PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 22, Fasc. 2 (2002), pp. 333–342

WEAK CONVERGENCE OF SOME RANDOMLY INDEXED EMPIRICAL PROCESSES

BY

LUISA BEGHIN (ROMA)

Abstract. In this paper, we shall be concerned with weak convergence of the randomly indexed versions of the standard and "independence" empirical processes, in the general framework of stochastic processes indexed by classes of functions and without any distributional assumption. We obtain, in the limit, some generalizations of well-known Gaussian fields such that those arising with a deterministic index are embedded as a special case.

Key words: Kac empirical process, random sample size, generalized P-Brownian fields, Donsker classes, independence empirical process.

1. INTRODUCTION

The aim of this work is to study the limiting behaviour of some empirical processes in the case where the sample size is itself a random variable. We shall be concerned with the randomly indexed versions of the standard and "independence" empirical processes in the general framework of stochastic processes indexed by classes of functions and without any distributional assumption.

The statistical motivation for introducing random sample sizes is that, in many applied circumstances (for example, when sampling from vegetable or animal species), the number of elements in the sample is not fixed a priori because of constraints in time, space or costs. Empirical processes with random sample size were considered in a pioneering paper by Kac [10], for the particular case of a Poisson index, and later by Csörgo [6] and many other authors. In particular, Nikitin [12] proved, in the one-dimensional context, the weak convergence of the Kac process to a generalization of the Brownian bridge, within a wider class of distributions of the random index, and he applied this result to the evaluation of the Bahadur efficiency for various goodness of fit statistics. Many other applications are treated in Gnedenko and Korolev [8], where the asymptotic behaviour of sums of a random number of variables is studied under wider circumstances.

We now report some well-known results, and we introduce some definitions and notation to be used throughout the paper. Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and $X_1, X_2, \ldots$ a sequence of i.i.d. random elements with values in $\mathcal{X}$. Let P be a common distribution of X_i. The empirical measure $\boldsymbol{P}_n$ of the sample $X_1, \ldots, X_n$ is defined by $\boldsymbol{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$, where δ_{X_i} is the Dirac measure at the observation X_i. Given a class $\mathcal{F}$ of measurable functions $f\colon \mathcal{X} \to \boldsymbol{R}$, the *empirical process indexed by* $\mathcal{F}$ will be determined by

$$(1)\quad \boldsymbol{G}_n f \doteq \sqrt{n}(\boldsymbol{P}_n - P)f = \frac{1}{\sqrt{n}}\sum_{i=1}^n (\delta_{X_i} - P)f = \frac{1}{\sqrt{n}}\sum_{i=1}^n \big(f(X_i) - Pf\big), \quad f \in \mathcal{F}.$$

Here we use the abbreviation $Qf \doteq \int f dQ$ for $f \in \mathcal{F}$ and a signed measure Q.

For this kind of processes the *modern theory of weak convergence* (which generalizes the classic results presented in Billingsley [4]) was developed on the notion of weak convergence for random elements that are not necessarily Borel measurable (Hoffman-Jørgensen [9]): for a complete exposition, see van der Vaart and Wellner [14] and Dudley [7]. This new approach yields the so-called "Donsker theorems", which provide general conditions on $\mathcal{F}$ under which

$$(2)\qquad \boldsymbol{G}_n = \sqrt{n}(\boldsymbol{P}_n - P) \Rightarrow \boldsymbol{G} \quad \text{in } l^\infty(\mathcal{F}),$$

where $l^\infty(\mathcal{F})$ denotes, as usual, the space of all bounded functions from a set $\mathcal{F}$ to $\boldsymbol{R}$, equipped with the supremum norm $\|z\|_{\mathcal{F}} \doteq \sup_{f\in\mathcal{F}} |z(f)|$, and $\Rightarrow$ means the weak convergence (see e.g. van der Vaart and Wellner [14], p. 81). If (2) holds, $\mathcal{F}$ is called a *Donsker class*. The limiting field $\{\boldsymbol{G}f\colon f \in \mathcal{F}\}$ is the so-called *P-Brownian bridge*: it is a tight Borel measurable element of $l^\infty(\mathcal{F})$ and is defined as a Gaussian field with zero mean and covariance function

$$(3)\qquad E\boldsymbol{G}f\boldsymbol{G}g = P(f - Pf)(g - Pg) = Pfg - PfPg.$$

The P-Brownian bridge can also be expressed as

$$(4)\qquad \boldsymbol{G}(f) \stackrel{d}{=} \boldsymbol{S}(f) - \boldsymbol{S}(1)Pf,$$

where $\boldsymbol{S}$ denotes the so-called *P-Brownian sheet*, which is a zero-mean Gaussian field with covariance $E(\boldsymbol{S}f\boldsymbol{S}g) = Pfg$, and $\boldsymbol{S}(1)$ the P-Brownian sheet evaluated at $f \equiv 1$.

