A real application of the filtered derivative with false discovery rate

MOHAMED ELMI [1]

[1] *Université de Djibouti, Faculté de Science , mahamedelmifr@yahoo.fr*

Résumé. Dans ce travail, nous donnons une application réelle de la méthode de dérivée filtrée avec le taux de fausses découvertes (FDqV). La FDqV utilise deux étapes, la première étape est la dérivée filtrée et la séconde étape utilise le taux de fausses découvertes pour éliminer les fausses alarmes et récupérer uniquement les vrais instants de ruptures. La domination de la FDqV par rapport á la dérivée filtrée avec p-value est clairement établie par le critère de l'erreur quadratique de la moyenne. Ici, nous utilisons des données fournis par EDF concernant des éolionnes implantés quelques part en France, nous détectons les instants de ruptures de la vitesse du vent sur une période donnée.

Mots-clés. Series Temporelles, Dérivée Filtrée, Taux de Fausses Découvertes

Abstract. In this work, we give a real application of the method of Filtered Derivative with False Discovery Rate (FDqV)[6] . This method use the Filtered Derivative (FD)[2;3] as step 1 and a step 2 which use the False Discovery Rate [1] for elimination the false alarms at the end of step 2 and keep only as possible all right change points. The power of FDqV is provide in [6] for the criteria mean integrate square error (MISE) than Filtered Derivative with p-value (FDpV)[5]. Here we use a data given by electricity of France (EDF). It concerns wind turbines are implanted somewhere in France and we want to detect the breaks of the wind speed over a period.

Keywords. Time series, Filtered Derivative, False Discovery Rate

1 Introduction

In the literature, it exists two change points : The off-line detection or change points analysis and the on-line detection or sequential change points. Different methods for change point detection such that the penalized least square error [8], the filtered derivative [3], the filtered derivative with p-value [5] and the filtered derivative and false discovery rate [6] are used in the literature. In this work, we give an real application the filtered derivative and false discovery rate method. The rest of this paper is structured as followed : Section 2 describes the filtered derivative and false discovery rate, section 3 gives an real application of this method.

2 Recall method for change point analysis: Filtered Derivative and False Discovery Rate (FDqV)

Model

Let $\mathbf{X}= (X_1, X_2, \ldots, X_n)$ a sequence of independent random variable indexed by the time t= $1, 2, \ldots, n$. We suppose it exists a segmentation $\tau = (\tau_1, \tau_2, \ldots, \tau_K)$ with $\tau_k \in \{1, 2, \ldots, n\}$ and $\tau_1 < \tau_2 < \ldots \tau_K$. K denotes the number of changes. By convention, for the calculus of the mean, we set $\tau_o = 1$ and $\tau_{K+1} = n$. In other words, for $k = 0, \ldots K$, for i $=\tau_k + 1, \ldots \tau_{k+1}$, we have $X_k \sim \mathcal{N}(\mu_k, \sigma_k)$, where $\mathcal{N}(\mu, \sigma)$ is a gaussian law with mean μ and standard deviation σ.

The Filtered Derivative and False Discovery Rate (FDqV)

The FDqV method is introduced by [6] for a time series. In fact, The Filtered Derivative with False Discovery Rate has two steps : The Filtered Derivative and the False Discovery Rate

Step 1: The Filtered Derivative

Step 1 is based on Filtered Derivative and select a set of potential change points, More precisely, we have

Computation of the filtered derivative function

Computation of the filtered derivative function, which is defined as the difference between the estimators of the mean computed in two sliding windows respectively at the right and at the left of the time t, both of size A, that is as the function:

$$FD(t, A) = \widehat{\mu}(t + 1, t + A) - \widehat{\mu}(t - A, t), \ for \ A < t < N - A \tag{1}$$

where

$$\widehat{\mu}(t + 1, t + A) := A^{-1} \sum_{j=t+1}^{t+A} X_j$$

denote the empirical mean of the variables X_j on the box $(t + 1, t + A)$. This method consists in filtering data by computing the estimators of the parameter μ before applying a discrete derivation. So this construction explains the name of the algorithm, so called Filtered Derivative method [2;3]. Next, remark that quantities $A \times FD(t, A)$ can be iteratively calculated by using

$$A \times FD(t + 1, A) = A \times FD(t, A) + X(t + 1 + A) - 2X(t + 1) + X(t - A). \tag{2}$$

