Известия НАН Армении, Математика, том 58, н. 3, 2023, стр. 3 – 13.

WEAKLY CONSISTENT OFFLINE CLUSTERING OF ARMA PROCESSES

G. L. ADAMYAN

Yerevan State University, Armenia
E-mails: *garik.adamyan@ysu.am; garikadamyan97@gmail.com*

Abstract. In this paper, we consider the problem of weakly consistent offline clustering of ARMA processes. Under the provided assumptions we derive a weakly consistent clustering algorithm of invertible ARMA processes according to their forecast functions. Using BIC penalized quasi-maximum likelihood estimate of the distance function the weak consistency of Algorithm 1 is proven when the target number of clusters is known. The theoretical lower bound of the clustering function is provided.

MSC2020 numbers: 62M10; 62H30.
Keywords: unsupervised learning; consistent clustering; ARMA processes.

1. Introduction

The clustering problem is an unsupervised learning problem for grouping similar observations. Due to their unsupervised nature, clustering algorithms have a broad range of use in numerous fields of study, including finance, biology, and robotics [**?**]. Cluster analysis of random vectors, where objects are sampled from high-dimensional joint distributions, is an extensive research area with rich literature. Although there are general definitions and approaches to the problem of clustering, the clustering of random processes requires a special approach because their observations (realizations, time series) are sampled from process distributions. These algorithms are also interesting because, unlike the clustering of random vectors, the clustering of random processes also allows studying the new dimensions of asymptotics, the asymptotics of the realizations.

In general, clustering algorithms can be classified into six groups: Partitioning, Hierarchical, Grid-based, Model-based, Density-based, and Multi-step clustering algorithms. For a comprehensive review of existing algorithms, we refer the reader to work [1], as we mainly focus on partitioning-based clustering algorithms. The typical partitioning-based clustering algorithm requires a similarity measure to measure similar samples, the target number of clusters, and some partitioning algorithms, to group similar samples. Relying on this methodology, there are many algorithms introduced in the literature, the main difference of which lies in the way of defining

the distance metric and changes in the partitioning algorithm [2]. In addition to the introduction of new algorithms, analyzing the asymptotic behavior of time series clustering algorithms is also a noteworthy direction.

In [2], the authors presented consistent clustering algorithms for ergodic and stationary processes in online and offline problem setups. In [3], the authors considered the problem of clustering of wide-sense stationary ergodic processes, the asymptotically consistent algorithms are presented for clustering these processes. The presented algorithm in the mentioned works is based on strongly consistent distance estimates, which ensure the strong consistency of clustering algorithms in the offline problem setting. In this paper, we consider weakly asymptotically consistent clustering of ARMA processes according to their forecasting functions. We construct an asymptotically consistent estimation procedure of the defined metric and prove weakly asymptotically consistency of the presented algorithm.

2. Problem setup

In this section, we formally define consistent clustering of ARMA processes in an offline setting. We start by defining ARMA(p,q) models. Let ϵ_t be a Gaussian white noise, then $X = \{X_t\}_{t=1}^{\infty}$ stochastic process is an ARMA(p, q) process if X is stationary and if for every t:

$$X_t = \epsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \tag{2.1}$$

where the polynomials $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - ... - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + ... + \theta_q z^q$ have no common factors. By standard Box-Jenkins notation (2.1) becomes.

$$\phi(B)X_t = \theta(B)\epsilon_t$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$ and $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q$ are the lag polynomials.

Definition 2.1 (Invertibility of ARMA)**.** *An ARMA(p, q) process X is invertible if there exist absolutely summable constants $\boldsymbol{\pi}_x = \{\pi_j\}_{j=0}^{\infty}$ such that $\epsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$ for all t.*

The invertibility condition is equivalent to the condition $\theta(z) = 1 - \theta_1 z - \theta_2 z^2 - ... - \theta_p z^q \neq 0,\ |z| \leq 1$.([5]:86). Let us denote by $\mathcal{L}$ the class of invertible ARMA models. The invertibility assumption ensures that X_t can be represented in terms of its past values according to the $AR(\infty)$ formulation.

$$\pi(B)X_t = \epsilon_t \tag{2.2}$$

where $\pi(B) = \theta(B)^{-1} * \phi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$. The coefficients of sequence $\boldsymbol{\pi}_x$ are determined by the following recursive equations ([5]:86):

$$\pi_j + \sum_{k=1}^{q} \theta_k \pi_{j-k} = -\phi_j, j = 0, 1, ... \tag{2.3}$$

where $\phi_0 := -1, \phi_j := 0$ for $j > p$, and $\pi_j := 0$ for $j < 0$. Having (2.2), we note that given initial values and known orders, any process $X \in \mathcal{L}$ is fully characterized by the sequence $\boldsymbol{\pi}_x$. Defined sequence also completely specifies the forecasting function $\mathcal{F} = \mathbb{E}[X_t | X_{t-1}, X_{t-2}, ...]$ of the processes X.

