Reducing statistical time-series problems to kernel-based binary classification

Daniil Ryabko

Reducing statistical time-series problems to kernel-based binary classification

Daniil Ryabko
INRIA Lille-Nord Europe,
daniil@ryabko.net

March 1, 2012

Abstract

It is shown how kernel-based methods developed for solving problems concerning i.i.d. data can be used for statistical analysis of highly-dependent time series. Specifically, the problems of time-series clustering, homogeneity testing and classification are addressed. The methods are based on a new distance between time-series distributions.

1 Introduction

The last decades have seen and strong and consistent development of kernel-based methods for binary classification, both in theory and in applications. As a result, efficient algorithms have been developed, a host of application-specific kernels have been proposed, and a wealth domain knowledge has been accumulated with regards to their usage in different applications.

Binary classification being one of the most conceptually simple learning problem, it is natural to try and use it as a building block for solving other, more complex or just different problems. In other words, one can try to obtain efficient algorithms for different learning problems by *reducing* them to binary classification, and in particular to kernel-based methods developed for classification. Indeed, this approach has been applied to many different problems, starting with multi-class classification, but also including some statistical problems such as homogeneity testing and change-point detection [5, 8]. However, all of these problems are formulated in terms of independent and identically distributed (i.i.d.) samples. This is also the assumption underlying the theoretical analysis of most of the classification algorithms.

In this work we consider learning problems that concern time series for which independence assumptions do not hold. Time series can exhibit arbitrary long-range dependence, and different time-series samples may be interdependent as well. The problems we consider are the three-sample problem, clustering, and homogeneity testing. For the first two problems the only assumption on the data

that we make is that the distributions generating the samples are stationary ergodic; this is one of the weakest assumptions used in statistics. For homogeneity testing we have to make some mixing assumptions in order to obtain consistency results (this is unavoidable, as is shown in [13]).

We show how the considered problems can be reduced to kernel-based binary classification methods (such as SVM). The results are asymptotically consistent algorithms (for the case of stationary ergodic distributions) as well finite-sample analysis (for the case of distributions satisfying certain mixing conditions).

The proposed approach is based on a new distance between time-series distributions (that is, between probability distributions on the space of infinite sequences), which we call *telescope distance*. This distance can be evaluated using kernel methods, and its finite-sample estimates are shown to be asymptotically consistent. Three main building blocks are used to construct the telescope distance. The first one is a distance on finite-dimensional marginal distributions. The distance we use is the following: $d_{\mathcal{H}}(P,Q) := \sup_{h\in\mathcal{H}} |\mathbf{E}_P h - \mathbf{E}_Q h|$ where P, Q are distributions and $\mathcal{H}$ is a set of functions. This distance can be estimated (for i.i.d. data) using kernel-based binary classification methods, and thus can be used to reduce various statistical problems to the classification problem. As it was mentioned, this approach was previously applied to such statistical problems as homogeneity testing and change-point estimation. However, these applications so far have only concerned i.i.d. data, whereas we want to work with highly-dependent time series. Thus the second building block are the recent results of [1, 2], that show that empirical estimates of $d_{\mathcal{H}}$ are consistent (under certain conditions on $\mathcal{H}$) for arbitrary stationary ergodic distributions. This, however, is not enough: evaluating $d_{\mathcal{H}}$ for (stationary ergodic) time-series distributions means measuring the distance between their finite-dimensional marginals. Finally, the third step to construct the distance is what we call *telescoping*. It consists in summing the distances for all the (infinitely many) finite-dimensional marginals with decreasing weights. A similar approach has been used in [12] to construct an empirical distance that is shown there to converge to the so-called distributional distance [4] between time-series distributions. The empirical distance of [12] is based on counting frequencies. Here we use this trick to construct a distance based on $d_{\mathcal{H}}$, thus harnessing kernel-based methods to solve statistical problems about time series.

We show that the resulting distance (telescope distance) indeed can be consistently estimated based on sampling, for arbitrary stationary ergodic distributions. Further, we show how this fact can be used to construct consistent algorithms for the considered problems on time series. While the main results of this work are theoretical, we argue that all the proposed methods are easily computable, and readily available classification methods (such as SVMs) can be utilized to this end. Finally, we also provide some topological analysis of the telescope distance proposed in this work as compared to the distributional distance, that was previously used for statistical analysis of time series. Namely, we show that the telescope distance is stronger.

The rest of the paper is organized as follows. The next section introduces the notation and some definitions. In Section 3 we introduce the telescope

distance between time-series distributions and establish some of its basic properties. Sections 4 and 5 are devoted to the problems of time-series classification (the three-sample problem) and clustering, respectively. In Section 6 we show what kind of finite-sample performance guarantees can be established for the estimates of the telescope distance and thus for the presented algorithms, if some additional assumptions are made on the mixing rates; it is also shown that in this case a solution to the problem of homogeneity testing can be obtained. Section 7 discusses computational issues. Finally, section 8 provides some topological analysis of the telescope distance, and Section 9 concludes.

