Détection statistique de rupture dans le cadre online

Nassim Sahki, Anne Gégout-Petit, Sophie Wantz-Mézières

Détection Statistique de Rupture dans le Cadre Online

Nassim Sahki [1] & Anne Gégout-Petit [1] & Sophie Wantz-Mézières [1]

[1] *Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France ;*
nassim.sahki@inria.fr ; anne.gegout-petit@univ-lorraine.fr ;
sophie.mezieres@univ-lorraine.fr

Résumé. Nous introduisons la version online de la statistique de CUSUM basée sur un test séquentiel du rapport de vraisemblance, que nous remplaçons par une fonction de score dans le cas non-paramétrique. La détection de rupture est basée sur une règle d'arrêt et la sélection d'un seuil de détection. Dans notre travail, nous proposons un seuil de détection instantané, et des nouvelles règles d'arrêt dans le but de contrôler les paramètres de détection donnés par le taux de fausse alarme instantané (IFAR), le temps moyen entre fausses alarmes (MTBFA) ainsi que le délai moyen de détection (ADD). Finalement, nous présentons des résultats de simulation par l'estimation des paramètres de détection.

Mots-clés. Détection de rupture online, statistique récursive de CUSUM, seuil de détection, règles d'arrêt.

Abstract. We introduce the online version of the CUSUM statistics based on a sequential test of the likelihood ratio, which we replace with a score function in the non-parametric case. Change-point detection is based on a stopping rule and the selection of a detection threshold. In our work, we propose an instantaneous detection threshold dependent on time and new stopping rules in order to control the detection parameters given by the instantaneous false alarm rate (IFAR), the mean time between false alarms (MTBFA) and the average detection delay (ADD). Finally, we present simulation results by estimating detection parameters.

Keywords. Online change-point detection, recursive CUSUM statistics, detection threshold, stopping rules.

1 Introduction

Dans la théorie de la détection de rupture, la configuration des approches diffère entre deux principaux cadres : "offline" et "online". Dans le cadre offline, l'ensemble des données est fixe, c'est-à-dire qu'il est observé et traité en une fois. Dans ce cas, on s'intéresse à la détection de toutes les ruptures, le plus précisément possible. Dans le cadre online, les données arrivent en temps réel, soit par point ou par lot. L'ajout et le traitement de

données sont effectués instantanément avant l'arrivée de nouvelles données. Dans ce cas, on s'intéresse à la détection la plus rapide de la plus récente rupture.

Nous nous intéressons ici à la détection de rupture dans le cadre online, et nous considérons que la rupture concerne soit la moyenne, soit la variance. La détection de rupture online repose sur le choix d'une statistique et du seuil qu'elle doit atteindre pour signaler une détection. Nous utilisons ici la méthode classique des sommes cumulées (CUSUM). L'idée est de tester séquentiellement l'existence de rupture par l'écriture récursive de la statistique de détection introduite par Page (1954). La méthode classique pour la sélection du seuil de détection se base sur les inégalités de Wald (1945).

Dans notre travail, nous proposons une méthode empirique pour la sélection du seuil de détection. Nous suggérons également de remplacer le seuil constant par un seuil instantané dépendant du temps (dynamique). Dans le but de contrôler des performances de détection, nous proposons de nouvelles règles d'arrêt pour la statistique de détection par correction de la règle d'arrêt utilisée basiquement par l'approche CUSUM. Une nouvelle configuration de la détection est aussi proposée par la réinitialisation de la procédure de détection lorsque la statistique revient à l'état initial.

2 Cadre de détection online

Soit une série d'observations $\{X_i\}_{i=1,...,n}$ séquentiellement observée jusqu'à l'instant n, non fixé. X_n est le dernier point observé et mis-à-jour dans l'ensemble des données. Dans l'état normal (inexistence de rupture), toutes les observations sont distribuées selon la fonction de densité $f(x)$. Si une rupture existe à l'instant $1 \leq k < n$ (supposé déterministe et inconnu), on note par $v = k$ la rupture détectée. Dans ce cas, les observations $X_1, X_2, ..., X_v$ sont distribuées selon la densité $f(x)$, et les observations $X_{v+1}, X_{v+2}, ..., X_n$ seront distribuées selon une densité $g(x) \neq f(x)$.

