Available online at www.sciencedirect.com

ScienceDirect

C. R. Acad. Sci. Paris, Ser. I 347 (2009) 425–428

COMPTES RENDUS MATHEMATIQUE

Statistics

Detection of change-points near the end points of long-range dependent sequences

Weilin Nie[a], Samir Ben Hariz[b], Jonathan Wylie[c], Qiang Zhang[c,1]

[a] *Department of Mathematics and statistics, Wuhan University, Wuhan, 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, Kowloon Tong, Hong Kong*

Received 5 December 2008; accepted 8 January 2009

Available online 17 March 2009

Presented by Philippe G. Ciarlet

Abstract

We consider a sequence of observations $(X_i)_{i=1,\dots,n}$ with a marginal distribution that is given by $\mathscr{L}(X_i) = P_n$ if $i \leqslant n\theta_n$ and $\mathscr{L}(X_i) = Q_n$ if $i > n\theta_n$. The parameter $0 < \theta_n < 1$ is the location of the change-point which must be estimated and may depend on the sequence length. We consider the general case in which the change-point can converge to one of the end-points of the interval [0, 1] as the sequence length n tends to infinity. The sequence can be long-range dependent, short-range dependent or independent and may be non-stationary. We study a class of non-parametric estimators and prove they are consistent and that the rate of convergence is $1/n$. We also deal with the case in which the distance between the distributions P_n and Q_n tends to zero as n tends to infinity. ***To cite this article: W. Nie et al., C. R. Acad. Sci. Paris, Ser. I 347 (2009).***

Résumé

Détection de rupture près des extrémités pour des suites fortements dépendantes. On considère une suite d'observations $(X_i)_{i=1,\dots,n}$ avec des lois marginales vérifiant : $\mathscr{L}(X_i) = P_n$ pour $i \leqslant n\theta_n$ et $\mathscr{L}(X_i) = Q_n$ pour $i > n\theta_n$. Le paramètre θ_n, qui peut dépendre de la taille n de la suite d'observations, désigne la localisation du changement dans la loi marginale. On s'intéresse ici à l'estimation de ce paramètre. On considère le cas général où la position de rupture θ_n peut converger vers l'une des deux extrémités de l'intervalle [0, 1] lorsque la longueur de la suite tend vers l'infini. La suite peut être fortement dépendante, faiblement dépendante ou indépendante, voire même non stationnaire. On étudie une classe d'estimateurs non-paramètriques. On prouve qu'ils sont consistants et que leur vitesse de convergence est de $1/n$. On traite aussi le cas où la distance entre les distributions P_n et Q_n tend vers 0 quand n tend vers l'infini. ***Pour citer cet article : W. Nie et al., C. R. Acad. Sci. Paris, Ser. I 347 (2009).***

E-mail address: shariz@univ-lemans.fr (S. Ben Hariz).

[1] This work was supported by the Research Grant Council of Hong Kong Special Administrative Region, China, Project Cityu 103807.

1631-073X/$ – see front matter © 2009 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.crma.2009.02.002

1. Introduction

The change-point problem involves a finite sequence of random variables that has a change in the underlying distribution at an unknown location in the sequence. Given a sequence, one must use the data to estimate the location of the change in distribution. Such problems occur widely and are critical in an enormous number of applications and the change-point problem has come to be seen as a classical problem in statistics. For a detailed background of the topic we refer the reader to the monograph by Csorgo and Horvath [4].

Most of the previous work has focussed on the case in which the number of observations before and after the change-point are of the same order of magnitude. However, in general, the change-point can occur at any point of the sequence. In particular, it is possible that the change-point is close to one of the end points of the sequence. In this Note, we study exactly this problem for sequences with very general dependence structures, namely long-range dependent, short-range dependent or independent.

Extensive work has been done for independent sequences in the case that the number of observations before and after the change-point are of the same order of magnitude. In a parametric setting, Hinkley [9] showed that the maximum likelihood estimator has a rate of convergence $\mathrm{O}_p(n^{-1})$. The non-parametric setting was considered by Carlstein [3] who considered a class of estimators and derived a rate of convergence. These results were later improved by Dumbgen [5] who considered a very general framework that includes a wide range of non-parametric estimators and showed that the rate of convergence is $\mathrm{O}_p(n^{-1})$. Dumbgen also derived the rate of convergence in the case in which the difference between the distributions before and after the change-point tends to zero. In a number of papers, Ferger [6–8] considered almost sure rates of convergence and probability bounds.

