ELECTRONIC RESEARCH ANNOUNCEMENTS
OF THE AMERICAN MATHEMATICAL SOCIETY
Volume 5, Pages 123–127 (October 19, 1999)
S 1079-6762(99)00070-0

CONSISTENCY OF THE BIC ORDER ESTIMATOR

IMRE CSISZÁR AND PAUL C. SHIELDS

(Communicated by Yitzhak Katznelson)

ABSTRACT. We announce two results on the problem of estimating the order of a Markov chain from observation of a sample path. First is that the Bayesian Information Criterion (BIC) leads to an almost surely consistent estimator. Second is that the Bayesian minimum description length estimator, of which the BIC estimator is an approximation, fails to be consistent for the uniformly distributed i.i.d. process. A key tool is a strong ratio-typicality result for empirical k-block distributions. Complete proofs are given in the authors' article to appear in The Annals of Statistics.

1. Introduction

Let $\mathcal{M}_k$ denote the class of Markov chains of order at most k, with values drawn from a finite set A, and let $\mathcal{M} = \bigcup_{k=0}^{\infty} \mathcal{M}_k$. An important problem is to estimate the order of a Markov chain from observation of a finite sample path. A popular method is the so-called Bayesian Information Criterion (BIC), first introduced by Schwarz, [12], which gives the estimator defined by

$$\hat{k}_{\mathrm{BIC}} = \hat{k}_{\mathrm{BIC}}(x_1^n) = \arg\min_k \left(-\log P_{\mathrm{ML}(k)}(x_1^n) + \frac{|A|^k(|A|-1)}{2} \log n \right), \tag{1.1}$$

where $P_{\mathrm{ML}(k)}(x_1^n)$ is the k-th order maximum likelihood, i.e., the largest probability given to x_1^n by processes in $\mathcal{M}_k$. Our first result is the following.

The BIC consistency theorem. *For any stationary, irreducible process in $\mathcal{M}$, $\hat{k}_{BIC}(X_1^n)$ is eventually almost surely equal to the order of the process.*

Our theorem does not assume a finite number of model classes. Claims of BIC consistency have been made by several authors; see, for example, [5, 1] as well as several papers listed in [1]. All these papers include the explicit or hidden assumption that there are only a finite number of model classes. In this case, however, as noted in [7, 5], $\log n$ can be replaced in the definition (1.1) by something much smaller, e.g., a suitable multiple of $\log\log n$. We also point out that Kieffer established consistency without an a priori bound on the order for the case when $|A|^k(|A|-1)$ is replaced in (1.1) by a more rapidly growing function of k, [8].

Received by the editors February 25, 1999.
1991 *Mathematics Subject Classification.* Primary 62F12, 62M05; Secondary 62F13, 60J10.
Key words and phrases. Bayesian information criterion, order estimation, ratio-typicality, Markov chains.
First author supported in part by a joint NSF-Hungarian Academy grant 92.
Second author supported in part by a joint NSF-Hungarian Academy grant INT-9515485.

©1999 American Mathematical Society

An essential tool in proving the consistency of the BIC order estimator is a strong ratio-typicality result we establish for stationary, irreducible Markov chains, a result that appears to be of independent interest. It says, loosely, that as long as k does not grow too rapidly with sample path length, the ratio of the empirical relative frequency of each k-block to its probability is uniformly close to 1. The empirical distribution of k-blocks in x_1^n, or the *k-type of x_1^n*, is defined by the formula

$$\hat{P}_n(a_1^k) = \frac{1}{n-k+1} N(a_1^k|x_1^n), \; a_1^k \in A^k, \tag{1.2}$$

where $N(a_1^k|x_1^n) = |\{i \in [1, n-k+1]\colon\; x_i^{i+k-1} = a_1^k\}|$ is the number of occurrences of a_1^k in x_1^n.

