Nonparametric Statistical Methods and Related Topics

A Festschrift in Honor of Professor P K Bhattacharya on the Occasion of His 80th Birthday

J Jiang ♦ G G Roussas ♦ F J Samaniego
editors

World Scientific

Chapter 8

U-Statistics Based on Higher-Order Spacings

David D. Tung and S. Rao Jammalamadaka

ALEKS Corporation, Irvine, California, and University of California, Santa Barbara

In this paper, we investigate the asymptotic theory for U-statistics based on higher-order sample spacings. The usual asymptotic theory for U-statistics does not fully apply here because spacings are dependent variables. However, under the null hypothesis, the higher-order spacings can be represented by conditionally independent Gamma random variables. We exploit this idea to derive the relevant asymptotic theory both under the null hypothesis $H_0 : F = F_0$ and under a sequence of close alternatives.

The generalized Gini mean difference of the higher-order sample spacings is a prime example of a U-statistic of this type. It is found that the Gini mean squared difference test based on these higher-order spacings is the asymptotically locally most powerful test in this class, and has the same efficacy as the generalized Greenwood statistic based on the sum of squares of the higher-order spacings.

Contents

1. Introduction

The basic goodness-of-fit problem consists of testing whether a given data set fits a specified distribution. In particular, consider independent and identically distributed random variables $X_1, X_2, \ldots, X_{n-1}$ from a continuous distribution function F defined on the real line $\mathbb{R}$. In the statistical

literature, much attention has been devoted to the nonparametric problem of simple goodness-of-fit, namely testing the null hypothesis

$$H_0 : F(x) = F_0(x),$$

where F_0 is a completely specified distribution function.

If F is assumed to be continuous as we shall do, by way of the probability integral transform, the support of F reduces to the unit interval $[0, 1]$, and this also permits us to equate F_0 with the Uniform($[0, 1]$) distribution. Thus, the goodness-of-fit problem reduces to one of testing uniformity, i.e. testing the null hypothesis

$$H_0 : F(x) = x \cdot I(0 \leq x \leq 1).$$

Let $X_{(1)}, X_{(2)}, \ldots, X_{(n-1)}$ denote the sample order statistics. Put $X_{(0)} \equiv 0$ and $X_{(n)} \equiv 1$, so that $0 = X_{(0)} \leq X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)} = 1$. The *higher-order sample spacings* are defined by the random variables

$$D_{km}^{(m)} = X_{(km)} - X_{((k-1)m)}, \qquad \text{for } k = 1, 2, \ldots, N = \lfloor n/m \rfloor. \tag{1}$$

If F is the Uniform($[0, 1]$) distribution, as under the null hypothesis, we use the notation $\{U_k\}$ for the sample observations and

$$T_{km}^{(m)} = U_{(km)} - U_{((k-1)m)}, \qquad \text{for } k = 1, 2, \ldots, N = \lfloor n/m \rfloor \tag{2}$$

for the *higher-order uniform spacings*. When $m = 1$, then $\{D_k^{(1)}\}$ or simply $\{D_k\}$ are called the *sample spacings*, and likewise $\{T_k^{(1)}\}$ or simply $\{T_k\}$ are called the *uniform spacings*. Tests based on higher-order spacings are studied here for testing the null hypothesis. Since we are mainly interested in asymptotic properties, we assume that $N = n/m$, without loss of generality.

Let $h : [0, \infty) \times [0, \infty) \to \mathbb{R}$ be a symmetric function in its arguments satisfying some regularity conditions. Consider the general test statistic

$$W_N(h) = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} h(nD_{im}^{(m)}, nD_{jm}^{(m)}) \tag{3}$$

which is a second-order U-statistic of the higher-order sample spacings and symmetric in the pairs $(D_{im}^{(m)}, D_{jm}^{(m)})$. An important example of such a statistic is the generalized Gini mean difference of the higher-order sample spacings, i.e.

$$G_N(r) = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \left| nD_{im}^{(m)} - nD_{jm}^{(m)} \right|^r, \qquad r > 0 \tag{4}$$

which is an average of absolute pairwise differences of the higher-order sample spacings to the r^{th} power. The special case of $G_N(1)$ is the Gini mean difference higher-order spacings test, and the special case of $G_N(2)$ will be called the Gini mean squared difference higher-order spacings test.

On the other hand, symmetric sum-functions of the higher-order spacings, i.e. general test statistics of the form

$$V_N(g) = \frac{1}{N}\sum_{k=1}^{N} g\left(nD_{km}^{(m)}\right), \tag{5}$$

where $g(\cdot)$ is a real-valued function satisfying some regularity conditions, are symmetric in $\{D_{km}^{(m)}\}$, and can also be thought of as first-order U-statistics of the higher-order sample spacings.

For the special case of $m = 1$ where $N = n$, note that the general test statistic $W_N(h)$ reduces to

$$W_n(h) = \frac{2}{n(n-1)}\sum_{1\le i<j\le n} h(nD_i, nD_j) \tag{6}$$

and the general test statistic $V_N(g)$ reduces to

$$V_n(g) = \frac{1}{n}\sum_{k=1}^{n} g(nD_k). \tag{7}$$

Likewise, the statistic $G_N(r)$ reduces to

$$G_n(r) = \frac{2}{n(n-1)}\sum_{1\le i<j\le n} |nD_i - nD_j|^r, \qquad r > 0 \tag{8}$$

the generalized Gini mean difference of the sample spacings.

