Gusztáv MORVAI and Benjamin WEISS:

Order Estimation of Markov Chains

IEEE Trans. Inform. Theory 51 (2005), no. 4, 1496–1497.

Abstract

We describe estimators $\chi_n(X_0, X_1, \ldots, X_n)$, which when applied to an unknown stationary process taking values from a countable alphabet $\mathcal{X}$, converge almost surely to k in case the process is a k-th order Markov chain and to infinity otherwise.

Keywords: Stationary processes, Markov chains, order estimation
*Mathematics Subject Classifications (2000)*62M05, 60G25, 60G10

1 Introduction

When faced with an unknown stationary and ergodic stochastic process $X_1, X_2, \ldots, X_n, \ldots$ one may try to determine various properties of this process from the successive observations up to time n. For example, one might try to estimate the entropy of the process. Several schemes of the form $g_n(X_1, \ldots, X_n)$ are known which will converge almost surely to the entropy of the process $\{X_n\}$ cf. Bailey [1], Csiszár and Shields [2], Csiszár [3], Ornstein and Weiss [8], [7], [9], Kontoyiannis, Algoet, Suhov and Wyner [6] and Ziv [10]. However, if one just wants to determine whether or not the process has positive entropy (often associated with the popular notion of chaos) then there is no sequence of two valued functions $e_n(X_1, \ldots, X_n) \in \{ZERO, POSITIVE\}$ with the property that almost surely, e_n stabilize at $ZERO$ for all zero entropy processes and at $POSITIVE$ for all positive entropy processes. (While this result does not appear explicitly in Ornstein amd Weiss [7], it can be readily established using a very simple variant of the construction given there in § 4.)

A similar situation obtains in testing for membership in the class of k-th order Markov chains. One can estimate the order of a Markov chain by e.g the method of Csiszár and Shields [2] or Csiszár [3]. They show that the minimum description length Markov estimator will converge almost surely to the correct order if the alphabet size is bounded a priori. Without this assumption they show that this is no longer true. To accomplish their goals they study the large scale typicality of Markov sample paths. A further negative result is that of Bailey [1] who showed that no two valued test exists for testing mixing Markov vs. not mixing Markov.

We will present a more direct estimator for the order of a Markov chain which also uses the fact that there are universal rates for the convergence of empirical k-block distributions in this class. Our approach enables us to dispense with the assumption that the alphabet size is bounded, indeed it may even be infinite, as long as there is a finite memory. In addition we will show that if the process is not a Markov chain then the estimate for the order will tend to infinity. This is in complete analogy with the entropy estimation that we mentioned earlier.

2 The Order Estimator

Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic time series taking values from a discrete (finite or countably infinite) alphabet $\mathcal{X}$. (Note that all stationary time series $\{X_n\}_{n=0}^{\infty}$ can be thought to be a two sided time series, that is, $\{X_n\}_{n=-\infty}^{\infty}$.) For notational convenience, let $X_m^n = (X_m, \ldots, X_n)$, where $m \leq n$. Note that if $m > n$ then X_m^n is the empty string.
Let $p(x_{-k}^0)$ and $p(y|x_{-k}^0)$ denote the distribution $P(X_{-k}^0 = x_{-k}^0)$ and the conditional distribution $P(X_1 = y|X_{-k}^0 = x_{-k}^0)$, respectively.
A discrete alphabet stationary time series is said to be a Markov chain if for some $K \geq 0$, for all $y \in \mathcal{X}$, $i \geq 1$ and $z_{-K-i+1}^0 \in \mathcal{X}^{K+i}$, if $p(z_{-K-i+1}^0) > 0$ then

$$p(y|z_{-K+1}^0) = p(y|z_{-K-i+1}^0).$$

The order of a Markov chain is the smallest such K.

