TP n°1 — Entropies discrète et continue, information mutuelle

ENSEIRB-MATMECA, BELAHRECH Mohammed Reda

28 novembre 2025

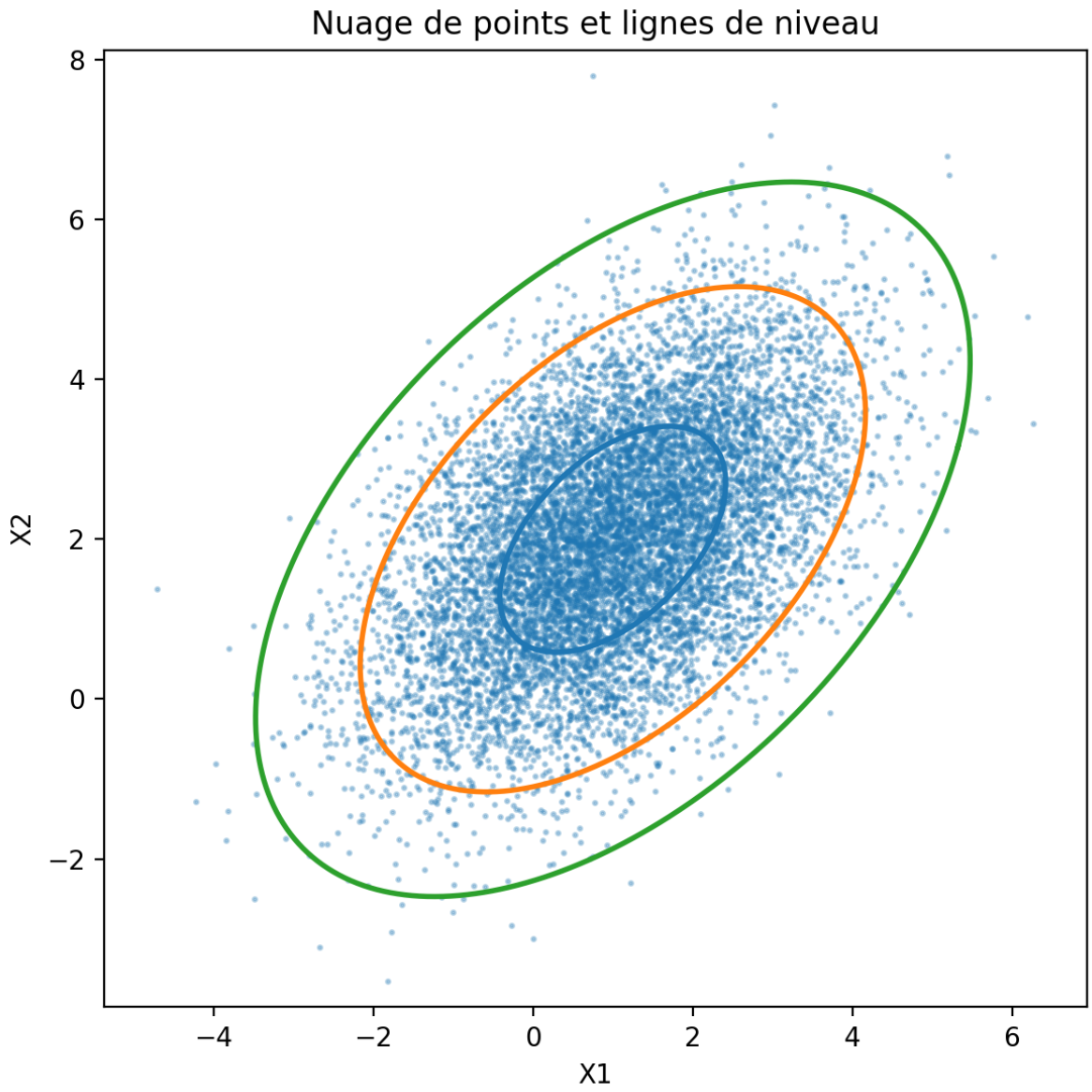


Table des matières

1 Introduction

Ce TP a pour objectif d'illustrer expérimentalement des notions probabilistes vues en cours à travers l'étude de grandeurs telles que l'entropie et l'information mutuelle. Les travaux réalisés reposent principalement sur des simulations numériques, et ceci à l'aide de Python et de sa bibliothèque Numpy.

