Strong Limit Theorems for Sums of Logarithms of $m$th Order Spacings

Magnus Ekström

Department of Mathematical Statistics
Umeå University
S-901 87 Umeå, Sweden

Abstract

Several strong limit theorems are proved for sums of logarithms of $m$th order spacings from general distributions. In all given results, the order of the spacings is allowed to increase to infinity with the sample size. These results provide a nonparametric strongly consistent estimator of entropy as well as a characterization of the uniform distribution on $[0,1]$. Furthermore, it is shown that Cressie’s (1976) goodness of fit test is strongly consistent against all continuous alternatives.  

Key words and phrases: Spacings, Strong limit theorems, Entropy estimation, Uniform distribution, Goodness of fit

1991 AMS subject classification: 60F15, 62G05, 94A17

1Research was supported by The Bank of Sweden Tercentenary Foundation.
Let $X_1, X_2, ..., X_n$ be an i.i.d. sample of random variables on $[0,1]$ with distribution $F(x)$. Denote the order statistics by $X(1) \leq X(2) \leq ... \leq X(n)$ and define $X(0) = 0$ and $X(n+1) = 1$. For an integer $m \geq 1$ the $m$th order spacings, sometimes called $m$th order gaps, are defined by

$$D_{i,n}^{(m)} = X_{(j+m)} - X_{(j)}, \quad j = 0, 1, ..., n - m + 1.$$ 

When $m = 1$, the $m$th order spacings reduce to simple spacings (or one-step spacings). A voluminous literature exists on simple spacings, see e.g. the reviews by Pyke (1965, 1972), or D'Agostino and Stephens (1986).

Consider the statistic

$$L_n^{(m)}(X) = \frac{1}{n - m + 2} \sum_{j=0}^{n-m+1} \log \left( \frac{n+1}{m} D_{i,n}^{(m)} \right).$$

The statistic $L_n^{(m)}$ has been popular for testing uniformity and was first investigated by Darling (1953). Attention has mostly been focused on the case $m = 1$, see the reviews mentioned above. For literature on testing uniformity using statistics like $L_n^{(m)}$ based on high order spacings, see Cressie (1976, 1978, 1979), Dudewicz and Van der Meulen (1981), Kuo and Rao (1981) and Hall (1986). Furthermore, a general method of estimating parameters in continuous univariate distributions, called the maximum spacing method, is based on $L_n^{(1)}$. This method was introduced by Cheng and Amin (1983) and independently by Ranneby (1984). That functions other than the log-function can be used to obtain consistent estimators is shown in Ranneby and Ekström (1997).

The asymptotic properties of statistics like $L_n^{(m)}$ have been investigated by many authors, mainly for the particular case where the underlying distribution is uniform on $[0,1]$, or for a sequence of restricted alternatives approaching the uniform distribution as $n$ increases.

For general distributions $F$, convergence in probability of $L_n^{(m)}$ is shown in Hall (1984) ($m$ fixed) and in Khashimov (1989) ($m \to \infty$ as $n \to \infty$). General almost sure convergence of $L_n^{(m)}$ is given in Van Es (1992). For $m$ even, $L_n^{(m)}$ is related to Vasicek's (1976) entropy estimator, for which Beirlant and Van Zuijlen (1985) give a strong limit theorem.

In the present paper several one- and two-sided strong limit theorems for $L_n^{(m)}$ will be established, where $m$ is allowed to increase with the sample size. In section 2, the main result is a strong limit theorem for $L_n^{(m)}$, $m = o(n)$, based on uniform spacings. Further, a characterization of the uniform distribution on $[0,1]$ is given. In section 3 several one-sided and two-sided strong limit results for statistics related to $L_n^{(m)}$, $m = o(n/\log n)$, will be shown, including some results for the statistic (1) multiplied on the right by the indicator function $I_B(X_{(j)})$, where $B \subseteq [0,1]$ is a measurable set. For the results for general continuous
distributions related to Van Es (1992), we will use less restrictive assumptions on the densities, e.g. we do not assume that the densities are bounded away from zero on their support. In Van Es (1992) the approach of Vasicek (1976), using Stieltjes sums, was used. Here we adopt a different approach, related to that of Shao and Hahn (1995) who gave results for $L_n^{(1)}$.

We add one remark to guard against misapprehension. To say $m = o(s(n))$ or $m = O(s(n))$, where $s(n) \to \infty$ as $n \to \infty$, does not exclude that $m = m(n)$ can be a bounded sequence.

