Inference on Markov chains parameters via Large Deviations ABC

Inferenza sui parametri di Catene di Markov mediante Large Deviations ABC

Cecilia Viscardi, Fabio Corradi, Michele Boreale, Antonietta Mira

Abstract We propose a method for Bayesian inference on the parameters governing the transition probabilities of finite state Markov chains. We address the difficulty of deriving the parameters' posterior distribution when the likelihood function is unavailable or computationally demanding to evaluate. The approach is an extension of the Large Deviations Approximate Bayesian Computation already proposed for i.i.d random variables. The method is developed by accommodating an information theoretic formulation of the Large Deviations Theory into Approximate Bayesian Computation (ABC). By contrast to the customary ABC, this approach avoids discarding parameter values having an (exponentially) small probability of producing simulation outcomes close to the observed data. We experimentally evaluate our method through a toy example.

Abstract *Proponiamo un metodo di inferenza Bayesiana per l'apprendimento dei parametri che governano le probabilità di transizione in catene di Markov a stati finiti qualora la funzione di verosimiglianza non è derivabile analiticamente ed una sua valutazione è computazionalmente costosa. In particolare, estendiamo alle catene di Markov un metodo di Approximate Bayesian Computation (ABC) basato sulla teoria delle Grandi Deviazioni proposto per variabili discrete i.i.d.. Il risultato è ottenuto integrando la teoria delle grandi deviazioni entro ABC. Questo metodo consente di non scartare le proposte di parametri che hanno una (esponenzialmente) piccola probabilità di produrre dati simulati simili a quelli osservati. Il metodo è illustrato attroverso un semplice esempio.*

Key words: ABC, Large deviations, Parametric Markov chains, Sample degeneracy, Method of Types.

1 Introduction and preliminary concepts

Parametric Markov chains (pMC) are discrete time Markov chains whose transitions probabilities are expressed as polynomials of real-valued parameters [4, 6]. Statistical methods for inferring the parameters governing such transition probabilities have been proposed, both from a classical and Bayesian viewpoint. Here we propose a method for deriving the parameters' posterior distributions when the evaluation of the joint probability function

of the Markovian sequence given the parameters is infeasible. In particular, we extend to finite state pMC the Large Deviations Approximate Bayesian Computation (LD-ABC) proposed in [13, 11] for i.i.d. discrete data. Generally speaking, Approximate Bayesian Computation (ABC) [9] is a class of likelihood-free methods allowing Bayesian inference when the likelihood function is intractable and only requiring the ability of simulating pseudo-data from a *simulator*, i.e., a probabilistic program reproducing the stochastic data generating process. The LD-ABC method represents a novel proposal for improving the ABC performances by mitigating the *sample degeneracy* problem in ABC. The method enhance the ABC likelihood resorting to an information theoretic formulation of Large Deviations Theory (LDT) based on the Method of Types [3].

Preliminary concepts

Let $\{X_t\}$ be a *stationary* parametric Markov process taking values in a finite set $\mathbb{A} \stackrel{\triangle}{=} \{a_1,...,a_k\}$ with cardinality k. For simplicity the elements of $\mathbb{A}$ will be hereafter denoted by their labels $\{1,...,k\}$. The Markov process can be characterized by its *doublet probability distribution* (dpd), P_θ, defined as a non-negative matrix of order $k \times k$ inducing a probability measure $\Pr\{\cdot,\cdot|\theta\}$ over $\mathbb{A}^2 \stackrel{\triangle}{=} \mathbb{A} \times \mathbb{A}$. Thus, denoted by $P_\theta(ij)$, the entries of P_θ are

$$P_\theta(ij) \stackrel{\triangle}{=} \Pr\{X_t = i, X_{t+1} = j|\theta\} \qquad \forall (i,j) \in \mathbb{A}^2$$

and sum to 1. The subscript θ indicates the dependence from the parameter (or vector of parameters) θ, object of our inference.