Csörgo [5], van der Vaart and Wellner [14], § 3.5, considered the direct analogue of $\boldsymbol{G}_n$ for a random sample size, i.e.

$$(5)\qquad \boldsymbol{G}_{\nu_n} \doteq \sqrt{\nu_n}(\boldsymbol{P}_{\nu_n} - P) = \frac{1}{\sqrt{\nu_n}}\sum_{i=1}^{\nu_n} (\delta_{X_i} - P),$$

where $\{\nu_n\}_{n\geq 1}$ denotes a sequence of non-negative integer-valued random variables (independent of X_i's). Those authors showed that $\boldsymbol{G}_{\nu_n}$ weakly converges

to the P-Brownian bridge $\boldsymbol{G}$, in $l^\infty(\mathscr{F})$, under the assumption that $\mathscr{F}$ is P-Donsker and that there exists a deterministic sequence $c_n \to \infty$ such that $v_n/c_n \xrightarrow{P} v$ for some non-negative random variable (r.v.) v. Note that the normalizing sequence in (5) is itself random. In the present paper we are interested, instead, in the asymptotic behaviour of processes normalized by the deterministic sequence $\{\lambda_n\}_{n\geq 1}$ to be defined as the expected value of the random index $\{v_n\}_{n\geq 1}$. More precisely, let $E(v_n) \doteq \lambda_n$, $\mathrm{Var}(v_n) \doteq \gamma_n^2$, and introduce the following

ASSUMPTIONS. $\{v_n\}_{n\geq 1}$ is a sequence of non-negative and integer-valued r.v.'s such that

(A) v_n is independent of X_i, $i = 1, 2, \ldots$;

(B) for all $n = 1, 2, \ldots$, $0 < \lambda_n$, $\gamma_n^2 < \infty$, and

$$\lim_{n\to\infty} \lambda_n = \infty, \qquad \lim_{n\to\infty} \frac{\gamma_n^2}{\lambda_n} = \beta \geq 0;$$

(C) either v_n is a degenerate r.v. for all n or $\gamma_n^2 > 0$, $v_n/\lambda_n \to 1$ in probability and $(v_n - \lambda_n)/\gamma_n \to Z \sim N(0, 1)$ in distribution ($\xrightarrow{d}$) as $n \to \infty$.

First, we shall focus on the process

$$\boldsymbol{G}^*_{v_n} \doteq \frac{1}{\sqrt{\lambda_n}}\Big(\sum_{j=1}^{v_n} \delta_{X_j} - \lambda_n P\Big) = \sqrt{\lambda_n}\left(\frac{1}{\lambda_n}\boldsymbol{N}_{v_n} - P\right), \qquad \boldsymbol{N}_{v_n} \doteq \sum_{j=1}^{v_n} \delta_{X_j}, \tag{6}$$

indexed by a collection $\mathscr{F}$ of measurable functions, for any probability measure P defined on $(\mathscr{X}, \mathscr{A})$ and for v_n satisfying (A)–(C). Let us stress that, in particular, $(\mathscr{X}, \mathscr{A})$ may be a product space, which implies that X_i's may be random vectors whose components need not to be independent. If we take $\mathscr{F}$ to be the class of indicator functions, $\boldsymbol{G}^*_{v_n}$ is hence the natural candidate for a general goodness of fit test, in the presence of a random sample size. The statistical importance of the empirical process, however, goes very much beyond that, as illustrated in part 3 of the book by van der Vaart and Wellner [14] or in van der Vaart [13]; their applications include, for instance, the asymptotic theory for M- and Z-estimators, the two-samples problem, testing for independence, and many other issues in parametric and non-parametric inference in the presence of i.i.d. observations.

A special case of (6) for $v_n \sim \text{Poisson}(n)$ is the functionally indexed version of the so-called *Kac empirical process*, i.e.

$$\boldsymbol{K}_{v_n} \doteq \frac{1}{\sqrt{n}}\Big(\sum_{j=1}^{v_n} \delta_{X_j} - nP\Big) = \sqrt{n}\left(\frac{1}{n}\boldsymbol{N}_{v_n} - P\right). \tag{7}$$

Klaassen and Wellner [11] show that if the class $\mathscr{F}$ is P-Donsker with finite envelope function and such that $\|P\|_{\mathscr{F}} = \sup\{|Pf|\colon f \in \mathscr{F}\} < \infty$, then

$$\boldsymbol{K}_{v_n} \Rightarrow \boldsymbol{S} \text{ in } l^\infty(\mathscr{F}) \quad \text{as } n \to \infty. \tag{8}$$

In this case $\mathscr{F}$ is called a *P-Kac class*. Our first result can also be viewed as an extension of (8) to more general sequences $\{v_n\}_{n\geq 1}$. For any distribution of v_n,

we define N_{ν_n}/λ_n to be a *modified Kac empirical distribution function* and we denote it by $P^*_{\nu_n}$ (in the sequel we use the asterisk to stress the non-random nature of the normalizing factor).