We are given a time series dataset with N samples $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$. We assume that each $\mathbf{x}_i$ is generated from one of the κ unknown ARMA process with unknown forecasting function $\mathcal{F}_k, k = 1, 2, ..., \kappa$, where $\kappa < N$. Note that time series samples may have arbitrary lengths, and we denote the length of $\mathbf{x}_i$ time series by n_i.

Definition 2.2 (Ground-truth $\mathcal{G}$)**.** *Let $\mathcal{G} = \mathcal{G}_1, ..., \mathcal{G}_k$ be a partitioning of the set $\{1, 2, ..., N\}$ into κ disjoint subsets $\mathcal{G}_k$, $\mathcal{G}_k \neq \emptyset$, $k = 1, 2, ..., \kappa$, such that the forecasting function of the process that generates $\mathbf{x}_i$, $i = 1, 2, ..., N$ is $\mathcal{F}_k$ for some $k = 1, 2, ..., \kappa$ if and only if $i \in \mathcal{G}_k$. We call $\mathcal{G}$ the ground-truth clustering.*

We denote by $X^{(k)}$ the underlying ARMA process for the cluster $\mathcal{G}_k$. From Definition 2.2, we need to note that given the same initial values the processes in the same cluster, will produce the same forecast.

The domain of the clustering function f is the finite set of samples $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ and a parameter κ (the number of target clusters) and the range is a set of partitions $f(\mathcal{D}, \kappa) := \{C_1, ..., C_\kappa\}$ of the index set $\{1, 2, ..., N\}$. Our goal is to find a clustering function f, which recovers the ground-truth partition. We call a clustering algorithm asymptotically consistent if it achieves this goal for long enough sequences $\mathbf{x}_i \in \mathcal{D}$, $i = 1, ..., N$: The following definitions represent the rigid formulation of the asymptotically consistent clustering.

Definition 2.3 (Consistency: offline settings)**.** *A clustering function f is consistent for a set of sequences $\mathcal{D}$ if $f(\mathcal{D}, \kappa) = \mathcal{G}$. Moreover, denoting by $n = min\{n_1, ..., n_N\}$, f is called strongly asymptotically consistent in the offline sense if with probability 1 $P(\exists n' \forall n > n' f(\mathcal{D}, \kappa) = \mathcal{G}) = 1$. We call it weakly asymptotically consistent if* $\lim_{n\to\infty} P(f(\mathcal{D}, \kappa) = \mathcal{G}) = 1$

It is worth noting that in the field of study of clustering of random processes, there is also another problem configuration, the online clustering of random processes. In the setting of the online problem, the number of realizations of random processes is not fixed, so the number of realizations and samples of each realization may

change over time. In this paper, our aim is to study the offline setting of random process clustering, we omit the definition of consistency of the online problem, referring the readers to the following works ([2], [3]).

As previously discussed, we are mainly focused on partitioning algorithms, which are based on dissimilarity measures. To construct an asymptotically consistent algorithm, we start by defining metric on ARMA processes.

Recalling the AR(∞) representation of the ARMA process, Piccolo in work [6] introduced metric on $\mathcal{L}$ as a measure of structural diversity between stochastic processes $X^{(1)}, X^{(2)} \in \mathcal{L}$. The metric function d_{PIC} on $\mathcal{L}$ is defined as

$$d_{PIC}(X^{(1)}, X^{(2)}) = \left\{ \sum_{j=0}^{\infty} (\pi_{1,j} - \pi_{2,j})^2 \right\}^{1/2} \tag{2.4}$$

where $\{\pi_{1,j}\}_{j=0}^{\infty}$ and $\{\pi_{2,j}\}_{j=0}^{\infty}$ is the $\boldsymbol{\pi}$ sequences for the $X^{(1)}$ and $X^{(2)}$ processes respectively. The d_{PIC} distance is well defined for all $X \in \mathcal{L}$ and can be computed even for processes with arbitrary orders and parameters. As for given ARMA process X the sequence $\{\pi_{x,j}\}_{j=0}^{\infty}$ fully characterize the forecasting function $\mathcal{F}$, therefore the defined distance between two ARMA processes, with given orders, is zero if, for the provided same set of initial values, the corresponding models produce the same forecasts [7]. Having this fact, if the $\mathbf{x}_i$ and $\mathbf{x}_j$ are two realizations of the two invertible ARMA processes $X^{(i)}$ and $X^{(j)}$, then if $i, j \in \mathcal{G}_k$ for some $k \in 1, ..., \kappa$, then corresponding distance between processes $d_{PIC}(X^{(i)}, X^{(j)}) = 0$.

