Electronic Journal of Statistics
Vol. 7 (2013) 1078–1093
ISSN: 1935-7524
DOI: 10.1214/13-EJS801

Asymptotics of a clustering criterion for smooth distributions

Karthik Bharath[*], **Vladimir Pozdnyakov**[†] **and Dipak K. Dey**[†]

Abstract: We develop a clustering framework for observations from a population with a smooth probability distribution function and derive its asymptotic properties. A clustering criterion based on a linear combination of order statistics is proposed. The asymptotic behavior of the point at which the observations are split into two clusters is examined. The results obtained can then be utilized to construct an interval estimate of the point which splits the data and develop tests for bimodality and presence of clusters.

AMS 2000 subject classifications: Primary 62F05, 62G30; secondary 60F17, 62M02.
Keywords and phrases: Clustering, trimmed means, CLT.

Received February 2013.

1. Introduction

In this article, we develop a general framework for univariate clustering based on the ideas in Hartigan (1978) for the case of observations from a population with smooth and invertible distribution function. Contrary to Hartigan's approach, which was based on a quadratic function of the observed data, our clustering criterion function possesses the advantage of being a linear combination of order statistics—in fact, it is a combination of trimmed sums and sample quantiles.

It is common in certain applications to assume that the data are taken from a population with smooth distribution function. One important example is modeling in continuous-time mathematical finance, wherein observations are typically increments from a continuous-time stochastic process, and therefore, have smooth distributions because of presence of Itô integral components. Keeping this in mind, we deviate from the Hartigan's framework and concentrate our attention on a function of the derivative of his *split function*. This approach permits us to obviate the existence of a finite fourth moment assumption imposed by Hartigan in the asymptotic investigation of his criterion function—a second moment assumption at the cost of an additional smoothness condition on our criterion function suffices. As an added benefit, this modification of Hartigan's approach provides us with the genuine possibility of extending our existing theory to more interesting scenarios involving dependent observations.

The notion of a "cluster" has several reasonable mathematical definitions. As in Hartigan (1978), we adopt a definition based on determining a point which

[†] University of Connecticut
[*] Ohio State University

splits the data into clusters via maximizing the between cluster sums of squares. The main results in this article involve the asymptotic behavior of this particular point. The theoretical properties of k-means clustering procedure for the univariate and the multivariate cases have been extensively investigated. Pollard (1981), Pollard (1982) proved strong consistency and asymptotic normality results in the univariate case. Serinko and Babu (1992) proved some weak limit theorems under non-regular conditions for the univariate case. With the intention of having a more robust procedure for clustering, García-Escudero, Gordaliza and Matrán (1999), and Cuesta-Albertos, Gordaliza and Matrán (1997) propose the trimmed k-means clustering and provide a central limit theorem for the multivariate case. Throughout the article we are primarily concerned with the case $k = 2$ on the real line. For extension of the split point approach to the case $k > 2$ we refer readers to the discussion in Hartigan (1978).

On a more practical note, our results enable us to construct an interval estimate of the point at which the data splits; this naturally allows us to develop simple tests for bimodality and presence (or absence) of clusters. Hypothesis tests for the presence (or absence) of clusters in a dataset has attracted considerable interest over the years. One of the earliest work in this area was by Engleman and Hartigan (1969); they developed a univariate method to test the null hypothesis of normally distributed cluster against the alternative of a two-component mixture of normals. Wolfe (1970) extended the work of Engleman and Hartigan (1969) to the multivariate normal setup using MLE techniques and applied his method to Fisher's Iris data. Motivated by applications in market segmentation, Arnold (1979) proposed a test for clusters based on examining the within-groups scatter matrix. A dataset generated from a large-scale survey of lifestyle statements was considered and the objective was to capture heterogeneity in the distribution of the responses to an appropriate questionnaire. Based on statistics concerning mean distances, minimum-within clusters sums and the resulting F-statistics, Bock (1985) presented several significance tests for clusters. In fact, he generalized some of the results in Hartigan (1978) to the multivariate setup. In similar spirit, we note that in our method, the point which splits the data is invariant to scaling and translation of the data. This permits us to examine the behavior of the point under the null hypothesis of "no cluster" and thereby construct a suitable test.

In Hartigan and Hartigan (1985), a popular test for unimodality, referred to as Dip test, was proposed and was applied to a dataset pertaining to the quality of 63 statistics departments. Indeed, their test did not possess good power against the specific bimodal alternative. More recently, Holzmann and Vollmer (2008) proposed a parametric test for bimodality based on the likelihood principle by using two-component mixtures. Their method was applied to investigate the modal structure of the cross-sectional distribution of per-capita log GDP across EU regions. Using the Kolmogorov-Smirnov and the Anderson-Darling statistics, Schwab, Podsiadlowski and Rappaport (2012), performed a test for bimodality in the distribution of neutron-star masses. They compared the empirical cumulative distribution function to the distribution functions of a unimodal normal and a bimodal two-component normal mixture. Our results enable us to construct, on identical lines as the test for clusters, a test for bimodality. The test statistic, again, is based on the point at which the data is split—the split point is the same for all unimodal distributions with finite second moment and can hence be used as the test statistic.