2 Strong limit theorems for $L_n^{(m)}$ based on uniform spacings

Throughout this section, $0 \equiv U(0) \leq U(1) \leq \ldots \leq U(n) \leq U(n+1) \equiv 1$ denotes the order statistics from an i.i.d. n-sample $U_1, U_2, \ldots, U_n$ of uniformly distributed random variables on $[0,1]$, and

$$L_n^{(m)}(U) = \frac{1}{n-m+2} \sum_{j=0}^{n-m+1} \log \left( \frac{n+1}{m} \left( U_{(j+m)} - U_{(j)} \right) \right).$$

Theorem 1 Let \( \{m_n\} \) be a sequence of positive integers such that $m_n = o(n)$. Then,

$$\lim_{n \to \infty} (L_n^{(m_n)}(U) - \psi(m_n) + \log m_n) = 0 \quad (a.s.),$$

where $\psi(x) = \frac{d}{dx} \log \Gamma(x)$ is the digamma function and $\Gamma$ the gamma function.

Remark. If $m_n$ is fixed or if $m_n \log n \to \infty$, Theorem 1 is a special case of Theorem 2 in Van Es (1992). For comments on Van Es' (1992) conditions, see the remark after Corollary 4.

For the proof of Theorem 1, we will use a well known relationship between uniform spacings and standard exponential random variables (Pyke (1965)), and a useful inequality for standard exponential random variables obtained by Van Es (1992), i.e. the two lemmas below.

Lemma 1 Let $Z_1, Z_2, \ldots, Z_{n+1}$ be independent standard exponential random variables. Set $W_j = Z_j / (Z_1 + \ldots + Z_{n+1})$, $j = 1, \ldots, n+1$. Then $(W_1, W_2, \ldots, W_{n+1})$ is distributed as the set of $n+1$ simple spacings determined by $n$ independent uniform random variables on $[0,1]$.

Proof. See Pyke (1965).
Lemma 2 Let $Z_1, ..., Z_{n+1}$ be independent standard exponential random variables. Then, if $m_n / \log n \to \infty$ as $n \to \infty$, for any $\varepsilon > 0$ and for $n$ large enough,

$$P \left( \sup_{0 \leq j \leq n-m_n+1} \left| \log \left( \frac{1}{m_n} \sum_{i=j+1}^{j+m_n} Z_i \right) \right| \geq \varepsilon \right) \leq \frac{2}{n^2}.$$ 

Proof. For any $\delta > 0$ and for $m_n$ such that $m_n / \log n \to \infty$, with $n$ large enough, Van Es (1992) obtained the inequality,

$$P \left( \sup_{0 \leq j \leq n-m_n+1} \left| \frac{1}{m_n} \sum_{i=j+1}^{j+m_n} Z_i - 1 \right| \geq \delta \right) \leq \frac{2}{n^2},$$

from which the lemma follows. \qed

Proof of Theorem 1. By the Borel-Cantelli lemma it suffices to show, for every positive $\delta$,

$$\sum_n P \left( \left| L_n^{(m_n)}(U) - \psi(m_n) + \log m_n \right| > \delta \right) < \infty. \quad (2)$$

For a sequence $Z_1, Z_2, ..., Z_{n+1}$ of i.i.d. standard exponentially distributed random variables, define

$$\alpha_{j,n} = \log \left( \frac{1}{m_n} \sum_{i=j+1}^{j+m_n} Z_i \right), \quad \beta_n = \log \left( \frac{1}{n+1} \sum_{i=1}^{n+1} Z_i \right).$$

Then, by Lemma 1,

$$P \left( \left| L_n^{(m_n)}(U) - \psi(m_n) + \log m_n \right| > \delta \right)$$

$$= P \left( \left| \frac{1}{n-m_n+2} \sum_{j=0}^{n-m_n+1} \alpha_{j,n} - \beta_n - \psi(m_n) + \log m_n \right| > \delta \right)$$

$$\leq P \left( \left| \frac{1}{n-m_n+2} \sum_{j=0}^{n-m_n+1} \alpha_{j,n} - \psi(m_n) + \log m_n \right| > \frac{\delta}{2} \right) + P \left( |\beta_n| > \frac{\delta}{2} \right).$$

An application of Lemma 2 with $m_n = n + 1$ yields $\sum_n (|\beta_n| > \delta/2) < \infty$. To check that $\sum_n P(\left| (n-m_n+2)^{-1} \sum_{j=0}^{n-m_n+1} \alpha_{j,n} - \psi(m_n) + \log m_n \right| > \delta/2) < \infty$, two separate cases will be considered.

