Hidden Markov Model estimation via Particle Gibbs

Stima di Hidden Markov Model tramite Particle Gibbs

Pierfrancesco Alaimo Di Loro, Enrico Ciminello and Luca Tardella

Abstract When the likelihood of the model is not explicitly available, standard Markov Chain Monte Carlo techniques may become impractical. This paper has the aim to investigate a combination of Sequential Monte Carlo and Metropolis-Hastings algorithm in the spirit of the pseudo-marginal approach. This produces algorithms known as Particle MCMC which are part of a powerful and flexible class of algorithms called Exact-Approximate MCMC. They establish a new paradigm in parameter estimation in the non-linear and non-Gaussian Hidden Markov Models (HMM). In this paper, the Particle Gibbs sampler is used to recover the parameter of an HMM applied to the time series of worldwide annual earthquakes of magnitude 7 or greater occurred in the 21st century.

Abstract *Quando la verosimiglianza non é trattabile analiticamente, allora le teniche Markov Chain Monte Carlo standard possono risultare inattuabili. Questo articolo ha l'obiettivo di analizzare una combinazione di tecniche Sequential Monte Carlo e dell'algorithmo Metropolis Hastings alla luce dell'approccio pseudo-marginale. Tale combinazione produce algoritmi Particle MCMC, parte di una classe molto flessibile nota come Exact-Approximate MCMC, che costituiscono un nuovo paradigma per la stima nell'ambito di Hidden Markov Model non lineari e non Gaussiani. In questo articolo é proposto l'utilizzo del Particle Gibbs per la stima di un modello HMM applicato alla serie annuale dei terremoti di magnitudo 7 o superiore avvenuti su scala mondiale durante il XXI secolo.*

Key words: MCMC, HMM, Bayesian, particle filters, time series, earthquake

Pierfrancesco Alaimo Di Loro
"La Sapienza" University of Rome, Piazzale Aldo Moro, 5, Rome, e-mail: pierfrancesco.alaimodiloro@uniroma1.it

Enrico Ciminello
"La Sapienza" University of Rome, Piazzale Aldo Moro, 5, Rome, e-mail: enrico.ciminello@uniroma1.it

Luca Tardella
"La Sapienza" University of Rome, Piazzale Aldo Moro, 5, Rome, e-mail: luca.tardella@uniroma1.it

Introduction

Interest in a process which can only be observed indirectly is a problem encountered in a variety of applications: biological sequences analysis [4], speech recognition [15] and time series analysis in general. Sometimes, the presence of a latent process depends on the theoretical framework, but often it is introduced for convenience.

Consider, for example, a Bayesian framework where the parameter of interest $\theta \in \Theta$ has a posterior density $p_y(\theta)$ that is not analytically available. The introduction of a latent variable $z \in \mathscr{Z}$ usually allows for an easier formulation and manipulation of the model. A typical estimation technique in such contexts is based on the *Gibbs Sampler* which samples alternatively from the conditionals $p_y(\theta|z)$ e $p_y(z|\theta)$. This sampling scheme can very often ease programming and lead to elegant algorithms. On the other hand, if $p_y(\theta)$ were known analytically or cheap to compute, it would often be possible to generate "more efficient" samples $\{\theta_i\}$ from a Markov chain by means of a classic *Metropolis-Hastings* (MH) algorithm. This led to the development of MCMC algorithms that try to combine possible computational efficiency of sampling directly from $p_y(\theta)$ and implementational ease of augmented schemes.

Often the likelihood, even if not analytically available, can be estimated in an unbiased way using Monte Carlo methods: this is the case of *Hidden Markov Models* (HMM, introduced in Section 1). *Sequential Monte Carlo* (SMC) allows to simulate from the unobservable process and get unbiased estimate of the needed density, paving the way to the application of approximated algorithms.

Section 2 focuses on *Particle MCMC (PMCMC)*, introduced in [2], which enables to perform parameter estimation for HMMs. These algorithms combine the ability of SMC to provide an unbiased estimate of the marginal likelihood of the process along with the pseudo-marginal approach [1], leading to a procedure that targets the true joint distribution of the parameters and the latent process.

In Section 3, a non-linear and non-Gaussian HMM is considered to model the annual time series of earthquakes of magnitude 7 or grater occurred worldwide during the 21st century. The Particle Gibbs sampler (PG), *Gibbs-style* version of the PMCMC, is used in order to recover the posterior distribution of the parameters and the resulting estimates are discussed in comparison to the ones obtained by [3] on the same data using the Particle Marginal Metropolis Hastings.