2 Notation and definitions

Let $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ be a measurable space (the domain). Time-series (or process) ddistributions are probability measures on the space $(X^{\mathbb{N}}, \mathcal{F}_{\mathbb{N}})$ of one-way infinite sequences (where $\mathcal{F}_{\mathbb{N}}$ is the sigma-algebra of $X^{\mathbb{N}}$). We use the abbreviation $X_{1..k}$ for $X_1, \dots, X_k$. All sets and functions introduced below (in particular, the sets $\mathcal{H}_k$ and their elements) are assumed measurable.

A distribution ρ is stationary if $\rho(X_{1..k} \in A) = \rho(X_{n+1..n+k} \in A)$ for all $A \in \mathcal{F}_{\mathcal{X}^k}$, $k, n \in \mathbb{N}$ (with $\mathcal{F}_{\mathcal{X}^k}$ being the sigma-algebra of X^k). A stationary distribution is called (stationary) ergodic if $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1..n-k+1} \mathbb{I}_{X_{i..i+k} \in A} = \rho(A)$ ρ-a.s. for every $A \in \mathcal{F}_{\mathcal{X}^k}$, $k \in \mathbb{N}$. (This definition, which is more suited for the purposes of this work, is equivalent to the usual one expressed in terms of invariant sets, see e.g. [4].)

3 Kernel-based distance between time-series distributions

We start with a kernel-based distance between distributions on $\mathcal{X}$, and then we will extend it to distributions on $\mathcal{X}^\infty$. For two probability distributions P and Q on $(\mathcal{X}, \mathcal{F})$ and a set $\mathcal{H}$ of measurable functions on $\mathcal{X}$, define the distance

$$d_{\mathcal{H}}(P, Q) := \sup_{h \in \mathcal{H}} |\mathbf{E}_P h - \mathbf{E}_Q h|.$$

This distance is variously known as probability metric with a ζ-structure [16, 11], integral probability metric [9] and maximum mean discrepancy [5]; here we just call it d.

We will be interested in the cases when $d_{\mathcal{H}}(P, Q) = 0$ implies $P = Q$. Note that in this case $d_{\mathcal{H}}$ is a metric (the rest of the properties are easy to see). Of particular interest to us are sets $\mathcal{H}$ that consist of indicator functions. In this case we can identify each $f \in \mathcal{H}$ with the set $\{x : f(x) = 1\} \subset \mathcal{X}$. Moreover, it is easy to check that $d_{\mathcal{H}}$ is a metric if and only if $\mathcal{H}$ generates $\mathcal{F}$. The latter property is often easy to verify directly. First of all, it trivially holds for the case when $\mathcal{H}$ is the set of halfspaces in a Euclidean $\mathcal{X}$. It is also easy to check that it holds for the most commonly used kernels (in the cases when feature space

is of the same or higher dimension than the input space), such as polynomial, RBF and Gaussian kernels. A general sufficient condition for $d_{\mathcal{H}}$ to be a metric is that $\mathcal{H}$ is a unit ball in a universal RKHS defined on the compact metric space $\mathcal{X}$ [5].

Based on $d_{\mathcal{H}}$ we can construct a distance between time-series probability distributions. For two time-series distributions ρ_1, ρ_2 we take the $d_{\mathcal{H}}$ between k-dimensional marginal distributions of ρ_1 and ρ_2 for each $k \in \mathbb{N}$, and sum them all up with decreasing weights.

Definition 1 (telescope distance D)**.** *For two time series distributions ρ_1 and ρ_2 on the space $(X^\infty, \mathcal{F}_\infty)$ and a sequence of sets of functions $\mathbf{H} = (\mathcal{H}_1, \mathcal{H}_2, \dots)$ define the* telescope distance

$$D_{\mathbf{H}}(\rho_1, \rho_2) \quad := \quad \sum_{k=1}^{\infty} w_k \sup_{h \in \mathcal{H}_k} |\mathbf{E}_{\rho_1} h(X_1, \dots, X_k) \; - \; \mathbf{E}_{\rho_2} h(Y_1, \dots, Y_k)|, \quad (1)$$

where w_k, $k \in \mathbb{N}$ is a sequence of positive summable real weights (e.g. $w_k = 2^{-k}$).

Lemma 1. *$D_{\mathbf{H}}$ is a metric if and only if $d_{\mathcal{H}_k}$ is a metric for every $k \in \mathbb{N}$.*

Proof. The statement follows from the fact that two process distributions are the same if and only if all their finite-dimensional marginals coincide. □

Further, introduce empirical estimates of the telescope distance. This estimate is biased, and, unlike in the i.i.d. case, the bias may depend on the distributions; however, it is $o(n)$ as will be shown shortly.