In section 2 we introduce the relevant constructs of our clustering framework: a theoretical criterion function and its zero followed by the empirical criterion function and its "zero". These quantities are of chief interest in this article. In section 3, we prove limit theorems for the empirical zero by examining the asymptotic behavior of the empirical criterion function and offer numerical verification of the limit results via simulation. Furthermore, we demonstrate the utility of our results on the popular `faithful` dataset pertaining to eruption times for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. Finally, in section 4 we highlight the salient features of our approach, note its shortcomings and comment on possible remedies and extensions.

2. Clustering criterion

In this section, based on Hartigan's approach, we propose an alternative clustering criterion and examine its properties. We first state our assumptions for the rest of the paper.

2.1. Assumptions

Let $W_1, W_2, \dots, W_n$ be i.i.d. random variables with cumulative distribution function F. We denote by Q the quantile function associated with F. We make the following assumptions:

$A1$. F is invertible for $0 < p < 1$ and absolutely continuous with respect to Lebesgue measure with density f.
$A2$. $E(W_1) = 0$ and $E(W_1^2) = 1$.
$A3$. Q is twice continuously differentiable at any $0 < p < 1$.

Note that owing to assumption $A1$, the quantile function Q is the regular inverse of F and not the generalized inverse.

2.2. Empirical cross-over function and empirical split point

Let us first consider the *split function* that was introduced in Hartigan (1978) for partitioning a sample into two groups. The *split function* of Q at $p \in (0, 1)$ is defined as

$$B(Q, p) = p(Q_l(p))^2 + (1-p)(Q_u(p))^2 - \left(\int_0^1 Q(q)dq\right)^2, \tag{2.1}$$

where

$$Q_l(p) = \frac{1}{p}\int_{q<p} Q(q)dq = \frac{1}{p}E[W_1 \mathbb{I}_{W_1 < Q(p)}]$$

and

$$Q_u(p) = \frac{1}{1-p}\int_{q\geq p} Q(q)dq = \frac{1}{1-p}E[W_1\mathbb{I}_{W_1\geq Q(p)}]$$

represent the conditional expectations of the random variables W_i up to and from $Q(p)$. Here $\mathbb{I}_A$ denotes the indicator function of a set A. In our case since $EW_1 = 0$ the last term in the definition of the split function is 0. A value p_0 which maximizes the split function is called the *split point*. It is seen that if Q is the regular inverse, as in our case, p_0 satisfies the equation

$$(Q_u(p_0) - Q_l(p_0))[Q_u(p_0) + Q_l(p_0) - 2Q(p_0)] = 0, \tag{2.2}$$

where the LHS is the derivative of $B(Q,p)$. Evidently, $(Q_u(p) - Q_l(p)) > 0$ for all $0 < p < 1$ and we hence, for our purposes, consider the *cross-over function*,

$$G(p) = Q_l(p) + Q_u(p) - 2Q(p), \tag{2.3}$$

for examining clustering properties. From a statistical perspective, we would like to work with the empirical version of (2.3). We deviate here from Hartigan's framework and consider the *empirical cross-over function*(ECF), defined in Bharath, Pozdnyakov and Dey (2012) as

$$G_n(p) = \frac{1}{k}\sum_{j=1}^{k} W_{(j)} - W_{(k)} + \frac{1}{n-k}\sum_{j=k+1}^{n} W_{(j)} - W_{(k+1)}, \tag{2.4}$$

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

$$G_n(p) = \frac{1}{n}\sum_{j=1}^{n} W_{(j)} - W_{(n)}, \tag{2.5}$$

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

The random quantity G_n, represents the empirical version of (2.3) and determines the split point for the given data. G_n is an L-statistic with irregular weights and hence not amenable for direct application of existing asymptotic results for L-statistics. Observe that

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

This simple observation captures the typical behavior of the empirical cross-over function. It starts positive and then at some point crosses the zero line. The index k at which this change occurs determines the datum $W_{(k)}$ at which the split occurs. In Bharath, Pozdnyakov and Dey (2012), it is shown that $G_n(p)$ is a consistent estimator of $G(p)$ for each $0 < p < 1$ and also that $\sqrt{n}(G_n(p) - G(p))$ is asymptotically normal.

We now introduce the *empirical split point in range* $[a,b]$, $0<a<b<1$, the empirical counterpart of the p_0 as

$$p_n = p_n(a,b) := \begin{cases} 0, \text{ if } G_n\left(\frac{k-1}{n}\right) < 0 \ \forall k \text{ such that } na < k < nb+1; \\ 1, \text{ if } G_n\left(\frac{k-1}{n}\right) > 0 \ \forall k \text{ such that } na < k < nb+1; \\ \frac{1}{n}\left[\max\{na<k<nb : G_n\left(\frac{k-1}{n}\right) G_n\left(\frac{k}{n}\right) \le 0\}\right], \text{ otherwise.} \end{cases}$$

The quantity p_n is our estimator of p_0, the true split point (when it is in the range). If p_n is equal to 0 or 1, we declare that the split point is outside the range. The asymptotic behavior of p_n can be used for the construction of test for the presence of clusters in the observations, or for the estimation of the true split point.

