Computational aspects of Bayesian spectral density estimation

Nicolas Chopin, Judith Rousseau, Brunero Liseo

COMPUTATIONAL ASPECTS OF BAYESIAN SPECTRAL DENSITY ESTIMATION

N. CHOPIN (CREST-ENSAE), J. ROUSSEAU (CREST-ENSAE AND UNIVERSITÉ PARIS DAUPHINE), AND B. LISEO (UNIVERSITA DI ROMA)

Abstract. Gaussian time-series models are often specified through their spectral density. Such models pose several computational challenges, in particular because of the non-sparse nature of the covariance matrix. We derive a fast approximation of the likelihood for such models. We use importance sampling to correct for the approximation error. We show that the variance of the importance sampling weights vanishes as the sample size goes to infinity. We show that the posterior is typically multi-modal, and derive a Sequential Monte Carlo sampler based on an annealing sequence in order to sample from the approximate posterior. Performance of the overall approach is evaluated on simulated and real datasets.

1. Introduction

Several models in the time series literature are defined through their spectral density. Assuming Gaussianity, one observes a vector $\boldsymbol{x}$ of length n, from the Gaussian distribution

$$\boldsymbol{x}|\mu, f \sim N\left(\mu\mathbf{1}, \boldsymbol{T}(f)\right)$$

where $\mathbf{1} = (1, \ldots, 1)'$, and $\boldsymbol{T}(f)$ is the $n \times n$ Toeplitz matrix associated to spectral density f, with entries $\boldsymbol{T}(f)(l, m) = \gamma_f(l - m)$,

$$\gamma_f(l) = \int_{-\pi}^{\pi} f(\lambda) e^{il\lambda}\, d\lambda, \quad l = -(n-1), \ldots, (n-1). \tag{1.1}$$

The models vary with respect to the specification of f. For instance, the FEXP parametrisation (Hurvich et al., 2002; Moulines and Soulier, 2003) assumes that

$$f(\lambda) = \frac{1}{2\pi} \left|1 - e^{-i\lambda}\right|^{-2d} g(\lambda), \quad g(\lambda) = \exp\left\{\sum_{j=0}^{k} \xi_j \cos(j\lambda)\right\}. \tag{1.2}$$

This specification conveniently separates the long-range behaviour, as determined by parameter $d \in [0, 1/2)$, and the short-memory part g. By taking k large enough, g may be arbitrarily close to any function in a certain regularity class. Rousseau et al. (2012) show that, for a well chosen prior with respect to $(k, d, \xi_1, \ldots)$, the corresponding posterior is consistent for both d and f, under semi-parametric settings; that is, assuming that the true spectral density belongs to a certain infinite-dimensional class of functions.

This FEXP parametrisation will be our running example. We note however that many other parametrisations of f are possible, which can also be tackled by the

Key words and phrases. FEXP, Long memory processes, Sequential Monte Carlo.

methodology developed in this paper. One may take $d = 0$, and obtain a non-parametric procedure for estimating the spectral density, under a short-memory assumption. One may replace g by another type of expansion, a spline regression, and so on. Finally, one may also consider a purely parametric model, such as an ARFIMA(p, q) model, given by:

$$f(\lambda) = \frac{1}{2\pi} \left|1 - e^{-i\lambda}\right|^{-2d} \left| \frac{1 + \sum_{j=1}^{q} \xi_j e^{-ij\lambda}}{1 - \sum_{j=1}^{p} \phi_j e^{-ij\lambda}} \right|^2 .$$

ARMA models may also be defined through their spectral density, but well-known specialised methods exist for such models, hence they fall outside the scope of this paper.

Whatever the specification of f, parametric or semiparametric, several computational difficulties arise regarding Bayesian inference for such models. First, the likelihood of the data

$$p(\boldsymbol{x}|\mu, f) = (2\pi)^{-n/2} \left|\boldsymbol{T}(f)\right|^{-1/2} \exp\left\{ -\frac{1}{2} \left(\boldsymbol{x} - \mu\mathbf{1}\right)^T \boldsymbol{T}(f)^{-1} \left(\boldsymbol{x} - \mu\mathbf{1}\right) \right\} \tag{1.3}$$

involves a determinant and a quadratic form which are expensive to compute, i.e. the cost is $O(n^3)$ if one uses off-the-shelf methods. Second, the entries of matrix $\boldsymbol{T}(f)$ itself, that is, the Fourier integrals (1.1) cannot be computed reliably using standard quadrature methods, because of the many oscillations of the integrand for large values of k. Third, Gibbs sampling is generally not feasible for posterior distributions associated to this likelihood. One usually resorts to the Metropolis-Hastings sampler (see e.g. Robert and Casella, 2004, Chap. 7), which is difficult to tune in order to obtain reasonable performance. This is quite problematic in this context: since the likelihood is expensive, performing several pilot runs in order to progressively tune the sampler may be a long and tedious process for the user. This problem is compounded if f is specified through a trans-dimensional prior; for instance, in the FEXP model above, if k is random. Then one needs to implement an algorithm for trans-dimensional sampling spaces, such as Green's algorithm (Green, 1995; Richardson and Green, 1997), also known as the reversible-jump sampler, which is harder to tune yet.

This paper proposes to address these problems in an unified manner, and is organised as follows. Section 2 discusses the exact computation the likelihood. It is seen that the cost of this operation is $O(n^3)$, but the constant in $O(n^3)$ is typically small, hence on modern hardware it remains possible to perform a reasonable number of such evaluations provided n is not too large.

Section 3 proposes a fast approximation of the likelihood, the cost of which is essentially $O(n)$. This approximation scheme motivates the following approach. In a first step, we perform Monte Carlo simulation of the approximate posterior, that is, the prior times the approximate likelihood. This step should cost $O(n)$, but typically with a large constant in front of n, due to the intensive nature of Monte Carlo algorithms. In a second step, but only if n is not too large, we correct for the approximation by doing importance sampling on a reasonable number of simulated samples; the cost of this second step is $O(n^3)$ but this time with a small constant in front of n^3. Thus, what we observe in practice is that the cost of this second step is in fact negligible with respect to the first step, at least for say $n \ll 10^4$. On the other hand, if n is large, we show that the importance sampling step becomes

superfluous, because the variance of the importance sampling weights converge to 0 as $n \to +\infty$ (under appropriate assumptions).

Section 4 discusses a Sequential Monte Carlo algorithm (SMC) for sampling from the approximate posterior; see Del Moral et al. (2006) for a general introduction to SMC. We shall see that the following advantages of SMC (relative to for instance MCMC) are particularly useful in this specific context. First, as explained above, one would like to apply the importance sampling step to a sample as small as possible, because the exact likelihood is expensive to compute. We shall see that our SMC sampler typically generates particles that are close to IID samples from its target distribution. Second, as also mentioned above, one has little prior information on the structure of the (approximate or exact) posterior. It is easy however to make SMC adaptive, by learning iteratively features of the target from the sample of simulated particles. Third, we observe in certain settings that semi-parametric models such as the FEXP model, may generate multimodal posteriors. The SMC sampler we propose is based on tempering ideas, and thus more able to escape from minor local modes.

Section 4 illustrates the proposed approach on simulated and real data.

We shall use the following notations: vectors and matrices are always in bold face, e.g. $\boldsymbol{x}$ and $\boldsymbol{\Sigma}$, the determinant, transpose, and trace of $\boldsymbol{\Sigma}$ are denoted respectively $|\boldsymbol{\Sigma}|$, $\boldsymbol{\Sigma}^T$ and $\mathrm{tr}(\boldsymbol{\Sigma})$.

2. Exact computation of the likelihood

2.1. **Marginalisation.** As a preliminary, we note that f is often parametrised in such a way that $f = \sigma^2 \bar{f}_{\boldsymbol{\theta}}$, where σ is a scale parameter, and $\boldsymbol{\theta}$ is the vector of all remaining parameters (except μ). For instance, in the FEXP case, we may set $\sigma^2 = \exp(\xi_0)$, $\boldsymbol{\theta} = (k, \boldsymbol{\theta}_k)$, $\boldsymbol{\theta}_k = (\mathrm{logit}(2d), \xi_1, \ldots, \xi_k)$, and

$$\bar{f}_{\boldsymbol{\theta}}(\lambda) = \frac{1}{2\pi} \left|1 - e^{-i\lambda}\right|^{-2d} \bar{g}_{\boldsymbol{\theta}}(\lambda), \quad \bar{g}_{\boldsymbol{\theta}}(\lambda) = \exp\left\{ \sum_{j=1}^{k} \xi_j \cos(j\lambda) \right\}. \tag{2.1}$$

(The function logit is defined as $\mathrm{logit}(x) = \log(x) - \log(1-x)$, and is used here to facilitate the construction of proposal distributions, see Section 4). It is then possible to marginalise out both μ and σ^2 from the likelihood, provided these two parameters are assigned the following standard $g-$prior distribution, independently from $\boldsymbol{\theta}$: $1/\sigma^2 \sim \mathrm{Gamma}(a, b)$, $\mu|\sigma^2 \sim N(m_\mu, \sigma^2/g_\mu)$. The marginal likelihood reads:

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\theta}) &= \int p(\boldsymbol{x}|\mu, \sigma^2, \boldsymbol{\theta}) p(\mu, \sigma^2)\, d\mu d\sigma^2 \\ &\propto \left| \boldsymbol{T}(\overline{f}_{\boldsymbol{\theta}}) + \frac{1}{g_\mu} \boldsymbol{E} \right|^{-1/2} \times \\ &\quad \left\{ b + \frac{1}{2} (\boldsymbol{x} - m_\mu \mathbf{1})^T \left(\boldsymbol{T}(\overline{f}_{\boldsymbol{\theta}}) + \frac{1}{g_\mu} \boldsymbol{E} \right)^{-1} (\boldsymbol{x} - m_\mu \mathbf{1}) \right\}^{-a-n/2}, \end{aligned} \tag{2.2}$$

where $\boldsymbol{E}$ is the $n \times n$ matrix filled with ones.

