ASYMPTOTIC STATISTICAL ANALYSIS OF STATIONARY ERGODIC TIME SERIES

Daniil Ryabko

INRIA Lille,
40, avenue Halley
Parc Scientifique de la Haute Borne
59650 Villeneuve d'Ascq, France
daniil@ryabko.net

ABSTRACT

It is shown how to construct asymptotically consistent efficient algorithms for various statistical problems concerning stationary ergodic time series. The considered problems include clustering, hypothesis testing, change-point estimation and others. The presented approach is based on empirical estimates of the distributional distance. Some open problems are also discussed.

1. INTRODUCTION

Statistical problems involving time-series data arise in a variety of modern applications, including biology, finance, network analysis, etc. These applications often dramatically violate traditional statistical assumptions imposed on time series. This applies not only to parametric models, but even to assumptions that are often considered non-parametric, for example that the data points are independent or that the time series have limited memory, or that the processes mix sufficiently fast and so on.

Here I summarize some recent work on statistical analysis of time series where the only assumption on the time series is that they are stationary ergodic. No independence or mixing-type assumptions are involved.

The considered problems are hypothesis testing, clustering, the two- and three-sample problems, and change point estimation. The main results establish asymptotically consistent algorithms for the considered problems. The consistency results follow from the simple fact that the so-called distributional distance [1] can be estimated based on sampling; this contrasts previous results that show that the $\bar{d}$ distance can not (in general) be estimated for stationary ergodic processes [2]. For more details on these results see [3, 4, 5, 6, 7].

2. PRELIMINARIES

Let A be an alphabet, and denote A^* the set of tuples $\cup_{i=1}^{\infty} A^i$. In this work we consider the case $A = \mathbb{R}$; extensions to the multidimensional case, as well as to more general spaces, are straightforward. Distributions, or (stochastic) processes, are measures on the space $(A^\infty, \mathcal{F}_{A^\infty})$, where $\mathcal{F}_{A^\infty}$ is the Borel sigma-algebra of A^∞. When talking about joint distributions of N samples, we mean distributions on the space $((A^N)^\infty, \mathcal{F}_{(A^N)^\infty})$.

For each $k, l \in \mathbb{N}$, let $B^{k,l}$ be the partition of the set A^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 $\mathcal{B} = \cup_{k=1}^{\infty} B^k$. The set $\{B \times A^\infty : B \in B^{k,l}, k, l \in \mathbb{N}\}$ generates the Borel σ-algebra on $\mathbb{R}^\infty = A^\infty$. For a set $B \in \mathcal{B}$ let $|B|$ be the index k of the set B^k that B comes from: $|B| = k : B \in B^k$.

We use the abbreviation $X_{1..k}$ for $X_1, \dots, X_k$. For a sequence $\mathbf{x} \in A^n$ 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.

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

A process ρ is *stationary* if

$$\rho(X_{1..|B|} = B) = \rho(X_{t..t+|B|-1} = B)$$

for any $B \in A^*$ and $t \in \mathbb{N}$. We further abbreviate $\rho(B) := \rho(X_{1..|B|} = B)$. A stationary process ρ is called *(stationary) ergodic* if the frequency of occurrence of each word B in a sequence $X_1, X_2, \dots$ generated by ρ tends to its a priori (or limiting) probability a.s.:

$$\rho(\lim_{n\to\infty} \nu(X_{1..n}, B) = \rho(B)) = 1.$$

Denote $\mathcal{E}$ the set of all stationary ergodic processes.

Definition 1 (distributional distance)**.** *The distributional distance is defined for a pair of processes ρ_1, ρ_2 as follows (e.g. [1])*

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

where $w_j = 1/j^2$.

(The weights in the definition are fixed for the sake of concreteness only; we could take any other summable sequence of positive weights instead.) In words, we are taking a sum over a series of partitions into cubes of decreasing volume (indexed by l) of all sets A^k, $k \in \mathbb{N}$, and count the differences in probabilities of all cubes in all these partitions. These differences in probabilities are weighted: smaller weights are given to larger k and finer partitions. It is easy to see that d is a metric. We refer to [1] for more information on this metric and its properties.

The methods below are based on *empirical estimates of the distance* d:

$$\hat{d}(X^1_{1..n_1}, X^2_{1..n_2}) = \sum_{m,l=1}^{\infty} w_m w_l \sum_{B \in B^{m,l}} |\nu(X^1_{1..n_1}, B) - \nu(X^2_{1..n_2}, B)|, \quad (1)$$

where $n_1, n_2 \in \mathbb{N}$, $\rho \in \mathcal{S}$, $X^i_{1..n_i} \in A^{n_i}$. Although the expression (1) involves taking three infinite sums, it will be shown below that it can be easily calculated (see Section 4).