Remark 1. Let us provide some intuition behind Hartigan's split function. The k-means clustering method for the case $k=2$ requires us to minimize (with respect to k^*) the following within group sum of squares:

$$\begin{aligned} W^* &= \sum_{i=1}^{k^*}\left(W_{(i)} - \frac{1}{k^*}\sum_{i=1}^{k^*} W_{(i)}\right)^2 + \sum_{i=k^*+1}^{n}\left(W_{(i)} - \frac{1}{n-k^*}\sum_{i=k^*+1}^{n} W_{(i)}\right)^2 \\ &= \sum_{i=1}^{n} W_{(i)}^2 - \frac{1}{k^*}\left(\sum_{i=1}^{k^*} W_{(i)}\right)^2 - \frac{1}{n-k^*}\left(\sum_{i=k^*+1}^{n} W_{(i)}\right)^2. \end{aligned}$$

That is, minimizing W^* is equivalent to maximizing

$$\frac{1}{k^*}\left(\sum_{i=1}^{k^*} W_{(i)}\right)^2 + \frac{1}{n-k^*}\left(\sum_{i=k^*+1}^{n} W_{(i)}\right)^2$$

or

$$\frac{k^*}{n}\left(\frac{1}{k^*}\sum_{i=1}^{k^*} W_{(i)}\right)^2 + \frac{n-k^*}{n}\left(\frac{1}{n-k^*}\sum_{i=k^*+1}^{n} W_{(i)}\right)^2,$$

which is basically an empirical version of Hartigan's split function (2.1) and k^*/n will be another version of empirical split point.

In this paper we proceed in parallel with Hartigan (1978) (his Theorem 1 and Theorem 2) and prove consistency and asymptotic normality of p_n under a uniqueness assumption. The theoretical conditions that guarantee the uniqueness of the split point is an open question. It is easy to see that for a unimodal symmetric distribution with a finite second moment, the split point is $1/2$, and for all unimodal symmetric light-tailed distributions that we checked the split point was unique. However, Hartigan (1978) gives an example of a unimodal symmetric heavy-tailed distribution for which every point in $(0,1)$ is a split point. For a bimodal distribution the split point p_0 is typically unique and $Q(p_0)$ lies between the cluster means.

The presented results can be employed for testing "no-clusters" hypothesis, testing bimodality, and estimation of the split point. The extension of this technique to the case of k clusters is discussed in Hartigan (1978). In our case, instead of one cross-over function one needs to introduce $k-1$ functions; the split point in this case will be a $(k-1)$-dimensional vector. To find this split point we then need to solve a system of $k-1$ equations. For instance, for partition of data into three groups we need to introduce two cross-over functions

$$G_{1n}\left[\frac{k_1-1}{n},\frac{k_2-1}{n}\right]=\frac{1}{k_1}\sum_{i=1}^{k_1}W_{(i)}-W_{(k_1)}+\frac{1}{k_2-k_1}\sum_{i=k_1+1}^{k_2}W_{(i)}-W_{(k_1+1)},$$

$$G_{2n}\left[\frac{k_1-1}{n},\frac{k_2-1}{n}\right]=\frac{1}{k_2-k_1}\sum_{i=k_1+1}^{k_2}W_{(i)}-W_{(k_2)}+\frac{1}{n-k_2}\sum_{i=k_2+1}^{n}W_{(i)}-W_{(k_2+1)},$$

and, respectively, one needs to solve (in an appropriate sense) the following system of equations:

$$G_{1n}\left[\frac{k_1-1}{n},\frac{k_2-1}{n}\right]=0,$$
$$G_{2n}\left[\frac{k_1-1}{n},\frac{k_2-1}{n}\right]=0.$$

We do not address the general $k>2$ cluster situation in this paper.

Finally, as stated in the Introduction, let us remark on the main technical difference between Hartigan's assumptions and ours. Since we deal here with the derivative of the split function, we need a *stronger* smoothness condition (the second derivative of G instead of the first one). In return, we work with trimmed means and a *weaker* moment condition suffices (the finite second moment instead of the fourth one as in Hartigan (1978)).

Remark 2. Notice that if for constants $\alpha>0$ and β and $i=1,\ldots,n$,

$$Z_i=\alpha W_i+\beta,$$

and we define G_n^z to be the ECF based on Z_i, then,

$$\begin{aligned}G_n^z\left(\frac{k-1}{n}\right)&=\frac{1}{k}\sum_{j=1}^{k}Z_{(j)}-Z_{(k)}+\frac{1}{n-k}\sum_{j=k+1}^{n}Z_{(j)}-Z_{(k+1)}\\&=\alpha\left[\frac{1}{k}\sum_{j=1}^{k}W_{(j)}-W_{(k)}+\frac{1}{n-k}\sum_{j=k+1}^{n}W_{(j)}-W_{(k+1)}\right]\\&=\alpha G_n\left(\frac{k-1}{n}\right),\end{aligned}$$

and therefore, G_n^z and G_n cross-over 0 at the same point. Thus assumption $A2$ is not restrictive since p_n is invariant to scaling and translation of the data; as it should be, since in a clustering problem scale and location changes of the data should not affect the clustering mechanism. We can hence quite safely assume that we are dealing with random variables with mean 0 and variance 1.

