Asymptotically consistent estimation of the number of change points in highly dependent time series

Azadeh Khaleghi AZADEH.KHALEGHI@CURIE.FR
CBIO Mines ParisTech, INSERM U900, Institut Curie, FRANCE

Daniil Ryabko DANIIL.RYABKO@INRIA.FR
SequeL-INRIA Lille - Nord Europe, FRANCE

Abstract

The problem of change point estimation is considered in a general framework where the data are generated by arbitrary unknown stationary ergodic process distributions. This means that the data may have long-range dependencies of an arbitrary form. In this context the consistent estimation of the number of change points is provably impossible. A formulation is proposed which overcomes this obstacle: it is possible to find the correct number of change points at the expense of introducing the additional constraint that the correct number of process distributions that generate the data is provided. This additional parameter has a natural interpretation in many real-world applications. It turns out that in this formulation change point estimation can be reduced to time series clustering. Based on this reduction, an algorithm is proposed that finds the number of change points and locates the changes. This algorithm is shown to be asymptotically consistent. The theoretical results are complemented with empirical evaluations.

1. Introduction

Change point estimation is a classical problem in statistics and machine learning (Brodsky & Darkhovsky, 1993; Basseville & Nikiforov, 1993) and has applications in a broad range of such domains as market analysis, bioinformatics, audio and video segmentation, fraud detection, only to name a few. The problem can be introduced as follows. A given sequence $\boldsymbol{x} := X_1, \dots, X_{\lfloor n\theta_1 \rfloor}, \dots, X_{\lfloor n\theta_\kappa \rfloor + 1}, \dots, X_n$ is formed as the concatenation of an (unknown) number $\kappa + 1$ of non-overlapping segments where $\theta_k \in (0, 1),\ k = 1..\kappa$. Each segment is generated by one of $r \leq \kappa$ unknown stochastic process distributions. The process distributions that generate every pair of consecutive segments are different. The index $\lfloor n\theta_k \rfloor$ where one segment ends and another starts is called a *change point*. The parameters $\theta_k,\ k = 1..\kappa$ that specify the change points $\lfloor n\theta_k \rfloor$ are unknown and to be estimated.

We consider highly dependent time series, making as little assumptions as possible on how the data are generated. In particular, the distributions that generate the data are unknown and can be arbitrary; the only assumption is that they are stationary ergodic. This means that we make no such assumptions as independence, finite memory or mixing. Moreover, the change refers to the change in time series distribution and can have an arbitrary form. In particular, it is not restricted to the change in the mean, moment, etc., and is not confined to the finite-dimensional marginals of any fixed size. For example, the change may concern only the form of the long-range dependence.

With no further assumptions or additional information, this general formulation of the problem does not allow for the correct estimation of the number of change points, even in the weakest asymptotic sense. Indeed, as shown by (Ryabko, 2010b) in the general setting of highly dependent time series, it is even impossible to distinguish between the case of 0 and 1 change point; this impossibility result holds even for discrete-valued time series.

Stricter statistical frameworks are usually considered in the literature, making it possible to find κ at the cost of making stronger assumptions on the time series and on the form of the change. These stronger assumptions typically result in that the speed of convergence of a certain empirical statistic is known, e.g. from concentration inequalities, which can then be used to identify the changes via thresholding. However, our objective in this paper is to seek a natural for-

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

mulation under which it is possible to consistently solve the change point problem without the need to impose stronger assumptions on the process distributions.

To this end, we propose a novel formulation of the problem, which, at the expense of a single additional parameter allows us to overcome the impossibility result described. Namely, we assume that the total number r of process distributions that generate the data is provided to the algorithm. This formulation and the additional parameter is motivated by applications. Indeed, first of all, the assumption that the data are highly dependent complies well with most real-world scenarios. Assumptions on the rates of convergence of certain statistics (such as mixing), and even more general assumptions made to that effect, are usually impossible to verify in practice. On the other hand, in many applications the number r of distributions is a natural parameter of the problem. The simple example where a pair of distributions alternate in generating a sequence with many change points, may, in many cases, correspond to a system whose behaviour over time alternates between *normal* and *abnormal* ($r = 2$). To a varying extent, this may be a suitable model for system performance management, video surveillance and fraud detection. The identification of coding versus non-coding regions in genomic data is another potential application for this formulation with $r = 2$. Another application concerns the problem of authorship attribution in a given text written collaboratively by a known number r of authors. Finally, in speech segmentation r may be the total number of speakers.

Main Results. We propose a natural formulation of the change point problem, as well as a nonparametric algorithm to find the number of change points and to estimate the changes in stationary ergodic time series. We demonstrate both theoretically and experimentally that our algorithm is asymptotically consistent in the general framework described. The asymptotic regime means that the error is arbitrarily small if the sequence is sufficiently long. In particular, the problem is “offline” and the sequence does not grow with time.

The novelty of our work lies not only in the algorithm proposed, but also in the problem formulation. We demonstrate that, despite theoretical impossibility results, if the total number r of process distributions that generate the data is provided, the problem of obtaining the correct number of change points admits a solution, without requiring the knowledge of any probabilistic characteristics of the distributions generating the data or of the form of the changes.

Moreover, we show that a consistent algorithm can be obtained by a reduction to the problem of time series clustering. Thus, our results establish a novel formal link between two classical unsupervised learning problems, namely clustering and change point analysis, potentially bringing a new insight to both communities.

The theoretical results are illustrated with experiments on synthetic data. To generate the data we have used a well-known family of distributions that, while being stationary ergodic, do not belong to any “simpler” class of processes, and in particular cannot be modeled by finite- or countable-state models (such as finite-state HMMs).

Methodology. Our approach is composed of two main steps. First, we use a so-called list-estimator (Khaleghi & Ryabko, 2012) to produce an exhaustive list of at least κ candidate estimates, whose first κ elements are guaranteed to be asymptotically consistent. Once sorted in increasing order, the resulting list induces a partitioning of the sequence into consecutive segments. At this point we group the resulting segments into r clusters. A candidate estimate is identified as *redundant* if it joins a pair of consecutive segments in the same cluster. Finally, we remove the redundant estimates from the list and provide the remaining estimates as output. The clustering procedure uses farthest-point initialization to designate r cluster centres, and then assigns each remaining point to the nearest centre. To measure the distance between the segments, empirical estimates (Ryabko & Ryabko, 2010) of the so-called distributional distance (Gray, 1988) are used. The consistency of the proposed method can be established using any list-estimator that is consistent under the considered framework; we take a list-estimator from (Khaleghi & Ryabko, 2012) as an example. Different clustering procedures can be used as well, although their consistency for stationary ergodic time series cannot be exploited directly. The main challenge is in that the clustering procedure is used on segments that are obtained as concatenations of sequences generated by *different* process distributions, rather than by a single distribution. This means that the consistency analysis has to be performed anew.