We define the consistent estimator of the metric d as follows.

Definition 2.4. *We say that $\widehat{d}(\mathbf{x}_i, \mathbf{x}_j)$ is strongly asymptotically consistent if.*

$$\widehat{d}(\mathbf{x}_i, \mathbf{x}_j) \xrightarrow{a.s.} d(X^{(i)}, X^{(j)})$$

and weakly asymptotically consistent if.

$$\widehat{d}(\mathbf{x}_i, \mathbf{x}_j) \xrightarrow{P} d(X^{(i)}, X^{(j)})$$

as $n \to \infty, n = min\{n_i, n_j\}$.

The asymptotic consistency of the estimate $\widehat{d}(\mathbf{x}_i, X^{(j)})$ is defined by the same analogy. The estimator of the distributional distance between the processes defined in [2], and an estimator of the distance between covariance structures defined in work [3] are examples of strictly asymptotically consistent estimators. Later, in section 3.1 we will introduce an example of a weakly consistent estimate.

In [2] authors showed that a simple algorithm that initializes the clusters using farthest-point initialization and then assigns each remaining point to the nearest

cluster is strongly asymptotically consistent. This is done by using a strongly asymptotically consistent estimate of distributional distance [4].

3. Main results

3.1. **Consistent estimation of autoregressive metric.** In addition to the listed properties, we can show that d_{PIC} has a computationally efficient, weakly consistent estimator $\widehat{d}_{PIC}$. Let $X^{(1)}, X^{(2)} \in \mathcal{L}$ be two invertible ARMA processes, with (p_1,q_1), $\beta^1 = (\phi^1_1, \phi^1_2, ..., \phi^1_{p_1}, \theta^1_1, \theta^1_2, ..., \theta^1_{q_1})$, $\boldsymbol{\pi}_1 = \{\pi_{1,j}\}_{j=0}^{\infty}$ and (p_2,q_2), $\beta^2 = (\phi^2_1, \phi^2_2, ..., \phi^2_{p_2}, \theta^2_1, \theta^2_2, ..., \theta^j_{q_2})$, $\boldsymbol{\pi}_2 = \{\pi_{2,j}\}_{j=0}^{\infty}$ orders, parameter vectors and associated coefficients respectively.

Defined distance d_{PIC} is defined under the assumption that orders of the processes $X^{(1)}$ and $X^{(2)}$ are known. Thus, we have not considered incorrect model specification and unit root problems. Let us consider samples $\mathbf{x}_1 = \{x_1, x_2, .., x_{n_1}\}$ and $\mathbf{x}_2 = \{x_1, x_2, .., x_{n_2}\}$ generated from the $X^{(1)}$ and $X^{(2)}$ ARMA processes. If the orders of the processes $X^{(1)}$ and $X^{(2)}$ are known, then, as an estimate of the d_{PIC}, it is natural to consider the Euclidean distance between truncated sequences $\boldsymbol{\pi}_i$ $(i = 1, 2)$ of the estimated parameters with maximum likelihood. This approach is intuitive, but it limits us to apply the estimated distance to a clustering problem defined earlier since it is impractical to assume that the orders of all underlying processes are known. Considering the mentioned problem, for constructing an asymptotically consistent estimator of d_{PIC}, our estimation procedure needs to include both the consistent estimation of the orders and the parameters of the processes. Despite the fact that the consistent time series model selection literature is expansive, the majority of works are based on the maximization of the penalized log-likelihood (or quasi-log likelihood). The asymptotic results which include the ARMA processes can be found in ([8], [9]). To construct such a procedure, we refer to the latest asymptotic results in [8], where authors derive sufficient conditions for asymptotically consistent estimation of the orders and parameters of the large class of affine causal random processes. This class includes ARMA or AR(∞) processes, as well as the GARCH or ARCH(∞), APARCH, ARMA-GARCH, and many other stochastic processes. Let us assume that $X = \{X_t\}_{t=1}^{\infty}$ is an ARMA(p^*, q^*) model with parameter vector $\beta^* = (\phi_1, \phi_2, ..., \phi_{p^*}, \theta_1, \theta_2, ..., \theta_{q^*})$, and time series samples $\mathbf{x} = \{x_1, x_2, .., x_n\}$ are generated with the model. Let $\mathcal{M}$ be a finite set of ARMA model structures, where each model $m \in \mathcal{M}$ has parameter space