First the case $m_n < n^{1/4}$. Here the high order moments of the $\alpha_{j,n}$'s will be used. As $Z_{j+1} + Z_{j+2} + ... + Z_{j+m_n}$ is $\Gamma(m_n, 1)$-distributed it follows, for all positive integers $m_n$ and $r, \beta_n$, that

$$E \left[ \alpha_{j,n}^r \right] = \frac{1}{\Gamma(m_n)} \int_0^{\infty} \left( \log \frac{x}{m_n} \right)^r x^{m_n-1} e^{-x} dx,$$
and especially that

\[ E[\alpha_{i,n}] = \psi(m_n) - \log m_n. \]

Clearly, there exist constants \( C_r \) such that, for \( m_n = 1 \),

\[ E[|\alpha_{i,n}|^r] \leq \int_0^\infty |\log x|^r e^{-x} \, dx \leq C_r \int_0^\infty x e^{-x} \, dx + \int_0^1 |\log x|^r \, dx \leq C_r + r!, \]

and, for \( m_n \geq 2 \),

\[
E[|\alpha_{i,n}|^r] \leq \frac{1}{\Gamma(m_n)} \int_0^\infty \left|\log \frac{x}{m_n}\right|^r x^{m_n-1} e^{-x} \, dx \\
\leq \frac{C_r}{\Gamma(m_n)} \int_0^\infty \left( \frac{x + m_n}{x} \right) x^{m_n-1} e^{-x} \, dx \\
= \frac{C_r}{\Gamma(m_n)} \left( \frac{\Gamma(m_n + 1)}{m_n} + m_n \Gamma(m_n - 1) \right) \leq 3C_r.
\]

Thus, \( \alpha_{i,n} \) has finite moments of all orders with an upper bound independent of \( n \). Moreover, for all positive integers \( r_1, r_2, \ldots, r_s \) and \( r = r_1 + \ldots + r_s \),

\[ E[\alpha_{j_1,n}^{r_1} \alpha_{j_2,n}^{r_2} \cdots \alpha_{j_s,n}^{r_s}] \leq E[\alpha_{j_1,n}^{r_1} \alpha_{j_2,n}^{r_2} \cdots \alpha_{j_s,n}^{r_s}] \leq r E[|\alpha_{0,n}|^r] \leq r(3C_r + r!). \tag{3} \]

Note that \( \{\alpha_{i,n}\}_{i=0}^{m_n} \) is a \((m_n - 1)\)-dependent sequence, that is \( \alpha_{i,n} \) and \( \alpha_{j,n} \) are independent for all \( i \) and \( j \) such that \(|i - j| > m_n - 1\). This together with the inequality (3) implies that there exist positive constants \( K \) and \( N \) such that, for all \( n > N \),

\[
E \left[ \left( \frac{1}{n-m_n+2} \sum_{j=0}^{n-m_n+1} \alpha_{i,n} - \psi(m_n) + \log m_n \right)^4 \right] \\
= O \left( n^{-2} \right) + \frac{1}{(n-m_n+2)^4} \left( 24 \sum_{i_1 < \ldots < i_4} E \left[ \prod_{j=1}^4 (\alpha_{i_j,n} - \psi(m_n) + \log m_n) \right] \right) \\
+ 12 \sum_{\substack{i_1 \neq i_2, i_1 \neq i_3 \neq i_3 \neq i_2 < i_3}} E \left[ (\alpha_{i_1,n} - \psi(m_n) + \log m_n) \prod_{j=1}^3 (\alpha_{i_j,n} - \psi(m_n) + \log m_n) \right] \\
\leq K \left( \frac{m_n}{n} \right)^2 + O \left( n^{-2} \right).
\]

Thus, by Chebyshev’s inequality, for some constant \( K_1 \), \( m_n < n^{1/4} \) and for \( n \) large enough, \( P(|(n-m_n+2)^{-1} \sum_{j=0}^{n-m_n+1} \alpha_{i,n} - \psi(m_n) + \log m_n| > \delta/2) \leq K_1/n^{3/2} \).

Next, consider the second case, \( m_n \geq n^{1/4} \). Since \( \psi(m) - \log m \to 0 \) as \( m \to \infty \), an application of Lemma 2 yields, if \( m_n \geq n^{1/4} \) and if \( n \) is large enough, \( P(|(n-m_n+2)^{-1} \sum_{j=1}^{n-m_n+1} \alpha_{j,n} - \psi(m_n) + \log m_n| > \delta/2) \leq 2/n^2 \). Thus, the inequality (2) holds and the theorem is proved. \( \square \)
Theorem 2 Let $X_1, X_2, ..., X_n$ be i.i.d. random variables on $[0,1]$ with distribution function $F(\cdot)$, and let $\{m_n\}$ be a sequence of positive integers such that $m_n = o(n/\log n)$. Then $F$ is the uniform distribution if and only if

$$\lim_{n \to \infty} \left( L_n^{(m_n)}(X) - \psi(m_n) + \log m_n \right) = 0 \quad \text{(a.s.)}.$$ 

Proof. See next section.

3 Strong limit theorems for statistics related to $L_n^{(m)}$ based on non-uniform spacings

The context of this section will generalize several results obtained for $L_n^{(1)}$ by Shao and Hahn (1995) to high order spacings.

