Modélisation probabiliste n-d et conditionnement probabiliste

G. Perrin

guillaume.perrin@univ-eiffel.fr

Année 2022-2023

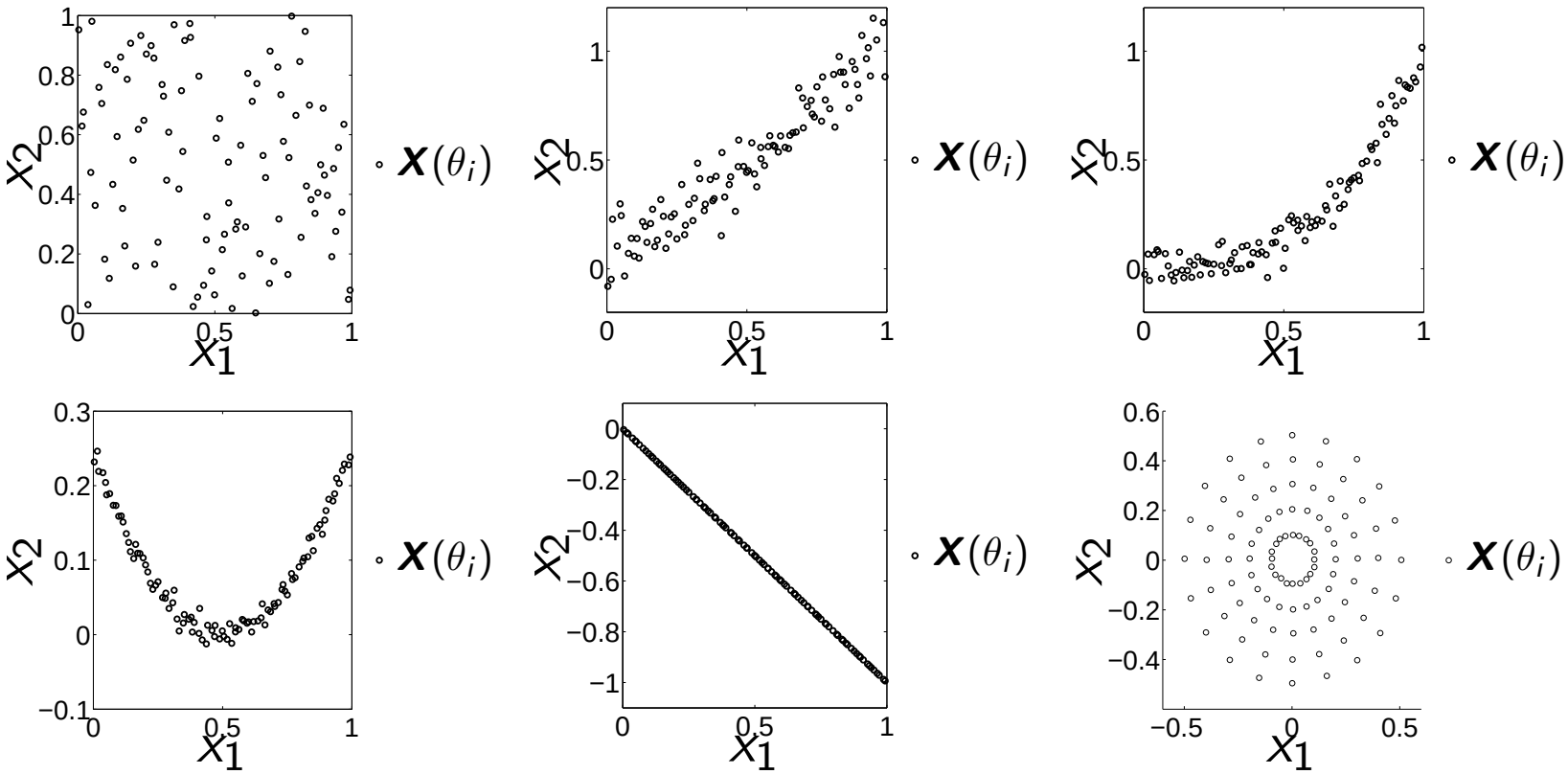





Plan de la séance

Partie 1 : introduction

- Précédemment, nous avons vu comment caractériser une **unique** variable aléatoire X (PDF, CDF, moments statistiques, quantiles...).
- Si $X_1, \ldots, X_d$ forment d v.a., alors on appelle $\boldsymbol{X} = (X_1, \ldots, X_d)$ **vecteur aléatoire**, aux dépendances potentiellement diverses.



