Decision theory for finding the order of a m-Markov Chain

Jan Lennartsson, Ph.D. student

7th September 2006

Abstract

The order estimation of multiple steps Markov chains is an open area of research. Different order estimators exhibit different properties. By simulation the new order estimator, the Generalized Maximal Fluctuation Criteria, introduced by Peres and Shields clearly out perform the other estimators in accuracy i.e. for x_1^n sampled from a k-Markov chain, $P(\hat{k}_{\mathrm{GMFC}}(x_1^n) = k)$ is generally larger than $P(\hat{k}(x_1^n) = k))$ for all other suggested order estimators.

1 Introduction

Multiple step Markov chains are a widely used tool in the analysis of sequenced data. This essay will focus on the problem "how much knowledge from the past enhances your ability to predict the next outcome?". This is called the order estimation problem and is a frequent topic of research.

The historically most common estimators in the literature are the AIC [Akaike, 1974] and BIC [Schwartz, 1978] but just recently Peres and Shields suggested a new order estimators, the Maximal fluctuation estimator, [Peres and Shields, 2006]. I will also include, what I suppose is, the most basic idea of estimating the order namely home made criteria based on the generalized likelihood ratio tests.

2 Multiple steps Markov chains

A k-Markov chain on a finite state space, S, is a discrete stochastic process on S that is only dependent on the previous k steps. Let X_n be a k-Markov chain on S then this is in mathematical notation:

$$P(X_n = s_n | X_{n-k-1} = s_{n-k-1}, ..., X_{n-1} = s_{n-1}) = P(X_n = s_n | F_{n-1}) \quad \forall n > k.$$

Where $s_i \in S$ and F_n denotes the filtration of $\{X_k : k \geq 1\}$ up till n.

To abbreviate and ease the comprehension let us define the following notation: let a_m^n denote the sequence $a_m, a_{m+1}, ..., a_n$ of $\{a_i : m \leq i \leq n\}$. Let the state space, S, be called the alphabet, elements, s, in S be called letters and compositions of letters $v = s_1^l$ be called words (of length l). Let also for some word, $v \in S^l$, $N(v|x_1^n) = |\{i \in [1, n-l] : x_i^{i+l} = v\}|$ denote the number occurrences of v in x_1^n.

Actually the class of k-Markov chains is no generalization of the ordinary Markov chains it is rather a subclass of Markov chains. To see this let $\{Y_n : n \geq 1\}$ be a stochastic process on S^k such that $Y_n = X_{n-k-1}^n$. Then

$$P(Y_n = s_{n-k}^n | Y_{n-1} = s_{n-k-1}^{n-1}) = P(X_n = s_n | X_{n-k-1}^{n-1} = s_{n-k-1}^{n-1}) = \\ P(X_n = s_n | F_{n-1}) = P(Y_n = s_{n-k}^n | F_{n-1})$$

and $\{Y_n : n \geq 1\}$ is a Markov Chain on the state space S^k.

Let now M_k be the vector space of all k-Markov processes with finite state space S, M_0 denote the space of i.i.d. processes (on S) and $P_{\mathrm{ML}(k)}(x_1^n)$ be the kth order maximum likelihood i.e. the largest probability given to the outcome x_1^n by a process in M_k.

Now we can start talk about the order of a process. The order of a process, $\{X_n : n \geq 1\} \in \cup_{k=0}^{\infty} M_k$, is the least k such that for some $l \geq 0$ then $\{X_n : n \geq l\} \in M_k$.

3 Order estimators

A order estimator is a function of the type $M_n^* : S^n \to N$ for $n \geq 1$ that in theory should give $M_n^*(x_1^n) = k$ for all realizations x_1^n of $\{X_i : i \in [1, n]\} \in M_k$. from now and in the sequel x (lower-case) denotes a sampled version of the stochastic variable X (upper-case). However this convergence result is hardly achievable for any estimator and has to be lightened to have a practical use. No well defined requirement for the estimators exist for finite data sets. Instead, a great effort of the research community has been invested in the area of limits of large data sets. An order estimator M_n^* is called consistent if

$$\lim_{n \to \infty} M_n^*(x_1^n) = k \text{ a.s.}$$

or equivalent $\lim_{n\to\infty} P(M_n^*(x_1^n) = k) = 1$. That an order estimator is consistent does not however give any information about the convergence rate and for practical use one would be more interested in a measure of the accuracy of the estimator on finite data sets. That is what one really wants is $P(M_n^*(x_1^n) = k)$ as a function of n for each estimator M_n^* and order k (and possibly size of alphabet $|S|$, transition probabilities $P(X_n = x_n | X_{n-k-1}^{n-1})$ etc.), but general ideas of how to find that lies in the future development of the order estimators. So the question of which order estimator is the best is still open and different purposes demands different estimators.