Theorem 3 Let $U_1, U_2, ..., U_n$ be an i.i.d. sequence of uniformly distributed random variables on $[0,1]$ and let $\{m_n\}$ be a sequence of positive integers such that $m_n = o(n/\log n)$. Then, for any measurable set $B \subseteq [0,1]$ and for any nondecreasing bounded function $G(\cdot)$ on $[0,1]$,

$$\lim_{n \to \infty} \frac{1}{n-m_n+2} \sum_{j \in J_B} \log \left( \frac{G(U(j+m_n)) - G(U(j))}{U(j+m_n) - U(j)} \right) \leq \int_B \log g(x) dx \quad \text{(a.s.),}$$

where $g(x) = G'(x)$ a.e., and $J_B = \{j : U(j) \in B, 0 \leq j \leq n-m_n+1\}$. If the index set is redefined as $J_B = \{j : [U(j), U(j+m_n)] \subseteq B, 0 \leq j \leq n-m_n+1\}$, the inequality above holds if $B$ is a finite union of intervals.

To show that the majority of the quotients $(G(U(j+m_n)) - G(U(j)))/(U(j+m_n) - U(j))$ are arbitrarily close to $g(U(j))$ when $n \to \infty$, the following two lemmas will be used.

Lemma 3 Let $U_1, U_2, ..., U_n$ be an i.i.d. sequence of uniformly distributed random variables on $[0,1]$. Then there is a constant $C$ such that

$$\lim_{n \to \infty} \frac{n}{\log \log n} \max_{1 \leq j \leq n+1} \left( U(j) - U(j-1) \right) \leq \log n \quad \text{(a.s.)}.$$


Lemma 4 Let $G(\cdot)$ be any nondecreasing bounded function on $[0,1]$ and let $\mathcal{M}_G$ be the set of all points where the derivative $g$ of $G$ exists. Further, let $\{a_n\}$ be a
sequence of positive real numbers such that $\lim_{n \to \infty} a_n = 0$. Then, given $\varepsilon, \delta > 0$, there exists an integer $N$ such that the set

$$A_N = \left\{ x \in [0, 1] \setminus \mathcal{M}_G : \sup_{h \in b(0,a_N)} \left| \frac{G(x + h) - G(x)}{h} - g(x) \right| \leq \varepsilon \right\},$$

where $b(0, a)$ denotes a ball with center 0 and radius $a$, has Lebesgue measure greater than $1 - \delta$.

Proof. The proof follows by Egorov’s theorem and is essentially the same as the proof of Lemma 3.1. in Shao and Hahn (1995). \qed

Proof of Theorem 3. We follow the same approach as in Shao and Hahn (1995), where this one-sided strong limit result is proved for the special case $m_n = 1$ and $B = [0,1]$.

Only the case when $J_B = \{ j : U(j) \in B, 0 \leq j \leq n-m_n+1 \}$ will be proved here. The derivation of the second part, where $J_B$ is redefined, is similar because if $B$ is a finite union of intervals then for some positive constant $l$

$$\text{card}\{ j : [U(j), U(j+m_n)] \subseteq B, 0 \leq j \leq n-m_n+1 \} \leq \text{card}\{ j : U(j) \in B, 0 \leq j \leq n-m_n+1 \} \leq \text{card}\{ j : [U(j), U(j+m_n)] \subseteq B, 0 \leq j \leq n-m_n+1 \} + lm_n.$$

Let $k_n = (m_n/n) \log n$. Then, by Lemma 3, there exists a constant $C$ and a sufficiently large (random) $N_1$ such that for all $n \geq N_1$,

$$\max_{0 \leq j \leq n-m_n+1} \left( U(j+m_n) - U(j) \right) < Ck_n \quad \text{(a.s.)},$$

and by Lemma 4, for any positive numbers $\varepsilon, \delta < 1$, there exists an integer $N$ such that

$$A_N = \left\{ x \in [0, 1] \setminus \mathcal{M}_G : \sup_{h \in b(0,Ck_n)} \left| \frac{G(x + h) - G(x)}{h} - g(x) \right| \leq \varepsilon \right\}$$

has Lebesgue measure greater than $1 - \delta$. Let $B_N = B \cap A_N$, $B_N^c = B \setminus B_N$, $J_N = \{ j : U(j) \in B_N, 0 \leq j \leq n-m_n+1 \}$, $J_N^c = \{0, 1, \ldots, n-m_n+1\} \setminus J_N$ and let $j_N^c = \text{card} J_N^c$. Then for $n$ large enough,

$$\sum_{j \in J_B} \log \left( \frac{G(U(j+m_n)) - G(U(j))}{U(j+m_n) - U(j)} \right)$$

$$= \sum_{j \in J_N} \log \left( \frac{G(U(j+m_n)) - G(U(j))}{U(j+m_n) - U(j)} \right) + \sum_{j \in J_N^c} \log \left( \frac{G(U(j+m_n)) - G(U(j))}{U(j+m_n) - U(j)} \right)$$