3. Main results

Let us start with the functional limit theorem for $U_n(p) = \sqrt{n}(G_n(p) - G(p))$ proved in Bharath, Pozdnyakov and Dey (2012).

Theorem 1. *Define*

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

Under assumptions A1-A3,

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

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 function given by

$$C(p,q) = Cov(U(p), U(q)) = Cov(\theta_p, \theta_q). \tag{3.1}$$

The next lemma states that the Gaussian process $U(p)$ allows a continuous modification. This fact will be employed, for example, to justify the usage of the mapping theorem (for instance, Billingsley (1968) and Pollard (1984)).

Lemma 1. *Under assumptions A1-A3, the centered Gaussian process $U(p), a \leq p \leq b$ with covariance function in (3.1) is continuous.*

Proof. First note that on the interval $[a,b]$ the functions (of p) $1/p$, $1/(1-p)$, $Q(p)$, $1/f(Q(p)) = Q'(p)$, $E[W_1 \mathbb{I}_{W_1 < Q(p)}]$ and $E[W_1^2 \mathbb{I}_{W_1 < Q(p)}]$ are continuously differentiable. Second, the functions $\max(p,q)$ and $\min(p,q)$ are (globally) Lipschitz continuous on $[a,b] \times [a,b]$. Therefore, the covariance function $C(p,q)$ is Lipschitz continuous on $[a,b] \times [a,b]$; i.e., there is a constant K such that for all p, q, p' and q' from $[a,b]$

$$|C(p,q) - C(p',q')| \leq K(|p-p'| + |q-q'|).$$

Therefore,

$$\begin{aligned} E[U(p) - U(q)]^2 &= C(p,p) + C(q,q) - 2C(p,q) \\ &\leq |C(p,q) - C(p,p)| + |C(p,q) - C(p,p)| \\ &\leq 2K|p-q|. \end{aligned}$$

By Theorem 1.4 from Adler (1990) we get that $U(p)$ is continuous. □

This immediately leads us to the following important consequence.

Corollary 1. *Under assumptions $A1 - A3$, as $n \to \infty$,*

$$\sup_{a \leq p \leq b} |G_n(p) - G(p)| \xrightarrow{P} 0.$$

Proof. Since the functional $\sup_{p\in[a,b]}|x(p)|$ is continuous on $C[a,b]$ equipped with the uniform metric, and the process $U(p)$ is continuous (that is, $U(p)\in C[a,b]$ with probability 1), by the mapping theorem (Pollard (1984), p. 70) we have

$$\sup_{a\le p\le b}\sqrt{n}\,|G_n(p)-G(p)|\Rightarrow\sup_{a\le p\le b}|U(p)|.$$

Therefore,

$$\sup_{a\le p\le b}|G_n(p)-G(p)|=\frac{1}{\sqrt{n}}\sup_{a\le p\le b}\sqrt{n}\,|G_n(p)-G(p)|\xrightarrow{P}0.$$

□

The empirical cross-over function is a step-function. The next lemma tells us that the jump at any $p\in(0,1)$ is $o_p(1/\sqrt{n})$.

Lemma 2. *Under assumptions A1 − A3, for $0<p<1$ and $\frac{k-1}{n}\le p<\frac{k}{n}$, as $n\to\infty$,*

$$\left|G_n\left(\frac{k-1}{n}\right)-G_n\left(\frac{k-2}{n}\right)\right|=o_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof.

$$\left|G_n\left(\frac{k-1}{n}\right)-G_n\left(\frac{k-2}{n}\right)\right|=\Bigg|\frac{1}{k}\sum_{i=1}^{k}W_{(i)}-W_{(k)}+\frac{1}{n-k}\sum_{k+1}^{n}W_{(i)}-W_{(k+1)}$$
$$-\frac{1}{k-1}\sum_{i=1}^{k-1}W_{(i)}+W_{(k-1)}-\frac{1}{n-k+1}\sum_{i=k}^{n}W_{(i)}+W_{(k)}\Bigg|.$$

Re-arranging terms, the RHS can written as

$$\Bigg|-\sum_{i=1}^{k-1}\frac{W_{(i)}}{k(k-1)}+\sum_{i=k+1}^{n}\frac{W_{(i)}}{(n-k)(n-k+1)}+W_{(k)}\left(\frac{n+1}{k(n-k+1)}\right)$$
$$+(W_{(k-1)}-W_{(k+1)})\Bigg|.$$

Observe that, by the law of large numbers for trimmed sums (see Stigler (1973)),

$$\frac{1}{(k-1)}\left|\sum_{i=1}^{k-1}W_{(i)}\right|\xrightarrow{P}\eta,$$

where η is a constant and hence, as $n\to\infty$,

$$\frac{1}{k(k-1)}\left|\sum_{i=1}^{k-1}W_{(i)}\right|=O_p\left(\frac{1}{n}\right)$$

and similarly,

$$\frac{1}{(n-k)(n-k+1)}\left|\sum_{i=k+1}^{n}W_{(i)}\right|=O_p\left(\frac{1}{n}\right).$$

