Annals of the Institute of Statistical Mathematics manuscript No.
(will be inserted by the editor)

Asymptotics of the Empirical Cross-over Function

Karthik Bharath · Vladimir Pozdnyakov · Dipak. K. Dey

Abstract We consider a combination of heavily trimmed sums and sample quantiles which arises when examining properties of clustering criteria and prove limit theorems. The object of interest, which we call the Empirical Cross-over Function, is an L-statistic whose weights do not comply with the requisite regularity conditions for usage of existing limit results. The law of large numbers, CLT and a functional CLT are proven.

Keywords Clustering · L-statistics · CLT · Functional CLT

PACS 62F05 · 62G30 · 62E20

1 Introduction

Suppose $W_1, W_2, \ldots, W_n$ for $n \geq 1$ are i.i.d random variables with distribution function F. If $W_{(1)} \leq W_{(2)} \leq \cdots \leq W_{(n)}$ are the order statistics, then, we define, for $0 < p < 1$, the *Empirical Cross-over Function* (ECF)

$$G_n(p) = \frac{1}{k}\sum_{j=1}^{k} W_{(j)} - W_{(k)} + \frac{1}{n-k}\sum_{j=k+1}^{n} W_{(j)} - W_{(k+1)} \qquad \text{for } \frac{k-1}{n} \leq p < \frac{k}{n}. \quad (1)$$

The function G_n is a special case of linear functions of order statistics $W_{(i)}$, $1 \leq i \leq n$, popularly referred to as L-statistics. L-statistics are usually represented as

$$L_n = \sum_{i=1}^{n} a_{i,n} W_{(i)}, \quad 1 \leq i \leq n, \quad (2)$$

where $a_{i,n}$ is a triangular array of constants, referred to as *weights*. A wide variety of limiting results on L-statistics have been derived over the years. We direct the interested

Karthik Bharath
Ohio State University; 1958 Neil Avenue; Columbus, OH-43210
E-mail: karthikbharath@gmail.com

Vladimir Pozdnyakov
University of Connecticut; 215 Glenbrook Road; Storrs, CT-06269

Dipak. K. Dey
University of Connecticut; 215 Glenbrook Road; Storrs, CT-06269

reader to Arnold et al. (2008) for a good source of results and relevant references. The asymptotic properties of these objects have been determined under suitable regularity conditions, albeit usually not too stringent, nevertheless disconcerting on occasions in practice. In this paper, we examine one such occasion, wherein we are faced with an L-statistic—the ECF—whose weights are not sufficiently smooth. As a consequence, asymptotic normality and a functional limit theorem do not follow readily.

Hartigan (1978), in his elegant paper derived asymptotic distributions of clustering criteria. He employed, what he referred to as the *split function*, in deriving the limiting results. The ECF, G_n, arises in a natural manner as the empirical counterpart of a certain functional of his split function when we are concerned with random variables having common invertible distribution function. The ECF is an interesting probabilistic object in its own right and being linear, offers an advantage over Hartigan's quadratic criterion function in terms of being amenable to extension to more interesting settings—namely clustering in higher dimensions and clustering of dependent observations.

The properties of the k-means clustering procedure for the univariate and the multivariate cases have been investigated extensively. Pollard (1981), Pollard (1982) proved strong consistency and asymptotic normality results in the univariate case. Serinko and Babu (1992) proved some weak limit theorem under non regular conditions for the univariate case. With the intention of having a more robust procedure for clustering, García-Escudero et al. (1999), Cuesta-Albertos et al. (1997) propose the trimmed k-means clustering and provide a central limit theorem for the multivariate case. In this paper, we prove consistency , a central limit theorem and also an invariance principle for our criterion function G_n, which is not in a form amenable for the usual representation of an L-statistic; nor, are its weights sufficiently smooth for the applicability of existing results.

2 Empirical Cross-over Function

In this section, we introduce the necessary constructs from clustering techniques from which we develop the ECF. Let $W_1, W_2, \ldots, W_n$ be i.i.d random variables with continuous cumulative distribution function F. We make the following assumptions.

*A*1. F is invertible for $0 < p < 1$ and absolutely continuous with density f.
*A*2. $E(W_1^2) < \infty$.
*A*3. For $0 < p < 1$, F is twice differentiable at $F^{-1}(p)$.

It is fairly common to encounter invertible distribution functions in applications. For example, models in finance possess strictly increasing distribution functions usually guaranteed by the additive 'sort of Gaussian' noise from the Ito integral component which smooths and removes both jumps and flat areas of the distribution function.

For $0 < p < 1$, consider the *split function* of F, as defined in Hartigan (1978),

$$B(F,p) = p\mu_l^2(p) + (1-p)\mu_u^2(p) - \left(\int_0^1 F^{-1}(q)dq\right)^2, \tag{3}$$

where

$$\mu_l(p) = \frac{1}{p}\int_{q\leq p} F^{-1}(q)dq = \frac{1}{p}\int_{-\infty}^{F^{-1}(p)} wdF,$$
$$\mu_u(p) = \frac{1}{1-p}\int_{q>p} F^{-1}(q)dq = \frac{1}{1-p}\int_{F^{-1}(p)}^{\infty} wdF.$$

Note that because of assumption $A1$ all the quantities are finite.

