Stochastic Processes and their Applications 42 (1992) 345-351
North-Holland

Convergence of changepoint estimators

Dietmar Ferger and Winfried Stute

Mathematical Institute, University of Giessen, Germany

Received 15 October 1990
Revised 30 April 1991

Let $X_1^n, \ldots, X_n^n$ be an array of independent random vectors such that $X_1^n, \ldots, X_{[n\theta]}^n$ have distribution function F, and $X_{[n\theta]+1}^n, \ldots, X_n^n$ have distribution function G with $F \neq G$. In this paper we propose an estimator θ_n of the changepoint θ and show that $n(\theta_n - \theta) = O(\ln n)$ with probability one.

AMS 1980 Subject Classifications: Primary 62G05; Secondary 60F15.

changepoint estimator * exponential tail bound * almost sure convergence

1. Introduction

Consider a triangular array $X_1^n, \ldots, X_n^n$, $n \geq 1$, of rowwise independent random vectors in $\mathbb{R}^d$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Suppose that for some $0 < \theta < 1$, $X_1^n, \ldots, X_{[n\theta]}^n$ have distribution function (d.f.) F, and $X_{[n\theta]+1}^n, \ldots, X_n^n$ have d.f. G, with $F \neq G$, both unknown. θ resp. $[n\theta]$ is called the changepoint for the underlying d.f. The problem of estimating θ has found much interest in the literature. A review of available nonparametric methods is contained in Csörgő and Horváth (1988b). The fundamental idea in all these approaches is to compare, for each $0 \leq t \leq 1$, the subsamples $X_1^n, \ldots, X_{[nt]}^n$ and $X_{[nt]+1}^n, \ldots, X_n^n$. E.g., Darkhovskh (1976) and Carlstein (1988) considered the empirical functions

$$ {}_th^n(x) = n^{-1} \sum_{i=1}^{[nt]} 1_{\{X_i^n \leq x\}}, \qquad h_t^n(x) = n^{-1} \sum_{i=[nt]+1}^{n} 1_{\{X_i^n \leq x\}}, $$

of the subsamples $X_1^n, \ldots, X_{[nt]}^n$ and $X_{[nt]+1}^n, \ldots, X_n^n$, respectively. For each fixed $0 \leq t \leq 1$ one may apply Kiefer's (1961) exponential bound together with Borel-Cantelli to get

$$ \sup_x |{}_th^n(x) - {}_th(x)| \to 0 \quad \text{and} \quad \sup_x |h_t^n(x) - h_t(x)| \to 0 $$

with probability one. Here,

$$ {}_th(x) = 1_{\{t \leq \theta\}} tF(x) + 1_{\{t > \theta\}} [\theta F(x) + (t - \theta) G(x)] $$

Correspondence to: Prof. Dr. Winfried Stute, Mathematisches Institut, Justus-Liebig-Universität, Arndtstrasse 2, W-6300 Giessen, Germany.

Work supported by the 'Deutsche Forschungsgemeinschaft'.

and

$$h_t(x) = 1_{\{t \leq \theta\}}[(\theta - t)F(x) + (1-\theta)G(x)] + 1_{\{t > \theta\}}(1-t)G(x).$$

Carlstein (1988) compared the vectors (appropriately weighted)

$${}_tD^n = ({}_th^n(X_i^n))_{1 \leq i \leq n} \quad \text{and} \quad D_t^n = (h_t^n(X_i^n))_{1 \leq i \leq n}$$

and chose θ_n to be that value t in $\Delta_n = \{1/n, 2/n, \ldots, (n-1)/n\}$ for which $S({}_tD^n, D_t^n)$ is maximal. Here S is a suitable norm on the n-dimensional Euclidean space. He proved that with probability one

$$|\theta_n - \theta| = o(n^{-\delta}) \quad \text{for each } \delta < \tfrac{1}{2}.$$

Dümbgen (1990) derived an in-probability statement, namely

$$|\theta_n - \theta| = O(n^{-1}) \quad \text{in probability.}$$

Csörgő and Horváth (1988a) introduced a U-statistic type process and investigated its large sample behavior for testing the 'hypothesis of no changepoint'. In this paper we use their approach to define an estimate θ_n of θ. A new exponential inequality for the tails of $\theta_n - \theta$ is proved which in particular yields

$$|\theta_n - \theta| = O(n^{-1} \ln n) \quad \text{with probability one.}$$

2. Main results

Now, let $K : \mathbb{R}^{2d} \to \mathbb{R}$ be a bounded (measurable) kernel. Set