6
\[ \leq \sum_{j \in J_N} \log \left( g(U_{(j)}) + \varepsilon \right) + \sum_{j \in J_N^c} \log \left( \frac{m_n G(1) - G(0)}{j_N(U_{(j+m_n)} - U_{(j)})} \right) \]

\[ \leq \sum_{j=0}^{n} \log \left( g(U_{(j)}) + \varepsilon \right) I_{B_N}(U_{(j)}) - m_n \log \varepsilon + j_N \log \left( \frac{G(1) - G(0)}{j_N/(n+1)} \right) \]

\[ - \sum_{j=0}^{n-m_n+1} \log \left( \frac{n+1}{m_n} \left( U_{(j+m_n)} - U_{(j)} \right) \right) I_{B_N}(U_{(j)}). \] (4)

By the strong law of large numbers,

\[ \lim_{n \to \infty} \frac{1}{n-m_n+2} \sum_{j=0}^{n} \log \left( g(U_{(j)}) + \varepsilon \right) I_{B_N}(U_{(j)}) = \int_{B_N} \log \left( g(x) + \varepsilon \right) dx \quad (a.s.) \]

(5)

and since \( j_N - 1 \leq \text{card} \{ j : U_j \in B_N^c, 1 \leq j \leq n \} \leq j_N + m_n \),

\[ \lim_{n \to \infty} \frac{j_N}{n-m_n+2} \log \left( \frac{G(1) - G(0)}{j_N/(n+1)} \right) = \mu(B_N^c) \log \left( \frac{G(1) - G(0)}{\mu(B_N^c)} \right) \quad (a.s.), \]

(6)

where \( \mu(\cdot) \) is the Lebesgue measure. An application of Cauchy-Schwarz inequality gives

\[ \frac{1}{n-m_n+2} \sum_{j=0}^{n-m_n+1} \left| \log \left( \frac{n+1}{m_n} \left( U_{(j+m_n)} - U_{(j)} \right) \right) \right| I_{B_N^c}(U_{(j)}) \]

\[ \leq \left( \frac{1}{n-m_n+2} \sum_{j=0}^{n} I_{B_N^c}(U_{(j)}) \right)^{1/2} \left( \frac{1}{n-m_n+2} \sum_{j=0}^{n-m_n+1} \log^2 \left( \frac{n+1}{m_n} \left( U_{(j+m_n)} - U_{(j)} \right) \right) \right)^{1/2}. \]

To obtain an upper bound for the second factor on the right hand side above, the notation of the proof of Theorem 1 will be used. Let \( \varepsilon_1 > 0 \) be arbitrary. Then

\[ \sum_n P \left( \frac{1}{n-m_n+2} \sum_{j=0}^{n-m_n+1} \log^2 \left( \frac{n+1}{m_n} \left( U_{(j+m_n)} - U_{(j)} \right) \right) - 2E[\alpha_{0,n}^2] > \varepsilon_1 \right) \]

\[ = \sum_n P \left( \frac{1}{n-m_n+2} \sum_{j=0}^{n-m_n+1} (\alpha_{j,n} - \beta_n)^2 - 2E[\alpha_{0,n}^2] > \varepsilon_1 \right) \]

\[ \leq \sum_n P \left( \frac{1}{n-m_n+2} \sum_{j=0}^{n-m_n+1} \alpha_{j,n}^2 - E[\alpha_{0,n}^2] + \beta_n^2 > \varepsilon_1 \right) \]

\[ \leq \sum_n P \left( \frac{1}{n-m_n+2} \sum_{j=0}^{n-m_n+1} \alpha_{j,n}^2 - E[\alpha_{0,n}^2] > \varepsilon_1 \frac{1}{4} \right) + \sum_n P \left( |\beta_n| > \frac{\sqrt{\varepsilon_1}}{2} \right). \] (7)

That the second sum on the right hand side is finite is shown in the proof of Theorem 1. Since \( \{\alpha_{j,n}\} \) is an \( (m_n-1) \)-dependent sequence it follows, using the approach used in the proof of Theorem 1, that for some constant \( K \) and if \( m_n < n^{1/4} \),
we have $P \left( (n-m_n+2)^{-1} \sum_{j=0}^{n-m_n+1} \alpha_{j,n}^2 - E[\alpha_{0,n}^2] > \varepsilon_1/4 \right) < K/n^{3/2}$. Furthermore since $E[\alpha_{0,n}^2] \to 0$ as $m_n \to \infty$ it follows from Lemma 1 that if $m_n \geq n^{1/4}$ and if $n$ is large enough, then $P \left( (n-m_n+2)^{-1} \sum_{j=0}^{n-m_n+1} \alpha_{j,n}^2 - E[\alpha_{0,n}^2] > \varepsilon_1/4 \right) < 2/n^2$.