The generalized Gini mean difference spacings test $G_n(1)$ was proposed in Jammalamadaka and Goria (2004) for testing goodness-of-fit. There they derive both the exact and asymptotic distribution of $G_n(1)$ under the null hypothesis, and show that it has good performance based on Monte Carlo powers. More generally, the asymptotic theory for $W_n(h)$ has been studied in Tung and Jammalamadaka (2010). They show that the asymptotically locally most powerful (ALMP) test for U-statistics based on the spacings is the Gini mean squared difference spacings test

$$G_n(2) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(nD_i - nD_j)^2 \tag{9}$$

which corresponds to the classical Greenwood statistic $\frac{1}{n}\sum_{k=1}^{n}(nD_k)^2$ based on the sum of squares of the spacings. The importance of the

Greenwood statistic is somewhat justified in view of the result established in Sethuraman and Rao (1970) that among the class of symmetric sum-functions of the spacings, i.e., for $V_n(g)$, the Greenwood statistic is the ALMP test.

There are asymptotically more efficient procedures than the Greenwood statistic if one takes into consideration test statistics based on higher-order spacings. Del Pino (1979) showed that among the class of symmetric sum-functions of the higher-order spacings, i.e., for $V_N(g)$, the generalized Greenwood statistic, i.e., $\frac{1}{N}\sum_{k=1}^{N}(nD_{km}^{(m)})^2$ based on the sum of squares of these higher-order spacings is the ALMP test, and asymptotically more efficient than the classical Greenwood statistic.

One perceived weakness of symmetric spacings tests is that they cannot discriminate alternatives converging to the null hypothesis at a rate faster than $n^{-1/4}$, and hence have poor asymptotic performance as compared to say the Kolmogorov–Smirnov test.

The asymptotic distribution for second-order U-statistics of the higher-order spacings under the null hypothesis is studied in the next section. Section 3 deals with their asymptotic behavior under a sequence of close alternatives. Section 4 contains results on the ALMP test for the class of U-statistics of the higher-order sample spacings.

2. The Asymptotic Null Distribution

In this section, we obtain the asymptotic distribution for second-order U-statistics of the higher-order uniform spacings, i.e., general test statistics of the form

$$W_N(h) = \frac{2}{N(N-1)} \sum_{1\le i<j\le N} h(nT_{im}^{(m)}, nT_{jm}^{(m)}) \tag{10}$$

under the null hypothesis. Since we are only interested in asymptotic properties, we shall assume that $N = n/m$, without loss of generality, and increases to infinity while m is fixed. Under the null hypothesis, the higher-order uniform spacings have the well-known conditional representation

$$\left(nT_{1m}^{(m)}, nT_{2m}^{(m)}, \ldots, nT_{Nm}^{(m)}\right) \simeq \left(S_1, S_2, \ldots, S_N \,\middle|\, \bar{S}_N = m\right), \tag{11}$$

where $S_1, S_2, \ldots, S_N$ are independent Gamma$(m, 1)$ random variables (cf. with Sobel and Wells (1990) and Rao and Sobel (1980)). Note that for

$m = 1$ where $N = n$, this reduces to

$$(nT_1, nT_2, \ldots, nT_n) \simeq \left(Z_1, Z_2, \ldots, Z_n \,\middle|\, \bar{Z}_n = 1 \right), \tag{12}$$

which is a conditional representation of the uniform spacings in terms of independent Exponential(1) random variables $Z_1, Z_2, \ldots, Z_n$. Here as elsewhere, we use $\simeq$ to denote the distributional equivalence of quantities on the left and right hand sides of the symbol.

There are at least two known approaches to deriving the asymptotic null distribution. One approach is by applying a conditional limit theorem for U-statistics due to Holst (1981, Theorem 6.2). A second approach is by way of the well-known Hoeffding decomposition for U-statistics, which connects the asymptotic theory for U-statistics of the higher-order uniform spacings with the asymptotic theory for symmetric sum-functions of the higher-order uniform spacings.

Let

$$U_N = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} h(S_i, S_j) \tag{13}$$

be a second-order U-statistic based on the independent Gamma$(m, 1)$ random variables $S_1, S_2, \ldots, S_N$, where the kernel $h : [0, \infty) \times [0, \infty) \to \mathbb{R}$ is a symmetric function with $\mathrm{Var}[h(S_1, S_2)] < \infty$.

For the case of independent and identically distributed random variables, the Hoeffding decomposition (cf. Lee (1990, Sec. 1.6)) asserts that a U-statistic of order k is a linear combination of uncorrelated U-statistics of order $1, 2, \ldots, k$. The case for $k = 2$ has been most studied and best understood. We state the Hoeffding decomposition for U_N in the following.

Lemma 1. *The Hoeffding decomposition of U_N has the form*

$$U_N = \theta + \frac{2}{N} \sum_{k=1}^{N} h^{(1)}(S_k) + \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} h^{(2)}(S_i, S_j) \tag{14}$$

where

$$\begin{aligned} \theta &= \mathbb{E}[h(S_1, S_2)] \\ g(t) &= \mathbb{E}[h(t, S_2)] = \mathbb{E}[h(S_1, S_2) \mid S_1 = t] \\ h^{(1)}(t) &= g(t) - \theta \\ h^{(2)}(s_1, s_2) &= h(s_1, s_2) - g(s_1) - g(s_2) + \theta. \end{aligned}$$

Moreover, the normalized U-statistic

$$\sqrt{N}(U_N - \theta) = \frac{2}{\sqrt{N}} \sum_{k=1}^{N} [g(S_k) - \theta] + N^{1/2} R_N \tag{15}$$

where

$$N^{1/2} R_N = \frac{2\sqrt{N}}{N(N-1)} \sum_{1 \le i < j \le N} h^{(2)}(S_i, S_j) = o_P(1). \tag{16}$$

The following result provides some useful identities for the expectations, variances, and covariances of the U-statistic kernels in the Hoeffding decomposition.

