PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 33, Fasc. 1 (2013), pp. 93–106

JOINT CLUSTER COUNTS FROM UNIFORM DISTRIBUTION

BY

ÉVA OSZTÉNYINÉ KRAUCZI (KECSKEMÉT)

Abstract. We consider a vector of numbers of clusters at different distance levels of n independent identically distributed random variables uniformly distributed on $[a, b]$. We prove asymptotic normality of this vector when the ends a, b are known or are estimated from the sample. Basing on these asymptotic results we propose new tests for uniformity, called *cluster tests*. We also present results of a simulation study showing empirical behaviour of these tests.

2000 AMS Mathematics Subject Classification: Primary: 62G10; Secondary: 62G20.

Key words and phrases: Asymptotic normality, clusters of independent uniform random variables, number of clusters, testing for uniformity.

1. INTRODUCTION

Let $U_1, U_2, \ldots$ be independent random variables, each uniformly distributed in the unit interval $[0, 1]$. For each $n \in \mathbb{N}$, let $U_{1,n} \leqslant \ldots \leqslant U_{n,n}$ be the order statistics pertaining to the sample $U_1, \ldots, U_n$. The elements of the sample are almost surely different, so $U_{1,n} < \ldots < U_{n,n}$ almost surely. Given a deterministic threshold $d_n \in (0, 1)$, the sequence $U_1, \ldots, U_n$ breaks up into nonempty disjoint clusters $C_{1,n}, \ldots, C_{K_n,n}$ at level d_n, where a random integer $K_n \in \{1, \ldots, n\}$ is the number of clusters. The distance between any two neighbouring elements of $C_{k,n} = \{U_{N_{0,n}+\ldots+N_{k-1,n}+1,n}, \ldots, U_{N_{1,n}+\ldots+N_{k,n},n}\}$ is not greater than d_n, $k = 1, \ldots, K_n$, where $N_{k,n} = |C_{k,n}|$ is a number of elements in $C_{k,n}$ and $N_{0,n} = 0$, and if $K_n > 1$, then $U_{N_{1,n}+\ldots+N_{k-1,n}+1,n} - U_{N_{1,n}+\ldots+N_{k-1,n},n} > d_n$ for $k = 2, \ldots, K_n$.

In the terminology of random graphs, the random variables $U_1, \ldots, U_n$ and the distance level d_n generate the random interval graph $\mathcal{G}_n = \mathcal{G}(U_1, \ldots, U_n; d_n)$. The vertex set of $\mathcal{G}_n$ is the set $\{1, \ldots, n\}$ representing $U_1, \ldots, U_n$. Between two different vertices i and j there is an edge if and only if $|U_i - U_j| < d_n$, where $i, j \in \{1, \ldots, n\}$. In this language the number of clusters K_n is the number of connected components.

Godehardt and Jaworski [3] studied a uniform model for random interval graphs on the unit interval, and derived an exact formula for the number of clusters K_n. Apart from the exact description of K_n, the question of the asymptotic behaviour of K_n naturally arises, thus from now on it is assumed that $d_n \to 0$. While asymptotic theorems were already derived by Godehardt and Jaworski in [3], the question was investigated further; for example, Csörgő and Wu [1] described all possible asymptotic distributions of K_n.

Godehardt and Jaworski [3] showed that if $n^2 d_n \to 0$, then $n - K_n \to 0$ almost surely, namely there are no edges in $\mathcal{G}_n$ if d_n is small enough. They studied the number of clusters of order l, $l \in \mathbb{N}$, and the size of a cluster containing a given element of the sample $U_1, \dots, U_n$ on further d_n's. Csörgő and Wu [1] derived the asymptotic distribution of K_n on these further d_n's.

All convergence relations are understood throughout the paper as $n \to \infty$ unless otherwise specified, and let $\xrightarrow{\mathcal{D}}$ denote convergence in distribution. Letting $\mathcal{N}(\mu, \sigma^2)$ denote a normal random variable with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$ and denoting by $\Phi(\cdot)$ the distribution function of $\mathcal{N}(0,1)$, Csörgő and Wu showed, in particular, the following theorem.

THEOREM 1.1 (Csörgő and Wu [1]). (i) *If $nd_n \to 0$ and $n^2 d_n \to \infty$, then*

$$\begin{aligned}\Delta_n &= \sup_{x\in\mathbb{R}} \left| P\left(\frac{K_n - ne^{-nd_n}}{\sqrt{ne^{-nd_n}(1-e^{-nd_n})}} \leqslant x \right) - \Phi(x) \right| \\ &= O\left(\sqrt{(nd_n + \varepsilon_n) \log \frac{1}{nd_n}} + \frac{\log(n\sqrt{d_n})}{n\sqrt{d_n}} \right),\end{aligned}$$

where $\varepsilon_n = \sqrt{(4\log n)/n}$, and so $(K_n - ne^{-nd_n})/(n\sqrt{d_n}) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$.

(ii) *If $0 < \liminf_n nd_n \leqslant \limsup_n nd_n < \infty$, then*

$$\sup_{x\in\mathbb{R}} \left| P\left(\frac{K_n - ne^{-nd_n}}{\sqrt{ne^{-2nd_n}(e^{nd_n} - 1 - n^2 d_n^2)}} \leqslant x \right) - \Phi(x) \right| = O\left(\frac{\log^{3/4} n}{n^{1/4}} \right),$$

and hence if $nd_n \to c \in (0,\infty)$, then $(K_n - ne^{-nd_n})/\sqrt{n} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2)$, where $\sigma^2 = e^{-2c}[e^c - 1 - c^2]$.