1 Hidden Markov Models

[6] describe *Hidden Markov Models* (HMM) as *the most successful statistical modeling ideas that have came up in the last forty years*. The partial bibliography by [5] gives a partial account of the wide scope of the domain: speech recognition [10], econometrics [12], computational biology [13], etc. The use of the hidden states makes such models generic enough to handle a variety of complex real-world time

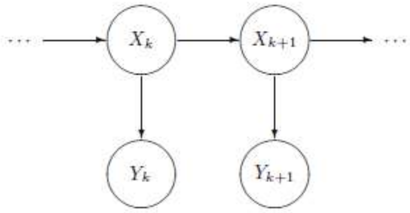

Fig. 1 *Hidden Markov Model* graph representation. Circles represent variables and arrows represent dependency relationships.

series, while the relatively simple formulation still allows for the use of efficient computational procedures.

An HMM is a process composed of a Markov Chain $\{X_t\}_{t=1}^{T} \subset \mathscr{X}^T$ with initial density $X_1 \sim p_\theta(\cdot)$ and transition probability $X_{t+1}|(X_t = x) \sim f_\theta(\cdot|x)$, where $\theta \in \Theta$ lives on a space of arbitrary dimension. This process $\{X_t\}$ is not observed directly, but through another stochastic process $\{Y_t\}_{t=1}^{T} \subset \mathscr{Y}^T$, whose observations are assumed to be independent conditionally on $X_1, X_2, \dots, X_T$ and to have conditional density $Y_t|\,(X_i = x_i)_{i=1}^{T} \sim p_\theta(\cdot|x_t)$. The dependence structure of an HMM can be represented by a graphical model as in Figure 1, where nodes (circles) in the graph correspond to the random variables, and the edges (arrows) represent the structure of the joint probability distribution.

When the space $\mathscr{X}$ of the hidden state X_t is discrete, the likelihood can be computed analytically while, if the space of X_t is continuous, it can be computed only when the models for X_t and Y_t are linear and Gaussian [11]. This stimulated the interest in alternative strategies that could be applied to more general frameworks. [9] proposed a first attempt of approximating the target distribution using a sequential version of the Monte Carlo importance sampling known as *Sequential Monte Carlo* (SMC). It is based on a recursive filtering approach, so that the received data can be processed sequentially rather then as a batch. Anyway, the choice of a suitable importance distribution in the form $q(x_{1:t}|y_{1:t})$ which is easy to sample from is not trivial. The procedure may be eased by the use of an auxiliary distribution that can be factored as follows:

$$q(x_{1:t}|y_{1:t}) = q(x_1|y_1)\prod_{i=1}^{t} q(x_i|x_{1:i-1}, y_{1:i}).$$

The sampling problem reduces to the one of recursively sample an arbitrarily large number N of *particles* (latent process samples) from univariate distributions of the form above. Approximations to the marginal likelihood and posterior densities of either parameters and latent states can be obtained from the set of the N weighted particles, where the weights are proportional to the condional likelihood $p_\theta(y_t|x_t)$.

2 Particle Markov Chain Monte-Carlo

PMCMC methods rely on a non-trivial and non-standard combination of MCMC and Sequential Monte Carlo methods, which takes advantage of the strength of both components. MCMC machineries requires either $p_\theta(\mathbf{y})$ to be analytically tractable or $p_\theta(z|y)$ to be sampled from. In the case of non-linear and non-gaussian HMMs, SMC may be used to produce likelihood estimates or to sample from $p_\theta(z|y)$ in order to yield flexible algorithms, which approximate the exact ones in the spirit of the pseudo-marginal approach [1]. PMCMC can be used to solve the same inferential problems of SMC and its extensions (IBIS by [7]; SMC^2 by [8]) but, in spite of its reliance on SMC methods, usually it is *much more robust and less likely to suffer the depletion problem* [2].