Related Work. In a typical formulation of the problem, the sequence is assumed to have a single change point and the samples within each segment are assumed to be generated i.i.d, the distributions have known forms and the change is in the mean; see e.g. (Csörgö & Horváth, 1998) for a comprehensive review. In the literature on non-parametric methods for dependent data, the nature of dependence is typically restricted. For example, strong mixing conditions is a constraint that is commonly imposed, see e.g. (Brodsky & Darkhovsky, 1993). More general settings have also been considered, e.g. (Giraitis et al., 1995; Carlstein & Lele, 1993), with the latter work considering stationary ergodic time series. However, all these works assume that the single-dimensional marginals are different before and after the change point; this assumption is in fact prevalent in the literature.

The multiple change point estimation problem is considerably more difficult than the analysis of a single change, even if the number of change points is known. Thus, it is not as widely explored as that concerning single change point analysis. For known κ and dependent observations satisfying mixing conditions, the problem has been addressed from a global optimization perspective (Lavielle, 1999; Lavielle & Teyssiere, 2007). For the general framework considered in this paper, the case where κ is known has been considered by (Ryabko & Ryabko, 2010) ($\kappa = 1$) and (Khaleghi & Ryabko, 2013) ($\kappa > 1$). However, if κ provided to these algorithms is incorrect, their behavior can be arbitrarily bad. The case of unknown κ in this general setting is considered by (Khaleghi & Ryabko, 2012), where a list of possibly more than κ candidate estimates is produced, but no attempt is made to estimate κ; the produced list is sorted such that its first κ elements converge to the true change points.

The problem of estimating the number of change points is nontrivial, even under these more restrictive assumptions. In such settings, this problem is usually addressed with penalized criteria; see, e.g. (Lebarbier, 2005; Lavielle, 2005). Such criteria necessarily rely on additional parameters on which the resulting number of change points depends. Note that the algorithm proposed in this work also requires an input parameter: the number r of distributions. However, as argued above, parameter has a natural interpretation in many real-world applications. The problem of clustering stationary ergodic time series is considered by (Ryabko, 2010a), where the goal is to group together those and only those sequences that are generated by the same process distribution. Here we reduce the change point problem to clustering in this formulation. The important difference, which makes the consistency result of (Ryabko, 2010a) not directly applicable, is that we have to deal with concatenations of sequences generated by different distributions, rather than with individual sequences each generated by a single distribution.

Organization. The remainder of this paper is organized as follows. In Section 2 we introduce some preliminary notation and definitions. In Section 3 we formalize the problem. In Section 4 we present our algorithm, state the main consistency result, and intuitively explaining why it holds; we also provide a brief discussion on its computational complexity. The proofs of the main consistency results are given in Section 7. Section 5 is dedicated to our experimental results, and some concluding remarks are provided in Section 6.

2. Preliminaries

Let $\mathcal{X}$ be a measurable space (the domain); in this work we let $\mathcal{X} = \mathbb{R}$ but extensions to more general spaces are straightforward. For a sequence $X_1, \dots, X_n$ we use the abbreviation $X_{1..n}$. Consider the Borel σ-algebra $\mathcal{B}$ on $\mathcal{X}^\infty$ generated by the cylinders $\{B \times \mathcal{X}^\infty : B \in B^{m,l}, m, l \in \mathbb{N}\}$, where the sets $B^{m,l}, m, l \in \mathbb{N}$ are obtained via the partitioning of $\mathcal{X}^m$ into cubes of dimension m and volume 2^{-ml} (starting at the origin). Let also $B^m := \cup_{l \in \mathbb{N}} B^{m,l}$. Process distributions are probability measures on the space $(\mathcal{X}^\infty, \mathcal{B})$. For $\boldsymbol{x} = X_{1..n} \in \mathcal{X}^n$ and $B \in B^m$ let $\nu(\boldsymbol{x}, B)$ denote the *frequency* with which $\boldsymbol{x}$ falls in B, i.e.

$$\nu(\boldsymbol{x}, B) := \frac{\mathbb{I}\{n \geq m\}}{n - m + 1} \sum_{i=1}^{n-m+1} \mathbb{I}\{X_{i..i+m-1} \in B\}. \quad (1)$$

A process ρ is *stationary* if for any $i, j \in 1..n$ and $B \in B^m$, $m \in \mathbb{N}$, we have $\rho(X_{1..j} \in B) = \rho(X_{i..i+j-1} \in B)$. A stationary process ρ is *stationary ergodic* if for all $B \in \mathcal{B}$ with probability 1 we have $\lim_{n \to \infty} \nu(X_{1..n}, B) = \rho(B)$. By virtue of the ergodic theorem, this definition can be shown to be equivalent to the standard definition, see, e.g. (Csiszar & Shields, 2004).

Definition 1 (Distributional Distance)**.** *The distributional distance between a pair of process distributions ρ_1, ρ_2 is defined as follows (Gray, 1988)*

$$d(\rho_1, \rho_2) := \sum_{m,l=1}^{\infty} w_m w_l \sum_{B \in B^{m,l}} |\rho_1(B) - \rho_2(B)|,$$

where w_m and w_l, $m, l \in \mathbb{N}$ are sequences of positive summable real weights; we let $w_j := 1/j^2$.

In words, we partition the sets $\mathcal{X}^m$, $m \in \mathbb{N}$ into cubes of decreasing volume (indexed by l) and take a weighted sum over the differences in probabilities of all the cubes in these partitions, where smaller weights are given to larger m and finer partitions.

Definition 2 (Empirical estimates of $d(\cdot, \cdot)$)**.** *Consider sequences $\boldsymbol{x}_i \in \mathcal{X}^{n_i}$ $n_i \in \mathbb{N}$, $i = 1, 2$, $\boldsymbol{x} \in \mathcal{X}^n$, $n \in \mathbb{N}$ and process distribution ρ. The empirical estimates of d are defined as follows.*

$$\hat{d}(\boldsymbol{x}, \rho) := \sum_{m=1}^{m_n} \sum_{l=1}^{l_n} w_m w_l \sum_{B \in B^{m,l}} |\nu(\boldsymbol{x}, B) - \rho(B)| \quad (2)$$

$$\hat{d}(\boldsymbol{x}_1, \boldsymbol{x}_2) := \sum_{m=1}^{m_n} \sum_{l=1}^{l_n} w_m w_l \sum_{B \in B^{m,l}} |\nu(\boldsymbol{x}_1, B) - \nu(\boldsymbol{x}_2, B)| \quad (3)$$

where m_n and l_n are any sequences of integers that go to infinity with n.