Lemma 2. *Let S_1, S_2 and S_3 be independent Gamma$(m, 1)$ random variables, and let g be defined as in Lemma 1. Then*

$$\begin{aligned}
\mathbb{E}[h(S_1, S_2)] &= \mathbb{E}[g(S_1)] \\
Cov[h(S_1, S_2), h(S_1, S_3)] &= Var[g(S_1)] \\
Cov[h(S_1, S_2), S_1] &= Cov[g(S_1), S_1] \\
&\quad Cov[h(S_1, S_2), (S_1 - m - 1)^2 + (S_2 - m - 1)^2] \\
&= 2 \cdot Cov[g(S_1), (S_1 - m - 1)^2].
\end{aligned}$$

The next result gives the asymptotic null distribution for symmetric sum-functions of the higher-order uniform spacings (cf. with Del Pino (1979)). We use the notation $N_1(\mu, \sigma^2)$ to denote the one-dimensional Normal distribution with mean μ and variance σ^2.

Lemma 3. *Under the null hypothesis, in the limit as $N \to \infty$,*

$$\frac{1}{\sqrt{N}} \sum_{k=1}^{N} [g(nT_{km}^{(m)}) - \mathbb{E}g(S_1)] \xrightarrow{D} N_1(0, \sigma^2(g)) \tag{17}$$

where

$$\sigma^2(g) = Var[g(S_1)] - \frac{Cov^2[g(S_1), S_1]}{Var[S_1]}. \tag{18}$$

The next result gives the asymptotic distribution for U-statistics based on the higher-order uniform spacings under the null hypothesis.

Theorem 1. *Under the null hypothesis, in the limit as $N \to \infty$,*

$$\sqrt{N} \left(\frac{2}{N(N-1)} \sum_{1 \le i < j \le N} h(nT_{im}^{(m)}, nT_{jm}^{(m)}) - \mathbb{E}[h(S_1, S_2)] \right) \xrightarrow{D} N_1(0, \sigma^2(h)), \tag{19}$$

where

$$\sigma^2(h) = 4(\sigma_1^2 - \sigma_{12}^2)$$
$$\sigma_1^2 = Cov[h(S_1, S_2), h(S_1, S_3)]$$
$$\sigma_{12}^2 = \frac{Cov^2[h(S_1, S_2), S_1]}{Var[S_1]}.$$

Proof. By Lemma 1, and the conditional representation of the higher-order uniform spacings (11), we have

$$\begin{aligned}
&\sqrt{N}\left(\frac{2}{N(N-1)}\sum_{1\le i<j\le N} h(nT_{im}^{(m)}, nT_{jm}^{(m)}) - \mathbb{E}[h(S_1, S_2)]\right) \\
&\simeq \left(\sqrt{N}(U_N - \theta)\,\middle|\, \bar{S}_N = m\right) \\
&\simeq \left(\frac{2}{\sqrt{N}}\sum_{k=1}^{N}[g(S_k) - \mathbb{E}g(S_1)]\,\middle|\, \bar{S}_N = m\right) + o_P(1) \\
&\simeq \frac{2}{\sqrt{N}}\sum_{k=1}^{N}[g(nT_{km}^{(m)}) - \mathbb{E}g(S_1)] + o_P(1) \xrightarrow{D} N_1(0, 4\sigma^2(g)).
\end{aligned}$$

The convergence in distribution to the $N_1(0, 4\sigma^2(g))$ distribution follows from Lemma 3 and Slutsky's Theorem. By Lemma 2, the asymptotic variance

$$4\sigma^2(g) = 4\left(\mathrm{Var}[g(S_1)] - \frac{\mathrm{Cov}^2[g(S_1), S_1]}{\mathrm{Var}[S_1]}\right) = 4(\sigma_1^2 - \sigma_{12}^2) = \sigma^2(h). \quad (20)$$

This completes the proof. □

3. The Asymptotic Distribution Under a Sequence of Close Alternatives

In this section, we derive the asymptotic distribution for second-order U-statistics of the higher-order sample spacings under a sequence of close alternatives. In order to study asymptotic efficiencies, one needs to obtain the asymptotic distribution of test statistics under a sequence of close alternatives (also called smooth alternatives), which converges to the null hypothesis. Thus, we specify the alternative hypothesis by a sequence of distribution functions $\{F_n(x) : n \ge 1\}$ that converges to the Uniform$([0, 1])$ distribution function, which corresponds to the null hypothesis, in the limit as $n \to \infty$.

For symmetric spacings tests, the appropriate sequence of close alternatives (cf. with Sethuraman and Rao (1970) and Rao and Sethuraman (1975)) is obtained by letting the distribution function

$$F_n(x) = x + \frac{L_n(x)}{n^{1/4}}, \qquad \text{for } 0 \le x \le 1 \tag{21}$$

where $L_n(0) = L_n(1) = 0$. We further assume that $L_n(x)$ is twice differentiable on the unit interval $[0,1]$ and that there exists a function $L(x)$ which is twice continuously-differentiable with $L(0) = L(1) = 0$ and

$$\begin{aligned} n^{1/4} \sup_{0\le x\le 1} |L_n(x) - L(x)| &= o(1) \\ n^{1/4} \sup_{0\le x\le 1} |L_n'(x) - l(x)| &= o(1) \\ n^{1/4} \sup_{0\le x\le 1} |L_n''(x) - l'(x)| &= o(1), \end{aligned}$$

where $l(x)$ and $l'(x)$ are respectively the first and second derivatives of $L(x)$. Note that $L(x) = \int_0^x l(u)\,du$ and $\int_0^1 l(u)\,du = 0$ by the fundamental theorem of calculus.