One way to think of $B(F,p)$ is, as the 'between cluster sum of squares', in the case where we are concerned with two clusters in one dimension. Therefore, the value of $p \in (0,1)$ maximizing this function, would determine the location at which data is split into two clusters. Let us denote that value as p_0 and p_0 is referred to as the *split point* in Hartigan (1978). As pointed out in Hartigan (1978), the conditions that guarantee the existence and uniqueness of the split point, are unclear. Determination of the requisite conditions, alone, is worthy of further investigation. However, for the purposes of this paper, those conditions and the split point itself are not important. When F is invertible, it is known that the split point p_0 solves

$$(\mu_u(p) - \mu_l(p))[\mu_u(p) + \mu_l(p) - 2F^{-1}(p)] = 0, \tag{4}$$

where the left side is the derivative of $B(F,p)$. Owing to the fact that $(\mu_u(p) - \mu_l(p)) > 0$ for all $0 < p < 1$, we are interested only in the zero of

$$G(p) = \mu_l(p) + \mu_u(p) - 2F^{-1}(p), \tag{5}$$

which we refer to as the *cross-over function*. The empirical version of the cross-over function represents the primary object of this paper. At this juncture, for better exposition, we recall the definition of the ECF; for $0 < p < 1$, we have

$$G_n(p) = \frac{1}{k}\sum_{j=1}^{k} W_{(j)} - W_{(k)} + \frac{1}{n-k}\sum_{j=k+1}^{n} W_{(j)} - W_{(k+1)} \tag{6}$$

for $\frac{k-1}{n} \leq p < \frac{k}{n}$ and

$$G_n(p) = \frac{1}{n}\sum_{j=1}^{n} W_{(j)} - W_{(n)}, \tag{7}$$

for $\frac{n-1}{n} \leq p < 1$, where $1 \leq k \leq n-1$.

Remark 1 Intuition about the ECF is useful here. The term 'cross-over' arises owing to the observation that

$$G_n\left(\frac{0}{n}\right) = W_{(1)} - W_{(1)} + \frac{1}{n-1}\sum_{j=2}^{n} W_{(j)} - W_{(2)} \quad \geq 0,$$
$$G_n\left(\frac{n-2}{n}\right) = \frac{1}{n-1}\sum_{j=1}^{n-1} W_{(j)} - W_{(n-1)} + W_{(n)} - W_{(n)} \quad \leq 0,$$

and the function crosses over 0 at some $1 \leq k \leq n-1$. If k^* is the index at which G_n crosses over, then $W_{(k^*)}$ represents the datum at which the data is split leading to the formation of two clusters. The term, $\frac{1}{k}\sum_{j=1}^{k} W_{(j)} - W_{(k)}$, can be thought of as a 'distance' between the mean of the first k observations, arranged in increasing order, and their maximum value; the term, $\frac{1}{n-k}\sum_{j=k+1}^{n} W_{(j)} - W_{(k+1)}$, represents the 'distance' between the mean of the last k observations and their minimum.

Remark 2 Let us again reiterate one important point here. The split function (3), cross-over function (5), and empirical empirical cross-over function (6) are designed for finding an optimal partition of data into *two* groups. The extension of this technique to the general case of three or more groups is discussed in Hartigan (1978). More specifically, if one needs to find an optimal partition of data into $K > 2$ clusters, instead of the split function (3) we need to introduce the following partition function:

$$B(F, p_1, \ldots, p_{K-1}) = \sum_{i=1}^{K} (p_i - p_{i-1})\mu_i^2 - \left(\int_0^1 F^{-1}(q)dq \right)^2,$$

where

$$\mu_i = \frac{1}{p_i - p_{i+1}} \int_{p_{i-1}}^{p_i} F^{-1}(q)dq,$$

and $0 = p_0 \leq p_1 \leq \cdots \leq p_{K-1} \leq p_K = 1$. Consequently, one needs to introduce $K - 1$ cross-over functions

$$\mu_{i+1} + \mu_i - 2F^{-1}(p_i),$$

and solve a system of $K - 1$ equations to find the optimal partition. We do not address the general case of $K > 2$ in this paper.

It is easy to see that for symmetrically distributed random variables 1/2 is a split point. Moreover, typically for light-tailed distributions (normal, uniform etc.) the split point is unique. As suggested in Hartigan (1978), this fact can be used to construct tests for the presence of clusters. However, finding conditions that guarantee uniqueness of the split point (or uniqueness of the solution of (4)) is a non-trivial task and is an open question. For instance, Hartigan (1978) gives an example of unimodal symmetric heavy-tailed distribution for which every p from $(0, 1)$ is a split point.