Marginalising out parameters usually improves the performance of the sampling algorithm. We note however that the approach developed in this paper would work with little modification for the unmarginalised likelihood $p(\boldsymbol{x}|\mu, \sigma^2, \boldsymbol{\theta})$. These

two likelihood functions suffer from the same computational difficulties, which are described in the next section.

2.2. **Computational difficulties associated to likelihood evaluation.** We review in this section the specific difficulties that arise when evaluating either the standard likelihood function (1.3), or the marginal version (2.2). First, both likelihood functions include some quadratic form $\boldsymbol{y}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$ involving the inverse of a $n \times n$ symmetric matrix $\boldsymbol{\Sigma}$; in (1.3), $\boldsymbol{\Sigma} = \boldsymbol{T}(f)$, and in (2.2), $\boldsymbol{\Sigma} = \boldsymbol{T}(\overline{f}_{\boldsymbol{\theta}}) + \frac{1}{g_\mu}\boldsymbol{E}$. Second, both functions involve the determinant of the same matrix $\boldsymbol{\Sigma}$. Third, in both cases, evaluating the entries of $\boldsymbol{\Sigma}$ requires computing simultaneously n Fourier integrals, see (1.1).

Our solution to the third difficulty is described in the two next Sections. Regarding the two first points, the most direct solution is to compute the Cholesky lower triangle of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}^{\boldsymbol{T}}$. Then, one obtains the determinant by taking the square of the product of the diagonal elements of $\boldsymbol{C}$, and one computes $\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} = \left(\boldsymbol{C}^{-1}\boldsymbol{y}\right)^T \boldsymbol{C}^{-1}\boldsymbol{y}$ as the norm of the solution (in $\boldsymbol{z}$) of the linear system $\boldsymbol{C}\boldsymbol{z} = \boldsymbol{y}$, which is quickly obtained by back-substitution. The Cholesky decomposition is a $O(n^3)$ operation.

For the sake of completeness, we mention briefly faster, but specialised, algorithms for solving directly the system $\boldsymbol{\Sigma}\boldsymbol{z} = \boldsymbol{y}$, in order to compute $\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$. Given that $\boldsymbol{\Sigma}$ is Toeplitz, one may use Levinson's algorithm (Levinson, 1949; Press et al., 2007, p. 96), which is $O(n^2)$. Alternatively, Chen et al. (2006) have developed a variant of the conjugate gradient method, based on a a particular preconditioned matrix. Their algorithm requires $O(\log^{3/2} n)$ iterations, each involving a FFT transform over $2n$ points. The overall cost is therefore $O(n \log^{5/2} n)$.

Unfortunately, these alternative approaches do not provide an evaluation of the determinant as a by-product. Chen et al. (2006), Holan et al. (2009) approximate the determinant by using a particular asymptotic approximation, which we describe later, but obviously this approach is not entirely satisfactory when used within a non-asymptotic, and in particular, a Bayesian, approach.

Perhaps more importantly, we note that the constant in front of the $O(n^3)$ cost of the Choleksy decomposition is typically very small, due to very efficient implementation in most scientific software. To give an order of magnitude, for $n = 10^3$, 1000 of such operations takes about one minute on the first author's computer. This observation underpins the strategy laid out in the introduction: to run some Monte Carlo algorithm so as to sample from an approximation of the posterior, then, provided n is not too large, to correct for the approximation using importance sampling on a moderate (possibly sub-sampled from the first step) Monte Carlo sample.

2.3. **Computing the Fourier coefficients.** In this section and the following, we consider the problem of evaluating simultaneously the n Fourier integrals defined in (1.1). As noted in the introduction, using standard quadrature would work very poorly, because the integrand in (1.1) strongly oscillates when n gets large. The solution we describe here seems well known in the numerical mathematics literature (e.g. Press et al., 2007, Chap. 13, Sect. 9), yet, to the best of our knowledge, it has not been used before in the time series literature. Instead, previous approaches (e.g. Chen et al., 2006; Holan et al., 2009) rely on more specific algorithms such as the splitting algorithm of Bertelli and Caporin (2002). The approach described

here has the same computational cost as such alternative approaches, that is, that of a FFT (Fast Fourier Transform), i.e. $O(n \log n)$. However, we find our approach slightly more convenient, for the following reasons: (a) this is a generic approach, which requires only pointwise evaluation of the spectral density, whereas the splitting algorithm requires exact expressions for the moving average coefficients (of the short memory part) which are specific to the considered class of spectral densities; for instance, it is unclear how one could use these methods if f would be specified through splines; (b) it is characterised by only one level of approximation (see below), whereas the splitting algorithm expresses first the moving averages coefficients as an infinite sum, which must be truncated, then plug these coefficients into another infinite sum, which must be truncated again, so assessing the numerical error is slightly more delicate; and (c) in long memory settings, the terms of these infinite sums are supposed to decay slowly, hence one may need to truncate to a large number of terms.

Let M a power of two, such that $M \geq 2n$. We explain first how to compute efficiently and simultaneously the Fourier integrals

$$\gamma_g(l) = \int_{-\pi}^{\pi} g(\lambda) e^{il\lambda} \, d\lambda$$

for a given bounded function g, and $0 \leq l \leq M/2$. (If $n < M/2$, simply discard the extra values.) If the spectral density f itself is bounded, that is, if it corresponds to a short memory process, then the following method may be used directly by taking $g = f$. If f corresponds to a long-memory process, then f diverges at 0, and a minor modification is required to use the following method, see next section.

The idea is to replace g by a linear interpolation $\widetilde{g}$:

$$\tilde{g}(\lambda) = \sum_{j=0}^{M-1} g_j \psi(\frac{\lambda - \lambda_j}{\Delta}) + g_0 \varphi_0(\frac{\lambda - \lambda_0}{\Delta}) + g_M \varphi_M(\frac{\lambda - \lambda_M}{\Delta})$$

where $\Delta = 2\pi/M$, $\lambda_j = -\pi + j\Delta$, $g_j = \tilde{g}(\lambda_j)$ $j = 0, \ldots, M$, and ψ is the linear interpolation kernel, i.e. $\psi(\lambda) = (1 - |\lambda|)^+$; φ_0 and φ_M are boundary corrections, the expression of which may be skipped for the rest of the discussion. Note $\tilde{g}$ is defined on the entire real-line, and is zero outside $[-\pi, \pi]$. Applying the operator $\int (\cdot) \, e^{il\lambda} \, d\lambda$ yields:

$$\begin{aligned}
\int_{-\pi}^{\pi} \tilde{g}(\lambda) e^{il\lambda} \, d\lambda &= \int_{-\infty}^{+\infty} \tilde{g}(\lambda) e^{il\lambda} \, d\lambda \\
&= \sum_{j=0}^{M-1} g_j \int_{-\infty}^{+\infty} \psi(\frac{\lambda - \lambda_j}{\Delta}) e^{il\lambda} \, d\lambda \\
&\quad + g_0 \int_{-\infty}^{+\infty} \varphi_0(\frac{\lambda - \lambda_0}{\Delta}) e^{il\lambda} \, d\lambda + g_M \int_{-\infty}^{+\infty} \varphi_M(\frac{\lambda - \lambda_M}{\Delta}) e^{il\lambda} \, d\lambda \\
&= \Delta(-1)^l \left\{ W(l\Delta) \sum_{j=0}^{M-1} g_j e^{ijl\frac{2\pi}{M}} + g_0 \alpha_0(l\Delta) + g_M \alpha_M(l\Delta) \right\}
\end{aligned}$$

where $W(\lambda) = \int_{-\infty}^{\infty} \psi(u) e^{i\lambda u} \, du$, $\alpha_0(\lambda) = \int_{-\infty}^{\infty} \varphi_0(u) e^{i\lambda u} \, du$, $\alpha_M(\lambda) = \int_{-\infty}^{\infty} \varphi_M(u - M) e^{i\lambda u} \, du$. The first sum may be computed using a FFT, in $O(M \log(M))$ time.

All the other functions admit close-form expressions, given by Press et al. (2007, Chap. 13, Sect. 9):

$$W(\lambda) = \frac{2(1-\cos\lambda)}{\lambda^2}, \quad \alpha_0 = \alpha_M = -\frac{W}{2}.$$

The scheme above may be adapted so as to rely on a cubic (rather than linear) interpolation. This point is interesting in settings where the spectral density is parametrised in terms of cubic splines. Then, the method described here becomes exact. Otherwise, the accuracy of the method is determined by the size of the grid, $M+1$. We follow Press et al. (2007) and takes M to be the smallest power of two such that $M \geq 2n$, and we observe in practice that it gives very accurate results (in the sense that larger value of M give sensibly the same values).