The next result gives, under a sequence of close alternatives, the asymptotic distribution for symmetric sum-functions of the higher-order spacings (cf. Del Pino (1979)).

Lemma 4. *Under the close alternatives* (21), *in the limit as* $N \to \infty$,

$$\frac{1}{\sqrt{N}} \sum_{k=1}^{N} [g(nD_{km}^{(m)}) - \mathbb{E}g(S_1)] \xrightarrow{D} N_1(\mu(g), \sigma^2(g)), \tag{22}$$

where

$$\begin{aligned} \mu(g) &= \frac{m^{-1/2}}{2} \left(\int_0^1 l^2(u)\,du \right) Cov[g(S_1), (S_1 - m - 1)^2] \\ \sigma^2(g) &= Var[g(S_1)] - \frac{Cov^2[g(S_1), S_1]}{Var[S_1]}. \end{aligned}$$

Let $0 = \xi_0 < \xi_1 < \xi_2 < \cdots < \xi_n < \xi_{N+1} = 1$ form a partition of the unit interval $[0,1]$, where

$$\xi_k = \frac{k}{N+1}, \qquad \text{for } k = 0, 1, 2, \ldots, N+1.$$

Under the close alternatives, the higher-order spacings $\{D_{km}^{(m)}\}$ are related to the uniform spacings $\{T_{km}^{(m)}\}$ by the relation

$$\begin{aligned} nD_{km}^{(m)} &= n[F_n^{-1}(U_{(km)}) - F_n^{-1}(U_{((k-1)m)})] \\ &= nT_{km}^{(m)} + \left(\frac{-l(\xi_k)}{n^{1/4}} + \frac{l^2(\xi_k) + L(\xi_k)l'(\xi_k)}{n^{1/2}}\right)(nT_{km}^{(m)}) + o_P(n^{-1/2}) \end{aligned} \tag{23}$$

where $o_P(\cdot)$ is uniform in k. This follows from the mean value theorem for differential calculus and a continuity argument found in Rao and Sethuraman (1975).

We assume that $h : [0,\infty)\times[0,\infty) \to \mathbb{R}$ is a symmetric function with first and second-order partial derivatives. We use the notation $h_x(x,y) = \frac{\partial h}{\partial x}$ and $h_y(x,y) = \frac{\partial h}{\partial y}$ to denote the first partial derivatives of h, and use $h_{xx}(x,y) = \frac{\partial^2 h}{\partial x^2}$, $h_{yy}(x,y) = \frac{\partial^2 h}{\partial y^2}$ and $h_{xy}(x,y) = \frac{\partial^2 h}{\partial x \partial y}$ to denote the second-order partial derivatives of h.

By using (23), we prove the following.

Lemma 5. *Under the close alternatives* (21), *in the limit as* $N \to \infty$,

$$\begin{aligned} &\frac{2\sqrt{N}}{N(N-1)} \sum_{1\le i<j\le N} [h(nD_{im}^{(m)}, nD_{jm}^{(m)}) - h(nT_{im}^{(m)}, nT_{jm}^{(m)})] \\ &\xrightarrow{\mathbb{P}} \frac{m^{-1/2}}{2}\left(\int_0^1 l^2(u)\,du\right) \cdot Cov[h(S_1,S_2), (S_1 - m - 1)^2 + (S_2 - m - 1)^2]. \end{aligned} \tag{24}$$

Proof. Using (23) in a two-dimensional Taylor expansion of $h(nD_{im}^{(m)}, nD_{jm}^{(m)})$ around $h(nT_{im}^{(m)}, nT_{jm}^{(m)})$ and summing over all $i < j$ gives

$$\begin{aligned} &\frac{2\sqrt{N}}{N(N-1)} \sum_{1\le i<j\le N} [h(nD_{im}^{(m)}, nD_{jm}^{(m)}) - h(nT_{im}^{(m)}, nT_{jm}^{(m)})] \\ &= -\frac{2m^{-1/4}N^{1/4}}{N(N-1)} \sum_{1\le i<j\le N} l(\xi_i)(nT_{im}^{(m)})\, h_x(nT_{im}^{(m)}, nT_{jm}^{(m)}) \\ &\quad - \frac{2m^{-1/4}N^{1/4}}{N(N-1)} \sum_{1\le i<j\le N} l(\xi_j)(nT_{jm}^{(m)})\, h_y(nT_{im}^{(m)}, nT_{jm}^{(m)}) \\ &\quad + \frac{2m^{-1/2}}{N(N-1)} \sum_{1\le i<j\le N} [l^2(\xi_i) + L(\xi_i)l'(\xi_i)](nT_{im}^{(m)})\, h_x(nT_{im}^{(m)}, nT_{jm}^{(m)}) \end{aligned}$$

$$+\frac{2m^{-1/2}}{N(N-1)}\sum_{1\le i<j\le N}[l^2(\xi_j)+L(\xi_j)l'(\xi_j)](nT_{jm}^{(m)})\,h_y(nT_{im}^{(m)},nT_{jm}^{(m)})$$

$$+\frac{2m^{-1/2}}{N(N-1)}\sum_{1\le i<j\le N}l(\xi_i)l(\xi_j)(nT_{im}^{(m)})(nT_{jm}^{(m)})\,h_{xy}(nT_{im}^{(m)},nT_{jm}^{(m)})$$

$$+\frac{m^{-1/2}}{N(N-1)}\sum_{1\le i<j\le N}l^2(\xi_i)(nT_{im}^{(m)})^2\,h_{xx}(nT_{im}^{(m)},nT_{jm}^{(m)})$$

$$+\frac{m^{-1/2}}{N(N-1)}\sum_{1\le i<j\le N}l^2(\xi_j)(nT_{jm}^{(m)})^2\,h_{yy}(nT_{im}^{(m)},nT_{jm}^{(m)})+o_P(1).$$