Recall that $\mu(B_{\delta}^c) \leq \delta$. Then, since the sums in (7) are finite, an application of the Borel-Cantelli lemma yields almost surely that

$$\lim_{n \to \infty} \frac{1}{n-m_n+2} \sum_{j=0}^{n-m_n+1} \log^2 \left( \frac{n+1}{m_n} (U_{(j+m_n)} - U_{(j)}) \right) \leq \lim_{n \to \infty} 2E[\alpha_{0,n}^2] \leq 6C_2$$

implying

$$\lim_{n \to \infty} \frac{1}{n-m_n+2} \sum_{j=0}^{n-m_n+1} \left| \log \left( \frac{n+1}{m_n} (U_{(j+m_n)} - U_{(j)}) \right) \right| I_{B_{\delta}^c}(U_{(j)}) \leq \sqrt{6C_2 \delta}.$$

This inequality, together with (4)-(6) yields the desired result in the limit as $\varepsilon, \delta \to 0$. □

**Corollary 1** Under the assumptions of Theorem 3, if $g(x)$ is bounded away from zero a.e., i.e. $g(x) \geq \rho > 0$ a.e.,

$$\lim_{n \to \infty} \frac{1}{n-m_n+2} \sum_{j \in J_B} \log \left( \frac{G(U_{(j+m_n)}) - G(U_{(j)})}{U_{(j+m_n)} - U_{(j)}} \right) = \int_B \log g(x) dx \quad (a.s.).$$

**Proof.** Using the notation of the proof of Theorem 3, for $n$ large enough and for $\varepsilon < \rho$,

$$\frac{1}{n-m_n+2} \left( \sum_{j \in J_N} \log \left( \frac{G(U_{(j+m_n)}) - G(U_{(j)})}{U_{(j+m_n)} - U_{(j)}} \right) + \sum_{j \in J_{B_{\delta}^c}} \log \left( \frac{G(U_{(j+m_n)}) - G(U_{(j)})}{U_{(j+m_n)} - U_{(j)}} \right) \right)$$

$$\geq \frac{1}{n-m_n+2} \sum_{j \in J_N} \log (g(U_{(j)})) - \varepsilon + \frac{1}{n-m_n+2} \sum_{j \in J_{B_{\delta}^c}} \log \rho$$

$$\to \int_{B_{\delta}^c} \log (g(x) - \varepsilon) dx + \mu(B_{\delta}^c) \log \rho, \quad (a.s.).$$

By letting $\varepsilon$ and $\delta$ tend to zero we obtain the desired result. □

By combining Theorems 1 and 3 gives the following corollary for

$$L_n^{(m)}(G(U)) = \frac{1}{n-m+2} \sum_{j=0}^{n-m+1} \log (G(U_{(j+m)}) - G(U_{(j)})) .$$

**Corollary 2** Under the assumptions of Theorem 3,

$$\lim_{n \to \infty} \left( L_n^{(m)}(G(U)) - \psi(m_n) + \log m_n \right) \leq \int_0^1 \log g(x) dx \quad (a.s.).$$
A function $F$ will be called a pseudo distribution if and only if there exists a constant $c$ such that $cF$ is a distribution function.

**Corollary 3** Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with distribution function $F_{\theta_0}$ and density function $f_{\theta_0}$. Let $\{m_n\}$ be a sequence of positive integers such that $m_n = o(n/\log n)$. Then, for any measurable set $A$ of real numbers and for any pseudo distribution $F_\theta$, with (pseudo) density $f_\theta$, almost surely,

$$
\lim_{n \to \infty} \frac{1}{n - m_n + 2} \sum_{j \in J_A} \log \left( \frac{F_\theta(X_{j+m_n}) - F_\theta(X_j)}{F_{\theta_0}(X_{j+m_n}) - F_{\theta_0}(X_j)} \right) \leq \int_A \log \left( \frac{f_\theta(x)}{f_{\theta_0}(x)} \right) dF_{\theta_0}(x),
$$

where $J_A = \{ j : X_{(j)} \in A, 0 \leq j \leq n - m_n + 1 \}$. If the index set is redefined as $J_A = \{ j : [X_{(j)}, X_{(j+m_n)}] \subseteq A, 0 \leq j \leq n - m_n + 1 \}$, the inequality above holds if $A$ is a finite union of intervals.