Let us denote by Δ^{k^2-1} the (k^2-1)-simplex, i.e., the set of possible dpd over $\mathbb{A}^2$, and by $\mathscr{M}(\mathbb{A}^2) \subset \Delta^{k^2-1}$ the set of the stationary dpd. Each $P_\theta \in \mathscr{M}(\mathbb{A}^2)$ is characterized by entries such that $\sum_{j\in\mathbb{A}} P_\theta(ij) = \sum_{j\in\mathbb{A}} P_\theta(ji)$, $\forall i \in \mathbb{A}$. This implies that the probability distribution over $\mathbb{A}$, $p_\theta \stackrel{\triangle}{=} \{p_\theta(i) = \sum_{j\in\mathbb{A}} P_\theta(ij),\ \forall j \in \mathbb{A}\}$, is invariant along the process and P_θ captures all the relevant information about it. In fact, the *state transition matrix* of the pMC, Q_θ, is the stochastic matrix of order $k \times k$ composed by entries retrieved from P_θ

$$q_\theta(ij) \stackrel{\triangle}{=} \Pr\{X_{t+1} = j|X_t = i,\theta\} = \frac{P_\theta(ij)}{p_\theta(i)} \quad \forall (i,j) \in \mathbb{A}^2$$

and p_θ is a (normalized) row eigenvector of Q_θ corresponding to eigenvalue 1:

$$(p_\theta Q_\theta)_j = \sum_{i\in\mathbb{A}} p_\theta(i) q_\theta(ij) = \sum_{i\in\mathbb{A}} P_\theta(ij) = p_\theta(j) \qquad \forall j \in \mathbb{A}.$$

2 The Method of Types for Markov chains and LDT

The Method of Types (MoT)[3] is a powerful tool shifting the focus from a vector of random variables to a lower dimensional vector: the *type*. Originally, MoT has been proposed for i.i.d. random variables and the 1st order type was defined as the empirical distribution of a sequence of random variables. Here we consider an extension: the 2nd order type which is suitable for observations modelled as Markov chains.

Given a Markov process $\{X_t\}$ with dpd P_θ and an observed sample path $x^n = x_1,...,x_n$ from the Markov process, the 2nd order type [3] is defined by

$$T^{(2)}_{x^n}(i,j) \stackrel{\triangle}{=} \frac{1}{n-1}\sum_{t=1}^{n-1} \mathbb{1}\{x_t = i, x_{t+1} = j\} \quad \forall (i,j) \in \mathbb{A}^2.$$

This type can be thought as a matrix of order $k \times k$ representing an empirical estimate of P_θ. An alternative definition is based on the *cyclic convention* that poses the $(n+1)$-th element of the path equal to x_1 and ensures the stationarity of the 2nd order type obtained from the n terms. This definition allows for establishing the MoT formulation of the Large Deviations principle for Markov chains.

LDT is concerned with probabilities of *rare events* going to zero with an exponential decay. A well-known result is the Sanov's Theorem (see[2, Th. 11.4.1]) which establishes the *rate function*, i.e., the function quantifying the probability of rare events, for sequences of i.i.d. random variables. Its analog for Markov chains can be found in the Donsker and Varadhan Theorem [5] and has been presented as an application of the MoT by Csiszár [3]. In what follows we let $D_c(\cdot||\cdot)$ be the *conditional relative entropy* (see [2] for a definition) and log be the logarithm to base 2.

Theorem 1. *Let $\{X_t\}$ be a Markov process taking values in the finite set $\mathbb{A}$, with stationary doublet probability distribution $P_\theta \in \mathscr{M}(\mathbb{A}^2)$ and let $X^n = X_1, ..., X_n$. If $E \subseteq \mathscr{M}(\mathbb{A}^2)$, then for each $\theta \in \Theta$*

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{T^{(2)}_{X^n} \in E|\theta\} = -\inf_{P\in E} D_c(P||P_\theta) = -D_c(E||P_\theta). \tag{1}$$

Proof. See [8] for a proof based on an easy counting approach.

3 ABC for finite state Markov Chains

Let x^n be an observed sequence from a Markov process $\{X_t\}$ taking values in $\mathbb{A}$ with stationary dpd P_θ. In Bayesian framework one is interested in computing the posterior distribution of $\theta \in \Theta$ given the data x^n and a prior distribution $\pi(\cdot)$ over Θ:

$$\pi(\theta|x^n) \propto \pi(\theta) \Pr\{X^n = x^n|\theta\}$$

where $\Pr\{X^n = x^n|\theta\} = p_\theta(x_1) \prod_{t=1}^{n-1} P_\theta(x_{t+1}, x_t)/p_\theta(x_t)$.

