Clustering piecewise stationary processes

Azadeh Khaleghi
Department of Mathematics & Statistics
Lancaster University, Lancaster, UK
a.khaleghi@lancaster.ac.uk

Daniil Ryabko
Fishlife Research
daniil@ryabko.net

***Abstract*—The problem of time-series clustering is considered in the case where each data-point is a sample generated by a piecewise stationary process. While stationary processes comprise one of the most general classes of processes in nonparametric statistics, and in particular, allow for arbitrary long-range dependencies, their key assumption of stationarity remains restrictive for some applications. We address this shortcoming by considering piecewise stationary processes, studied here for the first time in the context of clustering. It turns out that this problem allows for a rather natural definition of consistency of clustering algorithms. Efficient algorithms are proposed which are shown to be asymptotically consistent without any additional assumptions beyond piecewise stationarity. The theoretical results are complemented with experimental evaluations.**

***Index Terms*—stationary ergodic processes, unsupervised learning, clustering, consistency**

I. Introduction

Clustering involves breaking a dataset into disjoint subsets called clusters where the elements within the same cluster are somehow more similar to each other than to those in other clusters. This task is meant to help with making sense of the data that typically have complex structures and represent some unknown underlying phenomena to be inferred. Given the nature of the problem, it is desirable to make as little assumptions as possible about the underlying mechanisms that generate the data. We consider a setting where each data-point is a time series. Such sequential data are ubiquitous in modern applications involving, for example, user behaviour, social networks, as well as financial or biological data. The common features in these datasets are the abundance of data and the absence of precise models.

Statistical inference concerning time series is typically performed under the assumption that the observations are i.i.d, or that their distribution belongs to a specific model class. However, such assumptions undermine the possibly complex nature of the data which may possess long-range dependencies.

To address this shortcoming, one approach is to assume that the process distributions are stationary without requiring any conditions to hold on their memory. This allows for arbitrary long-range dependencies between the observations. Moreover, thanks to Birkhoff's ergodic theorem, under this assumption alone, the frequency of occurrence of events converge almost surely to their underlying probabilities, even though there is no guarantee on the speed of convergence. Various problems of statistical inference can be solved under this assumption alone [1]. In particular, [2] suggested to cluster stationary ergodic time-series samples based on the distribution that generates them, putting together those and only those samples whose distribution is the same. Making use of the fact that in this setting the target clustering has the so-called strict separation property see [3], it was shown that asymptotically consistent clustering is achievable under the assumption of stationarity alone.

While already a weak assumption, stationarity often breaks in applications. Simple real-world examples concerning user behaviour that exhibit this property include such events as changing a job, a mobile phone, or having a child. Typical past events may stop happening and events of new kind start occurring. In these cases, it is still possible to measure frequencies of events in-between the changes.

To allow for these generalizations, we introduce a setting where each time series is generated by a piecewise stationary process, so that each time series can broken into stationary segments. The segments' boundaries are arbitrary and unknown, as are the stationary distributions that generate the data within each segment. The only requirement imposed is stationarity. Thus, the data within each segment are not assumed to possess any independence, finite-memory or mixing properties.

A piecewise stationary distribution is identified with a "bag of distributions" corresponding to the distributions of the stationary segments. Thus locations of distributional change, as well as the order of the stationary distributions, are disregarded. As a result, two piecewise stationary distributions are considered equivalent if the set of stationary distributions of their segments coincide. The *clustering objective is to put together those and only those time-series samples whose distributions are equivalent in this sense.* An algorithm that achieves this objective is said to be consistent.

The *main result*, provided in Section IV, is an algorithm that, as we show, is asymptotically consistent under the only assumption that each time series segment is stationary ergodic. This algorithm relies on a novel distance between equivalence classes of piecewise stationary distributions. The distance is based on minimax distances between the distributions that generate the stationary segments. We show that this distance can be estimated consistently based on samples. This latter result in itself can be useful in future research concerning inference for piecewise stationary processes. In order to establish our results, we require a careful consideration of the joint distribution of piecewise stationary samples along with a rigorous formulation of the clustering problem. This is particularly important since, unlike in the vast majority of the literature on time series, in our setting different time-series samples are allowed to be dependent. The necessary formalism is introduced in Section III. The clustering algorithm proposed generalizes those on clustering stationary time series [2], [4]. It uses as sub-routines the algorithms for changepoint analysis developed in [5], [6]. Our results are theoretical, and their main appeal is in their generality. Yet, the proposed methods are shown to be computationally feasible. Experimental evaluations are provided in Section V.

While the literature on the related topic of changepoint analysis is vast, the existing work, with the exception of [1], [5]–[8], is mostly concerned with independent or mixing data, and also restricts the nature of the changes to single-dimensional marginals. To our knowledge, there are no prior attempts to consider piecewise stationary distributions in the context of clustering. A related problem outside of changepoint analysis that has been considered previously, albeit under much more restrictive assumptions, is that of prediction. For example, [9] considers time-series prediction concerning piecewise i.i.d. processes; see also [10] and references.

II. Preliminaries

We use the abbreviation $u..v$ for $\{u,\dots,v\},\ u \le v \in \mathbb{N}$. Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be a measurable space. In this work we let $\mathcal{X} = [0,1]$. However, the results can be readily generalized to $\mathbb{R}$.