**Proof.** The proof follows from Theorem 3 with $G(x) = F_\theta(F_{\theta_0}^{-1}(x))$, $F_{\theta_0}^{-1}(x) = \inf\{ t : F_{\theta_0}(t) \geq x \}$ since $F_{\theta_0}(X_1)$ is uniformly distributed and $F_\theta(X_1) \equiv F_\theta(F_{\theta_0}^{-1}(F_{\theta_0}(X_1)))$. □

The preceding corollary is used in Ekström (1997), where it plays a significant role in a proof of consistency of a so-called generalized maximum spacing (GMSP) estimator. GMSP-estimators are based on certain approximations of different information measures, e.g. the Kullback-Leibler information, Jeffreys’ divergence and the Hellinger distance, and can be regarded as alternatives to the maximum likelihood (ML) estimator. As for the ML-method, an unknown distribution $F_{\theta_0}$ should be estimated from an i.i.d. sample of random variables $X_1, X_2, \ldots, X_n$. Under the assumption that $F_{\theta_0}$ belongs to a family $\mathcal{F} = \{ F_\theta : \theta \in \Theta \}$ of distributions, absolutely continuous with respect to the Lebesgue measure, the GMSP estimate of $F_{\theta_0}$ based on the Kullback-Leibler information is the distribution in $\mathcal{F}$ that maximizes

$$
L_n^{(m_n)}(F_\theta(X)) = \frac{1}{n - m_n + 2} \sum_{j=0}^{n-m_n+1} \log \left( \frac{n+1}{m_n} \left( F_\theta(X_{j+m_n}) - F_\theta(X_j) \right) \right).
$$

The corollary above and Theorem 1, together with Jensen’s inequality, imply that the limit of this statistic is maximized if and only if $F_\theta(x) = F_{\theta_0}(x)$. Thus, for large samples it can be intuitively expected that those distributions which are most probable to have generated the data should be found when $L_n^{(m_n)}(F_\theta(X))$ is maximized.

The (original) maximum spacing method, as introduced by Cheng and Amin (1983) and Ranneby (1984), was based on simple spacings. Ranneby proposed the method as an alternative to the maximum likelihood method, and showed that both these methods can be obtained from approximations of the Kullback-Leibler information $I(f_\theta, f_{\theta_0}) = \int f_{\theta_0}(x) \log \left( f_{\theta_0}(x)/f_\theta(x) \right) dx$ (see Kullback and
Leibler (1951)). This basic idea, that an estimation method may be obtained from an approximation of the Kullback-Leibler information, is applicable even if high order spacings are used. It may be useful to consider \( L_n^{(m_n)} \) with \( m_n > 1 \), to overcome the difficulties with very small values. Because the effect on \( L_n^{(1)} \) of inaccurate measurements of the values \( D_{j,n}^{(1)} = X_{(j+1)} - X_{(j)} \) may be considerable, especially for small values, a small inaccuracy in \( D_{j,n}^{(1)} \) produces a big error in \( \log D_{j,n}^{(1)} \).

Corollary 3 is not only valuable for the (generalized) maximum spacing method, it can also be interpreted in terms of goodness of fit tests. Suppose \( X_1, X_2, ..., X_n \) is an i.i.d. sequence of random variables with distribution \( F \), belonging to a family \( \mathcal{F} \) of distributions, absolutely continuous with respect to the Lebesgue measure. Let the hypothesis to be tested be \( H_0 : F(x) = F_0(x) \), against the alternative \( H_A : F(x) \neq F_0(x) \). Under the null-hypothesis it follows, for \( m \) fixed, that

\[
V_{n,1}^{(m)}(F_0(X)) = \frac{1}{\sqrt{n - m + 2}} \left( \sum_{j=0}^{n-m+1} \log \left( (n+1) \left( F_0(X_{(j+m)}) - F_0(X_{(j)}) \right) \right) - \psi(m) \right)
\rightarrow N(0, \sigma^2)
\]

where \( \sigma^2 = (2m^2 - 2m + 1)\psi'(m) - 2m + 1 \) (see Cressie (1976) and Holst (1979)). Cressie (1976) also showed, for \( m = m_n \) growing to infinity such that \( m = o(n^{1/3}) \), that

\[
V_{n,2}^{(m)}(F_0(X)) = \sqrt{\frac{3m}{n}} \left( \sum_{j=0}^{n-m+1} \log \left( (n+1) \left( F_0(X_{(j+m)}) - F_0(X_{(j)}) \right) \right) - \psi(m) \right)
\rightarrow N(0, 1).
\]

Thus, using \( V_{n,i}^{(m)}(F_0(X)) \) as a test statistic, critical values can be obtained (for large samples) from a standard normal table. The null-hypothesis can therefore be rejected for large values of \( |V_{n,i}^{(m)}(F_0(X))| \).