Definition 2 (empirical telescope distance $\hat{D}$)**.** *For a pair of samples $X_{1..n}$ and $Y_{1..m}$ define* empirical telescope distance *as*

$$\hat{D}_{\mathbf{H}}(X_{1..n}, Y_{1..m}) := \sum_{k=1}^{\min\{m,n\}} w_k \sup_{h \in \mathcal{H}_k} \left| \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} h(X_{i..i+k-1}) - \frac{1}{m-k+1} \sum_{i=1}^{m-k+1} h(Y_{i..i+k-1}) \right|. \quad (2)$$

All the methods presented in this work are based on empirical telescope distance. The key fact is that it is an asymptotically consistent estimate of the telescope distance, that is, the latter can be consistently estimated based on sampling.

Theorem 1. *Let $\mathbf{H} = (\mathcal{H}_1, \mathcal{H}_2, \dots)$, $\mathcal{H}_k \subset \mathcal{X}^k$, $k \in \mathbb{N}$ be a sequence of separable sets of indicator functions of finite VC dimension such that $\mathcal{H}_k$ generates $\mathcal{F}_k$. Then, for every stationary ergodic time series distributions ρ_X and ρ_Y generating samples $X_{1..n}$ and $Y_{1..m}$ we have*

$$\lim_{n,m \to \infty} \hat{D}_{\mathbf{H}}(X_{1..n}, Y_{1..m}) = D_{\mathbf{H}}(\rho_X, \rho_Y) \quad (3)$$

Proof. As is established in [2], under the conditions of the theorem we have

$$\lim_{n\to\infty} \sup_{h\in\mathcal{H}_k} \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} h(X_{i..i+k-1}) = \sup_{h\in\mathcal{H}_k} \mathbf{E}_{\rho_X} h(X_1,\dots,X_k) \ \rho_X\text{-a.s.} \quad (4)$$

for all $k \in \mathbb{N}$, and likewise for ρ_Y. Fix an $\varepsilon > 0$. We can find a $T \in \mathbb{N}$ such that $\sum_{k>T} w_k \le \varepsilon$. Moreover, as follows from (4), for each $k = 1..T$ we can find an N_k such that

$$\left| \sup_{h\in\mathcal{H}_k} \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} h(X_{i..i+k-1}) \ - \ \sup_{h\in\mathcal{H}_k} \mathbf{E}_{\rho_X} h(X_{1..k}) \right| \ \le \ \varepsilon/T \quad (5)$$

Let $N_k := \max_{i=1..T} N_i$ and define analogously M for ρ_Y. Thus, for $n \ge N$, $m \ge M$ we have

$$\begin{aligned}
&\hat{D}_{\mathbf{H}}(X_{1..n}, Y_{1..m}) \\
&\le \sum_{k=1}^{T} w_k \sup_{h\in\mathcal{H}_k} \left| \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} h(X_{i..i+k-1}) - \frac{1}{m-k+1} \sum_{i=1}^{m-k+1} h(Y_{i..i+k-1}) \right| + \varepsilon \\
&\le \sum_{k=1}^{T} w_k \sup_{h\in\mathcal{H}_k} \left\{ \left| \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} h(X_{i..i+k-1}) - \mathbf{E}_{\rho_1} h(X_{1..k}) \right| \right. \\
&\qquad + |\mathbf{E}_{\rho_1} h(X_{1..k}) - \mathbf{E}_{\rho_2} h(Y_{1..k})| \\
&\qquad \left. + \left| \mathbf{E}_{\rho_2} h(Y_{1..k}) - \frac{1}{m-k+1} \sum_{i=1}^{m-k+1} h(Y_{i..i+k-1}) \right| \right\} + \varepsilon \\
&\le 3\varepsilon + D_{\mathbf{H}}(\rho_X, \rho_Y).
\end{aligned}$$

Since ε was chosen arbitrary the statement follows. □

4 Time-series classification

We start with a conceptually simple problem of time-series classification, also known as the three-sample problem. We are given three samples $X = (X_1, \dots, X_n)$, $Y = (Y_1, \dots, Y_m)$ and $Z = (Z_1, \dots, Z_l)$. It is known that X and Y were generated by different time-series distributions, whereas Z was generated by the same distribution as either X or Y. It is required to find our which one is the case. Both distributions are assumed to be stationary ergodic, but no further assumptions are made about them (no independence, mixing or memory assumptions).

Note that this problem is very different from the traditional problem of (binary) classification of i.i.d. samples. Indeed, in the latter problem one typically

assumes that a large number of (i.i.d.) training samples from each of the two classes is given, and possibly a large number of (i.i.d.) testing samples has to be classified. In contrast, in the-time series classification problem we are given just two training points — one from each class — and one testing point to classify.

The three sample problem for dependent time series has been addressed in [6] for Markov processes and in [14] for stationary ergodic time series. The latter work uses an approach based on the distributional distance, whose empirical estimates are based on counting frequencies.

Indeed, to solve this problem it suffices to have consistent estimates of some distance between time series distributions. Thus, we can use the telescope distance. The following statement is a simple corollary of Theorem 1.