Denote by $\Delta_{u,v} := \{[\frac{i_1}{2^v}, \frac{i_1+1}{2^v}) \times \cdots \times [\frac{i_u}{2^v}, \frac{i_u+1}{2^v}) : i_j \in 0..2^v-1,\ j \in 1..u\}$ the set of dyadic cubes in $\mathcal{X}^u,\ u \in \mathbb{N}$ of side-length 2^{-v}, and let $\mathcal{B}^{(u)} := \sigma(\{\Delta_{u,v},\ v \in \mathbb{N}\})$ be the Borel subsets of $\mathcal{X}^u,\ u \in \mathbb{N}$. Let $\mathcal{X}^{\mathbb{N}}$ be the set of all $\mathcal{X}$-valued infinite sequences equipped with the Borel σ-algebra $\mathcal{B} := \sigma(\{B \times \mathcal{X}^{\mathbb{N}} : B \in \Delta_{u,v},\ u,v \in \mathbb{N}\})$. Stochastic processes are probability measures on $(\mathcal{X}^{\mathbb{N}}, \mathcal{B})$. Take a sequence of random variables $\mathbf{x} := \langle X_t \rangle_{t\in\mathbb{N}}$ with joint distribution μ where for every $t \in \mathbb{N}$, $X_t : \mathcal{X}^{\mathbb{N}} \to \mathcal{X}$ is the coordinate projection of $\mathbf{a} := \langle a_t \rangle_{t\in\mathbb{N}} \in \mathcal{X}^{\mathbb{N}}$ onto its t^{th} element, i.e. $X_t(\mathbf{a}) = a_t$. For each $n \in \mathbb{N}$ and $B \in \mathcal{B}^{(u)},\ u \in \mathbb{N}$ define the empirical measure $\mu_n(\mathbf{x}, B) : \mathcal{X}^{\mathbb{N}} \to [0,1],\ n \in \mathbb{N}$ of B as $\mu_n(\mathbf{x}, B) := \frac{1}{n-u+1}\sum_{i=1}^{n-u+1} \mathbb{I}\{X_{i..i+u} \in B\}$ for $n \ge u$ and 0 otherwise, where $\mathbb{I}$ is the indicator function.

Definition 1: A process μ is stationary if $\mu(X_{1..u} \in B) = \mu(X_{1+j..u+j} \in B)$ for all $B \in \mathcal{B}^{(u)},\ u \in \mathbb{N}$ and $j \in \mathbb{N}$. A stationary process μ with corresponding sequence of random variables $\mathbf{x} = \langle X_t \rangle_{t\in\mathbb{N}}$ is (stationary) ergodic if for every $u \in \mathbb{N}$ and $B \in \mathcal{B}^{(u)}$ it holds that $\lim_{n\to\infty} \mu_n(\mathbf{x}, B) = \mu(B),\ \mu-$ a.s. This is equivalent to the standard definition involving triviality of invariant measurable sets, e.g. [11].

Joint process distributions. We simultaneously consider multiple samples $X_{1..n},\ n \in \mathbb{N}$ generated by different, possibly dependent stationary ergodic processes. To allow for this, we first define a distribution over a matrix of random variables, each row of which shall correspond to one of the samples. Next, we obtain each process as the *marginal* distribution of the corresponding row of the matrix. We have the following formulation. For a fixed $m \in \mathbb{N}$, let ρ be a measure on the space $(\mathcal{X}^{m\times\mathbb{N}}, \mathcal{B}^{\otimes m})$ where, $\mathcal{B}^{\otimes m} := \sigma(\{B_1 \times \cdots \times B_m : B_i \in \mathcal{B},\ i \in 1..m\})$. Define the matrix of $\mathcal{X}$-valued random variables $\mathbf{X} := \big(X_{i,j}\big)_{i\in 1..m, j\in\mathbb{N}}$ where $X_{i,j} : \mathcal{X}^{m\times\mathbb{N}} \to \mathcal{X},\ i \in 1..m, j \in \mathbb{N}$ are jointly distributed according to ρ, so that for $B \in \mathcal{B}^{\otimes m}$ we have $\Pr(\mathbf{X} \in B) = \rho(B)$. For each $i \in 1..m$, let $\mathbf{x}_i := \langle X_{i,j} \rangle_{j\in\mathbb{N}}$ and define the projection map $\pi_i \mapsto \mathbf{x}_i$. The marginal distribution μ_i of $\mathbf{x}_i$ is then defined as the distribution induced by ρ over the i^{th} row, i.e. $\mu_i := \rho \circ \pi_i^{-1}$. We denote by $\mathcal{M}(\rho) := \{\mu_i :\ i \in 1..m\}$ the set of marginal process distributions of ρ.

Definition 2 ([11]): A *distributional distance* between a pair of processes μ, μ' is defined as $d(\mu,\mu') := \sum_{u,v\in\mathbb{N}} w_u w_v \sum_{B\in\Delta_{u,v}} |\mu(B) - \mu'(B)|$ where $w_j = 2^{-j}$, $j \in \mathbb{N}$, or a summable sequence of positive weights.

