Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 136 (2006) 4119–4128
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Universal codes as a basis for nonparametric testing of serial independence for time series☆

Boris Ryabko[a,*], Jaakko Astola[b]

[a]*Siberian State University of Telecommunications and Computer Science, Russia*
[b]*Tampere University of Technology, Finland*

Received 26 August 2004; received in revised form 2 April 2005; accepted 24 July 2005
Available online 1 September 2005

Abstract

We address the problem of nonparametric testing of serial independence for time series and its generalization. More precisely, we consider a stationary and ergodic source p, which generates symbols $x_1 \dots x_t$ from some finite set A and a null hypothesis H_0 that p is a Markov source of order at most m, $(m \geqslant 0)$. The alternative hypothesis H_1 is that the sequence is generated by a stationary and ergodic source, which differs from the source under H_0. In particular, if $m = 0$ we have the null hypothesis H_0 that the sequence is generated by a Bernoulli source (i.e. the hypothesis that $x_1 \dots x_t$ are independent). In this paper some new tests that are based on so-called universal codes and universal predictors, are suggested.

MSC: 60G10; 60J10; 62M02; 62M07; 94A29

Keywords: Independence; Serial independence; Universal coding; Hypothesis testing; Information theory; Markov process; Random process; Prediction

☆ Research was supported by the joint project grant "Efficient randomness testing of random and pseudorandom number generators" of Royal Society, UK (Grant ref: 15995) and Russian Foundation for Basic Research (Grant no. 03-01-00495.).

* Corresponding author.
E-mail address: boris@ryabko.net (B. Ryabko).

0378-3758/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2005.07.008

1. Introduction

Nonparametric testing of independence in time series is very important in statistical applications. There is an extensive literature dealing with nonparametric independence testing. We mention only the well-known methods that are based on the chi-square tests (see for review Kendall and Stuart, 1961) and the classical papers of Hoeffding (1948) and Blum et al. (1961); quite a full review can also be found in Ghoudi et al. (2001).

In this paper, we consider a source (or process), which generates elements from a finite set A and the following two hypotheses: H_0 that the source is Markovian one of order not larger than m, $(m \geqslant 0)$, and the alternative hypothesis H_1 that the sequence is generated by a stationary and ergodic source, which differs from the source under H_0. The test should be based on a sample $x_1 \dots x_t$ generated by the source.

For example, the sequence $x_1 \dots x_t$ might be a DNA-string and one can consider the question about the depth of the statistical dependence.

We suggest a family of tests that are based on so-called universal predictors (or universal data compression methods). The Type I errors of the tests are not larger than a given $\alpha(\alpha \in (0, 1))$ for any source under H_0, whereas the Type II error for any source under H_1 tends to 0, when the sample size t grows.

The tests are based on results and ideas of Information Theory and, especially, on those of universal coding. Informally, the idea of the tests can be described as follows. Suppose that the source generates letters from an alphabet A and one wants to test H_0 (the source is Markovian of order $m, m \geqslant 0$). First we recall that there exist so-called universal codes which, loosely speaking can "compress" any sequence of length t generated by a stationary and ergodic source, to the length th_∞ bits, where h_∞ is the limiting Shannon entropy as t tends to infinity. Secondly, it is well known in Information Theory that h_∞ equals mth-order (conditional) Shannon entropy h_m, if H_0 is true, and h_∞ is strictly less than h_m if H_1 is true. So, the following test appears natural: compress the sample sequence $x_1 \dots x_t$ by a universal code and compare the length of the obtained file with th_m^*, where h_m^* is an estimate of h_m. If the length of the compressed file is significantly less than th_m^*, then the hypothesis H_0 should be rejected.

It is no surprise that the results and ideas of universal coding can be applied to some classical problems of mathematical statistics. In fact, methods of universal coding (and the closely connected universal prediction) extract information from observed data in order to compress (or predict) data efficiently in the case where the source statistics is unknown. Recently such a connection between universal coding and mathematical statistics was used by Csiszár and Shields (2000) for estimating the order of Markov sources and by Ryabko and Monarev (2005) for constructing efficient tests for randomness, i.e. for testing the hypothesis $\hat{\mathrm{H}}_0$ that a sequence is generated by a Bernoulli source and all letters have equal probabilities against $\hat{\mathrm{H}}_1$ that the sequence is generated by a stationary and ergodic source, which differs from the source under $\hat{\mathrm{H}}_0$.

The outline of the paper is as follows. The next part contains definitions and necessary information from the theory of universal coding and universal prediction. Part three is devoted to testing the above described hypotheses. All proofs are given in the appendix.