The second and third empirical fields considered here arise more directly when testing for independence (see Section 3 for details). They involve the case of a product space $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ and corresponding i.i.d. random elements $X_i = (X_{1i}, X_{2i})$, $i = 1, 2, \ldots$, having distribution P and marginals P_1 and P_2. We define

$$\text{(9)} \qquad \begin{aligned} Z^*_{\nu_n}(f_1, f_2) &\doteq \sqrt{\lambda_n}(P^*_{\nu_n} - P^*_{1,\nu_n} \times P_2)(f_1 \times f_2) \\ &= \sqrt{\lambda_n}\left(\frac{1}{\lambda_n}\sum_{i=1}^{\nu_n} \delta_{(X_{1i}, X_{2i})} - \frac{1}{\lambda_n}\sum_{i=1}^{\nu_n} \delta_{X_{1i}} \times P_2\right)(f_1 \times f_2). \end{aligned}$$

The field $Z^*_{\nu_n}$ is indexed by $\mathcal{F} = (\mathcal{F}_1 \times \mathcal{F}_2)$, by which we denote the class of all measurable functions $f_1 \times f_2 \colon \mathcal{X}_1 \times \mathcal{X}_2 \mapsto \boldsymbol{R}$ such that

$$(f_1 \times f_2)(x_1, x_2) = f_1(x_1) f_2(x_2),$$

where f_1 and f_2 belong to $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. Such a field arises, for instance, when testing for independence under the assumption that the first marginal is assumed to be unknown, and hence estimated from the data, and the second is known.

Finally, we introduce the empirical field $T^*_{\nu_n}$ which generalizes the so-called *independence empirical process* to the case of a random index. We define

$$\text{(10)} \qquad \begin{aligned} T^*_{\nu_n}(f_1 \times f_2) &\doteq \sqrt{\lambda_n}\left[(P^*_{\nu_n} - P^*_{1,\nu_n} \times P^*_{2,\nu_n}) - (P - P_1 \times P_2)\right](f_1 \times f_2) \\ &= \sqrt{\lambda_n}\left[\left(\frac{1}{\lambda_n}\sum_{j=1}^{\nu_n} \delta_{(X_{1j}, X_{2j})} - \frac{1}{\lambda_n^2}\sum_{j=1}^{\nu_n} \delta_{X_{1j}} \sum_{l=1}^{\nu_n} \delta_{X_{2l}}\right) - (P - P_1 \times P_2)\right](f_1 \times f_2), \end{aligned}$$

where P is any measure on $(\mathcal{X}_1 \times \mathcal{X}_2)$, and P_1 and P_2 are its marginals on $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. The indexing class of functions is again $\mathcal{F} = (\mathcal{F}_1 \times \mathcal{F}_2)$ and the second terms in (10) obviously vanish under the null hypothesis of independence. Again, this field is interesting for statistical applications, when testing for independence under the assumption that both marginals are unknown.

2. CONVERGENCE RESULTS

Our first result in this section is the following

THEOREM 2.1. *Under the assumptions* (A)–(C) *and for any P-Donsker class* $\mathcal{F}$ *such that* $\|P\|_{\mathcal{F}} < \infty$, *as* $n \to \infty$,

$$\text{(11)} \qquad G^*_{\nu_n} \Rightarrow G_\beta \quad \text{in } l^\infty(\mathcal{F}).$$

The limiting field G_β *is Gaussian, centered and with covariance*

$$\text{(12)} \qquad EG_\beta f G_\beta g = Pfg - (1-\beta) PfPg.$$

Proof. From (6) and (5) we have

$$G^*_{\nu_n} = \sqrt{\frac{\nu_n}{\lambda_n}}(G_{\nu_n} - G_{[\lambda_n]}) + \left(\sqrt{\frac{\nu_n}{\lambda_n}} - 1\right) G_{[\lambda_n]} + G_{[\lambda_n]} + \sqrt{\lambda_n}\left(\frac{\nu_n}{\lambda_n} - 1\right) P \tag{13}$$

([x] is the integer part of x).

The first summand converges to zero in probability by Theorem 3.5.3 of van der Vaart and Wellner [14], by assumption (C) and by the extended Slutsky's lemma (see van der Vaart and Wellner [14], p. 32).