2.4. **Computing the Fourier coefficients when f diverges at 0.** In this Section, we explain how to adapt the method above for computing Fourier integrals, in the situations where f is a long-range dependent spectral density,

$$f(\lambda) = \frac{1}{2\pi}\left|1-e^{-i\lambda}\right|^{-2d} g(\lambda)$$

where g is a bounded function, and $0 < d < 1/2$. In this case, f diverges at 0, hence it may not be well approximated by a piecewise linear function. To address this problem, we simply decompose f in two terms:

$$2\pi f(\lambda) = \left|1-e^{-i\lambda}\right|^{-2d} g(0) + \left|1-e^{-i\lambda}\right|^{-2d} \{g(\lambda)-g(0)\},$$

and compute each Fourier integral as a sum of two Fourier integrals corresponding to each terms. The Fourier integrals of the first term correspond to the autocovariance function of a fractionally integrated noise, which admits a close-form expression (Brockwell and Davis, 2009, Chap. 13):

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|1-e^{-i\lambda}\right|^{-2d} e^{il\lambda}\,d\lambda = \begin{cases} \frac{\Gamma(l+d)\Gamma(1-d)}{\Gamma(l-d+1)\Gamma(d)} & \text{if } l \geq 1, \\ \frac{\Gamma(1-2d)}{\Gamma(1-d)^2} & \text{if } l = 0. \end{cases}$$

The Fourier integrals of the second term may be computed directly using the method of the previous section: assuming g is differentiable at 0, the second term vanishes at 0, since $\left|1-e^{-i\lambda}\right|^{-2d}\{g(\lambda)-g(0)\} \sim g'(0)\lambda^{1-2d}$, with $1-2d \geq 0$.

3. Approximated likelihood

3.1. **Principle.** For convenience, we consider only the standard likelihood function (1.3) in this section, and furthermore we assume that $\mu = 0$, i.e. a model with zero mean. The main idea of our approximation scheme is to replace $\boldsymbol{T}(f)^{-1}$ by $\boldsymbol{T}(1/4\pi^2 f)$ in the quadratic form, yielding

$$\tilde{p}(\boldsymbol{x}|f) = (2\pi)^{-n/2}\left|\boldsymbol{T}(f)\right|^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{T}(\frac{1}{4\pi^2 f})\boldsymbol{x}\right\}.$$

It is easy to see that the quadratic form within the exponential may now be computed in $O(n\log n)$ time. We return to this point and other implementation aspects in the next section.

The idea of approximating $\boldsymbol{T}(f)^{-1}$ by $\boldsymbol{T}(1/4\pi^2 f)$ is related to the asymptotic theory of Toeplitz matrices. It is in fact a common technical tool in the asymptotic theory of long-memory processes (Dahlhaus, 1989; Rousseau et al., 2012), but to the best of our knowledge has not been used for computational reasons before.

Now assume that f is parametrised in some way, $f = f_{\boldsymbol{\theta}}$, and denote $p(\boldsymbol{x}|f) = p(\boldsymbol{x}|\boldsymbol{\theta})$, $\tilde{p}(\boldsymbol{x}|f) = \tilde{p}(\boldsymbol{x}|\boldsymbol{\theta})$. As explained in the introduction, our strategy boils down to sample from the approximate posterior $\pi_n(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{\theta})\tilde{p}(\boldsymbol{x}|\boldsymbol{\theta})$, then to perform importance sampling from the approximate posterior to the true posterior, that is, to assign some weight to any simulation from the approximated posterior, with a weight function defined as:

$$w_{\text{Corr}}(\boldsymbol{\theta}) \triangleq \frac{p(\boldsymbol{x}|\boldsymbol{\theta})}{\tilde{p}(\boldsymbol{x}|\boldsymbol{\theta})} = \exp\left\{-\frac{1}{2}\boldsymbol{x}^T\left[\boldsymbol{T}(f_{\boldsymbol{\theta}})^{-1} - \boldsymbol{T}(\frac{1}{4\pi^2 f})\right]\boldsymbol{x}\right\}.$$

The following theorem justifies this strategy.

Theorem 1. *Consider the FEXP model, as defined by* (1.2)*, and let* π_n *denote the approximated posterior distribution defined as* $\pi_n(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta})\tilde{p}(\boldsymbol{x}|\boldsymbol{\theta})$*, where* $p(\boldsymbol{\theta})$ *is some prior density with respect to parameter* $\boldsymbol{\theta}$*, then, under certain conditions (given in the Appendix) on the prior distribution* $p(\boldsymbol{\theta})$*, and the true distribution of* $\boldsymbol{x}$*, with associated spectral density* f_o*, one has*

$$\mathbb{E}^{\pi_n}\left[w_{Corr}(\boldsymbol{\theta})\right] = w^0_{Corr}\left\{1 + o_P(1)\right\}, \quad \mathbb{E}^{\pi_n}\left[w_{Corr}(\boldsymbol{\theta})^2\right] = \left(w^0_{Corr}\right)^2\left\{1 + o_P(1)\right\}$$

where $w^0_{Corr} = p(\boldsymbol{x}|f_o)/\tilde{p}(\boldsymbol{x}|f_o)$.

The proof and the technical conditions on the prior and the true spectral density are given in the Appendix. Because this theorem relies heavily on technical results of Rousseau et al. (2012), it is restricted to the semi-parametric FEXP model presented in the introduction, but with zero mean. We believe it could be extended to other classes of models with some extra effort.

In practical terms, this theorem says that the variance of the weights goes to zero as n goes to infinity, or, in other words, that the importance weights become nearly constant as n goes to infinity.

3.2. **Practical implementation.** In this section, we work out a practical implementation of the approximation scheme proposed in the previous section. We now turn our attention to the marginalised likelihood defined in (2.2).

First, we simplify the quadratic form by ignoring the uncertainty with respect to μ, i.e. by taking $m_\mu = \bar{\boldsymbol{x}}$, $g_\mu = +\infty$, so that the likelihood simplifies to

$$\tilde{p}(\boldsymbol{x}|\boldsymbol{\theta}) \propto \left|\boldsymbol{T}(\overline{f}_{\boldsymbol{\theta}})\right|^{-1/2}\left\{b + \frac{1}{2}\widetilde{\boldsymbol{x}}^T\boldsymbol{T}(\overline{f}_{\boldsymbol{\theta}})^{-1}\widetilde{\boldsymbol{x}}\right\}^{-a+n/2}, \quad \widetilde{\boldsymbol{x}} = \boldsymbol{x} - \bar{x}\mathbf{1}.$$

We observe that this particular approximation is quite accurate in practice, which relates to the fact, in long-memory scenarios, $\overline{x}$ is the standard estimator of μ, and typically converges faster than other features (e.g. d) of the model.

Second, as explained in the previous section, we replace the inverse of $\boldsymbol{T}(\bar{f}_{\boldsymbol{\theta}})$ by $\boldsymbol{T}(4\pi^2/\bar{f}_{\boldsymbol{\theta}})$ which leads to the following approximation of the quadratic form:

$$\widetilde{\boldsymbol{x}}^T\boldsymbol{T}(\overline{f}_{\boldsymbol{\theta}})^{-1}\widetilde{\boldsymbol{x}} \approx \widetilde{\boldsymbol{x}}\boldsymbol{T}(\frac{4\pi^2}{\overline{f}_{\boldsymbol{\theta}}})\widetilde{\boldsymbol{x}} = \sum_{j=0}^{n-1} c_j(\tilde{\boldsymbol{x}})\gamma_h(j) \tag{3.1}$$

where the coefficients $c_k(\tilde{\boldsymbol{x}})$ may be pre-computed, once and for all, from the data (denoting $\tilde{x}_i = x_i - \bar{x}$ the components of $\tilde{\boldsymbol{x}}$):

$$c_j(\tilde{\boldsymbol{x}}) = \begin{cases} \sum_{i=1}^n \tilde{x}_i^2 & \text{if } j = 0 \\ 2\sum_{i=1}^{n-k} \tilde{x}_i\tilde{x}_{i+j} & \text{if } j = 1, \ldots, n-1 \end{cases}$$

and the $\gamma_h(j)$' are the coefficients of the Toeplitz matrix $\boldsymbol{T}(4\pi^2/\overline{f}_{\boldsymbol{\theta}})$, that is, the Fourier integrals corresponding to function $h = 1/\left(4\pi^2\bar{f}\right)$. Using the method described in Section 2.3, one obtains a $O(n \log n)$ overall cost for evaluating the quadratic form.

Third, we approximate the determinant using a $O(1)$ asymptotic approximation, as explained in the next section.

