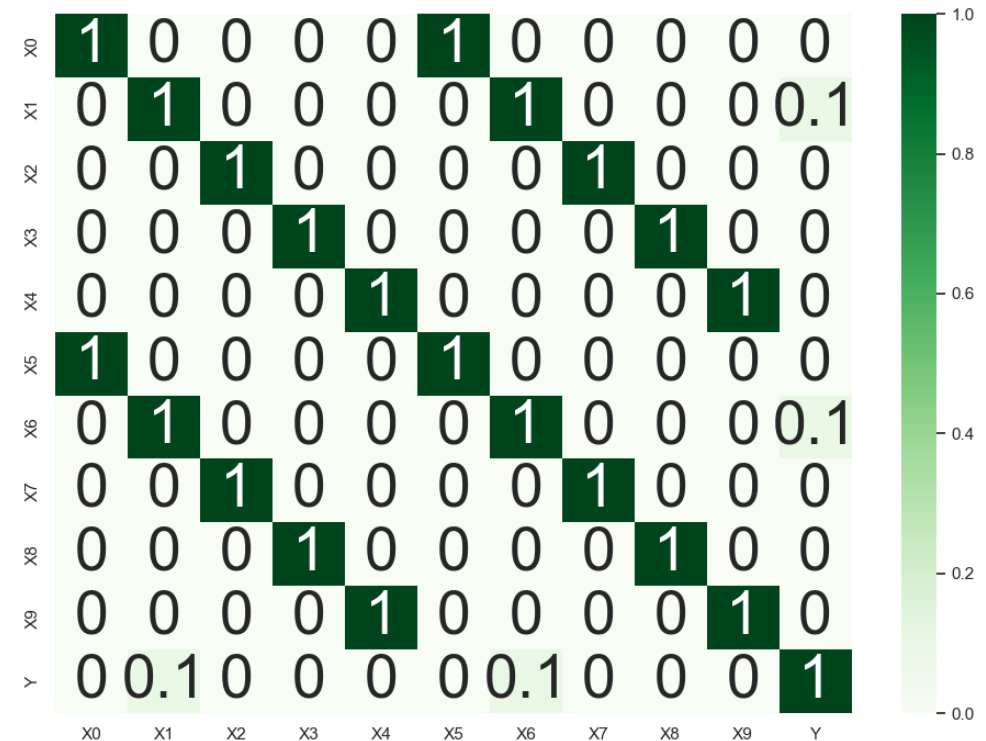
Mesures de corrélation complexe

- Avantages du SHGR :
 - ✓ Performant: détection de relations non-linéaire complexes
 - ✓ Robuste au bruit
 - ✓ Robuste aux hallucinations (en situation d'indépendance)
 - ✓ Robuste aux valeurs extrêmes (et aberrantes)
 - ✓ Temps de calcul
 - ✓ Test de significativité
 - ✓ Propriétés de Rényi
- Surpasse les compétiteurs classiques: ACE, RDC, MIC, etc.
- Permet d'améliorer les compétiteurs
- Accès aux transformations (visualisation graphique)

Mesures de corrélation complexe



Mesures de corrélation complexe





Merci !



alcoforado.renata@ufpe.br | *denys.pommeret@univ-amu.fr* | *samuel.stocksieker@univ-amu.fr*

- *Special thanks* to the anonymous insurer that provided the data.
- This work benefited from the support of the Research Chair ACTIONS under the aegis of the Risk Foundation, a joint initiative by BNP Paribas Cardif and by the Institut des Actuares and chair DIALog under the aegis of the Risk Foundation and CNP Assurances



Évaluez cet atelier