$$\mu = \int\int K(x,y)F(\mathrm{d}y)F(\mathrm{d}x), \qquad \tau = \int\int K(x,y)G(\mathrm{d}y)G(\mathrm{d}x)$$

and

$$\lambda = \int\int K(x,y)F(\mathrm{d}y)G(\mathrm{d}x).$$

Observe that both μ and τ equal zero if K is antisymmetric:

$$K(x,y) = -K(y,x),$$

as is $K(x,y) = \operatorname{sgn}(x-y)$ or, more generally,

$$K(x,y) = \psi(x-y), \quad \text{with } \psi \text{ skew-symmetric.}$$

The resulting estimate may be viewed as a robust version of the θ_n pertaining to $\psi = \mathrm{id}$ (when $d = 1$), which in this case leads to a successive comparison of the means of the two subsamples.

Boundedness of K is essential in order to guarantee the existence of the above integrals and (in proofs) the applicability of some standard exponential bounds for sums of independent random variables. For unbounded kernels some extra integrability conditions will be required. In place of the exponential bounds the Marcinkiewicz-Zygmund inequality may then serve as a substitute. We prefer to state the results for bounded K's, however, since in this case the conditions are completely carried by the given K rather than by the unknown F and G.

Put

$$r(t) = \int\int K(x, y)_t h(\mathrm{d}y) h_t(\mathrm{d}x), \quad 0 \leq t \leq 1.$$

It is easily seen that

$$r(t) = 1_{\{t \leq \theta\}}[t\mu(\theta - t) + t\lambda(1-\theta)] + 1_{\{t > \theta\}}[(1-t)\lambda\theta + (1-t)\tau(t-\theta)].$$

Moreover, r is continuous on $[0, 1]$ and differentiable at $t \neq \theta$. Under suitable assumptions on K, θ is the unique maximizer (resp. minimizer) of r. The following procedure for constructing θ_n is therefore obvious. Define

$$r_n(t) = \int\int K(x, y)_t h^n(\mathrm{d}y) h_t^n(\mathrm{d}x) = n^{-2} \sum_{i=nt+1}^{n} \sum_{j=1}^{nt} K(X_i^n, X_j^n)$$

on Δ_n, the empirical analogue of r. Set

$$\theta_n = \begin{cases} \arg\min r_n(t) & \text{if } \theta \text{ minimizes } r, \\ \arg\max r_n(t) & \text{if } \theta \text{ maximizes } r. \end{cases} \tag{2.1}$$

θ_n from (2.1) is related to Darkhovskh's (1976) estimator, if we put

$$K(x, y) = 1_{\{x \leq y\}}. \tag{2.2}$$

Apart from the fact that θ maximizes (resp. minimizes) r, we need the following assumption on r:

$$|r(t) - r(\theta)| \geq L|t - \theta|, \quad \text{some positive } L. \tag{2.3}$$

Condition (2.3) is satisfied if $r'(t)$, $t \neq \theta$, is bounded away from zero. For antisymmetric K,

$$r(t) = 1_{\{t \leq \theta\}} t\lambda(1-\theta) + 1_{\{t > \theta\}}(1-t)\lambda\theta.$$

So, if λ is positive, say, θ is the unique maximizer of r and (2.3) is obviously true. The quantity λ serves as a means to measure the 'distance' between F and G.

In a real data situation λ is unknown. Hence even for antisymmetric kernels it is not known whether θ minimizes or maximizes r. It follows from our bounds however (see Remark 1 in Section 3) that $r_n \to r$ uniformly with probability one. So r_n is likely to exhibit whether r is a U-type or hat-type function. In other words, if, e.g., r_n is U-shaped, we take for θ_n the minimizer of r_n.

Typically the kernel K cannot discriminate between all $F \neq G$, i.e., there may exist $F \neq G$ for which $\lambda = 0$. The situation is similar for two-sample linear rank statistics based on a given score function. In each case, specific knowledge of possible changes may help one to choose appropriate kernels. E.g., to detect changes in location we may take $K(x, y) = \psi(x - y)$, with ψ skew-symmetric and strictly increasing. For a change in scale we may take the same ψ and put $K(x, y) = \psi(x^2 - y^2)$.

Finally, these kernels also work if G is an ε-contaminated F:

$$G = (1 - \varepsilon)F + \varepsilon H.$$

In this case

$$\lambda = \varepsilon \int\int K(x, y)F(\mathrm{d}y)H(\mathrm{d}x),$$

so that the above conclusions apply to F and H if H results from F by a change in location or scale.