Thus, the computation of the whole function $t \longmapsto FD(t)$ for $t \in [A, n-A]$ requires $\mathcal{O}(n)$ operations and the storage of n real numbers. Let us point that the absolute value of filtered derivative $|FD|$ presents hats at the vicinity of the change points. Potential change points τ_k^*, for $k = 1, \ldots, K^*$, are selected as local maxima of the absolute value of the filtered derivative $|FD(t, A)|$ where moreover $|FD(\tau_k^*, A)|$ exceed a given threshold C_1. In [3;5], we have given the asymptotic distribution of the maximum $|FD|$ under the null hypothesis. Therefore, we can fix the error type at level $p_1^\star$, and then we can deduce the threshold C_1 corresponding to $\Pr(\max |FD(\tau_k, A)| > C_1) = p_1^*$. We can remark the existence of many local maxima in the vicinity of each right change point (see [3;5]for theoretical explanation). On the other hand, if there is no noise that is when $\sigma = 0$, we get hats of width $2A$ and hight $\mu_{k+1} - \mu_k$ at each change point τ_k,.

For this reason, we select as first potential change point τ_k^* the global maximum of the function $|FD_k(t, A)|$, then we define the function FD_{k+1} by putting to 0 a vicinity of width $2A$ of the point τ_k^* and we iterate this algorithm while $|FD_k(\tau_k^*, A)| > C_1$, see [4;5].

Step 2: The False Discovery Rate

A potential change point $\tau_k^\star$ can be an estimator of a right change point or a false alarm. We want to eliminate false detection in order to keep (as possible) only the right change points. In [6], we use as Step 2 multiple hypothesis tests. More formally, consider K hypothesis tests for all $1 \leq k \leq K$, $(H_{0,k}) : \widehat{\mu}_k = \widehat{\mu}_{k+1}$ versus $(H_{1,k}) : \widehat{\mu}_k \neq \widehat{\mu}_{k+1}$ where $\widehat{\mu}_k$'s are defined as in the model. For each hypothesis test, we calculate the p-values $p_1^\star, \ldots, p_{K^\star}^\star$ associated to each potential change point $\tau_1^\star, \ldots, \tau_{K^\star}^\star$. After the calculation of p-value, we use a Bonferroni type multiple testing procedure:

1. We tidy up p- value in the increasing order $p_{(1)}^* \leq \ldots \leq p_{(K^*)}^*$.
2. We choose a threshold q corresponding to the rate of false alarms or FDR.
3. We keep only the potential change points τ_i^* corresponding to a p-value $p_{(i)}^*$ such that $p_{(i)}^* \leq \dfrac{i}{K^*} q$.

For more details see [6].

Simulation

For $n = 10,000$, we have simulated one replication of a sequence of Gaussian random variables $(X_1, \ldots, X_n)$ with variance $\sigma^2 = 1$ and mean $\mu_t = f(t)$ where f is a piecewise constant function with seven change points at times $\tau = (2000, 2500, 3000, 4000, 7000, 8000, 9000)$ with means $\mu = (2.5, 2, 3, 4.5, 3, 3.5, 4, 5.5)$. We have made the following choices: $A = 250$, $K_{max} = 20$, $C_1 = 0.25$, and $q = 10^-6$.