(iii) *If $nd_n \to \infty$ and $ne^{-nd_n} \to \infty$, then*

$$\Delta_n = O\left(\frac{(nd_n)^{3/2}}{\sqrt{e^{nd_n}}} + \sqrt{\varepsilon_n nd_n \log(ne^{-nd_n})} + \sqrt{\frac{e^{nd_n}}{n}} \log(ne^{-nd_n}) \right),$$

where Δ_n is as in the case (i) *and $\varepsilon_n = \sqrt{(4\log n)/n}$ again, and so*

$$\frac{K_n - ne^{-nd_n}}{\sqrt{ne^{-nd_n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

In this paper we extend the results of Csörgő and Wu [1] to multivariate limit theorems. We also apply them for testing uniformity on a known and unknown interval.

In Section 2 we collect multivariate limit theorems. Subsection 2.1 aims at proving the multivariate version of Theorem 1.1, hoping that we can obtain more information about the sample. The next subsection contains the extension of this theorem to random variables with the uniform distribution on an arbitrary interval $[a, b]$, $a, b \in \mathbb{R}$, $a < b$. In Subsection 2.3 we investigate what happens to the asymptotic distribution when the sample comes from the uniform distribution on an unknown interval. Statistical applications are given in Section 3, where we also use the multivariate theorem to test for uniformity on the unit interval $[0, 1]$ and we perform some power investigation. We also suggest a test for uniformity on an unknown interval and present a simulation study to evaluate the power of this test.

2. THEORETICAL RESULTS

2.1. Asymptotic distribution of joint cluster counts from the uniform distribution on the unit interval. As was said above, Csörgő and Wu [1] showed that K_n is asymptotically normal for three cases of distance level rates of convergence to zero. Here we are interested in the joint behaviour of K_n's for a sequence of different distance levels coming from these three cases.

Set $J \geqslant 1$ and let $d_{n1} \leqslant d_{n2} \leqslant \ldots \leqslant d_{nJ}$ be distance levels satisfying one of the following conditions:

(i) $nd_{nj} \to 0$, $n^2 d_{nj} \to \infty$;

(ii) $0 < \liminf_n nd_{nj} \leqslant \limsup_n nd_{nj} < \infty$;

(iii) $nd_{nj} \to \infty$, $ne^{-nd_{nj}} \to \infty$.

Let $K_{nj}(d_{nj})$ be numbers of clusters as described in Section 1 corresponding to the distance levels d_{nj}, $j = 1, \ldots, J$. Set $m_{nj} = ne^{-nd_{nj}}$,

$$\sigma_{nj}^2 = e^{-2nd_{nj}}(e^{nd_{nj}} - 1 - n^2 d_{nj}^2), \tag{2.1}$$

and

$$\mathbf{K}_n = \frac{1}{\sqrt{n}}\left(\frac{K_{n1}(d_{n1}) - m_{n1}}{\sigma_{n1}}, \ldots, \frac{K_{nJ}(d_{nJ}) - m_{nJ}}{\sigma_{nJ}}\right). \tag{2.2}$$

Then we have

THEOREM 2.1. *Under the above notation, the assumptions* (i)–(iii), *and*

$$e^{-nd_{ni}-nd_{nj}}(e^{nd_{ni}} - 1 - n^2 d_{ni} d_{nj})/\sigma_{ni}\sigma_{nj} \to s_{ij} \tag{2.3}$$

for all $1 \leqslant i < j \leqslant J$, *we have*

$$\mathbf{K}_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma), \tag{2.4}$$

where $\Sigma = (s_{ij})$ *with* $s_{jj} = 1$ *for all* $j = 1, \ldots, J$ *and* s_{ij}, $i \neq j$, *given by* (2.3).

Note that the matrix Σ is, in general, nonnegative definite and the limiting normal distribution may be degenerate, i.e. concentrated on some linear subspace of $\mathbb{R}^J$.

P r o o f. We start with some general statements.

Let (g_n) be a sequence of measurable functions on $\mathbb{R}$ and $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables such that $E\big(g_n(X_r)\big)=0$, $E\big(g_n^2(X_r)\big)=1$, and $E\left(|g_n(X_r)|^3\right) = o(\sqrt{n})$. Then the triangular array $Z_{nr} = g_n(X_r)$, $r = 1, \ldots, n$, $n = 1, 2, \ldots$, fulfills the Liapunov condition, and hence $(Z_{n1} + \ldots + Z_{nn})/\sqrt{n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$.

Using the above statement and the Cramér–Wold device we have the following multivariate CLT.

PROPOSITION 2.1. *Let $J \geqslant 1$ be a natural number and $\{g_{nj}(x)$, $j = 1, \ldots, J$, $n = 1, 2, \ldots\}$ be a collection of measurable functions. Suppose $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables such that $E\big(g_{nj}(X_r)\big) = 0$, $E\big(g_{nj}^2(X_r)\big) = s_{jj} = 1$, $E\left(|g_{nj}(X_r)|^3\right) = o(\sqrt{n})$ for every j, and $E\big(g_{ni}(X_r)g_{nj}(X_r)\big) \to s_{ij}$ for every $i \neq j$. Then the triangular array $Z_{n1}, \ldots, Z_{nn}$, $n = 1, 2, \ldots$, of random vectors $Z_{nr} = \big(g_{n1}(X_r), \ldots, g_{nJ}(X_r)\big)$ in $\mathbb{R}^J$ satisfies*

$$(Z_{n1} + \ldots + Z_{nn})/\sqrt{n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma),$$

where $\Sigma = (s_{ij})$.

