BE - PROBAS/STATS - 2GE

Introduction

Sur un avion, l'usure des gaines d'isolation des cables d'alimentation électriques (cf Figure 1) peut engendrer un phénomène de conduction entre phases. Cela provoque l'apparition d'arcs qui se propagent le long des cables (phénomène appelé "arc tracking"). Ce phénomène est précédé de faibles perturbations des signaux électriques (qui se manifestent par des transitoires, comme illustré sur la Figure 2) qui doivent être détectés le plus tôt possible afin de changer les circuits avant que les équipements ne soient endommagés. Ces transitoires se manifestent comme des changements de variances qu'il convient de détecter. Cette phase de détection fait l'objet de ce projet.

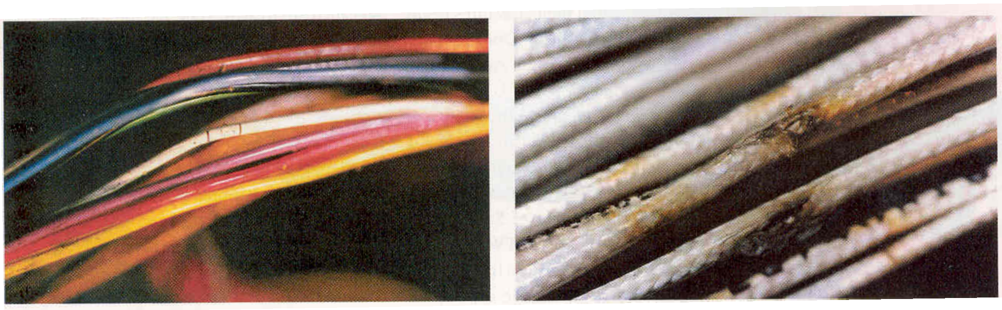

Figure 1: Usure des gaines d'isolation.

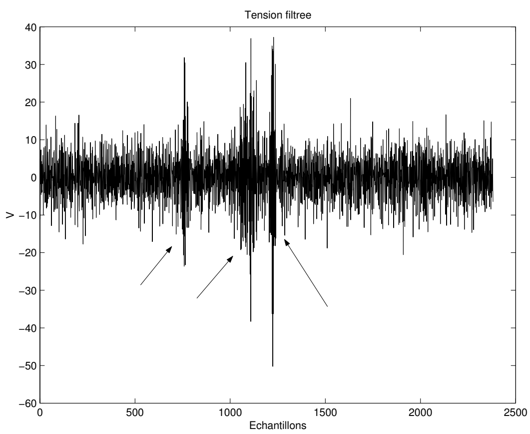

Figure 2: Transitoires visibles sur un signal de tension filtré.

Un modèle réaliste de signal transitoire noté $\boldsymbol{y} = (y_1, ..., y_N)^T$ est de considérer que les variables aléatoires $Y_1, ..., Y_n$ sont indépendantes de même loi normale de moyenne nulle et de variance σ^2. De manière à modéliser les sauts de variances, on suppose que $\sigma^2 = \sigma_0^2$ dans les zones de faible variance et $\sigma^2 = \sigma_1^2 > \sigma_0^2$ dans les zones de forte variance. Dans ce projet, on supposera pour simplifier qu'il n'y a qu'une seule zone de forte variance correspondant à $i = n_0 + 1, ..., n_1$ (de largeur notée $L = n_1 - n_0$) et que les deux autres zones ($i = 1, ..., n_0$ et $i = n_1 + 1, ..., N$) sont de faible variance. On suppose également dans un premier temps que les paramètres σ_0^2, N, n_0 et n_1 sont connus, ce qui signifie que le seul paramètre inconnu du modèle est $\theta = \sigma_1^2$.

Travail à effectuer

1. **Génération d'un signal test**
Écrire une fonction `Y = generer(`σ_0`,`σ_1`,`n_0`,`n_1`,N,K)` qui renvoie une matrice `Y` de taille $N \times K$, dont chaque colonne contient une réalisation du signal $\boldsymbol{y} = (y_1, \ldots, y_N)^T$. Les paramètres d'entrée sont

   - σ_0 et σ_1 : écart types de Y_i associées aux zones de faible et forte variances
   - n_0 et n_1 : limites (début et fin) de la zone de forte variance
   - N : nombre de points d'un signal observé
   - K : nombre de signaux observés

   Pour générer les réalisations du signal, on utilisera la fonction `randn(M,N)` de Matlab qui génère une matrice de taille $M \times N$ constituée de réalisations indépendantes d'une loi normale $\mathcal{N}(0,1)$ et on multipliera les éléments de cette matrice par une constante à déterminer de manière à obtenir des réalisations de lois normales $\mathcal{N}(0,\sigma_0^2)$ ou $\mathcal{N}(0,\sigma_1^2)$ répondant au problème posé. Tester cette fonction avec $N = 100$, $K = 500$, $n_0 = 40$, $n_1 = 60$, $\sigma_0^2 = 1$ et $\sigma_1^2 = 10$. Afficher une réalisation du signal $\boldsymbol{y}$ (c'est-à-dire une colonne de la matrice `Y`) et tracer ensuite la variance des colonnes de `Y` à l'aide de la fonction `var`.

2. **Estimation statistique**
Des calculs élémentaires permettent de montrer que l'estimateur du maximum de vraisemblance de θ est donné par :
$$\widehat{\theta}_{MV} = \frac{1}{n_1 - n_0} \sum_{i=n_0+1}^{n_1} Y_i^2.$$
On montre ensuite que

   (a) $\widehat{\theta}_{MV}$ est un estimateur non-biaisé :
$$E\left[\widehat{\theta}_{MV}\right] = \theta$$

   (b) la variance de $\widehat{\theta}_{MV}$ est définie par :
$$\text{var}\, \widehat{\theta}_{MV} = \frac{2\sigma_1^4}{n_1 - n_0}$$

   (c) $\widehat{\theta}_{MV}$ est un estimateur efficace, c'est-à-dire que sa variance est égale à la borne de Rao-Cramer des estimateurs non-biaisés de θ.