3. ASYMPTOTIC CONSISTENCY RESULTS

The consistency results are based on the following statement, which is quite easy to derive from the definition of ergodicity (or from Birkhoff's ergodic theorem).

Lemma 1 (**$\hat{d}$ is consistent**)**.** *Let $\rho_1, \rho_2 \in \mathcal{E}$ and let two samples $\mathbf{x}_1 = X^1_{1..n_1}$ and $\mathbf{x}_2 = X^2_{1..n_2}$ be generated by a distribution ρ such that the marginal distribution of $X^i_{1..n_i}$ ρ_i is stationary ergodic for $i = 1, 2$. Then*

$$\lim_{n_1, n_2 \to \infty} \hat{d}(X^1_{1..n_1}, X^2_{1..n_2}) = d(\rho_1, \rho_2) \ \rho\text{–}a.s.$$

3.1. The three-sample problem

The first problem we consider is the three-sample problem, also known as process classification. Let there be given three samples $X = (X_1, \dots, X_k), Y = (Y_1, \dots, Y_m)$ and $Z = (Z_1, \dots, Z_n)$. Each sample is generated by a stationary ergodic process ρ_X, ρ_Y and ρ_Z respectively. Moreover, it is known that either $\rho_Z = \rho_X$ or $\rho_Z = \rho_Y$, but $\rho_X \neq \rho_Y$. We wish to construct a test that, based on the finite samples X, Y and Z will tell whether $\rho_Z = \rho_X$ or $\rho_Z = \rho_Y$.

The proposed test chooses the sample X or Y according to whichever is closer to Z in $\hat{d}$. That is, we define the test $G(X, Y, Z)$ as follows. If $\hat{d}(X, Z) \leq \hat{d}(Y, Z)$ then the test says that the sample Z is generated by the same process as the sample X, otherwise it says that the sample Z is generated by the same process as the sample Y.

Theorem 1. *The described test makes only a finite number of errors with probability 1, when $|X|, |Y|$ and $|Z|$ go to infinity.*

The statement is easy to derive from Lemma 1. Note that X, Y, Z are not required to be independent. All we need is that the distributions are stationary ergodic (more formally, the distribution generating the three sequences is arbitrary except for the fact that the marginals are stationary ergodic).

3.2. Time-series clustering

A more general but closely related problem is time-series clustering. We are given N samples $\mathbf{x}_1, \dots, \mathbf{x}_N$, where each sample $\mathbf{x}_i$ is a string of length n_i of symbols from A: $\mathbf{x}_i = X^i_{1..n_i}$. Each sample is generated by one out of k different *unknown* stationary ergodic distributions $\rho_1, \dots, \rho_k \in \mathcal{E}$. Thus, there is a partitioning $I = \{I_1, \dots, I_k\}$ of the set $\{1..N\}$ into k *disjoint* subsets $I_j, j = 1..k$

$$\{1..N\} = \cup_{j=1}^k I_j,$$

such that $\mathbf{x}_j$, $1 \leq j \leq N$ is generated by ρ_j if and only if $j \in I_j$. The partitioning I is called the *target clustering* and the sets $I_i, 1 \leq i \leq k$, are called the *target clusters*. Given samples $\mathbf{x}_1, \dots, \mathbf{x}_N$ and a target clustering I, let $I(\mathbf{x})$ denote the cluster that contains $\mathbf{x}$.

It is required to partition the index set $\{1..N\}$ in such a way that as the length of each sequence grows the partitioning coincides with the target clustering from some time on with probability 1. Such an algorithm is called asymptotically consistent. In other words, when the sequences are long enough, we have to group together those and only those sequences that were generated by the same distributions.

This can be done as follows. The point $\mathbf{x}_1$ is assigned to the first cluster. Next, find the point that is farthest away from $\mathbf{x}_1$ in the empirical distributional distance $\hat{d}$, and assign this point to the second cluster. For each $j = 3..k$, find a point that maximizes the minimal distance to those points already assigned to clusters, and assign it to the cluster j. Thus we have one point in each of the k clusters. Next simply assign each of the remaining points to the cluster that contains the closest points from those k already assigned. One can notice that the described algorithm just one iteration of the k-means algorithm, with so-called farthest-point initialization and using the distance $\hat{d}$.

Theorem 2. *The described algorithm is strongly asymptotically consistent provided ρ_i is stationary ergodic for each $i = 1..k$.*

3.3. Change-point estimation

Next we consider the change-point problem. The sample $Z = (Z_1, \dots, Z_n)$ consists of two concatenated parts $X = (X_1, \dots, X_k)$ and $Y = (Y_1, \dots, Y_m)$, where $m = n - k$, so that $Z_i = X_i$ for $1 \leq i \leq k$ and $Z_{k+j} = Y_j$ for $1 \leq j \leq m$. The samples X and Y are generated independently by two different stationary ergodic processes with alphabet $A = \mathbb{R}$. The distributions of the processes are unknown. The value k is called the *change point*. It is assumed that k is linear in n; more precisely, $\alpha n < k < \beta n$ for some $0 < \alpha \leq \beta < 1$ from some n on.