Likewise, the second summand is negligible by assumptions (B) and (C), by (2) and again the extended Slutsky's lemma.

Finally, the third and fourth summands are independent; the former converges weakly to G in view of (B) and (2). For the latter we have

$$\sqrt{\lambda_n}\left(\frac{\nu_n}{\lambda_n} - 1\right) = \frac{\nu_n - \lambda_n}{\gamma_n} \frac{\gamma_n}{\sqrt{\lambda_n}} \xrightarrow{d} \sqrt{\beta}\, Z, \tag{14}$$

where $Z \sim N(0, 1)$, so that we get

$$G^*_{\nu_n} \Rightarrow G + \sqrt{\beta}\, ZP \stackrel{d}{=} G_\beta. \tag{15}$$

Since G_β is a tight element of $l^\infty(\mathscr{F})$, to check that the last equality holds, it is enough to look at the finite-dimensional distributions. As both fields are centered Gaussian, we just need to note that, for any $f, g \in \mathscr{F}$, the covariance function $E[G(f) + \sqrt{\beta}\, ZPf][G(g) + \sqrt{\beta}\, ZPg]$ coincides with (12), since Z is independent of G. ∎

From the previous result, we can obtain the following interesting particular cases:

(i) $\nu_n \sim \text{Poisson}(\lambda_n)$: $\beta = 1$, $G^*_{\nu_n} \Rightarrow S$;

(ii) $\nu_n = n$: $\beta = 0$, $G^*_{\nu_n} \Rightarrow G$;

(iii) $\nu_n \sim \text{Bin}(n, p)$: $\beta = 1 - p$, $G^*_{\nu_n} \Rightarrow G_{1-p}$.

Case (i) corresponds to a generalization of (8), while (ii) is the standard result for the deterministic case (see (2)). Case (iii) is of particular interest for the wide applicability of the binomial distribution. Finally, we remark that G_β is a generalization of the *P-Brownian bridge*, to which it reduces for $\beta = 0$ in (12).

We now study the weak convergence of $Z^*_{\nu_n}$, defined in (9) on the space $\mathscr{X} = (\mathscr{X}_1 \times \mathscr{X}_2)$ of the two-dimensional random element (X_1, X_2), whose components are assumed to be independent (under the null hypothesis).

THEOREM 2.2. *Let $\mathscr{F} = (\mathscr{F}_1 \times \mathscr{F}_2)$ be P-Donsker with $P = P_1 \times P_2$ and $\|P\|_{\mathscr{F}} < \infty$. Then, for any distribution of ν_n satisfying* (A)–(C),

$$Z^*_{\nu_n} \Rightarrow Z_P \quad \text{in } l^\infty(\mathscr{F}_1 \times \mathscr{F}_2),$$

where Z_P is the P-Kiefer–Müller process on $(\mathscr{F}_1 \times \mathscr{F}_2)$. This field is defined to be centered Gaussian, and with covariance

$$EZ_P(f_1 \times f_2) Z_P(g_1 \times g_2) = P_1 f_1 g_1 (P_2 f_2 g_2 - P_2 f_2 P_2 g_2). \tag{16}$$

Proof. By adding and subtracting $\sqrt{\lambda_n}(P_1 \times P_2)(f_1 \times f_2)$, we rewrite (9) as

$$\sqrt{\lambda_n}(P^*_{\nu_n} - P_1 \times P_2)(f_1 \times f_2) - \sqrt{\lambda_n}(P^*_{1,\nu_n} - P_1) f_1 P_2 f_2. \tag{17}$$

Since, by assumption, $\mathscr{F} = (\mathscr{F}_1 \times \mathscr{F}_2)$ is P-Donsker, we can apply Theorem 2.1, which, together with Lemma 1.4.4 of van der Vaart and Wellner [14], entails that $Z^*_{\nu_n}$ is asymptotically tight.

Then we can establish the weak convergence of (17) by analyzing the convergence of the finite-dimensional distributions of $Z^*_{\nu_n}$ as follows. We consider the k-dimensional vector $(Z^*_{\nu_n}(f'_1 \times f'_2), \ldots, Z^*_{\nu_n}(f^{(k)}_1 \times f^{(k)}_2))$ whose j-th element is defined as

$$\begin{aligned} Z^*_{\nu_n}(f^{(j)}_1 \times f^{(j)}_2) &= \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\nu_n} (\delta_{(X_{1i}, X_{2i})} - \delta_{X_{1i}} \times P_2)(f^{(j)}_1 \times f^{(j)}_2) \\ &= \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\nu_n} f^{(j)}_1(X_{1i}) [f^{(j)}_2(X_{2i}) - P_2 f^{(j)}_2] = \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\nu_n} \xi^{(j)}_i. \end{aligned}$$

