

Sequential importance sampling algorithm for Bayesian nonparametric long range inference

*Algoritmi di tipo Monte Carlo sequenziale per inferenze non parametriche
relative a processi stazionari a memoria lunga*

Brunero Liseo¹ and Judith Rousseau²

¹ Dip. di studi geoeconomici, linguistici, statistici e storici per l'analisi regionale,

Università di Roma "La Sapienza"

e-mail: brunero.liseo@uniroma1.it

² CEREMADE, Université Paris Dauphine

Riassunto: Si propone un algoritmo Monte Carlo sequenziale per la stima bayesiana non parametrica della densità spettrale di un processo gaussiano con memoria lunga.

Keywords: population Monte Carlo, posterior consistency.

1. Introduction

Let $\{X_t, t = 1, 2, \dots\}$ be a stationary Gaussian random process, with mean $EX_t = \mu$ and covariance $\gamma(\tau) = E(X_t - \mu)(X_{t+\tau} - \mu)$. Let $f(\lambda)$ be the corresponding spectral density. A stationary Gaussian process is said to be long-range dependent, if there exist a positive number C and a value d ($0 < d < 1/2$) such that $\lim_{\lambda \rightarrow 0} C\lambda^{2d}f(\lambda) = 1$. Alternatively, one can define a long memory process as one such that its spectral density $f(\lambda)$ can be written as the product of a slowly varying function $\tilde{f}(\lambda)$ and the quantity λ^{-2d} which causes the presence of a pole of $f(\lambda)$ at the origin. Interest in long-range dependent time series has increased enormously over the last fifteen years; Beran (1994) provides a comprehensive introduction and the book edited by Doukhan et al. (2003) explores in depth both theoretical aspects and various applications of long-range dependence analysis in several different disciplines. In a few and simple words, long memory processes have been used to describe the persistent correlation structures observed in various fields such as hydrology, finance, astronomy. These processes exhibit significantly large correlations between measurements separated by long time intervals, hence the name long-memory processes. The memory of these processes is characterized by a single parameter, d , which can be interpreted as a (fractional) order of integration of a series. The statistical properties of a long memory process can be quite different from those of a set of

i.i.d. observations. For example, the familiar sampling properties of averages are far from valid in a long range setup. Parametric inference for long memory processes is usually performed via ARFIMA (p, d, q) models: these procedures are plagued by several problems: in particular there are no reliable model selection criteria for choosing the orders of autoregressive and moving average components; this is a serious drawback because, if p and q are misspecified, the maximum likelihood estimates of the parameter of interest d are asymptotically biased. Then semiparametric approaches have been considered in the last years; we refer to Doukhan et al. (2003) for a general overview from a frequentist perspective and to Liseo et al. (2001) for a Bayesian approach, where a local Whittle approximation to the likelihood was used. In this paper we describe the computational aspects of a novel Bayesian nonparametric approach proposed in Rousseau and Liseo (2005), based on the actual likelihood. In the next section we briefly recall the model and state the main results. In §3 we describe the algorithm, which belongs to the relatively new family of sequential importance sampling algorithms, where the choice of the proposal density is driven by the past iterations of the algorithm itself.

2. Model and main results

We consider the entire spectral density $f(\cdot)$ as our object of interest, and we assume that our prior belief on f are such that it can be written as $f(\lambda) = |1 - e^{i\lambda}|^{-2d} \tilde{f}(\lambda)$, where $\log \tilde{f}(x) = \sum_{j=0}^K \theta_j \cos(jx)$ is the short memory component; we assume that the true log spectral density satisfies some regularity conditions and $\log \tilde{f}_0(x) = \sum_{j=0}^{\infty} \theta_{0j} \cos(jx)$: in other words, it is equal to its Fourier series expansion. A prior on f can then be expressed as a prior on the parameters $(d, K, \theta_0, \dots, \theta_K)$ in the form $p(K)\pi(d|K)\pi(\theta|d, K)$, where $\theta = (\theta_0, \dots, \theta_K)$, and K represents the (random) truncation order of the Fourier expansion of $\log \tilde{f}_0(x)$. Also, d is set independent of θ for given K and it is also independent of K itself. Let $\pi(d) > 0$ and let K be a priori Poisson distributed; conditionally on K , assume that $\pi(\theta|K)$ puts mass 1 on the set of θ 's such that $\sum_{j=0}^K |\theta_j| \leq A$, for some value A large but finite. A possible way to formalize it is to assume that, for given K , the quantity $S_K = \sum_j |\theta_j|$ has finite support; then, setting $V_j = |\theta_j|/S_K$, $j = 1, \dots, K$, one may consider a distribution on the set $\{z \in \mathbb{R}^K; z = (z_1, \dots, z_K), \sum z_i = 1, z_i \geq 0\}$, for example: $(V_1, \dots, V_K) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K)$. Since the variance of the $|\theta_j|$'s should be decreasing, we may assume that, for all j 's, $\alpha_j = O((1+j)^{-2})$. Alternatives are possible here: for example one can assume that $(V_1, \dots, V_K)$ follows a logistic normal density, which allows for a more flexible elicitation. This prior is called a FEXP prior, after Hurvich et al (1994) where a similar construction is used in a classical framework. Under

the above conditions on the prior, Rousseau and Liseo (2005) prove that the posterior of f is strongly consistent in terms of the distance $d(f_0, f) = \min\{D(f_0, f), D(f, f_0)\}$, D being the Kullback-Leibler divergence. Also the Bayes estimator $\hat{f}$ is almost surely consistent and so is the Bayes estimator $\hat{d}$. Results on the rate of convergence are also provided. In the next section we describe the algorithm to produce Monte Carlo approximation of the posterior of the quantities of interest. The evaluation of the posterior is difficult because the likelihood associated with a sample from a long memory process takes the form of a multivariate normal kernel whose covariance matrix has a Toeplitz structure: this matrix is time consuming to invert.