Definition 3: Consider a pair of marginals μ and $\mu' \in \mathcal{M}(\rho)$ with corresponding sequence of random variables $\mathbf{x}$ and $\mathbf{x}'$ respectively, where $\mu := \mu_i,\ \mu' := \mu_j$, and $\mathbf{x} := \langle X_{i,t} \rangle_{t\in\mathbb{N}},\ \mathbf{x}' := \langle X_{j,t} \rangle_{t\in\mathbb{N}}$ correspond to rows $i, j \in 1..m$ of $\mathbf{X}$. An empirical estimate of $d(\mu,\mu')$ can be given by $\widehat{d}_n(\mathbf{x},\mathbf{x}') := \sum_{u,v\in\mathbb{N}} w_u w_v \sum_{B\in\Delta_{u,v}} |\mu_n(\mathbf{x},B) - \mu_n(\mathbf{x}',B)|$ with $w_j,\ j \in \mathbb{N}$ as in Definition 2. Note that $\widehat{d}_n$ can be efficiently calculated with computational complexity $\mathcal{O}(n \log n)$ for $u_n := \log n,\ v_n := -\log(s_{\min})$, where $s_{\min}$ is the minimal non-zero difference between the union of all the elements of the two sequences $\mathbf{x}, \mathbf{x}'$, see [4].

Proposition 1 ([4]): If the marginals in $\mathcal{M}(\rho)$ are stationary ergodic, then $\lim_{n\to\infty} \widehat{d}_n(\mathbf{x}_s, \mathbf{x}_t) = d(\mu_s, \mu_t),\ \rho-$ a.s., for any $\mu \in \mathcal{M}(\rho)$ and $s,\ t \in 1..m$, where $\mathbf{x}_j := \langle X_{j,t} \rangle_{t\in\mathbb{N}}$ correspond to j^{th} row of $\mathbf{X}$ above and $\mu_j \in \mathcal{M}(\rho),\ j = s, t$.

III. Problem formulation

Piecewise stationary processes. We shall be dealing with multiple samples $\mathbf{y}$ of the form

$$Y_1, \dots, Y_{\tau_1}, Y_{\tau_1+1}, \dots, Y_{\tau_2}, \dots, Y_{\tau_\kappa}, \dots, Y_n, \tag{1}$$

where the (stationary) segments $Y_{\tau_i}, \dots, Y_{\tau_{i+1}}$ are generated by different, possibly dependent, stationary ergodic processes. To specify the distribution of the sample $\mathbf{y}$ we define a distribution on a matrix of random variables, each row of which shall correspond to a stationary segment of the sample. The set of *stationary-segment distributions* of (1) is the set of the distributions of the rows.

More formally, we specify a *Piecewise Stationary Process* as follows. Consider a measure ρ on $(\mathcal{X}^{\kappa\times\mathbb{N}}, \mathcal{B}^{\otimes\kappa+1})$ for some fixed $\kappa \in \mathbb{N}$ with set of marginals $\mathcal{M}(\rho) = \{\mu_i,\ i \in 1..\kappa+1\}$, where $\mu_i \neq \mu_{i+1},\ i \in 1..\kappa$ are stationary ergodic. Fix some $n \in \mathbb{N}$ and a sequence $\boldsymbol{\tau} := \langle \tau_i \rangle_{i\in 1..\kappa}$ with $\tau_1 < \tau_2 < \cdots < \tau_\kappa \in 1..n$. Define the mapping $c : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ as $c(j) \mapsto (t^*(j)+1, j - \tau_{t^*(j)})$ where $t^*(j) := \max_{i\in 0..\kappa+1} \tau_i \le j$ picks out the changepoint τ_i that is closest to $j \in \mathbb{N}$ from the left, with the convention that $\tau_0 := 0$ and $\tau_{\kappa+1} := n$. A *Piecewise Stationary Sample* of the form (1) generated by $(\rho, \boldsymbol{\tau})$ can be specified as a sequence of coordinate projections $Y_t : \mathcal{X}^n \to \mathcal{X},\ t \in 1..n$ such that for any $\ell \in 1..n,\ t_1, \dots, t_\ell \in 1..n$ and $B_i \in \mathcal{B}_{\mathcal{X}},\ i \in 1..\ell$ it holds that $\Pr(Y_{t_1} \in B_1, \dots, Y_{t_\ell} \in B_\ell) = \rho(X_{c(t_1)} \in$

$B_1, \dots, X_{c(t_\ell)} \in B_\ell)$. Thus, the distribution of each segment $Y_{\tau_i+1..\tau_{i+1}}$ is given by a stationary ergodic process $\mu_i,\ i \in 1..\kappa+1$. Since it is assumed that $\mu_i \neq \mu_{i+1},\ i \in 1..\kappa$, the indices $\tau_i,\ i \in 1..\kappa$ are called *changepoints*. The pair $(\rho, \boldsymbol{\tau})$ composed of the measure ρ and its corresponding sequence of changepoints $\boldsymbol{\tau}$ defines a piecewise stationary process.

Definition 4: A pair of piecewise stationary processes $(\rho, \boldsymbol{\tau})$ and $(\rho', \boldsymbol{\tau}')$ are considered equivalent if and only if they agree on their set of stationary-segment distributions, i.e.,

$$(\rho, \boldsymbol{\tau}) \sim (\rho', \boldsymbol{\tau}') \Leftrightarrow \mathcal{M}(\rho) = \mathcal{M}(\rho'). \tag{2}$$

Let $\mathcal{P}$ denote the set of all piecewise stationary processes. The equivalence relation defined above induces a partitioning of $\mathcal{P}$ into distinct classes $[(\rho, \boldsymbol{\tau})],\ (\rho, \boldsymbol{\tau}) \in \mathcal{P}$ where $[(\rho, \boldsymbol{\tau})] := \big\{(\rho', \boldsymbol{\tau}') \in \mathcal{P} : (\rho', \boldsymbol{\tau}') \sim (\rho, \boldsymbol{\tau})\big\}$, so that two piecewise stationary processes belong to the same class if and only if they are equivalent in the sense of (2). Let $\mathcal{C} := \{[(\rho, \boldsymbol{\tau})] : (\rho, \boldsymbol{\tau}) \in \mathcal{P}\}$ be the set of all such classes.

