Physica A 392 (2013) 1593–1601

Contents lists available at SciVerse ScienceDirect

Physica A

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

Markov chain order estimation with conditional mutual information

M. Papapetrou, D. Kugiumtzis *

Department of Mathematical, Physical and Computational Science, Faculty of Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 20 September 2012
Received in revised form 14 December 2012
Available online 26 December 2012

Keywords:
Order estimation
Markov chains
Conditional mutual information (CMI)
Randomization test
DNA

ABSTRACT

We introduce the Conditional Mutual Information (CMI) for the estimation of the Markov chain order. For a Markov chain of K symbols, we define CMI of order m, $I_c(m)$, as the mutual information of two variables in the chain being m time steps apart, conditioning on the intermediate variables of the chain. We find approximate analytic significance limits based on the estimation bias of CMI and develop a randomization significance test of $I_c(m)$, where the randomized symbol sequences are formed by random permutation of the components of the original symbol sequence. The significance test is applied for increasing m and the Markov chain order is estimated by the last order for which the null hypothesis is rejected. We present the appropriateness of CMI-testing on Monte Carlo simulations and compare it to the Akaike and Bayesian information criteria, the maximal fluctuation method (Peres–Shields estimator) and a likelihood ratio test for increasing orders using ϕ-divergence. The order criterion of CMI-testing turns out to be superior for orders larger than one, but its effectiveness for large orders depends on data availability. In view of the results from the simulations, we interpret the estimated orders by the CMI-testing and the other criteria on genes and intergenic regions of DNA chains.

1. Introduction

Let $\{x_t\}_{t=1}^N$ denote a symbol sequence generated by a Markov chain $\{X_t\}$, of an unknown order $L \geq 1$ in a discrete space of K possible states $A = \{a_1, \ldots, a_K\}$. The objective is to estimate L from the symbol sequence $\{x_t\}_{t=1}^N$ for a limited length N.

Many criteria for Markov chain order estimation have been proposed and evaluated in terms of their asymptotic properties. The Bayesian information criterion (BIC) was proposed to render consistency of the popular Akaike information criterion (AIC) [1–3]. However, BIC was found to perform worse than AIC for small sequence lengths, questioning the value of asymptotic properties in practical problems [2,4–6]. A more recent and general criterion than AIC and BIC is the efficient determination criterion (EDC), opting for a penalty function from a wide range of possible such functions [7]. Peres–Shields proposed in Ref. [8] the maximal fluctuation method, which compares transition probabilities for words of increasing lengths, and Dalevi and Dubhashi [9] modified it for practical settings and, instead of having to set a different threshold for each problem, they estimate the order from a sharp change in the transition probabilities. They found that the Peres–Shields (PS) estimator is simpler, faster and more robust to noise than other criteria like AIC and BIC [9]. Another method is that of global dependency level (GDL), also called relative entropy, using the f-divergence to measure the discrepancy between two probability distributions [10]. GDL was found consistent and more efficient than AIC and BIC on relatively small sequences. Finally, the method of Menendez et al. [11–13] makes likelihood ratio tests for increasing orders using the ϕ-divergence measures [14]. This procedure was found more powerful in tested cases than the existing chi-square and likelihood ratio procedures, and it has also been applied to DNA [13].

* Corresponding author. Tel.: +30 2310 995955; fax: +30 2310 995958.
E-mail addresses: mariapap@gen.auth.gr (M. Papapetrou), dkugiu@gen.auth.gr (D. Kugiumtzis).

0378-4371/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2012.12.017

Here, we follow a different approach and estimate the Markov chain order from sequential hypothesis testing for the significance of the conditional mutual information (CMI) for increasing orders m, denoted as $I_c(m)$. $I_c(m)$ is the mutual information of x_i and x_{i+m} conditioning on the intermediate variables of the chain, $x_{i+1}, \ldots, x_{i+m-1}$. A significant $I_c(m)$ indicates that the order of the Markov chain is at least m. Thus the repetition of the significance test of $I_c(m)$ for increasing m allows for the estimation of the Markov chain order L from the last order m for which the null hypothesis of zero CMI is rejected. We show that the significance bounds for $I_c(m)$ formed by means of appropriate resampling are more accurate than the approximate analytic bounds we derived based on previous analytic results on the bias of entropy [15]. We further compare the CMI testing with other criteria for order selection on simulated Markov chains and DNA sequences.