3.1 Akaike Information Criterion (AIC)

The first still standing order estimator is the AIC order estimator introduced by Akaike [Akaike, 1974]. AIC is based on the maximum likelihood estimator compensated with the number of independently adjusted parameters within the model. Actually the criteria is general and is highly applicable to other areas where the number of parameters is to be chosen. In the case of multiple steps Markov chains this becomes for sampled data set x_1^n:

$$\hat{k}_{\mathrm{AIC}}(x_1^n) = \arg\min_k - \log P_{\mathrm{ML}(k)}(x_1^n) + |S|^k.$$

The idea of AIC is to correct for the bias of the estimation of the entropy.

Let X_n be k_0-Markov with the alphabet S then $n^{-1} P_{\mathrm{ML}(k)}(x_1^n)$ is an estimate of the entropy,

$$E[\log(X; \text{for optimal } |S|^k \text{ parameters})],$$

but

$$E[\,E[\log(X;\text{for optimal } |S|^k \text{ parameters})] - n^{-1} \log\, P_{\mathrm{ML}(k)}(x_1^n))]\,] \approx -|S|^k/n$$

so $\hat{k}_{\mathrm{AIC}}$ is an unbiased minimal entropy estimate w.r.t. to the order of the process.

The AIC estimator is not consistent. To prove this let x_1^n be a sample set of $X_n \in M_{k_0}$ $(k_0 > 0)$ and choose a $k_{max} > k_0$ then for all $k \leq k_{max}$,

$$\lim_{n\to\infty} P_{\mathrm{ML}(k)}(x_1^n) = 0$$

so

$$\lim_{n\to\infty} \hat{k}_{\mathrm{AIC}}(x_1^n) \in \{0\} \cup (k_{max}, \infty]$$

and can not be a consistent estimate of the order of X_n.

3.2 Bayesian Information Criterion (BIC)

Schwartz introduced the BIC order estimator [Schwartz, 1978] but not until recently was the consistency established under general conditions [Csiszar and Shields, 2000].

The BIC estimator for multiple steps Markov chains of data set x_1^n is

$$\hat{k}_{\mathrm{BIC}}(x_1^n) = \arg\,\min_k - \log\, P_{\mathrm{ML}(k)}(x_1^n) + \frac{|S|^k(|S|-1)}{2}\log(n).$$

The complete proof of consistency for the general case is quite long but I will give you a proof of that for a irreducible aperiodic process $X_n \in M_{k_0}$

$$\hat{k}_{\mathrm{BIC}}(x_1^n) \notin [0, k_0) \quad \text{eventually a.s..}$$

That is I show that eventually a.s. no underestimation of the order will occur. I leave the part of long dependence structure and to complete the proof, like is done in [Csiszar and Shields, 2000] , one should also include a part proving that $\hat{k}_{\mathrm{BIC}}(x_1^n) \notin (k_0, n]$ eventually a.s..

Let $X_n \in M_{k_0}$ be an irreducible aperiodic process and let Q be a stationary probability measure of X_n. Then the ergodic theorem gives that for $k < k_0$

$$\lim_{n\to\infty} \frac{N(a_1^{k+1}|x_1^n)}{n-k} = \sum_{b_1^{k_0-k-1} \in S^{k_0-k-1}} Q([b_1^{k_0-k-1}\; a_1^{k+1}]) \quad \text{a.s.,} \tag{1}$$

where the sum is taken to be $Q(a_1^{k+1})$ if $k+1 = k_0$. I abbreviate the sum on the right hand side of (1) to $Q^*(a_1^{k+1})$.

