Journal of Statistical Planning and Inference 140 (2010) 2003–2017

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Multiple change-point estimation with U-statistics

Maik Döring

Institute for Applied Mathematics and Statistics, University of Hohenheim, Schloss Hohenheim, 70599 Stuttgart, Germany

ARTICLE INFO

Article history:
Received 21 November 2009
Accepted 26 January 2010
Available online 4 February 2010

Keywords:
Change-point estimator
U-statistic
Consistency
Rate of convergence

ABSTRACT

We consider a multiple change-point problem: a finite sequence of independent random variables consists of segments given by a known number of the so-called change-points such that the underlying distribution differs from segment to segment. The task is to estimate these change-points under no further assumptions on the within-segment distributions. In this completely nonparametric framework the proposed estimator is defined as the maximizing point of weighted multivariate U-statistic processes. Under mild moment conditions we prove almost sure convergence and the rate of convergence.

1. Introduction and main results

Change-point models have originally been developed in connection with applications in quality control, where a change from the in-control to out-of-control state has to be detected based on the available observation. Change-point models are being used in many fields, for instance, archaeology, econometrics, epidemiology, medicine and reliability. An overview of the existing contribution in the literature, including many examples can be found in monographs by Brodsky and Darkhovsky (1993, 2000), Basseville and Nikiforov (1993), Antoch et al. (1999), and Chen and Gupta (2000). For example, Braun and Mueller (1998) have used multiple change-point models for the segmentation of DNA sequences. These sequences are long chain-like molecules composed of four nucleic acids, or bases. The bases are adenine (A), guanine (G), cytosine (C) and thymine (T). The observations along the sequence take on one of the values of the DNA alphabet (A, C, G or T). Braun and Mueller suppose that there are segments within which the observations follow the same or nearly the same distribution, and between which observations have different distributions. Interest may lie in describing the structure of the sequence, in detecting segments which are anomalous (in the sense that they are either mistakenly included in the sequence under consideration or perhaps derive from some other organizational scheme), or in comparing structures between sequences.

We consider a change-point model for independent random variables, where the number of change points $q \in \mathbb{N}$ is assumed to be known. So let $X_{1,n}, \dots, X_{n,n}$, $n \in \mathbb{N}$ be a triangular array of row wise independent random variables defined on a common probability space $(\Omega, \mathfrak{A}, P)$ with values in a measurable space $(\mathfrak{X}, \mathfrak{F})$. The multiple change-point is given by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q) \in H$, where

$$H := \{\mathbf{t} \in \mathbb{R}^q : 0 < t_1 < \cdots < t_q < 1\}.$$

We denote by $[s] := \max(k \in \mathbb{Z}, k \le s)$ for $s \in \mathbb{R}$ and $[\mathbf{t}] := ([t_1], \dots, [t_q])$ for $\mathbf{t} = (t_1, \dots, t_q) \in \mathbb{R}^q$. Further, we assume that there exist measures ν_i for $0 \le i \le q$ such that

$$P \circ X_{j,n}^{-1} = \nu_i \quad \text{for} [n\theta_i] < j \le [n\theta_{i+1}], \;\; 0 \le i \le q \text{ where } \theta_0 = 0, \theta_{q+1} = 1.$$

E-mail address: maik.doering@uni-hohenheim.de

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.01.040

The goal is to estimate the unknown multiple change-point $\boldsymbol{\theta}$. Here, nothing is known about the underlying distributions ν_i except that $\nu_{i-1} \neq \nu_i$ for $1 \leq i \leq q$.

U-statistics in change-point analysis have been introduced by Csörgö and Horváth. They use U-statistics to detect a change in the distribution in a sequence of independent real valued random variables. Further details can be found in Csörgö and Horváth (1997). There are various approaches in the literature dealing with U-statistics in change point analysis. See for example Horváth and Hušková (2005), Orasch (2004), Ferger (1994, 1995, 2001), Gombay (2000, 2001) and Aly and Kochar (1997).

The basic idea of the proposed estimator for the unknown multiple change-point is the segmentation of the n observations in $q+1$ subsamples for any possible change point configuration $\mathbf{t} \in H$. The i-th sample consists of the random variables $X_{[nt_i]+1,n}, \ldots, X_{[nt_{i+1}],n}$ for $0 \leq i \leq q$, where $t_0=0$ and $t_{q+1}=1$. The estimator is defined as the maximizing point over H of a weighted $(q+1)$-sample U-statistic with a chosen kernel h of degree $\mathbf{m}=(m_0,\ldots,m_q)$, where $m_i \in \mathbb{N}_0$, $m := \sum_{i=0}^{q} m_i > 0$ and $h : \mathfrak{X}^m \to \mathbb{R}$ is a $\mathfrak{F}^m$-$\mathfrak{B}(\mathbb{R})$- measurable function. It is assumed that the kernel h is symmetrical in each of the m_i coordinates and a suitable integrable, i.e.

$h(x_{1,0},\ldots,x_{m_0,0}\ldots,x_{1,q},\ldots,x_{m_q,q}) = h(x_{\pi_0(1),0},\ldots,x_{\pi_0(m_0),0}\ldots,x_{\pi_q(1),q},\ldots,x_{\pi_q(m_q),q})$ for all $\mathbf{x} \in \mathfrak{X}^m$ and all permutations π_i of m_i coordinates.