Remark 1. *Despite the infinite summations, $\hat{d}$ can be calculated efficiently, (Ryabko, 2010a); its computational complexity is $\mathcal{O}(n \text{ polylog } n)$ for $m_n := \log n$, and $l_n := -\log s$, where $s := \min_{\substack{X_i \neq X_j \\ i,j \in 1..n}} |X_i - X_j|$. This choice* *of parameter is justified by (Khaleghi et al., 2012), and (Ryabko, 2010a); see also (Khaleghi & Ryabko, 2012).*

Proposition 1 ($\hat{d}(\cdot,\cdot)$ is asymptotically consistent (Ryabko & Ryabko, 2010)). *Let a pair of sequences $\boldsymbol{x}_i \in \mathcal{X}^{n_i}$, $i = 1, 2$ be generated by a distribution ρ whose marginals ρ_i, $i = 1, 2$ are stationary and ergodic. Then*

$$\lim_{n_i \to \infty} \hat{d}(\boldsymbol{x}_i, \rho_j) = d(\rho_i, \rho_j),\ i, j \in 1, 2,\ \rho - a.s., \quad (4)$$

$$\lim_{n_1, n_2 \to \infty} \hat{d}(\boldsymbol{x}_1, \boldsymbol{x}_2) = d(\rho_1, \rho_2),\ \rho - a.s. \quad (5)$$

3. Problem formulation

We formalize the problem as follows. The sequence $\boldsymbol{x} := X_1, \dots, X_n \in \mathcal{X}^n$, $n \in \mathbb{N}$, generated by an unknown arbitrary process distribution, is formed as the concatenation of $\kappa + 1$ of sequences $X_{1..\lfloor n\theta_1 \rfloor}, X_{\lfloor n\theta_1 \rfloor + 1..\lfloor n\theta_2 \rfloor}, \dots, X_{\lfloor n\theta_\kappa \rfloor + 1..n}$ where $\theta_k \in (0, 1)$, $k = 1..\kappa$, and κ are unknown. Each of the sequences $\boldsymbol{x}_k := X_{\pi_{k-1}+1..\pi_k}$, $k = 1..\kappa + 1$, $\pi_0 := 0$, $\pi_{\kappa+1} := n$ is generated by one of $r \leq \kappa + 1$ *unknown stationary ergodic* process distributions $\rho_1, \dots, \rho_r$. It is important to note that the number of change points κ is unknown and has to be estimated, but the number of different process distributions r is known and is provided to the algorithm.

To define the setting more formally, consider a matrix $\mathbf{X} \in (\mathcal{X}^\infty)^{\kappa+\mathbf{1}}$ of random variables generated by some (unknown) stochastic process distribution ρ such that **1.** the marginal distribution over every one of its rows is an *unknown* stationary ergodic process distribution, and **2.** the marginal distributions over the consecutive rows are different, so that every two consecutive rows are generated by different process distributions in $\{\rho_1, \dots, \rho_r\}$. The sequence $\boldsymbol{x} \in \mathcal{X}^n$, $n \in \mathbb{N}$ with κ change points is obtained as follows. First, the length $n \in \mathbb{N}$ is fixed, then for each $k = 1..\kappa + 1$ a segment $\boldsymbol{x}_k \in \mathcal{X}^{\lfloor n(\theta_k - \theta_{k-1}) \rfloor}$ is obtained as the first $\lfloor n(\theta_k - \theta_{k-1}) \rfloor$ elements of the k^{th} row of $\mathbf{X}$ with $\theta_0 := 0$, $\theta_{\kappa+1} := 1$. The individual segments $\boldsymbol{x}_k$, $k = 1..\kappa + 1$ are concatenated to produce $\boldsymbol{x} := \boldsymbol{x}_1, \dots, \boldsymbol{x}_{\kappa+1}$. Thus, there exists a ground-truth partitioning

$$\mathcal{G} := \{\mathcal{G}_1, \dots, \mathcal{G}_r\} \quad (6)$$

of the set $\{1..\kappa + 1\}$ into r disjoint subsets where for every $k = 1..\kappa + 1$ and $r' = 1..r$ we have $k \in \mathcal{G}_{r'}$ if and only if $\boldsymbol{x}_k$ is generated by $\rho_{r'}$. The parameters θ_k, $k = 1..\kappa$ specify the *change points* $\lfloor n\theta_k \rfloor$ which separate consecutive segments $\boldsymbol{x}_k, \boldsymbol{x}_{k+1}$ generated by different process distributions. The change points are *unknown* and to be estimated. The process distributions $\rho_1, \dots, \rho_r$ are completely unknown and may be dependent. Moreover, the means, variances, or more generally, the finite-dimensional marginal distributions of any fixed size before and after the change points are not required to be different. We consider the most general scenario where the process distributions are different.

Define the minimum separation of the change point parameters θ_k , $k = 1..\kappa$ as

$$\lambda_{\min} := \min_{k=1..\kappa+1} \theta_k - \theta_{k-1}. \quad (7)$$

Since the consistency properties we are after are asymptotic in n, we require that $\lambda_{\min} > 0$. Note that this condition is standard in the change point literature, although it may be unnecessary when simpler formulations of the problem are considered, for example when the samples are i.i.d. However, conditions of this kind are inevitable in the general setting that we consider, where the segments as well as the samples within each segment are allowed to be arbitrarily dependent: if the length of one of the sequences is constant or sub-linear in n, obtaining asymptotic consistency is not possible in this setting.

As discussed in the introduction, it is provably impossible (Ryabko, 2010b) to distinguish between the case of one and zero change points in the general framework. Hence, the number κ of change points cannot be estimated with no further information. This is the reason why we assume that the total number r of process distributions that generate the data is provided (while the number κ of change points remains unknown).

Thus, the *problem formulation* we consider can be described as follows. Given a sequence $\boldsymbol{x}$, a lower-bound on the minimum distance λ between the change points, and the total number r of process distributions, we seek a method that outputs the estimated number $\hat{\kappa}$ of change points and the estimated change point parameters $\hat{\theta}_1, \dots, \hat{\theta}_{\hat{\kappa}}$. We require the algorithm to be *asymptotically consistent* so that with probability 1 from some n on $\hat{\kappa} = \kappa$, and the estimates $\hat{\theta}_k$ satisfy

$$\lim_{n \to \infty} \sup_{k=1..\kappa} |\hat{\theta}_k(n) - \theta_k| = 0 \text{ a.s.} \quad (8)$$

Note that in the particular case of $r = \kappa + 1$ where all of the process distributions are different we have κ change points and thus we arrive to the formulation of the problem with known κ. More generally, r can be very different from $\kappa + 1$, and as discussed in the introduction, has a natural interpretation in many real-world applications.