Notice that

$$\hat{k}_{\mathrm{BIC}} \notin [0, k_0) \quad \text{eventually a.s.,}$$

is equivalent to the statement: for all $k < k_0$ there is a positive constant, C, such that

$$-\log\, P_{\mathrm{ML}(k)}(x_1^n) \geq -\log\, P_{\mathrm{ML}(k_0)}(x_1^n) + Cn, \quad \text{eventually a.s.} \tag{2}$$

Remember that $P_{\mathrm{ML}(k)}(x_1^n)$ is the maximal likelihood of x_1^n sampled from a process of M_k i.e.

$$\log P_{\mathrm{ML}(k)}(x_1^n) = \log P(X_1^k = x_1^k) \prod_{k+1}^{n} P(X_i = x_i | X_{i-k}^{i-1} = x_{i-k}^{i-1}).$$

Set $P(X_1^k = x_1^k) = 1$ then

$$\log P_{\mathrm{ML}(k)}(x_1^n) = \sum_{k+1}^{n} \log P(X_i = x_i | X_{i-k}^{i-1} = x_{i-k}^{i-1}) =$$
$$\sum_{a_1^{k+1} \in x_1^n} N(a_1^{k+1}|x_1^n) \log(P(X_{k+1} = a_{k+1}|X_1^k = a_1^k)) =$$
$$\sum_{a_1^{k+1} \in x_1^n} N(a_1^{k+1}|x_1^n) \log\Big(\frac{N(a_1^{k+1}|x_1^n)}{N(a_1^k|x_1^{n-1})}\Big).$$

Here the third equality comes from that the sum over the logarithmed transition probabilities are maximized when

$$P(X_{k+1} = a_{k+1}|X_1^k = a_1^k) = \frac{N(a_1^{k+1}|x_1^n)}{N(a_1^k|x_1^{n-1})}.$$

The ergodic theorem (see [Norris, 1997]) gives

$$\lim_{n\to\infty} -\frac{1}{n} \log P_{\mathrm{ML}(k)}(x_1^n) = -\sum_{a_1^{k+1} \in S^{k+1}} Q^*(a_1^{k+1}) \log \frac{Q^*(a_1^{k+1})}{Q^*(a_1^k)} \quad \text{a.s.},$$

and hence the limit is the conditional entropy, H_k, of $X_n \in M_k$ given X_1^k. It is well known that H_k is strictly greater than H_{k_0} if $k < k_0$ (see [Shields, 1996]). So there exist a $C > 0$ such that

$$\lim_{n\to\infty} -\frac{1}{n} \log P_{\mathrm{ML}(k)}(x_1^n) \geq \lim_{n\to\infty} -\frac{1}{n} \log P_{\mathrm{ML}(k_0)}(x_1^n) + C \quad \text{a.s}$$

and multiplying by n on both sides gives what was asked for.

3.3 Maximal Fluctuation Criterion (MFC) and Generalized Maximal Fluctuation Criterion (GMFC)

In contrast to the other here described order estimators the Maximal Fluctuation Criterion is designed for multiple steps Markov chains. Introduce the notation $\tau_m(v) = v_{l-m+1}^l$ for some $m \in [1, l]$ to be the m-suffix of the word $v \in S^l$. Then the *Peres-Shields Fluctuation* function of sampled data set x_1^n is defined as

$$\Delta_k(v) = \max_{a \in S} |N(va|x_1^n) - \frac{N(\tau_k(v)a|x_1^n)}{N(\tau_k(v)|x_1^n)} N(v|x_1^n)|.$$

If the true order is k or less, one expects this fluctuation to be small, otherwise large. The Maximal Fluctuation estimator is defined as

$$\hat{k}_{\mathrm{MFC}}(x_1^n) = \min\{k \geq 0 : \max_{k < |v| < \log\log(n)} \Delta_k(v) < n^{3/4}\}$$

(see [Peres and Shields, 2006]).

Actually the upper threshold on the length of the word, $\log\log(n)$, can be relaxed to a function that grows slower to infinity than $\log(n)$.

I will give you a proof of that for a process X_n in M_{k_0}

$$\hat{k}_{\mathrm{BIC}}(x_1^n) \notin [0, k_0) \quad \text{eventually a.s..}$$

Let

$$\delta_k(a_1^m) = N(a_1^m|x_1^n) - N(a_1^{m-1}|x_1^{n-1})\frac{N(\tau_k(a_1^m)|x_1^n)}{N(\tau_k(a_1^{m-1})|x_1^{n-1})},$$

and note that $\Delta_k = \max_{k<m<\log\log(n)} \max_{a_1^k \in S^k} \delta_k(a_1^k)$ of data set x_1^n. Suppose $X_n \in M_{k_0}$ and for $k < k_0$ choose $a \in S$ and $v \in S^{k_0}$ such that

$$P(X_{k_0+1} = a|X_1^{k_0} = v) > P(X_{k_0+1} = a|\tau_k(X_1^{k_0}) = \tau_k(v)).$$