Clustering Problem. Fix some $N \in \mathbb{N}$, which is the number of samples, and (unknown) sequence $\kappa_i \in \mathbb{N},\ i \in 1..N$ corresponding to the number of changepoints in each sample. Moreover, define (unknown) increasing sequences $\boldsymbol{\theta}_i := \langle \theta_j^{(i)} \rangle_{j \in 1..\kappa_i+1}$ with $\theta_1^{(i)} < \cdots < \theta_{\kappa_i+1}^{(i)} \in (0,1),\ i \in 1..N$. For any $n \in \mathbb{N}$, define the sequence $\boldsymbol{\tau}_i(n) := \langle \tau_j^i(n) \rangle_{j \in 1..\kappa_i},\ i \in \mathbb{N}$ where $\tau_j^i(n) := \lfloor n\theta_j^{(i)} \rfloor$, with the convention that $\theta_0^{(i)} = 0$ for all $i \in 1..N$.

The problem is formulated as follows. For a fixed $n \in \mathbb{N}$, we are given a set

$$\mathcal{S}(n) := \{\mathbf{y}_1, \dots, \mathbf{y}_N\} \tag{3}$$

of N piecewise stationary samples of the form (1), each of length $n_i := \lfloor n\theta_{\kappa_i+1} \rfloor$ generated by an unknown piecewise stationary process $(\rho_i, \boldsymbol{\tau}_i(n)),\ i \in 1..N$. Thus, each sample $\mathbf{y}_i,\ i \in 1..N$ has κ_i changepoints $\tau_j^i(n),\ j \in 1..\kappa_i$. It is assumed that each of N piecewise stationary processes that generate the samples belongs to one of m distinct classes $C_1, \dots, C_m \in \mathcal{C}$, which are unknown. Define the normalized minimum separation between the changepoints as

$$\alpha := \min_{i \in 1..N} \min_{k \in 1..\kappa_i+1} \theta_j^{(i)} - \theta_{j-1}^{(i)}. \tag{4}$$

We assume $\alpha > 0$, i.e. the segments are at least $n\alpha$ long.

Definition 5 (Ground-Truth Clustering): Let $\mathcal{G} := \{\mathcal{G}_1, \dots, \mathcal{G}_m\}$ be a partitioning of $1..N$ where for any $i \in 1..N$, it holds that $i \in \mathcal{G}_\ell$ for some $\ell \in 1..m$ if and only if $(\rho_i, \boldsymbol{\tau}_i(n)) \in C_\ell$. We call $\mathcal{G}$ the ground-truth clustering.

Thus, samples fall into the same *ground-truth* cluster if and only if their corresponding piecewise-stationary distributions are equivalent in the sense that they have the same set of stationary-segment distributions.

A clustering function f takes a set $\mathcal{S}$ of samples and the number m of target clusters to produce a partition $f(\mathcal{S}, m) \mapsto \{J_1, \dots, J_m\}$ of $1..N$, aiming to recover the ground-truth $\mathcal{G}$.

Definition 6 (Consistency): A clustering function f is consistent for a set of samples $\mathcal{S} = \mathcal{S}(n),\ n \in \mathbb{N}$ if $f(\mathcal{S}, m) = \mathcal{G}$. Moreover, f is called asymptotically consistent if with probability 1 it holds that $\lim_{n\to\infty} f(\mathcal{S}(n), m) = \mathcal{G}$.

Joint distribution of piecewise stationary samples. Observe that the problem requires us to simultaneously consider multiple samples, each generated by a piecewise stationary process. These samples can themselves be dependent. Formally, this is defined through the following construction. Consider the space $\mathcal{Y} := \mathcal{X}^{\kappa_1 \times \mathbb{N}} \times \cdots \times \mathcal{X}^{\kappa_N \times \mathbb{N}}$. Denote by $\mathcal{F} := \mathcal{B}_1 \otimes \cdots \otimes \mathcal{B}_N$ the product σ-algebra on $\mathcal{Y}$ where $\mathcal{B}_i := \sigma(\{B_1 \times \cdots \times B_{\kappa_i} : B_j \in \mathcal{B}\}),\ i \in 1..N$ is in turn the product σ-algebra on $\mathcal{X}^{\kappa_i \times \mathbb{N}}$. Let P be a probability measure on $(\mathcal{Y}, \mathcal{F})$. Consider a sequence $\mathbf{Z} := \langle \mathbf{Y}_i \rangle_{i \in 1..N}$ of infinite matrices of $\mathcal{X}$-valued random variables $Y_{s,t}^{(i)} : \mathcal{X}^{\kappa_i \times \mathbb{N}} \to \mathcal{X},\ s \in 1..\kappa_i,\ t \in \mathbb{N}, i \in 1..N$, which can be easily shown to be $\mathcal{F}$-measurable. Suppose that P is the distribution of $\mathbf{Z}$ so that $\Pr(\mathbf{Z} \in F) = P(F)$ for all $F \in \mathcal{F}$. For each $i \in 1..N$, define the projection $\widetilde{\pi}_i \mapsto \langle Y_{s,t}^{(i)} \rangle_{s \in 1..k_i,\ t \in \mathbb{N}}$. Then $\rho_i := P \circ \widetilde{\pi}_i^{-1},\ i \in 1..N$ is the measure of $\mathbf{Y}_i$. Our statements are made in terms of P.