The sequence x_1^n is called *(k, ϵ)-typical* for a process Q if $\hat{P}_n(a_1^k) = 0$, whenever $Q(a_1^k) = 0$, and

$$\left| \frac{\hat{P}_n(a_1^k)}{Q(a_1^k)} - 1 \right| < \epsilon, \text{ whenever } Q(a_1^k) > 0. \tag{1.3}$$

The typicality theorem. *For any stationary irreducible process $Q \in \mathcal{M}$, there exist positive numbers α and β such that eventually almost surely as $n \to \infty$, the sequence x_1^n is $(k, n^{-\beta})$-typical for every $k \le \alpha \log n$.*

Earlier typicality results in which block length grows with sample path length include the following.

1. Marton and Shields have obtained large deviations bounds for the variational distance between the empirical k-block distribution and the theoretical k-block distribution that are valid for all $k \le (\log n)(H + \epsilon)$, where H is the process entropy, [10]. Their results do yield our typicality theorem for $k = o(\log \log n)$, but this is not sufficient for our proof of the BIC consistency theorem.
2. Flajolet, Kirschenhofer, and Tichy have obtained similar results for the case when Q is the unbiased coin-tossing process, [6]. They consider longer blocks but do not give an error rate.

The Bayesian framework that underlies the BIC starts with a prior distribution $\{p_k\}$ on the possible orders, together with a prior distribution on $\mathcal{M}_k$, for each k; the latter defines a mixture distribution, that is, a weighted average of the processes in $\mathcal{M}_k$. The Bayesian order estimator chooses that k for which the product of p_k and the mixture probability of x_1^n is largest. We consider here the mixture of those k-th order Markov chains whose starting distribution is uniform on A^k, taking as prior the $|A|^k$-fold product of the $(\frac{1}{2}, \ldots, \frac{1}{2})$ Dirichlet distributions put on the transition probability matrices $Q(\cdot|\cdot)$. This mixture distribution $\mathrm{KT}_k(x_1^n)$, whose explicit form is given in (2.1), below, plays a distinguished role in the theory of universal coding; see Krichevsky and Trofimov, [9]. It can be shown that

$$-\log \mathrm{KT}_k(x_1^n) \approx -\log P_{\mathrm{ML}(k)}(x_1^n) + \frac{|A|^k(|A|-1)}{2} \log n, \tag{1.4}$$

as $n \to \infty$ with k fixed, which is the principal reason for the estimator (1.1).

In contrast to the BIC consistency theorem, for the Bayesian order estimator

$$\hat{k}_{\mathrm{KT}}(x_1^n) = \arg\min_k \Big(-\log p_k - \log \mathrm{KT}_k(x_1^n) \Big), \tag{1.5}$$

we prove the following.

The inconsistency theorem. *If $\{X_n\}$ is i.i.d. with X_n uniformly distributed on A, then $\hat{k}_{KT}(x_1^n) \to +\infty$, almost surely, provided the p_k of (1.5) are taken, as usual, to be slowly decreasing, say $p_k = ck^{-2}$.*

Similar phenomena have been established in other Bayesian settings by Diaconis and Freedman; see, for example, [4]. The order estimator $\hat{k}_{\mathrm{KT}}(x_1^n)$ is often introduced via the minimum description length (MDL) principle; see Rissanen, [11], or Barron, Rissanen, and Yu, [1]. It is interesting to compare our inconsistency theorem with an MDL-inspired result by Barron; see [1]. That result, specialized to the estimator $\hat{k}_{\mathrm{KT}}(x_1^n)$, can be formulated as follows. If the processes of $\mathcal{M}_k$ are parametrized by an $|A|^k(|A|-1)$ dimensional parameter, then for Lebesgue almost all choices of the parameter, the estimator $\hat{k}_{\mathrm{KT}}(x_1^n)$ is eventually almost surely equal to the correct order. Our inconsistency theorem shows that "for almost every choice of the parameter," is not vacuous.