   - Ecrire une fonction `[teta_est,BRC] = estimateur_mv(Y,`σ_1`,`n_0`,`n_1`)`, qui renvoie l'estimateur $\widehat{\theta}_{MV}$ pour chacune des K réalisations de $y = (y_1, \ldots, y_N)^T$, à partir de la matrice `Y` construite à la question 1, ainsi que la borne de Rao-Cramer des estimateurs non-biaisés de σ_1^2. On obtient alors K valeurs de $\widehat{\theta}_{MV}$, notées $\left(\widehat{\theta}_{MV}(k)\right)_{k=1,\ldots,K}$;
   - Donner alors la moyenne et la variance de $\left(\widehat{\theta}_{MV}(k)\right)_{k=1,\ldots,K}$, et comparer avec les valeurs théoriques ;

3. **Détection**
On cherche à étudier les performances d'un test statistique qui permet de détecter s'il y a un transitoire dans le signal observé entre les instants n_0 et n_1 (c'est-à-dire s'il y a une zone de forte variance entre ces instants). Les deux hypothèses sont alors définies par

$$\begin{aligned} H_0 : \quad & Y_i \sim \mathcal{N}(0, \sigma_0^2), \; i = n_0 + 1, ..., n_1 \\ H_1 : \quad & Y_i \sim \mathcal{N}(0, \sigma_1^2), \; i = n_0 + 1, ..., n_1 \end{aligned} \tag{1}$$

La statistique de test donnée par le théorème de Neyman-Pearson associée à ces deux hypothèses s'écrit

$$T(\boldsymbol{Y}) = \sum_{i=n_0+1}^{n_1} Y_i^2.$$

De plus, pour une probabilité de fausse alarme α, la région critique du test (zone de rejet de H_0) est définie par :

$$R_\alpha = \quad \left\{ \boldsymbol{y} \in \mathbb{R}^N | T(\boldsymbol{y}) > \lambda_\alpha \right\}. \tag{2}$$

Dans ce cas, le seuil de décision s'écrit

$$\lambda_\alpha = \sigma_0^2 F_L^{-1}(1 - \alpha) \tag{3}$$

où F_L^{-1} est l'inverse de la fonction de répartition d'une loi du chi2 à $L = n_1 - n_0$ degrés de liberté (une loi χ_L^2). On montre également que la probabilité de non-détection (ou risque de 2ième espèce) du test s'exprime sous la forme suivante

$$\beta = F_L\left(\frac{\lambda_\alpha}{\sigma_1^2}\right)$$

où F_L est la fonction de répartition d'une loi du chi2 à $L = n_1 - n_0$ degrés de liberté. On souhaite tracer les courbes théoriques de la puissance du test $\pi = 1 - \beta$ en fonction de la probabilité de fausse alarme α, puis retrouver ces courbes par simulations.

- En utilisant les fonctions `chi2inv` et `chi2cdf`, écrire une fonction

  `p = pi_theorique(`σ_0`,`σ_1`,`L`)`

  qui renvoie la puissance théorique p du test pour $\alpha \in \{0.01, 0.02, \ldots, 0.98, 0.99\}$ en fonction des paramètres σ_0^2,σ_1^2 et $L = n_1 - n_0$. Tracer les courbes obtenues pour $\sigma_0^2 = 1$,$\sigma_1^2 = 10$ et différentes valeurs de L ($L \in \{10, 15, 20\}$. Tracer les courbes obtenues pour $\sigma_0^2 = 1$, $L = 20$ et différentes valeurs de σ_1^2 ($\sigma_1^2 \in \{10, 6, 4\}$. Commenter les résultats obtenus.
- On cherche maintenant à retrouver ces résultats par simulation. Puisque σ_0 est un paramètre connu, on peut déterminer le seuil du test λ_α pour toute valeur de α en utilisant (3). Pour estimer la puissance du test, il suffit donc d'estimer la probabilité P[Rejeter H_0 $|H_1$ vraie]. Pour cela
  - Ecrire une fonction qui génère K réalisations de signaux associés à l'hypothèse H_1 du test (1).
  - Écrire une fonction

    `p = pi_estimee(`σ_0`,`σ_1`,`L`,`K`)`

    qui renvoie la puissance estimée $\hat{p}$ du test pour $\alpha \in \{0.01, 0.02, \ldots, 0.99\}$, en fonction de σ_0^2, σ_1^2, L, et du nombre de simulations `K`. La puissance sera estimée à l'aide des signaux associés à l'hypothèse H_1 générés à la question précédente. Superposer la courbe COR théorique obtenue avec la fonction `pi_theorique` et la courbe COR estimée $\hat{p}$ avec $\sigma_0^2 = 1$,$\sigma_1^2 = 4$,$L = 20$, et $K = 50000$. Commenter.

- *Loi Gaussienne* $\mathcal{N}(m, \sigma^2)$ $\sigma > 0$, $x \in \mathbb{R}$
  - *densité :* $f(x; m, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]$
  - *moyenne :* m
  - *variance :* σ^2
- *Loi du chi2 à* L *degrés de liberté* χ^2_L $L > 0$, $x > 0$
  - *définition : si* $X_1, ..., X_n$ *sont* n *variables aléatoires indépendantes de même loi normale* $\mathcal{N}(0, 1)$ *alors la loi de* $\sum_{i=1}^{n} X_i^2$ *est une loi du* χ^2_L