Remark 3 The function G_n is a linear combination of order statistics $W_{(i)}, 1 \leq i \leq n$ and hence an L-statistic. In the representation of an L-statistic L_n shown in (2), if the weights $a_{i,n}, 1 \leq i \leq n$, are of the form $\frac{1}{n} J\left(\frac{i}{n+1}\right)$, where $J(u), 0 < u < 1$, is the *weight function*, then it is possible to obtain an equivalent representation as

$$L_n = \frac{1}{n} \sum_{i=1}^{n} J\left(\frac{i}{n+1}\right) W_{(i)}.$$

The form of the weights $a_{i,n}$ represent the smoothness condition which guarantees asymptotic normality (See for e.g., Arnold et al. (2008), page 227 or Vaart (1998), page 318). Unfortunately, G_n cannot be represented in this form, since it has 'bad' weights, in the following sense; For $0 < p < 1$, we see that the order statistics $W_{(\lceil np \rceil)}$ and $W_{(\lceil np \rceil + 1)}$ have weights $\frac{1}{\lceil np \rceil} - 1$ and $\frac{1}{\lceil n(1-p) \rceil} - 1$, respectively. This clearly violates the smoothness condition rendering the usage of existing results inappropriate.

Remark 4 Observe that for a fixed $0 < p < 1$

$$\frac{1}{k} \sum_{j=1}^{k} W_{(j)} - W_{(k)} = \frac{1}{\lceil np \rceil} \sum_{j=1}^{\lceil np \rceil} W_{(j)} - W_{(\lceil np \rceil)},$$

$$\frac{1}{n-k} \sum_{j=k+1}^{n} W_{(j)} - W_{(k+1)} = \frac{1}{\lceil n(1-p) \rceil} \sum_{j=\lceil np \rceil + 1}^{n} W_{(j)} - W_{(\lceil np \rceil + 1)},$$

where $\lceil x \rceil$ represents the smallest integer not less than x. For a fixed $p \in (0,1)$, the sums shown above are *trimmed* sums. More precisely, since $\frac{\lceil np \rceil}{n} \to p$ and $\frac{\lceil n(1-p) \rceil}{n} \to 1-p$, they represent the case of *heavy* trimming; asymptotics for which are well known (see for e.g., Maller (1988) and Stigler (1973)). Unfortunately, the two order statistics, $W_{(\lceil np \rceil)}$ and $W_{(\lceil n(1-p) \rceil)}$, represent a formidable obstacle in the use of existing results for asymptotic normality of heavily trimmed sums. The function G_n is hence some sort of a combination of heavily trimmed sums and intermediate order statistics, and asymptotic results for such a combination, to our knowledge, are yet to developed.

3 Limit theorems for G_n

In this section, we prove the main results on the asymptotic behavior of the sample cross-over function G_n.

Theorem 1 *Under the assumptions A*1 *and A*2 *as* $n \to \infty$,

$$G_n(p) \xrightarrow{P} G(p).$$

Proof Because we only need to prove consistency for individual components of the ECF, it is a relatively easy exercise. However, for a purpose of completeness and in order to introduce notation and ideas that will be used in the proof of the subsequent theorem, we decided to provide a detailed proof of the law of large numbers for G_n.

For $0 < p < 1$, it is well known that $W_{(\lceil np \rceil)} \xrightarrow{P} F^{-1}(p)$ at points of continuity of F^{-1}. It is also the case that $W_{(\lceil np \rceil + 1)} \xrightarrow{P} F^{-1}(p)$, since the necessary condition for k_n-th order statistic $W_{(k_n)}$ to be consistent for $F^{-1}(p)$ is that $\frac{k_n}{n} \to p$ (see for instance, Vaart (1998)). Let us define

$$r_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{W_i < F^{-1}(p)},$$

where $\mathbb{I}_A$ is the indicator function of the set A. By the strong law of large numbers, $r_n \to p$ w.p.1. Now,

$$\frac{1}{k} \sum_{i=1}^{k} W_{(i)} = \frac{1}{\lceil np \rceil} \sum_{i=1}^{\lceil np \rceil} W_{(i)}.$$

Therefore,

$$\frac{1}{k} \sum_{i=1}^{k} W_{(i)} = \frac{1}{\lceil np \rceil} \left[\sum_{i=1}^{\lceil nr_n \rceil} W_{(i)} + \sum_{i=\lceil nr_n \rceil + 1}^{\lceil np \rceil} W_{(i)} \right].$$

It is clear here that if $\lceil nr_n \rceil + 1 > \lceil np \rceil$, the upper and lower limits of the second sum are interchanged with a negative sign.

