EDN: NTXBVR
УДК 517.24

On Special Empirical Processes of Independence in Presence of Covariates

Abduraxim A. Abdushukurov[*]
Moscow State University named after M. V. Lomonosov, Tashkent Branch
Tashkent, Uzbekistan
V. I. Romanovskiy Institute of Mathematics of Uzbekistan Academy of Sciences
Tashkent, Uzbekistan

Farkhad A. Abdikalikov[†]
Karakalpak State University named after Berdakh
Nukus, Uzbekistan
V. I. Romanovskiy Institute of Mathematics of Uzbekistan Academy of Sciences
Tashkent, Uzbekistan

Received 10.05.2022, received in revised form 15.08.2022, accepted 04.11.2022

Abstract. In this paper we investigate asymptotical properties of one class of empirical processes in case of presence of covariates for a class of measurable functions.

Keywords: empirical processes, metrical entropy, Gaussian processes.

Citation: A.A. Abdushukurov, F.A. Abdikalikov, On Special Empirical Processes of Independence in Presence of Covariates, J. Sib. Fed. Univ. Math. Phys., 2023, 16(1), 66–75. EDN: NTXBVR

1. Introduction and preliminaries

A special empirical processes of independence has been introduced in works of Abdushukurov and Kakadjanova [1, 2] in the case of indexing of empirical processes by class of measurable functions $\mathcal{F}$. The modern asymptotic theory of empirical processes indexed by a class $\mathcal{F}$ is actively developed and the current results of this theory allow us to establish uniform versions of the laws of large numbers and central limit theorems for empirical measures under the imposing of the entropy conditions for a class $\mathcal{F}$. These results are essentially generalization of classical theorems of Glivenko-Cantelli and Donsker [3, 4]. In applied mathematics, in order to generalize of Glivenko-Cantelli theorems for a class of sets Vapnik and Chervonenkis in 70-s years of the last centure made a significant contribution to the development of statistical (machine) learning theory (theory of Vapnik-Chervonenkis), which justifies the principle of minimizing of empirical risk (for details, see the monograph [5]).

In the papers of authors [1, 2] the limiting properties of generalized empirical processes of independence of random variables (r.v.-s) and events indexed by a class $\mathcal{F}$ were investigated. Here we extend this model to the regression case. The necessity of considering such processes stems from practical situation, where we are investigated in joint properties of the triple of

[*] a_abdushukurov@rambler.ru

[†] abdikalikovf@mail.ru

observed data: r.v.-s, event and covarate. Let us consider the sequence of observed triples $\{(Z_k,\, A_k,\, X_k)\,,\ k \geqslant 1\}$, where Z_k are positive random elements defined on a probability space $(\Omega, \mathcal{A}, P)$ with values in a measurable space $(\mathfrak{X}, \mathfrak{B})$. Events A_k have a common probability $p = \mathrm{P}\,(A_k) \in (0,1)$. For our analysis, we consider the observed data $(Z_1,\, \delta_1)\,, \ldots, (Z_n,\, \delta_n)$ at n fixed design points $0 \leqslant x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n \leqslant 1$ of covariate X, where $\delta_k = I\,(A_k)$ is an indicator variable of the event A_k. The observed r.v.-s at design points $x \in [0,1]$ are Z_x and δ_x. Here $\delta_x = 1$ denotes that event A_x occurs. Each pair $(Z_x,\, \delta_x)$ of sample induces a statistical model $(\mathfrak{X} \times \{0,1\}\,, \mathfrak{B} \times \{0,1\}\,, \mathcal{P}_x)$ for a given $X = x$, where distribution

$$\{\mathcal{P}_x\,(B \times D) = \mathrm{P}\,(Z \in B,\, \delta \in D/\, X = x)\,,\, B \in \mathfrak{B},\, D \subset \{0,\, 1\}\}\,,$$

for each Borel set B represented through subdistribution:

$$\mathcal{P}_x\,(B \times \{0,\,1\}) = \mathbb{Q}_x\,(B) = \mathbb{Q}_{0x}\,(B) + \mathbb{Q}_{1x}\,(B)\,,\ \mathbb{Q}_{mx}\,(B) = \mathcal{P}_x\,(B \times \{m\})\,,\ m = 0,1.$$