The proof of Theorem 2.1 is a straightforward application of Proposition 2.1 and the argument given in Section 2.2 of Csörgő and Wu [1]. To see this, denote by $Y_1, Y_2, \ldots$ a sequence of i.i.d. exponentially distributed random variables with $P(Y_1 > x) = 1 - F(x) = e^{-x}$ and put for $j = 1, \ldots, J$ (cf. the definition of V_{jn} in [1], p. 407)

$$Z_{nrj} = \big(nd_{nj}e^{-nd_{nj}}(1 - Y_r) - [\mathbf{I}(Y_r \leqslant nd_{nj}) - F(nd_{nj})]\big)/\sigma_{nj},$$

where $\mathbf{I}(A)$ is the indicator of the event A, d_{nj} are distance levels as in Theorem 2.1, and σ_{nj} are given by (2.1). Then $Z_{nrj} = g_{nj}(Y_r)$ with

$$g_{nj}(x) = \big(nd_{nj}e^{-nd_{nj}}(1 - x) - [\mathbf{I}(x \leqslant nd_{nj}) - F(nd_{nj})]\big)/\sigma_{nj}.$$

Consequently, $E\big(g_{nj}(Y_r)\big) = 0$, and for $i \leqslant j$ we have

$$\begin{aligned} E\big(g_{ni}(Y_r)g_{nj}(Y_r)\big) &= E\left(Z_{nri}Z_{nrj}\right) \\ &= e^{-nd_{ni}-nd_{nj}}(e^{nd_{ni}} - 1 - n^2 d_{ni}d_{nj})/\sigma_{ni}\sigma_{nj} \end{aligned}$$

and, in particular, $E\big(g_{nj}^2(Y_r)\big) = 1$. Moreover, by the triangle inequality,

$$\Big(E\big(|g_{nj}(Y_r)|^3\big)\Big)^{1/3} = \big(E(|Z_{nrj}|^3)\big)^{1/3} \leqslant \frac{nd_{nj}}{\sigma_{nj}} e^{-nd_{nj}} \big(E(|Y_r - 1|^3)\big)^{1/3} + \frac{[e^{-nd_{nj}}(1 - e^{-nd_{nj}})]^{1/3}}{\sigma_{nj}},$$

which, by the conditions (i)–(iii), implies $E\left(|g_{nj}(Y_r)|^3\right) = o(\sqrt{n})$ for every j. From (2.3) and Proposition 2.1 we get

$$(Z_{n1} + \ldots + Z_{nn})/\sqrt{n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma).$$

Taking into account the decompositions (2.8) and (2.15) in [1] we see that

$$\mathbf{K}_n = M_n + R_n, \quad \text{where } M_n \stackrel{\mathcal{D}}{=} (Z_{n1} + \ldots + Z_{nn})/\sqrt{n}.$$

The convergence of R_n to zero in probability follows from the proof of Theorem 2.1 in [1] and the fact that $(e^{nd_{nj}} - 1 - n^2 d_{nj}^2)/(e^{nd_{nj}} - 1) \to 1$ for both cases (i) and (iii). This completes the proof of Theorem 2.1 ∎

Now, suppose $J \geqslant 2$ and $0 \leqslant J_1 \leqslant J_2 \leqslant J$ are such that distance levels d_{nj} in (2.2) satisfy the condition (i) for $j \leqslant J_1$ and the condition (iii) for $j > J_2$. Moreover, assume additionally

(A1) for $i < j \leqslant J_1$ it follows that $\sqrt{d_{ni}/d_{nj}} \to s_{ij}$;

(A2) for $J_1 < j \leqslant J_2$ it follows that $nd_{nj} \to c_j$;

(A3) for $J_2 < i < j$ it follows that $n(d_{nj} - d_{ni}) \to -2 \log s_{ij}$.

Then (2.3) is satisfied with s_{ij} given by (A1) and (A3), and specified in Corollary 2.1 below. Thus, we have the following

COROLLARY 2.1. *Under the conditions* (i), (iii), *and* (A1)–(A3), *the relation* (2.4) *holds*, *where*

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{pmatrix}$$

is the block diagonal matrix with $s_{jj} = 1$, s_{ij} *as in* (A1) *and* (A3) *for the first and third group*, *while for* $J_1 < i < j \leqslant J_2$

$$s_{ij} = (e^{c_i} - 1 - c_i c_j)/\sqrt{(e^{c_i} - 1 - c_i^2)(e^{c_j} - 1 - c_j^2)}.$$

Csörgő and Wu illustrated Theorem 1.1 by giving well-behaving examples called *typical sequences*. We apply Corollary 2.1 to these typical sequences suggested by Csörgő and Wu in [1], hereby we choose the parameters to obtain a

diagonal covariance matrix. A typical sequence $\{d_n\}$ for the case (i) is $d_n = n^{-\alpha}$ for some $\alpha \in (1,2)$. In particular, we take $d_{nj} = n^{-\alpha_j}$ for $j \leqslant J_1$, with $\alpha_1 > \alpha_2 > \ldots > \alpha_{J_1}$ resulting in $s_{ij} = 0$ for $i < j \leqslant J_1$. Similarly, a typical sequence $\{d_n\}$ for the case (iii) is $d_n = \beta(\log n)/n$ for some $\beta \in (0,1)$. So we take $d_{nj} = \beta_j(\log n)/n$ for $j > J_2$, with $\beta_i < \beta_j$ for $J_2 < i < j$ resulting again in $s_{ij} = 0$. Finally, let $0 \leqslant J_2 - J_1 \leqslant 2$, which means that Σ_2 is at most a 2×2 matrix, and take $c_{J_2} = (e^{c_{J_1+1}} - 1)/c_{J_1+1}$ in the case $J_2 - J_1 = 2$. Under these special choices Corollary 2.1 reduces to the following one.