Partie 1 : introduction

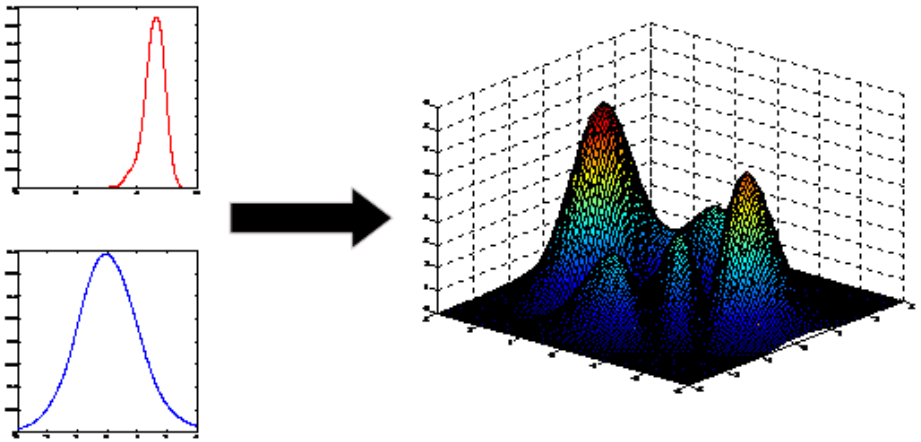
Exercice : corrélation, dépendance et causalité.

1. peut-on dire que deux variables **dépendantes** sont **corrélées** ?
2. peut-on dire que deux variables **corrélées** sont **dépendantes** ?
3. peut-on dire que la **corrélation** implique la **causalité** ?

Partie 1 : introduction

De manière plus générale, pour introduire les dépendances :

1. **Identification des dépendances**, à partir de tests expérimentaux (linéaires, monotones, fréquentielles, temporelles...)
2. **Modélisation des dépendances**, à travers la notion de **copule**.



Partie 1 : introduction

- La prise en compte des dépendances est **difficile** mais **primordiale** !

Plan de la séance

Partie 2 : modélisation probabiliste n-d

Comme pour les v.a., un **vecteur aléatoire** $\boldsymbol{X} = (X_1, \ldots, X_d)$ est caractérisé par sa fonction de répartition multidimensionnelle, $F_{\boldsymbol{X}}$, telle que :

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

Par construction, toute CDF vérifie les propriétés suivantes :

- $F_{\boldsymbol{X}}$ est monotone et non décroissante par rapport à toutes ses variables,
- $F_{\boldsymbol{X}}$ est continue à droite par rapport à tous ses variables,
- $0 \leq F_{\boldsymbol{X}}(\boldsymbol{x}) \leq 1$,
- $\lim_{x_1,\ldots,x_d \to +\infty} F_{\boldsymbol{X}}(\boldsymbol{x}) = 1$, $\lim_{x_i \to -\infty} F_{\boldsymbol{X}}(\boldsymbol{x}) = 0$, $1 \leq i \leq d$.

Partie 2 : modélisation probabiliste n-d

Si elle existe, la PDF multidimensionnelle $f_{\boldsymbol{X}}$ de $\boldsymbol{X}$ (ou densité jointe) est par ailleurs définie par :

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{D}^d) = \int_{\mathcal{D}^d} f_X(d\boldsymbol{x})d\boldsymbol{x},$$

où $\mathcal{D}^d$ est n'importe quel sous espace de $\mathbb{R}^d$ sur lequel $f_{\boldsymbol{X}}$ est bien définie. Par construction, on peut vérifier que :

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\partial^d F_{\boldsymbol{X}}}{\partial x_1 \cdots \partial x_d}(\boldsymbol{x}).$$

On nomme par ailleurs i^e marginale de $\boldsymbol{X}$, la fonction f_{X_i} telle que :

$$f_{X_i}(x_i) = \int_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d} f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x}.$$