By assumption, $\xi_1, \xi_2, \ldots$ are i.i.d. random vectors and $E\xi^{(j)}_i = 0$,

$$E\xi^{(j)}_i \xi^{(l)}_i = P_1 f^{(j)}_1 f^{(l)}_1 (P_2 f^{(j)}_2 f^{(l)}_2 - P_2 f^{(j)}_2 P_2 f^{(l)}_2) \doteq \Theta_{jl}.$$

Therefore, by the multivariate Central Limit Theorem and by (A)–(C), we get $\lambda_n^{-1/2} \sum_{i=1}^{\nu_n} \xi_i \xrightarrow{d} Y \sim N_k(0, \Theta)$, where Θ has Θ_{jl} as a generic element. ∎

We stress that the limiting field Z_P is independent of β: this implies that the empirical process $Z^*_{\nu_n}$ behaves in the same way, as $n \to \infty$, whatever the distribution of ν_n, and therefore it is asymptotically equivalent to the process arising in the presence of a deterministic sample size.

Finally, we consider the third empirical field $T^*_{\nu_n}$, defined in (10) and indexed by $(\mathscr{F}_1 \times \mathscr{F}_2)$. Recall that now, contrary to the previous theorems, P is an arbitrary measure on $\mathscr{X} = \mathscr{X}_1 \times \mathscr{X}_2$ with marginals P_1, P_2.

THEOREM 2.3. *Let ν_n satisfy* (A)–(C) *and let $\mathscr{F}_1 \times \mathscr{F}_2$ be P-Donsker; if, moreover, $\|P_1\|_{\mathscr{F}_1} < \infty$ and $\|P_2\|_{\mathscr{F}_2} < \infty$, we get*

$$T^*_{\nu_n} \Rightarrow T_{P,\beta} \quad \text{in } l^\infty(\mathscr{F}_1 \times \mathscr{F}_2). \tag{18}$$

The limiting field $T_{P,\beta}$ is defined on $(\mathscr{F}_1 \times \mathscr{F}_2)$ by

$$T_{P,\beta} \doteq T_P + \sqrt{\beta} Z [P - 2P_1 \times P_2], \tag{19}$$

where $Z \sim N(0, 1)$ is independent of T_P,

$$T_P(f_1 \times f_2) \doteq G_P((f_1 - P_1 f_1) \times (f_2 - P_2 f_2)), \tag{20}$$

and G_P is the P-Brownian bridge on $\mathscr{F} = (\mathscr{F}_1 \times \mathscr{F}_2)$.

Proof. By adding and subtracting $\sqrt{\lambda_n}(P_1 \times P^*_{2,\nu_n})(f_1 \times f_2)$, we obtain for (10)

$$
\begin{aligned}
(21)\quad & T^*_{\nu_n}(f_1 \times f_2) \\
&= \sqrt{\lambda_n}(P^*_{\nu_n} - P)(f_1 \times f_2) - \sqrt{\lambda_n}(P^*_{1,\nu_n} - P_1) f_1 P^*_{2,\nu_n} f_2 - P_1 f_1 \sqrt{\lambda_n}(P^*_{2,\nu_n} - P_2) f_2 \\
&= \sqrt{\lambda_n}(P^*_{\nu_n} - P)\{(f_1 - P_1 f_1) \times (f_2 - P_2 f_2) - [P_1 f_1 \times P_2 f_2]\} \\
&\quad - \sqrt{\lambda_n}(P^*_{1,\nu_n} - P_1) f_1 (P^*_{2,\nu_n} - P_2) f_2 .
\end{aligned}
$$

The second step can be verified by considering that the following equalities hold:

$$P^*_{\nu_n}(f_1 \times P_2 f_2) = P^*_{1,\nu_n} f_1 P_2 f_2 \quad \text{and} \quad P^*_{\nu_n}(P_1 f_1 \times f_2) = P_1 f_1 P^*_{2,\nu_n} f_2 .$$

On the other hand, we note that

$$P^*_{\nu_n}(P_1 f_1 \times P_2 f_2) \neq P(P_1 f_1 \times P_2 f_2),$$

since $P^*_{\nu_n}$ is not a probability measure (as P_n would be in the deterministic case).

The second term of (21) is asymptotically negligible. Indeed, from the assumption that $\mathscr{F}_1 \times \mathscr{F}_2$ is P-Donsker it follows that $\mathscr{F}_1$ and $\mathscr{F}_2$ are Donsker with respect to the corresponding marginal measure (see van der Vaart and Wellner [14], Theorem 2.10.1). We can hence apply Theorem 2.1 to prove that $\sqrt{\lambda_n}(P^*_{1,\nu_n} - P_1) \Rightarrow G_\beta$ in $l^\infty(\mathscr{F}_1)$. On the other hand, the uniform version of the law of large numbers also holds (see van der Vaart and Wellner [14], p. 82), so that $\|P^*_{2,\nu_n} f - P_2 f\|_{\mathscr{F}_2}$ converges to zero in outer probability.

