Projet M2-SSD
2022-2023

• Le projet s'effectue seul ou en binôme ou en trinôme.
• Les rapports doivent être rendus, au plutard, le 08 décembre à midi via la plateforme classroom
• Les fichiers sont déposés sous la forme d'un seul fichier **PDF**
• Les fichiers doivent être nommés : **Nom1.Nom2.Nom3**
• Les instructions proposées du logiciel R doivent être jointes au projet et mises au clair.

A. Données ciruclaires simulées

Les données réparties sur un cercle se produisent dans de nombreuses applications comme la biologie, la médecine, la géologie ou la géographie.. Il s'agit des données angulaires, puisqu'à chaque angle θ correspond un et un seul point du cercle (ici unitaire) $z = (\cos(\theta), \sin(\theta))$.

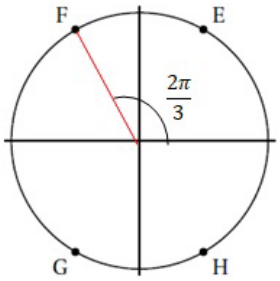

Figure 1 : 4 points sur le cercle unitaire : centré en $(0, 0)$ et de rayon 1. Le point F a pour coordonnées $(\cos(2\pi/3), \sin(2\pi/3))$, il est caractérisé par la donnée de l'angle $\theta = 2\pi/3$.

Exemples des données angulaires comprennent les directions quotidiennes du vent, les directions des courants océaniques, les directions de départ des animaux, des directions du plan de fracture osseuse ou les orientations des abeilles dans une ruche après des stimuli.

L'objectif de ce projet est d'étudier la trajectoire suivie par un animal après un stimulus. On observe cette direction aux instants $t_1, \cdots, t_n$. On suppose qu'à l'instant 0, l'animal occupe la position P_0 du plan. A l'instant t_1, cet animal occupe la position P_1. Cette position est obtenue apartir de la position initiale via la relation $P_1 = P_0\epsilon_1$ où ϵ_1 est un bruit et ainsi de suite, c'est-à-dire, sa position P_n à l'instant t_n est $P_n = P_{n-1}\epsilon_n$, P_{n-1} étant la position de l'animal à l'instant t_{n-1}. Les données permettant d'étudier la direction suivie par cet animal sont donc :

$$P_0,\ P_1, \cdots, P_n, \cdots$$

Loi uniforme sur le cercle

Une loi uniforme circulaire est une loi de probabilité sur le cercle unité dont la densité f est uniforme pour tous les angles $\theta \in [0, 2\pi]$,

$$f(z) = \frac{1}{2\pi},$$

z est un point du cercle unitaire, z est donc un point du plan de coordonnées $(\cos(\theta), \sin(\theta))$ pour $\theta \in [0, 2\pi]$.

Question (A1).

(A1) On rapelle que si U suit la loi uniforme sur l'intervalle $[0, 1]$ alors le couple aléatoire $(\cos(2\pi U), \sin(2\pi U))$ est de loi uniforme sur le cercle unitaire. Utiliser ce résultat pour simuler n points $z_1, \cdots, z_n$ uniformément distribués sur le cercle unitaire.

Loi de Cauchy sur le cercle $C^*(\varphi)$

Soit φ fixé dans $]0, 1[$. La loi de Cauchy sur le cercle, notée $C^*(\varphi)$, est une loi de densité, pour $z = (\cos(\theta), \sin(\theta))$

$$g(z) = \frac{1}{2\pi} \frac{1 - \varphi^2}{(\cos(\theta) - \varphi)^2 + \sin^2(\theta)}.$$

Question (A2). On admet le résultat suivant : si U suit la loi uniforme sur l'intervalle $[0, 1]$ alors le couple $W = (V_1, V_2)$ avec

$$V_1 = \frac{2\varphi + \cos(2\pi U) + \varphi^2 \cos(2\pi U)}{1 + \varphi^2 + 2\varphi \cos(2\pi U)}, \quad V_2 = \frac{\sin(2\pi U) - \varphi^2 \sin(2\pi U)}{1 + \varphi^2 + 2\varphi \cos(2\pi U)},$$

suit la loi $C^*(\varphi)$.

(A2) En utilisant ce résultat et en fixant une valeur de φ, simuler n points $w_1, \cdots, w_n$ issus de la loi $C^*(\varphi)$ (noter bien que ces points sont tous sur le cercle unitaire).