In many statistical applications (e.g. network analysis, epidemiological or genetic models) this probability is analytically intractable or computationally demanding to evaluate. In such cases one should resort to ABC whose key idea is to provide a conversion of samples from the prior distribution into samples from the posterior by rejecting those parameters that, given as input to the *simulator*, produce simulated observations, y^n, different from the observed data. Rejection ABC (R-ABC) displayed in Algorithm 1 produces samples form an *approximate posterior distribution* introducing three sources of approximation by 1) resorting to an arbitrary distance function $d(\cdot,\cdot)$; 2) introducing a positive tolerance parameter ε; 3) summarizing the observed and the simulated data through summary statistics $s_x = s(x^n)$ and $s_y = s(y^n)$ with $s : \mathbb{A}^n \to \mathscr{S}$. The output of the algorithm is a sample of pairs $(\theta^{(s)}, s_y^{(s)})$ from the following ABC joint posterior distribution

$$\tilde{\pi}(\theta, s_y|s_x) \propto \pi(\theta) \Pr\{s_Y = s_y|\theta\} \mathbb{1}\{d(s_y, s_x) \le \varepsilon\} \tag{2}$$

which, marginalising out s_y, i.e., simply discarding the simulates summaries, leads to the *marginal approximate posterior distribution*:

Algorithm 1 Rejection ABC (R-ABC)

for $s = 1,...,S$ **do**
  Draw $\theta^{(s)} \sim \pi$
  Generate $y \sim P(\cdot|\theta^{(s)})$ from the simulator
  Accept the pair $(\theta^{(s)}, s_y^{(s)})$ if $d(s_y^{(s)}, s_x) \leq \varepsilon$
end for

$$\tilde{\pi}(\theta|s_x) \propto \pi(\theta) \sum_{\mathscr{S}} \Pr\{s_Y = s_y|\theta\} \mathbb{1}\{d(s_y, s_x) \leq \varepsilon\} ds_y = \pi(\theta) \cdot \Pr\{d(s_y, s_x) \leq \varepsilon|\theta\}. \quad (3)$$

The indicator function in (3) does not enable to discriminate between pseudo-data equal to simulated data and pseudo-data just close enough. Thus, it is often replaced by a kernel function, which is a positive function of the distance $d(s_y, s_x)$, defined on a compact support and decaying continuously from 1 to 0.

Looking at (3), it is apparent that the probability $\Pr\{d(s_y, s_x) \leq \varepsilon|\theta\}$ represents the *approximate likelihood*. At each iteration s, $\Pr\{d(s_y, s_x) \leq \varepsilon|\theta\}$ is approximated pointwise by the indicator function or another kernel function defined on a compact support. This crude approximation causes a very large number of rejections leading to one of the major drawbacks of the ABC methods: the sample degeneracy (see [10, Ch. 4] for a discussion of the problem of sample degeneracy in ABC). This typically implies that ABC sampling schemes require a very large number of iterations to get a good approximation of the posterior distribution, especially in the tail area where $\Pr\{d(s_y, s_x) \leq \varepsilon|\theta\}$ is exponentially small. Here, we speculate that an improvement can be achieved employing a kernel function based on LDT, thus taking into account the exponential decay of $\Pr\{d(s_y, s_x) \leq \varepsilon|\theta\}$.

4 LD-ABC for Markov Chains

Let us consider the set $\Gamma_\varepsilon \stackrel{\triangle}{=} \{P \in \Delta^{k^2-1} : D_c(P||T_x^{(2)}) \leq \varepsilon\}$. Letting $y^m = y_1,...,y_m$ be a sample path from a pMC with dpd P_θ, from Theorem 1 follows that

$$\Pr\{T_{y^m}^{(2)} \in \Gamma_\varepsilon|\theta\} \approx 2^{-mD_c(\Gamma_\varepsilon||P_\theta)} \cdot c. \quad (4)$$

The following Theorem proves that $D_c(\Gamma_\varepsilon||T_{y^m}^{(2)}) \approx D_c(\Gamma_\varepsilon||P_\theta)$, as $m \to \infty$.

Theorem 2. *Let $\{Y_t\}$ be a Markov process taking values in the finite set $\mathbb{A}$ whose stationary dpd is $P_\theta \in \mathscr{M}(\mathbb{A}^2)$ and let $Y^m = Y_1,...,Y_m$. Then, under the measure induced by P_θ*

$$\lim_{m\to\infty} D_c(\Gamma_\varepsilon||T_{y^m}^{(2)}) = D_c(\Gamma_\varepsilon||P_\theta) \quad a.s. \quad (5)$$

Proof. See [11, Appendix D].