3.3. **Determinant approximation.** Approximations of determinants of the form $\left|\boldsymbol{T}(\bar{f}_{\boldsymbol{\theta}})\right|$ typically rely on asymptotic expansions. For instance, Whittle's approximation consists in replacing $\log\left|\boldsymbol{T}(\bar{f}_{\boldsymbol{\theta}})\right|/n$ by its limit,

$$\frac{1}{n}\log\left|\boldsymbol{T}(\bar{f}_{\boldsymbol{\theta}})\right| \to \int_{-\pi}^{\pi} \log\left\{2\pi\bar{f}_{\boldsymbol{\theta}}(\lambda)\right\}\, d\lambda.$$

Chen et al. (2006) use more refined asymptotic results on Toeplitz matrices (e.g. Böttcher and Silbermann, 1999, p. 177) to obtain a more accurate approximation. Using their approach, one obtains in the FEXP case

$$\log\left|\boldsymbol{T}(\bar{f})\right| \approx D_n(\overline{f}_{\boldsymbol{\theta}}) \triangleq d^2 \log n + \frac{1}{4}\sum_{j=1}^{k} j\xi_j^2 + d\sum_{j=1}^{k} j\xi_j + \log\frac{G(1-d)^2}{G(1-2d)}.$$

where G is Barnes' function (Adamchik, 2001). This approach is easily adapted to other time series models, such as ARFIMA; we refer to Chen et al. (2006) for more details and possible extensions to other classes of models.

3.4. **Further approximation.** The cost of the approximation described in the previous section is that of a FFT operation, that is $O(n \log n)$. This cost may be further reduced by remarking that the approximation of the quadratic form may be rewritten as (see e.g. Palma, 2007, Chap. 4):

$$\widetilde{\boldsymbol{x}}^T\boldsymbol{T}(\overline{f}_{\boldsymbol{\theta}})^{-1}\widetilde{\boldsymbol{x}} \approx \widetilde{\boldsymbol{x}}'\boldsymbol{T}(\frac{1}{4\pi^2\overline{f}_{\boldsymbol{\theta}}})\widetilde{\boldsymbol{x}} = \frac{n}{2\pi}\int_{-\pi}^{\pi}\frac{I(\lambda)}{\overline{f}_{\boldsymbol{\theta}}(\lambda)}\, d\lambda, \quad I(\lambda) = \left|\sum_{j=1}^{n}\tilde{x}_j e^{ij\lambda}\right|^2,$$

that is, $I(\lambda)$ is the periodogram of the (centred) dataset $\tilde{\boldsymbol{x}}$.

It is relatively easier to evaluate $I(\lambda)$ at the Fourier frequencies $\lambda_j = 2\pi j/n$, which suggests a further approximation, where this integral is replaced by a Riemann sum computed over the λ_j:

$$\frac{n}{2\pi}\int_{-\pi}^{\pi}\frac{I(\lambda)}{\overline{f}_{\boldsymbol{\theta}}(\lambda)}\, d\lambda \approx \sum_{j=1}^{n}\frac{I(\lambda_j)}{\overline{f}_{\boldsymbol{\theta}}(\lambda_j)}.$$

Computing simultaneously the $I(\lambda_j)$'s requires performing a FFT transform. The cost is $O(n \log n)$, but this needs to be done only once, for a given dataset. Then the approximate likelihood may be evaluated for many different values of $\boldsymbol{\theta}$, at a $O(n)$ cost. Since typically about $10^4 - 10^6$ such evaluations are performed when Monte Carlo sampling from the approximate posterior, one may for practical purposes ignore the pre-computation time, and consider this further approximation as a $O(n)$ operation.

To conclude, our final approximation takes the following form:

$$\tilde{p}(\boldsymbol{x}|\boldsymbol{\theta}) \propto D_n(\boldsymbol{\theta})^{-1/2} \left\{ b + \frac{1}{2} \sum_{j=1}^{n} \frac{I(\lambda_j)}{\overline{f}_{\boldsymbol{\theta}}(\lambda_j)} \right\}^{-a-n/2} . \tag{3.2}$$

This is very close in spirit to Whittle's approximation of the likelihood, which is based on the idea that the $I(\lambda_j)$'s are nearly independent, with variance $\overline{f}_{\boldsymbol{\theta}}(\lambda_j)$; see again e.g. Palma (2007, Chap. 4). Rigorously speaking, we have not been able to establish that this further approximation is valid in the sense defined in Section 3.1, that is, that the variance of an importance sampling step from the approximated to the true likelihood converges to zero. We merely observe empirically that this approximate likelihood is nearly indistinguishable numerically to the more principled approximation developed in the previous Sections, that is:

$$\tilde{p}(\boldsymbol{x}|\boldsymbol{\theta}) \propto D_n(\boldsymbol{\theta})^{-1/2} \left\{ b + \frac{1}{2} \widetilde{\boldsymbol{x}}' \boldsymbol{T} (\frac{1}{4\pi^2 \overline{f}_{\boldsymbol{\theta}}}) \widetilde{\boldsymbol{x}} \right\}^{-a+n/2} .$$

On the other hand, we also observe that the speed improvement brought by this further approximation is rather modest, which is in line with the respective theoretical costs of $O(n)$ and $O(n \log n)$. From now on, we do not distinguish between these two approximations. We only note that an additional advantage of (3.2) is that its gradient is easy to compute, which would make it possible to use Langevin-type MCMC moves.

4. Monte Carlo sampling

4.1. **Background.** This section discusses Monte Carlo sampling from the approximate posterior, that is,

$$\tilde{\pi}_n(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \tilde{p}(\boldsymbol{x}|\boldsymbol{\theta})$$

where $p(\boldsymbol{\theta})$ is some prior density defined with respect to parameter $\boldsymbol{\theta}$, and $\tilde{p}(\boldsymbol{x}|\boldsymbol{\theta})$ is the approximate likelihood defined in Section 3. Although this discussion is, as before, not specific to the FEXP model, we shall assume for notational simplicity that the considered model is parametrised as follows:

$$\boldsymbol{\theta} = (k, \boldsymbol{\theta}_k) \in \cup_{i \in \mathbb{N}} \{i\} \times \mathbb{R}^{i+1}.$$

For instance, in the FEXP case, we have seen that $\boldsymbol{\theta}_k = (\text{logit}(2d), \xi_1, \ldots, \xi_k)$. We shall also assume a nested structure for the $\boldsymbol{\theta}_{\boldsymbol{k}}'s$, that is, $\boldsymbol{\theta}_0 = (\theta_0)$, $\boldsymbol{\theta}_{k+1}^T = \left(\boldsymbol{\theta}_k^T, \theta_{k+1}\right)^T$. Again, in the FEXP case, $\theta_0 = \text{logit}(2d)$, $\theta_j = \xi_j$ for $j \geq 1$. Finally, we denote $p_k(\boldsymbol{\theta}_k) = p(\boldsymbol{\theta}_k|k)$, the prior of $\boldsymbol{\theta}_k$, conditional of k, $p_{k+1}(\theta_{k+1}|\boldsymbol{\theta}_{\boldsymbol{k}})$, the prior of component θ_{k+1}, conditional on the first k components of $\boldsymbol{\theta}_{k+1}$ being equal to those of vector $\boldsymbol{\theta}_k$, and $\tilde{p}(\boldsymbol{x}|k, \boldsymbol{\theta}_k)$ the approximate likelihood $\tilde{p}(\boldsymbol{x}|\boldsymbol{\theta})$ for $\boldsymbol{\theta} = (k, \boldsymbol{\theta}_k)$.

A common approach to this problem would be to alternate the steps described below as Algorithms 1 and 2, that is a random walk Metropolis step with respect to $\boldsymbol{\theta}_k$, conditional on k, and a birth-and-death step, that is, a particular instance of Green (1995)'s reversible jump sampler, which attempts at either incrementing (birth) or decrementing (death) k. In case a birth is proposed, the current vector $\boldsymbol{\theta}_k$ is completed with a draw from the conditional prior $p_{k+1}(\theta_{k+1}|\boldsymbol{\theta}_k)$. In case a death step is proposed, component θ_k is deleted from the current vector $\boldsymbol{\theta}_k$.

Of course, many variants of this basic MCMC strategy could be considered. However, such variants are unlikely to directly address the following two limitations

Algorithm 1 Gaussian random walk Metropolis step (conditional on k)

Input: $\boldsymbol{\theta} = (k, \boldsymbol{\theta}_k)$
Output: $\boldsymbol{\theta}' = (k, \boldsymbol{\theta}_k^{'})$
1. Sample $\boldsymbol{\theta}_k^{\star} \sim N(\boldsymbol{\theta}_k, \boldsymbol{\Sigma}_k)$.
2. With probability $1 \wedge r$, take $\boldsymbol{\theta}_k^{'} = \boldsymbol{\theta}_k^{\star}$, otherwise $\boldsymbol{\theta}_k^{'} = \boldsymbol{\theta}_k$, with

$$r = \frac{p_k(\boldsymbol{\theta}_k^{\star})\tilde{p}(\boldsymbol{x}|k, \boldsymbol{\theta}_k^{\star})}{p_k(\boldsymbol{\theta}_k)\tilde{p}(\boldsymbol{x}|k, \boldsymbol{\theta}_k)}.$$