By the ergodic theorem (see again [Norris, 1997]) there exist $\epsilon > 0$ such that,

$$N(v|x_1^{n-1}) > \epsilon n \quad \text{and} \quad \frac{N(va|x_1^n)}{N(v|x_1^{n-1})} - \frac{N(\tau_k(va)|x_1^n)}{N(\tau_k(v)|x_1^{n-1})} \geq \epsilon \quad \text{eventually a.s.}$$

This implies that $\delta_k(va) \geq \epsilon^2 n$ eventually a.s. and since definitely the inequality $\Delta_k \geq \delta_k(va)$ holds the conclusion that $\hat{k}_{\mathrm{BIC}} \geq k_0$ eventually a.s. is evident.

Although consistency of $\hat{k}_{\mathrm{MFC}}$ is well established it is harder to use this method in applications. The threshold $n^{3/4}$ is not sharp in practical use and although under estimation of the order eventually a.s. never happens it will "almost always" happen for finite applications. Where "almost always" is defined as all the applications I have so far come across. The maximal fluctuation method can however be modified to be highly functional for applications. Dalvei et al. suggested the (closely related) estimator

$$\hat{k}_{\mathrm{GMFC}}(x_1^n) = \arg\max_k \frac{\max_{k-1<|v|<f(n)} \Delta_{k-1}(v)}{\max_{k<|v|<f(n)} \Delta_k(v)}$$

(see [Dalevi et al., 2006]), where $f(n)$ is a function growing to infinity slower than $\log(n)$.

3.4 Generalized Likelihood Ratio Criterion (GLRC)

The likelihood ratio test (LRT) is a statistical test of the goodness-of-fit between two models [Rice, 1995]. A relatively simpler model is compared to a more complex model to see if the data set fits the complex model significantly better. The test is based on the likelihood ratio or rather the log likelihood difference between the two models. To formulate a criterion founded by the LRT we let the estimated order of a process be the largest k such that the improvement is significant.

Let $\xi_k = \log P_{\mathrm{ML}(k+1)}(x_1^n) - \log P_{\mathrm{ML}(k)}(x_1^n)$ and if $k < k_0$

$$2\xi_k \overset{D}{\approx} \chi_{df(k)^2}.$$

Where $df(k) = |S|^k(|S|-1)$ is the number of additional parameters. Now I define the Generalized Likelihood Ratio Criterion of data set x_1^n as:

$$\hat{k}_{\mathrm{GLRC}}(x_1^n) = \max_k\{2\xi_l \geq {\chi^2_{df(l)}}^{-1}(1-\alpha), \forall l \leq k\}.$$

Clearly, since there always is a probability, $\alpha > 0$, for each $l \leq k_0$ for the estimator to stop short, the $\hat{k}_{\mathrm{GLRC}}$ can not be consistent.

4 Applications

To get some information of how good the order estimators perform in applications, I simulated a process in M_k and estimated to what extent the order estimators replicated the true order k. That is I simulated $X_1^n \in M_k$ with alphabet $S = \{0, 1\}$ m times and for each simulation I estimated the order of the process with the four depict estimators. Then I estimated the accuracy, $P(M_n^*(x_1^n) = k)$ of estimator M_n^* by

$$\hat{P}(M_n^*(x_1^n) = k) = \frac{|\{i : \hat{k}(x_1^n(i)) = k, i = 1...m\}|}{m}.$$

The transition probabilities for $X_1^n \in M_k$ are the maximum likelihood estimated parameters from a data set consisting of precipitation data ($X_i = 1$ precipitation on day i and $X_i = 0$ no precipitation on day i).

This, the process described in the first paragraph, was repeated for different n and k:s to gain, at least some, information of which estimator that was most efficient i.e. "which estimator that demands least number of data points to achieve the highest accuracy of correct order estimation". Though the simulations were quite computer heavy I only used $m = 100$. The variability may been very high and is to be analyzed in the conclusions.