For the first term in (21) we get

$$
\begin{aligned}
&\sqrt{\frac{\nu_n}{\lambda_n}}\left[\frac{1}{\sqrt{\nu_n}} \sum_{j=1}^{\nu_n} \delta_{(X_{1j}, X_{2j})} - \sqrt{\nu_n} P\right](f_1 - P_1 f_1) \times (f_2 - P_2 f_2) \\
&\qquad + \sqrt{\lambda_n}\left(\frac{\nu_n}{\lambda_n} - 1\right)[P(f_1 - P_1 f_1) \times (f_2 - P_2 f_2) - (P_1 f_1 \times P_2 f_2)] \\
&\qquad \Rightarrow G_P[(f_1 - P_1 f_1) \times (f_2 - P_2 f_2)] + \sqrt{\beta} Z[P - 2P_1 \times P_2](f_1 \times f_2),
\end{aligned}
$$

in view of the assumption that $(\mathscr{F}_1 \times \mathscr{F}_2)$ is P-Donsker and by applying a similar argument to that in Theorem 2.1, i.e. by adding and subtracting,

$$\left(\sqrt{\frac{\nu_n}{\lambda_n}} - 1\right)\left(\frac{1}{\sqrt{[\lambda_n]}} \sum_{j=1}^{[\lambda_n]} \delta_{(X_{1j}, X_{2j})} - \sqrt{[\lambda_n]} P\right)(f_1 - P_1 f_1) \times (f_2 - P_2 f_2). \quad \blacksquare$$

The covariance of the limiting field is readily obtained as

$$
\begin{aligned}
(22)\quad & E T_{P,\beta}(f_1 \times f_2)\, T_{P,\beta}(g_1 \times g_2) \\
&\quad = E\{G_P[(f_1 - P_1 f_1) \times (f_2 - P_2 f_2)]\, G_P[(g_1 - P_1 g_1) \times (g_2 - P_2 g_2)]\} \\
&\qquad + \beta[P - 2P_1 \times P_2](f_1 \times f_2)[P - 2P_1 \times P_2](g_1 \times g_2)
\end{aligned}
$$

$$= P\{[(f_1 - P_1 f_1) \times (f_2 - P_2 f_2)]\,[(g_1 - P_1 g_1) \times (g_2 - P_2 g_2)]\}$$
$$- P[(f_1 - P_1 f_1) \times (f_2 - P_2 f_2)]\, P[(g_1 - P_1 g_1) \times (g_2 - P_2 g_2)]$$
$$+ \beta[P - 2P_1 \times P_2](f_1 \times f_2)[P - 2P_1 \times P_2](g_1 \times g_2).$$

Under the assumption that X_1 and X_2 are independent, the last line on the right-hand side of (22) is equal to $\beta P_1 f_1 P_2 f_2 P_1 g_1 P_2 g_2$, whereas the previous one vanishes, so that (22) reduces to

(23) $(P_1 f_1 g_1 - P_1 f_1 P_1 g_1)(P_2 f_2 g_2 - P_2 f_2 P_2 g_2) + \beta P_1 f_1 P_2 f_2 P_1 g_1 P_2 g_2.$

Again the result for the deterministic case can be obtained anew by putting simply $\beta = 0$ in the previous theorem. Indeed, under these circumstances, (18) and (19) give the convergence of the independence empirical process to the field $\boldsymbol{T}$ (see van der Vaart and Wellner [14], § 3.8).

3. STATISTICAL APPLICATIONS

The limiting fields obtained from $\boldsymbol{G}^*_{\nu_n}$ and $\boldsymbol{T}^*_{\nu_n}$ are generalizations of well-known Gaussian fields, depending upon the value of the parameter β. The intermediate case $\boldsymbol{Z}^*_{\nu_n}$ is, to some extent, different in that the limiting field does not depend on β, whatever the distribution of ν_n. A thorough analysis of these and related fields can be found in Beghin [1], where various inequalities for their maximal distributions are obtained.