The composite trapezoid rule asserts that there exists a number $c\in(0,1)$ for which

$$\int_0^1 l(u)\,du=\frac{1}{N+1}\sum_{k=1}^{N}l(\xi_k)-\frac{l''(c)}{12(N+1)^2}. \tag{25}$$

Since we have

$$\lim_{N\to\infty}N^{1/4}\left(\frac{1}{N+1}\sum_{k=1}^{N}l(\xi_k)-\int_0^1 l(u)\,du\right)=\lim_{N\to\infty}\frac{N^{1/4}\cdot l''(c)}{12(N+1)^2}=0, \tag{26}$$

then the first two terms on the RHS of (24) converge in probability to zero.

Since $(nT_{im}^{(m)},nT_{jm}^{(m)})\xrightarrow{D}(S_1,S_2)$, as $n\to\infty$, observe that both

$$\frac{2m^{-1/2}}{N(N-1)}\sum_{1\le i<j\le N}[l^2(\xi_i)+L(\xi_i)l'(\xi_i)](nT_{im}^{(m)})\,h_x(nT_{im}^{(m)},nT_{jm}^{(m)})$$

$$\xrightarrow{\mathbb{P}}m^{-1/2}\cdot\mathbb{E}[S_1\,h_x(S_1,S_2)]\int_0^1[l^2(x)+L(x)l'(x)]\,dx=0$$

and

$$\frac{2m^{-1/2}}{N(N-1)}\sum_{1\le i<j\le N}[l^2(\xi_j)+L(\xi_j)l'(\xi_j)](nT_{jm}^{(m)})\,h_y(nT_{im}^{(m)},nT_{jm}^{(m)})$$

$$\xrightarrow{\mathbb{P}}m^{-1/2}\cdot\mathbb{E}[S_2\,h_y(S_1,S_2)]\int_0^1[l^2(y)+L(y)l'(y)]\,dy=0$$

because from integration by parts

$$\int_0^1 L(u)\,l'(u)\,du=-\int_0^1 l^2(u)\,du.$$

Observe also that

$$\frac{2m^{-1/2}}{N(N-1)} \sum_{1\le i<j\le N} l(\xi_i) l(\xi_j) (nT_{im}^{(m)})(nT_{jm}^{(m)})\, h_{xy}(nT_{im}^{(m)}, nT_{jm}^{(m)})$$
$$\xrightarrow{\mathbb{P}} m^{-1/2} \left(\int_0^1 \int_0^1 l(x)\, l(y)\, dx\, dy \right) \cdot \mathbb{E}[S_1 \cdot S_2\, h_{xy}(S_1, S_2)] = 0$$

because

$$\int_0^1 \int_0^1 l(x)\, l(y)\, dx\, dy = \left(\int_0^1 l(x)\, dx \right) \left(\int_0^1 l(y)\, dy \right) = 0. \tag{27}$$

Moreover, we have

$$\frac{m^{-1/2}}{N(N-1)} \sum_{1\le i<j\le N} l^2(\xi_i) (nT_{im}^{(m)})^2\, h_{xx}(nT_{im}^{(m)}, nT_{jm}^{(m)})$$
$$\xrightarrow{\mathbb{P}} \frac{m^{-1/2}}{2} \left(\int_0^1 l^2(x)\, dx \right) \cdot \mathbb{E}[S_1^2\, h_{xx}(S_1, S_2)]$$

and

$$\frac{m^{-1/2}}{N(N-1)} \sum_{1\le i<j\le N} l^2(\xi_j) (nT_{jm}^{(m)})^2\, h_{yy}(nT_{im}^{(m)}, nT_{jm}^{(m)})$$
$$\xrightarrow{\mathbb{P}} \frac{m^{-1/2}}{2} \left(\int_0^1 l^2(y)\, dy \right) \cdot \mathbb{E}[S_2^2\, h_{yy}(S_1, S_2)]$$

with

$$\mathbb{E}[S_1^2\, h_{xx}(S_1, S_2) + S_2^2\, h_{yy}(S_1, S_2)]$$
$$= \mathrm{Cov}[h(S_1, S_2), (S_1 - m - 1)^2 + (S_2 - m - 1)^2].$$

This completes the proof. □

By combining Theorem 1 with Lemma 5, we have the following.