IV. Main Results

This section outlines our main results, with a focus on describing the proposed algorithm and explaining how and why it works. We refer to the longer version of the paper [12] for more detailed arguments and technical proofs.

We start by introducing a distance between pairs of equivalence classes of piecewise stationary distributions. Since these classes are distinguished by their sets of stationary-segment distributions, it is natural to require the distance between any pair of equivalence classes to be 0 if and only if they agree on these distributions. This leads to the following definition.

Definition 7: Let $C = [(\rho, \boldsymbol{\tau})]$ and $C' = [(\rho', \boldsymbol{\tau}')] \in \mathcal{C}$ be two classes of piecewise stationary processes. We define a distance between C and C' as

$$\delta(C, C') = \max_{\mu \in \mathcal{M}(\rho)} \min_{\mu' \in \mathcal{M}(\rho')} d(\mu, \mu') + \max_{\mu' \in \mathcal{M}(\rho')} \min_{\mu \in \mathcal{M}(\rho)} d(\mu', \mu)$$

where, $d(\cdot, \cdot)$ is given by Definition 2.

Proposition 2: The distance δ induces a metric on the set of equivalence classes of piecewise stationary processes.

See the longer version of the paper [12] for a proof.

Next, we present Algorithm 1 that estimates δ from piecewise stationary samples. The output is used in the clustering algorithm, namely, Algorithm 2. In order to estimate δ, Algorithm 1 relies on a so-called list-estimator, defined below, which is a function that takes a piecewise stationary sample with κ changepoints and produces an exhaustive list of at least κ candidate estimates.

Definition 8 (List-estimator): Given a parameter $\lambda \in (0,1)$, define a list-estimator as a function $\mathcal{L}_\lambda : \bigcup_{j \in \mathbb{N}} \mathcal{X}^j \to \mathbb{N}^{\lfloor 1/\lambda \rfloor}$ that takes any $\mathbf{x} \in \mathcal{X}^n,\ n \in \mathbb{N}$ and produces a list $\{\psi_i \in 1..n :\ i \in 1..\lfloor 1/\lambda \rfloor\}$ of indices to correspond to candidate changepoint estimates in $\mathbf{x}$.

Definition 9 (Consistent List-Estimator): Consider a sequence $\theta_1 < \cdots < \theta_\kappa \in (0,1)$ for some fixed $\kappa \in \mathbb{N}$. Let $\mathbf{y} := Y_{1..n},\ n \in \mathbb{N}$ be a sample of the form (1), generated by a piecewise stationary process $(\rho, \boldsymbol{\tau}(n))$, where $\boldsymbol{\tau}(n) := \langle n\theta_i \rangle_{i \in 1..\kappa}$.

Algorithm 1 Calculating an empirical estimate of δ

```
INPUT: $\mathbf{y} \in \mathcal{X}^{n_1},\ \mathbf{y}' \in \mathcal{X}^{n_2},\ \lambda \in (0,1)$
Obtain a sequence of candidate changepoints in $\mathbf{y}$ and $\mathbf{y}'$ respectively, using the method of [5].
    $\widehat{\boldsymbol{\tau}} \leftarrow \mathcal{L}_\lambda(\mathbf{y}) \quad \text{and} \quad \widehat{\boldsymbol{\tau}}' \leftarrow \mathcal{L}_\lambda(\mathbf{y}') \qquad (5)$
Generate sets $\mathcal{U}$ and $\mathcal{U}'$ of consecutive stationary-segments corresponding to $\mathbf{y}$ and $\mathbf{y}'$.
    $\mathcal{U} \leftarrow \{\overline{\mathbf{y}}_i := \mathbf{y}_{\psi_{i-1}..\psi_i},\ i \in 1..|\widehat{\boldsymbol{\tau}}|+1 : \langle\psi_i\rangle_{i\in 1..|\widehat{\boldsymbol{\tau}}|} = \widehat{\boldsymbol{\tau}},\ \psi_0 := 1,\ \psi_{|\widehat{\boldsymbol{\tau}}|+1} := n_1\} \qquad (6)$
    $\mathcal{U}' \leftarrow \{\overline{\mathbf{y}}'_i := \mathbf{y}_{\psi'_{i-1}..\psi'_i},\ i \in 1..|\widehat{\boldsymbol{\tau}}'|+1 : \langle\psi'_i\rangle_{i\in 1..|\widehat{\boldsymbol{\tau}}'|} = \widehat{\boldsymbol{\tau}}',\ \psi'_0 := 1,\ \psi'_{|\widehat{\boldsymbol{\tau}}'|+1} := n_2\} \qquad (7)$
Calculate an empirical estimate of $\delta$.
    $n \leftarrow \min\{\lambda n_1, \lambda n_2\}$
    $\delta(\mathbf{y},\mathbf{y}',\lambda) \leftarrow \max_{\overline{\mathbf{y}}\in\mathcal{U}} \min_{\overline{\mathbf{y}}'\in\mathcal{U}'} \widehat{d}_n(\overline{\mathbf{y}},\overline{\mathbf{y}}') + \max_{\overline{\mathbf{y}}'\in\mathcal{U}'} \min_{\overline{\mathbf{y}}\in\mathcal{U}} \widehat{d}_n(\overline{\mathbf{y}}',\overline{\mathbf{y}}) \qquad (8)$
OUTPUT: $\delta(\mathbf{y},\mathbf{y}',\lambda)$
```