Our interest is focused on hypothesis $\mathcal{H}$ of independence of Z_x and δ_x. It's easy to see that under validity of $\mathcal{H}$: $\mathbb{Q}_{1x}\,(B) = p_x \mathbb{Q}_x\,(B)$ and $\mathbb{Q}_{0x}\,(B) = (1 - p_x)\,\mathbb{Q}_x\,(B)$, for all $B \in \mathfrak{B}$, where $p_x = \mathbb{Q}_{1x}\,(\mathfrak{X})$. Let's introduce the signed measure

$$\{\Lambda_x\,(B) = \mathbb{Q}_{1x}\,(B) - p_x \mathbb{Q}_x\,(B)\,,\, B \in \mathfrak{B}\}\,,$$

which is equal to zero under hypothesis $\mathcal{H}$. Using this measure, we construct an empirical process for testing a hypothesis $\mathcal{H}$. In this regard, we introduce empirical analogues of the above measures for $B \in \mathfrak{B}$:

$$\mathbb{Q}_{xh}\,(B) = \sum_{i=1}^{n} \omega_{ni}\,(x; h_n)\, I\,(Z_i \in B) = \mathbb{Q}_{0xh}\,(B) + \mathbb{Q}_{1xh}\,(B)\,, \tag{1}$$

where

$$\mathbb{Q}_{mxh}\,(B) = \sum_{i=1}^{n} \omega_{ni}\,(x; h_n)\, I\,(Z_i \in B,\, \delta_i = m),\ \ m = 0,1,$$

and

$$\Lambda_{xh}\,(B) = \mathbb{Q}_{1xh}\,(B) - p_{xh} \mathbb{Q}_{xh}\,(B)\,,\ p_{xh} = \mathbb{Q}_{1xh}\,(\mathfrak{X})\,.$$

The nonparametric estimators above involve a sequence of smoothing weights $\{\omega_{ni}\,(x; h_n)\}$, depending on a positive bandwidth sequence $\{h_n,\, n \geqslant 1\}$, tending to zero as $n \to \infty$. In our present case of fixed design points, it is common to use the Gasser-Müller - type weights, given by

$$\omega_{ni}\,(x; h_n) = \frac{1}{C_n\,(x; h_n)} \int_{x_{i-1}}^{x_i} \frac{1}{h_n} k\left(\frac{x - z}{h_n}\right) dz \ \ (i = 1, \ldots, n),$$

$$C_n\,(x; h_n) = \int_{0}^{x_n} \frac{1}{h_n} k\left(\frac{x - z}{h_n}\right) dz.$$

Here $x_0 = 0$ and k is a known probability density function (kernel).

2. Asymptotic results

Under $B = (-\infty, t]$, let's define conditional distribution function (d.f.) and subdistribution functions for a given $X = x$:

$$G_x\,(t) = \mathbb{Q}_x\,((-\infty, t]) = \mathrm{P}\,(Z \leqslant t/\, X = x) = \mathrm{P}\,(Z_x \leqslant t)\,,$$

and

$$G_{mx}(t) = \mathbb{Q}_{mx}((-\infty, t]) = \mathrm{P}(Z \leqslant t,\, \delta = m/\, X = x) = \mathrm{P}(Z_x \leqslant t,\, \delta_x = m)\,, \quad m = 0, 1.$$

We will need the following additional notation. For the design points $x_1, \ldots, x_n$ we denote $\underline{\Delta}_n = \min\limits_{1\leqslant i\leqslant n}(x_i - x_{i-1})$ and $\overline{\Delta}_n = \max\limits_{1\leqslant i\leqslant n}(x_i - x_{i-1})$. For the kernel k we use the following assymptions the design points and the kernel (see, [6–8]):

(C1) $x_n \to 1$, $\overline{\Delta}_n = O\left(\frac{1}{n}\right)$, $\overline{\Delta}_n - \underline{\Delta}_n = o\left(\frac{1}{n}\right)$.

(C2) k is a probability density function with support $[-M, M]$ for some $M > 0$, $m_1(k) = \int\limits_{-\infty}^{\infty} yk(y)\,dy = 0$ and k is Lipschitz of order 1.

Note that $C_n(x; h_n) = 1$ for n sufficiently large since $x_n \to 1$ and k has finite support. This makes that in all proofs of asymptotic results we may take $C_n(x; h_n) = 1$.

Further we will need typical smoothness condition of $G_x(t)$ and $G_{mx}(t)$, $m = 0, 1$ and probability $p_x = G_{1x}(+\infty) = \lim\limits_{t\to+\infty} G_{1x}(t)$.

(C3) The second-order partial derivatives $\ddot{G}_x(t) = \dfrac{\partial^2}{\partial x^2}G_x(t)$, $\ddot{G}_{mx}(t) = \dfrac{\partial^2}{\partial x^2}G_{mx}(t)$ and $G''_x(t) = \dfrac{\partial^2}{\partial t^2}G_x(t)$ and $G''_{mx}(t) = \dfrac{\partial^2}{\partial t^2}G_{mx}(t)$, $m = 0, 1$, exist and are continuous for $0 \leqslant x \leqslant 1$ and $t \in R$.

(C4) The second-order partial derivatives $\ddot{p}_x = \dfrac{d^2}{dx^2}p_x$ exist and are continuous for $0 \leqslant x \leqslant 1$.