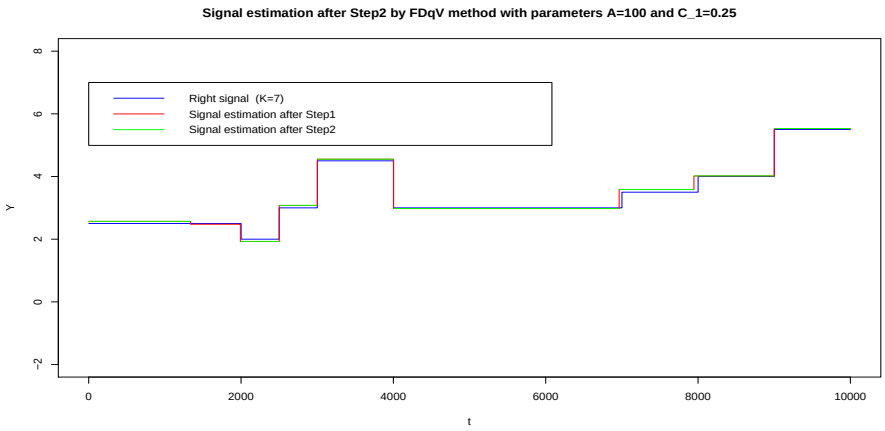

Figure 1: *Signal reconstruction after Step2 by FDqV method*

3 A real application of Filtered Derivative and False Discovery Rate

In this paragraph, we want to apply the FDqV-method for a real application. The data concerns the wind speed of the wind turbines. We have 50598 observations and we want to detect abrupt changes of the wind speed over the time which corresponds when the wind speed change. We take the parameters followings A=144, Kmax=20, $C_1 = 0.1$ and $q = 10^{-6}$. The figure 2 corresponds the signal speed wind , the figure 3 give us when the instant of potential changes are produced and the last figure is the reconstruction of the signal by our method.

Remark

The quality of the signal estimation by the FDqV- method (as the filtered derivative and the filtered derivative with p-value) depends strongly on the parameters optimization. This work will be published in [7].

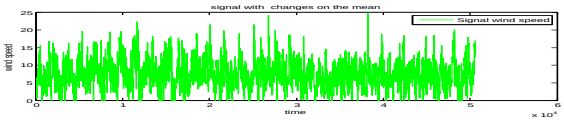

Figure 2: *The signal of wind speed*

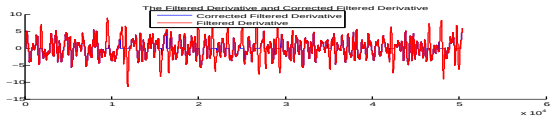

Figure 3: *The Filtered Derivative and Corrected Filtered Derivative*

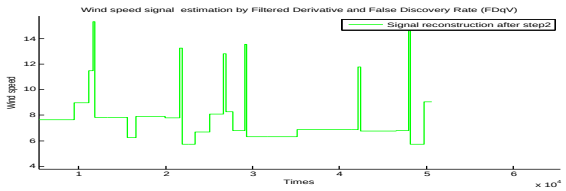

Figure 4: *Signal estimation by FDqV method*

Bibliographie

[1] Benjamini, Y and Hochberg, Y (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society. Series B. Methodological ,Vol 55 p 289–300.
[2] Benveniste, A. and Basseville, M. (1984), *Detection of abrupt changes in signals and dynamical systems: some statistical aspects*, Lecture Notes in Control and Inform. Sci. p 145–155.
[3] Basseville, M. and Nikiforov, I V. (1993), *Detection of abrupt changes: Theory and application*, Englewood Cliffs, NJ. p xxviii+528.
[4] Pierre, B, Mehdi, F, Arnaud, G (2011), Off-line detection of multiple change points by the filtered derivative with p-value method, Sequential Anal., Vol 30 n 1 p 172–207.
[5] Pierre, B, (2000), A local method for estimating change points: the "hat-function' ,Statistics. A Journal of Theoretical and Applied Statistics, Vol 34 n 3, p 215–235.
[6] ELMI, M (2014), Detection multiple change by filtered derivative and false discovery rate, International journal of statistics and probability, Vol 3 n 1 p 12–23.
[7] ELMI, M (2014),The parameters optimization of filtered derivative for change points analysis, to appear in International journal of statistics and probability, Vol 3 n 2.
[8] Lavielle, M. and Moulines, E (2000), Least-squares estimation of an unknown number of shifts in a time series, Journal of Time Series Analysis, Vol 21 p 33–59.