COROLLARY 2.2. *We have*

$$\mathbf{K}_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, I).$$

The diagonal form of the matrix Σ can be obtained for other sequences of the distance levels.

2.2. Asymptotic distribution of joint cluster counts from the uniform distribution on a given interval. Let $V_1, V_2, \ldots, V_n$ be independent random variables, each uniformly distributed on the interval $[a,b]$ with $a, b \in \mathbb{R},\ a < b$, known. Let $K_n^{a,b}(d_n)$ denote the number of clusters of (V_i) defined on $[a,b]$ according to the definition of $K_n^{0,1}(d_n) = K_n(d_n)$ on $[0,1]$ corresponding to a distance level d_n.

We are again interested in the joint behaviour of $K_n^{a,b}$'s for a sequence of different distance levels coming from the three cases. We extend Theorem 2.1 to this case.

Set $J \geqslant 1$ and let $d_{n1} \leqslant d_{n2} \leqslant \ldots \leqslant d_{nJ}$ be distance levels. Let us replace the condition (iii) given in Subsection 2.1 by

(iii′) $nd_{nj} \to \infty$, $ne^{-nd_{nj}/(b-a)} \to \infty$,

and suppose each d_{nj} satisfies one of the conditions: (i), (ii) or (iii′).

Let $K_{nj}^{a,b}(d_{nj})$ be numbers of clusters corresponding to the distance levels $d_{nj},\ j = 1, \ldots, J$. Set $m_{nj}^* = ne^{-nd_{nj}/(b-a)}$,

$$(\sigma_{nj}^*)^2 = e^{-2nd_{nj}/(b-a)}\big(e^{nd_{nj}/(b-a)} - 1 - [nd_{nj}/(b-a)]^2\big), \tag{2.5}$$

and

$$\mathbf{K}_n^{a,b} = \frac{1}{\sqrt{n}}\left(\frac{K_{n1}^{a,b}(d_{n1}) - m_{n1}^*}{\sigma_{n1}^*}, \ldots, \frac{K_{nJ}^{a,b}(d_{nJ}) - m_{nJ}^*}{\sigma_{nJ}^*}\right). \tag{2.6}$$

Then we have

THEOREM 2.2. *Under the above notation*, *the assumptions* (i), (ii), (iii′), *and*

$$e^{(-nd_{ni}-nd_{nj})/(b-a)}\left(e^{nd_{ni}/(b-a)} - 1 - \frac{n^2 d_{ni} d_{nj}}{(b-a)^2}\right)/\sigma_{ni}^*\sigma_{nj}^* \to s_{ij} \tag{2.7}$$

for all $1 \leqslant i < j \leqslant J$, *we have*

$$\mathbf{K}_n^{a,b} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma), \tag{2.8}$$

where $\Sigma = (s_{ij})$ *with* $s_{jj} = 1$ *for all* $j = 1, \ldots, J$.

Theorem 2.2 is an immediate corollary of Theorem 2.1 after applying a linear transformation of V_i's onto the interval $[0, 1]$. Of course, the transformation refers both to the sample and to the distance levels, thus $K_n^{a,b}(d_n) = K_n^{0,1}\big(d_n/(b-a)\big)$.

Replacing the conditions (A2) and (A3) given in Subsection 2.1 by

(A2′) for $J_1 < j \leqslant J_2$ it follows that $nd_{nj}/(b-a) \to c_j$;

(A3′) for $J_2 < i < j$ it follows that $n(d_{nj} - d_{ni})/(b-a) \to -2 \log s_{ij}$,

we see that (2.7) is satisfied. Thus we get the following analogue of Corollary 2.1.

COROLLARY 2.3. *Under the conditions* (i), (iii′) *and* (A1), (A2′), (A3′), *the relation* (2.8) *holds*, *where*

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{pmatrix} \tag{2.9}$$

is the block a diagonal matrix with $s_{jj} = 1$, s_{ij} *as in* (A1′) *and* (A3′) *for the first and third group*, *while for* $J_1 < i < j \leqslant J_2$

$$s_{ij} = (e^{c_i} - 1 - c_i c_j)/\sqrt{(e^{c_i} - 1 - c_i^2)(e^{c_j} - 1 - c_j^2)}. \tag{2.10}$$

2.3. Asymptotic distribution of joint cluster counts from the uniform distribution on an unknown interval. Let $V_1, \ldots, V_n$ be independent, uniformly distributed random variables on the interval $[a, b]$ with $a < b$ being unknown and let $V_{1,n}, \ldots, V_{n,n}$ be the ordered sample. We shall investigate a counterpart of Theorems 2.1 and 2.2 when the endpoints of the interval are estimated by $\hat{a} = V_{1,n}$ and $\hat{b} = V_{n,n}$.