Algorithm 2 Birth and death step

Input: $\boldsymbol{\theta} = (k, \boldsymbol{\theta}_k)$
Output: $\boldsymbol{\theta}' = (k', \boldsymbol{\theta}'_{k'})$
Constants: $\rho_{k\to k+1} = \rho_{k\to k-1} = 1/2$ for $k \geq 1$, $\rho_{0\to 1} = 1$, $\rho_{0\to -1} = 0$.
1. Let $\boldsymbol{\theta}^{\star} = (k^{\star}, \boldsymbol{\theta}^{\star}_{k^{\star}})$, where, with probability $\rho_{k\to k+1}$, $k^{\star} = k+1$, $\boldsymbol{\theta}^{\star}_{k^{\star}} = (\boldsymbol{\theta}_k, \theta_{k+1})$, and $\theta_{k+1} \sim p_{k+1}(\xi_{k+1}|\boldsymbol{\theta}_k)$ (birth step); otherwise, with probability $\rho_{k\to k-1} = 1 - \rho_{k\to k+1}$, $k^{\star} = k-1$, $\boldsymbol{\theta}^{\star}_{k^{\star}} = (\theta_0, \ldots, \theta_{k-1})$ (death step).
2. Set $\boldsymbol{\theta}' = \boldsymbol{\theta}^{\star}$ with probability $1 \wedge r$, otherwise $\boldsymbol{\theta}' = \boldsymbol{\theta}$, with

$$r = \frac{\rho_{k^{\star}\to k} p(k^{\star})\tilde{p}(\boldsymbol{x}|\boldsymbol{\theta}^{\star})}{\rho_{k\to k^{\star}} p(k)\tilde{p}(\boldsymbol{x}|\boldsymbol{\theta})}.$$

of standard MCMC: (a) calibration; and (b) local behaviour. Calibration refers to the difficulty of choosing certain tuning parameters, such as that the scales $\boldsymbol{\Sigma}_k$ in the random walk step. It is well known that the choice of such tuning parameters may have a critical impact on the performance of the algorithm. We shall see in our numerical study, Section (5), that tuning manually the $\boldsymbol{\Sigma}_k$'s is rather tedious. A more principled approach would be to use some form of adaptive MCMC sampling, see Andrieu and Thoms (2008) for a review, where one uses past samples to iteratively adapt the tuning parameters to the target distribution. However, adaptive MCMC is less straightforward to use in trans-dimensional settings (as an infinite number of tuning parameters $\boldsymbol{\Sigma}_k$ must be learnt), and also does not address the second difficulty, which we now comment upon.

The phrase "local behaviour" refers to the fact that MCMC samplers have difficulties exploring multi-modal posteriors, as they tend to get trapped in the attraction of meaningless modes. One may wonder if multi-modality is an issue for the classes of models considered in this paper.

In our simulations, we observed that the FEXP model may indeed generate multimodal posteriors, and we propose the following explanation. Figure 4.1 plots a spectral density of FEXP type, see (2.1), with $d = 0.4$, $k = 3$, and $(\xi_1, \xi_2, \xi_3) = (1, -1, 1)$. Overlaid are two other FEXP spectral densities, but with $k = 10$ in both cases, $d = 0.2$ (dotted line), $d = 0$ (dashed line), and coefficients ξ_j fit by least-squares on the Fourier frequencies $\pi j/100$, $j = 1, \ldots, 100$. These three densities are hard to distinguish except maybe in the close vicinity of 0. This indicates that, unless the sample size is very large, the posterior distribution may be multi-modal, with modes that may correspond to values of d which are very far from the true value, while k is set to a larger value so as to introduce additional components to compensate the bias in d.

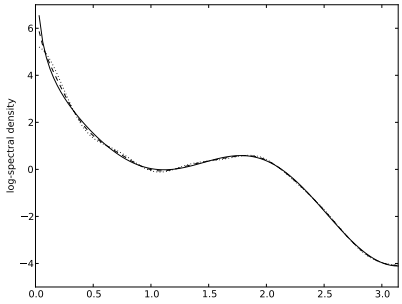

FIGURE 4.1. Graphical illustration of a source of multimodality for the FEXP model: FEXP log spectral densities, with $k = 3$, $d = 0.4$, $(\xi_1, \xi_2, \xi_3) = (1, -1, 1)$ (solid line), compared to $k = 10$, $d = 0$ (dashed line), and $d = 0.2$ (dotted line), and coefficients ξ_j adjusted by least-squares estimation.

Algorithm 3 Generic SMC sampler

All operations are for $j = 1, \ldots, N$.

(1) Sample $\boldsymbol{\theta}^j \sim \eta_0(\boldsymbol{\theta})$.
(2) For $t = 1, \ldots, T$, do:
   (a) Reweighing step: Compute and normalise weights
   $$w_t(\boldsymbol{\theta}^j) \propto \frac{\eta_t(\boldsymbol{\theta}^j)}{\eta_{t-1}(\boldsymbol{\theta}^j)}, \quad W_t^j = \frac{w_t(\boldsymbol{\theta}^j)}{\sum_{l=1}^N w_t(\boldsymbol{\theta}^l)}.$$
   (b) Resampling step: sample $\hat{\boldsymbol{\theta}}^j$ from the multinomial distribution that assigns probability W_t^l to value $\boldsymbol{\theta}_t^l$, $l = 1, \ldots, N$.
   (c) Move step: regenerate $\boldsymbol{\theta}^j$ as
   $$\boldsymbol{\theta}^j \sim K_t(\hat{\boldsymbol{\theta}}^j, d\boldsymbol{\theta})$$
   where K_t is a Markovian kernel that leaves η_t invariant.

4.2. **Sequential Monte Carlo.** Algorithm 3 describes a generic SMC sampler. Compared to the more general framework of Del Moral et al. (2006), this algorithm is specialised to the case where the sequence of distributions (η_t) is defined on a common sampling space, $\boldsymbol{\Theta}$, and also to a specific choice of the backward kernel, see Del Moral et al. (2006) for more details. The algorithm depends on the specification of a sequence of distributions (η_t) and a sequence of Markovian kernels (K_t). The former must be such that it is easy to sample from η_0, and that $\eta_T = \tilde{\pi}_n$, the target distribution (the approximate posterior, as defined in the previous section). The latter must be such that K_t leaves invariant η_t, and is typically a MCMC kernel.

For (η_t), we take a particular annealing sequence (Neal, 2001), that is, a geometric bridge between the prior and the approximate posterior

$$\eta_t(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \left\{\tilde{p}(\boldsymbol{x}|\boldsymbol{\theta})\right\}^{\gamma_t} \tag{4.1}$$

with $\gamma_0 = 0 < \ldots < \gamma_T = 1$. The weight function is then:

$$w_t(\boldsymbol{\theta}) = \tilde{p}(\boldsymbol{x}|\boldsymbol{\theta})^{\alpha_t}, \quad \alpha_t = (\gamma_t - \gamma_{t-1}).$$

We follow Jasra et al. (2011), Schäfer and Chopin (2011), and adjust dynamically the annealing coefficients γ_t by solving at iteration t, with respect to variable α_t, the equation

$$\frac{\left(\sum_{j=1}^N w_t(\boldsymbol{\theta}^j)\right)^2}{\sum_{j=1}^N w_t(\boldsymbol{\theta}^j)^2} = cN$$

for some fixed $c \in (0,1)$; in our simulations, we took the default value $c = 1/2$, and we used Brent's root-find algorithm (Press et al., 2007, Section 9.3) to solve numerically this equation. The left hand side is the effective sample size of Kong et al. (1994); it takes values in $[1, N]$, and is a convenient measure of the weight degeneracy.

For (K_t), we use M steps of the MCMC sampler described in the previous section, that is, we repeat M times the following sequence: Algorithm 1 (random walk Metropolis), then Algorithm 2 (birth-and-death). A big advantage of the SMC framework is that we can use the current particle system to calibrate these MCMC steps. Specifically, before each move step we set $\boldsymbol{\Sigma}_l = \tau_l \boldsymbol{S}_l$, $\tau_l = 2.38^2/(l+1)$, where $\boldsymbol{S}_l$ is the covariance matrix of those resampled particles $\hat{\boldsymbol{\theta}}^{(j)} = \left(\hat{k}^{(j)}, \boldsymbol{\theta}^{(j)}_{\hat{k}^{(j)}}\right)$ such that $\hat{k}^{(j)} = l$; in case this set is empty, we take instead $\boldsymbol{S}_l = \boldsymbol{I}_l$. This particular choice is motivated by the theory on optimal scaling of Hastings-Metropolis kernel, as developed in Roberts and Rosenthal (2001).

The only tuning parameters of this algorithm are the number of particles N, and the number of MCMC steps performed at each move step, M. In practice, taking M in the range $5 - 20$ seems to lead to reasonable performance, but of course, it may be adjusted more precisely by for instance looking at the acceptance rate of the MCMC steps in a preliminary run; see Section 5.

Again, one may propose many variants of this SMC sampler. For instance, one could also use adaptive reversible jump steps, where the jump probabilities $\rho_{k\to k+1}$, $\rho_{k\to k-1}$ (currently set to $1/2$) could be optimised using the particle sample, or even design more elaborate proposals based on a family of linear transforms, as in Green (2003), where the corresponding matrices may also be learnt from the particle sample. (We did some experiments with this strategy, but did not obtain significantly better results in the examples we looked at.) However, our main focus here is to develop a black box algorithm, which requires as little input from the user as possible, while giving reasonable performance, even in the presence of multimodality. Our numerical experiments seem to indicate this is the case, see Section 5.

