Asymptotics of a Clustering Criterion for Smooth Distributions

Vladimir Pozdnyakov

University of Connecticut

based on joint work with K. Bharath and D. Dey

ICORS 2013

Notation and assumptions

Let $W_1, W_2, \cdots, W_n$ 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.

k-means clustering

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

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

Hartigan's split function

The *split function* was introduced in Hartigan (Annals of Statistics, 1978) for partitioning a sample into two groups, and it is defined as

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

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

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

Split point

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

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

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*,

$$G(p) = Q_l(p) + Q_u(p) - 2Q(p), \tag{3}$$

for examining clustering properties.

Empirical cross-over function

From a statistical perspective, we would like to work with the empirical version of (3). We deviate here from Hartigan's framework and consider the *empirical cross-over function*(ECF), defined as

$$G_n(p) = \frac{1}{k}\sum_{j=1}^{k} W_{(j)} - W_{(k)} + \frac{1}{n-k}\sum_{j=k+1}^{n} W_{(j)} - W_{(k+1)}, \tag{4}$$

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

$$G_n(p) = \frac{1}{n}\sum_{j=1}^{n} W_{(j)} - W_{(n)}, \tag{5}$$

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

Empirical split point

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.

Functional CLT for $G_n(p)$

Consider $U_n(p) = \sqrt{n}(G_n(p) - G(p))$.

Theorem 1 *Define*

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

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

$$C(p,q) = Cov(U(p), U(q)) = Cov(\theta_p, \theta_q). \tag{6}$$

LLN for $G_n(p)$

The next lemma states that the Gaussian process $U(p)$ allows a continuous modification. This fact is employed, for example, to justify the usage of the mapping theorem.

Lemma 1 *Under assumptions A1-A3, the centered Gaussian process* $U(p), a \leq p \leq b$ *with covariance function in (6) is continuous.*

This immediately leads us to the following important consequence.

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

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

Consistency of p_n

An immediate consequence is 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 $n \to \infty$.

Normality of p_n

Now, under an additional assumption that $G'(p_0) < 0$ (cf. with Theorem 2 from Hartigan (1978)) one can establish asymptotic normality of p_n. This result is proved in three steps. First, we establish that p_n is in the $O_p(1/\sqrt{n})$ neighborhood of p_0. Then we 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)) is employed to get the CLT for p_n.

Normality of p_n: p_n is in the $O_p(1/\sqrt{n})$ neighborhood of p_0

Lemma 2 *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 $n \geq N$*

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

Normality of p_n: $G_n(p)$ is almost a line with slope $G'(p_0)$ in the neighborhood

Lemma 3 *Assume $A1-A3$ hold. Suppose that $G(p) = 0$ has a unique solution, p_0, and $G'(p_0) < 0$. Then for any $C > 0$*

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

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

$$G^{'}(p_0) = \frac{1}{p_0} \left[Q(p_0) - Q_l(p_0) \right] - \frac{1}{1 - p_0} \left[Q(p_0) - Q_u(p_0) \right] - 2Q'(p_0). \qquad (7)$$

Normality of p_n: connection between p_n and G_n

Lemma 4 *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 (7).

Main result: CLT for p_n

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 $n \to \infty$,*

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

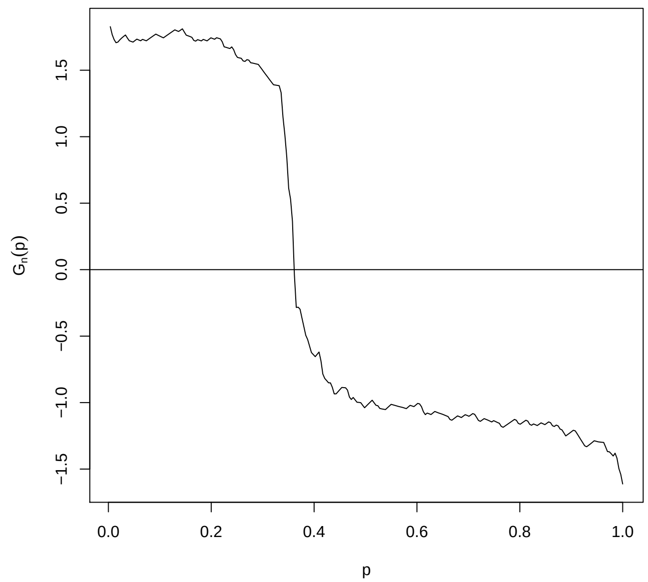
where θ_{p_0} is as defined in Theorem 1.

First example: Old Faithful geyser

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 Example: Old Faithful Geyser

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 $p_n = 97/272 \approx .357$.

First example: Old Faithful geyser

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, $Q(p_0)$, $f(Q(p_0))$, $Q_l(p_0)$, $Q_u(p_0)$,

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

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

$$.357 \pm .057.$$

Second example: Merton/Kou models

The popular Merton model for assets pricing X_t is a one-dimensional process given by

$$X_t = X_0 + \mu t + \sigma B_t + \sum_{k=0}^{P_t(\lambda)} J_k \quad 0 \leq t \leq T, \tag{8}$$

where the scalar $\mu \in \mathbb{R}$ represents the drift component of the process, $\sigma \in \mathbb{R}^+$, its spot volatility, B_t is the standard Brownian motion and the process P_t is a Poisson jump process with intensity λ with jumps sizes represented by i.i.d random variables J_k. It is assumed that B_t, $P_t(\lambda)$ and $\{J_k\}$ are independent.