In recent years, the importance of long-memory or long-range dependence (LRD) sequences has been realised in a number of applications. Examples include the analysis of telecommunication and financial data. There is no universally agreed definition of what exactly constitutes a LRD sequence, but in this Note, we will consider a sequence $(X_i)_{i=1,\dots,n}$ to be short-range dependent (SRD) if

$$\limsup_{n\to\infty} n^{-1}\mathbb{E}\left[\sum_{i=1}^{n}\bigl(X_i-\mathbb{E}[X_i]\bigr)\right]^2<\infty$$

and LRD otherwise. Several works have generalised the results for change-point detection for independent sequences to the SRD setting. There are a number of technical and theoretical challenges when dealing with LRD sequences that make the problem significantly harder than the SRD and independent cases and there are fewer results in this case. Kokoszka and Leipus [11] considered the parametric case for a change in the mean and gave a rate of convergence that gets worse as the dependence becomes more long-range. Horvath and Kokoszka [10] considered a similar problem to that considered by Kokoszka and Leipus [11] in which the size of the jump in the mean tends to zero as the length of the sequence increases. In this case they were able to determine the limiting distribution. Ben Hariz and Wylie [1] considered the parametric case and showed that the rate of convergence does not get worse as the dependence becomes more long range and showed that the $\mathrm{O}_p(n^{-1})$ rate of convergence for independent sequences is also achieved for both SRD and LRD sequences. Ben Hariz et al. [2] considered a very general class of dependence structures that includes independent, SRD and LRD sequences and makes no assumption about stationarity in the dependence structure. They proved the consistency of a class of estimators similar to that proposed by Dumbgen and showed that the rate of convergence for SRD and LRD sequences is the same as that for independent sequences, that is, $\mathrm{O}_p(n^{-1})$. They also derived the rate of convergence for the case in which the difference between the distributions before and after the change-point tends to zero and found that, in this case, the rate can depend on the dependence structure.

The simplest setting for change-point detection is when the distributions before and after the jump are fixed and the location of the change-point is also fixed. However, in the most general case, the difference between the distributions can tend to zero and the change-point location can tend to one of the end points of the sequence namely for sequential change-point estimation. As mentioned above, the case in which the difference between the distributions before and after the jump tends to zero has been widely studied for both independent and dependent sequences. However, in all of the references mentioned above the focus has been on the case in which the number of observations before and after the change-point are of the same order of magnitude. Yao et al. [13] studied independent sequences and considered the case in which the location of the change-point tends towards one of the end points, but the difference between the distributions before and after the jump does not tend to zero. They derived an almost sure $\mathrm{O}(n^{-1})$ rate under the condition that the smaller of the pre-change and post-change sample sizes was larger than a constant multiplied by $\log(n)$. A natural question that arises is whether a similar result can be obtained for dependent sequences. It is also of interest to consider the case when the location of the change-point tends towards one of the end points and the difference between the distributions before and after the jump simultaneously tend to zero as the sequence length increases. In this Note, we address exactly these questions.

We use a unified framework that can deal with independent, SRD and LRD sequences to consider the problem when the change-point tends towards one of the end points allowing for the possibility that the difference between the distributions before and after the jump can tend to zero. Under very weak conditions on the dependence structure, we show that the $\mathrm{O}_p(n^{-1})$ rate can be obtained under the condition that the smaller of the pre-change and post-change sample sizes tends to infinity, no matter how slowly. This condition is less restrictive than that considered by Yao et al. [13]. Moreover, the result applies to independent, SRD and LRD sequences. Detailed proofs of these results can be found in Nie et al. [12].

2. Main results

Let $(X_i)_{i=1,\dots,n}$ be a sequence in a measurable space E defined on $(\Omega, \mathcal{A}, \mathbb{P})$. The marginal distribution (which may depend on the sequence length n) is given by

$$\mathscr{L}(X_i) = \begin{cases} P_n, & i \leqslant k_n; \\ Q_n, & i > k_n, \end{cases}$$

where $k_n := n\theta_n$, and $0 < \theta_n < 1$ is the location of the change-point. This means that we assume first-order stationarity on either side of the change-point, but make no assumption about stationarity in the dependence structure of the sequence. We will consider the case in which the location of the change-point θ_n tends to either 0 or 1 as the sequence length n tends to infinity.

Given the sequence, $(X_i)_{i=1,\dots,n}$, we aim to estimate the location of the change-point θ_n using an estimator of the following general type:

$$\hat{\theta}_n = \frac{1}{n} \min\Big(\operatorname{argmax}_{1 \leqslant k < n} \big\{N(D_k)\big\}\Big), \tag{1}$$

where N is a (possibly random) semi-norm on the space $\mathcal{M}$ of signed finite measures on E,

$$D_k = \left[\frac{k}{n}\left(1 - \frac{k}{n}\right)\right]^{1-\gamma} \left(\frac{1}{k}\sum_{i=1}^{k} \delta_{X_i} - \frac{1}{n-k}\sum_{i=k+1}^{n} \delta_{X_i}\right), \tag{2}$$

and γ is a parameter satisfying $0 \leqslant \gamma < 1$.