It is required to estimate the change point k based on the sample Z.

Note that we do not assume that the single-dimensional marginals before and after the change point are different, as is done almost exclusively in the literature on this problem. We are in the most general situation where the time-series distributions are different, i.e. the change may be only in the long-range dependence.

For each t, $1 \leq t \leq n$, denote U^t the sample $(Z_1, \dots, Z_t)$ consisting of the first t elements of the sample Z, and denote V^t the remainder $(Z_{t+1}, \dots, Z_n)$.

Define the change point estimate $\hat{k} : A^* \to \mathbb{N}$ as follows:

$$\hat{k}(X_1, \dots, X_n) := \operatorname{argmax}_{t \in [\alpha n, n-\beta n]} \hat{d}(U^t, V^t).$$

The following theorem establishes asymptotic consistency of this estimator.

Theorem 3. *For the estimate $\hat{k}$ of the change point k we have*

$$\frac{1}{n}|\hat{k} - k| \to 0 \text{ a.s.}$$

where n is the size of the sample, and when $k, n-k \to \infty$ in such a way that $\alpha < \frac{k}{n} < \beta$ for some $\alpha, \beta \in (0,1)$ from some n on.

This result can be extended [7] to multiple change points and unknown α and β, although the algorithm becomes much more sophisticated.

3.4. Impossibility results: the two-sample problem and its implications

For the problems considered above we have relatively simple algorithms that are asymptotically consistent under most general assumptions. What is more, the proofs of consistency (although mostly omitted here) are quite simple as well. From this one can get the impression that asymptotic consistency results are very easy to obtain and probably they hold for all other interesting problems as well.

This is not the case. The first example is another classical statistical problem: homogeneity testing, also known as the two-sample problem. We are given two samples $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_n)$ generated by two stationary ergodic distributions ρ_X and ρ_Y. We want to tell whether they were generated by the same or by different distributions, that is, whether $\rho_X = \rho_Y$. We are willing to settle for a rather weak asymptotic result. Say a two-sample test $L(X, Y)$, that takes two samples and outputs 0 or 1, is asymptotically consistent if $\mathbf{E}L \to 1$ as $n \to \infty$ if $\rho_X = \rho_Y$ and $\mathbf{E}L \to 0$ otherwise. Moreover, we can further assume that the samples are binary-valued and there is no dependence between X and Y. This does not help:

Theorem 4. *There is no asymptotically consistent two-sample test.*

This result holds even if we additionally require ρ_X and ρ_Y to be B-processes [5], contrasting earlier results of Ornstein and Weiss for this class of processes [2]. The proof (omitted here) relies on a counterexample which is a limit of hidden Markov processes with a countably infinite state space, using a method similar to that of [8].

As a consequence of this negative result, we can also derive impossibility results for some generalizations of the problems considered above.

Corollary 1. *Under the assumptions of theorems 2 and 3 respectively, there is no asymptotically consistent clustering algorithm when the number of clusters is unknown, and there is no asymptotically consistent change-point detection algorithm.*

3.5. Hypothesis testing

Some of the problems considered above, as well as many other interesting problems, cab be formulated in the following way. Consider two sets H_0 and H_1 which are subsets of the set of all stationary ergodic processes, and let there be given a sample $X_1, \dots, X_n$ generated by a stationary ergodic process distribution ρ. We want to tell whether $\rho \in H_0$ or $\rho \in H_1$. The problem arises to characterize those pairs (H_0, H_1) for which this is possible in some asymptotic sense, that is, whether asymptotically consistent tests exists. It turns out that the distributional distance can be used to answer this question to a considerable extent.

To define the notion of consistency we use for this problem, recall that Type I error is said to occur if the test says "1" while the sample was generated by the distribution from H_0. Type II error occurs if the test says "0" while H_1 is true. In many practical situations, these errors may have very different meaning: for example, this is the case when H_0 is interpreted as that a patient has a certain ailment, and H_1 that he does not. In such cases, one may wish to treat the errors asymmetrically. Also H_0 can often be much simple than the alternative H_1, for example, H_0 can be a simple parametric family, or it may consist of just one process distribution, while H_1 can be the complement of H_0 to the set of all stationary ergodic processes.

Call a test *consistent* if, for any pre-specified *level* $\alpha \in (0,1)$, any sample size n and any distribution in H_0 *the probability of Type I error (the test says H_1) is not greater than* α, while for every distribution in H_1 and every α *the Type II error is made only a finite number of times with probability 1*, as the sample size goes to infinity.