Second example: Merton/Kou models

"Geometric" version is called the Kou's jump-diffusion model. It is a process defined by the stochastic differential equation

$$\frac{dS_t}{S(t-)} = \mu dt + \sigma dB_t + d\left(\sum_{i=1}^{P_t(\lambda)} (V_i - 1)\right), \tag{9}$$

where V_i are i.i.d non-negative random variables and all other quantities are as defined in the Merton model in (8). We observe either X_t or S_t only at n discrete equally spaced times: $0 \leq \Delta \leq 2\Delta \leq \cdots \leq n\Delta \leq T$ where $\Delta = T/n$.

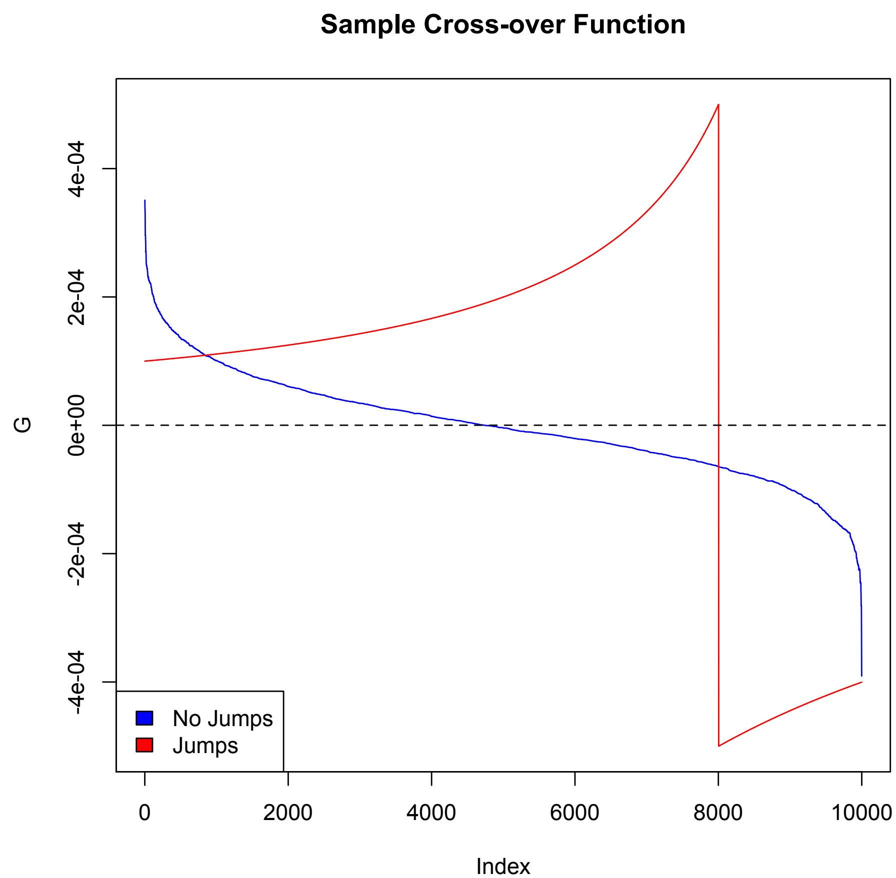
Second example: Merton/Kou models

Special Case

In this illustration, we use the notation of the Merton model. For the purposes of demonstrating the utility of our method, we consider a special case of (8) when all the jumps are of unknown constant size $h > 0$. The model in (8) consequently reduces to

$$X_t = X_0 + \mu t + \sigma B_t + hP_t(\lambda) \quad 0 \le t \le T. \tag{10}$$

Second example: Merton/Kou model



Second example: Merton/Kou model

We report the performance of our test and also provide a comparison with the test given in Ait-Sahalia and Jacod (2009) with $p = 4$, $k = 2$ and $\Delta_n = \frac{1}{n}$. We perform 10000 simulations with $\mu = 0$ and $\sigma = 1$ since both the tests do not depend on them. To recall, the null hypothesis is that the process X_t follows a Brownian motion with constant drift μ and constant volatility σ and the alternative hypothesis is that the X_t follows the Merton/Kou model.

Second example: Merton/Kou model

Here are the results:

	Rejection rate in simulations	
n	**Our test**	**Ait-Sahalia's test**
500	0.043	0.1037
1000	0.046	0.0776
5000	0.0492	0.0452
10000	0.0497	0.0418
25000	0.0482	0.0465
50000	0.0501	0.0505

From the table we can observe that our test requires fewer number of observations, as compared to Ait-Sahalia's test, to attain level α. However, this is not surprising since Ait-Sahalia's test is applicable under a very general setup for a large class of semimartingales. Our test, on the other hand, is testing for two specific models and expectedly performs better. Our claim, however, is, if one is interested in choosing between a Brownian motion with drift model and the Merton/Kou model, our test is a better alternative to Ait-Sahalia's general test.

References

- K. Bharath, V.I. Pozdnyakov, and D. Dey, Asymptotics of a Clustering Criterion for Smooth Distributions, *Electronic Journal of Statistics*, 7 (2013), 1078-1093.

- K. Bharath, V.I. Pozdnyakov, and D. Dey, Asymptotics of the Empirical Cross-over Function, *Annals of the Institute of Statistical Mathematics*, to appear.