This class of estimators considers each location as a candidate for the location of the change-point. For each candidate, the estimator, compares the difference between the empirical distributions before and after the candidate and selects the maximum weighted difference. The weighting factor $[k/n(1 - k/n)]^{1-\gamma}$ is for the purpose of penalizing candidates close to the end-points which give rise to large statistical errors in the empirical distributions.

In what follows we state the main results. The first result deals with the consistency of the estimator while the second result gives a rate of convergence.

Assumption 1. We assume that there exists a countable family of real valued functions $\mathcal{F}$ defined on E such that

A1: There exist constants $C > 0$ and $0 < \rho < 1$ such that

$$\sup_{f \in \mathcal{F}} \sup_{1 \leqslant k \leqslant n-m} \left| E\left[\left(\sum_{i=k}^{k+m} f(X_i) - E\big(f(X_i)\big)\right)^2\right]\right| \leqslant C m^{2-\rho} \|f\|^2. \tag{3}$$

A2: The norm N satisfies

$$N(\nu) \leqslant \sum_{f \in \mathcal{F}} d(f) \big|\nu(f)\big|, \tag{4}$$

where $\nu(f) \equiv \int_E f(x)\nu(\mathrm{d}x)$ and $d(f)$ are positive constants such that $\sum_{f \in \mathcal{F}} d(f)\|f\| < \infty$.

Theorem 1. *Under the conditions of Assumption* 1, *we assume that there exists a positive sequence b_n such that*

$$\mathbb{P}\big[N(P_n - Q_n) > b_n\big] \to 1 \quad \text{as } n \to \infty. \tag{5}$$

$$m_n \to +\infty \quad \text{as } n \to \infty, \tag{6}$$

$$b_n^{-1}\big[n^{-\rho/2-\gamma+1} m_n^{\gamma-1} + (m_n)^{\gamma-1} + (m_n)^{-\rho/2}\big] \to 0, \tag{7}$$

where $m_n = \min(n\theta_n, n(1-\theta_n))$. Then the estimator is consistent in probability

$$\hat{\theta}_n - \theta_n = \mathrm{o}_p(1). \tag{8}$$

The following theorem gives a rate of convergence of the estimator:

Theorem 2. *Under the assumptions of Theorem* 1, *we further assume that $2\gamma - 2 + \rho > 0$, then*

$$\hat{\theta}_n - \theta_n = \mathrm{O}_p\big(n^{-1} b_n^{-2/\rho}\big). \tag{9}$$

In particular, if $b_n > b > 0$, then

$$\hat{\theta}_n - \theta_n = \mathrm{O}_p\big(n^{-1}\big).$$

Remark. Theorem 2 gives the optimal rate of convergence, namely $1/n$, even when θ_n is near one of the end points provided that the sample size before and after the change tends to infinity, no matter how slowly. It is easy to show that if b_n is bounded away from zero and the condition $2\gamma - 2 + \rho > 0$ in Theorem 2 is satisfied, then condition (7) in Theorem 1 is automatically satisfied.

References

[1] S. Ben Hariz, J.J. Wylie, Rates of convergence for the change-point estimator for long-range dependent sequences, Statist. Probab. Lett. 73 (2005) 155–164.

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

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

[4] M. Csörgő, L. Horváth, Limit Theorems in Change-Point Analysis, Wiley, Chichester, 1997.

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

[6] D. Ferger, On the rate of almost sure convergence of Dumbgen's change-point estimators, Statist. Probab. Lett. 19 (1995) 27–31.

[7] D. Ferger, Exponential and polynomial tailbounds for change-point estimators, J. Statist. Plann. Inference 92 (2001) 73–109.

[8] D. Ferger, Boundary estimation based on set-indexed empirical processes, Nonparametric Statist. 16 (2004) 245–260.

[9] D.V. Hinkley, Inference about the change-point in a sequence of random variables, Biometrika 57 (1970) 1–17.

[10] L. Horváth, P. Kokoszka, The effect of long-range dependence on change-point estimators, J. Statist. Plann. Inference 64 (1997) 57–81.

[11] P. Kokoszka, R. Leipus, Change-point in the mean of dependent observations, Statist. Probab. Lett. 40 (1998) 385–393.

[12] W.L. Nie, S. Ben Hariz, J.J. Wylie, Q. Zhang, Detection of change-point near the end points of long-range dependent sequences, Preprint, 2008.

[13] Y.-C. Yao, D. Huang, R.A. Davis, On almost sure behavior of change-point estimators, in: E. Carlstein, H.-G. Müller, D. Siegmund (Eds.), Change-Point Problems, IMS, Hayward, CA, 1994, pp. 359–372.