The random sum

$$\begin{aligned} \frac{1}{\lceil np \rceil} \left| \sum_{i=\lceil nr_n \rceil + 1}^{\lceil np \rceil} W_{(i)} \right| &\leq \frac{1}{\lceil np \rceil} \sum_{i=\lceil nr_n \rceil + 1}^{\lceil np \rceil} |W_{(i)}| \\ &\leq \frac{1}{\lceil np \rceil} |\lceil np \rceil - \lceil nr_n \rceil| \left(|W_{(\lceil np \rceil)}| + |W_{(\lceil nr_n \rceil)}| \right). \end{aligned}$$

Recall that $r_n = p + O_p(n^{-1/2})$ and hence $|W_{\lceil np \rceil}|$ and $|W_{\lceil nr_n \rceil}|$ converge in probability to $F^{-1}(p)$(see Vaart (1998), page 308) and $|r_n - p| \xrightarrow{P} 0$. Consequently, we have that

$$\frac{1}{\lceil np \rceil} \sum_{i=\lceil nr_n \rceil + 1}^{\lceil np \rceil} W_{(i)} \xrightarrow{P} 0.$$

However,

$$\frac{1}{\lceil np \rceil} \sum_{i=1}^{\lceil nr_n \rceil} W_{(i)} = \frac{1}{\lceil np \rceil} \sum_{i=1}^{n} W_i \mathbb{I}_{W_i < F^{-1}(p)} \xrightarrow{P} \mu_l(p) \quad ,$$

by the law of large numbers of i.i.d random variables. As a consequence, $\frac{1}{k} \sum_{i=1}^{k} W_{(i)} \xrightarrow{P} \mu_l(p)$. In similar fashion we note that $\frac{1}{n-k} \sum_{i=k+1}^{n} W_{(i)} \xrightarrow{P} \mu_u(p)$ for all $0 < p < 1$ and $\frac{k-1}{n} \leq p < \frac{k}{n}$. Combining the above two convergences with the convergence of $W_{(\lceil np \rceil)}$ and $W_{(\lceil np \rceil + 1)}$ to their identical limits, we have that, $G_n(p) \xrightarrow{P} G(p)$ for each $0 < p < 1$.

Remark 5 It is worthwhile to note that the trimmed (at the random level) sum $\sum_{i=1}^{\lceil nr_n \rceil} W_{(i)}$ is exactly equal to the truncated sum $\sum_{i=1}^{n} W_i \mathbb{I}_{W_i < F^{-1}(p)}$, which is the sum of i.i.d. random random variables. This subtle relationship is greatly convenient in our proofs.

For ease of notation, let us define for $0 < p < 1$,

$$\begin{aligned}\theta_p =& \frac{1}{p} W_1 \mathbb{I}_{W_1 < F^{-1}(p)} - \frac{1}{p} F^{-1}(p) \mathbb{I}_{W_1 < F^{-1}(p)} \\ &+ \frac{1}{1-p} W_1 \mathbb{I}_{W_1 \geq F^{-1}(p)} - \frac{1}{1-p} F^{-1}(p) \mathbb{I}_{W_1 \geq F^{-1}(p)} \\ &+ \frac{2 \mathbb{I}_{W_1 < F^{-1}(p)}}{f(F^{-1}(p))},\end{aligned}$$

and $U_n(p) = \sqrt{n}(G_n(p) - G(p))$ for $0 < a \leq p \leq b < 1$.

Theorem 2 *Under assumptions $A1 - A3$ as $n \to \infty$,*

$$\sqrt{n} \left(G_n(p) - G(p) \right) \xrightarrow{d} N\left(0, \sigma_p\right),$$

where $\sigma_p = Var(\theta_p)$. Furthermore,

$$U_n(p) \Rightarrow U(p),$$

in the Skorohod space $D[a,b]$, $0 < a < b < 1$ equipped with the J_1 topology, where $U(p)$ is a Gaussian process with mean 0 and covariance given by

$$Cov(U(p), U(q)) = Cov(\theta_p, \theta_q).$$

Proof The trick used in proving the asymptotic normality of G_n is to consider mean-zero asymptotics of its individual components and by the use of *Bahadur's representation* for sample quantiles, rewrite G_n as a sum of i.i.d random variables and an error term, which goes to zero at an appropriate rate. This would then pave the way for the usage of standard results.

More specifically, first note that for $0 < p < 1$, and each $i = 1, \ldots, n$,

$$E\left(W_i \mathbb{I}_{W_i < F^{-1}(p)}\right) = p\mu_l(p)$$

and

$$E\left(W_i \mathbb{I}_{W_i \geq F^{-1}(p)}\right) = (1-p)\mu_u(p).$$

Observe that, for $\frac{k-1}{n} \leq p < \frac{k}{n}$,

$$\begin{aligned}
\sqrt{n}\left[\frac{1}{k}\sum_{i=1}^{k} W_{(i)} - np\mu_l(p)\right] &= \frac{\sqrt{n}}{\lceil np \rceil}\left[\sum_{i=1}^{\lceil np \rceil} W_{(i)} - np\mu_l(p)\right] \\
&= \frac{\sqrt{n}}{\lceil np \rceil}\left[\sum_{i=1}^{\lceil nr_n \rceil} W_{(i)} + \sum_{i=\lceil nr_n \rceil + 1}^{\lceil np \rceil} W_{(i)} - np\mu_l(p)\right] \\
&= \frac{\sqrt{n}}{\lceil np \rceil}\left[\sum_{i=1}^{\lceil nr_n \rceil} W_{(i)} - np\mu_l(p)\right] + \frac{\sqrt{n}}{\lceil np \rceil}\left[\sum_{i=\lceil nr_n \rceil + 1}^{\lceil np \rceil} \left(W_{(i)} - F^{-1}(p)\right)\right] \\
&\quad + \frac{\sqrt{n}}{\lceil np \rceil} F^{-1}(p)\left(\lceil np \rceil - \lceil nr_n \rceil\right).
\end{aligned}$$

