Available online at www.sciencedirect.com
ScienceDirect
C. R. Acad. Sci. Paris, Ser. I 346 (2008) 467–470
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistics

Change-point detection for continuous processes with high-frequency sampling

Guangming Wang [a], Samir Ben Hariz [b], Jonathan J. Wylie [c,1], Qiang Zhang [c]

[a] *School of Mathematics and Statistics, Wuhan University, Hubei, 430072, People's Republic of China*
[b] *Laboratoire de statistique et processus, département de mathématiques, université du Maine, avenue Olivier-Messiaen, 72085 Le Mans cedex 9, France*
[c] *Department of Mathematics, City University of Hong Kong, 83, Tat Chee Avenue, Kowloon, Hong Kong*

Received 29 November 2007; accepted 4 December 2007

Available online 3 March 2008

Presented by Philippe G. Ciarlet

Abstract

We consider the detection of a jump in a continuous process over a fixed time interval. We aim to locate the jump position via discrete observations and consider how increasing the frequency of the observations affects the accuracy of the detection process. We show that the classical cumulative-sum estimator fails, and propose a new estimator based on local information that we prove converges exponentially fast. ***To cite this article: G. Wang et al., C. R. Acad. Sci. Paris, Ser. I 346 (2008).***

Résumé

Détection de rupture pour des processus continus avec un échantillonnage à haute fréquence. Nous considérons la détection de saut dans un processus en temps continu sur un intervalle de temps fini et fixé. On cherche à localiser dans le temps la position de rupture via des observations discrètes. Nous étudions l'effet de croissance de la fréquence d'échantillonnage sur la précision de l'estimation. Nous montrons que la procédure classique avec les estimateurs à sommes cumulatives échoue et nous proposons une alternative basée sur une localisation de l'information. Nous prouvons que le nouvel estimateur converge avec une vitesse exponentielle. ***Pour citer cet article : G. Wang et al., C. R. Acad. Sci. Paris, Ser. I 346 (2008).***

1. Introduction

Sudden localized changes in data occur in a very broad range of applications that span image processing, signal processing, economics and the physical sciences. Given a finite set of observations, one typically needs to estimate the location at which changes occur. In this paper, we consider a continuous-time process with both dependent and

E-mail addresses: G.M.Wang@student.cityu.edu.hk (G. Wang), Samir.Ben_Hariz@univ-lemans.fr (S. Ben Hariz), mawylie@cityu.edu.hk (J.J. Wylie), mazq@cityu.edu.hk (Q. Zhang).

[1] The work of J.J. Wylie was supported by a grant from CityU (7001752).

1631-073X/$ – see front matter © 2008 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.crma.2008.02.005

independent increments over a fixed time interval with a single jump at an unknown location. Correlations are inherent in continuous processes and these correlations cause a number of challenging practical and theoretical difficulties.

The change-point problem has a wide literature that spans a number of fields. There are also a wide range of techniques including linear-model based approaches and nonparametric approaches. For a comprehensive review of the subject we refer the reader to Basseville and Nikiforov [1].

For uncorrelated processes or for independent data a number of important results have been obtained. Carlstein [4] introduced a class of nonparametric estimators based on cumulative sums for problems in which the distribution is stationary on both sides of the jump. Dumbgen [5] considered a more general class of estimators and proved an $\mathrm{O}_p(n^{-1})$ rate of convergence, where n is the number of observations.

Wang [8] proposed a method based on wavelets that considered a white-noise model in which the mean varies continuously on both sides of the jump. Wong et al. [9] considered jump detection in a heteroscedastic autoregressive model and proved the consistency of an estimator that used wavelets.

Nikiforov [6] considered generalized change detection problems for stochastic models. He provided a statistical method for both detecting and isolating changes in stochastic systems. He also considered the optimality of the approach in the sense of minimizing the worst mean detection delay. Nikiforov [7] made a detailed comparison between two strategies for both change detection and isolation. He compared optimal sequential and fixed-size sample strategies and showed that a simple nonsequential method is almost as efficient as an optimal sequential algorithm, but with less costly computation.

When there are heavy correlations in the noise the problem is significantly more challenging. Ben Hariz et al. [3] considered data with long-range dependence where correlations decay to zero algebraically or faster. They showed that the classical cumulative-sum (CUSUM) estimators also achieve the $\mathrm{O}_p(n^{-1})$ rate of convergence in this case.