Recall that a stationary process can be represented as a mixture of stationary ergodic processes, that is, as a measure on the set $\mathcal{E}$ (see, e.g., [1]). The set $\mathcal{E}$ is not closed with respect to the distributional distance, but the set $\mathcal{S}$ of all stationary process distributions is. The following theorem utilizes these facts. Its proofs relies in addition on some other nice properties of the metric space $(\mathcal{S}, d)$; see [6] for the proof and [1] for the properties of $(\mathcal{S}, d)$.

Theorem 5. *There exists a consistent test for H_0 against H_1 if H_0 has probability 1 with respect to ergodic decomposition of every distribution from the closure of H_0, where the closure is with respect to the distributional distance d. Conversely, if there is a consistent test H_0 against H_1 then H_1 has probability 0 with respect to ergodic decomposition of every distribution from the closure of H_0.*

The necessary and sufficient conditions coincide if H_1 is the complement of H_0 to the set $\mathcal{E}$ of all stationary ergodic process distributions:

Corollary 2. *There exists a consistent test for H_0 against $H_1 := \mathcal{E} \backslash H_0$ if and only if H_1 has probability 0 with respect to ergodic decomposition of every distribution from the closure of H_0.*

4. COMPUTATIONAL COMPLEXITY

While the definition of empirical distributional distance $\hat{d}$ involves taking infinite sums, in can be calculated not only in finite time but efficiently. To see this, first observe that in $\hat{d}$ all summands corresponding to $m > n$ (where n is the min length of x_1, x_2) are 0. In the sum over l (cube size) all the summands are the same from the point where each cube has at most one point in it. This already makes computations finite. Moreover, even though the number of cubes in $B^{m,l}$ is exponential in m and l, at most $2n$ cubes are non-empty and these are easy to track (across different values of cube size l) with a tree structure. Thus, $\hat{d}$ can be calculated as is (in a naive way) in time $O(n^2 s \log n)$ where s is the minimal non-zero distance between points. This can be further reduced: the summands for $m > \log n$ and for l such that each cube less than $\log n$ points have no chance to have consistent estimates and only contribute (a negligible part) to the error. Thus, it is only practical to truncate the sums at $\log n$; since all the theoretical results presented here are asymptotic in n, it is easy to check that they still hold with this modification of $\hat{d}$. The computational complexity of $\hat{d}$ becomes $O(n \operatorname{polylog} n)$. For more information on implementation of the resulting algorithms see [9]. The latter work also provides some empirical evaluations of the clustering algorithm described here, as well as theoretical results for the online version of this problem.

5. OUTLOOK

Here we mention some interesting open problems for future research. First, the characterisation of those hypotheses for which consistent tests exist is so far incomplete: the necessary and sufficient conditions coincide only in the case when H_1 is the complement of H_0 (cf. Theorem 5 and the corollary). Furthermore, one can consider other notions of consistency of tests, both weaker and stronger ones, such as requiring both probabilities of error to converge to 0, or requiring both errors to be bounded uniformly. An interesting statistical problem that we did not consider here is independence testing. Given two samples it is required to test whether they were generated independently or not. Given the negative result of Theorem 4, one could think that this problem is also impossible to solve. However, Theorem 5 implies that it is, in fact, possible. Finding an actual test (possibly using $\hat{d}$) is an interesting open problem.

6. ACKNOWLEDGMENTS

This work was supported by Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council and FEDER through the "Contrat de Projets Etat Region (CPER) 2007-2013," the ANR projects Explora (ANR-08-COSI-004) and Lampada (ANR-09-EMER-007), by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement 270327 (project CompLACS) add Pascal-2.

7. REFERENCES

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

[2] D.S. Ornstein and B. Weiss, "How sampling reveals a process," *Annals of Probability*, vol. 18, no. 3, pp. 905–930, 1990.

[3] D. Ryabko, "Clustering processes," in *Proc. the 27th International Conference on Machine Learning (ICML 2010)*, Haifa, Israel, 2010, pp. 919–926.

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

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

[6] D. Ryabko, "Testing composite hypotheses about discrete ergodic processes," *Test*, vol. 21, no. 2, pp. 317–329, 2012.

[7] A. Khaleghi and D. Ryabko, "Multiple change-point estimation in stationary ergodic time-series," Tech. Rep. arXiv:1203.1515v4, arxiv, 2012.

[8] B. Ryabko, "Prediction of random sequences and universal coding," *Problems of Information Transmission*, vol. 24, pp. 87–96, 1988.

[9] A. Khaleghi, D. Ryabko, J. Mary, and P. Preux, "Online clustering of processes," in *AISTATS*, 2012, JMLR W&CP 22, pp. 601–609.