The structure of the paper is as follows. In Section 2, CMI is defined and estimated on symbol sequences, an analytic significance limit of CMI is derived, and a randomization significance test is proposed, forming our method of CMI-testing for the estimation of the Markov chain order. Other methods for estimating the Markov chain order are briefly presented. In Section 3, we assess the efficiency of the proposed CMI-testing and compare it to other order selection criteria on simulations of Markov chains produced by randomly chosen transition probability matrices of different order, as well as transition probability matrices estimated on genes and intergenic regions of DNA sequence. In Section 4, we apply the CMI testing to the two DNA sequences and investigate the limitations of order estimation in terms of data size. Finally, concluding remarks are discussed in Section 5.

2. Conditional mutual information and Markov chain order estimation

First we define CMI in terms of mutual information and subsequently entropies. The Shannon entropy expresses the information (or uncertainty) of a random variable Xt

$$H(X) = -\sum_x p(x) \ln p(x),$$

where the sum is defined for all possible symbols (discrete values) $x \in A$, and $p(x)$ is the probability of x occurring in the chain. The definition of Shannon entropy is extended to a vector variable $\mathbf{X}_t = [X_t, X_{t-1}, \ldots, X_{t-m+1}]$ from a stationary Markov chain $\{X_t\}$, referred to as word of length m, and reads

$$H(\mathbf{X}_t) = -\sum_{x_t,\ldots,x_{t-m+1}} p(\mathbf{x}_t) \ln p(\mathbf{x}_t),$$

where $\mathbf{x}_t = \{x_t, x_{t-1}, \ldots, x_{t-m+1}\} \in A^m$, $p(\mathbf{x}_t)$ is the probability of a word $\mathbf{x}_t$ occurring in the chain, and the sum is over all possible words of K symbols and length m.

The mutual information (MI) of two random variables in the Markov chain being m time steps apart, denoted $I(m) = I(X_t; X_{t-m})$, is defined in terms of entropy as [16]

$$I(m) = H(X_t) + H(X_{t-m}) - H(X_t, X_{t-m}) = \sum_{x_t, x_{t-m}} p(x_t, x_{t-m}) \ln \frac{p(x_t, x_{t-m})}{p(x_t)p(x_{t-m})}. \tag{1}$$

While $I(1)$ quantifies the amount of information X_{t-1} carries about X_t and vice versa, $I(2)$ cannot be interpreted accordingly due to the presence of X_{t-1}, and the information of X_{t-2} about X_t, or part of it, may already be shared with X_{t-1}. Thus if we are after the genuine information of X_{t-2} about X_t, we need to account for the information of X_{t-1} about X_t. This is indeed desired when we want to estimate the memory of the process, i.e. the order of the Markov chain. The appropriate measure for this is the conditional mutual information (CMI). CMI of order m is defined as the mutual information of X_t and X_{t-m} conditioning on $X_{t-m+1}, \ldots, X_{t-1}$ [16]

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

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

2.1. Estimation of conditional mutual information

The estimation of CMI is given through the estimation of the joint probability and the conditional probabilities in (2) by the corresponding relative frequencies. Specifically, the maximum likelihood estimate (MLE) of $p(x_t, x_{t-1}, \ldots, x_{t-m+1})$ is

$$\hat{p}(x_t, x_{t-1}, \ldots, x_{t-m+1}) = \frac{n_{i_1,\ldots,i_m}}{K^m},$$

where $n_{i_1,\ldots,i_m}$ is the frequency of occurrence of a word $\{i_1, \ldots, i_m\} \in A^m$ in the symbol sequence $\{x_t\}_{t=1}^N$, defined as $n_{i_1,\ldots,i_m} = \sum_{t=m}^N I(x_t = i_1, \ldots, x_{t-m+1} = i_m)$, where I denotes the indicator function. Respectively, the MLE of the conditional probability $p(x_t|x_{t-1}, \ldots, x_{t-m})$ is

$$\hat{p}(x_t|x_{t-1}, \ldots, x_{t-m}) = \frac{n_{i_1,\ldots,i_m,i_{m+1}}}{n_{i_1,\ldots,i_m}}.$$