In what follows, we also use the notation

$$\left\|\dot{G}_x\right\| = \sup_{(t,x)\in[0,T]\times[0,1]}\left|\dot{G}_x(t)\right|, \quad \left\|\ddot{G}_x\right\| = \sup_{(t,x)\in[0,T]\times[0,1]}\left|\ddot{G}_x(t)\right|,$$
$$\|\dot{p}_x\| = \sup_{x\in[0,1]}|\dot{p}_x|\,, \quad \|\ddot{p}_x\| = \sup_{x\in[0,1]}|\ddot{p}_x|\,.$$

We denote a weighted estimates for $G_x(t)$ and $G_{mx}(t)$, $m = 0, 1$ obtained from (1) as

$$\begin{aligned} G_{xh}(t) &= \mathbb{Q}_{xh}((-\infty, t]) = \sum_{i=1}^{n}\omega_{ni}(x; h_n)\, I(Z_i \leqslant t), \\ G_{mxh}(t) &= \mathbb{Q}_{mxh}((-\infty, t]) = \sum_{i=1}^{n}\omega_{ni}(x; h_n)\, I(Z_i \leqslant t,\, \delta_i = m), \quad m = 0, 1, \end{aligned} \tag{2}$$

and by defenition $\omega_{ni}(x; h_n)$, $\omega_{n1}(x; h_n) + \cdots + \omega_{nn}(x; h_n) = 1$. Note that, when we put $\omega_{ni}(x; h_n) = 1/n$, $i = 1, \ldots, n$, then estimators (2) transformated to usual empirical estimator for $G_x(t)$ and $G_{mx}(t)$, $m = 0, 1$.

For a sufficient large n by condition (C1), we shall suppose that $C_n(x; h_n) \approx 1$. Hence, in future calculation in asymptotic results we will put $C_n(x; h_n) = 1$.

Now we give some asymptotic results for estimators (2) from works [6,8]. Let $T < T_{xG} = \inf\{t : G_x(t) = 1\}$.

Lemma 2.1 ([6]) *(Bias and variance). (a) Let conditions (C1)–(C3) satisfies and at $n \to \infty$, $h_n \to 0$, $nh_n \to \infty$. Then under $n \to \infty$*

$$\sup_{0\leqslant t\leqslant T}|EG_{xh}(t) - G_x(t)| = o(h_n)\,.$$

(b) Let conditions (C1)–(C3) satisfies and at $n \to \infty$, $h_n \to 0$. Then under $n \to \infty$

$$\sup_{0\leqslant t\leqslant T} |EG_{xh}(t) - G_x(t)| = O\left(h_n^2 + \frac{1}{n}\right).$$

In particular,

$$\sup_{0\leqslant t\leqslant T} \left| EG_{xh}\left(t\right) - G_x\left(t\right) - \frac{1}{2}m_2\left(k\right)\ddot{G}_x\left(t\right)h_n^2\right| = o\left(h_n^2\right) + O\left(\frac{1}{n}\right).$$

(c) Under conditions of (a) at $n \to \infty$

$$DG_{xh}\left(t\right) = \frac{1}{nh_n}G_x\left(t\right)\left(1 - G_x\left(t\right)\right)\|k\|_2^2 + o\left(\frac{1}{nh_n}\right),$$

where $m_2\left(k\right) = \int\limits_{-\infty}^{\infty} y^2 k\left(y\right)dy$ and $\|k\|_2^2 = \int\limits_{-\infty}^{\infty} k^2\left(y\right)dy$.

Lemma 2.2 ([6]) *(Pointwise strong consistency). Let conditions (C1)–(C3) satisfies and at $n \to \infty$, $h_n \to 0$, $\dfrac{\log n}{nh_n} = o(1)$. Then under $n \to \infty$ and $t \leqslant T$*

$$G_{xh}\left(t\right) \overset{a.s.}{\to} G_x\left(t\right).$$

Lemma 2.3 ([6]) *(Exponential estimator of Dworetzky–Kiefer–Wolfowitz). Let conditions (C1), (C2) satisfies and at $n \to \infty$, $nh_n \to \infty$.*

(a) For $\varepsilon > 0$ and large n such that

$$\varepsilon^2 \geqslant \frac{3}{2}\|k\|_2^2 \frac{1}{nh_n},$$

and for $T > 0$

$$P\left(\sup_{0\leqslant t\leqslant T} |G_{xh}\left(t\right) - EG_{xh}\left(t\right)| > \varepsilon\right) \leqslant 2d_0 nh_n \varepsilon \exp\left(-d_1 nh_n \varepsilon^2\right).$$

(b) Moreover, in condition (C3) hold for sufficient for $\varepsilon > 0$ and n such that

$$\varepsilon \geqslant \max\left\{\left(\sqrt{6}\|k\|_2(nh_n)^{-1/2}\right), \left(2\left\|\dot{G}_x\right\|\bar{\Delta}_n + 2m_2(k)\left\|\ddot{G}_x\right\|h_n^2\right)\right\},$$

then

$$P\left(\sup_{0\leqslant t\leqslant T} |G_{xh}\left(t\right) - G_x\left(t\right)| > \varepsilon\right) \leqslant \frac{1}{2}d_0 nh_n \varepsilon \exp\left(-\frac{1}{4}d_1 nh_n \varepsilon^2\right) \qquad (3)$$

where $d_0 = \dfrac{8e^2}{\|k\|_2^2}$ and $d_1 = \dfrac{4}{3\|k\|_2^2}$. From (3) by Borel–Cantelli lemma under $\varepsilon = \varepsilon_n = c(nh_n)^{-1/2}(\log n)^{1/2}$ we have