Theorem 2. *Under the close alternatives* (21), *in the limit as* $N \to \infty$,

$$\sqrt{N} \left(\frac{2}{N(N-1)} \sum_{1\le i<j\le N} h(nD_{im}^{(m)}, nD_{jm}^{(m)}) - \mathbb{E}[h(S_1, S_2)] \right)$$
$$\xrightarrow{D} N_1(\mu(h), \sigma^2(h)),$$

where

$$\mu(h) = \frac{m^{-1/2}}{2}\left(\int_0^1 l^2(u)\,du\right) \cdot Cov[h(S_1, S_2), (S_1 - m - 1)^2 + (S_2 - m - 1)^2]$$

$$\sigma^2(h) = 4(\sigma_1^2 - \sigma_{12}^2)$$

$$\sigma_1^2 = Cov[h(S_1, S_2), h(S_1, S_3)]$$

$$\sigma_{12}^2 = \frac{Cov^2[h(S_1, S_2), S_1]}{Var[S_1]}.$$

Proof. Put

$$\sqrt{N}\left(\frac{2}{N(N-1)}\sum_{1\le i<j\le N} h(nD_{im}^{(m)}, nD_{jm}^{(m)}) - \mathbb{E}[h(S_1, S_2)]\right)$$

$$= \sqrt{N}\left(\frac{2}{N(N-1)}\sum_{1\le i<j\le N} h(nT_{im}^{(m)}, nT_{jm}^{(m)}) - \mathbb{E}[h(S_1, S_2)]\right)$$

$$+ \frac{2\sqrt{N}}{N(N-1)}\sum_{1\le i<j\le N}[h(nD_{im}^{(m)}, nD_{jm}^{(m)}) - h(nT_{im}^{(m)}, nT_{jm}^{(m)})]$$

$$\xrightarrow{D} N_1(\mu(h), \sigma^2(h)).$$

The convergence in distribution follows from Slutsky's Theorem, where the first term on the RHS converges to the $N_1(0, \sigma^2(h))$ distribution by Theorem 1, and the second term converges in probability to $\mu(h)$ by Lemma 5. This completes the proof. □

4. The Asymptotically Locally Most Powerful Test

Recall that $\mu(g)$ and $\sigma^2(g)$ denote the asymptotic mean and asymptotic variance corresponding to the general test statistic

$$V_N(g) = \frac{1}{N}\sum_{k=1}^{N} g(nD_{km}^{(m)})$$

under the sequence of close alternatives. Here it is assumed that $V_N(g)$ has been normalized to have asymptotic mean zero and finite variance under the null hypothesis. The *Pitman asymptotic relative efficiency* (ARE) of

$V_N(g_1)$ relative to $V_N(g_2)$ is given by

$$\mathrm{ARE}(g_1, g_2) = \left(\frac{e^2(g_1)}{e^2(g_2)}\right)^2 = \frac{\left(\frac{\mu^2(g_1)}{\sigma^2(g_1)}\right)^2}{\left(\frac{\mu^2(g_2)}{\sigma^2(g_2)}\right)^2}. \tag{28}$$

The quantity

$$e^2(g) = \frac{\mu^2(g)}{\sigma^2(g)} \tag{29}$$

is called the *efficacy* of the test $V_N(g)$. A test with maximum efficacy is the asymptotically locally most powerful (ALMP) test. In order to find the ALMP test, for symmetric sum-functions of the higher-order spacings, against the close alternatives, one needs to find a function $g(\cdot)$ which maximizes

$$e(g) = \frac{m^{-1/2}\left(\int_0^1 l^2(u)\,du\right)\mathrm{Cov}[g(S_1), (S_1 - m - 1)^2]}{2\left\{\mathrm{Var}[g(S_1)] - m^{-1}\cdot\mathrm{Cov}^2[g(S_1), S_1]\right\}^{1/2}}. \tag{30}$$

As mentioned before, the importance of the generalized Greenwood statistic is somewhat justified by the next two results, which were established by Del Pino (1979). The generalized Greenwood statistic is the ALMP test among the class of symmetric sum-functions of the higher-order spacings.

Lemma 6. *The functional $e(g)$ is maximized by taking $g(t) = t^2$, which in turn gives*

$$\max e(g) = \sqrt{\frac{m+1}{2}}\left(\int_0^1 l^2(u)\,du\right).$$

Lemma 7. *For symmetric sum-functions of the higher-order spacings, the asymptotically locally most powerful (ALMP) test of the null hypothesis against the sequence of close alternatives is to reject the null hypothesis when*

$$\sum_{k=1}^{N}\left(nD_{km}^{(m)}\right)^2 > C(\alpha), \tag{31}$$

where the critical value $C(\alpha)$ is determined by the level of significance α. The asymptotic distribution of this optimal statistic under the sequence of

close alternatives (21) *is given by*

$$\frac{1}{\sqrt{N}} \sum_{k=1}^{N} \left[\left(nD_{km}^{(m)} \right)^2 - m(m+1) \right]$$
$$\xrightarrow{D} N_1 \left(m^{1/2}(m+1) \left(\int_0^1 l^2(u)\, du \right), 2m(m+1) \right).$$

The asymptotic distribution under the null hypothesis is obtained by taking $l(u) = 0$ *in the above.*

Recall that $\mu(h)$ and $\sigma^2(h)$ denote the asymptotic mean and asymptotic variance corresponding to the general test statistic

$$W_N(h) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} h(nD_{im}^{(m)}, nD_{jm}^{(m)}) \tag{32}$$

under the sequence of close alternatives. It is assumed that $W_N(h)$ has been normalized to have asymptotic mean zero and finite variance under the null hypothesis. The Pitman ARE of $W_N(h_1)$ relative to $W_N(h_2)$ is given by

$$\mathrm{ARE}(h_1, h_2) = \left(\frac{e^2(h_1)}{e^2(h_2)} \right)^2 = \frac{\left(\frac{\mu^2(h_1)}{\sigma^2(h_1)} \right)^2}{\left(\frac{\mu^2(h_2)}{\sigma^2(h_2)} \right)^2} \tag{33}$$

and the quantity

$$e^2(h) = \frac{\mu^2(h)}{\sigma^2(h)} \tag{34}$$

is the efficacy of the test $W_N(h)$.