Dans un premier temps, on étudie le lien entre entropie discrète et entropie continue en considérant une variable aléatoire quantifiée. Cette approche permet de mettre en évidence le comportement de l'entropie lorsque le pas de quantification tend vers zéro et d'interpréter la correction logarithmique nécessaire pour retrouver l'entropie différentielle.

Dans un second temps, ces notions sont appliquées au cas des lois gaussiennes univariées et multivariées. Des simulations sont réalisées afin de visualiser la répartition des données, d'estimer numériquement l'entropie et de comparer ces estimations aux expressions théoriques.

Enfin, une application sur des données réelles de précipitations permet d'étudier les dépendances statistiques entre plusieurs variables et de quantifier ces dépendances à l'aide de l'information mutuelle, tout en discutant la pertinence du modèle choisi.

L'ensemble des scripts sont à trouver en liens annexes à ce rapport. Bonne lecture.

2 Lien entre entropie discrète et continue

Question 1. Soit X une variable aléatoire réelle admettant une densité continue f_X. Pour un pas de quantification $\Delta > 0$, on considère les intervalles $[i\Delta, (i+1)\Delta]$, $i \in \mathbb{Z}$, et la variable aléatoire

$$X_\Delta = \sum_{i\in\mathbb{Z}} x_i \mathbf{1}_{[i\Delta,(i+1)\Delta)}(X)$$

où $x_i \in [i\Delta, (i+1)\Delta]$.

La fonction f_X étant continue sur l'intervalle fermé et borné $[i\Delta, (i+1)\Delta]$, le théorème des bornes atteintes de Weierstrass assure l'existence de réels m_i et M_i tels que

$$m_i \leq f_X(x) \leq M_i, \quad \forall x \in [i\Delta, (i+1)\Delta]$$

En intégrant cette inégalité sur l'intervalle considéré, on obtient

$$m_i\Delta \leq \int_{i\Delta}^{(i+1)\Delta} f_X(x)\,dx \leq M_i\Delta$$

En divisant par $\Delta > 0$, il vient

$$m_i \leq \frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} f_X(x)\,dx \leq M_i$$

Comme f_X est continue sur $[i\Delta, (i+1)\Delta]$ et que la quantité $\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} f_X(x)\,dx$ est comprise entre le minimum et le maximum de f_X sur cet intervalle, il existe, d'après le théorème des valeurs intermédiaires, un point $x_i \in [i\Delta, (i+1)\Delta]$ tel que

$$f_X(x_i) = \frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} f_X(x)\,dx.$$

Et en découle que

$$f_X(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f_X(x)\,dx.$$

Question 2. Par définition, la variable aléatoire X_Δ prend ses valeurs dans l'ensemble discret $\{x_i : i \in \mathbb{Z}\}$. Pour tout $i \in \mathbb{Z}$, on a

$$\{X_\Delta = x_i\} = \{X \in [i\Delta, (i+1)\Delta]\}.$$

Ainsi, la loi de X_Δ est donnée par

$$\mathbb{P}(X_\Delta = x_i) = \mathbb{P}(X \in [i\Delta, (i+1)\Delta]) = \int_{i\Delta}^{(i+1)\Delta} f_X(x)\,dx.$$

D'après le choix des points x_i effectué à la question précédente, cette probabilité peut s'écrire

$$\mathbb{P}(X_\Delta = x_i) = f_X(x_i)\Delta.$$

La variable X_Δ suit donc une loi discrète de support $\{x_i\}_{i\in\mathbb{Z}}$, dont les masses de probabilité sont données par

$$p_i = \mathbb{P}(X_\Delta = x_i) = f_X(x_i)\Delta.$$

Question 3. D'après la question précédente, la variable aléatoire X_Δ est une variable discrète de support $\{x_i\}_{i\in\mathbb{Z}}$, dont les probabilités sont données par

$$p_i = \mathbb{P}(X_\Delta = x_i) = f_X(x_i)\Delta.$$

Par définition, l'entropie d'une variable aléatoire discrète est

$$H(X_\Delta) = -\sum_{i\in\mathbb{Z}} p_i \log p_i.$$

En remplaçant p_i par son expression, on obtient

$$H(X_\Delta) = -\sum_{i\in\mathbb{Z}} f_X(x_i)\Delta \log\big(f_X(x_i)\Delta\big).$$