Now note that

$$\begin{aligned}
&\frac{\sqrt{n}}{\lceil np \rceil}\left|\sum_{i=\lceil nr_n \rceil + 1}^{\lceil np \rceil} W_{(i)} - F^{-1}(p)\right| \\
&\qquad \leq \frac{\sqrt{n}}{\lceil np \rceil}\left|\lceil np \rceil - \lceil nr_n \rceil\right| \max\left(\left|W_{(\lceil np \rceil)} - F^{-1}(p)\right|, \left|W_{(\lceil nr_n \rceil)} - F^{-1}(p)\right|\right).
\end{aligned}$$

By the same argument used in the proof of Theorem 1, $|W_{(\lceil np \rceil)} - F^{-1}(p)| \xrightarrow{P} 0$ and $|W_{(\lceil nr_n \rceil)} - F^{-1}(p)| \xrightarrow{P} 0$. By the central limit theorem for i.i.d random variables, $\sqrt{n}|p - r_n|$ is asymptotically normal and hence bounded in probability. Consequently,

$$\frac{\sqrt{n}}{\lceil np \rceil}\left|\sum_{i=\lceil nr_n \rceil + 1}^{\lceil np \rceil} W_{(i)} - F^{-1}(p)\right| \xrightarrow{P} 0 \quad .$$

Next, recall that

$$\begin{aligned}
\sum_{i=1}^{\lceil nr_n \rceil} W_{(i)} &= \sum_{i=1}^{n} W_i \mathbb{I}_{W_i < F^{-1}(p)}, \\
nr_n &= \sum_{i=1}^{n} \mathbb{I}_{W_i < F^{-1}(p)}.
\end{aligned}$$

Therefore,

$$\sqrt{n}\left[\frac{1}{\lceil np\rceil}\sum_{i=1}^{\lceil np\rceil} W_{(i)} - np\mu_l(p)\right] = \frac{1}{p}\Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(W_i\mathbb{I}_{W_i<F^{-1}(p)} - p\mu_l(p)) + \frac{1}{\sqrt{n}}F^{-1}(p)\sum_{i=1}^{n}(p - \mathbb{I}_{W_i<F^{-1}(p)})\Big] + o_p(1) = \sqrt{n}\bar{\xi} + o_p(1),$$

where

$$\xi_i = \frac{1}{p}\left[W_i\mathbb{I}_{W_i<F^{-1}(p)} - F^{-1}(p)\mathbb{I}_{W_i<F^{-1}(p)} - (p\mu_l(p) - pF^{-1}(p))\right]$$

are i.i.d random variables for $i = 1, \cdots, n$ and $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$.

Using a similar argument, we can claim that

$$\sqrt{n}\left[\frac{1}{\lceil n(1-p)\rceil}\sum_{i=\lceil np\rceil+1}^{n} W_{(i)} - n(1-p)\mu_u(p)\right] = \sqrt{n}\bar{\tau} + o_p(1),$$

where

$$\tau_i = \frac{1}{1-p}\left[W_i\mathbb{I}_{W_i\geq F^{-1}(p)} - F^{-1}(p)\mathbb{I}_{W_i\geq F^{-1}(p)} - ((1-p)\mu_u(p) - (1-p)F^{-1}(p))\right]$$

are i.i.d random variables and $\bar{\tau} = \frac{1}{n}\sum_{i=1}^{n}\tau_i$. That takes care of the two trimmed sums.

Next, we turn our attention to the two quantiles $W_{(k)}$ and $W_{(k+1)}$ or equivalently $W_{(\lceil np\rceil)}$ and $W_{(\lceil np\rceil+1)}$ for $\frac{k-1}{n} \leq p < \frac{k}{n}$. Using the Bahadur representation for sample quantiles (see Bahadur (1966)), justified by assumptions $A1$ and $A3$, we have

$$\sqrt{n}\left(W_{(\lceil np\rceil)} - F^{-1}(p)\right) = \sqrt{n}\left(W_{(\lceil np\rceil+1)} - F^{-1}(p)\right) = \sqrt{n}\bar{\kappa} + o_p(1),$$

where $\bar{\kappa} = \frac{1}{n}\sum_{i=1}^{n}\kappa_i$, and

$$\kappa_i = \frac{p - \mathbb{I}_{W_i<F^{-1}(p)}}{f(F^{-1}(p))}$$

are i.i.d random variables.