Theorem 2.1. *Assume that θ is the unique maximizer (minimizer) of r, and that (2.3) is satisfied. Also let K be a bounded kernel. Then there exist positive constants C_0 and C_1 such that for all $\varepsilon > 0$ (and $n \geq n_0 = n_0(F, G, \theta, K)$),*

$$\mathbb{P}(n|\theta_n - [n\theta]/n| \geq 4\varepsilon L^{-2}) \leq C_0 n^2 \exp[-C_1 \varepsilon].$$

C_0 is universal, while C_1 may depend on K.

From Theorem 2.1 and Borel–Cantelli we immediately get:

Corollary 2.2. *With probability one,*

$$n|\theta_n - \theta| = \mathrm{O}(\ln n). \qquad \square$$

For the special case (2.2), Darkhovskh (1976) proved $\theta_n \to \theta$ in probability.

3. Proofs

We shall only consider the case when θ and θ_n maximize r and r_n, respectively. Along with r_n and r, define

$$\begin{aligned}\bar{r}_n(t) &\equiv \mathbb{E} r_n(t) \\ &= 1_{\{nt \leq [n\theta]\}} \left\{ \frac{t\mu([n\theta] - nt)}{n} + \frac{t\lambda(n - [n\theta])}{n} \right\} \\ &\quad + 1_{\{nt > [n\theta]\}} \left\{ \frac{[n\theta]\lambda(1 - t)}{n} + \frac{(1 - t)\tau(nt - [n\theta])}{n} \right\},\end{aligned}$$

for $t \in \Delta_n$. The following lemma turns out to be crucial.

Lemma 3.1. *There exist positive constants C_0, C_1 such that for each $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{s<t; s,t \in \Delta_n} \frac{\sqrt{n}|r_n(t) - r_n(s) - \bar{r}_n(t) + \bar{r}_n(s)|}{\sqrt{t-s}} \geq \varepsilon \right) \leq C_0 n^2 \exp[-C_1 \varepsilon^2].$$

Proof. For $s < t$, we have, omitting the upper index n,

$$n^2[r_n(t) - r_n(s)] = \sum_{i=nt+1}^{n} \sum_{j=ns+1}^{nt} K(X_i, X_j) - \sum_{i=ns+1}^{nt} \sum_{j=1}^{ns} K(X_i, X_j).$$

A similar expansion holds for $\bar{r}_n$. We shall only bound the second term, the analysis for the first being similar. Introduce

$$R_1(y) = \int K(x, y) F(\mathrm{d}x) \quad \text{and} \quad R_2(y) = \int K(x, y) G(\mathrm{d}x).$$

Then

$$\mathbb{E}[K(X_i, X_j) \mid X_1, \ldots, X_{ns}] = H_i(X_j),$$

where $H_i = R_1$ for $i \leq [n\theta]$ and $H_i = R_2$ for $i > [n\theta]$. Set

$$S = S(s, t)) = \frac{1}{n\sqrt{n(t-s)}} \sum_{i=ns+1}^{nt} \sum_{j=1}^{ns} [K(X_i, X_j) - H_i(X_j)].$$

For any $h > 0$, Markov's inequality yields

$$\mathbb{P}(S \geq \varepsilon) \leq \exp[-h\varepsilon\sqrt{n(t-s)}] \mathbb{E}\left\{ \exp\left[h \sum_{i=ns+1}^{nt} \frac{1}{n} \sum_{j=1}^{ns} [\cdots] \right] \right\}.$$

By independence,

$$\mathbb{E}[\{\cdots\} \mid X_1 = x_1, \ldots, X_{ns} = x_{ns}] = \mathbb{E}\left[\exp\left(h \sum_{i=ns+1}^{nt} \delta_i \right) \right]$$

$$= \prod_{i=ns+1}^{nt} \mathbb{E}[\exp h\delta_i],$$

where

$$\delta_i = n^{-1} \sum_{j=1}^{ns} [K(X_i, x_j) - H_i(x_j)], \quad ns+1 \leq i \leq nt,$$

are independent and centered random variables. Assume w.l.o.g. that K is bounded by 1 in absolute values. Then each δ_i is bounded by $c = 2$. From inequality (4.16) in Hoeffding (1963),

$$\mathbb{E}(\exp h\delta_i) \leq \exp(2h^2), \tag{3.1}$$

so that integrating out gives

$$\mathbb{P}(S \geq \varepsilon) \leq \exp[-h\varepsilon\sqrt{n(t-s)} + 2n(t-s)h^2].$$

