On Hypotheses Testing for Ergodic Processes

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

Boris Ryabko
Institute of Computational Technologies
of Siberian Branch of Russian Academy of Science,
Siberian State University of Telecommunications and Informatics,
Novosibirsk, Russia; boris@ryabko.net

***Abstract*—We address three problems of statistical analysis of time series: goodness-of-fit (or identity) testing, process discrimination, and the change point problem. For each of the problems we construct a test that is asymptotically accurate for the case when the data is generated by stationary ergodic processes. All problems are solved in a similar way by using empirical estimates of the distributional distance between the processes.**

I. Introduction

In this work we consider three problems of statistical analysis of time series: goodness-of-fit (or identity) testing, process discrimination, and the change point problem. For all three problems we obtain asymptotically accurate tests not making any other assumptions about the distributions generating the data rather than they are stationary ergodic. The problems are solved in a similar way, namely by evaluating empirical distributional distance: a weighted sum of frequencies of all k-tuples for all natural k.

The first problem we consider is testing whether a data sample was generated by a known stationary ergodic process ρ versus it was generated by a stationary ergodic process which is different from ρ. This is called goodness-of-fit or identity testing (while the hypothesis $\{\rho\}$ is called simple). For the case when ρ has finite memory [11] gives a solution to this problem: a test that can be based on an arbitrary universal code. It was noted in [12] that an asymptotically accurate test for the general case of stationary ergodic processes exists. A generalization of this problem is when both H_0 and H_1 are composite [10]. Here we propose a concrete and simple asymptotically accurate goodness-of-fit test, which demonstrates the proposed approach: to use empirical distributional distance for hypotheses testing. The Type I error of the test is fixed and the Type II error goes to 0 with probability 1.

The second problem is process discrimination: we are given three samples $X = (X_1, \dots, X_k)$, $Y = (Y_1, \dots, Y_m)$ and $Z = (Z_1, \dots, Z_n)$ generated by ergodic processes with distributions ρ_X, ρ_Y and ρ_Z. It is known that $\rho_X \neq \rho_Y$ but either $\rho_Z = \rho_X$ or $\rho_Z = \rho_Y$. It is required to test which one is the case. This problem for the case of dependent time series was considered for example in [8]. It is closely related to many important problems in machine learning and statistics, such as classification. Apparently no asymptotically accurate procedure for process discrimination has been known so far for the general case of stationary ergodic processes. Here we propose a simple test that converges almost surely to the correct answer. The test answers ρ_X or ρ_Y according to whether X or Y is closer to Z in empirical distributional distance.

Finally we consider the change point problem. It is another classical problem of mathematical statistics, with vast literature on both parametric (see e.g. [2] for an introduction) and non-parametric (see e.g. [5]) methods for solving it. There are also many variants and generalizations of this problem. In this work we consider the following setting: a sample $Z_1, \dots, Z_n$ is given, where $Z_1, \dots, Z_k$ are generated according to some distribution ρ_X and $Z_{k+1}, \dots, Z_n$ are generated according to some distribution ρ_Y which is different from ρ_X. It is known that the distributions ρ_X and ρ_Y are stationary ergodic, but nothing else is known about them. The case when the stationary distributions before and after the change are known (possibly up to a parametrization) was considered in many works, for example [1], [6], [9]. As for non-parametric methods, so far most literature on change point problem for dependent time series is concerned with detecting changes in the mean, or in the form of the single-dimensional distribution. Thus, in [5] the change point problem is considered for distributions satisfying certain strong mixing conditions and having different single-dimensional marginals. Apparently the most general case is considered in the recent work [3], namely, it is assumed that the time series before and after the change point are first-order stationary and have different single-dimensional marginals. Again, we solve the change point problem by considering empirical distributional distance which results in an asymptotically accurate change point estimate for the case of stationary ergodic distributions before and after the change.

II. Preliminaries

We are considering (stationary ergodic) processes over an alphabet A. For simplicity, we concentrate on finite alphabets, but the results can be extended to the case when A is the set of real numbers, real vectors, or, in general, a complete separable metric space. We use the symbol A^* for $\cup_{i=1}^{\infty} A^i$. Elements of A^* are called words or tuples. For a word B the symbol $|B|$ stands for the length of B. Denote B_i the ith element of A^*, enumerated in such a way that the elements of A^i appear before the elements of A^{i+1}, for all $i \in \mathbb{N}$.

Denote $\nu(X, B)$ the frequency of occurrence of the word B in the word $X \in A^*$: it is defined as

$$\frac{1}{|X| - |B| + 1} \sum_{i=1}^{|X|-|B|+1} I_{\{(X_i, \dots, X_{i+|B|-1}) = B\}} \text{ if } |X| \geq |B|,$$

978-1-4244-2271-5/08/$25.00 ©2008 IEEE

where $X = (X_1, \dots, X_{|X|})$, and as 0 otherwise. For example, $\nu(0001, 00) = 2/3$.

