BE4 de Probabilités/Statistique

(Compléments)

1 Lecture de code barre

Dans votre supermarché préféré, les produits sont identifiés au moyen d'un code barre. Celui-ci est composé de bandes blanches et noires de largeurs variables. Ce code est lu en caisse par un stylo optique qui permet de déterminer le numéro auquel est associé le produit. Le stylo saisit une ligne traversant le code barre. Il s'agit ensuite de traiter le signal ainsi obtenu afin de déterminer la largeur des différentes bandes blanches et noires.
Lorsque la lecture est parfaite, le signal échantillonné $y(n)$ est constitué de points d'amplitudes A_0 (pour les bandes blanches) et A_1 (pour les bandes noires). Un tel signal est représenté sur la figure 1.
Cependant, en pratique, le signal est perturbé par un bruit dû à la lecture imparfaite du stylo optique. On obtient alors le signal représenté sur la figure 1. Le problème consiste alors à déterminer les sauts d'amplitude (égalements appelés "ruptures") dans ce signal bruité.

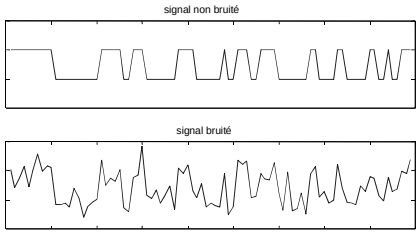

Fig. 1 : exemple de signal idéal et de signal bruité.

On supposera dans toute la suite du problème que le bruit suit une loi normale $\mathcal{N}\left(0,\sigma^2\right)$ de variance σ^2 connue. Pour simplifier, on suppose qu'il n'y a qu'une seule rupture. Le signal $(y(n))_{n=1,\dots,N}$ à traiter peut alors s'écrire :

$$y(n) = \begin{cases} A_0 + b(n), & \text{pour } n = 1,\dots,R \\ A_1 + b(n), & \text{pour } n = R+1,\dots,N \end{cases}$$

où $b(n)$ suit une loi $\mathcal{N}\left(0,\sigma^2\right)$. On suppose de plus que $A_1 > A_0$. On cherche à estimer l'instant de rupture R.

1.1 Estimateur du maximum de vraisemblance

Le signal $y(n)$ suit une loi $\mathcal{N}\left(A_0,\sigma^2\right)$ pour $n = 1,\dots,R$, et une loi $\mathcal{N}\left(A_1,\sigma^2\right)$ pour $n = R+1,\dots,N$. L'estimateur du maximum de vraisemblance $\widehat{R}$ de R_0 s'obtient en maximisant la loi du vecteur $(y(n))_{n=1,\dots,N}$ par rapport à R. On montre alors que $\widehat{R}$ est solution du problème de minimisation suivant :

$$\min_{1\leq R\leq N-1} f(R) = \sum_{n=1}^{R} \left(y(n) - A_0\right)^2 + \sum_{n=R+1}^{N} \left(y(n) - A_1\right)^2$$

QUESTIONS :

1. Ecrire une fonction Matlab `y=ruptures(A0,A1,R,N,sigma2)` qui renvoie le vecteur $(y(n))_{n=1,\dots,N}$ en fonction de A_0, A_1, R, N, et σ^2. Afficher le vecteur obtenu avec $A_0 = 0$, $A_1 = 1$, $R = 50$, $N = 100$, et $\sigma^2 = 0.1$.

2. Ecrire une fonction Matlab `f=f_R(y,A0,A1)` qui calcule le vecteur $(f(R))_{R=1,\dots,N-1}$ (utiliser la fonction `sum` pour chaque R).

3. Ecrire alors une fonction `R_est=estimation(A0,A1,R,N,sigma2)`qui trace $f(R)$ et déterminer $\widehat{R}$ grâce à la commande `min`.