$$\Theta(m) = \{\beta[m] = (\phi_1, \phi_2, ..., \phi_{p_i}, \theta_1, \theta_2, ..., \theta_{q_i}), \;\; m = (p_i, q_i)\}.$$

We suppose that the target model structure $m^* = (p^*, q^*) \in \mathcal{M}$. Our goal is to construct a consistent estimation procedure for the target model structure m^* and the parameters vector β^* using the given trajectory $\mathbf{x}$ and the candidate models set $\mathcal{M}$. The following theorem shows the consistency of the quasi-log likelihood estimation [8].

Theorem 3.1 (Consistent estimation of the model)**.** *If $\mathbf{x} = \{x_1, x_2, .., x_n\}$ is a time series sample generated from the stochastic process X with model structure $m^* = (p^*, q^*)$ and parameter vector β^*, then under standard regularity assumptions the BIC penalized quasi-log likelihood estimation of the true model is asymptotically consistent.*

$$\widehat{m}, \widehat{\beta}[\widehat{m}] = \underset{m \in \mathcal{M}, \beta \in \Theta[m]}{\operatorname{argmin}} \left(-2\widehat{L}_n(\widehat{\beta}[m]) + (p_m + q_m) log(n) \right) \tag{3.1}$$

$$P(\widehat{m} = m^*) \xrightarrow[n \to \infty]{} 1, \quad \widehat{\beta}[\widehat{m}] \xrightarrow[n \to \infty]{P} \beta^*$$

where $\widehat{L}_n(\widehat{\beta}(m))$ is the quasi-log likelihood of the model m.

Suppose $X^{(1)}$ and $X^{(2)}$ are two ARMA processes with unknown orders and parameters and the samples $\mathbf{x}_1 = \{x_1^1, x_2^1, .., x_{n_1}^1\}$ and $\mathbf{x}_2 = \{x_1^2, x_2^2, .., x_{n_2}^2\}$ are generated from the $X^{(1)}$ and $X^{(2)}$ processes. Having Theorem 3.1 it is easy to construct a weakly consistent estimator for the d_{PIC}.

Let us denote the empirical estimates of d_{PIC} as follows

$$\widehat{d}_{PIC}(\mathbf{x}_1, \mathbf{x}_2) = \left\{ \sum_{j=1}^{\tau_n} \left(\widehat{\pi}_{1,j} - \widehat{\pi}_{2,j} \right)^2 \right\}^{1/2}$$

$$\widehat{d}_{PIC}(\mathbf{x}_1, X^{(1)}) = \left\{ \sum_{j=1}^{\tau_n} \left(\widehat{\pi}_{1,j} - \pi_{1,j} \right)^2 \right\}^{1/2}$$

where τ_n goes infinity with $min(n_1, n_2)$, $\{\widehat{\pi}_{i,j}\}_{j=1}^{\tau}$ are given by (2.3) and parameters vectors $\widehat{\beta^i}$ estimated by (3.1). Despite the fact that $\widehat{d}_{PIC}$ is a continuous function over estimated $\{\widehat{\pi}_{i,j}\}_{j=1}^{\tau}$ vectors, then by the continuous mapping theorem, the $\widehat{d}_{PIC}$ is weakly asymptotically consistent. The following proposition below concludes the discussion above.

Proposition 3.1. *If the maximum orders of the processes $X^{(1)}$ and $X^{(2)}$ are known, then under the standard regularity assumptions the estimator $\widehat{d}_{PIC}(\mathbf{x}_1, \mathbf{x}_2)$ and $\widehat{d}_{PIC}(\mathbf{x}_1, X^{(2)})$ are weakly asymptotically consistent.*

$$\widehat{d}_{PIC}(\mathbf{x}_1, \mathbf{x}_2) \xrightarrow{P} d_{PIC}(X^{(1)}, X^{(2)})$$

$$\widehat{d}_{PIC}(\mathbf{x}_1, X^{(2)}) \xrightarrow{P} d_{PIC}(X^{(1)}, X^{(2)})$$

as $n \to \infty$, $n := min\{n_i, n_j\}$.