4. Main theoretical results

In this section we present our algorithm and informally explain how it works. Theorem 1 establishes its consistency.

The proposed algorithm relies on a so-called list estimator: a procedure that, given $\boldsymbol{x}$ and λ, outputs a (long, exhaustive) list of change point estimates, without attempting

Algorithm 1 A change point estimator for known r

```
input: $\boldsymbol{x} \in \mathcal{X}^n$, $\lambda \in (0, \lambda_{\min}]$, Number $r$ of process distributions
Initialize: $\psi_0 \leftarrow 0$, $\psi_{|\Upsilon|+1} \leftarrow n$
1. Obtain a list of candidate estimates via a consistent list-estimator:
      $\Upsilon \leftarrow \Upsilon(\boldsymbol{x}, \lambda)$
2. Sort the list in increasing order:
      $\{\psi_i : i = 1..|\Upsilon|\} \leftarrow \textbf{sort}(\{n\hat{\theta} : \hat{\theta} \in \psi\})$
   (so that $i < j \Leftrightarrow \psi_i < \psi_j,\ i, j \in 1..|\Upsilon|$.)
   Generate a set $\mathcal{S}$ of consecutive segments:
      $\mathcal{S} \leftarrow \{\widetilde{\boldsymbol{x}}_i := X_{\psi_{i-1}+1..\psi_i} : i = 1..|\Upsilon| + 1\}$   (9)
3. Partition $\mathcal{S}$ into $r$ clusters:
   i. Initialize $r$ farthest segments as cluster centres
      $c_1 \leftarrow 1$
      $c_j \leftarrow \underset{i=1..|\Upsilon|}{\operatorname{argmax}} \min_{i'=1}^{j-1} \hat{d}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_{c_{i'}}),\ j = 2..r$   (10)
   ii. Assign every segment to the closest cluster
      $T(\widetilde{\boldsymbol{x}}_i) \leftarrow \operatorname{argmin}_{j=1..r} \hat{d}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_{c_j}),\ i = 1..|\Upsilon|$
4. Eliminate redundant estimates:
   $\mathcal{C} \leftarrow \{1..|\Upsilon|\}$
   for $i = 1..|\Upsilon|$ do
      if $T(\widetilde{\boldsymbol{x}}_i) = T(\widetilde{\boldsymbol{x}}_{i+1})$ then
         $\mathcal{C} \leftarrow \mathcal{C} \setminus \{i\}$
      end if
   end for
5. Obtain $\kappa$ and the estimates for $\theta_k,\ k = 1..\kappa$
      $\hat{\kappa} \leftarrow |\mathcal{C}|,\ \hat{\theta}_i := \frac{1}{n}\psi_i,\ i \in \mathcal{C}$
   return: $\hat{\kappa}, \hat{\theta}_i,\ i \in \mathcal{C}$
```

to estimate the number of change points in the sequence. More precisely, a list-estimator is defined as follows.

Definition 3 (List-estimator). *A list-estimator Υ is a function that, given a sequence $\boldsymbol{x} \in \mathcal{X}^n$ of length $n \in \mathbb{N}$ and a parameter $\lambda \in (0, 1)$, produces a list $\Upsilon(\boldsymbol{x}, \lambda) := (\hat{\theta}_1(n), \ldots, \hat{\theta}_{|\Upsilon|}(n)) \in (0, 1)^{|\Upsilon|}$ of some $|\Upsilon| \geq \kappa$ estimates. Let $(\hat{\theta}_{\mu_1}, \hat{\theta}_{\mu_2}, \ldots, \hat{\theta}_{\mu_\kappa}) :=$ **sort**$(\hat{\theta}_1, \ldots, \hat{\theta}_\kappa)$ be the first κ elements of $\Upsilon(\boldsymbol{x}, \lambda)$, sorted in increasing order of value. We call Υ* asymptotically consistent *if for every $\lambda \in (0, \lambda_{\min}]$ with probability 1 we have $\lim_{n\to\infty} \sup_{k=1..\kappa} |\hat{\theta}_{\mu_k}(n) - \theta_k| = 0$.*

An example of a consistent list-estimator is provided in (Khaleghi & Ryabko, 2012).

The key idea of the proposed algorithm is to have a consistent list-estimator, such as that presented in (Khaleghi & Ryabko, 2012), produce a list of at least κ change point estimates, and then cluster the segments induced by the candidate estimates to identify the redundant estimates in the list.

More specifically, Algorithm 1 works as follows. First, a consistent list-estimator is used to obtain an initial set of change-point candidates. The estimates are sorted in *increasing order* to produce a set $\mathcal{S}$ of consecutive non-overlapping segments of $\boldsymbol{x}$. The set $\mathcal{S}$ is then partitioned into r clusters. We use the following clustering procedure. First, a total of r cluster centres are obtained as follows. The first segment $\boldsymbol{x}_1$ is chosen as the first cluster centre. Iterating over $j = 2..r$ a segment is chosen as a cluster centre if it has the highest minimum distance from the previously chosen cluster centres. Once the r cluster centres are specified, the remaining segments are assigned to the closest cluster. In each cluster, the change point candidate that joins a pair of consecutive segments of $\boldsymbol{x}$ is identified as *redundant* and is removed from the list. Once all of the redundant candidates are removed, the algorithm outputs the remaining candidate estimates.

Theorem 1. *[Algorithm 1 is asymptotically consistent.] With probability 1, from some n on we have $\hat{\kappa} = \kappa$, and the estimates $\hat{\theta}_k$ satisfy $\lim_{n\to\infty} \sup_{k=1..\kappa} |\hat{\theta}_k(n) - \theta_k| = 0$, provided that $\lambda \in (0, \lambda_{\min}]$, and the correct number r of process distributions are given.*

The proof is given in Section 7. A sketch of the proof follows. Since a consistent list-estimator Υ is used in the first step, for large enough n, an initial set of possibly more than κ estimated parameters is generated, that contains κ elements which are arbitrarily close to the true change point parameters. (Since κ is unknown, the fact that the correct estimates are listed first is irrelevant; all we can use here is that they are somewhere in the list.) Therefore, if $\boldsymbol{x}$ is long enough, the largest portion of each segment in $\mathcal{S}$ is generated by a single process distribution. Since the initial change point candidates are at least $n\lambda$ apart, the lengths of the segments in $\mathcal{S}$ are linear functions of n. Thus, we can show that for large enough n, the empirical estimate of the distributional distance between a pair of segments in $\mathcal{S}$ converges to 0 if and only if the same process distribution generates most of the two segments. Given the total number r of process distributions, for long enough $\boldsymbol{x}$ the clustering algorithm groups together those and only those segments in $\mathcal{S}$ that are generated by the same process distribution. This lets the algorithm identify and remove the redundant candidates. By the consistency of Υ, the remaining estimates converge to the true change point parameters.