By analogy with the previous notation, for given $J \geqslant 1$ and distance levels $d_{n1} < \ldots < d_{nJ}$, set $\hat{m}_{nj} = ne^{-nd_{nj}/(\hat{b}-\hat{a})}$,

$$\hat{\sigma}_{nj}^2 = e^{-2nd_{nj}/(\hat{b}-\hat{a})}\big(e^{nd_{nj}/(\hat{b}-\hat{a})} - 1 - [nd_{nj}/(\hat{b}-\hat{a})]^2\big),$$

and

$$\hat{\mathbf{K}}_n = \frac{1}{\sqrt{n}}\left(\frac{\hat{K}_{n1}(d_{n1}) - \hat{m}_{n1}}{\hat{\sigma}_{n1}}, \ldots, \frac{\hat{K}_{nJ}(d_{nJ}) - \hat{m}_{nJ}}{\hat{\sigma}_{nJ}}\right). \tag{2.11}$$

THEOREM 2.3. *If the assumptions* (i), (ii), (iii′) *are satisfied and* (2.7) *holds*, *then*

$$\hat{\mathbf{K}}_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma) \tag{2.12}$$

with Σ *as in* (2.8).

Before proving Theorem 2.3 we state some useful lemmas. The first one follows immediately from the Slutsky lemma while the second one is a well-known property of estimators $\hat{a}$ and $\hat{b}$.

LEMMA 2.1. *Let $X_1, X_2, \ldots$ be a sequence of random vectors in $\mathbb{R}^J$, and $l_n \in \mathbb{R}^J$ and $s_n \in \mathbb{R}$ be deterministic norming sequences. If $(X_n - l_n)/s_n \xrightarrow{\mathcal{D}} X$ and for some random vectors L_n and random variables S_n we have $(L_n - l_n)/s_n \xrightarrow{P} 0$ and $S_n/s_n \xrightarrow{P} 1$, then $(X_n - L_n)/S_n \xrightarrow{\mathcal{D}} X$.*

LEMMA 2.2. *For every $\alpha < 1$ we have*

$$n^\alpha(b - V_{n,n}) \xrightarrow{P} 0 \quad \textit{and} \quad n^\alpha(V_{1,n} - a) \xrightarrow{P} 0.$$

Now, we are ready to prove Theorem 2.3.

Proof of Theorem 2.3. According to Theorem 2.2 and Lemma 2.1 it is enough to prove that

$$\frac{\hat{m}_{nj} - m^*_{nj}}{\sqrt{n}\sigma^*_{nj}} \xrightarrow{P} 0 \quad \text{and} \quad \frac{\hat{\sigma}^2_{nj}}{(\sigma^*_{nj})^2} \xrightarrow{P} 1. \tag{2.13}$$

Since $\hat{b} - \hat{a} < b - a$ a.s. and, by the Lagrange theorem, for every $x \leqslant y$ it follows that $|e^{-x} - e^{-y}| \leqslant e^{-x}|x - y|$, we obtain

$$\begin{aligned}|\hat{m}_{nj} - m^*_{nj}| &= n|e^{-nd_{nj}/(\hat{b}-\hat{a})} - e^{-nd_{nj}/(b-a)}| \\ &\leqslant n^2 d_{nj} \frac{|\hat{b} - b| + |\hat{a} - a|}{(\hat{b} - \hat{a})(b - a)} e^{-nd_{nj}/(b-a)}.\end{aligned}$$

Hence

$$\left|\frac{\hat{m}_{nj} - m^*_{nj}}{\sqrt{n}\sigma^*_{nj}}\right| \leqslant \frac{\sqrt{n}|\hat{b} - b| + \sqrt{n}|\hat{a} - a|}{(\hat{b} - \hat{a})} \varphi\left(\frac{nd_{nj}}{b - a}\right),$$

where $\varphi(x) = x/\sqrt{e^x - 1 - x^2}$. As $\varphi(x)$ is a bounded function on $(0, \infty)$, by an application of Lemma 2.2 with $\alpha = 1/2$ we obtain the first relation in (2.13).

To prove the second relation, set $\psi(x) = x/(1 - e^{-x} - x^2 e^{-x})$ and observe that $\psi(x)$ is a Lipschitz function on $(0, \infty)$ and $\psi(x) > 1$ on this interval. Now,

$$\begin{aligned}&\frac{\hat{\sigma}^2_{nj}}{(\sigma^*_{nj})^2} - 1 = \\ &= \frac{b - a}{\hat{b} - \hat{a}} \exp\left\{-nd_{nj}\left(\frac{1}{\hat{b} - \hat{a}} - \frac{1}{b - a}\right)\right\} \frac{\psi\big(nd_{nj}/(b - a)\big) - \psi\big(nd_{nj}/(\hat{b} - \hat{a})\big)}{\psi\big(nd_{nj}/(\hat{b} - \hat{a})\big)} \\ &\qquad + \frac{b - a}{\hat{b} - \hat{a}} \exp\left\{-nd_{nj}\left(\frac{1}{\hat{b} - \hat{a}} - \frac{1}{b - a}\right)\right\} - 1.\end{aligned}$$

Because in all considered cases $nd_{nj} < \log n$ for large n, by Lemma 2.2 it follows that $nd_{nj}\big((\hat{b}-\hat{a})^{-1} - (b-a)^{-1}\big)$ converges in probability to zero, which by the properties of the function $\psi(x)$ implies that the second relation in (2.13) also holds. This completes the proof. ∎

As in the previous subsection we get the following corollary.

COROLLARY 2.4. *Under the conditions* (i), (iii′) *and* (A1), (A2′), (A3′), *the relation* (2.12) *holds*, *where* Σ *is given by* (2.9).

In particular, *if* $J = 1$ *and* (i) *or* (A2′), *or* (iii′) *holds, then*

$$\frac{\hat{K}_{n1} - \hat{m}_{n1}}{\sqrt{n}\hat{\sigma}_{n1}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1). \tag{2.14}$$

Obviously, (2.14) corresponds to Theorem 1.1.

