Examen

Master parcours SSD - UE Statistique Computationnelle

Septembre 2019

Consignes : *Cet examen compte pour la moitié de l'UE. Il prend la forme d'un devoir maison pour lequel vous devrez me rendre :*

1. *un rapport décrivant les analyses effectuées et commentant les résultats obtenus,*
2. *le(s) script(s)* `R` *correspondant,*

Il est à réaliser pour le **25 novembre** *et à me transmettre par email à* **pierre.mahe@biomerieux.com**.

1 Exercice 1 - simulation par inversion

On considère la fonction densité suivante, définie pour $a > 0$, $b > 0$:

$$f(x) = \begin{cases} \frac{ab^a}{x^{a+1}} & \text{si } x \geq b, \\ 0 & \text{sinon.} \end{cases}$$

1. Représenter cette densité pour $b = 2$ et $a \in \{1, 2, 3\}$.
2. Implémenter une procédure d'inversion pour simuler une variable aléatoire selon cette densité.
3. Simuler un échantillon et comparer graphiquement la densité obtenue empiriquement à la densité théorique.

On détaillera les calculs et le raisonnement suivi.

2 Exercice 2 - approximation de la fonction de répartition de la loi normale centrée réduite par approche MC

On cherche à approximer par une approche Monte-Carlo la fonction de répartition de la loi normale centrée réduite :

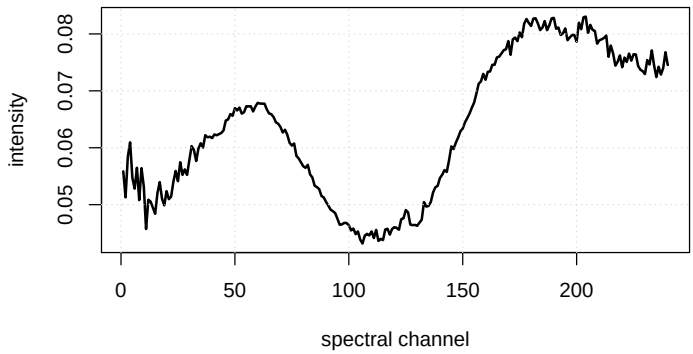
$$\phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) dt,$$

pour des valeurs $x > 0$.

1. Proposer une méthode basée sur la loi uniforme, et de préférence la loi $\mathcal{U}(0, 1)$.
2. Proposer une méthode basée sur la loi normale.
3. Comparer les valeurs obtenues par les deux procédures à la valeur réelle donnée par `R` via la fonction `pnorm` pour des valeurs croissantes de $x \in [0.1, 3]$. On considérera un nombre de tirages $n = 1000$.
4. Comparer la variance et les intervalles de confiance des deux estimateur pour les valeurs croissantes de $x \in [0.1, 3]$ et interpréter les résultats obtenus.
5. Comment modifier la procédure si $x < 0$?

3 Exercice 3 - bootstrap & ACP

Charger le jeu de données `spectra.Rdata`. Il contient une matrice `X` de taille 546×240 contenant 546 spectres définis par 240 canaux, tel que l'exemple représenté ci-dessous.

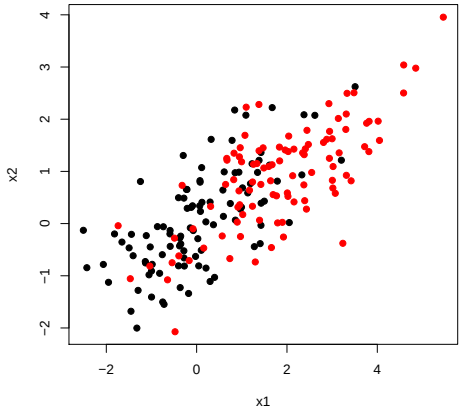


On souhaite réaliser une analyse en composantes principales (ACP) de ce jeu de données et associer un intervalle de confiance aux proportions de variance expliquées par les composantes principales, en utilisant une approche bootstrap. Pour cela :

1. Représenter (sur une même figure) un échantillon aléatoire de 50 spectres pour en apprécier leur variabilité, et commenter le résultat obtenu.
2. Effectuer une ACP et représenter les proportions de variance expliquées par les 10 premières composantes principales : quelle proportion de la variance totale est expliquée par les deux premières composantes principales ? Représenter sous la forme de spectres les deux axes principaux (i.e., les opérateurs linéaires permettant de passer de l'espace des canaux aux deux premières composantes principales) : sont-ils cohérents avec vos observations de la question 1 ?
3. Implémenter une procédure bootstrap pour calculer les intervalles de confiance associés aux proportions de variance expliquées par les 10 premières composantes principales, par la méthode de votre choix (e.g., quantile, basique, ...). Proposer une représentation graphique de vos résultats.
4. Enfin, modifier votre procédure pour pouvoir représenter la variabilité induite par le bootstrap sur les axes principaux de la question 2 et commenter les résultats obtenus.

4 Exercice 4 : tests par permutation

Charger le jeu de données `permutation.Rdata`. Il contient une matrice `X` de taille 200×2 représentée ci-dessous, les 100 premières lignes de la matrice correspondant aux points noirs, et les 100 dernières aux points rouges.



On souhaite tester si on est capable de détecter une différence entre ces deux populations, i.e., si on est capable de détecter une différence significative entre les deux nuages de points, stockés dans les 100 premières et les 100 dernières lignes de la matrice. On considère pour cela une statistique basée sur une information de voisinage :

$$T = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(y_{n(x_i)} = y_i),$$

où $n(x_i) \in \{1, ..., i-1, i+1, .., n\}$ est l'indice du plus proche voisin du point x_i, $y_i = 1$ si x_i appartient aux premier ensemble de points (les points noirs) et $y_i = 0$ sinon, et la fonction $\mathbf{1}(.)$ vaut 1 si son argument est vérifié et 0 sinon.

1. Que mesure cette statistique ?
2. Appppliquer une procédure par permutation pour $B = 1000$ tirages et évaluer la p-valeur obtenue. La différence entre les deux nuages de points est-elle significative ?
3. Reproduire cette analyse en tirant aléatoirement un ensemble de $n = 10, 20, ..., 90$ points parmi chaque population. Comment la p-valeur évolue t'elle ? Représenter les résultats sous la forme d'un graphique.
4. Enfin, reproduire cette seconde analyse 10 fois et représenter la variabilité dans les p-valeurs obtenues en fonction du nombre de points considérés. A partir de quelle taille d'échantillon est-on capable de détecter une différence significative en considérant que la médiane des p-valeurs obtenues sur les 10 répétitions doit être (et rester) inférieure à 0.05 ?