Moreover, since $W_{(k)} \xrightarrow{P} Q(p)$, for $\frac{k-1}{n} \leq p < \frac{k}{n}$, we have that

$$\frac{(n+1)}{k(n-k+1)}|W_{(k)}| = O_p\left(\frac{1}{n}\right).$$

Suppose $M_n = \sup_{1\leq k\leq n}(W_{(k)} - W_{(k-1)})$, then from Devroye (1981), we have that

$$M_n = O_p\left(\frac{\log n}{n}\right),$$

and therefore

$$(W_{(k-1)} - W_{(k+1)}) = O_p\left(\frac{\log n}{n}\right).$$

It is hence the case that the RHS is $o_p\big(\frac{1}{\sqrt{n}}\big)$. This concludes the proof. □

Now, we are ready to prove consistency of p_n. As in Hartigan (1978) (Theorem 1) we require a uniqueness condition.

Theorem 2. *Assume $A1 - A3$ hold. Suppose that $G(p) = 0$ has a unique solution, p_0. Then for any $0 < a < p_0 < b < 1$*

$$p_n \xrightarrow{P} p_0,$$

as $n \to \infty$.

Proof. Note that by Cauchy-Schwarz inequality we have

$$[EW_1\mathbb{I}_{W_1\geq Q(p)}]^2 \leq E[W_1^2\mathbb{I}_{W_1\geq Q(p)}]P[W_1 \geq Q(p)].$$

Since the second moment of W_1 is finite, we obtain that $B(Q, 0+) = B(Q, 1-) = 0$. That is, a nonnegative continuously differentiable split function $B(Q, p)$ has a unique maximum at p_0, and, as a result, $G(p)$ does change sign at p_0. Choose a, b such that $0 < a < p_0 < b < 1$. Because G is continuous on $[a, b]$ we get that $G(p) > 0$ for $a \leq p < p_0$ and $G(p) < 0$ for $b \geq p > p_0$. Moreover, for any $\delta > 0$ there exists an $\epsilon > 0$ and $0 < \delta' < \delta$ such that

$$G(p) > \epsilon \text{ for } a \leq p < p_0 - \delta',$$

and

$$G(p) < -\epsilon \text{ for } b \geq p > p_0 + \delta'.$$

By Corollary 1, as $n \to \infty$

$$P\left(\sup_{a\leq p\leq b}|G_n(p) - G(p)| < \frac{\epsilon}{2}\right) \to 1,$$

and therefore,

$$P\left(\inf_{a\leq p<p_0-\delta'} G_n(p) > \frac{\epsilon}{2} \text{ and } \sup_{b\geq p>p_0+\delta'} G_n(p) < -\frac{\epsilon}{2}\right) \to 1.$$

Using the result from Lemma 2, we obtain

$$P(p_n \in [p_0 - \delta', p_0 + \delta']) \to 1.$$

Note that since

$$P(p_n \in [p_0 - \delta, p_0 + \delta]) \geq P(p_n \in [p_0 - \delta', p_0 + \delta']),$$

we finally have

$$P(p_n \in [p_0 - \delta, p_0 + \delta]) \to 1.$$

□

Now, under an additional assumption that $G'(p_0) < 0$ (cf. with Theorem 2 from Hartigan (1978)) we will establish asymptotic normality of p_n. This result will be proved in three steps. First, we will establish that p_n is in the $O_p(1/\sqrt{n})$ neighborhood of p_0. Then we will show that in this neighborhood $G_n(p)$ can be adequately approximated by a line with slope $G'(p_0)$. Finally, an approach based on Bahadur's general method (see p. 95, Serfling (1980)) will be employed to get the CLT for p_n.

Lemma 3. *Assume $A1-A3$ hold. Suppose that $G(p) = 0$ has a unique solution, p_0, and $G'(p_0) < 0$. If a, b are such that $0 < a < p_0 < b < 1$, then for any $\delta > 0$ there exist N and $C > 0$ such that for all $n \geq N$*

$$P\left(|p_n - p_0| \leq \frac{C}{\sqrt{n}}\right) > 1 - \delta.$$

Proof. Fix arbitrary $\delta > 0$. Using Theorem 1 and mapping theorem we have

$$\sup_{a \leq p \leq b} \sqrt{n}\,|G_n(p) - G(p)| \Rightarrow \sup_{a \leq p \leq b} |U(p)|.$$

Therefore, for any $\delta > 0$ there exist N' and $C' > 0$ such that for all $n > N'$ we have

$$P\left(\sup_{a \leq p \leq b} |G_n(p) - G(p)| < \frac{C'}{\sqrt{n}}\right) > 1 - \delta. \tag{3.2}$$

By the same argument as in Theorem 2, p_0 is a unique split point, $0 < p_0 < 1$, $G(p) > 0$ for $p < p_0$ and $G(p) < 0$ for $p > p_0$. Assumption $G'(p_0) < 0$ tells us that in a neighborhood of p_0 the function $G(p)$ behaves like a line. Taking this into account we get that there exist $N > N'$ and $C > 0$ such that for all $n > N$

$$G(p) > \frac{2C'}{\sqrt{n}} \text{ for } a \leq p < p_0 - \frac{C}{\sqrt{n}},$$

and

$$G(p) < -\frac{2C'}{\sqrt{n}} \text{ for } b \geq p > p_0 + \frac{C}{\sqrt{n}}.$$