Lemma 2.4 ([6]) *(Rate of strong uniform consistency). Let conditions (C1)–(C3) satisfies and at $n \to \infty$, $\dfrac{nh_n^5}{\log n} = O(1)$. Then under $n \to \infty$*

$$\sup_{0\leqslant t\leqslant T} |G_{xh}\left(t\right) - G_x\left(t\right)| \overset{a.s.}{=} O\left(\left(\frac{\log n}{nh_n}\right)^{1/2}\right).$$

For a measure G_x and a class $\mathcal{F}$ of Borel measurable functions $f : \mathfrak{X} \to R$, we introduce the integral over $\mathfrak{X}$

$$G_x f = \int_{\mathfrak{X}} f dG_x, \ \ f \in \mathcal{F},$$

which is expectation by measure G_x of function f. Let us introduce the following $\mathcal{F}$ indexed extensions of (1) for $f \in \mathcal{F}$:

$$\int_{\mathfrak{X}} f \, dG_{xh} = \sum_{i=1}^{n} \omega_{ni}\left(x; h_n\right) f\left(Z_i\right) = G_{0xh} f + G_{1xh} f,$$

where

$$\begin{aligned} G_{0xh} f &= \sum_{i=1}^{n} \omega_{ni}\left(x; h_n\right)\left(1-\delta_i\right) f\left(Z_i\right), \\ G_{1xh} f &= \sum_{i=1}^{n} \omega_{ni}\left(x; h_n\right) \delta_i f\left(Z_i\right). \end{aligned} \tag{4}$$

Introduce the empirical processes under the validity of $\mathcal{H}$,

$$\left(\frac{nh_n}{p_{xh}(1-p_{xh})}\right)^{1/2} \left(\Lambda_{xh} - \Lambda_x\right) f = A_{1xh} f - p_x \cdot A_{xh} f - G_x f \cdot A_{1xh} 1 - R_{xh}\left(f\right), \ \ f \in \mathcal{F} \tag{5}$$

where

$$\begin{aligned} A_{xh} f &= \left(\frac{nh_n}{p_{xh}(1-p_{xh})}\right)^{1/2} \int_{\mathfrak{X}} f \, d\left(G_{xh} - G_x\right), \\ A_{1xh} f &= \left(\frac{nh_n}{p_{xh}(1-p_{xh})}\right)^{1/2} \int_{\mathfrak{X}} f \, d\left(G_{1xh} - G_{1x}\right), \\ R_{xh}\left(f\right) &= \left(\frac{nh_n}{p_{xh}(1-p_{xh})}\right)^{1/2} \left(p_{xh} - p_x\right) \int_{\mathfrak{X}} f \, d\left(G_{xh} - G_x\right). \end{aligned} \tag{6}$$

In order to considering the uniform variants of the Glivenko-Cantelli theorem and the Donsker theorem we need some notations from bracketing entropy theory. Let $\mathcal{L}_q\left(\mathbb{Q}\right)$ be the space of functions $f : \mathfrak{X} \to R$ with norm

$$\|f\|_{\mathbb{Q},q} = \left(\mathbb{Q}|f|^q\right)^{1/q} = \left(\int_{\mathfrak{X}} |f|^q d\mathbb{Q}\right)^{1/q}.$$

To determine the complexity or entropy of a set of a set of Borel measurable functions $\mathcal{F}$ it is necessary to define a concept of ε-brackets in $\mathcal{L}_q\left(\mathbb{Q}\right)$. So ε-bracket in $\mathcal{L}_q\left(\mathbb{Q}\right)$ is a pairs of functions $\varphi, \psi \in \mathcal{L}_q\left(\mathbb{Q}\right)$ such that $\mathbb{Q}\left(\varphi(Z) \leqslant \psi(Z)\right) = 1$ and $\|\psi - \varphi\|_{\mathbb{Q},q} \leqslant \varepsilon$, that is $\mathbb{Q}(\psi - \varphi)^q \leqslant \varepsilon^q$. Function $f \in \mathcal{F}$ is covered by bracket $[\varphi, \psi]$ if $\mathbb{Q}\left(\varphi(Z) \leqslant f(Z) \leqslant \psi(Z)\right) = 1$. Not that functions φ and ψ may not belong to the set $\mathcal{F}$ but they must have finite norms. The bracketing number $N_{[\,]}\left(\varepsilon, \mathcal{F}, \mathcal{L}_q\left(\mathbb{Q}\right)\right)$ is the minimum number of ε-brackets in $\mathcal{L}_q\left(\mathbb{Q}\right)$ needed to cover the set $\mathcal{F}$ [3,4]:

$$N_{[\,]}\left(\varepsilon, \mathcal{F}, \mathcal{L}_q\left(\mathbb{Q}\right)\right) = \min \begin{cases} k : \text{ for some } f_1, \ldots, f_k \in \mathcal{L}_q\left(\mathbb{Q}\right), \\ \mathcal{F} \subset \bigcup_{i,j} [f_i, f_j] : \|f_j - f_i\|_{\mathbb{Q},q} \leqslant \varepsilon. \end{cases}$$