We note that the MDL principle also leads to non-Bayesian estimators, replacing, e.g., mixture distributions by normalized maximum likelihoods; see [1]. In our context this means replacing $\mathrm{KT}_k(x_1^n)$ by

$$P_{\mathrm{ML}(k)}(x_1^n) \Big/ \sum_{y_1^n \in A^n} P_{\mathrm{ML}(k)}(y_1^n).$$

The inconsistency theorem also holds for the resulting order estimator.

2. Sketches of the proofs

Fix a stationary, irreducible Markov chain of order k_0 and distribution Q. In the sequel we use the notation $N_n(a_1^m) = N(a_1^m|x_1^n)$. The proof of the typicality theorem utilizes the fact that for any $m > k_0$, the sequence

$$\{N_n(a_1^m) - N_{n-1}(a_1^{m-1})Q(a_m|a_{m-k_0}^{m-1})\colon n \geq m\}$$

is a martingale. A direct calculation of the variance and an application of Kolmogorov's inequality for martingales then give

$$\mathrm{Prob}\left\{\max_{n \leq 2^i}\left|N_n(a_1^m) - N_{n-1}(a_1^{m-1})Q(a_m|a_{m-k_0}^{m-1})\right| > \lambda\right\} < \frac{2^i Q(a_1^m)}{\lambda^2}.$$

This, in turn, for suitable $\alpha > 0$ and $\beta > 0$ leads to a summable bound on the probability of the event that

$$\max_{2^{i-1} < n \leq 2^i}\left|\frac{\hat{P}_n(a_1^k)}{Q(a_1^k)} - \frac{\hat{P}_{n-(k-k_0)}(a_1^{k_0})}{Q(a_1^{k_0})}\right| > 2^{-i\beta}$$

occurs for some $k \in (k_0, \alpha \log 2^i]$ and $a_1^k \in A^k$ with $Q(a_1^k) > 0$. The typicality theorem follows because the assertion is clearly true for $k = k_0$, by the law of the iterated logarithm for Markov chains.

Next, we use a counting argument known in information theory as the method of types, to bound the probability of the event that both $\hat{k}_{\mathrm{BIC}}(x_1^n) = k$ and x_1^n is $(k+1, n^{-\beta})$-typical. One role of typicality in our argument is that the number of different $(k+1)$-types that typical sequences x_1^n can have admits a better upper bound than the number of all possible $(k+1)$-types. It can be shown that for the α and β of the typicality theorem and n^* and k^* large enough, the probability that

x_1^n is both $(k+1, n^{-\beta})$-typical and $\hat{k}_{\mathrm{BIC}} = k$, is upper bounded by n^{-2} for every $n \geq n^*$ and $k^* \leq k \leq \alpha \log n$. It follows that, eventually almost surely, $\hat{k}_{\mathrm{BIC}}(x_1^n)$ cannot belong to the interval $(k^*, \alpha \log n)$.

The key to ruling out larger values of k is that the estimate (1.4) substantially overestimates $-\log \mathrm{KT}_k(x_1^n)$, for large n and k. Indeed, using the explicit formula

(2.1)

$$\mathrm{KT}_k(x_1^n) = \frac{1}{|A|^k} \prod_{a_1^k \in x_1^{n-1}} \left[\frac{\prod_{a_{k+1}\,:\, a_1^{k+1} \in x_1^n} (N_n(a_1^{k+1}) - \frac{1}{2})(N_n(a_1^{k+1}) - \frac{3}{2}) \cdots (\frac{1}{2})}{(N_{n-1}(a_1^k) - 1 + \frac{|A|}{2})(N_{n-1}(a_1^k) - 2 + \frac{|A|}{2}) \cdots (\frac{|A|}{2})} \right],$$

it can be shown that

$$\log \mathrm{KT}_k(x_1^n) \geq \log P_{\mathrm{ML}(k)}(x_1^n) - \frac{|A|^k(|A|-1)}{2} \log^+ \frac{n}{|A|^k} - C|A|^k, \tag{2.2}$$

for a suitable constant C, independent of n and k. This easily leads to $\hat{k}_{\mathrm{BIC}}(x_1^n) = o(\log n)$, a.s. The BIC consistency theorem follows, since it is known (see [5]) that $\hat{k}_{\mathrm{BIC}}(x_1^n) \notin [0, k_0) \cup (k_0, k^*]$, eventually a.s.