Algorithm 2 Clustering piecewise stationary samples

```
INPUT: sequences $\mathcal{S} := \{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$, number $m$ of target clusters, parameter $\lambda$
Initialize $m$ points as cluster-centres
$c_1 \leftarrow 1$
$C_1 \leftarrow \{c_1\}$
for $\ell = 2..m$ do
    $c_\ell \leftarrow \min\{\text{argmax}_{i=1..N} \min_{j=1..l-1} \delta(\mathbf{y}_i, \mathbf{y}_{c_j}, \lambda)\}$, where $\delta$ is given by Algorithm 1
    $C_\ell \leftarrow \{c_\ell\}$
Assign the remaining points to appropriate clusters:
for $i = 1..N$ do
    $k \leftarrow \text{argmin}_{j\in\bigcup_{\ell=1}^m C_\ell} \delta(\mathbf{y}_i, \mathbf{y}_j, \lambda)$
    $C_\ell \leftarrow C_\ell \cup \{i\}$
OUTPUT: clusters $C_1, C_2, \cdots, C_m$
```

Denote by $\psi_1(n) \leq \cdots \leq \psi_{\lfloor 1/\lambda\rfloor}(n)$ the candidate estimates produced by a list-estimator $\mathcal{L}_\lambda(\mathbf{y})$ for some $\lambda \in (0, \lambda^*]$ where $\lambda^* := \min_{i\in 1..\kappa+1} \theta_i - \theta_{i-1}$ with $\theta_0 := 0,\ \theta_{\kappa+1} := 1$, is the minimum normalized distance between the changepoints of $\mathbf{y}$. We say that $\mathcal{L}_\lambda$ is consistent if with probability 1 it holds that $\lim_{n\to\infty} \max_{k\in 1..\kappa} \min_{i\in 1..\lfloor 1/\lambda\rfloor} |\frac{1}{n}\psi_i(n) - \theta_i| = 0$ and $\min_{i\in 1..\lfloor 1/\lambda\rfloor+1} \psi_i(n) - \psi_{i-1}(n) \geq n\lambda,\ \psi_0 := 0$.

In words, a *consistent* list-estimator takes a piecewise stationary sample of length $n \in \mathbb{N}$ whose stationary segments are of lengths $\mathcal{O}(n)$ and, for large enough n, produces a list of candidate estimates which contains a good approximation for each of the true changepoints. Note that it is not required for a list-estimator to find the number of changepoints κ, but among the candidate estimates that it outputs there should be κ estimates corresponding to the true, unknown changepoints.

An example of a consistent list-estimator is provided in [5]. Note that the algorithm in [5] establishes a stronger property than that required by Definition 9: specifically, it sorts the list in such a way that its first κ elements estimate the true changepoints of $\mathbf{y}$. We shall not require this feature here: it is enough to have a list of *arbitrary order* that includes a correct estimate for each changepoint. For simplicity, assume that the candidate estimates are sorted in increasing order.

Given two piecewise stationary samples $\mathbf{y}$ and $\mathbf{y}'$ and a parameter $\lambda \in (0,1)$ to specify a lower-bound on the minimum normalized length of the stationary segments, Algorithm 1 works as follows. First, a consistent list-estimator, e.g. that of [5], is applied to each sample to identify a set of stationary segments in each. An empirical estimate of δ is then obtained as a minimax empirical distributional distance $\widehat{\delta}$ given by (8) between the stationary segments identified. As follows from Proposition 3 below, Algorithm 1 can consistently estimate δ.

Proposition 3 (δ can be estimated consistently.): Consider the samples $\mathbf{y}, \mathbf{y}'$ generated by a distribution P with piecewise stationary marginals $(\rho, \boldsymbol{\tau})$ and $(\rho', \boldsymbol{\tau}')$; the lengths of the samples are parameterized by n. Let the estimate $\widehat{\delta}_n(\mathbf{y},\mathbf{y}') := \delta(\mathbf{y},\mathbf{y}',\lambda)$ be obtained as the output of Algorithm 1 when provided with $\mathbf{y},\ \mathbf{y}'$ and any $\lambda \in (0, \alpha]$ as input, where α is given by (4). Then $\lim_{n\to\infty} \widehat{\delta}_n(\mathbf{y},\mathbf{y}') = \delta(C, C'), \quad P-$ a.s., where $C := [(\rho, \boldsymbol{\tau})]$ and $C' := [(\rho, \boldsymbol{\tau}')]$ are the equivalence classes containing the processes that generate $\mathbf{y}$ and $\mathbf{y}'$ respectively. *See the longer version of the paper [12] for a proof.*