The number $H_q\left(\varepsilon\right) = \log N_{[\,]}\left(\varepsilon, \mathcal{F}, \mathcal{L}_q\left(\mathbb{Q}\right)\right)$ is called the metric entropy of class $\mathcal{F}$ in $\mathcal{L}_q\left(\mathbb{Q}\right)$. The metric entropies of class $\mathcal{F}$ in $\mathcal{L}_q\left(\mathbb{Q}_m\right)$, $m = 0, 1$ is we denoted by $H_{mq}(\varepsilon) =$

$= \log N_{m[\,]}(\varepsilon, \mathcal{F}, \mathcal{L}_q(\mathbb{Q}_m))$. Integrals of metric entropies are

$$J_{m[\,]}^{(q)}(\delta) = J_{m[\,]}(\delta, \mathcal{F}, \mathcal{L}_q(\mathbb{Q}_m)) = \int_0^\delta (H_{mq}(\varepsilon))^{1/2} d\varepsilon, \quad 0 < \delta \leqslant 1, \quad m = 0, 1.$$

Let us recall the important properties of numbers $N_{[\,]}(\cdot)$. They tend to $+\infty$ when $\varepsilon \downarrow 0$. However, for the Donsker theorems they should converge to $+\infty$ not very fast. This rate of convergence is measured by integrals $J_{m[\,]}^{(q)}(\delta)$. For example, for a class $\mathcal{F}$ of monotone functions $f : \mathfrak{X} \to [0, 1]$ and each measure $\mathbb{Q}_m$ one has

$$H_{mq}(\varepsilon) \leqslant k_0 \varepsilon^{-1},$$

where k_0 is depends only on q. In particular, for a class $\mathcal{F}$ of indicators $\mathcal{F} = \{I(-\infty, t], t \in R\}$ entropy is $H_{m1}(\varepsilon) \sim |\log \varepsilon|$ and at $n \to \infty$

$$J_{m[\,]}^{(2)}(\delta_n) = \int_0^{\delta_n} \left[\log O\left(\varepsilon^{-1}\right)\right]^{1/2} d\varepsilon = O\left(\delta_n^{1/2}\right) \to 0, \ \delta_n \downarrow 0.$$

In future we can investigate the relation (5) and its summands (6). Next lemma is useful in estimating of convergence to zero of remainder term $R_{xh}(f)$ in (6).

Lemma 2.5 ([8]) *Assume (C1), (C2) and (C4), $h_n \to 0$.*

(a) For $\varepsilon > 0$ and n sufficiently large such that

$$\varepsilon \geqslant 2 \|\dot{p}_x\| \bar{\Delta}_n + 2m_2(k) \|\ddot{p}_x\| h_n^2$$

we have

$$P(|p_{xh} - p_x| > \varepsilon) \leqslant 2 \exp\left(-dnh_n \frac{\varepsilon^2}{1 + \varepsilon/6}\right),$$

where d is some absolute constant.

(b) If $\dfrac{\log n}{nh_n} \to 0$, then $p_{xh} - p_x \to 0$ a.s.

(c) If $\dfrac{nh_n^5}{\log n} = O(1)$, then $p_{xh} - p_x = O\left((nh_n)^{-1/2} (\log n)^{1/2}\right)$ a.s.

Now we prove that two-dimensional vector field $(A_{xh}f, A_{1xh}g)$, f, $g \in \mathcal{F}$ weakly converges to corresponding Gaussian field uniformly with respect to space $l^\infty(\mathcal{F}) \times l^\infty(\mathcal{F})$ for every class of measurable functions $\mathcal{F}$. This is necessary for investigating of expansion (5).

