Simulation Modelling Practice and Theory 61 (2016) 1–13

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

journal homepage: www.elsevier.com/locate/simpat

Markov chain order estimation with parametric significance tests of conditional mutual information

Maria Papapetrou, Dimitris Kugiumtzis *

Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 23 January 2015
Received in revised form 3 November 2015
Accepted 6 November 2015
Available online 30 November 2015

Keywords:
Symbol sequence
Markov chain order
Conditional mutual information
Significance test
DNA

ABSTRACT

Besides the different approaches suggested in the literature, accurate estimation of the order of a Markov chain from a given symbol sequence is an open issue, especially when the order is moderately large. Here, parametric significance tests of conditional mutual information (CMI) of increasing order m, $I_c(m)$, on a symbol sequence are conducted for increasing orders m in order to estimate the true order L of the underlying Markov chain. CMI of order m is the mutual information of two variables in the Markov chain being m time steps apart, conditioning on the intermediate variables of the chain. The null distribution of CMI is approximated with a normal and gamma distribution deriving analytic expressions of their parameters, and a gamma distribution deriving its parameters from the mean and variance of the normal distribution. The accuracy of order estimation is assessed with the three parametric tests, and the parametric tests are compared to the randomization significance test and other known order estimation criteria using Monte Carlo simulations of Markov chains with different order L, length of symbol sequence N and number of symbols K. The parametric test using the gamma distribution (with directly defined parameters) is consistently better than the other two parametric tests and matches well the performance of the randomization test. The tests are applied to genes and intergenic regions of DNA sequences, and the estimated orders are interpreted in view of the results from the simulation study. The application shows the usefulness of the parametric gamma test for long symbol sequences where the randomization test becomes prohibitively slow to compute.

1. Introduction

Symbol sequences are directly observed on real-world processes, such as DNA sequences and on-line transaction logs, but can also be derived from discretization of time series. Sequence analysis, initially developed mostly for biological applications [1], has expanded with regard to both applications and methodologies, and sequence mining techniques are constantly being developed [2]. Here however, we concentrate on a classical and fundamental problem that regards the memory of the underlying mechanism to a symbol sequence. In the presence of association in symbol sequences, the first step of the analysis is to assume a Markov chain and estimate the order of the Markov chain.

There are many Markov chain order estimators proposed and assessed in the literature. The Bayesian information criterion (BIC) and the Akaike information criterion (AIC) are the two oldest and best known order estimators based on maximum

* Corresponding author.
E-mail addresses: mariapap@auth.gr (M. Papapetrou), dkugiu@auth.gr (D. Kugiumtzis).

http://dx.doi.org/10.1016/j.simpat.2015.11.002
1569-190X/

likelihood [3–5]. Another estimator is given by the maximal fluctuation method proposed by Peres–Shields [6] and modified by Dalevi and Dubhashi [7], who found that the Peres–Shields (PS) estimator is simpler, faster and more robust to noise than other criteria like AIC and BIC [7]. Other order estimation schemes include the method of Menéndez et al. [8], which uses the ϕ-divergence measures [9], the method of global dependency level (GDL), also called relative entropy [10], and the efficient determination criterion (EDC) [11]. Based on the information-related measures, and specifically the conditional mutual information (CMI), we recently proposed the order estimation by means of randomization significance tests for CMI at increasing orders [12]. In a somewhat similar way, Pethel et al. [13] propose a randomization test for the examined Markov chain order using the Chi-squared statistic.

In the approach of [12] we made no assumption on the distribution of CMI. Here we propose the order estimation with parametric tests, approximating the null distribution of CMI by normal and gamma distributions. We follow the bias correction and the approximation for the variance in [14,15] and approximate the distribution of mutual information with Gaussian distribution as an obvious possible choice [16,17]. We also consider the result in Goebel et al. [18] that the statistic of mutual information (MI), and subsequently CMI, follows gamma distribution. Finally, we consider a second gamma approximation with shape and scale parameter derived from the mean and variance approximations of the normal distribution. We implement the three parametric significance tests for CMI and compare them to the randomization test of [12], as well as other known Markov chain order estimation methods. Further, we attempt to assess the Markov chain order of DNA sequences and infer for short and long range correlation on the basis of the parametric and randomization CMI testing. A systematic investigation of long range correlation of DNA sequences using the CMI approach is reported in [19].