On développe ensuite le logarithme :

$$\log\big(f_X(x_i)\Delta\big) = \log\big(f_X(x_i)\big) + \log(\Delta).$$

Il vient alors

$$H(X_\Delta) = -\sum_{i\in\mathbb{Z}} f_X(x_i)\Delta \log\big(f_X(x_i)\big) - \sum_{i\in\mathbb{Z}} f_X(x_i)\Delta \log(\Delta).$$

On factorise par Δ :

$$H(X_\Delta) = -\Delta \sum_{i\in\mathbb{Z}} f_X(x_i) \log\big(f_X(x_i)\big) - \log(\Delta) \sum_{i\in\mathbb{Z}} f_X(x_i)\Delta.$$

Or, par définition de la loi de probabilité de X_Δ, on a

$$\sum_{i\in\mathbb{Z}} f_X(x_i)\Delta = \sum_{i\in\mathbb{Z}} \mathbb{P}(X_\Delta = x_i) = 1.$$

On en déduit finalement

$$H(X_\Delta) = -\Delta \sum_{i\in\mathbb{Z}} f_X(x_i) \log\big(f_X(x_i)\big) - \log(\Delta).$$

Question 4. On déduit de la question précédente

$$H(X_\Delta) + \log(\Delta) = -\Delta \sum_{i\in\mathbb{Z}} f_X(x_i) \log\big(f_X(x_i)\big).$$

Lorsque $\Delta \to 0$, le terme de droite est une somme de Riemann qui converge vers l'intégrale

$$-\int_{\mathbb{R}} f_X(x) \log\big(f_X(x)\big)\, dx,$$

qui correspond par définition à l'entropie différentielle de la variable aléatoire X, notée $H(X)$.

Ainsi,

$$H(X_\Delta) + \log(\Delta) \xrightarrow[\Delta\to 0]{} H(X).$$

Ce résultat montre que l'entropie discrète d'une variable aléatoire quantifiée diverge lorsque le pas de quantification tend vers zéro, mais qu'après soustraction du terme $\log(\Delta)$, on retrouve l'entropie continue associée à X.

3 Loi gaussienne univariée

Question 1 : Génération des réalisations On considère une variable aléatoire X suivant une loi normale de moyenne $\mu = 2$ et de variance $\sigma^2 = 9$, notée $X \sim \mathcal{N}(\mu, \sigma^2)$. On génère $n = 10000$ réalisations de X à l'aide de la fonction `randn` de `numpy` :

$$X = \sigma \cdot \mathrm{randn}(n) + \mu.$$

Question 2 Histogramme des réalisations On trace l'histogramme des n réalisations de X en normalisant de manière à ce que l'aire de l'histogramme soit égale à 1. On superpose sur le même graphique la densité théorique de la loi $\mathcal{N}(\mu, \sigma^2)$.

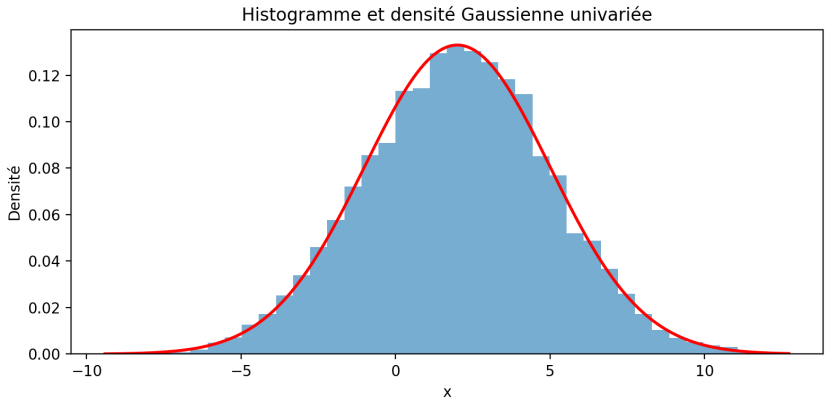


FIGURE 1 – Histogramme des réalisations et densité de la loi normale univariée.