Partie 2 : modélisation probabiliste n-d

Le paramétrage de la CDF multidimensionnelle de $\boldsymbol{X} = (X_1, \ldots, X_d)$ s'effectue généralement en trois temps :

- paramétrage des d CDF unidimensionnelles F_{X_i} de X_i,
- introduction du vecteur $\boldsymbol{U} = (U_1, \ldots, U_d) = (F_{X_1}(X_1), \ldots, F_{X_d}(X_d))$,
- paramétrage de la relation de dépendance entre les composantes de $\boldsymbol{U}$ dans l'hypercube $[0,1]^d$ (objectif de cette séance), à travers l'introduction d'une fonction copule C, telle que :

$$C(u_1, \ldots, u_d) = \mathbb{P}(U_1 \leq u_1, \ldots, U_d \leq u_d).$$

Exercices :

1. Vérifier que les composantes de $\boldsymbol{U}$ sont uniformément distribuées sur $[0,1]$.
2. Calculer $F_{\boldsymbol{X}}(\boldsymbol{x})$ en fonction de C et F_{X_i}.

Partie 2 : modélisation probabiliste n-d

D'un point de vue formel, on dira que la fonction C de $[0,1]^d$ dans $[0,1]$ est un copule ssi :

- $C(\boldsymbol{u}) = 0$ si $\prod_{i=1}^{d} u_i = 0$ (la fonction s'annule si l'une de ses composantes est nulle),
- $C(1,\ldots,1,u,1,\ldots,1) = u$,
- C est d-non-décroissante.

Pour $d = 2$, cela se traduit par :

- $C(u,0) = C(0,u) = 0$,
- $C(u,1) = C(1,u) = u$,
- $C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0$ pour tout $0 \leq u_1 \leq u_2 \leq 1$ et $0 \leq v_1 \leq v_2 \leq 1$.

Partie 2 : modélisation probabiliste n-d

Théorème de Sklar

- Toute fonction de répartition $F_{\boldsymbol{X}}$ peut s'exprimer à partir de ses marginales F_{X_i} et d'un copule C, tel que :

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = C(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)).$$

- Si les marginales F_{X_i} sont continues, la fonction copule est **unique**.

La réciproque est également vraie : si C est un copule, et F_{X_i} définissent des CDF unidimensionnelles, alors $C(F_{X_1}(x_1), \ldots, F_{X_d}(x_d))$ caractérise une fonction de répartition.

Partie 2 : modélisation probabiliste n-d

Théorème de Fréchet-Hoeffding

Pour tout copule C et tout $(u_1, \ldots, u_d) \in [0,1]^d$,

$$W(u_1, \ldots, u_d) \leq C(u_1, \ldots, u_d) \leq M(u_1, \ldots, u_d),$$

$$W(u_1, \ldots, u_d) = \max\left\{1 - d + \sum_{i=1}^{d} u_i, 0\right\}, \quad M(u_1, \ldots, u_d) = min\left\{u_1, \ldots, u_d\right\}.$$

Exercices :

1. Montrer que M est un copule. Interpréter la relation de dépendance entre les grandeurs.
2. Montrer que si $d = 2$, W est également un copule. Pour $d > 2$, on peut seulement affirmer qu'il existe un copule $\widehat{C}$ (pouvant varier) tel que $W(\boldsymbol{u}) = \widehat{C}(\boldsymbol{u})$.