3. STATISTICAL RESULTS AND SIMULATIONS

3.1. Test statistics. First consider the simple null hypothesis asserting that a sample $X_1, \ldots, X_n$ has the uniform distribution on $[0,1]$. Given $J \geqslant 1$ and distance levels $d_{n1} \leqslant \ldots \leqslant d_{nJ}$, each satisfying one of the conditions (i), (ii) or (iii) and such that (2.3) holds. Consider the statistic

$$C = \mathbf{K}_n^T \Sigma^{-1} \mathbf{K}_n,$$

where $\mathbf{K}_n$ is given by (2.2) and Σ is as in Theorem 2.1. Then from (2.4) it follows that under the null hypothesis

$$C \xrightarrow{\mathcal{D}} \chi^2_J, \tag{3.1}$$

where χ^2_k is a random variable with chi-square distribution with k degrees of freedom. So, C defines the upper-tailed test for uniformity called here the *cluster test* and denoted by C. This means that asymptotic critical values of this test are given by quantiles of the chi-square distribution with J degrees of freedom. Since the convergence in (3.1) is very slow, we propose rather to use empirical critical values (see Subsection 3.2).

Now, consider the composite null hypothesis asserting that a sample comes from the family of all uniform distributions on $\mathbb{R}$. As a test statistic one can use $\hat{\mathbf{K}}_n^T \Sigma^{-1} \hat{\mathbf{K}}_n$, where $\hat{\mathbf{K}}_n$ is given by (2.11). Here, we propose another solution based on the random transform of the data into the unit interval. To this end, let $Y_1, Y_2, \ldots$ be i.i.d. exponentially distributed random variables with mean one (cf. the proof of Theorem 2.1) and set $S_k = Y_1 + \ldots + Y_k$. It is well known that

$$\left(\frac{S_1}{S_{n+1}}, \ldots, \frac{S_n}{S_{n+1}}\right) \stackrel{\mathcal{D}}{=} (U_{1,n}, \ldots, U_{n,n}), \tag{3.2}$$

where $U_{1,n}, \ldots, U_{n,n}$ are the order statistics of the uniform $[0,1]$ sample $U_1, \ldots, U_n$. The following well-known lemma is a consequence of (3.2).

LEMMA 3.1. *Let $V_1, \ldots, V_n$ be i.i.d. uniformly distributed random variables over the interval $[a,b]$ and let $V_{1,n}, \ldots, V_{n,n}$ be the corresponding order statistics. Then for each fixed n*

$$\left(\frac{V_{2,n}-V_{1,n}}{V_{n,n}-V_{1,n}}, \ldots, \frac{V_{n-1,n}-V_{1,n}}{V_{n,n}-V_{1,n}}\right) \stackrel{\mathcal{D}}{=} (U_{1,n-2}, \ldots, U_{n-2,n-2}), \tag{3.3}$$

where on the right-hand side of (3.3) *we have the order statistics of the uniform $[0,1]$ sample $U_1, \ldots, U_{n-2}$.*

Given $J \geqslant 1$ and distance levels $d_{n1} \leqslant \ldots \leqslant d_{nJ}$, each satisfying one of the conditions (i), (ii) or (iii) and such that (2.3) holds. Let

$$\tilde{\mathbf{K}}_{n-2} = \frac{1}{\sqrt{n}}\left(\frac{\tilde{K}_{n-2,1}(d_{n1}) - m_{n-2,1}}{\sigma_{n-2,1}}, \ldots, \frac{\tilde{K}_{n-2,J}(d_{nJ}) - m_{n-2,J}}{\sigma_{n-2,J}}\right)$$

(cf. (2.2)) be a vector of normalized numbers of clusters of the randomly transformed sample $(V_{2,n}-V_{1,n})/(V_{n,n}-V_{1,n}), \ldots, (V_{n-1,n}-V_{1,n})/(V_{n,n}-V_{1,n})$. Now, for testing the composite null hypothesis we take the statistic

$$C_{mod} = \tilde{\mathbf{K}}_{n-2}^T \Sigma^{-1} \tilde{\mathbf{K}}_{n-2},$$

where again Σ is as in Theorem 2.1. Then from (3.3) and Theorem 2.1 it follows that under the null hypothesis

$$C_{mod} \xrightarrow{\mathcal{D}} \chi^2_J. \tag{3.4}$$

Thus, C_{mod} defines the upper-tailed test, called here the *modified cluster test* and denoted by C_{mod}. Since the convergence in (3.4) is slow, as for C we also propose to use empirical critical values.

3.2. Critical values. The exact null distributions of the test statistics C and C_{mod} are very difficult to evaluate. Moreover, Csörgő and Wu [1] showed that for typical distance levels (see the end of Subsection 2.1) the fastest rate of convergence of $K_{nj}(d_{nj})$ to the normal distribution is $O(n^{-1/4}\log n)$. So, the convergence for each j and probably also jointly is very slow. This can be easily seen in Table 1. Thus we provide simulated critical values both for C and C_{mod}. After preliminary simulations reported in Subsection 3.3 below, we decided to take $J = 6$, $J_1 = J_2 - J_1 = 2$, and distance levels described at the end of Subsection 2.1 with $\alpha_1 = 1.9, \alpha_2 = 1.1, c_1 = 1, c_2 = e-1, \beta_1 = 0.1, \beta_2 = 0.9$. Such a choice implies that Corollary 2.2 holds, i.e. $\Sigma = I$. In Table 1 we present simulated critical values corresponding to significance levels $0.1, 0.05$, and 0.01. The last row, corresponding to $n = \infty$, contains asymptotic critical values (the same for both tests).