Computational Complexity. It is easy to see that the algorithm can be implemented efficiently. The initial $|\Upsilon| \le 1/\lambda$ change point candidates are obtained via the algorithm of (Khaleghi & Ryabko, 2012) which as shown by the authors has complexity $\mathcal{O}(n^2 \operatorname{polylog} n)$, where n corresponds to the length of the sequence. The clustering procedure only requires $r|\Upsilon|$ pairwise distance calculations; apart from that, the remaining calculations are of order $\mathcal{O}(r(|\Upsilon|+1))$. Thus, by Remark 1, the resource complexity of Algorithm 1 is $\mathcal{O}(n^2 \operatorname{polylog} n)$.

5. Experimental results

We evaluate our method using synthetically generated data. The data are generated using stationary ergodic process distributions that do not belong to any "simpler" general class of time series, and cannot be approximated by finite state models. Also, the single-dimensional marginals of all distributions are the same throughout the generated sequence.

To generate a sequence $\boldsymbol{x} \in \mathbb{R}^n,\ n \in \mathbb{N}$ with κ change points we proceed as follows. For every $k \in 1..\kappa+1$ we use the so-called *ergodic rotation* to generate the segment $\boldsymbol{x}_k := X_{\lfloor n\theta_{k-1}\rfloor+1}, \dots, X_{\lfloor n\theta_k \rfloor} \in \mathbb{R}^{n_k}$ with $n_k := \lfloor n\theta_k \rfloor - \lfloor n\theta_{k-1}\rfloor$, where $\theta_0 := 0$ and $\theta_{\kappa+1} := 1$. More specifically, to generate the segment $\boldsymbol{x}_k,\ k \in 1..\kappa+1$ we proceed as follows.

1. Fix a parameter $\alpha_k \in (0,1)$ and two uniform distributions $\mathcal{U}_1$ and $\mathcal{U}_2$.
2. Let a_0 be drawn uniformly at random from $[0,1]$.
3. For each $i = 1..n_k$, shift a_{i-1} to the right by α_k and remove the integer part, i.e. $a_i = a_{i-1} + \alpha_k \mod 1$. Also draw $x_i^{(j)}$ from $\mathcal{U}_j,\ j = 1,2$.
4. Set $X_{i+\lfloor n\theta_{k-1}\rfloor}$ as

$$X_{i+\lfloor n\theta_{k-1}\rfloor} = \mathbb{I}\{r_i \le 0.5\}x_i^{(1)} + \mathbb{I}\{r_i > 0.5\}x_i^{(2)}.$$

If α_k is irrational this procedure produces a real-valued stationary ergodic time series $\boldsymbol{x}_k$. These processes are commonly used as examples of stationary ergodic processes that are not B-processes, see e.g. (Shields, 1996). We simulate $\alpha_k,\ k = 1..\kappa+1$ by a `longdouble` with a long mantissa. For the purpose of our experiment, we fixed three parameters $\alpha_1 := 0.12..,\ \alpha_2 := 0.14..,\ \alpha_3 := 0.16..$ (with long mantissae) to correspond to $r := 3$ different process distributions. To produce $\boldsymbol{x} \in \mathbb{R}^n$ we generated $\kappa := 3$ change point parameters $\theta_k,\ k = 1..\kappa$ with minimum separation $\lambda_{\min} := 0.1$. Every segment $\boldsymbol{x}_k$ of length was generated with α_k where $k := k' \mod r,\ k' = 0..\kappa+1$ (so that the first and the last segments, namely $\boldsymbol{x}_1$ and $\boldsymbol{x}_4$, were deliberately generated by the same process distribution), and using uniform distributions $\mathcal{U}_1$ and $\mathcal{U}_2$ over $[0, 0.7]$ and $[0.3, 1]$ respectively. We used the algorithm proposed by (Khaleghi & Ryabko, 2012) as the list-estimator $\Upsilon(\cdot,\cdot)$.

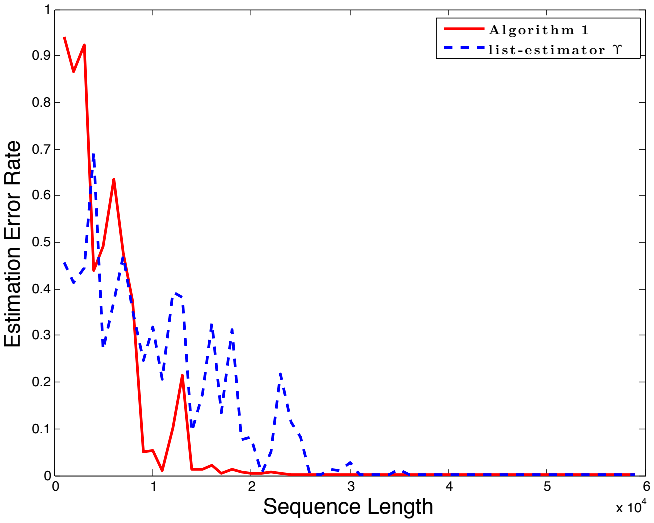

Figure 1. Average (over 40 runs) error rates of Alg $1(\boldsymbol{x}, r, \lambda)$ and the list-estimator $\Upsilon(\boldsymbol{x}, \lambda)$ of (Khaleghi & Ryabko, 2012) as a function of the sequence length n for $\boldsymbol{x} \in \mathbb{R}^n$; for Υ, the error is calculated on its first κ elements.

As shown in Figure 5 the estimation error rate of Alg$1(\boldsymbol{x}, \lambda, r)$ with $\lambda := 0.6\lambda_{\min}$ converges to 0 as n ranges over $1000..60000$. In each run, the error is calculated as

$$\mathbb{I}\{|\mathcal{C}| \ne \kappa\} + \mathbb{I}\{|\mathcal{C}| = \kappa\} \sum_{k=1}^{\kappa} |\hat{\theta}_k - \theta_k|.$$

For illustration, we also plot the performance of the list estimator of (Khaleghi & Ryabko, 2012). Since the reference method does not attempt to estimate κ, we calculate its error on the first κ elements of its output list.

6. Conclusion

We have presented an asymptotically consistent algorithm to find the number of change points, and locate the changes in highly dependent time series. While in the general framework considered, it is provably impossible to obtain the number of change points, we have managed to tackle the problem under a natural formulation, namely, under the assumption that the correct number of process distributions that generate the data is provided.

In this framework, rates of convergence are provably impossible to obtain, and the algorithms developed for this framework are forced not to rely on rates of convergence.