The Particle marginal Metropolis-Hastings

Recalling the standard decomposition for the posterior density $p(\theta, x_{1:T}|y_{1:T}) = p(\theta|y_{1:T})p_\theta(x_{1:T}|y_{1:T})$ it would be natural to suggest for the MH update a proposal of the form $q(\theta^*, x^*_{1:T}|\theta, x_{1:T}) = q(\theta^*|\theta)p_{\theta^*}(x^*_{1:T}|y_{1:T})$, for which the proposed $x^*_{1:T}$ is perfectly adapted to the proposed θ^*, that is sampled from an arbitrarily chosen proposal distribution $q(\cdot|\theta)$. The resulting MH acceptance ratio depends on the marginal likelihood $p_\theta(y_{1:T})$ that, in an HMM context, may be replaced by its SMC counterpart $\hat{p}_\theta(y_{1:T})$. This leads to the *Particle Marginal Metropolis Hastings* (PMMH).

Particle Marginal MH

1. **Initialization**. Choose $\theta^{(0)}$ and run the SMC. Compute $\hat{p}_{\theta^{(0)}}(y_{1:T})$ and sample $x^{(0)}_{1:T}$ from $\left\{x^{(i)}_{1:T}\right\}_{i=1}^{N}$ with weights $\left\{\omega^{(i)}_T\right\}_{i=1}^{N}$.
2. **Recursions**. For each $j \in \{1, \ldots, M\}$:
   a. propose θ^* from $q(\cdot|\theta^{(j-1)})$;
   b. run the SMC to get $\hat{p}_{\theta^*}(y_{1:T})$ and sample a new hidden path $x^*_{1:T}$;
   c. accept θ^* and $x^*_{1:T}$ in the chain $\left(\theta^{(j)}, x^{(j)}_{1:T}\right)$ with probability:

$$\alpha = 1 \wedge \frac{p\left(\theta^*\right)\hat{p}_{\theta^*}\left(y_{1:T}\right)q\left(\theta^{(j-1)}|\theta^*\right)}{p\left(\theta^{(j-1)}\right)\hat{p}_{\theta^{(j-1)}}\left(y_{1:T}\right)q\left(\theta^*|\theta^{(j-1)}\right)}.$$

   Set $\theta^{(j)} = \theta^{(j-1)}$ and $x^{(j)}_{1:T} = x^{(j-1)}_{1:T}$ otherwise.

The PMMH update leaves $p(\theta, x_{1:T}|y_{1:T})$ invariant and, under weak assumptions, ergodic [2]. The goodness of the resulting chain depends only on the more or less accurate choice of the proposal density. The problem is that such a choice is theoretically arbitrary, and there is a really low number of guidance on its appropriate formulation.

The Particle Gibbs

The *Particle Gibbs* consists of using a *Gibbs-style* update to sample iteratively $\theta \sim p(\theta|x_{1:T}, y_{1:T})$ and $X_{1:T} \sim p_\theta(x_{1:T}|y_{1:T})$. It is often possible to easily sample from $p(\theta|x_{1:T}, y_{1:T})$, and thus the issue of designing a proposal density for θ is encompassed. Sampling directly from $p_\theta(x_{1:T}|y_{1:T})$ is usually not feasible and, when replacing it by sampling from the SMC approximation $\hat{p}_\theta(x_{1:T}|y_{1:T})$, the convergence properties of the *Gibbs Sampler* (GS) to the target density $p_{x_{1:T}}(\theta|y_{1:T})$ do not hold anymore. A valid particle approximation to the GS requires the use of a special type of PMCMC update, called the *conditional SMC* update [2].

Particle Gibbs

1. **Initialization**. Set the arbitrary starting points $\theta^{(0)}$, $x_{1:T}^{(0)}$ and $B_{1:T}^{(0)}$.
2. **Recursions**. For $j \in \{1, \ldots, M\}$:

   a. sample $\theta^{(j)} \sim p(\cdot|x_{1:T}^{(j-1)}, y_{1:T})$;
   b. sample $x_{1:T}^{(j)}$ running a conditional SMC.

Stationarity and ergodicity of the resulting chain with respect to the target density are ensured in [2].

3 Applications

The earthquakes data used in [3] has been already analyzed by [14]. It describes the number of annual earthquakes with a magnitude of 7 or over (on the Richter scale) occurred worldwide along the 21st century. Usually this kind of data would be modeled using the Poisson distribution. However, the series is affected by over-dispersion and, furthermore, it presents significantly positive auto-correlations making unrealistic any hypothesis of independence. These features of the data motivated [14] to use a time-dependent parameter in the Poisson that may be resumed as follows:

$$\begin{cases} Y_t|(X_t = x_t) \sim Pois(\exp(\gamma + x_t)) \\ X_t|(X_{t-1} = x_{t-1}) \sim N(\phi x_{t-1}, \tau^2). \end{cases}$$

[3] applied the PMMH to get Bayesian estimates of the parameters: the results are exposed in Table 3. However, the principal drawback of the PMMH application is that the choice of the proposal for the random walk may play a key role in determining the quality of the final chains. [3] themselves were aware that better results could be achieved by a more careful choice of the proposal density, allowing for a better exploration of the parameters' space.

