Journal of Statistical Planning and Inference 147 (2014) 84–94

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

On a clustering criterion for dependent observations

Karthik Bharath *

Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, United States

ARTICLE INFO

Article history:
Received 22 April 2013
Received in revised form
9 November 2013
Accepted 15 November 2013
Available online 25 November 2013

Keywords:
Clustering
β-Mixing
Truncated sums
LLN
CLT
Functional CLT

ABSTRACT

A univariate clustering criterion for stationary processes satisfying a β-mixing condition is proposed extending the work of Bharath et al. (2013) to the dependent setup. The approach is characterized by an alternative sample criterion function based on truncated partial sums which renders the framework amenable to various interesting extensions for which limit results for partial sums are available. Techniques from empirical process theory for mixing sequences play a vital role in the arguments employed in the proofs of the limit theorems.

1. Introduction

The purpose of this paper is two-fold: we propose a clustering criterion for observations from a smooth, invertible distribution function which is noticeably simpler than the one recently proposed by Bharath et al. (2013); we then demonstrate the power of its simplicity by proving limit theorems for clustering constructs under a dependence setup extending their work from the i.i.d. case. A pleasant by-product of the proposed framework is that the proofs of the limit results, in some cases, follow along similar lines to the ones in Bharath et al. (2013).

Hartigan (1978) considered the statistical underpinnings of the k-means clustering framework and derived asymptotic distributions of a suitably defined criterion function and its maximum. Given a set of observations (data), he defined a cluster to be the subset of the observations which are grouped together on the basis of a point which splits the data by maximizing the between-group sums of squares; in other words, he considered a criterion function which was based on maximizing the between cluster sums of squares as opposed to minimizing the within cluster sums of squares. Bharath et al. (2013) too resorted to maximizing the between cluster sums of squares but deviated from Hartigan's framework and instead considered the zero of a certain function of the derivative of Hartigan's criterion function. Contrary to Hartigan's setup, which required the existence of a fourth moment for asymptotic results, they proved results for their clustering constructs under a second moment assumption with an added smoothness condition on the criterion function. Both works considered i.i.d. data and primarily the case $k=2$, viz. two clusters.

We extend the approach used in Bharath et al. (2013) to the case of dependent data satisfying a β-mixing condition and reprove all results under a dependent setup using techniques from empirical processes. We achieve this by employing an alternative sample version of the criterion function used in their paper which, by virtue of its definition, renders itself

* Tel.: +1 6077274172.
E-mail address: karthikbharath@gmail.com

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.11.005

amenable to a variety of interesting scenarios. The circumscription of the proposed framework to the β-mixing case is not necessary and is artificial; our intention is to highlight the versatility of our definition of the sample criterion function and its deployment in investigating asymptotic behavior of the k-means clustering criterion for dependent observations. The results presented in this paper ought to be viewed as a demonstration of the framework's amenability to sequences having a variety of dependence structures for which limit results for partial sums exist. The results presented in this paper are also valid to sequences which satisfy a ϕ-mixing condition since it represents a stronger condition in the sense that a ϕ-mixing sequence is also β-mixing.

Interest in β-mixing sequences has increased in recent years with attempts to develop stability bounds for learning algorithms. In many machine learning applications the i.i.d. assumption is not valid since the observations obtained from the learning algorithms usually display inherent temporal dependence. In fact, Vidyasagar (2003) argues that β-mixing is "just the right" assumption for the analysis of weakly dependent sample points in machine learning, in particular since several learning results then carry over to the non-i.i.d. case. Another important application of β-mixing sequences is in modeling scenarios involving aperiodic and positively recurrent Markov chains; i.e. if $\{Y_i\}_{i \geq 1}$ is an aperiodic positive recurrent Markov chain, then it satisfies the β mixing condition (see Rio, 2000, p. 139). This fact has been employed in several econometric applications; for an overview see Doukhan (1994). In a similar vein, the condition of ϕ-mixing is also a fairly common assumption while analyzing stability of learning algorithms such as Support Vector Regression (SVR) (Vapnik, 1998), Kernel Ridge Regression (KRR) (Saunders et al., 1998) and Support Vector Machines (SVM) (Cortes and Vapnik, 1995). Since ϕ-mixing represents a stronger condition than β-mixing, stability bounds for β-mixing usually lead to good bounds for ϕ-mixing too. In a recent paper, Mohri and Rostamizadeh (2010) proved stability bounds for stationary ϕ-mixing and β-mixing processes with applications in the above-mentioned algorithms. It is therefore of considerable interest, at least in machine learning applications, to move over from the i.i.d. setup to the β-mixing or ϕ-mixing cases when weak dependence which decays over time is a reasonable modeling assumption. The results in this paper can conceivably be employed while analyzing clustering properties of such sequences.

In Section 2 we review some of the pertinent constructs from Bharath et al. (2013) and Hartigan (1978) and define the quantities of interest in this paper: the theoretical and sample criterion functions and their respective zeroes. In Section 3 we state our assumptions regarding the conditions on the distribution function F and rate of decay of the β-mixing coefficients for the rest of the paper. Then, in Section 4, we examine the nature of the criterion function as a statistical functional which induces a functional on the space of cádlág functions and note some of the difficulties involved in directly using existing results. In Section 5, in anticipation of the arguments employed in the main results, we prove a few preparatory lemmas and state two results from the literature which play an important role in our proofs. Section 6 comprises the main results of this paper. We first examine the sample criterion function and prove a weak convergence result; this then provides us with several useful corollaries which will then assist us in examining the sample split point. However, since the "zero" of our sample criterion function is similar to the one in Bharath et al. (2013), the proofs of statements regarding its asymptotic behavior follow along almost identical lines as the ones for the empirical split point in Bharath et al. (2013); for ease of exposition and interests of brevity, we direct the interested reader to their paper for the details of the proofs. Finally, in Section 7, we provide numerical verification of the limit theorems via a simulation exercise.

2. Preliminaries

Let us first consider the criterion function that was introduced in Hartigan (1978) for partitioning a sample into two groups. For a distribution function F, if Q is the quantile function associated with F, then the *split function* of Q at $p \in (0,1)$ is defined as

$$B(Q,p) = p\left[\frac{1}{p}\int_{q<p} Q(q)\,dq\right]^2 + (1-p)\left[\frac{1}{1-p}\int_{q>p} Q(q)\,dq\right]^2 - \left[\int_0^1 Q(q)\,dq\right]^2. \tag{2.1}$$

A value $p_0 \in (0,1)$ which maximizes the split function is called the *split point*. When F is invertible (Q is the unique inverse), Bharath et al. (2013) considered a different criterion function defined as, for $0<p<1$,

$$G(p) = \frac{1}{p}\int_0^p Q(q)\,dq + \frac{1}{1-p}\int_p^1 Q(q)\,dq - 2Q(p),$$

which is a function of the derivative of $B(Q,p)$ with respect to p. The zero of G coincides with the maximum of the parabolic split function B. The linear statistical functional (of F) G proffers a simple way to determine the split point when F is invertible as opposed to Hartigan's parabolic split function which accommodated a general F.

If $X_{(i)},\ i=1,\dots,n$ are order statistics corresponding to i.i.d. real-valued observations X_i obtained from F, then the empirical counterpart of G, christened, the *Empirical Cross-over Function* (ECF), was defined in Bharath et al. (2013) as

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