Theorem 2.1 *Let us consider conditions (C1)–(C4) and the class $\mathcal{F}$ measurable functions f such that*

$$\mathcal{F} \subset \mathcal{L}_2(\mathbb{Q}_{mx}) \ \text{and} \ J_{m[\,]}^{(2)}(1) < \infty, \quad m = 0, 1. \tag{7}$$

Then for $n \to \infty$ sequence of random vector field $(A_{xh}f, A_{1xh}g)$, f, $g \in \mathcal{F}$ weakly converge in $l^\infty(\mathcal{F}) \times l^\infty(\mathcal{F})$ to the Gaussian field $(A_x f, A_{1x} g)$, f, $g \in \mathcal{F}$ with zero mean and covariance structure

$$\begin{aligned}
\operatorname{cov}(A_x f, A_x g) &= \|k\|_2^2 \{G_x fg - G_x f\, G_x g\}, \\
\operatorname{cov}(A_{1x} f, A_{1x} g) &= \|k\|_2^2 \{G_{1x} fg - G_{1x} f\, G_{1x} g\}, \\
\operatorname{cov}(A_x f, A_{1x} g) &= \|k\|_2^2 \{G_{1x} fg - G_x f\, G_{1x} g\}.
\end{aligned} \tag{8}$$

Proof. Consider the first condition in (7). Then for the fixed $f \in \mathcal{F}$ it follows that $\mathbb{Q}_{mx} f^2 < \infty$, $m = 0, 1$, and hence $\mathbb{Q}_x f^2 = \mathbb{Q}_{0x} f^2 + \mathbb{Q}_{1x} f^2 < \infty$. For every such Donsker class $\mathcal{F}$ with the second condition in (7) the sequences $A_{xh} f$ and $A_{1xh} g$ are asymptotically tight (see, Lemma 1.3.8 in [3]). There exist a tight Borel measurable version of Gaussian processes $A_x f$ and $A_{1x} g$, that is, the Gaussian processes with zero mean and jointly covariance (8). Tightness and measurability of limiting process $A_x f$ and $A_{1x} g$ are equivalent to the existence of versions of all sample paths $f \to A_x f$, $g \to A_{1x} g$ uniformly bounded and uniformly continuous with respect to the corresponding mean square metrics (see, [3], p. 226)

$$E(A_x f - A_x g)^2 = \sigma^2_{\mathbb{Q}_x}(f) + \sigma^2_{\mathbb{Q}_x}(g) + \sigma^2_{\mathbb{Q}_x}(f - g),$$

$$E(A_{1x} f - A_{1x} g)^2 = \sigma^2_{\mathbb{Q}_{1x}}(f) + \sigma^2_{\mathbb{Q}_{1x}}(g) + \sigma^2_{\mathbb{Q}_{1x}}(f - g),$$

where $\sigma^2_{\mathbb{Q}_x}(f) = \mathbb{Q}_x (f - \mathbb{Q}_x f)^2$, $\sigma^2_{\mathbb{Q}_{1x}}(f) = \mathbb{Q}_{1x} (f - \mathbb{Q}_{1x} f)^2$.

On the other hand, the considered vector-field is the normalized sum of independent and identically distributed random vectors

$$(A_{xh} f, A_{1xh} g) = (nh_n)^{-1/2} \sum_{i=1}^{n} (\omega_{ni}(x; h_n)(f(Z_i) - \mathbb{Q}_x f), \omega_{ni}(x; h_n)(\delta_i g(Z_i) - \mathbb{Q}_{1x} g)), \quad (9)$$

then by the multivariate central limit theorem the marginals of the sequence of vector-fields converge to the marginals of a Gaussian vector-valued field with zero mean and covariance matrix defined by structure (8). Vector-field (9) is element of product-space $l^\infty(\mathcal{F}) \times l^\infty(\mathcal{F})$, and it also induces tight sequences of distributions in product-space by Lemma 1.4.3 [3]. The limiting value of covariance structure of vector (9) is coincides with covariance structure (8). These arguments complete the proof of Theorem 2.1. □

Remark 2.1. Consider formulas (8). At $g \equiv 1$ for $f \in \mathcal{F}$ we have $A_{1x} 1 \equiv p_x$ and hence

$$\operatorname{cov}(A_x f, A_{1x} 1) = G_{1x} f - G_x f \, G_{1x} 1 = G_{1x} f - p_x G_x f = \Lambda_x f. \quad (10)$$

Because covariance (10) is zero under validity of hypothesis $\mathcal{H}$, then Gaussian fields $\{A_x f, f \in \mathcal{F}\}$ and normal r.v. $A_{1x} 1$ with variance $p_x(1 - p_x)$, are independent.

Remark 2.2. By Lemmas 2.1-2.4 and Lemma 2.5, by consistency of G_{xh} for G_x and p_{xh} for p_x we see that remainder $R_{xh}(f)$ tends to zero as $n \to \infty$: $|R_{xh}(f)| = o(1)$ in probability.

Now we study normalized empirical process (5) without remainder term $R_{xh}(f)$, which tends to zero as $n \to \infty$. Let's denote

$$\Delta_{xh} f = (nh_n)^{1/2} (\Lambda_{xh} - \Lambda_x) f = (p_{xh}(1 - p_{xh}))^{1/2} \{A_{1xh} f - p_x \cdot A_{xh} f - G_x f \cdot A_{1xh} 1\}.$$

This process is the intermediate random field plays a supporting role in the study of basic process (5) which property of weak convergence to a corresponding Gaussian process is contained in the following statement.