5. Numerical experiments

5.1. **Settings.** The first part of this numerical study focus on the following simulated example. We sampled a long time series from an ARFIMA$(1, d, 1)$ model, with $d = 0.45$, and AR (resp. MA) coefficient -0.9 (resp. -0.2), and applied the FEXP model described in the Introduction. We find this example to be challenging, because the spectral density of the simulated process has a particular shape which is not easily approximated by a FEXP spectral density, at least for small values of k; see e.g. Fig. 5.4 and additional comments around this Figure. (From a modelling perspective, the presence of an autoregressive root close to one makes

it difficult to determine whether the persistence in the data is really characteristic of long range behaviour.) Thus, and as described more properly in Rousseau et al. (2012), the FEXP model described in the introduction must be understood as a semi-parametric procedure, which makes it possible to consistently estimate the true spectral density (and related quantities, e.g. the long-range coefficient d), provided this spectral density belongs to some infinite-dimensional class of functions (of a certain regularity). In particular, one expects the posterior of k to shift towards infinity as n goes to infinity, which adds to the computational difficulty. The second part of the study considers a real dataset, see Section 5.6.

Except in the consistency study (Section 5.4), the dataset consists of the $n = 10^4$ first values of the simulated time-series. In the consistency study, different sample sizes are compared, by again taking the n first points for different values of n. Except in the prior sensitivity section (Section 5.5), we always take the following prior: independently, $d \sim \text{Uniform}[0, 1/2]$, $k \sim \text{Geometric}(1/5)$ (with support $\{0, 1, 2, \ldots\}$), $\xi_j \sim N(0, 100j^{-2\beta})$, $\mu|\sigma \sim N(0, \sigma^2/g_\mu)$, $1/\sigma^2 \sim \text{Gamma}(a, b)$, with hyper-parameters $\beta = 1$, $a = b = 0.5$, $g_\mu = 0.1$. Note that this prior falls slightly outside the class of prior distributions that ensures consistency, according to Rousseau et al. (2012). In fact, we found it interesting to see whether the conditions to ensure consistency of Rousseau et al. (2012) may be too strong; our numerical study seems to indicate it may indeed be the case.

All the simulations presented in this Section were performed on a standard 3GHz desktop computer, without resort to any form of parallelisation, and implemented in the Python language (using the NumPy and SciPy libraries); the programs are available on the first author's web-page.

5.2. **Performance of SMC.** This section is concerned with the performance of the SMC sampler, regarding sampling the approximate posterior. Note that that the final correction step (i.e. the importance sampling step from the approximate to the true posterior) is therefore not performed; see Section 5.4 for an evaluation of the performance of the correction step.

We run the SMC sampler 10 times, for $N = 1000$ particles, and $M = 5$ MCMC steps; CPU time is 20 minutes per run. Variability over these ten runs is small for posterior expectations of the ξ_j's (not shown) and d (see top left plot of Figure 5.1), but is a bit larger for the posterior distribution of k (see same plot). The right plot of Figure 5.2 reveals that the acceptance rate of the birth and death step is is often below 10%, which suggests $M = 5$ steps may not be enough to successfully move the particles with respect to component k. In that plot, the $x-$axis gives the annealing coefficient; that is γ_t at iteration t, which goes from $\gamma_0 = 0$ to $\gamma_T = 1$, see (4.1). Note that this axis is a non-linear function of elapsed CPU time, as the sequence γ_t grows slowly at the beginning, and progressively accelerates, whereas the cost per iteration t should be roughly constant.

We run again the SMC sampler with $N = 1000$ particles, but this time with $M = 20$ steps; CPU time is then 1 hour 20 minutes. We obtain very satisfactory results; see the box-plots in Figure 5.1. In particular, regarding the posterior expectation $\hat{d}$ (with respect to the approximated posterior) of component d, we observe that the empirical variance of the estimates of $\hat{d}$ obtained from the ten runs is very close to $\text{Var}^{\tilde{\pi}_n}(d)$, that is, the variance would be obtained from $N = 1000$ independent and identically distributed simulations from the posterior. Figure 5.1 also reports superimposed weighted histograms obtained from the second set of runs, and a

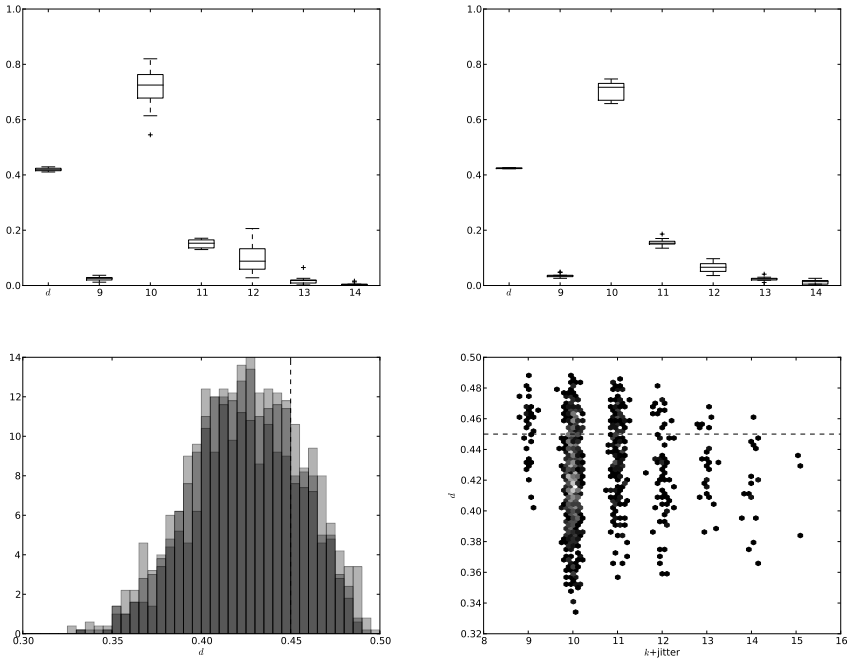

FIGURE 5.1. Top row: box-plots, over 10 repeated runs, of the SMC estimates of the posterior expectation of d and the posterior probabilities of $k = 9, \ldots, 14$ (Left: $(N, M) = (10^3, 5)$; Right: $(N, M) = (10^3, 20)$); Bottom Left: marginal posterior of d, as estimated by superimposed (with transparency effects) weighted histograms obtained by 3 first SMC runs; Bottom Right: hexagonal binning plot of k versus d, from the first SMC run; a $N(0, 0.1^2)$ jitter is added to k, and colour is proportional to the sum of the weights of the particles falling in each hexagon.

particle approximation of the bivariate posterior distribution of (k, d). (Some noise was added to k for the sake of visualisation.)

5.3. **Performance of MCMC.** This section compares the performance of our SMC sampler with standard MCMC. As the previous section, this comparison is in terms of sampling from the approximate posterior, and no correction step is performed at the end of the algorithm. The MCMC sampler we use is the one described in Section 4.1.

We simulate several MCMC chains with different tuning parameters, but always of size $N = 1.6 \times 10^5$, which gives a running time per chain of about 50 minutes (as compared to 20 minutes for the first sequence of SMC runs).

We start by 10 runs, with $\boldsymbol{\Sigma}_k$, the scale of the random walk in the $\boldsymbol{\theta}_k$-space, set to $\tau \boldsymbol{I}_{k+1}$ (τ times the identity matrix of rank $k+1$); after some pilot runs, we take $\tau = 0.015$ so as to obtain an acceptance rate for the random walk step close to 25%.

One chain seems to alternate between two local modes, one around $(k, d) = (20, 10^{-3})$, the other around $(k, d) = (30, 10^{-3})$; see first row of Figure 5.3. The

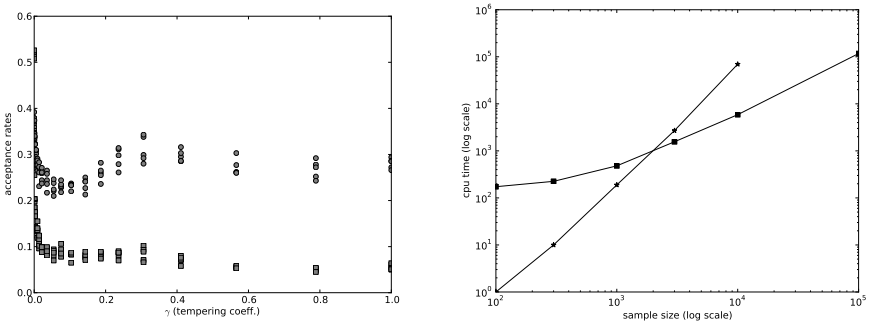

FIGURE 5.2. Left: acceptance rates of the random walk step (circles) and the birth and death step (squares) versus the tempering coefficient γ_t, which progresses from 0 to 1 during the algorithm ($N = 1000$, $M = 5$). Right: CPU times in seconds for the SMC sampler targeting the approximate posterior (squares) and the correction step (stars) versus the sample size of the data.

nine other chains do not get trapped in these local modes; however their mixing properties seem quite poor; see the MCMC traces for d the middle row plot. Note how the darkest trace never seems to visit the region $d > 0.45$, which accounts for 20% of the posterior mass, according to our SMC runs.

These results are hardly satisfactory. We also experimented with $\mathbf{\Sigma}_k$ set to τ times the prior covariance matrix of $\boldsymbol{\theta}_k$, but this did not give better results (not shown). Finally, we implemented a (crudely) adaptive version of our random walk Metropolis sampler, for the posterior conditional on $k = 10$ (no reversible jump steps), where, during a burn-in period of 5×10^4 iterations, the covariance matrix of the random walk proposal is re-calibrated every 10^3 iterations, using past samples; specifically $\mathbf{\Sigma}_k$ is set to $(2.38^2/(k+1))\boldsymbol{S}_t$ (see Roberts and Rosenthal, 2001), where $\boldsymbol{S}_t$ is the empirical covariance matrix computed on the t first iterations. One sees that learning the correlation structure of the posterior (conditional on k) significantly improve the mixing of the chain, although the auto-correlation function (with respect to d, and computed post burn-in) decays slowly, see bottom row of 5.3. However, learning the posterior covariance matrix, conditional on k, for each k in a given range, would of course be a bit wasteful.