Theorem 2. *Let the samples $X = (X_1, \dots, X_n)$, $Y = (Y_1, \dots, Y_m)$ and $Z = (Z_1, \dots, Z_l)$ be generated by stationary ergodic distributions ρ_X, ρ_Y and ρ_Z, with $\rho_X \neq \rho_Y$ and either (i) $\rho_Z = \rho_X$ or (ii) $\rho_Z = \rho_Y$. Assume that the sets $\mathcal{H}_k \subset \mathcal{X}^k$, $k \in \mathbb{N}$ are separable sets of indicator functions of finite VC dimension such that $\mathcal{H}_k$ generates $\mathcal{F}_k$. A test that declares (i) if $\hat{D}_{\mathbf{H}}(Z, X) \leq \hat{D}_{\mathbf{H}}(Z, Y)$ and (ii) otherwise makes only finitely many errors with probability 1 as $n, m, l \to \infty$.*

It is straightforward to extend this theorem to more than two classes; in other words, instead of X and Y one can have an arbitrary number of samples generated by different stationary ergodic distributions.

5 Clustering time series

We are given N samples $X^1 = (X_1^1, \dots, X_{n_1}^1), \dots, X^N = (X_1^N, \dots, X_{n_N}^N)$ generated by k different stationary ergodic time-series distributions $\rho_1, \dots, \rho_k$. The number k is known, but the distributions are not. It is required to group the N samples into k groups (clusters), that is, to output a partitioning of $\{X_1..X_N\}$ into k sets. While there may be many different approaches to define what is a good clustering (and in general, deciding what is a good clustering is a difficult problem), for the problem of classifying time-series samples there is a natural choice: those samples should be put together that were generated by the same distribution. Thus, define *target clustering* as the partitioning in which those and only those samples that were generated by the same distribution are placed in the same cluster. A clustering algorithm is called *asymptotically consistent* if with probability 1 there is an n' such that the algorithm produces the target clustering whenever $\max_{i=1..N} n_i \geq n'$. (This setup is after [12].)

Again, to solve this problem it is enough to have a metric between time-series distributions that can be consistently estimated, and our approach here is to go with the telescope distance, and thus to use $\hat{D}$.

The clustering problem is relatively simple if the target clustering has what is called the *strict separation property* [3]: every two points in the same cluster are closer to each other than to any point from a different cluster. The following statement is an easy corollary of Theorem 1.

Theorem 3. *Assume that the sets $\mathcal{H}_k \subset \mathcal{X}^k$, $k \in \mathbb{N}$ are separable sets of indicator functions of finite VC dimension, and such that $\mathcal{H}_k$ generates $\mathcal{F}_k$. If the distributions $\rho_1, \dots, \rho_k$ generating the samples $X^1 = (X_1^1, \dots, X_{n_1}^1), \dots, X^N = (X_1^N, \dots, X_{n_N}^N)$ are stationary ergodic, then with probability 1 from some $n := \max_{i=1..N} n_i$ on the target clustering has the strict separation property with respect to $\hat{D}_{\mathbf{H}}$.*

With the strict separation property at hand, it is easy to find asymptotically consistent algorithms. We will give a couple of simple examples, but the theorem below can be extended to many other distance-based clustering algorithms.

The *single linkage* algorithm works as follows. The distance between clusters is defined as the distance between closest points in these clusters. First, put each point into a separate cluster. Then, merge the two closest clusters; repeat the last step until the total number of clusters is k.

The *farthest point* clustering works as follows. Assign $c_1 := X^1$ to the first cluster. For $i = 2..k$, find the point X^j, $j \in \{1..N\}$ that maximizes the distance $\min_{t=1..i} \hat{D}_{\mathbf{H}}(X^j, c_t)$ (to the points already assigned to clusters) and assign $c_i := X^j$ to the cluster i. Then assign each of the remaining points to the nearest cluster. The following statement is a corollary of Theorem 3.

Theorem 4. *Under the conditions of Theorem 3, single linkage and farthest point clusterings are asymptotically consistent.*

Note that we do not require the samples to be independent; the joint distributions of the sample may be completely arbitrary, as long as the marginal distribution of each sample is stationary ergodic.

6 Speed of convergence

The results established so far are asymptotic out of necessity: they are established under the assumption that the distributions involved are stationary ergodic, which is too general to allow for any meaningful finite-time performance guarantees. Moreover, some statistical problems, such as homogeneity testing or clustering when the number of clusters is unknown, are provably impossible to solve under this assumption [13].

Therefore, while it is interesting to be able to establish consistency results under the most general assumptions (stationary ergodic distributions), it is also interesting to look what results can be obtained under stronger assumptions. Moreover, since it is usually not known in advance whether the data at hand satisfies given assumptions or not, it appears important to have methods that have both asymptotic consistency in the general setting and finite-time performance guarantees under stronger assumptions.