An alternative solution would be the application of a PG sampler in place of the PMMH as it does not require the design of any proposal distribution. The full conditionals of the parameters can be easily evaluated and they all turn out to belong to known parametric families. The obtained estimates are presented along with the

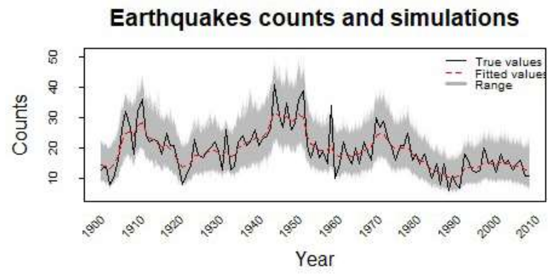

Fig. 2 Comparison of the observed counts with the PMCMC marginal predictive simulated trajectories and the average estimate.

Table 1 Estimates and ESS using PMMH [3] and PG.

Particle Marginal MH		
Parameters	Mean	ESS
ϕ	0.88	640
τ^2	0.02	683
γ	2.93	442

Particle Gibbs		
Parameters	Mean	ESS
ϕ	0.897	1232
τ^2	0.023	936
γ	2.783	555

effective sample sizes of the chains in Table 3. They are really close to the ones by [3] but outplay them in terms of *Effective Sample Size* (*ESS*). Finally the average trajectory, plotted against the observed counts, is presented in Figure 2 along with the whole range of estimated paths. This plot shows relatively good fit with the series of counts totally contained in the posterior predictive simulations range.

4 Concluding remarks

The aim of this work was to introduce the Particle MCMC methodology and to provide a comparison between the PMMH and the PG algorithm. In the analyzed application, if the random walk of the PMMH were implemented with an independent proposal, then the chains of the parameters would exhibit stickiness, leading to non satisfactory behaviours. [3] had to perform an accurate tuning of the covariance matrix of the proposal density in order to achieve well-behaved chains, without any guarantee that results could have been further improved.

The application of the PG spares the tuning and provides good results, characterized by fast convergence and good mixing of the chains. We may conclude that, whenever the model is simple enough and the correlation between parameters and latent states is not strong, the application of the PG is encouraged as its implementation is straightforward and estimation usually is more accurate.

References

[1] Christophe Andrieu and Gareth O Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, pages 697–725, 2009.

[2] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

[3] Jack Baker and Paul Fearnhead. Exact approximate Markov chain Monte Carlo. 2015.

[4] Ewan Birney. Hidden Markov models in biological sequence analysis. *IBM Journal of Research and Development*, 45(3.4):449–454, 2001.

[5] Olivier Cappé. Ten years of HMMs. *URL: www.tsi.enst.fr/~cappe/docs/hmmbib.html*, 2001.

[6] Olivier Cappé, Eric Moulines, and Tobias Rydén. Inference in hidden markov models. In *Proceedings of EUSFLAT Conference*, pages 14–16, 2009.

[7] Nicolas Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.

[8] Nicolas Chopin, Pierre E Jacob, and Omiros Papaspiliopoulos. SMC2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):397–426, 2013.

[9] JE Handschin. Monte Carlo techniques for prediction and filtering of non-linear stochastic processes. *Automatica*, 6(4):555–563, 1970.

[10] Frederick Jelinek. *Statistical methods for speech recognition*. MIT press, 1997.

[11] Rudolph Emil Kalman et al. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[12] Chang-Jin Kim, Charles R Nelson, et al. State-space models with regime switching: classical and Gibbs-sampling approaches with applications. *MIT Press Books*, 1, 1999.

[13] Timo Koski. *Hidden Markov models for bioinformatics*, volume 2. Springer Science & Business Media, 2001.

[14] Roland Langrock. Some applications of nonlinear and non-Gaussian state–space modelling by means of hidden Markov models. *Journal of Applied Statistics*, 38(12):2955–2970, 2011.

[15] Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.