We are now in a situation where for $0 < p < 1$, $\sqrt{n}(G_n(p) - G(p))$ has been expressed as sums of i.i.d random variables along with an error term which is $o_p(1)$. That is,

$$\sqrt{n}\left(G_n(p) - G(p)\right) = \sum_{i=1}^{n}\frac{Z_i}{\sqrt{n}} + o_p(1),$$

where $Z_i = \xi_i + \tau_i - 2\kappa_i$ are i.i.d random variables. The advantage of this representation lies in the fact that we are now allowed to examine G_n without having to concern ourselves with the correlations between its individual components. The representation ensures that the effect of the correlations is of order as that of the error term or smaller

and can hence be safely disregarded. Consequently, by the central limit theorem for i.i.d random variables

$$\sqrt{n}\left(G_n(p) - G(p)\right) \xrightarrow{d} N\left(0, \sigma_p\right).$$

We can now turn our attention to the functional limit of the process U_n. Since we are interested in the behavior of G_n for $0 < p < 1$ and in particular the point at which it crosses zero, we restrict ourselves to examining the behavior of U_n in the closed interval $[a, b]$ where a and b are constants bounded away from 0 and 1 respectively. Notice that U_n is a natural random element of the Skorohod space $D[a, b]$. It is straightforward to note that by virtue of our representation of $\sqrt{n}(G_n(p) - G(p))$, for each p, as *a sum of i.i.d. random variables plus an error term of order* $o_p(1)$, by the central limit theorem for random vectors, we have that

$$\sqrt{n}\left(U_n(p_1) - U(p_1), \cdots, U_n(p_k) - U(p_k)\right) \xrightarrow{d} N\left(\mathbf{0}, \Sigma\right),$$

where k is a finite positive integer and for $i, j = 1, \ldots, k$, $\Sigma = \left(\sigma_{ij}\right)$ with

$$\sigma_{ij} = \begin{cases} Var(\theta_{p_i}) & \text{if } i = j \\ Cov(\theta_{p_i}\theta_{p_j}) & \text{if } i \neq j. \end{cases}$$

Now, if we can show that the sequence U_n is tight, we then have the required convergence to U (see Billingsley (1968), for the necessary arguments). We set about proving tightness in an indirect way, as opposed to the usual method of showing that U_n concentrates on a compact set in $D[a, b]$ with high probability. Consider the components of U_n

$$\begin{aligned}
U_1^n &= \sqrt{n}\left(\frac{1}{\lceil np \rceil}\sum_{i=1}^{\lceil np \rceil} W_{(i)} - \frac{1}{p}\int_0^p F^{-1}(q)dq\right), \\
U_2^n &= \sqrt{n}\left(W_{(\lceil np \rceil)} - F^{-1}(p)\right), \\
U_3^n &= \sqrt{n}\left(\frac{1}{\lceil n(1-p) \rceil}\sum_{i=\lceil np \rceil + 1}^{n} W_{(i)} - \frac{1}{1-p}\int_p^1 F^{-1}(q)dq\right), \\
U_4^n &= \sqrt{n}\left(W_{(\lceil np \rceil + 1)} - F^{-1}(p)\right).
\end{aligned}$$

It is interesting that for every U_i^n for $i = 1, \ldots, 4$ the functional CLT is an established result. However, the weak convergence of the individual components U_i^n does not automatically guarantee weak convergence for the sum of the components. But at this point we need only the tightness. Since the sum of compact sets is a compact set again, it is easy to show that if each component is tight then it is indeed true that the sum is tight with respect to the Skorohod metric on $D[a, b]$. Now, note that U_2^n and U_4^n are quantile processes and converge weakly to a Gaussian process (see p. 308, Vaart (1998)) in $D[a, b]$. Using the result from Kasahara and Maejima (1992), we can claim that U_1^n and U_3^n also converge weakly to a limit process in $D[a, b]$. This proves that each U_i^n is relatively compact for each i. Now, since $D[a, b]$ is complete and separable with respect to the Skorohod metric (see p.115, Billingsley (1968)), using the converse of Prohorov's theorem (see p.37 Billingsley (1968)) we can claim that each U_i^n for $i = 1, \ldots, 4$ is tight and, therefore, $U_n = U_1^n + U_2^n + U_3^n + U_4^n$ is tight in $D[a, b]$ equipped with the J_1-topology.

We now provide verification of our asymptotic results regarding consistency and asymptotic normality by considering two examples. In both the examples we first generate 1000 random variables $T_n = \sqrt{n}\,(G_n(0.5) - G(0.5))$ and obtain the simulated mean and the variance. In order to verify asymptotic normality, we generate again 100 random variables T_n. This is done for different samples sizes n and results are tabulated.

Example 1 If $W_1, W_2, \ldots, W_n$ are i.i.d $N(0,1)$, then it can be ascertained quite easily that $G(0.5) = 0$ and $\sigma = 2\pi - 4 \approx 2.2831$. The numbers tabulated below offer satisfactory evidence about the accuracy of our results.

