Statistique pour données fonctionnelles. Exercices - Logiciel R[1]

Commencer par charger le package `fda` .

Exercice 1.

La méthode de simulation de trajectoires du mouvement brownien à partir de la décomposition de Karhunen-Loève vue en cours n'est pas la plus classique. On utilise plutôt généralement le fait que si W est un brownien indexé par $\mathbb{R}_+$, ses accroissements sont indépendants et stationnaires, de loi gaussienne : si on se donne $\{t_0 = 0 < t_1 < t_2 < \cdots < t_p\}$, les variables aléatoires réelles $W(t_1), W(t_2) - W(t_1), \ldots, W(t_p) - W(t_{p-1})$ sont indépendantes, et $W(t_j) - W(t_{j-1}) \sim \mathcal{N}(0, t_j - t_{j-1})$. Utiliser cette caractérisation pour écrire une fonction R simulant n trajectoires du mouvement brownien restreint à $[0, 1]$, et tracer les réalisations obtenues.

Exercice 2.

Écrire, pour chacune des courbes suivantes, une fonction R permettant de simuler un échantillon de n courbes aléatoires discrétisées en p points :

1. $X(t) = \xi_0 \cos(2\pi t) + \xi_1 \sin(4\pi t) + \xi_3 (t - 0.5)(t - 0.25)$ avec $\xi_1, \xi_2, \xi_3 \sim \mathcal{U}_{[0,1]}$, indépendantes.
2. $X(t) = \xi_0 + \xi_1 t + \xi_2 \exp(t) + \sin(\xi_3 t)$ avec $\xi_1 \sim \mathcal{U}_{[0,100]}$, $\xi_2 \sim \mathcal{U}_{[-30,30]}$, $\xi_3 \sim \mathcal{U}_{[0,10]}$, et $\xi_4 \sim \mathcal{U}_{[1,3]}$ indépendantes.

Pour chacune des fonctions, illustrer le résultat obtenu.

Exercice 3.

1. Construire une base de splines cubiques avec 21 noeuds équirépartis dans l'intervalle $[0, 2]$. Faire différents essais pour les tracer toutes, ou seulement certaines, et tracer également pour certaines fonctions, leurs dérivées jusqu'à l'ordre 3.
2. Construire une courbe aléatoire X exprimée comme combinaison linéaire de la base précédente, avec pour coefficients des variables indépendantes de même loi gaussienne centrée réduite. Tracer l'objet obtenu, ainsi que sa dérivée.
3. Construire maintenant un échantillon de 10 courbes comme la précédente, simulées à partir de variables indépendantes. Les représenter d'abord toutes sur le même graphique, puis représenter sur un autre graphique les 2 premières courbes ainsi que leur somme.

Exercice 4.

On considère les données `growth` du package `fda`. Proposer un code R permettant d'effectuer le lissage par moindres carrés des données de croissance des filles, dans une base de 12 B-splines d'ordre 6 sur l'intervalle $[1, 18]$. Tracer sur le même graphique les données brutes et la courbe lissée pour un individu de l'échantillon.

1. Enseignant : G. Chagny, `gaelle.chagny@univ-rouen.fr`.

Exercice 5.

1. Générer un vecteur t de 101 points équirépartis entre 0 et 2. Définir, pour tout point de t, les valeurs de la fonction x suivante

$$x(t) = \begin{cases} (t-0.5)^2 \text{ pour } t \leq 1, \\ 0.25(t-1) \text{ pour } t > 1 \end{cases}$$

   Tracer le résultat obtenu.
2. Créer une base de B-splines quadratiques (ordre 3) avec points de rupture 0, 0.5,1, 1.5 et 2, pour effectuer le lissage par moindres carrés de ces observations. Comment prendre en compte la discontinuité en $t = 1$?
3. À l'aide de la fonction `smooth.basis`, créer un objet fonctionnel `ylisse` avec les données de la question 1. et la base créée à la question 2. Améliorer les résultats obtenus en enlevant l'observation correspondant à $t = 1$.
4. Évaluer, à l'aide de la fonction `eval.fd` l'objet fonctionnel `ylisse` en 200 points pour créer un vecteur `yeval`. Comparer les résultats donnés par les calculs suivants :
   (i) `2*ylisse+4` et `2*yeval+4` ;
   (ii) `ylisse2` et `yeval2` ;
   (iii) `sqrt(ylisse)` et `sqrt(yeval)` ;

Exercice 6.

On considère les données `melanoma` du package `fda`.

1. Pour i allant de 1 à 35, créer une base de Fourier à i fonctions de base sur l'intervalle de temps des données, effectuer le lissage des données à l'aide de la fonction `smooth.basis`, et récupérer la valeur du critère gcv correspondante.
2. Choisir le nombre de fonctions de base minimisant le critère gcv, et superposer sur un même graphique le lissage correspondant et les données brutes.

Exercice 7. Étude des données `medfly`. Il s'agit de données concernant le nombre d'oeufs pondus par 50 mouches méditerranéennes des fruits pendant 25 jours.

1. Charger le fichier de données `medfly.Rdata` sur l'espace *MyCourse* du cours, représenter les données `medfly$eggcount`.
2. Proposer un lissage des données par moindres carrés pénalisés dans une base de splines cubiques sur $[0, 25]$, avec noeuds en chaque entier. On ajustera le paramètre λ de la pénalité usuelle par validation croisée généralisée, en choisissant dans l'ensemble $\{e^{-10}, e^{-9}, \ldots, e^{9}, e^{10}\}$.
3. Représenter les fonctions lissées, superposer sur le même graphique la fonction moyenne empirique.
4. Calculer et représenter la covariance empirique des données.
5. Effectuer l'ACP de ces données. Combien de fonctions propres sont nécessaires pour expliquer 90% de la variabilité ? Les représenter.

Exercice 8. Étude des données `pinch` du package `fda`. Il s'agit de 20 enregistrements de l'évolution au cours d'une manipulation de pincement de la force exercée entre le pouce et l'index d'un sujet.

1. Représenter les données `pinchraw` et `pinch`. Quelle est la différence entre les deux ?
2. Proposer un lissage des données par moindres carrés pénalisés dans une base de splines cubiques sur $[0, 150]$, avec noeuds en chaque entier. On ajustera le paramètre λ de la pénalité usuelle par validation croisée généralisée, en choisissant dans l'ensemble $\{e^{-10}, e^{-9}, \ldots, e^{9}, e^{10}\}$.
3. Effectuer l'ACP de ces données. Combien de fonctions propres sont nécessaire pour expliquer 90% de la variabilité ? Les représenter.