Partie 2 : modélisation probabiliste n-d

Copule indépendant

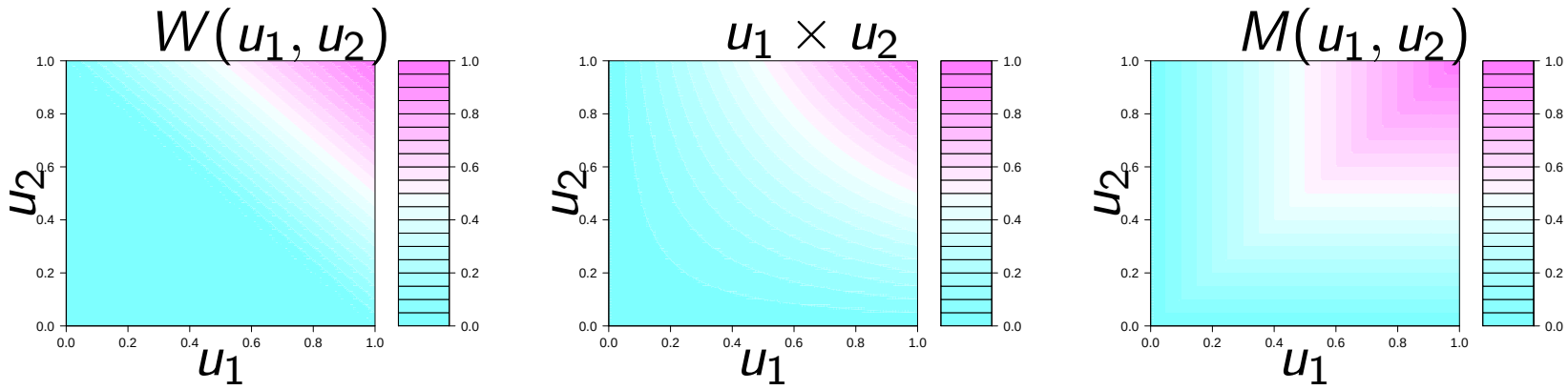
Les composantes de $\boldsymbol{X}$ sont indépendantes ssi $C(\boldsymbol{u}) = \prod_{i=1}^{d} u_i$

Démonstration :

- Si les composantes de $\boldsymbol{X}$ sont indépendantes, alors $\mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d) = \prod_{i=1}^{d} \mathbb{P}(X_i \leq x_i)$, si bien que $C(\boldsymbol{u}) = \prod_{i=1}^{d} u_i$.
- Réciproquement, si $C(\boldsymbol{u}) = \prod_{i=1}^{d} u_i$, alors $\mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d) = \prod_{i=1}^{d} \mathbb{P}(X_i \leq x_i)$ et on en déduit que les composantes de $\boldsymbol{X}$ sont indépendantes.

Partie 2 : modélisation probabiliste n-d

Copule indépendant et bornes de Fréchet-Hoeffding :



Commenter les différences entre les trois copules 2D représentés.

Partie 2 : modélisation probabiliste n-d

Copule gaussien

Si $\boldsymbol{X}$ est un vecteur aléatoire gaussien de moyenne $\boldsymbol{\mu}$ et de matrice de covariance $[R]$, alors son copule est défini par :

$$C(u_1, \ldots, u_d) = \Phi\left(\phi^{-1}(u_1), \ldots, \phi^{-1}(u_d)\right),$$

$$\phi(x) = \int_{-\infty}^{x_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy,$$

$$\Phi(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det([R])}} \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T [R]^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right).$$

Partie 2 : modélisation probabiliste n-d

Copules archimédiens - caractéristiques

- Les copules archimédiens forment une classe de copules.
- (+) La plupart des copules archimédiens présentent une expression explicite (ce qui n'est pas le cas pour le copule gaussien).
- (+) La popularité de ces copules provient du fait qu'ils permettent de modéliser la dépendance entre les composantes de $\boldsymbol{X}$ pour n'importe quelle valeur de d, à partir d'un unique paramètre, nommé θ.
- (-) Pour ce type de copules, les dépendances entre composantes de $\boldsymbol{X}$ présentent les mêmes structures.

Partie 2 : modélisation probabiliste n-d

Copules archimédiens - définition

La fonction C est un copule archimédien si elle peut s'écrire sous la forme $C(x_1, \ldots, x_d; \theta) = \psi^{[-1]}\left(\psi(x_1; \theta) + \ldots + \psi(x_d; \theta); \theta\right)$, où ψ est une fonction positive, continue, convexe, strictement décroissante sur $[0, 1]$, telle que $\psi(1; \theta) = 0$ et telle que sa fonction inverse, ψ^{-1}, est $d-$monotone.