We use the symbol $\mathcal{S}$ for the set of all stationary ergodic processes on A^∞. It is known (see e.g. [4]) that for any process $\rho \in \mathcal{S}$ generating a sequence $X_1, X_2, \dots$ the frequency of occurrence of each word B tends to its limiting probability a.s.:

$$\nu((X_1, \dots, X_n), B) \to \rho(B) = \rho((X_1, \dots, X_{|B|}) = B)$$

as $n \to \infty$, where $\rho((X_1, \dots, X_{|B|}) = B)$ is the limiting probability for the process ρ.

The distributional distance is defined for a pair of processes ρ_1, ρ_2 as follows [7]:

$$d(\rho_1, \rho_2) = \sum_{i=1}^{\infty} w_i |\rho_1(B_i) - \rho_2(B_i)|,$$

where w_i are summable positive real weights (e.g. $w_k = 2^{-k}$). It is easy to see that d is a metric.

Define empirical distributional distance as

$$\hat{d}(X, Y) = \sum_{i=1}^{\infty} w_i |\nu(X, B_i) - \nu(Y, B_i)|.$$

Similarly, we can define the empirical distance when only one of the process measures is unknown:

$$\hat{d}(X, \rho) = \sum_{i=1}^{\infty} w_i |\nu(X, B_i) - \rho(B_i)|,$$

where ρ is a process and X is a sample.

Lemma 1: Let two samples $X = (X_1, \dots, X_k)$ and $Y = (Y_1, \dots, Y_m)$ be generated by stationary ergodic processes ρ_X and ρ_Y respectively. Then

(i) $\lim_{k,m\to\infty} \hat{d}(X, Y) = d(\rho_X, \rho_Y)$ a.s.

(ii) $\lim_{k\to\infty} \hat{d}(X, \rho_Y) = d(\rho_X, \rho_Y)$ a.s.

Proof: For any $\varepsilon > 0$ we can find such an index J that $\sum_{i=J}^{\infty} w_i < \varepsilon/2$. Moreover, for each j we have $\nu(X, B_j) \to \rho_X(B_j)$ a.s., so that

$$|\nu(X, B_j) - \rho(B_j)| < \varepsilon/(4Jw_j)$$

from some K_j on. Let $K := \max_{j<J} K_j$ (K depends on the realization $X_1, X_2, \dots$). Define analogously M for the sequence $(Y_1, \dots, Y_m, \dots)$. Thus for $k > K$ and $m > M$ we have

$$\begin{aligned}
&|\hat{d}(X, Y) - d(\rho_X, \rho_Y)| = \\
&\left| \sum_{i=1}^{\infty} w_i \big(|\nu(X, B_i) - \nu(Y, B_i)| - |\rho_X(B_i) - \rho_Y(B_i)| \big) \right| \\
&\le \sum_{i=1}^{\infty} w_i \big(|\nu(X, B_i) - \rho_X(B_i)| + |\nu(Y, B_i) - \rho_Y(B_i)| \big) \\
&\le \sum_{i=1}^{J} w_i \big(|\nu(X, B_i) - \rho_X(B_i)| + |\nu(Y, B_i) - \rho_Y(B_i)| \big) + \varepsilon/2 \\
&\le \sum_{i=1}^{J} w_i (\varepsilon/(4Jw_i) + \varepsilon/(4Jw_i)) + \varepsilon/2 = \varepsilon,
\end{aligned}$$

which proves the first statement. The second statement can be proven analogously. ■

III. Goodness-of-fit Test

For a given stationary ergodic process measure ρ and a sample $X = (X_1, \dots, X_n)$ we wish to test whether the sample was generated by ρ versus it was generated by a stationary ergodic source different from ρ. Thus H_0 is "the sample X was generated by ρ", and H_1 is "the sample X was generated by a stationary ergodic source $\rho' \ne \rho$".

Define the set B_δ^n as the set of all tuples of size n that are at least δ-far from ρ:

$$B_\delta^n := \{X \in A^n : \hat{d}(X, \rho) \ge \delta\}.$$

For a given confidence level α define the critical region C_α of the test as $C_\alpha^n := B_\gamma^n$ where

$$\gamma := \max\{\delta : \rho(B_\delta^n) \le \alpha\}.$$

As usual, the test rejects H_0 at confidence level α if $(X_1, \dots, X_n) \in C_\alpha^n$ and accepts it otherwise.

