Stat Non Paramétrique - M2 MMA - IMB - Université Paris Cité Mini Projet noté - Octobre 2022
Anne Sabourin anne.sabourin@u-paris.fr,

Mini-Projet : Estimation de densité par noyau

Consignes

Ce projet est à effectuer en binôme ou individuellement.

L'ensemble de votre travail devra être rassemblé dans un seul dossier portant votre nom de famille, et archivé au format .zip (de type `Dupont.zip`). Si vous travaillez en binôme, un seul dossier doit être rendu, le nom du dossier doit contenir les deux noms de famille (`Dupont_Tartempion.zip`). Ce dossier doit contenir

- Un Notebook R ou Python **Compilé** sous forme html ou pdf, obtenu avec Rstudio ou jupyter notebook, contenant les réponses rédigées au questions, les figures et résultats numériques. Tous les résultats doient êre commentés et interprétés. Certains jeux de données de `R` seront utilisés donc il est recommandé d'utiliser `R`.
- la source du notebook (un fichier `.Rmd` ou `.ipynb`) selon l'outil utilisé, exécutable sans modification de la part du correcteur. Pour que les résultats expérimentaux soient reproductibles vous devez fixer la graine de votre générateur de nombres aléatoires avant de générer des données, par exemple dans `R` :

```
set.seed(1)
```

Les dossiers doivent être déposés sur le moodle avant le mercredi 9 novembre 2022, 23h59. Chaque heure de retard coûte deux points (sur vingt).

En cas de problème d'ordre technique, vous pouvez envoyer un mail à l'adresse : `anne.sabourin@u-paris.fr`, avec en objet l'intitulé "projet SNP".

Notations

n est le nombre de données utilisées pour l'estimation, notées $X_1, \dots, X_n$. $K : \mathbb{R} \to \mathbb{R}$ est le noyau gaussien ; On note f la densité que l'on cherche à estimer, $\hat{f}_h$ la densité estimée grâce à l'estimateur à noyau de fenêtre h associé à K. On rappelle que

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K[(x - X_i)/h] \tag{1}$$

avec $K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.

Exercice 1 (Implémentation du noyau gaussien)**.**

1. Ecrire une fonction `densitygauss(X, h, Grid, plot = TRUE)`
   - prenant en arguments : le vecteur `X` des données, une fenêtre `h`, une grille d'évaluation du noyau `Grid`, et un paramètre logique `plot`, indiquant si l'estimateur doit être tracé automatiquement ou non.
   - renvoyant : une liste de deux éléments `x` et `y`, où `x = Grid` est la grille passée en argument `y` est l'estimateur à noyau $\hat{f}_h$ évalué en chaque point de `x`.
2. Simuler $n = 500$ données suivant une loi de mélange de deux lois Beta comme dans le TP1. Afficher un premier estimateur par défaut de la densité en utilisant un estimateur déjà codé (sous `R` on pourra utiliser la fonction `density` et la méthode `plot` associée).
3. Comparez visuellement l'estimateur de densité par défaut et le vôtre en superposant les deux graphes. Commentez.

Anne Sabourin anne.sabourin@u-paris.fr,

Exercice 2 (Sélection de fenêtre par validation croisées)**.** Le but de cet exercice est d'implémenter la méthode de validation croisée pour la sélection de fenêtre avec le noyau gaussien. Pour l'application on utilisera (i) les données simulées de l'exercice précédent , (ii) les données `CO2` disponibles dans `R`. On doit commencer par implémenter l'estimateur du risque quadratique intégré par validation croisée vu en cours. On rappelle que le principe de la méthode consiste à choisir $h \in \mathcal{H} = \{h_1, \dots, h_M\}$ qui minimise le critère

$$\hat{\phi}(h) = \underbrace{\int_{\mathbb{R}} \hat{f}_h(x)^2 \, \mathrm{d}x}_{A(h)} - 2 \underbrace{\frac{1}{n(n-1)h} \sum_{1 \le i,j \le n, i \neq j} K\Big(\frac{X_i - X_j}{h}\Big)}_{B(h)}.$$

le terme $B(h)$ ne pose pas de problème particulier mais il faut calculer le terme $A(h)$. On va voir dans les questions qui suivent que $A(h)$ a une expression explicite en fonction des X_i, X_j, dans le cas particulier du noyau gaussien.