While the downside is that the asymptotic results obtained for these methods cannot be strengthened, the advantage is that the rate-free methods designed for this framework are applicable to a much broader range of situations. At the same time, it may be interesting to derive the rates of convergence of the proposed algorithm under stronger assumptions (e.g. i.i.d., mixing, etc.). We conjecture that our algorithm is close to optimal in such settings as well (although it clearly cannot be optimal in parametric settings). Another interesting question concerns the time series distance used in the algorithms. We establish our consistency results using some properties specific to the (empirical estimates of the) distributional distance. It is possible that other distances, for example the telescope distance of (Ryabko & Mary, 2012), can replace the distributional distance used in our method. These questions may be addressed in future work.

7. Proofs

The proof of Theorem 1 relies on Lemma 1 which is borrowed from (Khaleghi & Ryabko, 2013), and on Lemma 2 which is stated and proven below.

Lemma 1 (Khaleghi & Ryabko, 2013)**.** *Let $\boldsymbol{x} = X_{1..n}$ be generated by a stationary ergodic process ρ. For $\alpha \in (0,1)$ the following statements hold with ρ-probability 1:*

(i) For every $T \in \mathbb{N}$ we have

$$\lim_{n\to\infty} \sup_{|b_2-b_1|\geq \alpha n} \sum_{\substack{B\in B^{m,l},\\ m,l\in 1..T}} |\nu(X_{b_1..b_2}, B) - \rho(B)| = 0.$$

(ii) $\lim_{n\to\infty} \sup_{|b_2-b_1|\geq \alpha n} \hat{d}(X_{b_1..b_2}, \rho) = 0.$

We introduce the following additional notation. Consider the set $\mathcal{S}$ of segments specified by Line (9) in Algorithm 1. For every $\widetilde{\boldsymbol{x}}_i := X_{\psi_{i-1}..\psi_i} \in \mathcal{S},\ i = 1..|\Upsilon|+1$ define $\widetilde{\rho}_i$ as the process distribution that generates the largest portion of $\widetilde{x}_i$. That is, let $\widetilde{\rho}_i := \rho_j$ where j is such that $K \in \mathcal{G}_j$ with

$$K := \operatorname*{argmax}_{k\in\mathcal{G}_{r'}} |\{\psi_{i-1}{+}1, \dots, \psi_i\} \cap \{n\theta_{k-1}{+}1, \dots, n\theta_k\}|,$$

where $\mathcal{G}_j,\ j = 1..r$ are given by (6).

Lemma 2. *Let $\boldsymbol{x} \in \mathcal{X}^n,\ n \in \mathbb{N}$ be a sequence with κ change points with minimum separation $\lambda_{\min}$ for some $\lambda_{\min} \in (0,1)$. Assume that the distributions that generate $\boldsymbol{x}$ are stationary and ergodic. Let $\mathcal{S}$ be the set of segments specified by (9) in Algorithm 1. For all $\lambda \in (0, \lambda_{\min})$ with probability 1 we have $\lim_{n\to\infty} \sup_{\boldsymbol{x}_i\in\mathcal{S}} \hat{d}(\widetilde{\boldsymbol{x}}_i, \widetilde{\rho}_i) = 0$.*

Proof. Fix an $\varepsilon \in (0, \lambda/2)$. Define $\pi_k := \lfloor n\theta_k \rfloor,\ k = 1..\kappa$. Since the initial list Υ of change point candidates are produced by a *consistent* list-estimator $\Upsilon(\boldsymbol{x}, \lambda)$, (see Definition 3), there exists an index-set $\mathcal{I} := \{\mu_1, \dots, \mu_\kappa\} \in \{1..|\Upsilon|\}^\kappa$ and some N_0 such that for all $n \geq N_0$ we have

$$\sup_{k=1..\kappa} \frac{1}{n}|\psi_{\mu_k} - \pi_k| \leq \varepsilon. \tag{11}$$

Moreover, the initial candidates are at least $n\lambda$ apart so that

$$\inf_{i\in 1..|\Upsilon|+1} \psi_i - \psi_{i-1} \geq n\lambda \tag{12}$$

where $\psi_0 := 0$ and $\psi_{|\Upsilon|+1} := n$. Let $\mathcal{I}' := \{1..|\Upsilon|\} \setminus \mathcal{I}$. Denote by

$$\mathcal{S}_1 := \{\widetilde{x}_i := X_{\psi_{i-1}+1..\psi_i} \in \mathcal{S} : \{i, i-1\} \cap \mathcal{I} = \varnothing\}$$

the subset of the segments in $\mathcal{S}$ whose elements are formed by joining pairs of *consecutive elements* of $\mathcal{I}'$ and let $\mathcal{S}_2 := \mathcal{S} \setminus \mathcal{S}_1$ be its complement. Let the true change points that appear immediately to the left and to the right of an index $j \in 1..n-1$ be given by $\mathcal{L}(j) := \max_{k\in 0..\kappa+1} \pi_k \leq j$ and $\mathcal{R}(j) := \min_{k\in 0..\kappa+1} \pi_k > j$ respectively, with $\pi_0 := 0,\ \pi_{\kappa+1} := n$ where equality occurs when j is itself a change point. We have two cases. **1.** Consider $\widetilde{x}_i := X_{\psi_{i-1}+1..\psi_i} \in \mathcal{S}_1$. By definition, $\widetilde{x}_i$ cannot contain a true change point for $n \geq N_0$ since otherwise, either $i-1$ or i would belong to $\mathcal{I}$ contradicting the assumption that $\widetilde{x}_i \in \mathcal{S}_1$. Therefore, for all $n \geq N_0$ we have $\widetilde{\rho}_i = \rho$ where $\rho \in \{\rho_1, \dots, \rho_r\}$ is the process that generates $X_{\mathcal{L}(\psi_{i-1})..\mathcal{R}(\psi_{i-1})}$. By (12) and hence part (ii) of Lemma 1, there exists some $N_i \geq N_0$ such that for all $n \geq N_i$ we have $\hat{d}(\widetilde{\boldsymbol{x}}_i, \widetilde{\rho}_i) := \hat{d}(X_{\psi_{i-1}..\psi_i}, \rho) \leq \varepsilon$. Let $N' := \max_{i \text{ s.t. } \widetilde{\boldsymbol{x}}_i\in\mathcal{S}_1} N_i$. For all $n \geq N'$ we have

$$\sup_{\widetilde{x}_i\in\mathcal{S}_1} \hat{d}(\widetilde{x}_i, \widetilde{\rho}_i) \leq \varepsilon. \tag{13}$$