The structure of the paper is as follows. First, in Section 2, CMI is defined and estimated on symbol sequences. Parametric significance tests for CMI of increasing orders are presented, approximating the null distribution of CMI by the normal and gamma distributions. In Section 3, we assess the efficiency of the parametric tests in estimating the Markov chain orders and compare them to other known methods. In Section 4, we apply the parametric and randomization tests to DNA sequences, and in Section 5, the results are discussed and the main conclusions are drawn.

2. Conditional mutual information and Markov chain order estimation

We start with the definition of entropy, mutual information (MI) and conditional mutual information (CMI) for Markov chains. Let $\{x_t\}$ denote a symbol sequence generated by a Markov chain $\{X_t\}, t \geqslant 1$, of an unknown order $L \geqslant 1$ in a discrete space of K possible states $A = \{a_1, \ldots, a_K\}$, $p(x_t)$ the probability of $x_t \in A$ occurring in the chain, $\mathbf{X}_t = [X_t, X_{t-1}, \ldots, X_{t-m+1}]$ a vector (word) of m successive variables of the Markov chain and $p(\mathbf{x}_t)$ the probability of a word $\mathbf{x}_t = \{x_t, x_{t-1}, \ldots, x_{t-m+1}\} \in A^m$ occurring in the chain. The entropy of a random variable of the Markov chain X_t is $H(X_t) = -\sum_{x_t} p(x_t) \ln p(x_t)$ and the entropy of a word $\mathbf{X}_t$ is $H(\mathbf{X}_t) = -\sum_{x_t,\ldots,x_{t-m+1}} p(\mathbf{x}_t) \ln p(\mathbf{x}_t)$. The MI of two random variables in the Markov chain being m time steps apart is [20]

$$\begin{aligned} I(m) &= I(X_t; X_{t-m}) = H(X_t) + H(X_{t-m}) - H(X_t, X_{t-m}) \\ &= \sum_{x_t, x_{t-m}} p(x_t, x_{t-m}) \ln \frac{p(x_t, x_{t-m})}{p(x_t) p(x_{t-m})} \end{aligned}$$

and quantifies the amount of information for the one variable given the other variable.

The fundamental property of a Markov chain of order L is

$$p(X_t | X_{t-1}, X_{t-2}, \ldots, X_{t-L}, X_{t-L-1}, \ldots) = p(X_t | X_{t-1}, X_{t-2}, \ldots, X_{t-L}), \tag{1}$$

meaning that the distribution of the variable X_t of the Markov chain at time t is determined only in terms of the preceding L variables of the chain. It is noted that $I(m)$ for $m > L$ may not drop to zero due to the existence of MI between the intermediate variables. Thus for estimating L we consider CMI that accounts for the intermediate variables. CMI of order m is defined as the mutual information of X_t and X_{t-m} conditioning on $X_{t-m+1}, \ldots, X_{t-1}$ [20]

$$\begin{aligned} I_c(m) &= I(X_t; X_{t-m} | X_{t-1}, \ldots, X_{t-m+1}) \\ &= -H(X_t, \ldots, X_{t-m}) + H(X_{t-1}, \ldots, X_{t-m}) + H(X_t, \ldots, X_{t-m+1}) - H(X_{t-1}, \ldots, X_{t-m+1}) \\ &= \sum_{x_t, \ldots, x_{t-m}} p(x_t, \ldots, x_{t-m}) \ln \frac{p(x_t | x_{t-1}, \ldots, x_{t-m})}{p(x_t | x_{t-1}, \ldots, x_{t-m+1})}. \end{aligned} \tag{2}$$

CMI coincides with MI for successive random variables in the chain, $I_c(1) = I(1)$.

From the Markov chain property in (1), for $m > L$ the logarithmic term in the sum of (2) is zero and thus $I_c(m) = 0$. On the other hand, for $m \leqslant L$, we expect in general the two variables m time steps apart be dependent given the $m-1$ intermediate variables, and $I_c(m) > 0$. It is possible that $I_c(m) = 0$ for $m < L$, but not for $m = L$, as then the Markov chain order would not be L. So, increasing the order m, we expect in general when $I_c(m) > 0$ and $I_c(m+1) = 0$ to have $m = L$. To account for complicated and rather unusual cases where $I_c(m+1) = 0$ occurs for $m+1 < L$, we can extend the condition $I_c(m) > 0$ and $I_c(m+1) = 0$ to require also $I_c(m+2) = 0$, and even further up to some maximum order.