5.4. **Consistency study.** In this section, we show how both the approximate and the true posterior evolve as the sample size grows. The goal is to assess both the asymptotic properties of our FEXP semi-parametric procedure, and the effect of the correction step on inference. Results are obtained from a single SMC run.

Figures 5.4 and 5.5 summarise posterior inference for the sample sizes, from top to bottom of Figure 5.4, $n = 100$, 300, 1000, 3000, and, from top to bottom of Figure 5.5, $n = 10^4$ and $n = 10^5$. On the left (resp. right) side, the marginal posterior distributions of d (resp. the 80% confidence bands for the spectral density) are compared; light grey is the approximate posterior, and dark grey is the true posterior. (Transparency effects are used.) One sees that the effect of the correction step becomes unnoticeable very quickly as far as the estimation of the spectral density (away from 0) is concerned. The effect on the marginal posterior of d takes

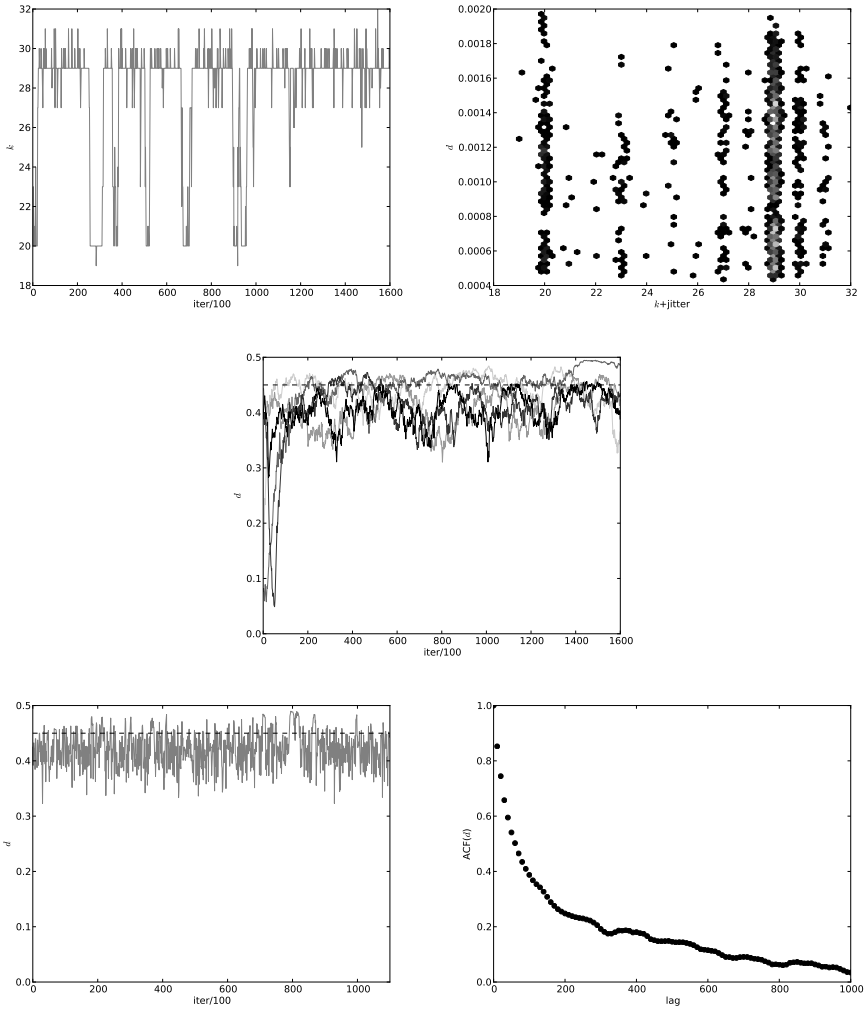

FIGURE 5.3. Top row: one MCMC run alternating between two minor modes (Left: trace of k, Right: scatter-plot of (k, d), with a jitter term added to k as in Fig.). Middle row: traces of d for five out of the nine remaining MCMC chains. Bottom row: adaptive MCMC; trace of d (left) and ACF plot (right), computed post burn-in.

a bit more time to vanish, but is already negligible for n around 3000; in fact, the effective sample size of the correction step is already above 900 (out of 1000 particles) for $n = 3000$.

The right panel of Figure 5.2 compares the CPU cost of the SMC sampler and the correction step. One sees that, unless n is of order of 10^4 or more, the overall cost of the complete procedure is dominated by that of the SMC sampler. For $n \gg 10^4$, the correction step becomes quickly too expensive, because of the $O(n^3)$ cost, and was not performed for $n = 10^5$. (Which is why the dark grey plots are

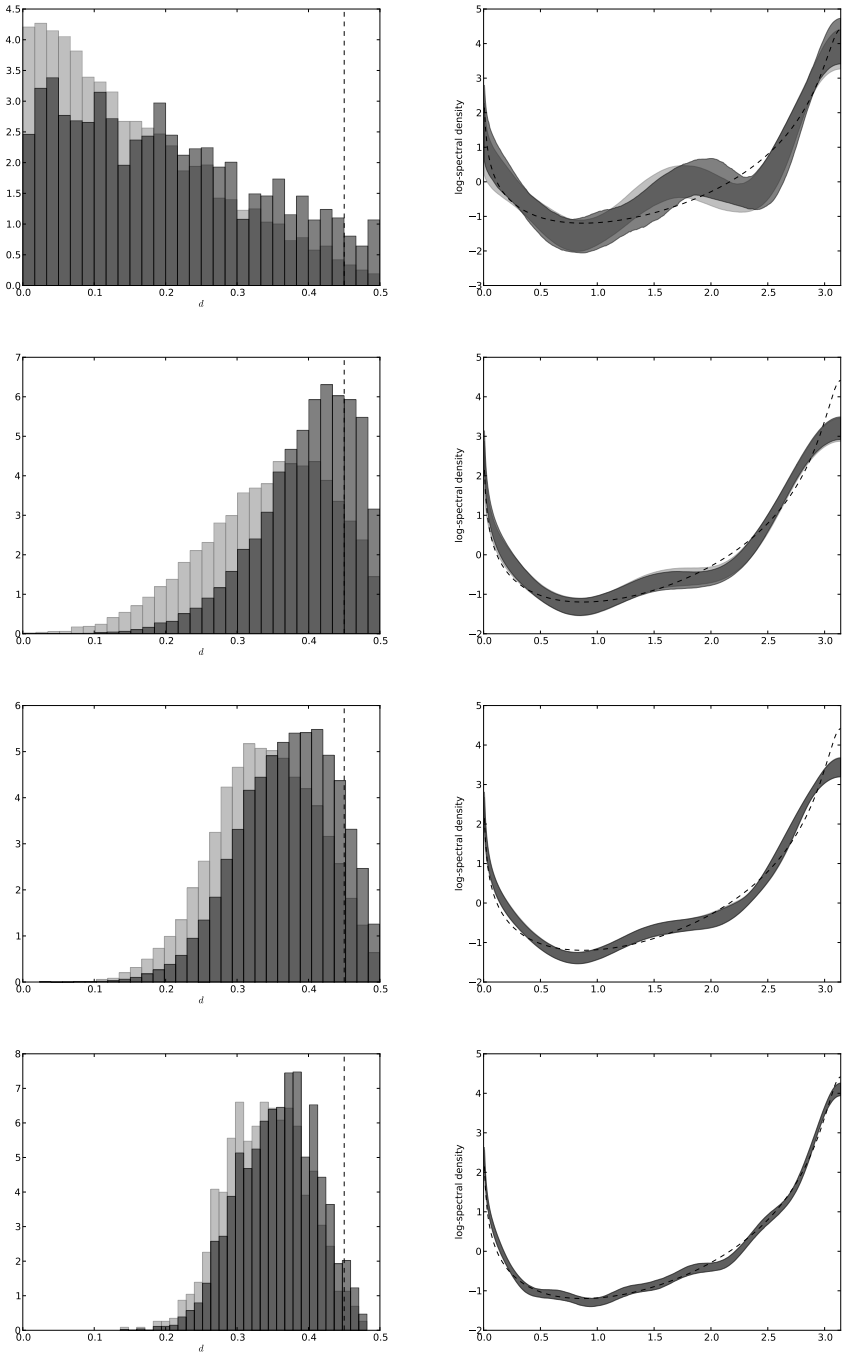

FIGURE 5.4. Consistency study for the ARFIMA data, Left: marginal posterior of d, approximate (light grey histogram), or exact (dark grey histogram); Right: 80% confidence bands for the log-spectral density (same color code), true spectral density (dashed line). From top to bottom, sample size is 100, 300, 1000, 3000.