We now present some statistical applications to independence testing procedures in the case of a random sample size. The statistical problem can be defined as follows. Let (X_{1i}, X_{2i}), $i = 1, \ldots, \nu_n$, be a sample from the two-dimensional r.v. (X_1, X_2) with distribution function $F(t_1, t_2)$. We are interested in testing the null hypothesis

$$H_0\colon F(t_1, t_2) = F(t_1) F_2(t_2) \text{ for any } t_1, t_2 \in [0, 1].$$

In the case where both marginals are known, we will use the two-dimensional version of $\boldsymbol{G}^*_{\nu_n}$, which can be obtained by choosing $\mathscr{F} = \{1_{(0,t]}\colon t \in [0, 1]^2\}$. By the inverse transform argument, we can assume, without any loss of generality, that P is uniform on $[0, 1]^2$; the sample space is $([0, 1]^2, \mathscr{B}_{[0,1]^2})$, so that (6) reduces to

$$\boldsymbol{G}^*_{\nu_n}(t_1, t_2) = \sqrt{\lambda_n}\left(\frac{1}{\lambda_n}\sum_{i=1}^{\nu_n} 1_{\{X_{1i} \leqslant t_1, X_{2i} \leqslant t_2\}} - t_1 t_2\right). \tag{24}$$

As well known, the class of indicator functions on $[0, 1]^2$ is P-Donsker with $P = \mathrm{Unif}[0, 1]^2$, so that we can apply Theorem 2.1. Therefore, for ν_n satisfying the assumptions (A)–(C), (24) weakly converges, in $l^\infty(\mathscr{F})$, to the two-dimensional Gaussian field $\boldsymbol{G}_\beta$, which, in this case, can be viewed as a generalization

of the *pinned Brownian sheet*. From (12) its covariance function is readily obtained as

$$(25) \qquad E\boldsymbol{G}_\beta(t_1, t_2)\boldsymbol{G}_\beta(s_1, s_2) = (t_1 \wedge s_1)(t_2 \wedge s_2) - (1-\beta)t_1 s_1 t_2 s_2.$$

Moreover, for the common case where $\beta < 1$ (see examples (ii) and (iii) in Section 2), $\boldsymbol{G}_\beta$ is equal in distribution to the restriction on $[0, 1]^2$ of the field

$$(26) \qquad \boldsymbol{G}'_u(t_1, t_2) \doteq \boldsymbol{S}(t_1, t_2) - \frac{t_1 t_2}{u^2}\boldsymbol{S}(u, u), \qquad 0 < t_1, t_2 < u,$$

for $u \doteq 1/\sqrt{1-\beta}$. The field (26) is the two-dimensional analogue of the so-called *Brownian bridge of length u* (see Nikitin [12], Beghin and Orsingher [3]): indeed, $\boldsymbol{G}'_u$ is a.s. equal to zero at the point (u, u) which is outside of the unit square. It can hence be viewed as a Brownian sheet $\boldsymbol{S}(t_1, t_2)$ conditioned upon $\{\boldsymbol{S}(u, u) = 0\}$; the classical Brownian sheet is obtained as a special case for $u \to +\infty$ (or equivalently $\beta \to 1^-$ for $\boldsymbol{G}_\beta$) (example (i) of Section 2). On the other hand, for $u \to 1^+$ (or $\beta \to 0^+$) we obtain again the pinned Brownian sheet which vanishes at (1, 1) (example (ii) in Section 2). Many other circumstances are intermediate, for instance the binomial case (example (iii)). For $\beta > 1$, $\boldsymbol{G}_\beta$ is however well defined, but the representation (26) is no more valid.

The two-dimensional version of the second empirical process $\boldsymbol{Z}^*_{\nu_n}$ is obtained by choosing, for (9), $(\mathscr{X}_i, \mathscr{A}_i) = ([0, 1], \mathscr{B}_{[0,1]})$ and $\mathscr{F}_i = \{1_{(0,t]}: t \in [0, 1]\}$, $i = 1, 2$. Again we assume, without any loss of generality, that $P_2 = \mathrm{Unif}[0, 1]$; from Theorem 2.2 we hence obtain the weak convergence of

$$(27) \qquad \boldsymbol{Z}^*_{\nu_n}(t_1, t_2) = \sqrt{\lambda_n}\left(\frac{1}{\lambda_n}\sum_{i=1}^{\nu_n} 1_{\{X_{1i} \leq t_1, X_{2i} \leq t_2\}} - \frac{1}{\lambda_n}\sum_{i=1}^{\nu_n} 1_{\{X_{1i} \leq t_1\}} t_2\right)$$

to the two-dimensional *Kiefer–Müller process*, i.e. the zero-mean Gaussian process with covariance

$$(28) \qquad E\boldsymbol{Z}(t_1, t_2)\boldsymbol{Z}(s_1, s_2) = (t_1 \wedge s_1)[(t_2 \wedge s_2) - t_2 s_2].$$

As well known, $\boldsymbol{Z}$ vanishes a.s. along the line $t_2 = 1$. We stress again that this limiting field does not depend on β, and hence on the distribution of ν_n.