From (4) and Th. 2 follows that by setting the 2nd order type as summary statistics and the conditional relative entropy as divergence measure, the probability $\Pr\{d(s_y, s_x) \leq \varepsilon|\theta\}$ can be e approximated by $2^{-mD_c(\Gamma_\varepsilon||P_\theta)}$. Meaning that, the indicator function in (2) may be replaced by the following kernel:

$$K_\varepsilon(T_{y^m}^{(2)}) \stackrel{\triangle}{=} \begin{cases} 1 & \text{if } D_c(T_{y^m}^{(2)}||T_{x^n}^{(2)}) \leq \varepsilon \\ 2^{-mD_c(\Gamma_\varepsilon||T_{y^m}^{(2)})} & \text{if } D_c(T_{y^m}^{(2)}||T_{x^n}^{(2)}) > \varepsilon \end{cases}. \quad (6)$$

Hence, the joint and the marginal ABC approximate posterior distributions become:

$$\tilde{\pi}(\theta, T_{y^m}^{(2)} | T_{x^n}^{(2)}) \propto \pi(\theta) P_\theta(T_y^{(2)}) K_\varepsilon(T_{y^m}^{(2)}) \tag{7}$$

$$\tilde{\pi}(\theta | T_{x^n}^{(2)}) \propto \pi(\theta) \sum_{T_{y^m}^{(2)} \in \mathscr{T}(m,2)} P_\theta(T_{y^m}^{(2)}) K_\varepsilon(T_{y^m}^{(2)}) \tag{8}$$

where $\mathscr{T}(m,2)$ is the set of the 2nd order types of sequences of length m from Markov processes taking values in $\mathbb{A}$.

In order to sample from (7) we present both an Importance Sampling (IS) and a MCMC scheme displayed in Alg.2 and Alg.3, respectively. Both the algorithms draw parameter values from a proposal distribution on the parametric space, $q(\cdot)$, and avoid implicit rejections involving the proposed kernel in the evaluation of the importance weights or of the acceptance ratio. We refer the reader to [10, Ch. 4] for the a description of the standard IS-ABC and MCMC-ABC algorithms.

Algorithm 2 LD-IS-ABC

for $s = 1, ..., S$ **do**
  Draw $\theta^{(s)} \sim q$
  Draw $y^{(s)} \sim P(\cdot|\theta^{(s)})$ and compute $T_{y^{(s)}}^{(2)}$
  if $D_c(T_{y^{(s)}}^{(2)} || T_x^{(2)}) \leq \varepsilon$ **then**
    Set $\omega_s = \dfrac{\pi(\theta^{(s)})}{q(\theta^{(s)})}$
  else
    $\omega_s = 2^{-nD(\Gamma_\varepsilon || T_{y^{(s)}}^{(2)})} \dfrac{\pi(\theta^{(s)})}{q(\theta^{(s)})}$
  end if
end for

Algorithm 3 LD-MCMC-ABC

for $s = 1, ..., S$ **do**
  Draw $\theta^* \sim q(\theta^{(s-1)}, \theta^*)$
  Draw $y^* \sim P(\cdot|\theta^*)$ and compute $T_{y^*}^{(2)}$
  Draw $u \sim \text{Unif}[0,1]$
  if $u < \min\left\{1, \dfrac{\pi(\theta^*) K_\varepsilon(T_{y^*}^{(2)}) q(\theta^*, \theta^{(s-1)})}{\pi(\theta^{(s-1)}) K_\varepsilon(T_{y^{(s-1)}}^{(2)}) q(\theta^{(s-1)}, \theta^*)}\right\}$ **then**
    Assign $(\theta^{(s)}, T_{y^{(s)}}^{(2)}) \leftarrow (\theta^*, T_{y^*}^{(2)})$
  else
    Assign $(\theta^{(s)}, T_{y^{(s)}}^{(2)}) \leftarrow (\theta^{(s-1)}, T_{y^{(s-1)}}^{(2)})$
  end if
end for

5 Toy example

We consider a time series $X^{60} = X_1, ..., X_{60}$ from an AR(1) process taking values in $\mathbb{A} = \{1,2,3\}$. Specifically, we consider the AR(1) process dealt with in [1], where

$$X_t = \begin{cases} X_{t-1} & \text{with probability } \lambda \\ \delta_t & \text{with probability } 1-\lambda \end{cases}$$