2. Take $\widetilde{x}_i := X_{\psi_{i-1}..\psi_i} \in \mathcal{S}_2$. Observe that, by definition $\mathcal{I} \cap \{i, i-1\} \neq \varnothing$ so that either $i-1$ or i belong to $\mathcal{I}$. We prove the statement for the case where $i-1 \in \mathcal{I}$. The case where $i \in \mathcal{I}$ is analogous. We start by showing that $[\psi_{i-1}, \psi_i] \subseteq [\pi - \varepsilon n, \pi' + \varepsilon n]$ for all $n \geq N_0$ where $\pi := \operatorname{argmin}_{\pi_k, k=1..\kappa} \frac{1}{n}|\pi_k - \psi_{i-1}|$ and $\pi' := \mathcal{R}(\pi)$. Since $i-1 \in \mathcal{I}$, by (11) for all $n \geq N_0$ we have $\frac{1}{n}|\pi - \psi_{i-1}| \leq \varepsilon$. We have two cases. Either $i \in \mathcal{I}$ so that by (11) for all $n \geq N_0$ we have $\frac{1}{n}|\psi_i - \pi'| \leq \varepsilon$, or $i \in \mathcal{I}'$ in which case $\psi_i < \pi'$. To see the latter statement assume by way of contradiction that $\psi_i > \pi'$ where $\pi' \neq n$; (the statement trivially holds for $\pi' = n$). By the consistency of $\Upsilon(\boldsymbol{x}, \lambda)$ there exists some $j > i-1 \in \mathcal{I}$ such that $\frac{1}{n}|\psi_j - \pi'| \leq \varepsilon$ for all $n \geq N_0$. Moreover, by (11) and (12) for all $n \geq N_0$ the candidates indexed by $\mathcal{I}'$ have linear distances from the true change points, that is,

$$\begin{aligned} &\inf_{k\in 1..\kappa,\ i\in\mathcal{I}'} |\pi_k - \psi_i| \qquad (14)\\ &\geq \inf_{k\in 1..\kappa,\ i\in\mathcal{I}', j\in\mathcal{I}} |\psi_i - \psi_j| - |\pi_k - \psi_j| \geq n(\lambda - \varepsilon). \end{aligned}$$

Thus, from (11) and (14) we obtain that $\psi_i - \psi_j \geq \lambda - 2\varepsilon > 0$. Since the initial estimates are sorted in increasing order, this implies $j \leq i$ leading to a contradiction. Thus we have $[\psi_{i-1}, \psi_i] \subseteq [\pi - \varepsilon n, \pi' + \varepsilon n]$. Therefore, $\widetilde{\rho}_i = \rho$ where ρ is the process distribution $\rho \in \{\rho_1, \dots, \rho_r\}$ that generates $X_{\pi..\pi'}$. To show that $\hat{d}(\widetilde{x}_i, \rho) \leq \varepsilon$ we proceed as follows. There exists some T such that $\sum_{m,l=T}^{\infty} w_m w_l \leq \varepsilon$. It is easy to see that by (14), and the assumption that $\lambda \in (0, \lambda_{\min}]$, (where $\lambda_{\min}$ is given by (7)), the segment $X_{\pi..\min\{\psi_i, \pi'\}}$ has length at least $n\lambda$, i.e.

$$\min\{\psi_i, \pi'\} - \pi \geq n\lambda.$$

Therefore, by part (i) of Lemma 1, there exists some $N_i \geq N_0$ such that for all $n \geq N_i$ we have

$$\sum_{m,l=1}^{T} w_m w_l \sum_{B \in B^{m,l}} |\nu(X_{\pi..\min\{\psi_i,\pi'\}}, B) - \rho(B)| \leq \varepsilon. \tag{15}$$

Using the definition of $\nu(\cdot,\cdot)$ given by (1), for every $B \in B^{m,l},\ m, l \in 1..T$ we have

$$\begin{aligned}
&(1 - \frac{m-1}{\psi_i - \psi_{i-1}})|\nu(\widetilde{x}_i, B) - \rho(B)| \\
&\leq \frac{\min\{\psi_i, \pi'\} - \pi - m + 1}{\psi_i - \psi_{i-1}} |\nu(X_{\pi..\min\{\psi_i,\pi'\}}, B) - \rho(B)| \\
&+ \frac{\mathbb{I}\{\psi_i \geq \pi'\}(\psi_i - \pi')}{\psi_i - \psi_{i-1}} + \frac{|\psi_i - \pi|}{\psi_i - \psi_{i-1}} \qquad (16)
\end{aligned}$$

Increase N_i if necessary to have $T/(\lambda N_i) \leq \varepsilon$, and let $n \geq N_i$. Recall that $\sum_{m,l=1}^{n} w_m w_l \leq 1$, and observe that $|\nu(\cdot,\cdot) - \rho(\cdot)| \leq 1$. By (11), (12), (15) and (16) we have

$$\begin{aligned}
&\hat{d}(\widetilde{x}_i, \widetilde{\rho}_i) \\
&\leq \sum_{m,l=1}^{T} w_m w_l \sum_{B \in B^{m,l}} (1 - \frac{m-1}{\psi_i - \psi_{i-1}})|\nu(\widetilde{\boldsymbol{x}}_i, B) - \rho(B)| \\
&\quad + \frac{m-1}{\psi_i - \psi_{i-1}} + \varepsilon \leq 3\varepsilon(1 + 1/\lambda) \qquad (17)
\end{aligned}$$

Let $N'' := \max_{i \text{ s.t. } \widetilde{\boldsymbol{x}}_i \in \mathcal{S}_2}$. By (17) for all $n \geq N''$ we have

$$\sup_{\widetilde{\boldsymbol{x}}_i \in \mathcal{S}_2} \hat{d}(\widetilde{\boldsymbol{x}}_i, \widetilde{\rho}_i) \leq 3\varepsilon(1 + 1/\lambda). \tag{18}$$

Finally, by (13) and (18) for all $n \geq \max\{N', N''\}$ we have $\sup_{\widetilde{\boldsymbol{x}}_i \in \mathcal{S}} \hat{d}(\widetilde{\boldsymbol{x}}_i, \widetilde{\rho}_i) \leq 3\varepsilon(1 + 1/\lambda)$. Since the choice of ε is arbitrary, this proves the statement. □

Proof of Theorem 1. Let $\delta := \min_{i \neq j \in 1..r} d(\rho_i, \rho_j)$ denote the minimum distance between the distinct distributions that generate $\boldsymbol{x}$. Fix an $\varepsilon \in (0, \delta/4)$. Recall that the list-estimator $\Upsilon(\boldsymbol{x}, \lambda)$ is consistent for all $\lambda \in (0, \lambda_{\min}]$ (see Definition 3). Therefore, there exists some N_1 such that for all $n \geq N_1$ the first κ elements of the list of candidate estimates that it produces, converge to the true change point parameters. Here, the only important message is that, for all $n \geq N_1$ the consistent estimates are somewhere within the list Υ. That is for all $n \geq N_1$ there exists a set of indices $\{\mu_k :\ k = 1..\kappa\} \subseteq 1..|\Upsilon|$ such that

$$\sup_{k \in 1..\kappa} |\hat{\theta}_{\mu_k} - \theta_k| \leq \varepsilon. \tag{19}$$