In order to estimate the order we need to define some explicit statistics.
For $k \geq 0$ let $\mathcal{S}_k$ denote the support of the distribution of X_{-k}^0 as

$$\mathcal{S}_k = \{x_{-k}^0 \in \mathcal{X}^{k+1} : p(x_{-k}^0) > 0\}.$$

Define

$$\Delta_k = \sup_{1 \leq i} \sup_{(z_{-k-i+1}^0, x) \in \mathcal{S}_{k+i}} \left| p(x|z_{-k+1}^0) - p(x|z_{-k-i+1}^0) \right|.$$

We will divide the data segment X_0^n into two parts: $X_0^{\lceil \frac{n}{2} \rceil - 1}$ and $X_{\lceil \frac{n}{2} \rceil}^n$. Let $\mathcal{S}_{n,k}^{(1)}$ denote the set of strings with length $k+1$ which appear at all in $X_0^{\lceil \frac{n}{2} \rceil - 1}$. That is,

$$\mathcal{S}_{n,k}^{(1)} = \{x_{-k}^0 \in \mathcal{X}^{k+1} : \exists k \leq t \leq \lceil \frac{n}{2} \rceil - 1 : X_{t-k}^t = x_{-k}^0\}.$$

For a fixed $0 < \gamma < 1$ let $\mathcal{S}_{n,k}^{(2)}$ denote the set of strings with length $k+1$ which appear more than $n^{1-\gamma}$ times in $X_{\lceil \frac{n}{2} \rceil}^n$. That is,

$$\mathcal{S}_{n,k}^{(2)} = \{x_{-k}^0 \in \mathcal{X}^{k+1} : \#\{\lceil \frac{n}{2} \rceil + k \leq t \leq n : X_{t-k}^t = x_{-k}^0\} > n^{1-\gamma}\}.$$

Let

$$\mathcal{S}_k^n = \mathcal{S}_{n,k}^{(1)} \bigcap \mathcal{S}_{n,k}^{(2)}.$$

For notational convenience, let $C(x|z_{-k+1}^0 : [n_1, n_2])$ denote the empirical conditional probability of $X_1 = x$ given $X_{-k+1}^0 = z_{-k+1}^0$ from the samples $(X_{n_1}, \ldots, X_{n_2})$, that is,

$$C(x|z_{-k+1}^0 : [n_1, n_2]) = \frac{\#\{n_1 + k \le t \le n_2 : X_{t-k}^t = (z_{-k+1}^0, x)\}}{\#\{n_1 + k - 1 \le t \le n_2 - 1 : X_{t-k+1}^t = z_{-k+1}^0\}}$$

where $0/0$ is defined as 0.
We define the empirical version of Δ_k as follows:

$$\hat{\Delta}_k^n = \max_{1 \le i \le n} \max_{(z_{-k-i+1}^0, x) \in \mathcal{S}_{k+i}^n} \left| C(x|z_{-k+1}^0 : [\lceil \frac{n}{2} \rceil, n]) - C(x|z_{-k-i+1}^0 : [\lceil \frac{n}{2} \rceil, n]) \right|.$$

Observe, that by ergodicity, for any fixed k,

$$\liminf_{n \to \infty} \hat{\Delta}_k^n \ge \Delta_k \quad \text{almost surely.} \tag{1}$$

We define an estimate χ_n for the order from samples X_0^n as follows. Let $0 < \beta < \frac{1-\gamma}{2}$ be arbitrary. Set $\chi_0 = 0$, and for $n \ge 1$ let χ_n be the smallest $0 \le k_n < n$ such that $\hat{\Delta}_{k_n}^n \le n^{-\beta}$.

THEOREM. *If the stationary and ergodic time series $\{X_n\}$ taking values from a discrete alphabet happens to be a Markov chain with any finite order then χ_n equals to the order eventually almost surely, and if it is not Markov with any finite order then $\chi_n \to \infty$ almost surely.*

Application: Let $M > 0$ be arbitrary. The goal is to decide if the discrete alphabet stationary and ergodic time series is a Markov chain with order less than M or not. One may use χ_n and say YES if $\chi_n < M$ and say NO otherwise. By the Theorem, eventually, the answer will be correct.