with mixing weight $\lambda \in [0,1]$. δ_t is a discrete random variable taking values in $\mathbb{A}$ with probabilities $\theta \stackrel{\triangle}{=} (\theta_1, \theta_2, \theta_3) \in \Delta^2$. Our aim is approximating the posterior distributions of the four parameters $\theta_1, \theta_2, \theta_3$ and λ. We assume that $(\theta_1, \theta_2, \theta_3)$ is a priori distributed as a $Dirichlet(1,1,1)$ and λ as a $Beta(1,1)$. In such a case, despite the complexity of the likelihood function, samples from the true posterior distributions can be obtained through the Importance Sampling scheme, here taken as a benchmark (see [1] for a detailed discussion of the likelihood evaluation and sampling schemes). We ran both R-ABC and LD-ABC with $m = 120$, $\varepsilon = 0.005$ and $S = 100,000$. Figure 1 shows the posterior distributions approximated by the two algorithms and Table 1 displays the $\widehat{MSE}$ and $\widehat{MISE}$, computed by averaging the squared errors for the posterior mean and the integrated squared errors over 100 reruns for both the algorithms. We can see that the LD-ABC outperforms the standard ABC both in terms of point estimates and of posterior distributions approximation. Finally, we evaluate the effects on sample degeneracy looking at the Effective Sample Size (ESS)

(see e.g. [7]): LD-ABC achieves an ESS of 4619 versus the 11 values accepted by R-ABC.

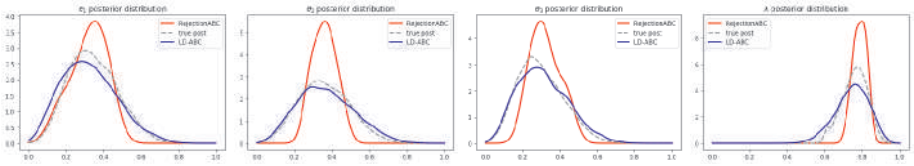

Fig. 1 Posterior distributions with $m = 120$ and $\varepsilon = 0.005$. The red lines are the posterior densities approximated via R-ABC and the blue lines via LD-ABC. The dashed grey lines are benchmarks obtained by IS.

Table 1 Squared errors and integrated squared errors averaged over 100 runs.

		$m = 120$, $\varepsilon = 0.005$			
		θ_1	θ_2	θ_3	λ
$\widehat{MSE}$	LD	$4.56 \cdot 10^{-4}$	$0.76 \cdot 10^{-4}$	$1.66 \cdot 10^{-4}$	$1.54 \cdot 10^{-4}$
	R	$13.59 \cdot 10^{-4}$	$16.35 \cdot 10^{-4}$	$8.99 \cdot 10^{-4}$	$6.63 \cdot 10^{-4}$
$\widehat{MISE}$	LD	0.0780	0.028	0.0274	0.1922
	R	0.2575	0.3162	0.3679	1.0681

References

1. Angers, J.F., Biswas, A. & Maiti, R. (2016). Bayesian Forecasting for time series of categorical data. The Journal of Forecasting, 36(3), 217-229.
2. Cover, T. M., & Thomas, J. A. (2006). Elements of information theory. John Wiley & Sons.
3. Csiszár, I. (1998). The method of types [information theory]. IEEE Transactions on Information Theory, 44(6), 2505-2523.
4. Daws, C. (2004). Symbolic and parametric model checking of discrete-time Markov chains. In International Colloquium on Theoretical Aspects of Computing (pp. 280-294). Springer.
5. Donsker,M. D., Varadhan,S. R. S. Asymptotic evaluation of certain Markov process expectations for large time I-III (1975-76). Comm. Proc. App. Math.,(28), l-47, 279-301.
6. Lanotte, R., Maggiolo-Schettini, A., & Troina, A. (2007). Parametric probabilistic transition systems for system design and analysis. Formal Aspects of Computing, 19(1), 93-109.
7. Liu JS (2008) Monte Carlo strategies in scientific computing. Springer Science & Business Media
8. Natarajan, S. (1985). Large deviations, hypotheses testing, and source coding for finite Markov chains. IEEE Transactions on Information Theory, 31(3), 360-365.
9. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular biology and evolution, 16(12), 1791-1798.
10. Sisson, S. A., Fan, Y., & Beaumont, M. (2018). Handbook of approximate Bayesian computation. Chapman and Hall\CRC.
11. Viscardi, C.(2021). Approximate Bayesian Computation and Statistical Applications to Anonymized Data: an Information Theoretic Perspective. PhD thesis.
12. Viscardi, C., Boreale, M. & Corradi, F. (2020). Improving ABC via Large Deviations Theory. Book of Short Papers SIS 2020, 673-678.
13. Viscardi, C., Boreale, M. & Corradi, F. (2021). Weighted Approximate Bayesian Computation via Sanov's Theorem . Computational Statistics, accepted for publication.