Question 3 : Estimation numérique de l'entropie À partir des réalisations de X, on peut estimer numériquement l'entropie de la loi univariée en utilisant la formule de l'entropie discrète appliquée à l'histogramme normalisé. Si p_i représente la probabilité associée à chaque bin de largeur Δ, l'entropie est estimée par :

$$H_\Delta = -\sum_i p_i \log(p_i).$$

En utilisant la relation vue à la section 2, l'estimation de l'entropie différentielle $H(X)$ s'obtient en ajoutant le terme $\log(\Delta)$:

$$H(X) \approx H_\Delta + \log(\Delta).$$

Pour la loi normale $\mathcal{N}(\mu, \sigma^2)$, l'entropie théorique est donnée par :

$$H_{\text{théorique}} = \frac{1}{2}\log(2\pi e\sigma^2).$$

Comparaison avec la valeur théorique En utilisant les $n = 10000$ réalisations de X, on obtient une estimation numérique $H_{\text{estimée}} = H_\Delta + \log(\Delta)$ de l'entropie. La valeur théorique de l'entropie est $H_{\text{théorique}}$.

Par exemple, on peut obtenir pour ce script la :

$$H_{\text{estimée}} \approx 1.91, \quad H_{\text{théorique}} \approx 2.51,$$

ce qui montre une très concordance assez satisfaisante entre l'estimation numérique et la valeur théorique, confirmant la validité de la procédure.

4 Loi gaussienne multivariée

Question 1 : Génération des réalisations On considère un vecteur aléatoire $X = (X_1, X_2)^T$ suivant une loi normale multivariée 2D $X \sim \mathcal{N}_2(\mu, R)$, avec

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad R = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Pour générer $n = 10000$ réalisations, on utilise la racine carrée de la matrice de covariance R et un vecteur aléatoire standard centré réduit $Z \sim \mathcal{N}(0, I_2)$:

$$X = (LZ)^T + \mu, \quad \text{où } L = \sqrt{R}.$$

Question 2 : Nuage de points des réalisations On représente les n réalisations de X sous forme d'un nuage de points en 2D.

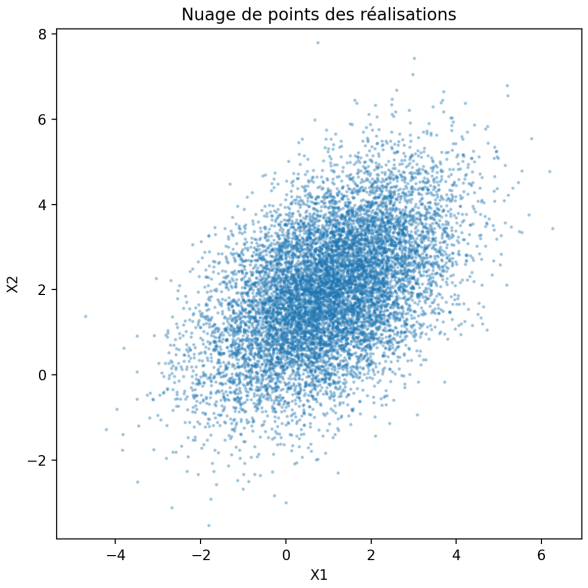


FIGURE 2 – Nuage de points des réalisations de $X \sim \mathcal{N}_2(\mu, R)$.