- ψ est appelée fonction **génératrice**.
- $\psi^{[-1]}$ est appelée pseudo-inverse de ψ, et vérifie :

$$\psi^{[-1]}(x; \theta) \begin{cases} = \psi^{-1}(x; \theta) \text{ si } 0 \leq x \leq \psi(0; \theta) \ , \\ = 0 \text{ sinon.} \end{cases}$$

Partie 2 : modélisation probabiliste n-d

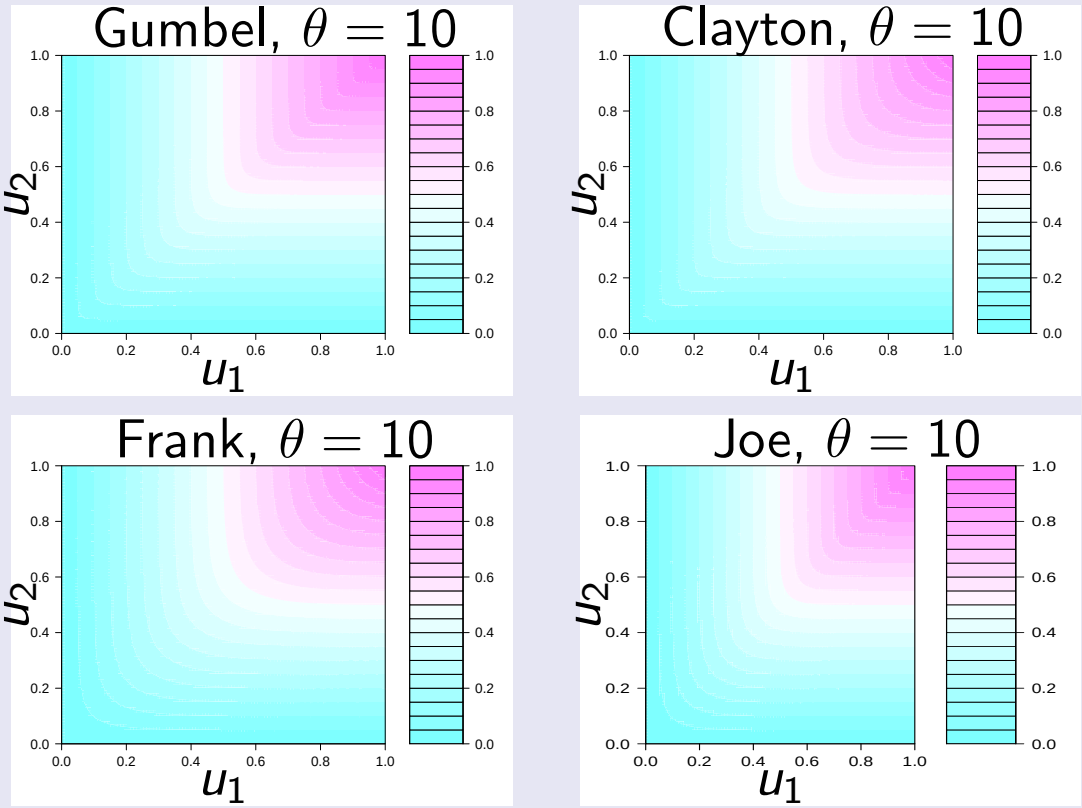
Copules archimédiens - exemples

- **Gumbel** : $\psi(x;\theta) = (-\log(x))^\theta$, $\psi^{-1}(x;\theta) = \exp(-x^{1/\theta})$, $\theta \in [1, +\infty[$.
- **Clayton** : $\psi(x;\theta) = \frac{1}{\theta}(x^{-\theta} - 1)$, $\psi^{-1}(x;\theta) = (1+\theta x)^{-1/\theta}$, $\theta \in [-1, +\infty[\setminus \{0\}$.
- **Frank** : $\psi(x;\theta) = -\log\left(\frac{\exp(-\theta x)-1}{\exp(-\theta)-1}\right)$, $\psi^{-1}(x;\theta) = -\frac{1}{\theta}\log(1+\exp(-x)(\exp(-\theta)-1))$, $\theta \in \mathbb{R} \setminus \{0\}$.
- **Joe** : $\psi(x;\theta) = -\log\left(1-(1-x)^\theta\right)$, $\psi^{-1}(x;\theta) = 1-(1-\exp(-x))^{1/\theta}$, $\theta \in [1, +\infty[$.