A key reason why this result holds is that for a large enough n the (consistent) list-estimator closely approximates *all* of the change-points in $\mathbf{y}$, partitioning it into a set $\mathcal{U}$ of (mostly) stationary segments. Thus it holds that $\lim_{n\to\infty} \max_{\overline{\mathbf{y}}\in\mathcal{U}} \widehat{d}_n(\overline{\mathbf{y}}, \mu^*(\overline{\mathbf{y}})) = 0$, and at least one sample per each of the stationary marginals of $\mathbf{y}$ is present in $\mathcal{U}$. An analogous argument can be given for $\mathbf{y}'$. Putting these together, Proposition 3 can be shown. The estimates of δ are in turn used to construct a clustering method outlined in Algorithm 2. The algorithm starts by initializing the clusters using farthest-point initialization of [13], and then assigns the remaining samples to the nearest cluster. Theorem 1 establishes that the algorithm is asymptotically consistent in the sense of Definition 6.

Theorem 1 (Algorithm 2 is asymptotically consistent.): Let $f(\mathcal{S}(n), m) = \text{Alg}\,2(\mathcal{S}(n), m, \lambda)$ be the output of Algorithm 2 when provided with the set $\mathcal{S}(n)$ of piecewise stationary samples (3), along with the correct number m of target clusters and some $\lambda \in (0, \alpha]$, where α is given by (4). It holds that $\lim_{n\to\infty} f(\mathcal{S}(n), m) = \mathcal{G},\ P-$ a.s., i.e. Algorithm 2 is asymptotically consistent. The computational complexity of Algorithm 2 is $\mathcal{O}(mN(n^2 \log n + \lambda^{-2} n \log n))$.
See the longer version of the paper [12] for a proof.

V. Experiments

In this section we provide some experimental evaluations of our clustering method using synthetically generated data. Our objective here is to showcase the generality of the proposed method. We generate the synthetic data according to stationary ergodic processes that do not belong to any "simpler" class.

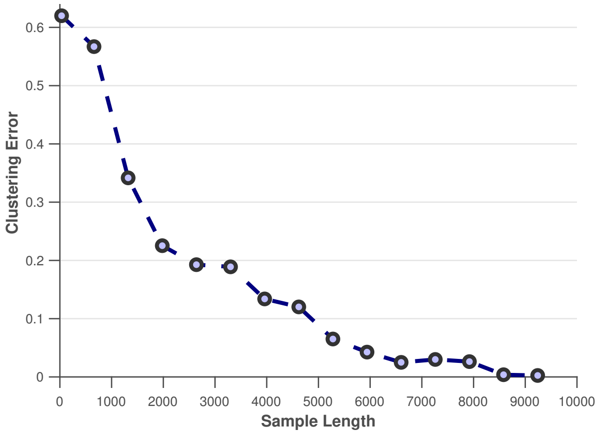

Fig. 1: Average error as a function of min sample length.

More specifically, we consider an ergodic rotation, see, e.g. [14], which is an example of a stationary ergodic process that is not a B-process, and that cannot be modelled by a hidden Markov model with a finite or countably infinite set of states. To generate a stationary ergodic sample $\mathbf{x}$ we proceed as follows. First, we fix some parameter $\beta \in (0,1)$. Then, we select $r_0 \in [0,1]$; for each $i = 1..n$ we obtain r_i by shifting r_{i-1} by β to the right, and removing the integer part, i.e. $r_i := r_{i-1} + \beta - \lfloor r_{i-1} + \beta \rfloor$. The sequence $\mathbf{x}$ is obtained from $r_i,\ i \in 1..n$ by thresholding at 0.5 i.e., for each $i \in 1..n$ we set $z_i = 1$ if $r_i > 0.5$ and $z_i = 0$ otherwise. If β is irrational then $\mathbf{z}$ forms a binary sample from a stationary ergodic process. We simulate β by a `longdouble` with a long mantissa. We obtain a piecewise stationary sample by first generating stationary samples as above, and then concatenating them in such a way that consecutive stationary samples are generated by rotation processes with different parameters. To carry out experiments concerning Algorithm 2 we fixed two sets of parameters $\boldsymbol{\beta}_1 := \{\beta_1^{(1)} = 0.121..,\ \beta_2^{(1)} = 0.141..,\ \beta_3^{(1)} = 0.161..\}$ and $\boldsymbol{\beta_2} := \{\beta_1^{(2)} = 0.141.., \beta_2^{(2)} = 0.15..\}$. Each set was used to parameterize the "bag of stationary disributions" corresponding to each cluster. The two sets were allowed to intersect, specifically, we let $\beta_2^{(1)} = \beta_1^{(2)}$. We generated one nonstationary sample with $\kappa = 2$ changepoints from the first process, using $\boldsymbol{\beta}_1$ as parameters for ergodic rotations, and two nonstationary samples with $\kappa = 1$ changepoint from the second one using $\boldsymbol{\beta}_2$. The objective was to ensure that the first nonstationary sample was separated from the second two samples. We set the clustering error to 0 if all three samples were correctly labeled and to 1 otherwise. Figure 1 shows the clustering error of Algorithm 2, averaged over 1000 repetitions, as a function of sample length n; the average error was calculated as the frequency of incorrect clustering. The parameter λ was set to 75% of the true value of α.

VI. Conclusion

We have introduced a novel setting for clustering time series. Our framework is more general than those considered in the literature, and allows for provably consistent algorithms. In this section we analyze the conditions of the main theorem and list possible extensions and generalizations.