Then by (3.2) we find that for all $n > N$

$$P\left(\inf_{a\le p<p_0-C/\sqrt{n}} G_n(p) > \frac{C'}{\sqrt{n}} \text{ and } \sup_{b\ge p>p_0+C/\sqrt{n}} G_n(p) < -\frac{C'}{\sqrt{n}}\right) > 1-\delta.$$

Therefore,

$$P\left(|p_n - p_0| \le \frac{C}{\sqrt{n}}\right) > 1-\delta.$$

□

Lemma 4. *Assume* $A1-A3$ *hold. Suppose that* $G(p) = 0$ *has a unique solution,* p_0*, and* $G'(p_0) < 0$*. Then for any* $C > 0$

$$\sup_{p\in I_n} \sqrt{n}\,|G_n(p) - G_n(p_0) - G'(p_0)(p-p_0)| \overset{P}{\to} 0, \text{ as } n\to\infty,$$

where $I_n = [p_0 - \frac{C}{\sqrt{n}}, p_0 + \frac{C}{\sqrt{n}}]$*, and*

$$G^{'}(p_0) = \frac{1}{p_0}\left[Q(p_0) - Q_l(p_0)\right] - \frac{1}{1-p_0}\left[Q(p_0) - Q_u(p_0)\right] - 2Q'(p_0). \tag{3.3}$$

Proof. Since the second derivative of $G(p)$ is uniformly continuous on $p_0 - C/\sqrt{n} \le p \le p_0 - C/\sqrt{n}$ we have

$$\begin{aligned} G(p) - G(p_0) &= (p-p_0)G'(p) + O((p-p_0)^2) \\ &= (p-p_0)G'(p) + O(1/n) \end{aligned}$$

It is hence sufficient to show that

$$\sup_{p\in I_n} \sqrt{n}\,|[G_n(p) - G(p)] - [G_n(p_0) - G(p_0)]| \overset{P}{\to} 0,$$

or that for any $\epsilon > 0$ and $\delta > 0$ there exists N such that for all $n > N$

$$P\left(\sup_{p\in I_n} \sqrt{n}\,|[G_n(p) - G(p)] - [G_n(p_0) - G(p_0)]| > \epsilon\right) < \delta.$$

Take arbitrary $\delta' > 0$. The functional $\sup_{p\in[p_0-\delta',p_0+\delta']} |x(p) - x(p_0)|$ is continuous on $C[a,b]$ equipped with the uniform metric. Therefore, Theorem 1 and the mapping theorem informs us that

$$\begin{aligned} &\sup_{p_0-\delta'\le p\le p_0+\delta'} \sqrt{n}\,|[G_n(p) - G(p)] - [G_n(p_0) - G(p_0)]| \\ &\quad\Rightarrow \sup_{p_0-\delta'\le p\le p_0+\delta'} |U(p) - U(p_0)|\,. \end{aligned}$$

Since for all sufficiently large n we have

$$\begin{aligned} &\sup_{p\in I_n} \sqrt{n}\,|[G_n(p) - G(p)] - [G_n(p_0) - G(p_0)]| \\ &\le \sup_{p_0-\delta'\le p\le p_0+\delta'} \sqrt{n}\,|[G_n(p) - G(p)] - [G_n(p_0) - G(p_0)]| \quad \text{a.s.,} \end{aligned}$$

it is indeed the case that

$$P\Big(\sup_{p\in I_n}\sqrt{n}\,|[G_n(p)-G(p)]-[G_n(p_0)-G(p_0)]|>\epsilon\Big)$$
$$\leq P\Big(\sup_{p_0-\delta'\leq p\leq p_0+\delta'}\sqrt{n}\,|[G_n(p)-G(p)]-[G_n(p_0)-G(p_0)]|>\epsilon\Big).$$

As a consequence,

$$\limsup_n P\Big(\sup_{p\in I_n}\sqrt{n}\,|[G_n(p)-G(p)]-[G_n(p_0)-G(p_0)]|>\epsilon\Big)$$
$$\leq P\Big(\sup_{p_0-\delta'\leq p\leq p_0+\delta'}|U(p)-U(p_0)|>\epsilon\Big).$$

Because the Gaussian process $U(p)$ is continuous, $\sup_{p_0-\delta'\leq p\leq p_0+\delta'}|U(p)-U(p_0)|\to 0$ with probability 1 as $\delta'\to 0$, and, therefore, it converges to 0 in probability. Choosing δ' small enough we can make

$$\limsup_n P\Big(\sup_{p\in I_n}\sqrt{n}\,|[G_n(p)-G(p)]-[G_n(p_0)-G(p_0)]|>\epsilon\Big)<\delta.$$

□

Lemma 5. *Assume A1 − A3 hold. Suppose that $G(p)=0$ has a unique solution, p_0, and $G'(p_0)<0$. If a,b are such that $0<a<p_0<b<1$ then as $n\to\infty$*

$$p_n = p_0 - \frac{G_n(p_0)}{G'(p_0)} + o_p(n^{-1/2}),$$

where $G'(p)$ is as defined in (3.3).