Theorem 1: The Type I error of the test is not greater than α, and the Type II error goes to zero as the sample size n tends to infinity: $\rho(C_\alpha^n) \le \alpha$ for any n, while $\lim_{n\to\infty} \rho'(C_\alpha^n) = 1$ for all $\rho' \in \mathcal{S} \backslash \{\rho\}$.

Note that using appropriate randomization in the definition of C_α^n we can make the Type I error exactly α.

Proof: The first statement holds by construction. Let the sample X be generated by $\rho' \in \mathcal{S}$, $\rho' \ne \rho$. Put $\delta = d(\rho, \rho')/2$. By Lemma 1 we have $\hat{d}(X, \rho) > \delta$ from some n on a.s., so $\rho'(B_\delta^n) \to 1$. At the same time $\rho(B_\delta^n) \to 0$, so that $\rho(B_\delta^n) < \alpha$ from some n on, whence $B_\delta^n \subset C_\alpha^n$, so that $\rho'(C_\alpha^n) \to 1$. ■

IV. Process discrimination

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 $\rho_X \ne \rho_Y$. 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 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 $\hat{d}(X, Z) \ge \hat{d}(Y, Z)$ then we conclude that the sample Z is generated by the same process as the sample X (formally $G(X, Y, Z) = 1$), otherwise we say that the sample Z is generated by the same process as the sample Y (or $G(X, Y, Z) = 2$).

Theorem 2: The suggested test makes a finite number of errors if k, m and n go to infinity. In other words: if $\rho_X = \rho_Z$ then $G(X, Y, Z) = 1$ from some k, m, n on with probability 1; otherwise $G(X, Y, Z) = 2$ from some k, m, n on with probability 1.

Proof: From the fact that d is a metric and from Lemma 1 we conclude that $\hat{d}(X, Z) \to 0$ (with probability 1) if and only

if $\rho_X = \rho_Z$. So, if $\rho_X = \rho_Z$ then by assumption $\rho_Y \neq \rho_Z$ and $\hat{d}(X,Z) \to 0$ a.s. while

$$\hat{d}(Y,Z) \to d(\rho_Y, \rho_Z) \neq 0.$$

Thus in this case $\hat{d}(Y,Z) > \hat{d}(X,Z)$ from some k, m, n on with probability 1 and $G(X,Y,Z) = 1$. The opposite case is analogous. ■

V. Change point problem

The sample $Z = (Z_1, \ldots, Z_n)$ consists of two concatenated parts $X = (X_1, \ldots, X_k)$ and $Y = (Y_1, \ldots, Y_m)$, 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 over a finite alphabet A. The distributions of the processes are unknown. The value k is called the *change point*; it is also unknown, but 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 a.s.).

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

For each t, $1 \leq t \leq n$, denote U^t the sample $(Z_1, \ldots, Z_t)$ consisting of the first t elements of the sample Z and V^t the remainder $(Z_{t+1}, \ldots, Z_n)$. Define the estimate $\hat{k}$ as follows:

$$\hat{k} = \mathrm{argmax}_{t \in [\sqrt{n}, n-\sqrt{n}]} \hat{d}(U^t, V^t).$$

In should be noted that the term $\sqrt{n}$ in this definition can be replaced by any $o(n)$ function that goes to infinity with n; this, in particular, does not affect the theorem below. Alternative approaches used in the literature on the change point problem are to introduce weights near the ends of the sample, or to specify linear bounds for the change point.

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

$$|\hat{k} - k| = o(n) \text{ a.s.}$$

when $k, m \to \infty$ in such a way that $\alpha < \frac{k}{n} < \beta$ for some $\alpha, \beta \in (0,1)$ from some n on.

Proof: To prove the statement, we will show that for every γ, $0 < \gamma < 1$ with probability 1 the inequality $\hat{d}(U^t, V^t) > \hat{d}(X,Y)$ holds for each t such that $\sqrt{n} \leq t < \gamma k$ possibly except for a finite number of times. Thus we will show that linear γ-underestimates occur only a finite number of times, and for overestimate it's analogous. Fix some γ, $0 < \gamma < 1$ and $\varepsilon > 0$. Let J be big enough to have $\sum_{i=J}^{\infty} w_i < \varepsilon/2$ and also big enough to have an index $j < J$ for which $\rho_X(B_j) \neq \rho_Y(B_j)$. Since empirical frequencies converge to the limiting probabilities a.s., we can find (for almost all sequences) such $M_\varepsilon \in \mathbb{N}$ that $|\nu(Y, B_i) - \rho_Y(B_i)| \leq \varepsilon/2J$ for all $m > M_\varepsilon$ and for each i, $1 \leq i \leq J$, and also to have $|B_j|/m < \varepsilon/J$. Moreover, since the distribution of the sample $X_s, X_{s+1}, \ldots, X_k$, where s is chosen independently of the sample, is governed by the same stationary ergodic process as $X_1, \ldots, X_k$, we can find such K_ε that for all $k > K_\varepsilon$ for all i, $1 \leq i \leq J$ we will have $|\nu(U^t, B_i) - \rho_X(B_i)| \leq \varepsilon/2J$ for each $t \geq \sqrt{n}$ and $|\nu((X_s, X_{s+1}, \ldots, X_k), B_i) - \rho_X(B_i)| \leq \varepsilon/2J$ for each $s \leq \gamma k$. So for each $t \in [\sqrt{n}, \gamma k]$ for the difference between the frequency of B_j in V_s and its probability we will have