Estimated accuracy	$k = 2$	$k = 3$	$k = 4$
$\hat{P}(\hat{k}_{\text{AIC}} = k)$	0.68	0.88	0.99
$\hat{P}(\hat{k}_{\text{BIC}} = k)$	0.85	0.89	0.92
$\hat{P}(\hat{k}_{\text{GMFC}} = k)$	0.89	0.92	0.94
$\hat{P}(\hat{k}_{\text{GLRC}} = k)$	0.80	0.88	0.89

Table 1: Estimated accuracy of order estimators for data set x_1^{5000} simulated from $X_1^n \in M_k$.

Estimated accuracy	$n = 1000$	$n = 5000$	$n = 10000$
$\hat{P}(\hat{k}_{\text{AIC}} = 3)$	0.93	0.93	0.93
$\hat{P}(\hat{k}_{\text{BIC}} = 3)$	0.89	0.92	0.94
$\hat{P}(\hat{k}_{\text{GMFC}} = 3)$	0.94	0.96	0.92
$\hat{P}(\hat{k}_{\text{GLRC}} = 3)$	0.86	0.88	0.89

Table 2: Estimated accuracy of order estimators for data set x_1^n simulated from $X_n \in M_3$.

5 Summary

Clearly, the Generalized Maximal Fluctuation Criterion perform best. It scores the best accuracy for moderate ($n = 1000$, $n = 5000$) sizes of data sets when considering a 3-Markov process and also for moderate ($k = 2$, $k = 3$) length of the Markov chain.

My home made order estimator is a poor straggler and achieves the lowest accuracy for all cases but $n = 1000$, $m = 2$ where the AIC estimator only has a accuracy of $\frac{2}{3}$:s and hardly is applicable.

Columns 2 of both tables should represent same accuracies but clearly they do not. This is due to only using $m = 100$ realizations of X_1^n to estimate the accuracy. By an application of the Central Limit Theorem (CLT) this variations can be explained. Let $\xi_i = 1_{\hat{k}(x_1^n(i))=k}$ and $\hat{\xi} = \frac{1}{m}\sum_{i=1}^m \xi_i$, then ξ_i is i.i.d. and ξ is (obviously) a sum of i.i.d. variables. Expected value,

$$E[\xi_i] = E[1_{\hat{k}(x_1^n)=k}] = P(\hat{k}(x_1^n) = k),$$

and second moment,

$$E[\xi_i^2] = E[(1_{\hat{k}(x_1^n(i))=k})^2] = E[1_{\hat{k}(x_1^n(i))=k}] = P(\hat{k}(x_1^n(i)) = k)$$

gives

$$VAR(\xi_i) = E[\xi_i^2] - E[\xi_i]^2 = P(\hat{k}(x_1^n(i)) = k)(1 - P(\hat{k}(x_1^n(i)) = k)).$$

By using CLT

$$\frac{\hat{\xi} - E[\xi_i]}{\sqrt{VAR(\xi_i)}/m} \overset{D}{\approx} N(0,1)$$

With the assumption (from tables) that $P(\hat{k}(x_1^n) = k) \approx 0.9$ a 90% confidence interval on the accuracy, $P(\hat{k}(x_1^n) = k)$, is found by

$$\hat{\xi} \pm Z_{90\%}\sqrt{VAR(\hat{\xi})/m} \approx \hat{\xi} \pm 0.05,$$

which explains the different results displayed in column 2 of the tables.

References

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, AC-19, No. 6.

[Csiszar and Shields, 2000] Csiszar, I. and Shields, P. (2000). The consistency of the bic markov order estimator. *The Annals of Statistics*, 28, No. 6:1601–1619.

[Dalevi et al., 2006] Dalevi, D., Dubhashi, D., and Hermansson, M. (2006). A new order estimator for fixed and variable length Markov models with applications to dna sequence similarity. *Statistical Applications in Genetics and Molecular Biology*, 5, Issue 1, Article 8.

[Norris, 1997] Norris, J. (1997). *Markov Chains.* Cambridge U P.

[Peres and Shields, 2006] Peres, Y. and Shields, P. (2006). Two new markov order estimators. *Electronic print*, arXiv:math.ST0506080.

[Rice, 1995] Rice, J. (1995). *Mathematical statistics and data analysis.* Duxbury Press.

[Schwartz, 1978] Schwartz, G. (1978). Estimating the dimension of a model. *The annals of Statistics*, 6:416–464.

[Shields, 1996] Shields, P. (1996). *The Ergodic Theory of Discrete Sample paths.* American Mathematics Society.