3. The algorithm

The practical implementation of a nonparametric approach must take into account the fact that the computation of the likelihood in this context is very expensive since both the determinant and the inverse of a large Toeplitz matrix must be computed at each evaluation. Then one should prefer to use a Monte Carlo approximation which is as easy as possible to handle and with fast convergence properties. Our final preference was for a modification of the adaptive Monte Carlo algorithm, proposed and discussed in Douc et al. (2005), which can also be used in the presence of variable dimension parametric spaces as is the case here. The main advantage of adaptive Monte Carlo algorithms is that they do not rely upon asymptotic convergence of the samplers but, rather, they should be considered as an evolution of the importance sampling schemes, with the advantage of an on-line adaptation of the weights given to the proposals, which are allowed to be more than one. We now describe the main features of the proposed algorithm. Denote by (K, η_k) the global parameter, with $\eta_k = (\theta_0, \dots, \theta_k, d)$; here K is a positive integer which determines the size of the parameter vector. First one has to define a set of proposals, which we denote by $Q_h(\cdot, \cdot)$, $h = 1, \dots, H$; they represent H possible different kernels, which we assume are all dominated by the same dominating measure, as the posterior π_x . Denote by $q_h(\cdot, \cdot)$ the corresponding densities. Then one has to perform T different iterations of an importance sampling approximation, each based on N proposed values. The novel feature is that, at each t , the weights of the sampled values are calibrated in terms of the previous iterations. In practice one should work with a large value of N , and a small value (5 to 10) of T . Douc et al.(2005) show that, in few iterations, the algorithm will converge to the optimal mixture of proposals, at least in a Kullback-Leibler distance sense. Our algorithm follows quite closely the one described in Douc et al. (2005). The only significant difference is that we have to deal with the variable dimension of the

parameter space; then we must be able to propose a set of possible “moves” to subspaces with a different value of k . Then, at each iteration t , ($t = 1, \dots, T$) and for each sample point j , ($1 \leq j \leq N$), we propose a new value $K'_{j,t}$ from a distribution on the set of integers, conditional on the previous value of K_j ; then, conditionally on $K'_{j,t}$ and on the value of η_j^{t-1} , propose a new value $\eta_{j,K'_j}^{(t)}$. A full description of the possible moves is too long: we sketch the ideas behind the strategy. For a fixed $1 \leq j \leq N$, at the t -th iteration of the algorithm, one draws a new value $K'_{j,t}$, according the following proposals for K' :

$$K_{j,t} = K_{j,t-1} + \xi_{j,t}, \text{ where } \xi_{j,t} | K_{j,t-1} \sim p_1 \text{Po}(\lambda_1) + (1 - p_1) \text{NePo}_{K_{j,t-1}}(\lambda_2);$$

here $0 < p_1 < 1$ and NePo_k denotes a truncated Poisson over the set $\{-k, \dots, -1, 0\}$. At each iteration the proposed value $K_{j,t}$ may be either (i) less than, (ii) equal to, or (iii) larger than $K_{j,t-1}$; (i) if $K_{j,t} < K_{j,t-1}$ then $\theta_{K_{j,t}+1}^{(t)} = \dots = \theta_{K_{j,t-1}}^{(t)} = 0$ and

$$(\theta_0^{(t)}, \dots, \theta_{K_{j,t}}^{(t)}) = (\theta_0^{(t-1)}, \dots, \theta_{K_{j,t}}^{(t-1)}) + \varepsilon_{K_{j,t}} \quad (1)$$

and $\varepsilon_{K_{j,t}}$ is a $K_{j,t}$ -dimensional symmetric proposal. (ii) If $K_{j,t} = K_{j,t-1}$ then the new point is draw according to (1), without eliminating any parameter. (iii) If $K_{j,t} > K_{j,t-1}$, then the first $K_{j,t-1}$ are drawn according to (1), while the latest components are drawn from the same kernel proposal ε , although centered on an easy-to-calculate point estimate obtained from a simplified version of the estimator presented in Hurvich et al. (2002). Notice a feature of the algorithm which makes it different from the its main alternative, that is the reversible jump: even when the algorithm proposes a change of dimension, one “must” propose a global move for all the components of the parameter, in order to satisfy the positivity condition on the proposal required by Douc et al (2005).

References

- Beran J. (1994) *Statistics for long-memory processes*, Chapman and Hall, New York.
- Douc R., Guillin A., Marin J.M., Robert C.P. (2005) Convergence of adaptive sampling schemes, *Tech. Report, Ceremade, University Paris Dauphine*
- Doukhan P., Oppenheim G., Taqqu M.S. (2003) *Theory and applications of long-range dependence* Birkhäuser Boston Inc., Boston.
- Hurvich C.M., Moulines E., Soulier P. (2002) The FEXP estimator for potentially non-stationary linear time series, *Stoc. Proc. and their Appl.* , 97, 2, 307–340.
- Liseo B., Marinucci D., Petrella L. (2001) Bayesian semiparametric inference on long-range dependence, *Biometrika*, 88,4, 1089–1104.
- Rousseau J., Liseo B. (2005) Bayesian nonparametric estimation of the spectral density of a Gaussian time series (submitted).