It is a noteworthy observation that for any $X^{(i)}, X^{(j)} \in \mathcal{L}$ and $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}$ the distance d_{PIC} and their empirical estimate $\widehat{d}_{PIC}$ satisfy triangle equations.

$$
\begin{aligned}
& d_{PIC}\left(X^{(i)}, X^{(j)}\right) \leq \widehat{d}_{PIC}\left(X^{(i)}, \mathbf{x}_i\right) + \widehat{d}_{PIC}\left(\mathbf{x}_i, X^{(j)}\right) \\
& \widehat{d}_{PIC}\left(\mathbf{x}_i, X^{(i)}\right) \leq \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_j\right) + \widehat{d}_{PIC}\left(\mathbf{x}_j, X^{(i)}\right) \\
& \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_j\right) \leq \widehat{d}_{PIC}\left(\mathbf{x}_i, X^{(i)}\right) + \widehat{d}_{PIC}\left(\mathbf{x}_j, X^{(i)}\right)
\end{aligned} \tag{3.2}
$$

Algorithm 1 Clustering ARMA models

Require: $\mathcal{D}$, κ, (p_{max}, q_{max})
  Estimate $\widehat{m}^i$,$\widehat{\beta}^i$ ***and*** $\{\widehat{\pi}_{i,j}\}_{j=1}^{\tau}$ ***sequences:***
  for $i = 1..N$ **do**
    $\widehat{m}^i, \widehat{\beta}^i \leftarrow \underset{m=(p_k,q_k)\in[0,p_{max}]*[0,q_{max}]}{\operatorname{argmin}} \left(-2\widehat{L}_{n_i}(\widehat{\beta}[m]) + (p_k + q_k)log(n_i)\right)$
    $\{\widehat{\pi}_{i,j}\}_{j=1}^{\tau} \leftarrow$ *Compute truncation of* π *seq*
  end for
  Initialize κ***-farthest points as cluster-centres:***
  $c_1 \leftarrow 1$
  $C_1 \leftarrow \{c_1\}$
  for $k = 2..\kappa$ **do**
    $c_k \leftarrow \underset{i=1.N}{\operatorname{argmax}} \min_{j=1..k-1} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_j}\right)$ ▷ where ties are broken arbitrarily
    $C_k \leftarrow \{c_k\}$
  end for
  Assign the remaining points to closest centres:
  for $i = 1..N$ **do**
    $k \leftarrow \operatorname{argmin}_{j \in \bigcup_{k=1}^{\kappa} C_k} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_j\right)$
    $C_k \leftarrow C_k \cup \{i\}$
  end for
  OUTPUT: clusters $C_1, C_2, ..., C_k$

3.2. **Weakly Consistency of Algorithm 1.** In this chapter, we show the weak consistency of Algorithm 1. The Algorithm 1 for each sample $\mathbf{x}_i$ estimates ARMA models from the candidate models and computes the truncated $\boldsymbol{\pi}$ sequences for each estimated model. And then, the estimated $\widehat{d}_{PIC}$ is used in the algorithm proposed in [2]. The algorithm initializes the clusters using farthest-point initialization and then assigns each remaining point to the nearest cluster. As stated in Proposition 3.1, the consistent estimation of the ARMA model requires including the true model structure m^* in the candidate models. We can weaken this condition, by demanding that maximum orders of underlying κ ARMA processes in the ARMA dataset $\mathcal{D}$ need to be known.

Theorem 3.2. *Assuming that maximum orders (p_{max}, q_{max}) of underlying ARMA processes and the target number of clusters κ are known, then Algorithm 1 is weakly asymptotically consistent. Moreover, for the given $\eta \in (0,1)$ there exists n, such that if $n_{\min} = \min_{i \in 1..N} n_i > n$, then*

$$P(f((\mathcal{D}, \kappa)) = \mathcal{G}) \geq (1 - (N - \kappa)(4 - 4\eta))(4\eta - 3)^{\kappa - 1}$$

Proof. Let us fix $\eta \in (0,1)$. Denote by δ the minimum nonzero distance between the underlying unknown ARMA processes:

$$\delta := \min_{k \neq k' \in 1..\kappa} d_{PIC}\left(X^{(k)}, X^{(k')}\right)$$