In order to find the ALMP test, for U-statistics of the higher-order spacings, against the sequence of close alternatives, we need to find a function h which maximizes the functional

$$e(h) = \frac{m^{-1/2} \left(\int_0^1 l^2(u)\, du \right) \cdot \mathrm{Cov}[h(S_1, S_2), (S_1 - m - 1)^2 + (S_2 - m - 1)^2]}{4 \left\{ \mathrm{Cov}[h(S_1, S_2), h(S_1, S_3)] - m^{-1} \cdot \mathrm{Cov}^2[h(S_1, S_2), S_1] \right\}^{1/2}}. \tag{35}$$

Lemma 8. *The functional* $e(h)$ *is maximized by taking the symmetric function* $h(s_1, s_2) = (s_1 - s_2)^2$, *which in turn gives*

$$\max e(h) = \sqrt{\frac{m+1}{2}} \left(\int_0^1 l^2(u)\, du \right) \tag{36}$$

which corresponds to that of the generalized Greenwood statistic.

Proof. It is enough to find a function h which maximizes the numerator in (35). By the Cauchy–Bunyakovsky–Schwarz Inequality, we have

$$
\begin{aligned}
& e(h) \\
&= \frac{m^{-1/2}\left(\int_0^1 l^2(u)\,du\right)\cdot \mathrm{Cov}[h(S_1,S_2),(S_1-m-1)^2+(S_2-m-1)^2]}{4\left\{\mathrm{Cov}[h(S_1,S_2),h(S_1,S_3)]-m^{-1}\cdot \mathrm{Cov}^2[h(S_1,S_2),S_1]\right\}^{1/2}} \\
&\le m^{-1/2}\left(\int_0^1 l^2(u)\,du\right) \\
&\qquad \times \frac{\sqrt{\mathrm{Var}[h(S_1,S_2)]}\sqrt{\mathrm{Var}[(S_1-m-1)^2+(S_2-m-1)^2]}}{4\left\{\mathrm{Cov}[h(S_1,S_2),h(S_1,S_3)]-m^{-1}\cdot \mathrm{Cov}^2[h(S_1,S_2),S_1]\right\}^{1/2}}.
\end{aligned}
\tag{37}
$$

The inequalities become equalities if and only if $h(s_1,s_2)=a[(s_1-m-1)^2+(s_2-m-1)^2]+b$, for some real numbers $a\neq 0$ and b. In this particular case, the functional $e(h)$ attains the upper bound in (37), i.e.

$$
\begin{aligned}
\max e(h) &= \frac{m^{-1/2}\left(\int_0^1 l^2(u)\,du\right) a\cdot \mathrm{Var}[(S_1-m-1)^2+(S_2-m-1)^2]}{4\sqrt{a^2\cdot \mathrm{Var}(S_1-m-1)^2-m^{-1}\cdot a^2\cdot \mathrm{Cov}^2[(S_1-m-1)^2,S_1]}} \\
&= \frac{m^{-1/2}\cdot\left(\int_0^1 l^2(u)\,du\right) a\cdot 4\cdot \mathbb{E}[S_1^2]}{4\sqrt{2m(m+1)a^2}} \\
&= \frac{m^{-1/2}m(m+1)\cdot\left(\int_0^1 l^2(u)\,du\right)}{\sqrt{2m(m+1)}} \\
&= \sqrt{\frac{m+1}{2}}\left(\int_0^1 l^2(u)\,du\right).
\end{aligned}
$$

On the other hand, since $\mathrm{Cov}[S_1S_2,(S_1-m-1)^2+(S_2-m-1)^2]=0$, the maximum of the functional $e(h)$ is attained by taking $h(s_1,s_2)=(s_1-s_2)^2$,

and directly from (35) we have

$$\begin{aligned}
&\max e(h) \\
&= \frac{m^{-1/2}\left(\int_0^1 l^2(u)\,du\right)\cdot \mathrm{Cov}[(S_1-S_2)^2,(S_1-m-1)^2+(S_2-m-1)^2]}{4\left\{\mathrm{Cov}[(S_1-S_2)^2,(S_1-S_3)^2]-m^{-1}\cdot \mathrm{Cov}^2[(S_1-S_2)^2,S_1]\right\}^{1/2}} \\
&= \frac{m^{-1/2}\left(\int_0^1 l^2(u)\,du\right)\cdot 2\cdot \mathbb{E}[S_1^2+S_2^2]}{4\sqrt{2m(3+m)-4m}} \\
&= \frac{4m^{1/2}(m+1)\left(\int_0^1 l^2(u)\,du\right)}{4\sqrt{2m(m+1)}} \\
&= \sqrt{\frac{m+1}{2}}\left(\int_0^1 l^2(u)\,du\right).
\end{aligned}$$

This completes the proof. □

By combining Theorem 2, and Lemma 8, we have the following.

Theorem 3. *For U-statistics of the higher-order spacings, the asymptotically locally most powerful (ALMP) test of the null hypothesis against the sequence of close alternatives is to reject the null hypothesis when*

$$\sum_{1\le i<j\le N}\left(nD_{im}^{(m)}-nD_{jm}^{(m)}\right)^2 > C(\alpha) \tag{38}$$

where the critical value $C(\alpha)$ is determined by the level of significance α. The asymptotic distribution of this optimal statistic under the sequence of close alternatives (21) is given by

$$\begin{aligned}
&\sqrt{N}\left(\frac{2}{N(N-1)}\sum_{1\le i<j\le N}\left(nD_{im}^{(m)}-nD_{jm}^{(m)}\right)^2-2m\right) \\
&\xrightarrow{D} N_1\left(2m^{1/2}(m+1)\left(\int_0^1 l^2(u)\,du\right),8m(m+1)\right).
\end{aligned}$$

The asymptotic distribution under the null hypothesis is obtained by taking $l(u)=0$ in the above.