3 Proof of the Theorem

Proof: If the process is a Markov chain, it is immediate that for all k greater than or equal the order, $\Delta_k = 0$. For k less than the order $\Delta_k > 0$. If the process is not a Markov chain with any finite order then $\Delta_k > 0$ for all k. Thus by (1) if the process is not Markov then $\chi_n \to \infty$ and if it is Markov

then χ_n is greater or equal the order eventually almost surely. We have to show that χ_n is less or equal the order eventually almost surely provided that the process is a Markov chain.
Assume that the process is a Markov chain with order k. Let $n \geq k$. We will estimate the probability of the undesirable event as follows:

$$P(\hat{\Delta}_k^n > n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil}) \leq$$
$$\sum_{i=1}^{n} P(\max_{(z^0_{-k-i+1},x)\in\mathcal{S}^n_{k+i}} \left| C(x|z^0_{-k+1} : [\lceil \frac{n}{2} \rceil, n]) - C(x|z^0_{-k-i+1} : [\lceil \frac{n}{2} \rceil, n]) \right| > n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil}).$$

We can estimate each probability in the sum as the sum of two terms:

$$\begin{aligned} & P(\max_{(z^0_{-k-i+1},x)\in\mathcal{S}^n_{k+i}} \left| C(x|z^0_{-k+1} : [\lceil \frac{n}{2} \rceil, n]) - C(x|z^0_{-k-i+1} : [\lceil \frac{n}{2} \rceil, n]) \right| > n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil}) \\ \leq \; & P(\max_{(z^0_{-k-i+1},x)\in\mathcal{S}^n_{k+i}} \left| C(x|z^0_{-k+1} : [\lceil \frac{n}{2} \rceil, n]) - p(x|z^0_{-k+1}) \right| > 0.5 n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil}) \\ + \; & P(\max_{(z^0_{-k-i+1},x)\in\mathcal{S}^n_{k+i}} \left| p(x|z^0_{-k+1}) - C(x|z^0_{-k-i+1} : [\lceil \frac{n}{2} \rceil, n]) \right| > 0.5 n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil}). \end{aligned}$$

We overestimate these probabilities. For any $m \geq 0$ and x^0_{-m} define $\sigma_i^m(x^0_{-m})$ as the time of the i-th ocurrence of the string x^0_{-m} in the data segment $X^n_{\lceil \frac{n}{2} \rceil}$, that is, let $\sigma_0^m(x^0_{-m}) = \lceil \frac{n}{2} \rceil + m - 1$ and for $i \geq 1$ define

$$\sigma_i^m(x^0_{-m}) = \min\{t > \sigma_{i-1}^m(x^0_{-m}) : X^t_{t-m} = x^0_{-m}\}.$$

Now

$$\begin{aligned} & P(\max_{(z^0_{-k-i+1},x)\in\mathcal{S}^n_{k+i}} \left| C(x|z^0_{-k+1} : [\lceil \frac{n}{2} \rceil, n]) - C(x|z^0_{-k-i+1} : [\lceil \frac{n}{2} \rceil, n]) \right| > n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil}) \\ \leq \; & P(\max_{(z^0_{-k+1},x)\in\mathcal{S}^{(1)}_{n,k}} \sup_{j > n^{1-\gamma}} \\ & \left| \frac{1}{j} \sum_{r=1}^{j} 1_{\{X_{\sigma_r^{k-1}(z^0_{-k+1})} = x\}} - p(x|z^0_{-k+1}) \right| > 0.5 n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil}) \\ + \; & P(\max_{(z^0_{-k-i+1},x)\in\mathcal{S}^{(1)}_{n,k+i}} \sup_{j > n^{1-\gamma}} \\ & \left| \frac{1}{j} \sum_{r=1}^{j} 1_{\{X_{\sigma_r^{k+i-1}(z^0_{-k-i+1})} = x\}} - p(x|z^0_{-k+1}) \right| > 0.5 n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil}) \end{aligned}$$

Since both $\mathcal{S}^{(1)}_{n,k}$ and $\mathcal{S}^{(1)}_{n,k+i}$ depend solely on $X_0^{\lceil \frac{n}{2} \rceil}$ we get