To our best knowledge, previous works have focussed on processes with independent increments, weakly dependent structure, or at least with correlations structures that tend to zero. They hence do not apply to situations in which there is a finite interval and the sampling frequency tends to infinity.

There are a number of important applications related to high-frequency sampling in which correlations between neighboring points do not decay to zero. For this class of problems we show that the traditional cumulative-sum and wavelet-based estimators may not yield consistent estimation. We propose an alternative estimator that is based on local averages and show that the estimator performs best when it uses localized information. For Gaussian data we derive a bound for the probability of the estimator selecting an incorrect change-point location and show that this probability tends to zero exponentially fast as the sampling rate tends to infinity. Detailed proofs of these results can be found in Ben Hariz et al. [2].

2. Framework and methodology

We consider the following model

$$Y_t = \delta \mathbb{1}_{\{t>\theta\}} + X_t, \quad t \in [0, T], \tag{1}$$

where δ is the size of the jump and $\theta \in (0, T)$ is the location of the jump. We assume that $(X_t)_{0\leqslant t\leqslant T}$ is a random process with a constant mean, which, without loss of generality, is taken to be zero. We also assume that the process (X_t) satisfies the following regularity condition

$$\exists \alpha > 0, C > 0, \text{ such that for } s, t \in [0, T], \quad \mathbb{E}(X_s - X_t)^2 \leqslant C|s-t|^{\alpha}. \tag{2}$$

This is a typical assumption to ensure the continuity of the process.

Observe that here the main concern is the local fluctuations of X rather than the long-range correlations of X. This is in sharp contrast with the common assumption that the random part of the data is either a white noise process, a weakly dependent process or a process with correlations that tend to zero as the distance between points tends to infinity.

Clearly, the process can only be observed at a finite number of grid points. For the sake of simplicity we consider the case $T = 1$ with uniformly distributed grid points at $t_1 = T/n, \ldots, t_k = kT/n, \ldots, t_n = T$. Given the observations $(Y_{t_i})_{1\leqslant i\leqslant n}$, we aim to estimate within which interval $[\hat{\theta}, \hat{\theta} + 1/n[$ the change-point lies. Here $\hat{\theta}$ is determined by a localized version of the cumulative-sum estimator

$$\hat{\theta} = \frac{1}{n} \min\Bigl(\arg \max_{1\leqslant k<n} \bigl\{|U_k|\bigr\}\Bigr), \tag{3}$$

with

$$U_k = \frac{1}{k-k_l+1}\sum_{i=k_l}^{k} Y_i - \frac{1}{k_u-k}\sum_{i=k+1}^{k_u} Y_i, \quad k=1,\dots,n-1, \tag{4}$$

where $k_l = \max(1, k-L+1)$ and $k_u = \min(k+L, n)$.

The implementation of these estimators is straight-forward, namely for a given continuous process, we sample the data uniformly at the points $t_i = i/n$, $i = 1,2,\dots,n$, and consider each interval $[t_i, t_{i+1}[$, $i = 1,2,\dots,n-1$, as a candidate for the segment of the process within which the change-point lies. We take a window with $2L$ observations with an obvious modification near the end points of the data set (see Eq. (4)), and determine the difference between the mean of Y based on the data from the first L points and that from the second L points. The algorithm identifies the interval which gives the maximum of the absolute value of the difference as the segment where the change point lies. The classical CUSUM estimator takes the whole set as the window size, i.e., L is equal to n.

3. Main results

We have performed numerical tests of processes with correlation functions that satisfy (2). The simulations show that the classical CUSUM change-point estimator that one obtains by setting $L = n$ in (4) does not yield consistent results. The results also show that smaller values of L yield more accurate estimators. This suggests that choosing $L = 1$ will be best choice since it gives the smallest estimation error. Therefore, in the rest of this Note, we will only consider the case of $L = 1$ and develop some surprising theoretical results for this apparently simple choice of L.

Before progressing to the next section in which we present our main theoretical results, we provide some simple intuitive explanation about why the conventional approach based on whole data set, that is $L = n$, fails and why $L = 1$ is a better choice.