1. (question préliminaire) Vérifiez que pour tous réels x, μ_1, μ_2, on a

$$(x-\mu_1)^2 + (x-\mu_2)^2 = 2(x - (\mu_1+\mu_2)/2)^2 + \frac{1}{2}(\mu_1 - \mu_2)^2. \tag{2}$$

2. En développant l'expression $\hat{f}_h(x)^2$ à partir de l'expression (1), en exploitant l'identité (2) et en utilisant le fait que pour tout $\mu \in \mathbb{R}$ fixé, $\int K_h(x-\mu)\,\mathrm{d}x = 1$, montrer que

$$A(h) = \frac{1}{2n^2 h\sqrt{\pi}} \sum_{1 \le i,j \le n} e^{-\frac{(X_i - X_j)^2}{4h^2}}. \tag{3}$$

3. Ecrire une fonction `evalA(X,h)` prenant en argument un vecteur de données X de taille n et une fenêtre h, et renvoyant la valeur de $A(h)$.
4. Ecrire une fonction `evalB(X,h)` prenant en argument un vecteur de données X de taille n et une fenêtre h, et renvoyant la valeur de $B(h)$.
5. à partir des deux questions précédentes écrire une fonction `critereCV(X,h)` prenant en argument un vecteur de données X de taille n et une fenêtre h, et renvoyant la valeur de $\hat{\phi}(h)$.
6. En utilisant une grille $\mathcal{H}$ bien choisie (en deux étapes, une étape exploratoire et une étape pour raffiner le pas), donner une estimation `hopt` de la fenêtre optimale pour le jeu de données simulées. Comparer avec la valeur de h sélectionnée automatiquement avec la fonction déjà implémentée. Tracer la courbe de densité estimée avec votre fonction `densitygauss` et la fenêtre `hopt`. Superposer avec la courbe obtenue avec la méthode par défaut, par exemple `density`.
7. Répéter la question précédente avec le jeu de donnés `CO2` disponible dans `R`. On utilisera la colonne `uptake` de cette base de données.

Exercice 3 (Selection de fenêtre, compromis biais-variance, données simulées)**.** Traiter la question 3. de la feuille de TP 1., sur les mêmes données simulées, avec $n = 500$, en considérant uniquement le noyau gaussien. Comparer le h^* (fenêtre optimale minimisant le compromis biais variance) obtenu en utilisant la connaissance de la loi, avec la valeur `hopt` obtenue par validation croisée sur un seul jeu de données.

Exercice 4 ((bonus) : estimation d'ensembles 'masse-volume')**.** Un ensemble 'masse-volume' (MV-set) au niveau $\alpha \in]0,1[$, noté Ω^*_α, pour une densité donnée f, est défini comme l'ensemble de 'volume' (ou longueur dans le cas 1D) minimal parmi les ensembles de probabilité (pour la loi f) supérieure ou égale à α. Autrement dit, en notant λ la mesure de lebesgue sur $\mathbb{R}$,

$$\Omega^*_\alpha = \underset{A \text{ mesurable}}{\arg\min} \ \lambda(A) \text{ sous contrainte } \int_A f(x)\,\mathrm{d}x \ge \alpha.$$

On peut montrer que sous des conditions faibles, Ω^*_α est l'intérieur d'une courbe de niveau de la densité, c.a.d.

$$\Omega^*_\alpha = \{x \in \mathbb{R} : f(x) \ge t_\alpha\}$$

où t_α est un niveau tel que $\int_{\{x: f(x) \ge t_\alpha\}} f(x)\,\mathrm{d}x = \alpha = \mathbb{P}\left(Z \in \{x : f(x) \ge t_\alpha\}\right) = \alpha$.

Un exemple d'application des MV-sets est la détection d'anomalie : une donnée est décrétée anormale lorsqu'elle se trouve en dehors d'un MV-set de niveau α proche de 1.

Anne Sabourin anne.sabourin@u-paris.fr,

- On considère les données `CO2$uptake` et l'estimateur de la densité $\hat{f}_{hopt}$ construit à l'exercice 2.
- Pour tout niveau t fixé on estime $\mathbb{P}(Z \in \{x : f(x) \geq t_\alpha\})$ par sa version empirique, sachant les données $X_1, \ldots, X_n$

$$\hat{P}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{f}_{hopt}(X_i) > t\}$$

1. En utilisant le principe d'estimation ci-dessus, estimer un MV-set de niveau $\alpha = 0.8$ sous la forme d'une union d'intervalles.