The right-hand side is minimized for $h = \varepsilon/4\sqrt{n(t-s)}$ yielding

$$\mathbb{P}(S \geq \varepsilon) \leq \exp[-\tfrac{1}{8}\varepsilon^2].$$

Similarly, for $\mathbb{P}(S \leq -\varepsilon)$. It remains to bound

$$T = T(s,t) = \frac{1}{n\sqrt{n(t-s)}} \sum_{i=ns+1}^{nt} \sum_{j=1}^{ns} [H_i(X_j) - \mathbb{E}H_i(X_j)].$$

Observing

$$|T| \leq \left| \frac{1}{\sqrt{ns}} \sum_{j=1}^{ns} \rho_j \right|,$$

with

$$\rho_j = \frac{1}{n(t-s)} \sum_{i=ns+1}^{nt} [H_i(X_j) = \mathbb{E}H_i(X_j)]$$

being independent, bounded and centered, application of (3.1) to ρ_j yields, similar to before,

$$\mathbb{P}(|T| \geq \varepsilon) \leq 2\exp[-\tfrac{1}{8}\varepsilon^2].$$

Since the cardinality of Δ_n is $n-1$, this completes the proof of the lemma. □

Corollary 3.2. *For each* $\varepsilon > 0$,

$$\mathbb{P}(\sqrt{n}|r_n(\theta_n) - r_n([n\theta]/n) - \bar{r}_n(\theta_n) + \bar{r}_n([n\theta]/n)| \geq \varepsilon\sqrt{|\theta_n - [n\theta]/n|})$$
$$\leq C_0 n^2 \exp[-C_1\varepsilon^2]. \quad \square$$

We are now in the position to give:

Proof of Theorem 2.1. Since θ_n maximizes r_n on Δ_n,

$$0 \leq r_n(\theta_n) - r_n([n\theta]/n)$$
$$= r_n(\theta_n) - r_n([n\theta]/n) - \bar{r}_n(\theta_n) + \bar{r}_n([n\theta]/n) + \bar{r}_n(\theta_n) - \bar{r}_n([n\theta]/n).$$

By Corollary 3.2, up to an event of probability less than or equal to $C_0 n^2 \exp(-C_1\varepsilon^2)$, the first sum is less than $\varepsilon\sqrt{|\theta_n - [n\theta]/n|n^{-1}}$. Check that, as a consequence of (2.3), for some finite D and all $n \geq n_0$,

$$\bar{r}_n(\theta_n) - \bar{r}_n([n\theta]/n) \leq -L|\theta_n - [n\theta]/n| + D|\theta_n - [n\theta]/n|/n$$
$$\leq -\tfrac{1}{2}L|\theta_n - [n\theta]/n|.$$

Conclude that

$$|\theta_n - [n\theta]/n| < (2/L)\varepsilon\sqrt{|\theta_n - [n\theta]/n|n^{-1}}.$$

Replace ε^2 by ε to get the assertion of the theorem. □

Remark 1. Similar to Lemma 3.1 one may show that

$$\mathbb{P}\left(\sup_{t\in\Delta_n}\sqrt{n}|r_n(t)-\bar{r}_n(t)|\geq\varepsilon\right)\leq C_0 n\exp[-C_1\varepsilon^2].$$

By Borel-Cantelli we thus get with probability one,

$$\sup_{t\in\Delta_n}|r_n(t)-\bar{r}_n(t)|=O(\sqrt{(\ln n)/n}).$$

Since

$$\sup_{T\in\Delta_n}|\bar{r}_n(t)-r(t)|=O(n^{-1}),$$

we obtain with probability one,

$$\sup_{t\in\Delta_n}|r_n(t)-r(t)|=O(\sqrt{(\ln n)/n})=o(1).$$

References

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

M. Csörgő and L. Horváth, Invariance principles for changepoint problems, J. Multivariate Anal. 27 (1988a) 151-168.

M. Csörgő and L. Horváth, Nonparametric methods for changepoint problems, in: P.R. Krishnaiah and C.R. Rao, eds., Handbook of Statistics, Vol. 7 (Elsevier, Amsterdam, 1988b) pp. 403-425.

B.S. Darkhovskh, A non-parametric method for the a posteriori detection of the "disorder" time of a sequence of independent random variables, Theory Probab. Appl. 21 (1976) 178-183.

L. Dümbgen, The asymptotic behavior of some nonparametric changepoint estimators, Dissertation, Univ. Heidelberg (Heidelberg, 1990).

W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Statist. Assoc. 58 (1963) 13-30.

J. Kiefer, On large deviations of the empiric D.F. of vector chance variables and a law of the iterated logarithm, Pacific J. Math. 11 (1961) 649-660.