The classical change-point problem corresponds to the situation in which the data is zero correlated. In this case, choosing larger value of L will reduce the statistical errors in the estimates for the empirical means in (4). If one considers correlated processes in which the correlations decay to zero a similar argument holds. These processes correspond to situations in which the frequency is fixed, but the time over which the process is recorded tends to infinity. For the class of problems considered in this Note, the time over which the process is recorded is fixed, but the sampling frequency increases. As one increases the frequency the correlations among all sampled data will not tend to zero. These correlations imply that the statistical errors in (4) do not decay. In this situation, as n tends to infinity, the data between the neighboring points becomes almost perfectly correlated. This is due to the fact the noise between neighboring points becomes almost perfectly correlated. Therefore the difference between data from the neighboring points will be very small except across the change point where the difference will be dominated by the jump. Therefore $L = 1$ will be the best choice when we consider high frequency correlated data. As one increases L, the noise of the data within the window containing the $2L$ observations becomes less correlated, and consequently the performance of the estimator deteriorates.

In the following theorem, we state an exponential bound for the probability of missing the change-point location and give some examples:

Theorem 3.1. *Assume that (X_t) is a Gaussian process. Then we have*

$$\mathbb{P}\left(\theta \notin \left[\hat{\theta}, \hat{\theta} + \frac{1}{n}\right]\right) \leqslant 2(n-1)\mathbb{P}\left(N(0,1) \geqslant \frac{\delta}{\sqrt{\varepsilon_n}}\right), \tag{5}$$

where $\varepsilon_n := \max_{i \neq k}(\mathrm{Var}(Z_i - Z_k), \mathrm{Var}(Z_i + Z_k))$, $Z_i := X_{i/n} - X_{(i+1)/n}$ and $k := [n\theta]$.

In particular,

(i) *If there exist constants $\alpha > 0$, $C > 0$, such that for $s, t \in [0, T]$, $\mathbb{E}(X_s - X_t)^2 \leqslant C|s-t|^\alpha$, then $\varepsilon_n \leqslant 4Cn^{-\alpha}$.*

(ii) *If $\mathbb{E}(X_s - X_t)^2 \leqslant C|s-t|^\alpha$ and for $s_1 \leqslant s_2 \leqslant t_1 \leqslant t_2$, $|\mathbb{E}(X_{s_2} - X_{s_1})(X_{t_2} - X_{t_1})| \leqslant C_1|s_2 - s_1|^{\alpha_1}|t_2 - t_1|^{\alpha_2}$ for $\alpha_1 + \alpha_2 > \alpha$, then $\varepsilon_n \leqslant 2Cn^{-\alpha} + \mathrm{o}(n^{-\alpha})$.*

Remark. If X is weak stationary and $|\mathrm{corr}(X_s, X_t) - 1| \leqslant C|s-t|^\alpha$, then (i) is satisfied. Condition (ii) is satisfied when (X_t) has independent increments.

In summary, we have considered the detection of change-points for continuous processes over a fixed time interval. We have shown that as the frequency of observations increases the classical CUSUM estimator fails. We have proposed an alternative, extremely simple, estimator. For Gaussian processes with natural conditions to control the regularity of the process, we have derived a bound for the probability of missing the change-point and shown that our estimator has exponentially fast convergence.

References

[1] M. Basseville, I.V. Nikiforov, Detection of Abrupt Changes: Theory and Application, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[2] S. Ben Hariz, G.M. Wang, J.J. Wylie, Q. Zhang, Jump detection for discretely observed continuous processes, Preprint, 2007.

[3] S. Ben Hariz, J.J. Wylie, Q. Zhang, Optimal rate of convergence for nonparametric change-point estimators for non-stationary sequences, Ann. Statist. 35 (2007) 1802–1826.

[4] E. Carlstein, Nonparametric change-point estimation, Ann. Statist. 16 (1988) 188–197.

[5] L. Dumbgen, The asymptotic behavior of some nonparametric change-point estimators, Ann. Statist. 19 (1991) 1471–1495.

[6] I.V. Nikiforov, A generalized change detection problem, IEEE Trans. Inform. Theory 41 (1) (1995) 171–187.

[7] I.V. Nikiforov, Two strategies in the problem of change detection and isolation, IEEE Trans. Inform. Theory 43 (2) (1997) 770–776.

[8] Y. Wang, Jump and sharp cusp detection by wavelets, Biometrika 82 (2) (1995) 385–397.

[9] H. Wong, W. Ip, Y. Li, Detection of jumps by wavelets in a heteroscedastic autoregressive model, Statist. Probab. Lett. 52 (4) (2001) 365–372.