In this section we will look at the speed of convergence of $\hat{D}$ under certain mixing conditions, and use it to construct solutions for the problems of homogeneity and clustering with an unknown number of clusters, as well as to establish finite-time performance guarantees for the methods presented in the previous sections.

6.1 β-mixing

A stationary distribution on the space of one-way infinite sequences $(\mathcal{X}^{\mathbb{N}}, \mathcal{F}_{\mathbb{N}})$ can be uniquely extended to a stationary distribution on the space of two-way infinite sequences $(\mathcal{X}^{\mathbb{Z}}, \mathcal{F}_{\mathbb{Z}})$ of the form $\ldots, X_{-1}, X_0, X_1, \ldots$.

Definition 3 (β-mixing coefficients)**.** *For a process distribution ρ define the mixing coefficients*

$$\beta(\rho, k) := \sup_{\substack{A\in\sigma(X_{-\infty..0}),\\ B\in\sigma(X_{k..\infty})}} |\rho(A\cap B) - \rho(A)\rho(B)|$$

where $\sigma(..)$ denotes the sigma-algebra of the random variables in brackets.

When $\beta(\rho, k) \to 0$ the process ρ is called absolutely regular; this condition is much stronger than ergodicity.

6.2 Speed of convergence of $\hat{D}$

Assume that a sample $X_{1..n}$ is generated by a distribution ρ that is uniformly β-mixing with coefficients $\beta(\rho, k)$ Assume further that $\mathcal{H}_k$ is a set of indicator functions with a finite VC dimension d_k, for each $k \in \mathbb{N}$.

The general tool that we use to obtain performance guarantees in this section is the following bound that can be obtained from the results of [7].

$$q_n(\rho, \mathcal{H}_k, \varepsilon) := \rho\Big(\sup_{h\in\mathcal{H}_k} |\frac{1}{n-k+1} \sum_{i=1}^{n-k+1} h(X_{i..i+k-1}) - \mathbf{E}_{\rho_1} h(X_{1..k})| > \varepsilon\Big)$$
$$\leq n\beta(\rho, t_n - k) + 8t_m^{d_k+1} e^{-l_n\varepsilon^2/8}, \quad (6)$$

where t_n and l_n are any integers in $1..n$. The parameters t_n, l_n should be set according to the values of β in order to optimize the bound. The bound (6) is a finite-sample version of the asymptotic result (4) that we used in the proof of Theorem 1.

One can use similar bounds for classes of finite Pollard dimension [10] or more general bounds expressed in terms of covering numbers, such as those given in [7]. Here we consider classes of finite VC dimension only for the ease of the exposition and for the sake of continuity with the previous section (where it was necessary).

Furthermore, for the rest of this section we assume geometric β-mixing distributions, that is, $\beta(\rho, t) \leq \gamma^t$ for some $\gamma < 1$. Letting $l_n = t_n = \sqrt{n}$ the bound (6) becomes

$$q_n(\rho, \mathcal{H}_k, \varepsilon) \leq n\gamma^{\sqrt{n}-k} + 8n^{(d_k+1)/2} e^{-\sqrt{n}\varepsilon^2/8}. \quad (7)$$

Lemma 2. *Let two samples $X_{1..n}$ and $Y_{1..m}$ be generated by stationary distributions ρ_X and ρ_Y whose β-mixing coefficients satisfy $\beta(\rho_., t) \leq \gamma^t$ for some*

$\gamma < 1$. *Let H_k, $k \in \mathbb{N}$ be some sets of indicator functions on $\mathcal{X}^k$ whose VC dimension d_k is finite and non-decreasing with k. Then*

$$P(|\hat{D}_{\mathbf{H}}(X_{1..n}, Y_{1..m}) - D_{\mathbf{H}}(\rho_X, \rho_Y)| > \varepsilon) \leq 2\Delta(\varepsilon/4, n') \tag{8}$$

where $n' := \min\{n_1, n_2\}$, the probability is with respect to $\rho_X \times \rho_Y$ and

$$\Delta(\varepsilon, n) := -\log \varepsilon (n\gamma^{\sqrt{n}+\log(\varepsilon)} + 8n^{(d_{-\log\varepsilon}+1)/2} e^{-\sqrt{n}\varepsilon^2/8}) \tag{9}$$

Proof. Note that $\sum_{k=-\log\varepsilon/2}^{\infty} w_k < \varepsilon/2$. Using this and the definitions 1 and 2 of $D_{\mathbf{H}}$ and $\hat{D}_{\mathbf{H}}$ we obtain

$$P(|\hat{D}_{\mathbf{H}}(X_{1..n_1}, Y_{1..n_2}) - D_{\mathbf{H}}(\rho_X, \rho_Y)| > \varepsilon) \leq \sum_{k=1}^{-\log(\varepsilon/2)} (q_n(\rho_X, \mathcal{H}_k, \varepsilon/4) + q_n(\rho_Y, \mathcal{H}_k, \varepsilon/4)),$$

which together with (7) implies the statement. □

6.3 Homogeneity testing

Given two samples $X_{1..n}$ and $Y_{1..m}$ generated by distributions ρ_X and ρ_Y respectively, the problem of homogeneity testing (or the two-sample problem) consists in deciding whether $\rho_X = \rho_Y$. A test is called (asymptotically) consistent if its probability of error goes to zero as $n' := \min\{m, n\}$ goes to infinity. In general, for stationary ergodic time series distributions, there is no asymptotically consistent test for homogeneity [13], so stronger assumptions are in order.