Proof. Consider the line $G_n(p_0)+G'(p_0)(p-p_0)$. Let random variable p^* be the solution of

$$G_n(p_0)+G'(p_0)(p-p_0)=0,$$

that is,

$$p^* = p_0 - \frac{G_n(p_0)}{G'(p_0)}. \tag{3.4}$$

From Theorem 1, we know that

$$G_n(p_0)-G(p_0)=G_n(p_0)=O_p(n^{-1/2})$$

and we hence have that $p^*=p_0+O_p(n^{-1/2})$. By Lemma 3 p_n is also in a $O_p(1/\sqrt{n})$ neighborhood of p_0. By Lemma 4 in $O_p(1/\sqrt{n})$ neighborhood of p_0 uniformly

$$G_n(p)=G_n(p_0)+G'(p_0)(p-p_0)+o_p(n^{-1/2}),$$

by which we can claim that $p_n=p^*+o_p(n^{-1/2})$. The result follows by substituting for p^* in (3.4). □

This immediately give us the final result.

Theorem 3. *Assume $A1 - A3$ hold. Suppose that $G(p) = 0$ has a unique solution, p_0, and $G'(p_0) < 0$. If a, b are such that $0 < a < p_0 < b < 1$ then as $n \to \infty$,*

$$\sqrt{n}(p_n - p_0) \Rightarrow N\left(0, \frac{Var(\theta_{p_0})}{G'^2(p_0)}\right),$$

where θ_{p_0} is as defined in Theorem 1.

3.1. Numerical verification

In this section we provide verification of our results regarding the asymptotic normality of p_n along the lines of Table 1 in Hartigan (1978). Since our split point p_0 coincides with Hartigan's split point (maximum of $B(Q,p)$), it is to be expected that our empirical split point p_n behaves asymptotically similar to his. Hartigan verifies his results when observations are obtained from a $N(0,1)$ population—a population with smooth distribution function; we do the same and note that the asymptotic mean and the variance of p_n agree with his.

It is a simple exercise to ascertain that, for the normal case, the split point p_0 is 0.5, $G'^2(0.5) \approx 3.34$ and $Var(\theta_{0.5}) = 2\pi - 4 \approx 2.283$. Consequently, we observe that the asymptotic variance of $\sqrt{n}(p_n - 0.5)$ is approximately 0.69. The table below corroborates our theoretical results.

Table 1
Simulated mean and variance of $\sqrt{n}(p_n - 0.5)$ for different sample sizes for the normal case. 1000 simulations were performed for each sample size

Sample sizes	$n = 100$	$n = 300$	$n = 500$	$n = 1000$
Simulated Mean	0.506	0.504	0.501	0.502
Simulated Variance	0.614	0.646	0.700	0.691

3.2. An example: Confidence interval estimation

We demonstrate here how Theorem 3 can be employed to construct approximate confidence intervals (CI) for a theoretical split point. We consider a classical example of bimodal distribution—the variable "eruption" in the data set `faithful` available in `R` package MASS. The data set contains 272 measurements of the duration of eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

First, we plot the ECF for the variable "eruption"; the plot is given in Figure 1. We can see that $G_n(\cdot)$ is generally a decreasing function that crosses zero line once, far away from 0 and 1: the end-points of its domain which is the $(0,1)$ interval. Thus our point estimate of theoretical split point is $p_n = 97/272 \approx .357$.

Now, to construct an approximate CI for p_0 we need to estimate $Var(\theta_{p_0})/G'^2(p_0)$. A straightforward (but rather tedious) calculation shows that this quantity explicitly depends on the following terms: p_0, $Q(p_0)$, $f(Q(p_0))$, $Q_l(p_0)$,

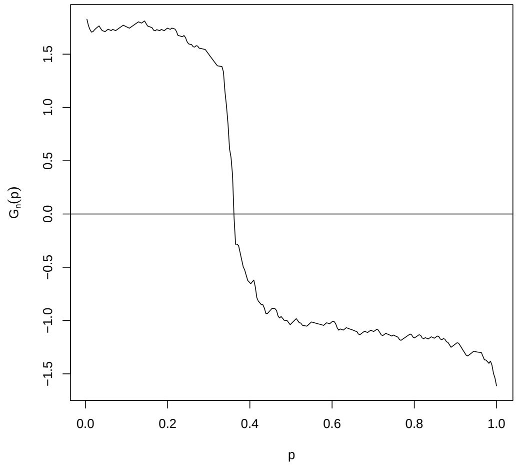

FIG 1. *Empirical Cross-over Function $G_n(p)$ for data set* `faithful`.