2.1 Hypothèses de détection

La question de la détection de rupture est posée comme un test statistique entre l'hypothèse nulle $H_{0,n}$: "aucune rupture ne s'est produite, $v = +\infty$" contre l'hypothèse alternative $H_{1,n}$: "Il existe un instant $1 \leq k < n$ où une rupture $v = k$ s'est produite". La détection de rupture online consiste à tester séquentiellement l'hypothèse $H_{0,n}$ contre $H_{1,n}$ pour chaque nouvelle observation x_n :

$$\begin{cases} H_{0,n} : v = +\infty & x_i \sim f(x_i) & \forall\ i = 1, ..., n \\ H_{1,n} : v = k\ , 1 \leq k < n & x_i \sim f(x_i) & \forall\ i = 1, ..., k \\ & x_i \sim g(x_i) & \forall\ i = (k+1), ..., n \end{cases} \tag{1}$$

On définit le log du rapport de vraisemblance entre la distribution après rupture et avant rupture pour chaque instant i comme suit :

$$L_i = \log(\Lambda_i) = \log\left(\frac{g(X_i)}{f(X_i)}\right), \qquad 1 \leq i \leq n$$

2.2 Statistique récursive de CUSUM

La statistique de détection CUSUM (Page (1954)) est donnée à l'instant $n \geq 1$ par :

$$V_n = \max_{1 \leq k \leq n} \prod_{j=k+1}^{n} \Lambda_j \qquad n \geq 1 \tag{2}$$

Pour une détection séquentielle de rupture, l'approche CUSUM maximise itérativement le log du rapport de vraisemblance. La statistique de CUSUM online est donnée à l'instant n par l'écriture récursive :

$$W_n = \max\{0, \log(V_n)\} \ = \max\{0, W_{n-1} + L_n\}, \quad n \geq 1 \ ; \quad W_0 = 0 \tag{3}$$

Dans la pratique, toutefois, les deux distributions f et g ne sont pas toujours connues. Dans ce cas, toute approche basée sur la vraisemblance est inutile. Il est donc proposé de remplacer le log du rapport de vraisemblance $L_n = \log(\Lambda_n)$, par une fonction de score $S_n = S_n(X_1, ..., X_n)$ que nous développerons pas ici (détails dans l'article Tartakovsky et al. (2013)).

2.3 Règle d'arrêt

La procédure de détection séquentielle est basée sur la règle d'arrêt :

$$C_h = \min\{n \geq 1 : W_n \geq h\}, \qquad h \geq 0 : \text{seuil de détection constant.}$$

La procédure calcule récursivement la statistique de détection (W_n) pour chaque nouvelle observation, et lorsque la statistique dépasse un certain seuil de détection h défini à l'avance, la procédure marque un temps d'arrêt (déclenchement d'alarme) pour signaler qu'une rupture s'est produite.

2.3.1 Temps d'arrêt

Soit T l'instant d'alarme défini par : $T = \min_{1 \leq i \leq n} \{W_i \geq h\}, \qquad n \geq 1$

T est une variable aléatoire qui a les propriétés d'un temps d'arrêt.

— Si $T = v$: détection précise de la rupture v.
— Si $T > v$: la rupture v est détectée avec un retard $(T - v)$.
— Si $T < v$: la rupture v ne s'est pas encore produite. On dit que la procédure a déclenché une fausse alarme.

2.3.2 Critères d'évaluation

Afin d'évaluer la performance de la détection, nous définissons théoriquement les paramètres d'évaluation. On note par $\mathbb{P}_v[.], \mathbb{E}_v[.]$, la probabilité et l'espérance sous $\mathrm{H}_{1,n}$ (lorsque la rupture est produite en v).

Temps moyen entre fausses alarmes : Page (1954) et Lorden (1976) proposent une métrique qui capture le nombre moyen d'observations avant le déclenchement d'une fausse alarme (MTBFA "Mean Time Between False Alarm").

$$MTBFA = \mathbb{E}_\infty [T] \tag{4}$$

Taux de fausse alarme instantané : est défini à partir du MTBFA.

$$\alpha = \frac{1}{\mathbb{E}_\infty [T]} \tag{5}$$

Délai moyen de détection : L'ADD "Average Detection Delay" est évalué sous $H_{1,n}$, pour quantifier la vitesse de détection.

$$ADD = \mathbb{E}_v [T - v |\ T \geq v] \tag{6}$$

2.3.3 Sélection du seuil de détection

La méthode classique pour sélectionner un seuil de détection constant h repose sur l'inégalité de Wald (1945) : $MTBFA \geq e^h$. D'après Egea-Roca et al (2017), le seuil de détection h est sélectionné après avoir fixé α en respectant : $h \leq -\log(\alpha)$.

3 Contributions

3.1 Méthode empirique pour la sélection du seuil de détection

Nous supposons connaître le comportement de la statistique de détection sous l'hypothèse $H_{0,n}$: "état normal sans rupture". L'idée de notre méthode est d'effectuer un grand nombre de simulations de cette statistique sous $H_{0,n}$, et de construire le seuil de détection à partir de la loi empirique de la statistique (quantile empirique). La statistique ne le dépasse pas avec un niveau de confiance choisi à l'avance.