Since there are a finite number of segments in the set $\mathcal{S}$ (specified by (9) in Algorithm 1), by Lemma 2, there exists some N_2 such that for all $n \geq N_2$ we have $\sup_{\widetilde{\boldsymbol{x}}_i \in \mathcal{S}} \hat{d}(\widetilde{\boldsymbol{x}}_i, \widetilde{\rho}_i) \leq \varepsilon$. Hence, applying the triangle inequality on $\hat{d}$, for all $n \geq N_2$ we have

$$\sup_{\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_j \in \mathcal{S},\ \widetilde{\rho}_i = \widetilde{\rho}_j} \hat{d}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_j) \leq 2\varepsilon. \tag{20}$$

$$\inf_{\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_j \in \mathcal{S},\ \widetilde{\rho}_i \neq \widetilde{\rho}_j} \hat{d}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_j) \geq \delta - 2\varepsilon. \tag{21}$$

By (20) and (21), for all $n \geq N_2$, the segments $\widetilde{x}_i,\ \widetilde{x}_{i+1} \in \mathcal{S}$ with $\widetilde{\rho}_i = \widetilde{\rho}_{i+1}$ are closer to each other (in the empirical estimate of the distributional distance) than to the rest of the segments. By (21) for all $n \geq N_2$ and every $j \in 2..r$ we have $\max_{i \in 1..|\mathcal{S}|} \min_{j' \in 2..j-1} \hat{d}(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_{c_j}) \geq \delta - 2\varepsilon$ where, as specified by Algorithm 1, $c_1 := 1$ and $c_j,\ j = 2..r$ are given by (10). Hence, for all $n \geq N_2$, the cluster centres $\widetilde{x}_{c_j}, j = 1..r$ are each generated by a different process distribution. That is, $\widetilde{\rho}_{c_j} \neq \widetilde{\rho}_{c_{j'}}$ for $j \neq j' \in 1..r$. On the other hand, the rest of the segments are each assigned to the closest cluster, so that by (20) for all $n \geq N$ we have

$$T(\widetilde{\boldsymbol{x}}_i) = T(\widetilde{\boldsymbol{x}}_{i'}) \Leftrightarrow \widetilde{\rho}_i = \widetilde{\rho}_{i'}. \tag{22}$$

Let $N := \max N_i,\ i = 1, 2$. It remains to show that for all $n \geq N$, all of the redundant estimates namely, $\hat{\theta}_i,\ i \neq \mu_k,\ k = 1..\kappa$ are removed in the last step of the algorithm, so that for all $n \geq N$ there exists an index $i \in 1..|\Upsilon|$ in $\mathcal{C}$, if and only if it corresponds to a consistent estimate in Υ. To this end, we note that by (19) and (22) for all $n \geq N$ and all $i \in \mathcal{C}$ we have $\widetilde{\rho}_i \neq \widetilde{\rho}_{i+1}$ so that $\mathcal{C} = \{\mu_k :\ k = 1..\kappa\}$ for all $n \geq N$. By (19) and noting that $\hat{\kappa} := |\mathcal{C}|$ the statement follows. □

Acknowledgments

This work is supported by the French Ministry of Higher Education and Research, by FP7/2007-2013 under grant agreements 270327 (CompLACS) and 216886 (PASCAL-2), by the French National Research Agency (project Lampada ANR-09-EMER-007), by the Nord-Pas-de-Calais Regional Council and FEDER through CPER 2007-2013, and by an INRIA Ph.D. grant to Azadeh Khaleghi. *Most of the work has been done while Azadeh Khaleghi was a PhD student at INRIA Lille - Nord Europe.*

References

Basseville, M. and Nikiforov, I.V. *Detection of abrupt changes: theory and application*. Prentice Hall information and system sciences series. Prentice Hall, 1993.

Brodsky, B.E. and Darkhovsky, B.S. *Nonparametric methods in change point problems*. Mathematics and its applications. Kluwer Academic Publishers, 1993.

Carlstein, E. and Lele, S. Nonparametric change point estimation for data from an ergodic sequence. *Teor. Veroyatnost. i Primenen.*, 38:910–917, 1993.

Csiszar, I. and Shields, P.C. Notes on information theory and statistics. In *Foundations and Trends in Communications and Information Theory*, 2004.

Csörgö, Miklós and Horváth, Lajos. *Limit Theorems in Change-Point Analysis (Wiley Series in Probability & Statistics)*. Wiley, January 1998.

Giraitis, L, Leipus, R, and Surgailis, D. The change point problem for dependent observations. *JStat Plan and Infer*, pp. 1–15, 1995.

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

Khaleghi, A. and Ryabko, D. Locating changes in highly-dependent data with unknown number of change points. In *Neural Information Processing Systems (NIPS)*, Lake Tahoe, Nevada, United States, 2012.

Khaleghi, A. and Ryabko, D. Nonparametric multiple change point estimation in highly dependent time series. In *ALT'13*, Singapore, 2013. Springer.

Khaleghi, A., Ryabko, D., Mary, J., and Preux, P. Online clustering of processes. In *AI & Stats*, pp. 601–609, La Palma, Canary Islands, 2012.

Lavielle, M. Using penalized contrasts for the change point problem. *Signal Processing*, 85(8):1501 – 1510, 2005.

Lavielle, Marc. Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications*, 83(1):79–102, 1999.

Lavielle, Marc and Teyssiere, Gilles. Adaptive detection of multiple change points in asset price volatility. In *Long memory in economics*, pp. 129–156. Springer, 2007.

Lebarbier, E. Detecting multiple change points in the mean of gaussian process by model selection. *Signal Processing*, 85(4):717 – 736, 2005.

Ryabko, D. Clustering processes. In *ICML 2010*, pp. 919–926, Haifa, Israel, 2010a.

Ryabko, D. Discrimination between B-processes is impossible. *Journal of Theoretical Probability*, 23(2):565–575, 2010b.

Ryabko, D and Mary, J. Reducing statistical time series problems to binary classification. In *NIPS*, pp. 2069–2077, United States, 2012.

Ryabko, D. and Ryabko, B. Nonparametric statistical inference for ergodic processes. *IEEE Trans. on Info. Theory*, 56(3):1430–1435, 2010.

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