Fix $\epsilon \in (0, \delta/4)$. The fact that the $\widehat{d}_{PIC}(\mathbf{x}_i, \mathbf{x}_j)$ distance estimate is weakly asymptotically consistent, implies for large enough n_{min} we can write

$$P(\max_{\substack{k \in 1..\kappa \\ i \in \mathcal{G}_k}} \widehat{d}_{PIC}\left(\mathbf{x}_i, X^{(k)}\right) \leq \varepsilon) > \eta \tag{3.3}$$

Having (3.3) and applying the triangle inequality for large enough n_{min} we have

$$\begin{aligned}
&P\left(\max_{\substack{k \in 1..\kappa \\ i,j \in \mathcal{G}_k}} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_j\right) < 2\varepsilon\right) \geq P\left(\max_{\substack{k \in 1..\kappa \\ i,j \in \mathcal{G}_k}} \widehat{d}_{PIC}\left(\mathbf{x}_i, X^{(k)}\right) + \widehat{d}_{PIC}\left(\mathbf{x}_j, X^{(k)}\right) < 2\varepsilon\right) \\
&\geq P\left(\{\max_{\substack{k \in 1..\kappa \\ i \in \mathcal{G}_k}} \widehat{d}_{PIC}\left(\mathbf{x}_i, X^{(k)}\right) < \varepsilon\} \cap \{\max_{\substack{k \in 1..\kappa \\ j \in \mathcal{G}_k}} \widehat{d}_{PIC}\left(\mathbf{x}_j, X^{(k)}\right) < \varepsilon\}\right) \\
&\geq P(\max_{\substack{k \in 1..\kappa \\ i \in \mathcal{G}_k}} \widehat{d}_{PIC}\left(\mathbf{x}_i, X^{(k)}\right) \leq \varepsilon) + P(\max_{\substack{k \in 1..\kappa \\ j \in \mathcal{G}_k}} \widehat{d}_{PIC}\left(\mathbf{x}_j, X^{(k)}\right) \leq \varepsilon) - 1 \geq 2\eta - 1
\end{aligned} \tag{3.4}$$

The inequality (3.4) states that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are samples from the same ground truth clusters then for large enough n_{min} we have $\widehat{d}_{PIC}(\mathbf{x}_i, \mathbf{x}_j) < 2\varepsilon$ with probability not less than $2\eta - 1$. Then having the definition of the δ, the inequality (3.4), and the triangle inequalities (3.2), we can find bounds for the distance between samples from the different clusters. In particular, for all large enough n_{min} we have

$$P\left(\min_{\substack{i \in \mathcal{G}_k \\ j \in \mathcal{G}_{k'} \\ k \neq k' \in 1..\kappa}} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_j\right) > \delta - 2\varepsilon\right) \geq \tag{3.5}$$

$$\geq P\left(\min_{\substack{i \in \mathcal{G}_k \\ j \in \mathcal{G}_{k'} \\ k \neq k' \in 1..\kappa}} \{d_{PIC}\left(X^{(k)}, X^{(k')}\right) - \widehat{d}_{PIC}\left(X^{(k)}, \mathbf{x}_i\right) -\right.$$

$$- \widehat{d}_{PIC}\left(X^{(k')}, \mathbf{x}_j\right)\} > \delta - 2\varepsilon\Big) \geq P\left(\max_{\substack{i\in\mathcal{G}_k \\ j\in\mathcal{G}_{k'} \\ k\neq k'\in 1..\kappa}} \{\widehat{d}_{PIC}\left(X^{(k)}, \mathbf{x}_i\right) + \right.$$

$$\left.\widehat{d}_{PIC}\left(X^{(k')}, \mathbf{x}_j\right)\} \leq 2\varepsilon\right) \geq 2\eta - 1.$$

Algorithm 1 initializes the clusters using farthest-point initialization $c_1 := 1$ and the c_k-th sample will be assigned with $c_k := \underset{i=1,N}{\operatorname{argmax}} \min_{j=1..k-1} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_j}\right)$ $k = 2, ..., \kappa$. To prove weak consistency first we need to show that assigned cluster centers $\mathbf{x}_{c_1}, \mathbf{x}_{c_2}, ..., \mathbf{x}_{c_\kappa}$ asymptotically are from different ground-truth clusters. Let us denote by $I(\mathbf{x}_i)$ the index of target cluster of the sample $\mathbf{x}_i$ and by $\widehat{I}(\mathbf{x}_i)$ the predicted index of cluster $\widehat{I}(\mathbf{x}_i) := \operatorname{argmin}_{j\in\bigcup_{k=1}^{\kappa} C_k} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_j\right)$ by Algorithm 1. If we denote the event of assigning first j clusters' centers with samples from different ground-truth clusters by $A(j) := (I(\mathbf{x}_{c_1}), I(\mathbf{x}_{c_2}), ..., I(\mathbf{x}_{c_j})$ are not equal), where $\mathbf{x}_{c_1}, \mathbf{x}_{c_2}, ..., \mathbf{x}_{c_j}$ are assigned by Algorithm 1, then from (3.4) and (3.5) we can estimate the probability that all κ clusters center is assigned correctly.