Now let Q denote the i.i.d. process with $Q(a) = |A|^{-1}, a \in A$. Two ideas are used to establish the inconsistency theorem. The first is that if no k-block occurs in x_1^{n-1} more than once, then $\mathrm{KT}_k(x_1^n) = |A|^{-n}$. This follows easily from the representation (2.1). The second is that if α is large enough and $k = \alpha \log n$, the probability that a k-block appears more than once in x_1^{n-1} is upper bounded by n^{-2}. The two ideas together show that $\hat{k}_{\mathrm{KT}}(x_1^n) > 0$, eventually almost surely. Using a bound similar to (2.2), it can be shown that $\hat{k}_{\mathrm{KT}}(x_1^n) \neq k$, eventually almost surely, for any $k > 0$, which combines with the above to show that $\hat{k}_{\mathrm{KT}}(x_1^n) \to \infty$, almost surely.

References

1. A. Barron, J. Rissanen, and B. Yu, *The minimum description length principle in coding and modeling,* IEEE Trans. Inform. Th. 44 (1998), 2743–2760. MR **99h**:94032
2. I. Csiszár and J. Körner, *Information Theory. Coding theorems for discrete memoryless systems*, Akadémiai Kiadó, Budapest, 1981. MR **84e**:94007
3. I. Csiszár and P. Shields, *The consistency of the BIC order estimator,* Ann. Statis., submitted.
4. P. Diaconis and D. Freedman, *Nonparametric binary regression: a Bayesian approach,* Ann. Statist. 21 (1993), 2108–2137. MR **94i**:62001
5. L. Finesso, *Estimation of the order of a finite Markov chain,* in *Recent Advances in the Mathematical Theory of Systems, Control, and Network Signals, Proc. MTNS-91,* H. Kimura and S. Kodama, Eds., Mita Press, 1992, pp. 643–645. MR **93g**:93005
6. P. Flajolet, P. Kirschenhofer, and R. F. Tichy, *Deviations from uniformity in random strings,* Probab. Th. Rel. Fields 80 (1988), 139–150. MR **90a**:11087
7. D. Haughton, *On the choice of a model to fit data from an exponential family,* Ann. Statist. 16 (1988), 342–355. MR **89e**:62036
8. J. Kieffer, *Strongly consistent code-based identification and order estimation for constrained finite-state model classes,* IEEE Trans. Inform. Th. 39 (1993), 803–902.
9. R. E. Krichevsky and V. K. Trofimov, *The performance of universal encoding,* IEEE Trans. Inform. Th. 27 (1981), 199–207. MR **83e**:94030
10. K. Marton and P. Shields, *Entropy and the consistent estimation of joint distributions,* Ann. Probab. 22 (1994), 960–977. (Correction, Ann. Probab. 24 (1996), 541–545.) MR **95g**:94004; MR **97c**:94004

11. J. Rissanen, *Stochastic complexity in statistical inquiry*, World Scientific, Singapore, 1989. MR **92f:**68076
12. G. Schwarz, *Estimating the dimension of a model,* Ann. Statist. 6 (1978), 461–464. MR **57:**7855

A. Rényi Institute of Mathematics, Hungarian Academy of Sciences, POB 127, 1364 Budapest, Hungary
E-mail address: csiszar@math-inst.hu

Mathematics Department, The University of Toledo, Toledo, OH 43606
E-mail address: paul.shields@utoledo.edu