$$
\begin{aligned}
&P(\max_{(z^0_{-k-i+1},x)\in\mathcal{S}^n_{k+i}} \left| C(x|z^0_{-k+1} : [\lceil \frac{n}{2} \rceil, n]) - C(x|z^0_{-k-i+1} : [\lceil \frac{n}{2} \rceil, n]) \right| > n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil}) \\
&\leq \sum_{(z^0_{-k+1},x)\in\mathcal{S}^{(1)}_{n,k}} \sum_{j=\lceil n^{1-\gamma} \rceil}^{\infty} P(\left| \frac{1}{j} \sum_{r=1}^{j} 1_{\{X_{\sigma_r^{k-1}(z^0_{-k+1})} = x\}} - p(x|z^0_{-k+1}) \right| \\
&\quad > 0.5 n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil}) \\
&+ \sum_{(z^0_{-k-i+1},x)\in\mathcal{S}^{(1)}_{n,k+i}} \sum_{j=\lceil n^{1-\gamma} \rceil}^{\infty} P(\left| \frac{1}{j} \sum_{r=1}^{j} 1_{\{X_{\sigma_r^{k+i-1}(z^0_{-k-i+1})} = x\}} \right. \\
&\quad \left. - p(x|z^0_{-k+1}) \right| > 0.5 n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil}).
\end{aligned}
$$

Each of these represents the deviation of an empirical count from its mean. The variables in question are independent since whenever the block z^0_{-k+1} occurs the next term is chosen using the same distribution $p(x|z^0_{-k+1})$. Thus by Hoeffding's inequality (cf. Hoeffding [5] or Theorem 8.1 of Devroye et. al. [4]) for sums of bounded independent random variables and since the cardinality of both $\mathcal{S}^{(1)}_{n,k}$ and $\mathcal{S}^{(1)}_{n,k+i}$ is not greater than $(n+2)/2$, we have

$$
\begin{aligned}
&P(\max_{(z^0_{-k-i+1},x)\in\mathcal{S}^n_{k+i}} \left| C(x|z^0_{-k+1} : [\lceil \frac{n}{2} \rceil, n]) - C(x|z^0_{-k-i+1} : [\lceil \frac{n}{2} \rceil, n]) \right| > n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil}) \\
&\leq 2 \frac{n+2}{2} \sum_{j=\lceil n^{1-\gamma} \rceil}^{\infty} 2 e^{-2n^{-2\beta} j}.
\end{aligned}
$$

Thus

$$
P(\hat{\Delta}^n_k > n^{-\beta} | X_0^{\lceil \frac{n}{2} \rceil}) \leq n(n+2) 4 e^{-2n^{-2\beta+1-\gamma}}.
$$

Integrating both sides we get

$$
P(\hat{\Delta}^n_k > n^{-\beta}) \leq n(n+2) 4 e^{-2n^{-2\beta+1-\gamma}}.
$$

The right hand side is summable provided $2\beta + \gamma < 1$ and the Borel-Cantelli Lemma yields that $P(\hat{\Delta}^n_k \leq n^{-\beta}$ eventually$) = 1$. Thus $\chi_n \leq k$ eventually almost surely provided the process is Markov with order k. The proof of the Theorem is complete.

References

[1] D. H. Bailey, *Sequential Schemes for Classifying and Predicting Ergodic Processes.* Ph. D. thesis, Stanford University, 1976.

[2] I. Csiszár and P. Shields, "The consistency of the BIC Markov order estimator," *Annals of Statistics.*, vol. 28, pp. 1601-1619, 2000.

[3] I. Csiszár, "Large-scale typicality of Markov sample paths and consistency of MDL order estimators ," *IEEE Transactions on Information Theory*, vol. 48, pp. 1616-1628, 2002.

[4] L Devroye, L. Györfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[5] W. Hoeffding, "Probability inequalities for sums of bounded random variables ," *Journal of the American Statistical Association*, vol. 58, pp. 13-30, 1963.

[6] I. Kontoyiannis, P. Algoet, Yu.M. Suhov, A.J. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with application to English text," *IEEE Transactions on Information Theory*, vol. 44, pp. 1319–1327, 1998.

[7] D. S. Ornstein and B. Weiss, "How sampling reveals a process," *The Annals of Probability*, vol. 18, pp. 905–930, 1990.

[8] D. S. Ornstein and B. Weiss, "Entropy and data compression schemes," *IEEE Transactions on Information Theory*, vol. 39, pp. 78–83, 1993.

[9] D. S. Ornstein and B. Weiss, "Entropy and recurrence rates for stationary random fields," *IEEE Transactions on Information Theory*, vol. 48, pp. 1699–1697, 2002.

[10] J. Ziv, " Coding theorems for individual sequences. *IEEE Transactions on Information Theory*, vol. 24, pp. 405–412, 1978.