Firstly for $l < \kappa$,

(3.6)

$$P\left(\underset{i=1,N}{\operatorname{argmax}} \min_{j=1..l-1} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_j}\right) \notin \{I(\mathbf{x}_{c_1}), I(\mathbf{x}_{c_2}), ..., I(\mathbf{x}_{c_{l-1}}\} | A(l-1)\right) =$$

$$= P\left(\max_{i\notin\{\mathcal{G}_1,..,\mathcal{G}_l\}} \min_{j=1..l-1} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_j}\right) > \max_{i\in\{\mathcal{G}_1,..,\mathcal{G}_l\}} \min_{j=1..l-1} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_j}\right)\right) \geq$$

$$= P\left(\{\max_{i\notin\{\mathcal{G}_1,..,\mathcal{G}_l\}} \min_{j=1..l-1} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_j}\right) > \delta - 2\epsilon\} \cap \right.$$

$$\left. \cap \{\max_{i\in\{\mathcal{G}_1,..,\mathcal{G}_l\}} \min_{j=1..l-1} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_j}\right)\} \leq 2\epsilon\right)$$

$$\geq P\left(\max_{i\notin\{\mathcal{G}_1,..,\mathcal{G}_l\}} \min_{j=1..l-1} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_j}\right) > \delta - 2\epsilon\right) +$$

$$+ P\left(\max_{i\in\{\mathcal{G}_1,..,\mathcal{G}_l\}} \min_{j=1..l-1} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_j}\right) < 2\epsilon\right) - 1 \geq$$

$$\geq (2\eta - 1) + P\left(\max_{i\in\{\mathcal{G}_1,..,\mathcal{G}_l\}} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_{I(x_i)}}\right) < 2\epsilon\right) - 1 \geq 4\eta - 3.$$

The inequality (3.6) states that if samples $\mathbf{x}_{c_1}, \mathbf{x}_{c_2}, ..., \mathbf{x}_{c_{l-1}}$ are chosen from different clusters, then Algorithm 1 will assign as a l-th cluster center, sample from the different cluster with probability at least $4\eta - 3$ for $l = 2, ..., \kappa$. Then, from (3.6) we can estimate the probability that all assigned cluster centers are from different

clusters.

$$P(A(\kappa)) = P\left(\{I(\mathbf{x}_{c_\kappa}) \notin \{I(\mathbf{x}_{c_1}), I(\mathbf{x}_{c_2}), ..., I(\mathbf{x}_{c_{\kappa-1}})\}\} \cap A(\kappa-1)\right) \geq$$
$$\geq P\left(\{I(\mathbf{x}_{c_\kappa}) \notin \{I(\mathbf{x}_{c_1}), I(\mathbf{x}_{c_2}), ..., I(\mathbf{x}_{c_{\kappa-1}})\}\} | A(\kappa-1)\right) \cdot P\left(A(\kappa-1)\right) \geq$$
$$\geq (4\eta - 3) \cdot P(A(\kappa-1)) \geq (4\eta-3)^{\kappa-1} \tag{3.7}$$

The last inequality is true because of (3.6) and the fact that in the first step, $\mathbf{x}_{c_1}$ is chosen properly with probability 1.

To complete the proof we need to show the weak convergence of the clustering function. Having the $A(\kappa)$ event, we can define the indicator function of misclustering the sample $\mathbf{x}_i$.

$$W_i = \begin{cases} 0, & \text{if } I(\mathbf{x}_i) = I(\mathbf{x}_{c_{\widehat{I}(\mathbf{x}_i)}}) \\ 1, & \text{if } I(\mathbf{x}_i) \neq I(\mathbf{x}_{c_{\widehat{I}(\mathbf{x}_i)}}) \end{cases}$$

In other words, the value of the random variable W_i is 0 if the index of the target cluster of the sample $\mathbf{x}_i$ coincides with the index of the target cluster of the nearest centroid, and $W_i = 1$ otherwise.