From Theorem 3, under a sequence of close alternatives, the Gini mean squared difference higher-order spacings test

$$G_N(2)=\frac{2}{N(N-1)}\sum_{1\le i<j\le N}\left(nD_{im}^{(m)}-nD_{jm}^{(m)}\right)^2 \tag{39}$$

is the ALMP test for U-statistics of the higher-order spacings. This also means that $G_N(2)$ is the best test among the generalized Gini mean difference higher-order spacings tests $G_N(r), r > 0$. Moreover, by Lemma 8, the Gini mean squared difference higher-order spacings test has efficacy

$$e^2(h) = \frac{m+1}{2}\left(\int_0^1 l^2(u)\,du\right)^2$$

which is the same as that of the generalized Greenwood statistic.

We compare the efficacies of the Gini mean squared difference higher-order spacings test, and the Kullback-Leibler divergence higher-order spacings test

$$\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\left[\left(nD_{km}^{(m)}\right)\log\left(nD_{km}^{(m)}\right) - m\cdot\psi(m+1)\right] \tag{40}$$

which under the alternatives, has the asymptotic Normal distribution, with asymptotic mean

$$\mu(g) = \frac{m^{1/2}}{2}\left(\int_0^1 l^2(u)\,du\right) \tag{41}$$

and asymptotic variance

$$\sigma^2(g) = -m(m+1)[m\cdot\psi^2(m+1) - 2m\cdot\psi(m+1)\cdot\psi(m+2) + m\cdot\psi^2(m+2) - \psi'(m+2)].$$

Here, the function $\psi(m) = \frac{\Gamma'(m)}{\Gamma(m)} = \mathbb{E}[\log(S_1)]$ is the digamma function, and its derivative

$$\psi'(m) = \sum_{n=0}^{\infty}(m+n)^{-2} = \mathrm{Var}[\log(S_1)]$$

is the trigamma function. Some notable constants are

$$\psi(1) = -\gamma = -0.57721\ldots = \mathbb{E}[\log(Z_1)]$$

which is the negative of the Euler–Mascheroni constant, and $\psi'(1) = \pi^2/6$.

It will be convenient to define the modified efficacy of a test as

$$e_M^2(\cdot) = \frac{e^2(\cdot)}{\left(\int_0^1 l^2(u)\,du\right)^2}.$$

The table lists, for various values of m, the modified efficacies of the higher-order Gini mean squared difference test, the Kullback–Leibler divergence based on higher-order spacings, and the Pitman's ARE of the Gini mean

squared difference test relative to the Kullback–Leibler divergence. It is seen that the Gini mean squared difference higher-order spacings test has the most efficacy and is more Pitman efficient. As expected from earlier literature, the efficacies of both tests increase with m for any finite m, but the latter test catches up with the former with increasing m.

Table 1. Modified Efficacies for the Gini Mean Squared Difference Higher-Order Spacings Test and the Kullback–Leibler Divergence Higher-order Spacings Test with Pitman ARE.

m	Gini Mean Squared Difference	Kullback–Leibler	Pitman ARE
1	1.000	0.862461	1.344377
2	1.500	1.3528	1.229463
3	2.000	1.84786	1.171445
4	2.500	2.34489	1.136672
5	3.000	2.84292	1.113559
10	5.500	5.33849	1.061423
20	10.500	10.336	1.031986
30	15.500	15.3351	1.021622

5. Conclusion

We derived the general asymptotic theory for U-statistics based on higher-order spacings and found the Asymptotically Locally Most Powerful test in this class under a sequence of close alternatives. The results established here extend those obtained in Tung and Jammalamadaka (2010) to the case of higher-order spacings.

References

Del Pino, G. E. (1979). On the asymptotic distribution of k-Spacings with applications to goodness-of-fit tests, *The Annals of Statistics* **7**, 5, pp. 1058–1065.

Holst, L. (1981). Some conditional limit theorems in exponential families, *The Annals of Probability* **9**, 5, pp. 818–830.

Jammalamadaka, S. R. and Goria, M. N. (2004). A test of goodness of fit based on Gini's index of spacings, *Statistics and Probability Letters* **68**, pp. 177–187.

Lee, A. J. (1990) *U-Statistics* (Marcel Dekker, Statistics: Textbooks and Monographs).

Rao, J. S. and Sethuraman, J. (1975). Weak convergence of empirical distribution functions of random variables subject to perturbations and scale factors, *The Annals of Statistics* **3**, 2, pp. 299–213.

Rao, J. S. and Sobel, M. (1980) Incomplete Dirichlet integrals with applications to ordered uniform spacings, *Journal of Multivariate Analysis* **10**, 4, pp. 603–610.

Sethuraman, J. and Rao, J. S. (1970). Pitman Efficiencies of Tests Based in Spacings, in M. L. Puri (ed.) *Nonparametric Techniques in Statistical Inference*, Cambridge University Press, pp. 405–415.

Sobel, M. and Wells, M. (1990). Dirichlet integrals and moments of gamma distribution order statistics, *Statistics and Probability Letters* **9**, pp. 431–437.

Tung, D. D. and Jammalamadaka, S. R. (2010). U-statistics based on spacings. Submitted for publication.