TABLE 1. Critical points of cluster tests C and C_{mod}
for different sample sizes and different significance levels, 200,000 MC runs

C				C_{mod}			
n	0.10	0.05	0.01	n	0.10	0.05	0.01
20	13.05	15.69	21.86	20	16.29	19.32	25.90
50	12.36	15.26	22.32	50	13.33	16.07	22.65
100	12.71	15.74	23.34	100	13.03	15.96	23.22
200	12.77	15.98	23.68	200	12.91	16.00	23.63
500	12.59	15.68	23.24	500	12.65	15.71	23.19
1000	12.34	15.38	22.64	1000	12.29	15.34	22.34
				∞	10.65	12.59	16.81

3.3. Optimal choice of J and distance levels. To see how the dimension J and a choice of distance levels influence powers attained by cluster tests we compared several cases with J between 2 and 6, $J_2 - J_1 \leqslant 2$, and typical distance levels such that Corollary 2.2 applies. The results are shown in Table 2. In all cases we used empirical critical values which are given in the third column of Table 2. We present empirical powers for two specific alternatives g_1 and g_2 described in Subsection 3.4. It is easy to see that C attains the highest power for $J = 6$ and other parameters as specified in Subsection 3.2. The behaviour of C_{mod} is quite similar.

TABLE 2. Empirical critical values ($u_{0.05}$) and empirical powers (in %) for the cluster test C under alternatives g_1 and g_2 for different dimensions J and different typical distance levels. Significance level 0.05, $n = 100$, 200,000 MC runs

J	α	c	β	$u_{0.05}$	g_1 with $\varrho = 3/2$	g_2 with $\varrho = 0.9$, $j = 5$
2	1.5	–	0.5	6.52	6	14
2	–	1	0.5	6.75	9	56
2	1.5	1	–	6.20	13	70
2	1.3	1	–	6.06	14	77
2	1.1	1	–	6.41	16	84
3	1.5	1	0.5	8.38	10	61
3	1.1	1	0.9	10.68	14	85
4	1.1	1	0.9			
	–	1.7	–	11.96	14	88
4	1.1	0.5	0.9			
	–	1.3	–	12.34	16	88
4	1.9	1	0.1			
	–	–	0.9	12.01	13	85
6	1.9	0.5	0.1			
	1.1	1.3	0.9	16.65	16	87
6	1.9	1	0.1			
	1.1	1.7	0.9	15.74	16	89

3.4. Power comparision. We simulated powers of the new tests C and C_{mod} with parameters specified in Subsection 3.2 for significance level 0.05, sample size $n = 100$, and for $200{,}000$ MC runs for each case. We considered five continuous alternative distributions on $[0, 1]$. All alternative distributions are identified by their density functions (g_1, g_2, g_3, g_4) or by their quantile function (G_5^{-1}). The list of alternatives is as follows:

1. $g_1(t) = \begin{cases} 2^{\varrho-1}\varrho t^{\varrho} & \text{if } 0 \leqslant t < 1/2, \\ 2^{\varrho-1}\varrho(1-t)^{\varrho} & \text{if } 1/2 \leqslant t \leqslant 1, \end{cases}$ where $\varrho > 0$;

2. $g_2(t) = 1 + \varrho\cos(\pi j t)$, where $\varrho \in [-1, 1]$;

3. $g_3(t) = c(\theta^{(j)}) \exp\left\{\sum_{k=1}^{j} \theta_k b_k(t)\right\}$, where b_k are the Legendre polynomials on $[0, 1]$ and $\theta^{(j)} = (\theta_1, \ldots, \theta_j)$;

4. contamination of the uniform distribution with beta distribution: $g_4(t) = 1 - \varrho + \varrho\Gamma(p+q)/\big(\Gamma(p) + \Gamma(q)\big)t^{p-1}(1-t)^{q-1}$, where $\varrho \in [0, 1]$;

5. $G_5^{-1}(t) = 1/2 + \big(t - (1-t)^{\varrho}\big)/2$, where $\varrho > 0$.

We compare the new tests C and C_{mod} with the data driven smooth test N_{T1} introduced in Inglot and Ledwina [5] which proved to be a strong test for uniformity. We took powers of N_{T1} from Tables 2–4 in [5]. Overall, the data driven smooth test N_{T1} appears to be the best and the cluster test and the modified cluster test give uniformly poorer performance, except for highly oscillating alternatives where the cluster tests are almost equally well or perform better than N_{T1}, for example for g_2 with $\varrho = 1.00$ and $j = 10$ the power of both tests is 100% and 99% (see Table 3).

For better illustration of the cluster tests performance we draw power functions of the three compared tests for alternative g_1 (Figure 1) and for alternative 5 (Figure 2). The procedure ran for 300 values of the parameters chosen from the interval $[0, 3]$, significance level 0.05, and $n = 100$. For alternative 5 the case $\varrho = 0$ corresponds to the uniform distribution on $[0, 1/2]$, while the case $\varrho = 1$ corresponds to the uniform distribution in $[0, 1]$, thus according to the modified cluster test the sample is contained in the class of all the uniform laws, see Figure 2.