If the first κ samples are from different clusters, then from (3.3) and (3.4), the probability of including $\mathbf{x}_i$ sample in the right cluster.

$$P(W_i = 0 | A(\kappa)) \geq P\left(\{\min_{\substack{c_j \in \{1,2,...,\kappa\} \\ I(\mathbf{x}_i) \neq I(x_{c_j})}} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_j}\right) > \delta - 2\epsilon\} \cap \right.$$
$$\left. \{\widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_{I(x_i)}}\right) < 2\epsilon\}\right) \geq P\left(\{\min_{\substack{j=1,N \\ I(\mathbf{x}_i) \neq I(\mathbf{x}_j)}} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_j}\right) > \delta - 2\epsilon\} \cap \right.$$
$$\left. \{\max_{\substack{j=1,N \\ I(\mathbf{x}_i) = I(\mathbf{x}_j)}} \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_{c_j}\right) < 2\epsilon\}\right) \geq 4\eta - 3. \tag{3.8}$$

If we denote by $W = \sum_{i=\kappa+1}^{N} W_i$ the number of misclustered samples, except the $\mathbf{x}_{c_1}, \mathbf{x}_{c_2}, ..., \mathbf{x}_{c_\kappa}$ then the probability of including all samples in their ground truth clusters, having that $\mathbf{x}_{c_1}, \mathbf{x}_{c_2}, ..., \mathbf{x}_{c_\kappa}$ samples are from different clusters can be estimated:

$$P(W = 0 | A(\kappa)) \geq 1 - \mathbb{E}[W | A(\kappa)] = 1 - \sum_{i=\kappa+1}^{N} \mathbb{E}[W_i | A(\kappa)] =$$
$$= 1 - \sum_{i=\kappa+1}^{N} P(W_i = 1 | A(\kappa)) \geq (1 - (N-\kappa)(4-4\eta)) \tag{3.9}$$

Therefore, for large enough n_{min}, the probability of recovering ground-truth clustering by Algorithm 1

$$P(f((\mathcal{D},\kappa)=\mathcal{G}))=P(\{W=0\}\cap\{A(\kappa)\})=P(W=0|A(\kappa))*P(A(\kappa))\geq$$
$$\geq(1-(N-\kappa)(4-4\eta))(4\eta-3)^{\kappa-1}.$$

Список литературы

[1] S. Aghabozorgi, S. A. Shirkhorshidi, T. Ying Wah, "Time-series clustering – A decade review", Jour. Information Systems, **53**, 16 – 38 (2015). DOI=10.1016/j.is.2015.04.007,

[2] A. Khaleghi, D. Ryabko, J. Mary, P. Preux, "Consistent algorithms for clustering time series", Journal of Machine Learning Research, **17(3)**, 1 – 32 (2016).

[3] Q. Peng, N. Rao, R. Zhao, "Covariance-based dissimilarity measures applied to clustering wide-sense stationary ergodic processes", Journ. Machine Learning, **108(12)**, 2159 — 2195 (2019). DOI=https://doi.org/10.1007/s10994-019-05818-x.

[4] R. Gray, Probability, Random Processes, and Ergodic Properties, Springer Verlag (1988).

[5] P. J. Brockwell, R. A. Davis, Introduction to Time Series and Forecasting, Springer (2002).

[6] D. Piccolo, "A distance measure for classifying Arima models", Journal of Time Series Analysis, **11(2)**, 153–164 (1990). DOI=https://doi.org/10.1111/j.1467-9892.1990.tb00048.x.

[7] M. Corduas and D. Piccolo, "Time series clustering and classification by the autoregressive metric", Computational Statistics Data Analysis, **52(4)**, 1860 -– 1872 (2008). DOI=https://doi.org/10.1016/j.csda.2007.06.001.

[8] , J.-M. Bardet, K. Kamila, and W. Kengne, "Consistent model selection criteria and goodness-of-fit test for common time series models", Electronic Journal of Statistics, **14(1)**, 2009 – 2052 (2020). DOI=https://doi.org/10.1214/20-ejs1709.

[9] E. J. Hannan, "The estimation of the order of an ARMA process", The Annals of Statistics, **8(5)**, 1071 – 1081 (1980). DOI=https://doi.org/10.1214/aos/1176345144.

Поступила 19 октября 2022

После доработки 10 января 2023

Принята к публикации 20 февраля 2023