Necessity of Conditions. In Theorem 1 we require that the correct number of clusters m, as well as a lower-bound on the minimum normalized length of the stationary segments, be provided. The former requirement is necessary: as shown in [15], there is no consistent two-sample test for stationary time series. Hence, without knowing the number of clusters, it is impossible to determine whether two samples generated by (single-piece and thus, of course, also piecewise) stationary distributions belong to the same or different clusters. Moreover, the lower-bound λ on the minimum normalized distance between changepoints is due to the corresponding condition of the changepoint estimation algorithm of [6]. We conjecture that this requirement is necessary. In contrast, as established by [8], it is possible to estimate the changepoints of a piecewise stationary sample consistently without knowing $\lambda \in (0, \alpha]$, provided that the number of changes is known. Thus, an analogous (though less practical) result can be obtained for the problem considered here: Algorithm 1 would obtain the changepoint estimates using the algorithm of [8], and proceed without further modifications. Hence, the knowledge of λ can be traded for that of the number of changepoints per sample.

Finite-time guarantees. In the context of stationary ergodic time series, fundamental results establish the impossibility of obtaining any finite-time guarantees on the error of the resulting algorithms: already the speed of convergence of frequencies of events to their underlying probabilities may be arbitrarily slow [14]. Thus, additional assumptions beyond stationarity and ergodicity are necessary to allow for finite-time guarantees. While this falls beyond the scope of the present paper, it would be interesting to consider the problem of clustering piecewise stationary mixing or even piecewise i.i.d. time series. Such assumptions would also make it possible to construct clustering algorithms that achieve consitency without knowing the correct number of clusters; see also [4] for a brief consideration of the problem concerning stationary mixing processes. Note that due to long-range dependencies, a rigorous finite-time analysis in this setting will require a careful consideration of random times, see, e.g. [16].

Extensions. An interesting generalization would be to an online setting whereby the samples grow over time, and new samples can be added at every time step. This has the potential to address a range of applications that involve growing bodies of data. The corresponding problem for stationary time series is addressed in [4]. Piecewise stationarity presents new challenges in this respect. Specifically, we may have infinitely many changepoints, and the number of stationary marginals per piecewise stationary process can also be infinite. An important challenge here would be to ensure robustness with respect to the case where two samples are generated by equivalent piecewise stationary distributions yet they consistently appear to be different in finite-time, since different subsets of their stationary marginals are revealed after any given number of time-steps. This task remains a challenge even when the sets of stationary marginals are assumed to be finite.

References

[1] D. Ryabko, *Asymptotic Nonparametric Statistical Analysis of Stationary Time Series*. Springer, 2019.

[2] D. Ryabko, "Clustering processes," in *Proceedings of the 27th International Conference on Machine Learning*, pp. 919–926, 2010.

[3] M. Balcan, A. Blum, and S. Vempala, "A discriminative framework for clustering via similarity functions," in *Proceedings of the 40th annual ACM symposium on Theory of computing*, pp. 671–680, ACM, 2008.

[4] A. Khaleghi, D. Ryabko, J. Mary, and P. Preux, "Consistent algorithms for clustering time series," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 94–125, 2016.

[5] A. Khaleghi and D. Ryabko, "Locating changes in highly dependent data with unknown number of change points," in *Advances in Neural Information Processing Systems 25*, pp. 3095–3103, 2012.

[6] A. Khaleghi and D. Ryabko, "Asymptotically consistent estimation of the number of change points in highly dependent time series," in *Proceedings of the 31st International Conference on Machine Learning*, vol. 32, pp. 539–547, 2014.

[7] D. Ryabko and B. Ryabko, "Nonparametric statistical inference for ergodic processes," *IEEE Transactions on Information Theory*, vol. 56, no. 3, pp. 1430–1435, 2010.

[8] A. Khaleghi and D. Ryabko, "Nonparametric multiple change point estimation in highly dependent time series," *Theoretical Computer Science*, vol. 620, pp. 119–133, 2016.

[9] F. M. Willems, "Coding for a binary independent piecewise-identically-distributed source," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 2210–2217, 1996.

[10] A. Gyorgy, T. Linder, and G. Lugosi, "Efficient tracking of large classes of experts," *IEEE Transactions on Information Theory*, vol. 58, no. 11, pp. 6709–6725, 2012.

[11] R. Gray, *Probability, Random Processes, and Ergodic Properties*. Springer Verlag, 1988.

[12] A. Khaleghi and D. Ryabko, "Clustering piecewise stationary processes," *arXiv preprint arXiv:1906.10921*, 2019.

[13] I. Katsavounidis, C.-C. J. Kuo, and Z. Zhang, "A new initialization technique for generalized Lloyd iteration," *IEEE Signal Processing Letters*, vol. 1, pp. 144–146, 1994.

[14] P. Shields, *The Ergodic Theory of Discrete Sample Paths*. AMS Bookstore, 1996.

[15] D. Ryabko, "Discrimination between B-processes is impossible," *Journal of Theoretical Probability*, vol. 23, no. 2, pp. 565–575, 2010.

[16] S. Grünewälder and A. Khaleghi, "Approximations of the restless bandit problem," *Journal of Machine Learning Research*, vol. 20, no. 1, pp. 514–550, 2019.