The last empirical field we have analyzed arises when both marginals are unknown: we specify $(\mathscr{X}_i, \mathscr{A}_i) = ([0, 1], \mathscr{B}_{[0,1]})$, $\mathscr{F}_i = \{1_{(0,t]}: t \in [0, 1]\}$ and $P_i = \mathrm{Unif}[0, 1]$, $i = 1, 2$. Then (10) becomes

$$(29) \quad \boldsymbol{T}^*_{\nu_n}(t_1, t_2) = \sqrt{\lambda_n}\left(\frac{1}{\lambda_n}\sum_{i=1}^{\nu_n} 1_{\{X_{1i} \leq t_1, X_{2i} \leq t_2\}} - \frac{1}{\lambda_n^2}\sum_{i=1}^{\nu_n} 1_{\{X_{1i} \leq t_1\}} \sum_{l=1}^{\nu_n} 1_{\{X_{2l} \leq t_2\}}\right).$$

From Theorem 2.3 we infer that (30) converges weakly, in $l^\infty(\mathscr{F}_1 \times \mathscr{F}_2)$, to the zero-mean Gaussian field $\boldsymbol{T}_\beta$ with covariance

$$(30) \quad E\boldsymbol{T}_\beta(t_1, t_2)\boldsymbol{T}_\beta(s_1, s_2) = [(t_1 \wedge s_1) - t_1 s_1][(t_2 \wedge s_2) - t_2 s_2] + \beta t_1 s_1 t_2 s_2.$$

We have

$$T_\beta(t_1, t_2) = T(t_1, t_2) - \sqrt{\beta}\, t_1 t_2 S(1, 1)$$
$$= S(t_1, t_2) - t_1 S(1, t_2) - t_2 S(t_1, 1) + (1 - \sqrt{\beta})\, t_1 t_2 S(1, 1),$$

so that T_β can be viewed as a straightforward generalization of the well-known *tucked Brownian sheet* T, which corresponds to the special case for $\beta = 0$. T vanishes a.s. on the lines $t_i = 1$, $i = 1, 2$, while T_β does not share this property.

The distributions of some functionals of the previous fields are exploited in Beghin and Nikitin [2], when evaluating the asymptotic Bahadur efficiency of some independence test statistics.

REFERENCES

[1] L. Beghin, *On the maximum of some conditional and integrated Gaussian fields*, Technical Reports Dip. Statistica, Probabilità, Stat. Appl. 12, Univ. di Roma "La Sapienza" 2001, submitted for publication.

[2] L. Beghin and Ya. Yu. Nikitin, *Approximate asymptotic Bahadur efficiency of independence tests when the sample size is random*, J. Italian Statistical Society 8 (1) (1999), pp. 1–23.

[3] L. Beghin and E. Orsingher, *On the maximum of the generalized Brownian bridge*, Lithuanian Math. J. 39 (2) (1999), pp. 200–213.

[4] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York 1968.

[5] S. Csörgo, *On weak convergence of the empirical process with random sample size*, Acta Sci. Math. 36 (1974), pp. 14–25.

[6] S. Csörgo, *Strong approximations of empirical Kac processes*, Ann. Inst. Statist. Math. 33 (3) (1981), pp. 417–423.

[7] R. M. Dudley, *Uniform Central Limit Theorems*, Cambridge Stud. Adv. Math., New York 1999.

[8] B. V. Gnedenko and V. Yu. Korolev, *Random Summation: Limit Theorems and Applications*, CRC Press, New York 1996.

[9] J. Hoffman-Jørgensen, *Stochastic Processes in Polish Spaces*, Various Publ. Ser. 39, Aarhus Univ., 1991.

[10] M. Kac, *On deviations between theoretical and empirical distributions*, Proc. Nat. Acad. Sci. U.S.A. 35 (1949), pp. 252–257.

[11] C. A. Klaassen and J. A. Wellner, *Kac empirical processes and the bootstrap*, in: *Proceedings of the Eight International Conference on Probability in Banach Spaces*, M. Hahn and J. Kuelbs (Eds.), 1992, pp. 411–429.

[12] Ya. Yu. Nikitin, *Limit distributions and comparative asymptotic efficiency of the Kolmogorov–Smirnov statistics with random index*, J. Soviet Math. 2 (1981), pp. 1042–1049.

[13] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, Cambridge 1998.

[14] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, Springer, New York 1996.

Dipartimento di Statistica, Probabilità e Statistiche Applicate
Università di Roma "La Sapienza"
Piazzale Aldo Moro 5, 00185 Roma (Italy)
E-mail: luisa.beghin@uniromal.it

Received on 20.3.2001;
revised version on 5.10.2002