Simulation des données

On rappelle que si $P = (a, b)$ et $\epsilon = (c, d)$ sont deux points du cercle unitaire alors $P\epsilon$ est aussi un point du cercle unitaire de coordonnées :

$$P\epsilon = (ac - db, ad + bc).$$

On note,

— P_0 est une variable aléatoire de loi uniforme sur le cercle unitaire

— $\epsilon_1, \cdots, \epsilon_n$ sont des variables aléatoires indépendantes toutes de loi $C^*(\varphi)$

— P_0 est indépendante de $(\epsilon_1, \cdots, \epsilon_n)$.

On définit donc le modèle,

$$\begin{aligned} P_1 &= P_0\epsilon_1 \\ P_2 &= P_1\epsilon_2 \\ &.. \\ &.. \\ P_n &= P_{n-1}\epsilon_n. \end{aligned}$$

On note par,

$$\mathcal{S}_n = \{P_1, \cdots, P_n\}.$$

Questions.

(A3) Simuler cet ensemble et répresenter $\mathcal{S}_n$ pour des différentes valeurs de n.

(A4) Vérifier, par des illustrations graphiques, que $\mathcal{S}_n$ s'approche d'un ensemble déterministe $\mathcal{M}$ que l'on déterminera.

B. Distance de Haussdord et bande de confiance

On voudrait mesurer la vitesse v_n avec laquelle le nuage de points $\mathcal{S}_n$ s'approche de $\mathcal{M}$ lorsque n devient grand en utilisant la "distance de Haussdorff" entre deux ensembles : $d_H(\mathcal{S}_n, \mathcal{M})$. On rappelle que,

$$d_H(\mathcal{S}_n, \mathcal{M}) = \max\left(\max_{x\in\mathcal{M}} \min_{1\le i\le n} d(x, X_i), \max_{1\le i\le n} \min_{x\in\mathcal{M}} d(x, X_i)\right),$$

d étant la distance naturelle sur le plan.

(B1) Faire varier n, et pour chaque n donner la valeur de $d_H(\mathcal{S}_n, \mathcal{M})$, présenter les résultats sous la forme d'un tableau.
(On pourra utiliser la fonction *distFct* de *TDA package*).

(B2) Représenter sur un même graphique la courbe de $n \to d(\mathcal{S}_n, \mathcal{M})$ avec celle de $n \to v_n$. On prendra d'abord $v_n = 1/\sqrt{n}$, ensuite $v_n = (\log n/n)^{1/2}$ et finalement $v_n = 1/n$. Conclure.

(B3) Répresenter l'histogramme de la distribution de $d_H(\mathcal{S}_n, \mathcal{M})$. Que peut être la loi asymptotique de $d_H(\mathcal{S}_n, \mathcal{M})$?.

C. Données bruitées

La plupart des grandeurs sont mesurées avec des erreurs. Les modèles de convolutions sont des modèles additifs qui permettent de relier une mesure observée x à sa valeur réelle p via une erreur err :

$$x = p + err.$$

Les v.a. $(P_i)_{1\leq i\leq n}$ ne sont donc pas observées, en réalité, mais à la place on observe $X_1, \cdots, X_n$ données par le modèle, dit modèle de convolution,

$$X_i = P_i + \sigma\alpha_i, \;\; i = 1, \cdots, n.$$

avec $\alpha_1, \cdots, \alpha_n$ sont i.i.d. et α_i est un vecteur gaussien centrée 2-dimensionnels (à valeurs dans $\mathbb{R}^2$). On suppose aussi que (P_i) et (α_i) sont indépendantes.

Question (C1) :

(C1) Simuler et représenter le nuage des points $X_1, \cdots, X_n$ pour une valeur de σ que vous choisissez et pour des différentes valeurs de n (les données $P_1, \cdots, P_n$ étant celles simulées lors de la question (**A3)).**

D. Lien avec la régression non paramétrique

On note par X_i^1 et Y_i les coordonnées de X_i pour tout $i = 1, \cdots, n$

$$X_i = (X_i^1, Y_i).$$

On cherche à étudier la relation liant Y_i à X_i^1 en posant un modèle non paramétrique de la forme,

$$Y_i = r(X_i^1) + e_i,$$

(e_i) est une suite d'erreur centrées à valeurs réelles qu'on modélisera par une loi normale centrée.

Questions.

(D1) Représenter graphiquement l'estimateur à noyau de la fonction r à l'aide des observations $(X_i)_{1\leq i\leq n}$, (les points $X_i = (X_i^1, Y_i)$ étant simulées à la question (**C1**)).

(D2) Conclure en rapport avec l'ensemble $\mathcal{M}$.