4. On cherche à évaluer la précision de $\widehat{R}$. L'instant de rupture sera choisi tel que $R = N/2$. Puisque cet instant dépend du nombre de points N, la précision de l'estimation doit se faire relativement à N. En d'autres termes, on doit considérer l'estimateur normalisé $\widehat{R}\,/N$. Ecrire une fonction `[m,v]=moyenne_variance_R(A0,A1,`
qui renvoie la moyenne et la variance de $\widehat{R}/N$ calculées sur 100 réalisations. Tester cette fonction avec $A_0 = 0$, $A_1 = 1$, $\sigma^2 = 10$, et $N = 10$, puis $N = 1000$ (cela peut être un peu long pour $N = 1000$, faire autre chose en attendant !...). Commenter.

1.2 Test de Neyman-Pearson

On veut déterminer si la rupture a lieu au point R_0 ou au point R_1, avec $R_1 > R_0$, à l'aide du test :

$$\begin{array}{ll} H_0 : & R = R_0 \\ H_1 : & R = R_1 \end{array}$$

On peut alors montrer que la statistique de test associée à ce problème s'écrit :

$$T\left(Y_1, \dots, Y_N\right) = \frac{1}{L} \sum_{n=R_0+1}^{R_1} Y_n \ ,$$

avec $L = R_1 - R_0$. $T\left(Y_1, \dots, Y_N\right)$ est donc la moyenne du signal entre les instants de ruptures supposés. Etant donné que $R_1 > R_0$, la loi de T sous chaque hypothèse est :

$$\begin{array}{ll} H_0 : & T \sim \mathcal{N}\left(A_1, \sigma^2/L\right) \\ H_1 : & T \sim \mathcal{N}\left(A_0, \sigma^2/L\right) \end{array}$$

Bien noter que la moyenne est A_1 sous H_0, et A_0 sous H_1, avec $A_1 > A_0$. Pour une probabilité de fausse alarme α, la région critique (zone de rejet de H_0) est donnée par :

$$R_\alpha = \left\{(y_1, \dots, y_N) \in \mathbb{R}^N \,|T\left(y_1, \dots, y_N\right) < \lambda_\alpha \right\}$$

Dans ce cas, le seuil de décision s'écrit :

$$\lambda_\alpha = \sqrt{\frac{\sigma^2}{L}} \Phi^{-1}(\alpha) + A_1 \ ,$$

et la probabilité de non-détection s'écrit :

$$\beta = 1 - \Phi\left(\sqrt{\frac{L}{\sigma^2}}\left(\lambda_\alpha - A_0\right)\right).$$

QUESTIONS :

1. Ecrire à l'aide des fonctions `normcdf` et `norminv` une fonction `p= pi_theorique(A0,A1,R0,R1,N,sigma2)` qui renvoie la puissance théorique π du test pour $\alpha = 0.01, 0.02, 0.03, \dots, 0.98, 0.99$, en fonction de A_0, A_1, R_0, R_1, N, et σ^2. Afficher le vecteur obtenu avec $A_0 = 0$, $A_1 = 1$, $R_0 = 50$, $R_1 = 55$, $N = 100$, et $\sigma^2 = 0.5$.

2. Montrer que β est une fonction de L, α, et $RSB = \frac{A_1 - A_0}{\sigma}$. Ecrire deux fonctions `cor_L(A0,A1,R0,N,sigma2)` et `cor_RSB(A0,A1,R0,R1,N)` qui donnent la courbe $\pi(\alpha)$ pour différentes valeurs de L, puis pour différentes valeurs de RSB. Commenter.

3. On cherche maintenant à retrouver ces résultats par simulations. Ecrire une fonction `p=pi_estimee(A0,A1,R0,R1` qui renvoie la puissance estimée $\widehat{\pi}$ du test pour $\alpha = 0.01, 0.02, 0.03, \ldots, 0.98, 0.99$, en fonction de A_0, A_1, R_0, R_1, N, σ^2, et du nombre de simulations K (cette fonction utilisera bien sûr la fonction `ruptures` définie à la question 2.1.1). Superposer à la première courbe les résultats obtenus avec $A_0 = 0$, $A_1 = 1$, $R_0 = 50$, $R_1 = 55$, $N = 100$, $\sigma^2 = 0.5$, et $K = 500$.