$Q_u(p_0)$,

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

We estimate these terms as follows:

$$p_0 \approx p_n, \quad Q(p_0) \approx W_{(98)},$$

$$Q_l(p_0) \approx \frac{1}{98} \sum_{i=1}^{98} W_{(i)}, \quad Q_u(p_0) \approx \frac{1}{272-98} \sum_{i=99}^{272} W_{(i)},$$

$$B_l(p_0) \approx \frac{1}{98} \sum_{i=1}^{98} W_{(i)}^2, \quad B_u(p_0) \approx \frac{1}{272-98} \sum_{i=99}^{272} W_{(i)}^2.$$

Finally, $f(Q(p_0))$ is estimated by $\hat{f}(W_{(98)})$, where $\hat{f}$ comes from the standard `R` function `density`. As a result, for instance, the 95% confidence interval for a theoretical split point p_0 is given by

$$.357 \pm .057.$$

4. Discussion

Admittedly, the definition of the empirical split point "in the range $[a, b]$" might appear a bit artificial. But we still believe the results can be useful in practical applications. For instance, as with the `faithful` data, if we know that the distribution at hand is bimodal, and we want to estimate a split point between two clusters, it is safe to assume that the split point is in the range between two modes.

It turns out that the behavior of the cross-over function when it is close to 0 or 1 can be rather complicated; under some natural assumptions, it can be shown that $\limsup_{p\uparrow 1} G(p) < 0$. For example, it is true if W_1 is bounded from above. When $Q(1-) = +\infty$, the following condition

$$\limsup_{x\to\infty} \frac{E(W_1 \mathbb{I}_{W_1 \geq x})}{xP(W_1 \geq x)} < 2 \tag{4.1}$$

is sufficient for $\limsup_{p\uparrow 1} G(p) < 0$. It is easy to see that, for instance, distributions with regularly varying tails and $EW_1^{2+\epsilon} < \infty$, for $\epsilon > 0$, satisfy (4.1). However, it is possible to construct a distribution with a "bumpy" tail for which $\limsup_{p\uparrow 1} G(p) \geq 0$. Consequently, it is suggestive that any extension of definition of p_n to the entire interval $(0, 1)$ will require some additional assumptions.

Acknowledgements

We are thankful to the Associate Editor and referee for careful reading of the manuscript, thoughtful criticisms and very helpful suggestions that allowed us to significantly improve the presentation of our results. We also want to thank Zhiyi Chi for discussions on the behavior of the cross-over function around 0 and 1.

References

Adler, R. J. (1990). An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes. *Lecture Notes-Monograph Series* **12**. MR1088478

Arnold, S. J. (1979). A Test for Clusters. *Journal of Marketing Research* **16** 545-551.

Bharath, K., Pozdnyakov, V. and Dey, D. K. (2012). Asymptotics of Empirical Cross-over Function. *Unpublished manuscript*. Arxiv:1112.3427v3.

Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley and Sons, New York. MR0233396

Bock, H. H. (1985). On Some Significance Tests in Cluster Analysis. *Journal of Classification* **2** 77-108. MR0800515

Cuesta-Albertos, J. A., Gordaliza, A. and Matrán, C. (1997). Trimmed k-means: An Attempt to Robustify Quantizers. *Annals of Statistics* **25** 553-576. MR1439314

Devroye, L. (1981). Laws of Iterated Logarithm for Order Statistics of Uniform Spacings. *Annals of Probability* **9** 860-867. MR0628878

Engleman, L. and Hartigan, J. A. (1969). Percentage Points of a Test for Clusters. *Journal of the American Statistical Association* **64** 1647-1648.

García-Escudero, L. A., Gordaliza, A. and Matrán, C. (1999). A Central Limit Theorem for Multivariate Generalized Trimmed k-means. *Annals of Statistics* **27** 1061-1079. MR1724041

Hartigan, J. (1978). Asymptotic Distributions for Clustering Criteria. *Annals of Statistics* **6** 117-131. MR0461780

Hartigan, J. A. and Hartigan, P. M. (1985). A Dip Test of Unimodality. *Annals of Statistics* **13** 70-84. MR0773153

Holzmann, H. and Vollmer, S. (2008). A Likelihood Ratio Test for Bimodality in Two-component Mixtures with Application to Regional Income Distribution in the EU. *Advances in Statistical Analysis* **92** 57-69. MR2389571

Pollard, D. (1981). Strong Consistency for K-Means Clustering. *Annals of Statistics* **9** 135-140. MR0600539

Pollard, D. (1982). A Central Limit Theorem for k-means Clustering. *Annals of Statistics* **10** 919-926. MR0672292

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York. MR0762984

Schwab, J., Podsiadlowski, P. H. and Rappaport, S. (2012). Further Evidence for the Bimodal Distribution of Neutron-Star Masses. *The Astrophysical Journal* **719** 722-727.

Serfling, R. (1980). *Approximation Theorems for Mathematical Statistics*. John Wiley, New york. MR0595165

Serinko, R. J. and Babu, G. J. (1992). Weak Limit Theorems for Univariate k-means Clustering under Nonregular Conditions. *Journal of Multivariate Analysis* **49** 188-203.

Stigler, S. M. (1973). The Asymptotic Distribution of the Trimmed Mean. *Annals of Statistics* **1** 472-477. MR0359134

Wolfe, J. H. (1970). Pattern Clustering by Multivariate Mixture Analysis . *Multivariate Behavioral Research* **5** 329-350.