Homogeneity testing is one of the classical problems of mathematical statistics, and one of the most studied ones. Huge literature exits on homogeneity testing for i.i.d. data, and for dependent processes as well. We do not attempt to survey this literature here. Kernel-based methods (based on $d_{\mathcal{H}}$) for this problems were used in [5]. Our contribution to this line of research is to show that kernel-based methods can be used (via the telescope distance) for this problem in the case of dependent (β-mixing) processes.

It is easy to see that under the mixing conditions of Lemma 1 a consistent test for homogeneity exists, and finite-sample performance guarantees can be obtained. It is enough to find a sequence $\varepsilon_n \to 0$ such that $\Delta(\varepsilon_n, n) \to 0$ (see (9). Then the test can be constructed as follows: say that the two distributions $X_{1..n}$ and $Y_{1..m}$ were generated by the same distribution if $\hat{D}_{\mathbf{H}}(X_{1..n}, Y_{1..m}) < \varepsilon_{n'}$, where $n' := \min\{n, m\}$; otherwise say that they were generated by different distributions. The following statement is an immediate consequence of Lemma 2.

Theorem 5. *Under the conditions of Lemma 2 the probability of Type I error (the distributions are the same but the test says they are different) of the described test is upper-bounded by $4\Delta(\varepsilon/8, n')$. The probability of Type II error (the distributions are different but the test says they are the same) is upper-bounded by $4\Delta(\delta - \varepsilon/8, n')$ where $\delta := 1/2 D_{\mathbf{H}}(\rho_X, \rho_Y)$.*

The optimal choice of ε_n may depend on the speed of increase of d_k (the VC dimension of $\mathcal{H}_k$); however, for most natural cases (recall that $\mathcal{H}_k$ are also parameters of the algorithm) this growth is polynomial so the main term to control is $e^{-\sqrt{n}\varepsilon^2/8}$.

For example, if $\mathcal{H}_k$ is the set of halfspaces in $\mathcal{X}^k = \mathbb{R}^k$ then $d_k = k + 1$ and one can chose $\varepsilon_n := n^{1/8}$. The resulting probability of Type I error decreases as $\exp(-n^{1/4})$.

6.4 Clustering with a known or unknown number of clusters

If the distributions generating the samples satisfy mixing conditions, then we can augment Theorems 3 and 4 with finite-sample performance guarantees.

Theorem 6. *Let the distributions $\rho_1, \dots, \rho_k$ generating the samples $X^1 = (X^1_1, \dots, X^1_{n_1}), \dots, X^N = (X^N_1, \dots, X^N_{n_N})$ satisfy the conditions of Lemma 2. Define $\delta := \min_{i,j=1..N, i\neq j} D_{\mathbf{H}}(\rho_i, \rho_j)$ and $n := \min_{i=1..N} n_i$. Then with probability at least*

$$1 - N(N-1)\Delta(\delta/4, n)/2$$

the target clustering of the samples has the strict separation property. In this case single linkage and farthest point algorithms output the target clustering.

Proof. Note that a sufficient condition for the strict separation property to hold is that for every one out of $N(N-1)/2$ pairs of samples the estimate $\hat{D}_{\mathbf{H}}(X^i, X^j)$ $i, j = 1..N$ is within $\delta/4$ of the $D_{\mathbf{H}}$ distance between the corresponding distributions. It remains to apply Lemma 2 to obtain the first statement, and the second statement is obvious (cf. Theorem 4). □

For stationary ergodic distributions since it is (in general) not possible to tell whether two samples have been generated by the same distribution or by different ones (homogeneity testing), it is also impossible to have a consistent clustering algorithm when the number of clusters k is unknown. Again, the situation changes if the distributions satisfy the mixing conditions. Indeed, it is easy to see that there exists a consistent clustering algorithm in this case, with unknown k.

Such an algorithm can be obtained as follows: assign to the same cluster all samples that are at most ε_n-far from each other, where the threshold ε_n is selected the same way as for homogeneity testing: $\varepsilon_n \to 0$ and $\Delta(\varepsilon_n, n) \to 0$. The optimal choice of this parameter depends on the choice of $\mathcal{H}_k$ through the speed of growth of the VC dimension d_k of these sets.