$$\begin{aligned}
&\left|\nu(V^s, B_j) - \frac{(1-\gamma)k\rho_X(B_j) + m\rho_Y(B_j)}{(1-\gamma)k + m}\right| \\
&\quad \leq \left|\frac{(1-\gamma)k\nu((X_s, \ldots, X_k), B_j) + m\nu(Y, B_j)}{(1-\gamma)k + m} - \right. \\
&\quad \left. \frac{(1-\gamma)k\rho_X(B_j) + m\rho_Y(B_j)}{(1-\gamma)k + m}\right| + \frac{|B_j|}{m + \gamma k} \\
&\quad \leq 3\varepsilon/J,
\end{aligned}$$

for $k > K_\varepsilon$ and $m > M_\varepsilon$ (from the definitions of K_ε and M_ε). Hence

$$\begin{aligned}
&|\nu(X, B_j) - \nu(Y, B_j)| - |\nu(U^s, B_j) - \nu(V^s, B_j)| \\
&\geq |\nu(X, B_j) - \nu(Y, B_j)| \\
&\quad - \left|\nu(U^s, B_j) - \frac{(1-\gamma)k\rho_X(B_j) + m\rho_Y(B_j)}{(1-\gamma)k + m}\right| - 3\varepsilon/J \\
&\geq |\rho_X(B_j) - \rho_Y(B_j)| \\
&\quad - \left|\rho_X(B_j) - \frac{(1-\gamma)k\rho_X(B_j) + m\rho_Y(B_j)}{(1-\gamma)k + m}\right| - 4\varepsilon/J \\
&\quad = \delta_j - 4\varepsilon/J,
\end{aligned}$$

for some δ_j that depends only on k/m and γ. Summing over different B_i we will have

$$\hat{d}(X,Y) - \hat{d}(U^s, V^s) \geq w_j\delta_j - 5\varepsilon,$$

for all n such that $k > K_\varepsilon$ and $m > M_\varepsilon$, which is positive for small enough ε. ■

References

[1] R. Bansal, P. Papantoni-Kazakos. An Algorithm for Detecting a Change in a Stochastic Process. IEEE Trans. on Information Theory, vol. IT-32 No. 2 (1986), pp. 227–235.

[2] M. Basseville, I. Nikiforov. Detection of Abrupt Changes: Theory and Applications. Prentice Hall, 1993.

[3] S. Ben Hariz, J. Wylie and Q. Zhang. Optimal Rate of Convergence for Nonparametric Change-Point Estimators for Nonstationary Sequences. The Annals of Statistics, Vol. 35, No. 4 (2007), pp. 1802-1826.

[4] P. Billingsley, Ergodic theory and information. Wiley, New York, 1965.

[5] B. Brodsky, B. Darkhovsky. Nonparametric Methods in Change-Point Problems. Kluwer Academic Pablishers, 1993.

[6] Cheng-Der Fuh, Asymptotic Operating Characteristics of an Optimal Change Point Detection in Hidden Markov Models. The Annals of Statistics, Vol. 32 No. 5 (2004), pp. 2305-2339.

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

[8] M. Gutman. Asymptotically Optimal Classification for Multiple Tests with Empirically Observed Statistics. IEEE Trans. Information Theory, vol. 35 no. 2 (1989), pp. 402–408.

[9] M. Lavielle, Detection of multiple changes in a sequence of dependent variables, Stochastic Processes and their Applications, 83 (1999), pp. 79–102.

[10] A. Nobel. Hypothesis testing for families of ergodic processes, Bernoulli, vol. 12 no. 2 (2006), pp. 251–269.

[11] B. Ryabko, J. Astola. Universal Codes as a Basis for Time Series Testing, Statistical Methodology vol. 3 (2006), pp. 375–397.

[12] P. Shields, The Interactions Between Ergodic Theory and Information Theory. IEEE Trans. on Information Theory, vol. 44, no. 6 (1998), pp. 2079–2093.