Example 2 In this example, we consider $W_1, W_2, \ldots, W_n$ to be i.i.d. exponential random variables with mean 1. This represents the archetypal case of a skewed distribution and we again check for the accuracy of our results. In this case, $G(0.5) = 2(1 - \ln 2) \approx 0.6137$ and $\sigma = 8(1 - \ln 2) \approx 2.4548$. The numbers in the tables below provide further corroborative evidence for our limiting results.

Table 1 Simulated means and variances for different sample sizes.

Random variables	$N(0,1)$			$Exp(1)$		
Sample sizes	$n = 100$	$n = 1000$	$n = 10000$	$n = 100$	$n = 1000$	$n = 10000$
Simulated Mean	-0.017	0.018	0.002	-0.041	-0.014	0.0019
Simulated Variance	2.407	2.324	2.296	2.491	2.463	2.452

Table 2 p-values for Kolmogorov-Smirnov test for normality

Random variables	$N(0,1)$	$Exp(1)$
$n = 100$	0.751	0.8786
$n = 1000$	0.12	0.2174
$n = 10000$	0.391	0.9955

Example 3 In this example, for the density $.25\mathcal{N}(-2,1) + .75\mathcal{N}(2,1)$, we will now illustrate how, using Theorem 2, one can construct a Confidence Interval (CI) for $G(p_0)$: the value of the cross-over function at the true split point p_0; indeed, this effectively implies a CI for 0 since $G(p_0) = 0$. We first plot the density and the cross-over function. The plot is in Figure 1.

Upon solving $G(p) = 0$ numerically, we obtain $p_0 \approx 0.265$. Now, suppose we have $W_1, \ldots, W_n$ i.i.d. from $.25N(-2,1) + .75N(2,1)$. A straightforward calculation shows that the variance, σ_{p_0}, of $\sqrt{n}(G_n(p_0) - G(p_0)) = \sqrt{n}G_n(p_0)$ explicitly depends on the following quantities: p_0, $F^{-1}(p_0)$, $f(F^{-1}(p_0))$, $\mu_l(p_0)$, $\mu_u(p_0)$,

$$\frac{1}{p_0}E[W_1^2\mathbb{I}_{W_1<F^{-1}(p_0)}] \text{ and } \frac{1}{1-p_0}E[W_1^2\mathbb{I}_{W_1\geq F^{-1}(p_0)}].$$

These quantities computed numerically yield $\sigma_{p_0} \approx 120.1$. Asymptotic normality of the deviation of $G_n(p_0)$ from 0 provides us with an approximate $100(1-\alpha)\%$ CI for $G(p_0) = 0$, for $0 < \alpha < 1$. We simulate 1000 $\sqrt{n}G_n(p_0)$ random variables for different

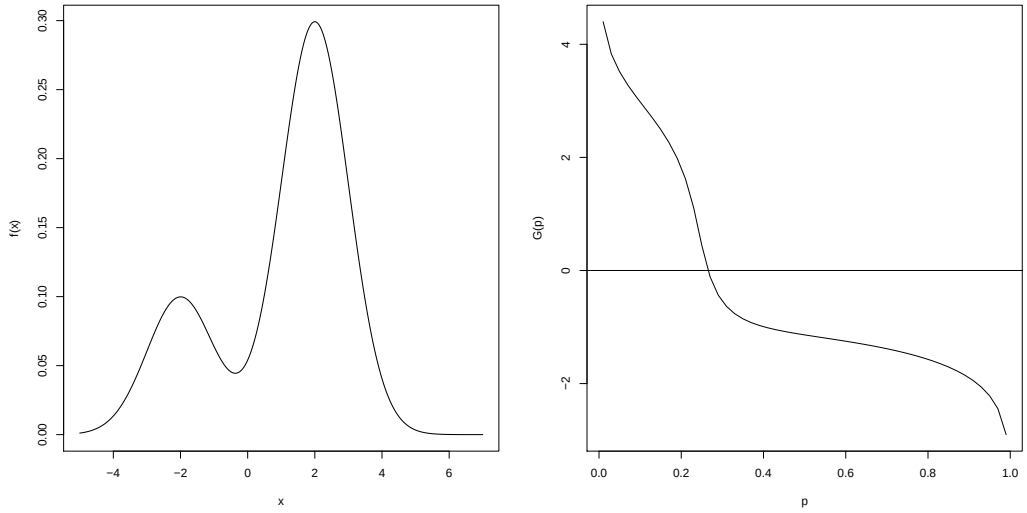

Fig. 1 Density of $.25N(-2,1) + .75N(2,1)$ and the corresponding Cross-over function $G(p)$

samples sizes n, construct the 95% CI and check the proportion amongst them which contain $G(p_0) = 0$. The results tabulated below offer satisfactory evidence regarding the validity of our results.