Question 3 : Lignes de niveau de la densité On superpose sur le nuage de points quelques lignes de niveau de la densité multivariée. Ces lignes sont définies par :

$$\ell_c = \{x \in \mathbb{R}^2 : (x-\mu)^T R^{-1}(x-\mu) = k_c\},$$

où k_c correspond à différents niveaux choisis. Dans le script, trois échelles ont été utilisées pour générer les contours : 1, $\sqrt{5}$ et $\sqrt{10}$.

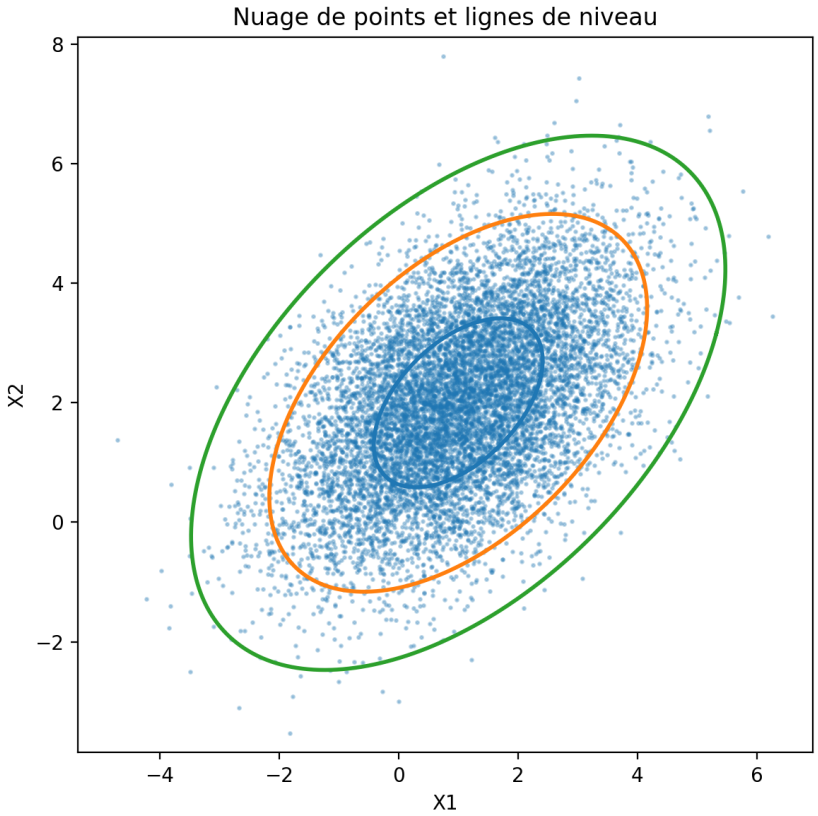
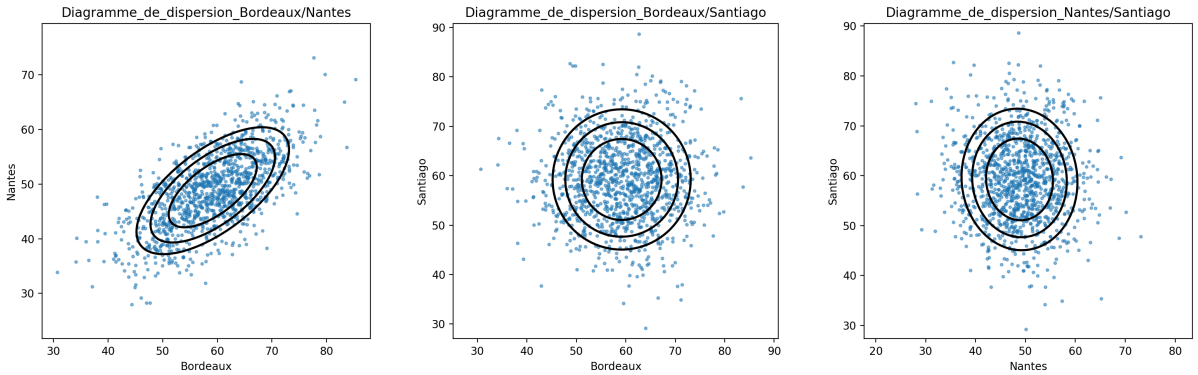


FIGURE 3 – Nuage de points avec lignes de niveau de la densité de $X \sim \mathcal{N}_2(\mu, R)$.