Theorem 7. *Given N samples generated by k different stationary distributions ρ_i, $i = 1..k$ (unknown k) all satisfying the conditions of Lemma 2, the probability of error (misclustering at least one sample) of the described algorithm is upper-bounded by*

$$2N(N-1)\max\{\Delta(\varepsilon/8, n), \Delta(\delta - \varepsilon/8, n)\}$$

where $\delta := \min_{i,j=1..k, i\neq j} D_{\mathbf{H}}(\rho_i, \rho_j)$ *and* $n = \min_{i=1..N} n_i$*, with* n_i*,* $i = 1..N$ *being lengths of the samples.*

7 Computational issues

While the main results of this paper are theoretical, all the methods presented can be implemented relatively easily. One of the advantages of the presented approach is that it allows one to reuse the methods already implemented for solving the binary classification problem for solving the problems considered here. Specifically, in order to compute the distance $\hat{D}_{\mathbf{H}}(\rho_X, \rho_Y)$ between two samples X and Y, one can use an SVM (e.g. [15]) to compute each of the summands in (2). To do this, an SVM has to be trained on the samples $X_{i..i+k-1}$, $i = 1..n-k+1$, $Y_{j..j+k-1}$, $j = 1..m-k+1$, and evaluated on the same data. While in (2) the number of summands is $n' = \min\{m, n\}$, it can be replaced with any $\gamma_{n'}$ such that $\gamma_{n'} \to \infty$, without affecting any asymptotic consistency results. A practically viable choice is $\gamma_n = \log n$; in fact, there is no reason to chose faster growing γ_n since the estimates for higher-order summands will not have enough data to converge (which, again, has no impact on the consistency properties of $\hat{D}$). Thus, the computation of $\hat{D}_{\mathbf{H}}$ can be realized at the cost of training and testing $\log n$ SVMs.

8 Comparison with the distributional distance

The results presented in the previous sections are all based on the property that the telescope distance on process distributions can be consistently estimated based on sampling, if any stationary ergodic distributions are chosen to generate the samples. It is interesting to compare this metric to other metrics that have this property. The so-called distributional distance [4] has this property, and it was used in [14, 12] to construct solutions to various statistical problems, including process classification and clustering; thus, it is a natural candidate for comparison.

Here we will show that the telescope distance is stronger in the topological sense. Since in fact both the telescope distance and the distributional distance (to be introduced shortly) are families of distances (the telescope distance depends on the sequence $\mathbf{H}$), we will fix a simple natural choice of each of these metrics. In general, different choices of parameters produce topologically non-equivalent metrics; it is easy to check that the analysis in this section extends to many other natural choices, but the general case is outside of scopes of this paper.

Thus, for the purpose of this section, let us fix $\mathcal{X} = \mathbb{R}$ and let H_k^0 be the set of halfspaces in $\mathcal{X}^k$. Denote $\mathbf{H}^0 := (\mathcal{H}_k^0 : k \in \mathbb{N})$. Let also $w_k := 2^{-k}$, so that the telescope distance $d_{\mathbf{H}}$ is defined precisely (cf. definition 1). Clearly, these $\mathcal{H}_k$ satisfy all the conditions of the theorems of Sections 4 and 5.

Next we introduce the distributional distance. The version below is the

one used in [12] for clustering time series. This distance is based on counting frequencies of cells in different partitionings of $\mathcal{X}^k$, for each $k \in \mathbb{N}$, and summing them up with weights.

More formally, for each $k, l \in \mathbb{N}$, let $B^{k,l}$ be the partition of the set $\mathcal{X}^k$ into k-dimensional cubes with volume $h_l^k = (1/l)^k$ (the cubes start at 0). Moreover, define $B^k = \cup_{l\in\mathbb{N}} B^{k,l}$ and $\mathcal{B} = \cup_{k=1}^{\infty} B^k$. Note that the set $\{B \times \mathcal{X}^\infty : B \in B^{k,l}, k, l \in \mathbb{N}\}$ generates the Borel σ-algebra on $\mathbb{R}^\infty = \mathcal{X}^\infty$. For a sequence $\mathbf{x} \in \mathcal{X}^n$ and a set $B \in \mathcal{B}$ denote $\nu(\mathbf{x}, B)$ the frequency with which the sequence $\mathbf{x}$ falls in the set B.

$$\nu(\mathbf{x}, B) \quad := \quad \begin{cases} \frac{1}{n-|B|+1} \sum_{i=1}^{n-|B|+1} I_{\{(X_i,\ldots,X_{i+|B|-1})\in B\}} & \text{if } n \geq |B|, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 4 (distributional distance)**.** *The distributional distance is defined for a pair of processes ρ_1, ρ_2 as follows*

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

where $w_j = 2^{-j}$.

One can use other partitions $B^{m,l}$ and other weights, obtaining different versions of the distance. See [4] for a general treatment.

A metric d_1 is called *stronger* than a metric d_2 if any sequence that converges in d_1 also converges in d_2.