TABLE 3. Empirical powers (in %) for N_{T1}, C, and C_{mod}.
Alternatives g_2, g_3 and g_4, significance level 0.05, $n = 100$, $200{,}000$ MC runs

Alternative	ϱ	j	p	q	θ	N_{T1}	C	C_{mod}
g_2	0.45	1				78	15	12
g_2	0.60	4				71	34	29
g_2	0.75	7				81	62	54
g_2	1.00	10				75	100	99
g_3		2			$(-0.2, -0.3)$	73	12	9
g_3		5			$(0, 0, 0, 0, 0.4)$	76	22	18
g_3		8			$(0, 0, 0, 0, 0, 0, 0, -0.5)$	90	42	36
g_4	0.25		2.0	10.0		73	16	15
g_4	0.50		0.8	1.5		61	10	9
g_4	0.10		0.1	0.1		68	36	26

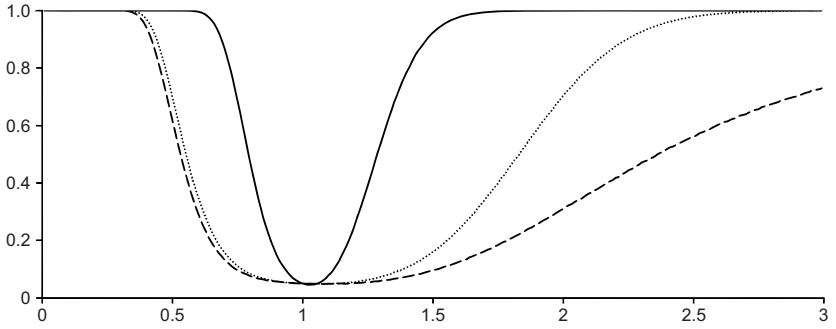

FIGURE 1. Empirical powers of N_{T1} (thick line), C (dotted line), and C_{mod} (dashed line) as functions of ϱ for alternative g_1. Significance level 0.05, $n = 100$, $200{,}000$ MC runs

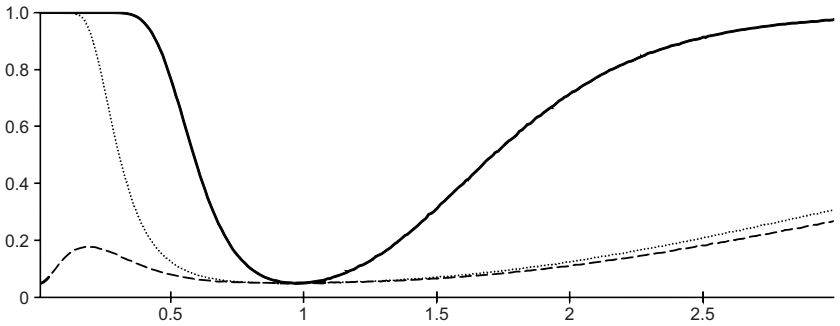

FIGURE 2. Empirical powers of N_{T1} (thick line), C (dotted line), and C_{mod} (dashed line) as functions of ϱ for alternative 5. Significance level 0.05, $n = 100$, $200{,}000$ MC runs

The consistency of both cluster tests is a difficult question because one should prove a kind of the Csörgő and Wu theorem for a nonuniformly distributed sample. From simulations not reported here in detail it follows that the new tests seem to be consistent since the increasing of a sample size results in a greater power. For example, for the test C and alternative g_2 with $\varrho = 0.80$ and $j = 8$ the powers for $n = 20, 50$, and 100 are $26\%, 48\%$, and 75%, respectively, while for the same alternative with $\varrho = 1.00$ and $j = 12$ the powers for $n = 20, 50$, and 100 are $31\%, 85\%$, and 100%, respectively.

The conclusion is that the cluster tests perform worse than other procedures unless some highly oscillating alternatives. Samples coming from these alternatives are inherently well-clusterable, such as those associated with periodic density functions.

Acknowledgments. I am grateful to Professor Sándor Csörgő for posing the problem, and Gyula Pap for his help during the work. I also thank the referee for many insightful comments.

REFERENCES

[1] S. Csörgő and W. B. Wu, *On the clustering of independent uniform random variables*, Random Structures Algorithms 25 (4) (2004), pp. 396–420.
[2] R. B. D'Agostino and M. A. Stephens, *Goodness-of-fit Techniques*, Marcel Dekker, New York 1986.
[3] E. Godehardt and J. Jaworski, *On the connectivity of a random interval graph*, Random Structures Algorithms 9 (1996), pp. 137–161.
[4] T. Inglot and A. Janic-Wróblewska, *Data driven chi-square test for uniformity with unequal cells*, J. Stat. Comput. Simul. 73 (8) (2003), pp. 545–561.
[5] T. Inglot and T. Ledwina, *Towards data driven selection of a penalty function for data-driven Neyman's test*, Linear Algebra Appl. 417 (2003), pp. 124–133.
[6] W. C. M. Kallenberg and T. Ledwina, *Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests*, Ann. Statist. 23 (5) (1995), pp. 1594–1608.
[7] T. Ledwina, *Data-driven version of Neyman's smooth test of fit*, J. Amer. Statist. Assoc. 89 (427) (1994), pp. 1000–1005.
[8] M. Schader and F. Schmid, *Power of tests for uniformity when limits are unknown*, J. Appl. Stat. 24 (2) (1997), pp. 193–205.
[9] M. A. Stephens, *EDF statistics for goodness of fit and some comparisons*, J. Amer. Statist. Assoc. 69 (1974), pp. 730–737.

GAMF
College of Kecskemét
Izsáki út 10. Kecskemét, Hungary-6000
E-mail: osztenyine.eva@gamf.kefo.hu

Received on 12.1.2011;
revised version on 20.2.2013