Seuil de détection instantané

La statistique de détection CUSUM est récursive et positive ($W_n \geq 0, \forall n \geq 1$). Même sur des observations sans rupture, l'évolution de la statistique dépend du temps : elle croit progressivement en fonction de l'accumulation des observations. Nous suggérons donc de

remplacer le seuil de détection constant par un seuil instantané qui dépend de l'instant de chaque nouvelle observation.

Pratiquement, voici les étapes que nous avons effectuées pour construire le seuil de détection instantané ($h_.$) pour la statistique W_n de CUSUM :

1. Sous $H_{0,n}$: simuler B séries de n observations $\{X_i^j\}_{i=1,..,n;\ j=1,..,B}$.
2. Pour chaque série $\{X_i^j\}_{i=1,..,n}$, calculer le maximum de la statistique jusqu'à l'instant i : $\mathrm{M}_i^j = \max_{1\leq k\leq i} W_k$.
3. Définir le seuil instantané comme le quantile empirique d'ordre (1-α) :
   $h_i = \mathbf{q}_{(1-\alpha)}\left[\left(M_i^j\right)_{j=1,..,B}\right],\quad i = 1,..,n.$

3.2 Réinitialisation de la procédure

Nous proposons une nouvelle configuration de la méthode de détection qui consiste à réinitialiser la procédure lorsque la statistique de détection revient à la valeur initiale. Nous considérons l'instant de la réinitialisation comme le début de test des données suivantes en déplaçant le seuil de détection à l'instant de chaque réinitialisation. Voici les deux configurations proposées :

Configuration sans réinitialisation : le seuil de détection instantané reste fixe tout au long de test de détection et quelque soit le comportement de la statistique de détection.

Configuration avec réinitialisation : le seuil de détection instantané se déplace à chaque fois que la statistique W_n de CUSUM revient à 0.

3.3 Règle d'arrêt corrigée

Pour signaler l'existence d'une rupture à un instant donné $n \geq 1$, la procédure de détection détaillée formellement dans la suite se base sur :

1. Calcul de la statistique de détection à l'instant $n \geq 1$ d'une façon récursive. La formule de la statistique W_n est donnée plus bas.
2. Vérification de la règle d'arrêt selon le choix de l'un des deux types suivants :

   Règle d'arrêt classique : la procédure signale l'existence d'une rupture lorsque la statistique de détection dépasse le seuil de détection instantané. C'est la règle classique de procédure CUSUM.

   Règle d'arrêt corrigée : nous proposons une nouvelle règle d'arrêt par une modification de la règle classique. La procédure corrigée signale l'existence d'une rupture lorsque la statistique de détection dépasse le seuil de détection instantané pendant un temps c ≥ 1, c : étant un paramètre à fixer à l'avance (c $= 1 \Leftrightarrow$ règle d'arrêt classique).

La méthode de détection séquentielle que nous proposons pour évaluer l'approche CUSUM, est donnée par le modèle suivant :

Statistique de CUSUM : $W_n = \max\{0, W_{n-1} + L_n\}, \quad n \geq 1 \; ; \quad W_0 = 0$

— **Règle d'arrêt classique** : $C_{(h.)} = \min_{n\geq 1}\{n; W_n \geq h_{n-k}\}$

— **Règle d'arrêt corrigée** : $C_{(h.)} = \min_{n\geq 1}\{n + c - 1; \bigcap_{j=n}^{n+c-1}(W_j \geq h_{j-k})\}$

$$k = \begin{cases} 0 & : \textbf{Sans réinitialisation} \\ \max\limits_{1\leq i\leq n}\{i-1; W_i = 0\} & : \textbf{Avec réinitialisation}. \end{cases}$$

Nous présenterons des résultats de simulations.

Bibliographie

Page, Ewan S (1954), Continuous inspection schemes, *Biometrika*, 41, 1/2, 100–115, JSTOR.

Wald, A., D. J. (1945). Sequential tests of statistical hypotheses, *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186.

Alexander G Tartakovsky, Aleksey S Polunchenko, and Grigory Sokolov (2013), Efficient computer network anomaly detection by changepoint detection methods, *IEEE Journal of Selected Topics in Signal Processing*, 7(1) :4–11.

Lorden, Gary and others (1971), Procedures for reacting to a change in distribution, *The Annals of Mathematical Statistics*, 42, 1897–1908. Institute of Mathematical Statistics.

Egea-Roca, Daniel and Seco-Granados, Gonzalo and López-Salcedo, José A (2017), Comprehensive overview of quickest detection theory and its application to GNSS threat detection, *Gyroscopy and Navigation*, vol. 8, no. 1, pp. 1–14, Springer.