Theorem 8. *$D_{\mathbf{H}^0}$ is stronger than D_{dd}.*

Proof. Fix any $\varepsilon > 0$ and find a $T \in \mathbb{N}$ such that $\sum_{m,l>T} w_m w_l < \varepsilon$. Let ρ_i, $i \in \mathbb{N}$ be a sequence of process measures that converges in $D_{\mathbf{H}^0}$. Note that, since any cube can be obtained by intersecting finitely many halfspaces, for any cube $B \in B^k$ we have

$$|\rho_i(B) - \rho_j(B)| \leq 2^k d_{\mathcal{H}_k}(\rho_i, \rho_j) \leq 2^{2k} D_{\mathbf{H}^0}(\rho_i, \rho_j), \tag{10}$$

where the second inequality follows from the definition of $D_{\mathbf{H}^0}$. Therefore, the sequence $\rho_i(B)$ converges for every $B \in \mathcal{B}$. It follows that for each $k \in \mathbb{N}$ we can find a cube $A_k \in B^k$ such $\rho_j(A_k) > 1 - \varepsilon$ for all $j > j_k$; let $J := \max_{i\leq T} j_i$. Using (10) and the definition of the partitions $B^{k,l}$ we can derive

$$\sum_{B\in B^{k,l}, B\subset A_k} |\rho_i(B) - \rho_j(B)| \leq 2^{4kl} D_{\mathbf{H}^0}(\rho_i, \rho_j) \tag{11}$$

for any $i, j \geq J$. From the fact that the sequence $\rho_i, i \in \mathbb{N}$ converges in $D_{\mathbf{H}}$, one can see that, increasing J if necessary, we can obtain $D_{\mathbf{H}^0}(\rho_i, \rho_j) \leq 2^{-4T^2}\varepsilon$ for

all $i, j \geq J$. From this and (11) we obtain

$$D_{dd}(\rho_i, \rho_j) \leq \sum_{m,l=1}^{T} w_m w_l \sum_{B \in B^{m,l}, B \subset A_m} |\rho_i(B) - \rho_j(B)| + 3\varepsilon$$
$$\leq 2^{4T^2} D_{\mathbf{H}^0}(\rho_i, \rho_j) + 3\varepsilon \leq 4\varepsilon,$$

for all $i, j > J$, which means that the sequence $\rho_i, i \in \mathbb{N}$ converges in D_{dd}. $\square$

9 Outlook and conclusion

We have presented a distance between process distributions that can be employed to harness known kernel methods for the binary classification problem (such as SVM) to solve various statistical problems that concern highly-dependent time series. The time-series problems that we have considered include unsupervised learning problems (such as clustering), and thus are seemingly very far from binary classification of i.i.d. data, which is used as a building block. It is clear that the presented approach is not limited to the problems considered here, but can be used in any distance-based method for solving other learning problems. For example, one can use nearest neighbors-based approaches along with the telescope distance to solve regression or (multi-class) classification problems, where each data-point is a time-series sample.

The "telescoping" trick used in the definition of the distance $D_{\mathbf{H}}$ is not limited to the kernel-based distance $d_{\mathcal{H}}$. First of all, it is also used in the definition of the distributional distance, a distance that can be applied to clustering time series as in [12]. Moreover, any metric d' between distributions on $\mathcal{X}^k$ can be "telescoped:" just substitute d' for $d_{\mathcal{H}}$ in (1). The resulting distance between process distributions can be used to solve the problems we have considered here, as long as it can be consistently estimated based on sampling (as in Theorem 1). However, the latter property may not be easy to establish, especially for the case of arbitrary stationary ergodic processes.

References

[1] Terrence M. Adams and Andrew B. Nobel. Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. *The Annals of Probability*, 38:1345–1367, 2010.

[2] Terrence M. Adams and Andrew B. Nobel. Uniform approximation and bracketing properties of VC classes. *Bernoulli*, to appear.

[3] M.F. Balcan, A. Blum, and S. Vempala. A discriminative framework for clustering via similarity functions. In *STOC*, 2008.

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

[5] Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schlkopf, and Alexander J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.

[6] M. Gutman. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Transactions on Information Theory*, 35(2):402–408, 1989.

[7] R.L. Karandikara and M. Vidyasagar. Rates of uniform convergence of empirical means with mixing processes. *Statistics and Probability Letters*, 58:297–307, 2002.

[8] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*, VLDB'04, pages 180–191, 2004.

[9] Alfred Muller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[10] D. Pollard. *Convergence of Stochastic Processes*. Springer, 1984.

[11] S. T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. Wiley, 1991.

[12] D. Ryabko. Clustering processes. In *Proc. the 27th International Conference on Machine Learning (ICML 2010)*, pages 919–926, Haifa, Israel, 2010.

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

[14] D. Ryabko and B. Ryabko. Nonparametric statistical inference for ergodic processes. *IEEE Transactions on Information Theory*, 56(3):1430–1435, 2010.

[15] V.N. Vapnik. *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, 1998.

[16] V. M. Zolotarev. Probability metrics. *Theory of Probability and Its Applications.*, 28(2):264–287, 1983.