There exists a real number $1 \leq p < \infty$ such that

$$M_p := \max_{\substack{\mathbf{k}\in\mathbb{N}_0^{q+1} \\ \sum_{i=0}^{q} k_i = m}} \int_{\mathfrak{X}^m} |h(x_1,\ldots,x_m)|^p \prod_{i=0}^{q} \prod_{j=\left(\sum_{l=0}^{i-1} k_l\right)+1}^{\left(\sum_{l=0}^{i-1} k_l\right)+k_i} \nu_i(dx_j) < \infty.$$

We adopt the conventions that $\sum_{i\in\emptyset} a_i = 0$ and $\prod_{i\in\emptyset} a_i = 1$. We define for $n \in \mathbb{N}$ and $\mathbf{t} \in H_{\mathbf{m},n}$, where

$$H_{\mathbf{m},n} := \{\mathbf{t} \in H : m_i \leq [nt_{i+1}]-[nt_i] \text{ for } 0 \leq i \leq q \text{ where } t_0 = 0, t_{q+1} = 1\},$$

a $(q+1)$-sample U-statistic $U_{n,\mathbf{t}}(h)$ by

$$U_{n,\mathbf{t}}(h) := \prod_{i=0}^{q} \binom{[nt_{i+1}]-[nt_i]}{m_i}^{-1} \sum_{1 \leq j_1^0 < \cdots < j_{m_0}^0 \leq [nt_1]} \cdots \sum_{[nt_i] \leq j_1^i < \cdots < j_{m_i}^i \leq [nt_{i+1}]} \cdots \sum_{[nt_q]+1 \leq j_1^q < \cdots < j_{m_q}^q \leq n} h\left(X_{j_1^0,n},\ldots,X_{j_{m_0}^0,n},\ldots,X_{j_1^q,n},\ldots,X_{j_{m_q}^q,n}\right),$$

where $t_0=0$ and $t_{q+1}=1$. We endow this multivariate U-process with a weight function $w : \mathbb{R}^q \to \mathbb{R}$ defined for an $\boldsymbol{\alpha} \in \mathbb{R}^{q+1}$ with $\frac{1}{2} < \alpha_i$, $0 \leq i \leq q$ by

$$w(\mathbf{t}) := \begin{cases} \prod_{i=0}^{q} (t_{i+1}-t_i)^{\alpha_i}, & \mathbf{t} \in H, \\ 0 & \text{otherwise}, \end{cases}$$

where $t_0=0$ and $t_{q+1}=1$. We will see that in the context of change-point estimation it is also worthwhile to work with weight functions. They are used to overcome boundary effects, which typically occur when any of the distances $\theta_{i+1}-\theta_i$ for $0 \leq i \leq q$, where $\theta_0 = 0$ and $\theta_{q+1} = 1$, is close to zero. Csörgö and Horváth originally introduced weight functions to improve the power of their tests. We define a sequence of stochastic processes $(\rho_n)_{n\in\mathbb{N}}$ with $\rho_n = \{\rho_n(\mathbf{t}) : \mathbf{t} \in \mathbb{R}^q\}$ by

$$\rho_n(\mathbf{t}) := \begin{cases} w\left(\frac{[nt_1]}{n},\ldots,\frac{[nt_q]}{n}\right) U_{n,\mathbf{t}}(h), & \mathbf{t} \in H_{\mathbf{m},n}, \\ 0 & \text{otherwise}. \end{cases}$$

For $n \in \mathbb{N}$ we introduce a class of estimators for the unknown change-point $\boldsymbol{\theta}$. By definition they are maximizing points of the weighted U-process $|\rho_n|$ in the region $G_n := \{(k_1/n,\ldots,k_q/n) \in H_{\mathbf{m},n} : k_i \in \mathbb{N}_0, 1 \leq i \leq q\}$, that is

$$\hat{\boldsymbol{\theta}}_n := \operatorname*{argmax}_{\mathbf{t}\in G_n} |\rho_n(\mathbf{t})|.$$

The estimator $\hat{\boldsymbol{\theta}}_n$ depends on the chosen kernel, $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(h)$. The effect of the distribution ν_i can be traced back to integrals of the kernel h. The quality of the estimator depends on the chosen kernel h and its order of integration p.

The case $q=1$ and $\mathbf{m}=(1,1)$ has been investigated by Ferger (1994, 1995, 2001), whereas $q=1$ and $\mathbf{m}=(2,2)$ has been treated in Döring (2004) for the special kernel $h(x_1, x_2, y_1, y_2) = g(x_1, x_2) - g(y_1, y_2)$, where $g : \mathfrak{X}^2 \to \mathbb{R}$ is a measurable and antisymmetric function, i.e. $g(x_1, x_2) = -g(x_2, x_1)$. Orasch (2004) analyses tests which are designed for the detection of multiple change-points. He considers the case $m_i=1$, $0 \leq i \leq q$ with the kernel $h(x_0,\ldots,x_q) = \sum_{i=0}^{q-1} \sum_{j=i+1}^{q} g(x_i, x_j)$, where $g : \mathfrak{X}^2 \to \mathbb{R}$ is a measurable, symmetric or antisymmetric function.

Choices for the kernel are $h(x_0,\ldots,x_q) = \sum_{i=1}^{q} g(x_{i-1}, x_i)$ with $m_i=1$, $0 \leq i \leq q$, where g is a kernel function from a single change point problem, see for example Ferger (2001). Another choice of an appropriate kernel is a kernel, which is the sum