Suppose that the true underlying distribution is \( F_* \neq F_0, F_* \in \mathcal{F} \). Then by Corollary 3 and Theorem 1,

\[
\lim_{n \to \infty} \frac{V_{n,1}^{(m)}(F_0(X))}{\sqrt{n - m + 2}} \leq \int_{-\infty}^{\infty} \log \left( \frac{f_0(x)}{f_*(x)} \right) dF_*(x) \quad (a.s.),
\]

and thus, by Jensen's inequality,

\[
\lim_{n \to \infty} \left| V_{n,1}^{(m)}(F_0(X)) \right| \geq \lim_{n \to \infty} -\sqrt{n - m + 2} \int_{-\infty}^{\infty} \log \left( \frac{f_0(x)}{f_*(x)} \right) dF_*(x) = \infty \quad (a.s.),
\]

where \( f_0 \) and \( f_* \) are the densities corresponding to \( F_0 \) and \( F_* \), respectively. The case when \( m \) grows to infinity is similar. Consequently, the tests are strongly consistent against all continuous alternatives.
Theorem 4 Let $X_1, X_2, ..., X_n$ be i.i.d. random variables on $[0,1]$ with distribution function $F$. Suppose that $\{m_n\}$ is a sequence of positive integers such that $m_n = o(n/\log n)$. Then,

$$\lim_{n \to \infty} \left( L_n^{(m_n)}(X) - \psi(m_n) + \log m_n \right) \leq \int_0^1 \log h(x) dx \quad (a.s.),$$

where $h(x) = \frac{d}{dx} F^{-1}(x)$ a.e., $F^{-1}(x) = \inf\{t : F(t) \geq x\}$. Moreover, if $F$ has a density $f$, then $\int_0^1 \log h(x) dx = -\int_0^1 f(x) \log f(x) dx$.

Proof. If $F$ is continuous then $X_1 \overset{d}{=} F^{-1}(F(X_1))$ and $F(X_1)$ is uniformly distributed, so the inequality in the theorem follows as a direct consequence of Corollary 2 with $G(x) = F^{-1}(x)$. Finally if $F$ has a density $f$, then $\frac{d}{dx} F^{-1}(x) = 1/f(F^{-1}(x))$ and the equality $\int_0^1 \log h(x) dx = -\int_0^1 f(x) \log f(x) dx$ follows by a change of variables. \qed

Remark. The integral $-\int f(x) \log f(x) dx$ is known as the entropy of an absolutely continuous distribution $F$ with density $f$, see Shannon (1948). For a review of methods, including methods based on spacings, to estimate entropy and their statistical properties, see Dudewicz and Van der Meulen (1987). See also Correa (1995), who introduces a different estimator of entropy based on spacings.

Proof of Theorem 2. The result follows by Theorem 1 and Theorem 4, together with Jensen's inequality. \qed

By assuming that $h(x)$ in Theorem 4 is bounded away from zero, using the ideas of Corollary 1, Theorem 4 becomes a two-sided strong limit theorem.

Corollary 4 Under the assumptions of Theorem 4, if $h(x)$ is bounded away from zero a.e., i.e. $h(x) \geq \rho > 0$ a.e.,

$$\lim_{n \to \infty} \left( L_n^{(m_n)}(X) - \psi(m_n) + \log m_n \right) = \int_0^1 \log h(x) dx \quad (a.s.).$$

If $F$ has a density $f$, then the equality holds if $f$ is bounded away from infinity. In this case $\int_0^1 \log h(x) dx = -\int_0^1 f(x) \log f(x) dx$.

Remark. For $m_n$ even, $L_n^{(m_n)}(X)$ is related to Vasicek's (1976) entropy estimator. Vasicek (1976) shows, by rewriting his estimator using Stieltjes sums, that his entropy estimator is weakly consistent. Beirlant and Van Zuijlen (1985) show, by using their Glivenko-Cantelli strong limit theorem for the distribution of uniform $m$th order spacings, that Vasicek's entropy estimator is strongly consistent under the assumptions that the variance of $X_1$ is finite and that $m_n \to \infty$ and $m_n = O(n^{1-\varepsilon})$ for some $0 < \varepsilon < 1$. Also, by using Vasicek's (1976) approach,
Van Es (1992) give a strong limit theorem for a statistic that has $L_n^{(m_n)}(X)$ as a special case. However Van Es (1992) imposes more restrictive assumptions on the density $f$, i.e. that it is bounded away from infinity and zero on its support, which is assumed to be an interval. He considers both the case when $m_n$ is fixed and the case when $m_n/\log n \to \infty$ and $m_n/n \to 0$.

Remark. It should be observed that Theorem 4 and Corollary 4 hold for distributions on intervals other than the $[0,1]$. In the proofs it is sufficient that $F^{-1}$ be bounded, and this holds for any distribution $F$ defined on any bounded interval.

References


