TP 11 (Contrôle continu n°2) : Simulation, intervalles de confiance, tests

Bonjour à tous ! Aujourd'hui, même format, même concept : vous pouvez utiliser votre travail des TPs précédents, l'aide de R ainsi qu'internet.

Le TP suivant vous est donné sous la forme d'un fichier .pdf, contenant le sujet, ainsi qu'un notebook qui tient lieu de "fiche réponse" : sur celui-ci, des zones sont laissées libres pour inclure votre code et vos réponses écrites. A la fin du contrôle, vous devrez enregistrer votre notebook complété par vos soins, le renommer sous la forme prenom_nom.Rmd, et me l'envoyer par mail dans le temps imparti à l'adresse suivante : `luca.ganassali@inria.fr`.

Attention, la qualité du code est évaluée : le manque de clarté, l'absence de commentaires, les erreurs faisant bugguer le kernel seront pénalisées.

Bon courage ! :)

Exercice 1 : Intervalles de confiance pour les lois exponentielles (10 points) On considère un échantillon de n variables $X_1, \ldots, X_n$ i.i.d. de loi exponentielle de paramètre $t > 0$. On considère l'estimateur de t suivant :

$$\hat{t} := \frac{n}{\sum_{i=1}^{n} X_i}.$$

1. Comment cet estimateur $\hat{t}$ a été obtenu selon vous ?

2. On donne le résultat suivant :

$$\sqrt{n}\frac{\hat{t} - t}{t} \xrightarrow{(d)} \mathcal{N}(0, 1),$$

cette convergence ayant lieu quand n tend vers $+\infty$.

A l'aide d'une méthode de votre choix, illustrer ce résultat. *Indication: on prendra garde à fournir une illustration claire et complète.*

3. A l'aide du résultat précédent, déterminer un intervalle de confiance pour le paramètre t de niveau $1 - \alpha$. *Indication : selon la méthode, il y a plusieurs expressions possibles. Je ne vous en demande qu'une seule, bien justifiée.*

S'agit-t-il d'un intervalle exact/par excès ? asymptotique/non-asymptotique ?

4. Ecrire une fonction `IC(X,alpha)` qui prend en argument le seuil d'erreur `alpha`, l'échantillon `X`, et qui renvoie l'intervalle de confiance établi plus haut pour le paramètre t de niveau $1 - \alpha$.

5. Les jeux de données data1_guichet1.csv, data1_guichet2.csv, et data1_guichet3.csv contiennent des mesures de temps d'attente de clients à trois guichets différents. Un de ces guichets semble avoir des temps d'attente significativement distincts des deux autres. On cherche à repérer lequel.

5.(a). Afficher, pour chaque groupe, l'histogramme des données. Commenter sa forme.

5.(b). A l'aide des questions précédentes, tester si les paramètres des lois sont significativement différents . Répondre à la question posée.

6. Les jeux de données data2_guichet1.csv et data2_guichet2.csv contiennent eux aussi des mesures de temps d'attente de clients à deux guichets différents. On cherche à repérer si l'un est significativement plus rapide que l'autre.

6.(a). Afficher, pour chaque groupe, l'histogramme des données. Commenter sa forme.

6.(b). Les moyennes des deux échantillons sont-elles significativement différentes ? *Indication : si vous voyez plusieurs méthodes pour cette question, choisissez-en une seule.*

* * *

Exercice 2 : Clustering dans un modèle de mélange (10 points, et jusqu'à + 3 pts pour la partie bonus) Dans tout cet exercice, les variables considérées sont à valeurs dans $\mathbb{R}^2$. On introduit quatre paramètres : $\mu^{(1)}, \mu^{(2)}$ sont des éléments de $\mathbb{R}^2$, σ est un réel positif, et $p \in [0,1]$.

On considère le modèle suivant : Soit $X_1^{(1)}, \dots, X_n^{(1)}$ (resp. $X_1^{(2)}, \dots, X_n^{(2)}$) des variables i.i.d. de loi normale multivariée $\mathcal{N}\left(\mu^{(1)}, \sigma^2 I_2\right)$ (resp. $\mathcal{N}\left(\mu^{(2)}, \sigma^2 I_2\right)$), où I_2 désigne la matrice identité de taille 2×2.

Les X_i sont échantillonnées indépendamment pour tout i comme suit : avec probabilité p, $X_i = X_i^{(1)}$, et avec probabilité $1-p$, $X_i = X_i^{(2)}$. Si $X_i = X_i^{(1)}$ (resp. $X_i = X_i^{(2)}$), on dira que X_i appartient au groupe 1 (resp. au groupe 2).

1. Simuler un échantillon $(X_i)_{1 \le i \le n}$ de taille $n = 500$, avec les paramètres suivants : $\mu^{(1)} = \binom{-2}{1}, \mu^{(2)} = \binom{0.3}{-3}, \sigma = 0.8$ et $p = 0.6$.