Theorem 2.2 *Under conditions (C1)–(C4) and (7). Then for $n \to \infty$ we have*

$$\Delta_{xh} f \Rightarrow \Delta_x f \ \text{in} \ l^\infty(\mathcal{F}), \quad (11)$$

where $\{\Delta_x f, f \in \mathcal{F}\}$ is a Gaussian fields with zero mean and with covariance

$$\operatorname{cov}(\Delta_x f, \Delta_x g) = \|k\|_2^2 \, p_x(1 - p_x) \{G_x fg - G_x f \, G_x g\}. \quad (12)$$

Proof. Let us consider process $\Delta_{xh}f$, which is zero mean Gaussian by Theorem 2.1. We then consider only covariance

$$\operatorname{cov}(\Delta_{xh}f, \Delta_{xh}g) = \|k\|_2^2\, p_x(1-p_x)\left\{\sum_{j=1}^{9} \mathrm{C}_j\right\}, \tag{13}$$

where

$$\begin{aligned} &\mathbb{C}_1 = G_{1x}fg - G_{1x}fG_{1x}g, && \mathbb{C}_2 = -p_x(G_{1x}fg - G_xfG_{1x}g),\\ &\mathbb{C}_3 = -(1-p_x)G_xfG_{1x}g, && \mathbb{C}_4 = -p_x(G_{1x}fg - G_xgG_{1x}f),\\ &\mathbb{C}_5 = p_x^2(G_xfg - G_xfG_xg), && \mathbb{C}_6 = p_xG_xf(G_{1x}g - p_xG_xg),\\ &\mathbb{C}_7 = -(1-p_x)G_xgG_{1x}f, && \mathbb{C}_8 = p_xG_xg(G_{1x}f - p_xG_xg),\\ &\mathbb{C}_9 = p_x(1-p_x)G_xfG_xg. \end{aligned} \tag{14}$$

Now adding of all elements (14) by formula (13) we obtain (12). Theorem 2.2 is proved. □

Thus, statistics for testing of hypothesis $\mathcal{H}$ one can construct from normalized process as a some functional

$$\left\{\left(\frac{nh_n}{p_{xh}(1-p_{xh})}\right)^{1/2} \|k\|_2^{-1}\left(\Lambda_{xh} - \Lambda_x\right)f,\ f \in \mathcal{F}\right\}. \tag{15}$$

3. Application to random censoring

Let us consider a right random censoring model, where $Z_i = \min\{T_i, C_i\}$, $A_i = \{T_i \leqslant C_i\}$. Here r.v.-s T_i and C_i denote life times and censoring times, which is independent at fixed design points $0 \leqslant x_1 \leqslant x_2 \leqslant \dots \leqslant x_n \leqslant 1$. Hence at each design points x_i, there is a r.v. C_i such that we only observe the pair (Z_i, δ_i), where $\delta_i = I\left(A_i\right)$. Furthermore, we suppose that d.f.-s F_{x_i} and K_{x_i} of r.v.-s T_i and C_i are continuous and $F_{x_i}(0) = K_{x_i}(0) = 0$. Consequently we have that the d.f.

$$H_{x_i}(t) = \mathrm{P}\left(Z_i \leqslant t/\, X = x_i\right) = 1 - \left(1 - F_{x_i}(t)\right)\left(1 - K_{x_i}(t)\right).$$

Subdistributions defined as

$$\mathbb{Q}_{0x_i}(B) = \mathrm{P}\left(Z_i \in B, \delta_i = 0/\, X = x_i\right) = \mathrm{P}\left(C_{x_i} \in B \cap [0, T_{x_i}]\right) = \int_B \left(1 - F_{x_i}(t)\right) K_{x_i}(dt),$$

$$\mathbb{Q}_{1x_i}(B) = \mathrm{P}\left(Z_i \in B, \delta_i = 1/\, X = x_i\right) = \mathrm{P}\left(T_{x_i} \in B \cap [0, C_{x_i}]\right) = \int_B \left(1 - K_{x_i}(t)\right) F_{x_i}(dt).$$

As in the situation without covariates, we can also define in this model a Koziol-Green type sub-model by assuming that for a given design point x, the conditional survival function of C_x is some power of the conditional survival function of T_x: for $t \geqslant 0$,

$$1 - K_x(t) = \left(1 - F_x(t)\right)^{\beta_x},$$

where $\beta_x > 0$ and is allowed to depend on the covariate x. We note that here

$$\mathrm{P}\left(\delta_x = 1\right) = \int_0^\infty \left(1 - K_x(t)\right) dF_x(t) = \int_0^\infty \left(1 - F_x(t)\right)^{\beta_x} dF_x(t),$$

$$\mathrm{P}\left(\delta_x = 0\right) = \int_0^\infty \left(1 - F_x(t)\right) dK_x(t) = \beta_x \int_0^\infty \left(1 - F_x(t)\right)^{\beta_x} dF_x(t),$$

and hence $\beta_x = \dfrac{\mathrm{P}(\delta_x = 0)}{\mathrm{P}(\delta_x = 1)}$.