Sample size n	**Proportion containing** $G(p_0)$
100	0.988
1000	0.953
10000	0.949

4 Concluding Remarks

Despite being an L-statistic, the asymptotic properties of the ECF cannot be studied using existing machinery owing to the fact that its weights are not smooth. Asymptotic results from heavily trimmed sums are inapplicable to our problem due the presence of the two order statistics, $W_{(k)}$ and $W_{(k+1)}$, with unfriendly weights. The centered ECF, however, can be expressed as a sum of i.i.d random variables and an error term, which goes to zero at an appropriate rate, by the use of a subtle trick involving truncated sums and the Bahadur representation for sample quantiles. Owing to this, the CLT follows immediately and what remains is to show that the centered process satisfies the tightness condition for the functional CLT.

Note that the ECF is invariant with respect to shift in the distribution of W's, but is linear with respect to scaling. If we introduce statistic p_n (the empirical split point) that 'solves', in some appropriate sense, the equation

$$G_n(p) = 0,$$

then this statistic is invariant with respect to both shifting and scaling (as it should be, because the clustering problem is invariant with respect to linear transformations), and potentially can be used to design a clustering test.

The asymptotics of p_n is the next natural question, which is the focus of the work in Bharath et al. (2013). The Central Limit Theorem for G_n, proved in this paper, constitutes a very important step towards the solution of determining the asymptotic behavior of p_n. According to a general plan outlined in Serfling (1980) p. 95, we can conjecture that

$$p_n \approx p_0 - G_n(p_0)/G'(p_0),$$

where p_0 is a theoretical split point. However, the rigorous proof of this statement requires significant efforts.

On a slightly different note, from a theoretical perspective, investigation into the second-order asymptotic properties of the ECF would perhaps offer an improvement in the rates of convergence of the ECF to the cross-over function. In a recent paper Gribkova and Helmers (2006) establish the validity of the Edgeworth expansion for a studentized heavily trimmed-mean under no assumptions on the distribution function F. Their results, for instance, could be used to refine our asymptotic results concerning the ECF. Furthermore, one could conceivably consider an ECF defined using lightly trimmed sums (intermediate trimming) and intermediate sample quantiles as opposed to the central ones. This allows for more flexibility in developing a sample counterpart to the crossover function since one need not necessarily employ only the sample quantiles–for instance, $W_{(\lceil np \rceil)}$ could be replaced with $W_{(np_n)}$ where $p_n/n \to 1$ or to 0. In this regard, results in Gribkova and Helmers (2012) would be particularly useful in establishing refined second-order asymptotic results.

References

Barry C Arnold, N Balakrishnan, and H N Nagaraja. *A First Course in Order Statistics*. Society of Industrial and Applied Mathematics (SIAM), Philadelphia, 2008.

R R Bahadur. A Note on Quantiles in Large Samples. *Ann.Math.Statist.*, 38:577–580, 1966.

K Bharath, V Pozdnyakov, and D K Dey. Asymptotics of a Clustering Criterion for Smooth Distributions. *Forthcoming in Electronic Journal of Statistics. arxiv:1205.2123v1*, 2013.

P Billingsley. *Convergence of Probability Measures*. John Wiley and Sons, New York, 1968.

J A Cuesta-Albertos, A Gordaliza, and C Matrán. Trimmed k-means: An Attempt to robustify Quantizers. *Ann.Statist.*, 25:553–576, 1997.

L A García-Escudero, A Gordaliza, and C Matrán. A Central Limit Theorem for Multivariate Generalized Trimmed k-means. *Ann.Statist.*, 27:1061–1079, 1999.

N Gribkova and R Helmers. The Empirical Edgeworth Expansion for a Studentized Trimmed Mean. *Math.Methods Statist*, 15:61–87, 2006.

N Gribkova and R Helmers. On a Bahadur-Kiefer Representation of Von Mises Statistics Type of Intermediate Sample Quantiles. *Probablity and Mathematical Statistics*, 32:255–279, 2012.

J Hartigan. Asymptotic Distributions for Clustering Criteria. *Ann.Statist.*, 6:117–131, 1978.

Y Kasahara and M Maejima. Limit Theorems for Trimmed Sums . *Journal of Theoretical Probability*, 5:617–628, 1992.

R A Maller. Asymptotic Normality of Trimmed Sums in Higher Dimensions. *Ann.Probab.*, 16:1608–1622, 1988.

David Pollard. Strong Consistency for K-Means Clustering. *Ann.Statist.*, 9:135–140, 1981.

David Pollard. A Central Limit Theorem for k-means Clustering. *Ann.Statist.*, 10:919–926, 1982.

R Serfling. *Approximation Theorems for Mathematical Statistics*. John Wiley, New york, 1980.

R J Serinko and G J Babu. Weak Limit Theorems for Univariate k-means Clustering under Nonregular Conditions. *J. Multivariate Anal.*, 49:188–203, 1992.

S M Stigler. The Asymptotic Distribution of the Trimmed Mean. *Ann.Statist.*, 1:472–477, 1973.

A W Van Der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK, 1998.