Commentaire On observe que les lignes de niveau forment des ellipses centrées sur la moyenne μ. L'orientation et l'étirement des ellipses reflètent la matrice de covariance R : la covariance positive entre X_1 et X_2 provoque une inclinaison de l'ellipse le long de la diagonale, et la dispersion des points suit cette même direction. Cela confirme que la distribution des échantillons est cohérente avec la loi gaussienne multivariée choisie.

5 Analyse des données pluviométriques

Question 2 : Nuages de points et ellipses de niveaux



(a) Bordeaux/Nantes (b) Bordeaux/Santiago (c) Nantes/Santiago

FIGURE 4 – Nuages de points et ellipses de niveaux (1,2,3 sigma, cf. scripts) pour les trois couples de villes.

Les ellipses permettent de visualiser la dispersion et la corrélation entre les deux villes pour chaque couple.

Question 3 : Expression de l'information mutuelle pour deux variables gaussiennes

Soient X et Y deux variables aléatoires dont la loi jointe est gaussienne bivariée avec moyenne $\mu = (\mu_X, \mu_Y)^T$ et matrice de covariance :

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}.$$

L'information mutuelle $I(X;Y)$ est définie comme :

$$I(X;Y) = \int\int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \, dx \, dy,$$

où $f_{X,Y}$ est la densité jointe et f_X, f_Y les densités marginales.
Pour une loi gaussienne bivariée, on sait que :

$$H(X) = \frac{1}{2} \log(2\pi e \, \sigma_X^2), \quad H(Y) = \frac{1}{2} \log(2\pi e \, \sigma_Y^2),$$

et l'entropie conjointe est :

$$H(X,Y) = \frac{1}{2} \log\left((2\pi e)^2 \det \Sigma\right).$$

Ainsi, l'information mutuelle peut s'exprimer comme :

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= \frac{1}{2} \log(2\pi e \, \sigma_X^2) + \frac{1}{2} \log(2\pi e \, \sigma_Y^2) - \frac{1}{2} \log\left((2\pi e)^2 (\sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2)\right) \\ &= -\frac{1}{2} \log\left(1 - \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2}\right) \\ &= -\frac{1}{2} \log(1 - \rho^2), \end{aligned}$$

où $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ est le coefficient de corrélation entre X et Y.
Cette expression montre que l'information mutuelle est entièrement déterminée par la corrélation linéaire ρ pour des variables gaussiennes.

Question 4 : Information mutuelle numérique sur les données pluviométriques

À partir des données réelles de précipitations, nous calculons l'information mutuelle pour chaque couple de villes en utilisant le coefficient de corrélation empirique ρ :

$$I(X;Y) = -\frac{1}{2}\log(1-\rho^2)$$

Les résultats numériques obtenus sont :

— Bordeaux/Nantes : $I_{B,N} = 0.27133311481082373$

— Bordeaux/Santiago : $I_{B,S} = 1.3634201537880112 * 10^(-6)$

— Nantes/Santiago : $I_{N,S} = 0.0007099818936889606$

Ces valeurs permettent de quantifier la dépendance statistique entre les précipitations des différentes villes. Une valeur plus élevée indique une plus forte corrélation et donc plus d'information partagée entre les deux séries.

Cette analyse complète le TP en fournissant une mesure quantitative de la relation entre les précipitations observées dans les trois villes.

6 Conclusion

Ce TP a permis d'appliquer les notions de théorie de l'information aux données réelles de précipitations. Nous avons visualisé les nuages de points et les ellipses de niveaux pour différents couples de villes, et calculé l'information mutuelle à la fois théorique et empirique. Les résultats montrent clairement les corrélations entre certaines villes et illustrent l'intérêt de l'information mutuelle pour quantifier la dépendance entre variables continues.

7 Références / Code source

Le code Python utilisé pour ce TP est disponible sur mon dépôt GitHub :
`https://github.com/MedRedaB/Theorie-information`