Indication : les variables $X_i^{(1)}$ et $X_i^{(2)}$ ont beau être dans $\mathbb{R}^2$, leurs coordonnées sont indépendantes dans ce modèle. On pourra donc simuler les variables coordonnées par coordonnées, et utiliser la commande `cbind`*.*

2. Afficher dans le plan le nuage de points des données simulées précédemment. Commenter ce que vous observez. Pourquoi ce modèle est-il appelé "modèle de mélange" selon vous ?

3. Que se passe-t-il dans le modèle lorsque le paramètre σ augmente ?

On suppose dans toute la suite que $\mu^{(1)}$ et $\mu^{(2)}$ sont connus. On cherche à retrouver dans quel groupe tombent les variables X_i, pour tout $1 \le i \le n$. Pour ce faire, on propose la méthode suivante : si $\|X_i - \mu^{(1)}\| \le \|X_i - \mu^{(2)}\|$, on décide que X_i appartient au groupe 1, et au groupe 2 sinon.

4. Ecrire une fonction `grouper(X,mu1,mu2)` qui prend en argument un échantillon X, $\mu^{(1)}$ et $\mu^{(2)}$, et qui renvoie un vecteur de même taille que X contenant les groupes inférés. *Indication : on pourra utiliser la fonction définie ci-dessous.*

```
norme = function(x){
  sqrt(sum(x^2))
}
```

5. Simuler un nouvel échantillon X de taille $n = 300$, avec les paramètres suivants : $\mu^{(1)} = \binom{-1}{1}$, $\mu^{(2)} = \binom{1.2}{-0.5}$, $\sigma = 2$ et $p = 0.2$, en gardant en mémoire les groupes des variables.

Tester la fonction `grouper` sur ce jeu de données. Quel est le pourcentage des données dont le groupe est bien retrouvé ? A quel pourcentage doit-on le comparer pour juger ? Commenter.

6. Exécuter le code suivant. Que fait-il ? Commenter.

```
plot(X,col=1+groupes)
plot(X,col=1+groupes_estims)
```

7. Afficher le ratio du nombre d'éléments classés dans le groupe 1 par notre algorithme sur la taille totale de l'échantillon. Commenter le résultat obtenu.

8. Compléter le squelette du code ci-dessous, qui affiche les performances de l'algorithme précédent pour les mêmes valeurs de $\mu^{(1)} = \binom{-1}{1}, \mu^{(2)} = \binom{1.2}{-0.5}$ et $p = 0.2$, mais en faisant varier σ dans $[0, 15]$. Commenter la courbe observée.

```
n = 5000 # taille de chaque échantillon
N = 10 # nombre de simulations pour chaque valeur de sigma
p = 0.2
mu1 = ### A COMPLETER
mu2 = ### A COMPLETER

pas_sigmas = 0.4 # pas d'évolution de sigma
sigmas = ### A COMPLETER

nb_sigmas = length(sigmas)

mean_performances = replicate(nb_sigmas,0)

for (i in 1:nb_sigmas){
  sigma = ### A COMPLETER
  performances = replicate(N,0)
  for (j in 1:N){
    ### On simule un échantillon X : A COMPLETER


    ### On infère les groupes de X : A COMPLETER
    groupes_estims =

    ### on enregistre la performance : A COMPLETER
    performances[j] =
  }
  ### on moyenne les performances sur les N échantillons : A COMPLETER
  mean_performances[i] =
}

plot(sigmas,mean_performances,col="blue",main="Performance de l'algorithme en fonction de sigma")
lines(sigmas,mean_performances, col="blue", lty = 2)
```

* * *

Fin de l'exercice 2 : partie bonus (jusqu'à + 3 pts) *Dans cette partie, toute trace de recherche pertinente, même non aboutie, sera prise en compte, à condition que l'éxécution du code ne fasse pas bugguer le kernel.*

9. (code) Quel(s) autre(s) paramètre(s) que σ, à votre avis, impacte le plus la performance de l'algorithme ? Illustrer ce résultat en adaptant la question 8.

10. (maths) On cherche des conditions sous lesquelles l'algorithme retrouve les bons groupes *sans aucune erreur*, avec probabilité tendant vers 1 quand $n \to +\infty$. Pour cela, on donne le résultat suivant : si $Y \sim \mathcal{N}(\mu_0, \sigma^2 I_2)$, alors pour tout $\mu \in \mathbb{R}^2$, on a

$$\mathbb{P}\left(\|Y - \mu\| < \|Y - \mu_0\|\right) \leq \frac{2\sigma}{\|\mu - \mu_0\|} \exp\left(-\frac{\|\mu - \mu_0\|^2}{8\sigma^2}\right).$$

Donner une condition (suffisante) faisant intervenir n et la quantité

$$\rho := \frac{\|\mu^{(2)} - \mu^{(1)}\|}{\sigma}$$

pour que l'algorithme ne fasse aucune erreur, avec probabilité tendant vers 1 quand $n \to +\infty$.

11. (culture) Dans le cas où $\mu^{(1)}$ et $\mu^{(2)}$ sont inconnus, comment pourrait-on faire selon vous pour retrouver les groupes des variables ?