By this extra assumption the estimator in this sub-model has a simpler form than in the general model and is given by

$$\hat{F}_{xh}(t) = 1 - (1 - H_{xh}(t))^{\gamma_{xh}},$$

where $H_{xh}(t) = \sum\limits_{i=1}^{n} \omega_{ni}(x; h_n) I(Z_i \leqslant t)$ and $\gamma_{xh} = \sum\limits_{i=1}^{n} \omega_{ni}(x; h_n) \delta_i$ are Stone type estimators for $H_x(t) = \mathrm{P}(Z_x \leqslant t)$ and $\gamma_x = \dfrac{1}{1+\beta_x} = \mathrm{P}(\delta_x = 1)$. This estimator has been studied more extentively by Veraverbeke and Cadarso-Suarez [7]. These authors noted the superiority of methods for estimating and the testing in Koziol–Green proportional hazards model and methods are based on $\hat{F}_{xh}$ rather than on the product-limit estimator of Kaplan–Meier [10] or relative risk power estimator of Abdushukurov [9]. Hence the question arises as to when the advantages of the Koziol–Green model can be used. In other words, there is now a need for testing of validity of composite hypothesis described by by relation (10). But this relation is equivalent to hypothesis $\mathcal{H}$ on independence of r.v.-s Z_x and δ_x in sample. Let us consider the following special normalized empirical process, special Kolmogorov-type statistics, obtained from (15): $\sup\limits_{|t|<\infty} \left|\Delta^0_{xh}(t)\right|$, where

$$\Delta^0_{xh}(t) = \left(\frac{nh_n}{p_{xh}(1-p_{xh})}\right)^{1/2} \|k\|_2^{-1} \left(H_{1xh}(t) - p_{xh} H_{xh}(t)\right), \quad |t| < \infty, \tag{16}$$

where $H_{1xh}(t) = \sum\limits_{i=1}^{n} \omega_{ni}(x; h_n) I(Z_i \leqslant t, \delta_i = 1)$. Then we have consequence of Theorem 2.2: if $\mathcal{H}$ holds, then as $n \to \infty$

$$\Delta^0_{xh}(t) \Rightarrow \mathbb{B}(H_x(\cdot)), \tag{17}$$

where $\{\mathbb{B}(y),\, 0 \leqslant y \leqslant 1\}$ is a Brownian bridge. Note that these statistics based on convergence (17) are consistent. Moreover, by Theorem 2.2 one can consider more general classes of statistics using $\mathcal{F}$- indexed processes that are more flexible in application than (16).

References

[1] A.A.Abdushukurov, L.R.Kakadjanova, A class of special empirical processes of independence, *J. Sib. Fed. Univ. Math. Phys.*, **8**(2015), no. 2, 125–133.

[2] A.A.Abdushukurov, L.R.Kakadjanova, Sequential empirical process of independence, *J. Sib. Fed. Univ. Math. Phys.*, **11**(2018), no. 5, 634–643.
DOI: 10.17516/1997-1397-2018-11-5-634-643

[3] A.W.Van der Vaart, J.A.Wellner, Weak convergence and empirical processes, Springer, 1996.

[4] A.W.Van der Vaart, Asymptotic Statistics, Cambridge University Press, 1998.

[5] V.N.Vapnik, Statistical learning theory, Wiley, New York, 1998.

[6] I. Van Keilegom, N.Veraverbeke, Estimation and bootstrap with censored data in fixed design nonparametric regression, *Annals of the Institute of Statistical Mathematics*, **49**(1997), 467–491.

[7] N.Veraverbeke, C.Cadarso-Suarez, Estimation of the conditional distribution in a conditional Koziol-Green model, *Test*, **9**(2000), 97–122.

[8] R.Breakers, Regression problems with partially informative or dependent censoring, Limdburgs University Centrum, 2004.

[9] A.A.Abdushukurov, Nonparametric estimation of the distribution function based on relative risk function, *Commun. Statist.: Th and Math.*, **27**(1998), no.8, 1991–2012.

[10] E.L.Kaplan, P.L.Meier, Nonparametric estimation from incomplete observations, *J.A.S.A.*, **53**(1958), 457–481.

Специальные эмпирические процессы независимости в присутствии ковариат

Абдурахим А. Абдушукуров
Филиал Московского государственного университета имени М. В. Ломоносова в г. Ташкенте
Ташкент, Узбекистан
Институт математики имени В. И. Романовского АН РУз
Ташкент, Узбекистан
Фархад А. Абдикаликов
Каракалпакский государственный университет
Нукус, Узбекистан
Институт математики имени В. И. Романовского АН РУз
Ташкент, Узбекистан

Аннотация. В работе исследуются асимптотические свойства одного класса эмпирических процессов при наличии ковариат для определенного класса измеримых функций.

Ключевые слова: эмпирические процессы, метрическая энтропия, гауссовские процессы.