Exercice : montrer que le copule indépendent est un copule archimédien. Calculer sa fonction génératrice et son pseudo inverse.

Partie 2 : modélisation probabiliste n-d

Copules archimédiens - exemples



Commenter les différences entre les quatre copules 2D représentés.

Partie 2 : modélisation probabiliste n-d

Copules empiriques

- Supposons que l'on dispose de N réalisations de $\boldsymbol{X}$, $\left\{\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}\right\}$, dont les CDF marginales sont connues et écrites F_i.
- On peut alors définir :

$$U_j^{(i)} := F_j(X_j^{(i)}) \approx \widehat{U}_j^{(i)} := \frac{1}{N} \sum_{n=1}^{N} 1_{X_j^{(n)} \leq X_j^{(i)}}.$$

- On peut alors définir l'approximation empirique du copule C par :

$$C(u_1, \ldots, u_d) \approx \widehat{C}(u_1, \ldots, u_d) := \frac{1}{N} \sum_{n=1}^{N} 1_{\widehat{U}_1^{(n)} \leq u_1, \ldots, \widehat{U}_d^{(n)} \leq u_d}.$$

En pratique, il faut que N soit très grand (et que d soit petit) pour que l'approximation empirique soit pertinente...

Plan de la séance

Partie 2 : conditionnement statistique

Théorème de Bayes

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Conditionnement par une variable aléatoire

Si X et Y sont deux variables aléatoires, alors $(Y|X = x)$ est également une variable aléatoire. Notons alors f_X, f_Y et $f_{Y|X=x}$ leurs PDFs, ainsi que $f_{(X,Y)}$ la loi jointe de (X, Y) (on se limite au cas où ces fonctions existent). On déduit alors :

$$f_{Y|X=x}(y) = \begin{cases} 0 \text{ si } f_X(x) = 0, \\ \dfrac{f_{(X,Y)}(x,y)}{f_X(x)} \text{ sinon.} \end{cases}$$

Attention : $Y|X$ désigne l'application $x \mapsto Y|X = x$. Ainsi, $\mathbb{E}[Y|X]$ correspond à l'application $x \mapsto \mathbb{E}[Y|X = x]$ ($\leftrightarrow$ **espérance conditionnelle**).

Partie 2 : conditionnement statistique

Quelques propriétés associées au conditionnement statistique

- Si X et Y sont indépendants, alors $f_{Y|X=x} = f_Y$.
- Pour toute fonction g, $\mathbb{E}[g(Y)|Y] = g(Y)$.
- Espérance totale : $\mathbb{E}[g(X,Y)] = \mathbb{E}[\mathbb{E}[g(X,Y)|X]]$.
- Variance totale : $\mathrm{Var}(g(X,Y)) = \mathbb{E}(\mathrm{Var}(g(X,Y)|X)) + \mathrm{Var}(\mathbb{E}(g(X,Y)|X))$.
- Inégalité de Jensen : si g est convexe, alors $g(X, \mathbb{E}[Y|X]) \leq \mathbb{E}[g(X,Y)|X]$.

Exercice : prouver les trois premières propriétés.

Partie 2 : conditionnement statistique

Lien entre tirage conditionné et génération de réalisations associées à un copule

- Remarquons que si X et Y sont deux v.a., et si $X^{(1)}$ est une réalisation de X, et $Y^{(1)}$ est une réalisation de $Y|X = X^{(1)}$, alors $(X^{(1)}, Y^{(1)})$ est une réalisation particulière de (X, Y).
- De manière générale, la génération de réalisations indépendantes d'un vecteur $\boldsymbol{U}$ de composantes uniformément distribuées sur $[0, 1]$ et de copule C passe par la généralisation en dimension d de cette approche basée sur des tirages conditionnés.
- Ceci justifie l'importance des copules explicites, dont les lois conditionnées sont des lois faciles à générer...

Exercice : soient X et Y deux v.a. gaussiennes centrées réduites de covariance ρ. Expliquer comment générer des réalisations de (X, Y) par tirage conditionné.

Plan de la séance