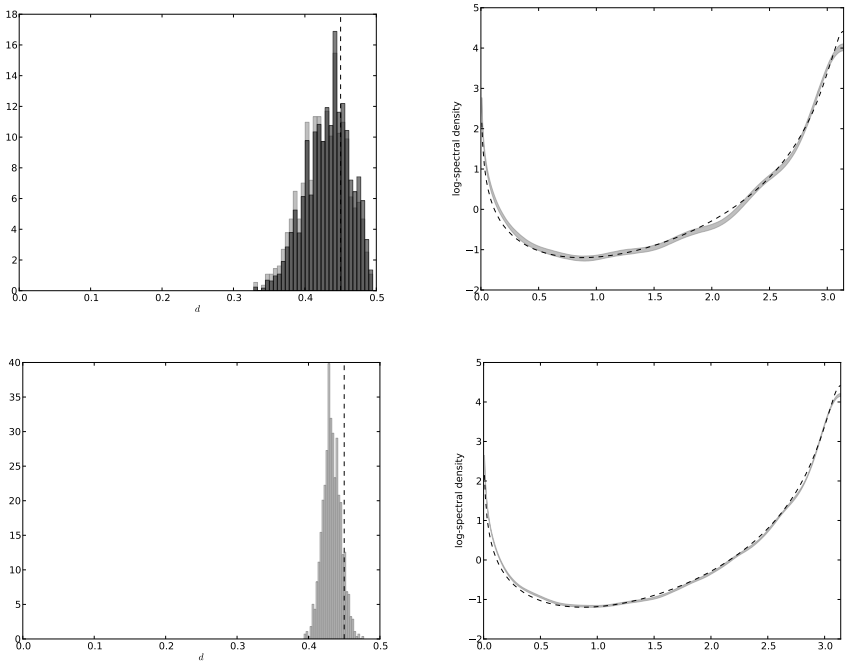

FIGURE 5.5. Consistency study for the ARFIMA data, same caption as Fig. 5.4, sample size is 10^4 (top) and 10^5 (bottom).

missing in the bottom row of Figure 5.5.) Fortunately, one sees that for $n \gg 10^4$ the correction step seems to be superfluous.

5.5. **Prior sensitivity analysis.** We have seen in Section 5.1 that the prior for the FEXP coefficients ξ_j was set to $N(0, 100j^{-2\beta})$, with $\beta = 1$; this hyper-parameter is related to assumptions on the smoothness of the true spectral density (the larger β, the smoother the true spectral density); see Rousseau et al. (2012). In this section, we consider briefly the effect of β on posterior inference.

Figure 5.6 gives the same types of posterior plots as in the previous sections, i.e. 80% confidence bands for the spectral density, and marginal posterior for the vector (k, d), for $\beta = 0$ (Left side) and $\beta = 2$ (right side). This must be compared with the top right plot of Figure 5.5 and the bottom right plot of 5.1, for which $\beta = 1$. One sees that the choice of β has a strong impact on the posterior marginal distribution of k, but a rather moderate impact on either d or the spectral density itself, except maybe on the right edge of the spectral density plots (for λ close to π). Since k is essentially a nuisance parameter, one sees that the choice of β does not seem to be too critical for inferential purposes. On the other hand, it is interesting to note the impact of β on the computational difficulty to explore the posterior. We observe that, for $\beta = 2$, it is even more difficult for a MCMC sampler to escape local modes such as those shown in the top row of Figure 5.3 (corresponding MCMC traces not shown).

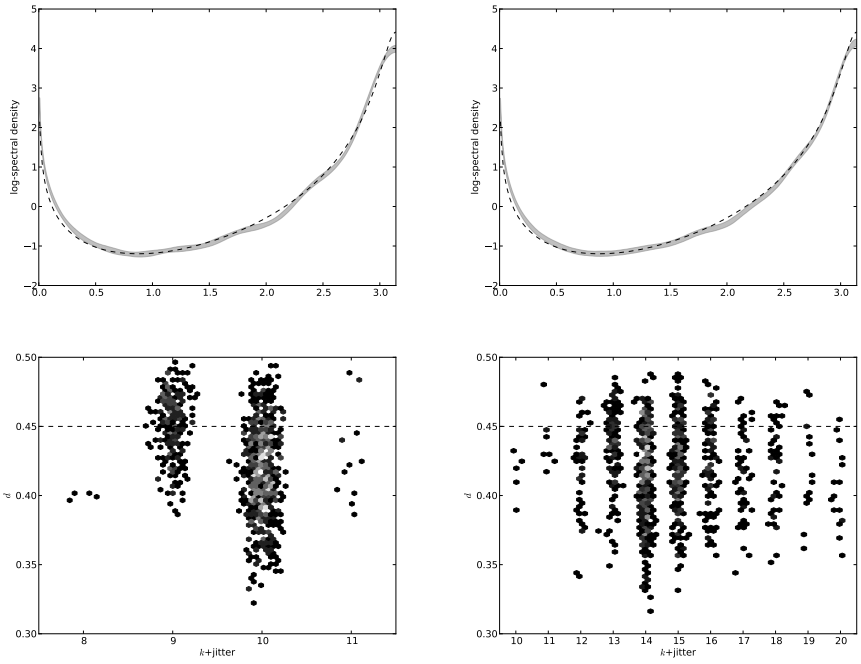

FIGURE 5.6. Prior sensitivity analysis: 80% posterior confidence bands for log-spectral density (top), hexagonal bin plot of $(k + \text{jitter}, d)$, where jitter is $N(0, 0.1^2)$ noise (bottom), for two different priors: $\beta = 0$, i.e. $\xi_j \sim N(0, 100)$ (left); and $\beta = 2$, i.e. $\xi_j \sim N(0, 100j^{-4})$ (right).

5.6. **Real data study: Ethernet Traffic.** In a preliminary study, we applied our methodology to the popular Nile data set, but found the example not to be very challenging, either from a computational point of view ($n = 663$), or from an inferential point of view (an ARFIMA$(0, d, 0)$ model, or in other words, a fractional Gaussian noise model, seems to fit the data well). More details may be obtained from the first author.

In this paper, we consider instead the Ethernet Traffic dataset of Leland et al. (1994), which can be found in the `longmemo R` package. This is a time-series of length $n = 4000$, which records the number of packets passing through a particular network per time unit. (For convenience, we divided the data values by 1000, in order to use the same prior as for the simulated dataset.) The right side of Fig. 5.7 plots the empirical spectrum of this time series. Although the empirical spectrum is an asymptotically biased estimator of the true spectral density under long-range dependence, the bias is typically small for frequencies sufficiently far away from 0 (e.g. Moulines and Soulier, 2003). Thus comparing the empirical spectrum with a given estimator of the spectral density provides at least some guidance on the performance of the said estimator.

Interestingly, we find these data to be quite challenging for frequentist parametric procedures. The `FEXPest` command of the `longmemo R` package, which computes

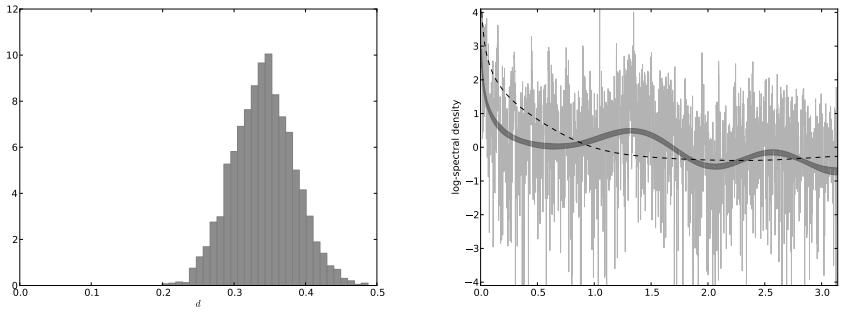

FIGURE 5.7. Ethernet data, Left: marginal posterior distribution of d from the FEXP semi-parametric model (as represented by a weighted histogram obtained from the SMC sampler); Right: Bayesian 80% confidence band for the spectral density (dark grey), spectral density corresponding to ML estimation of an ARFIMA model, with orders $p = q = 3$ (dashed line).

Beran (1993)'s estimator for a parametric FEXP model, returns the estimate $\hat{d} = 0.52$, although $d \in (0, 1/2)$. The `fracdiff` command from the `fracdiff` package, which does maximum likelihood estimation for ARFIMA(p, q) models, returns $\hat{d} = 0.3$ for $p = q = 3$ but the estimated spectral density, plotted as a dashed line in the right panel of Figure 5.7 does not seem to fit the spectrum of the data even for higher frequencies. (According to the same procedure, the orders $p = q = 3$ give the smallest AIC among ARFIMA(p, p) models, with $p \leq 10$).

In contrast, our Bayesian semi-parametric procedure seems to fit the data rather well; see again the right side of Figure 5.7. We find strong evidence in favour of long-range dependence, as evidenced by the marginal posterior distribution of d in the left panel of Figure 5.7.

From a computational point of view, we mention briefly that we observe the same things as in the previous section; that is, that the correction step has a negligible impact on the results, and that the output SMC sampler shows little variability when run several times.

ACKNOWLEDGEMENTS

N. Chopin and J. Rousseau are supported by the ANR grant "Bandhit" of the French Ministry of research.

APPENDIX: PROOF OF THEOREM 1

The proof borrows several intermediate results from Rousseau et al. (2012). Recall that the model is $\boldsymbol{x} \sim N(0, \boldsymbol{T}(f_{\boldsymbol{\theta}}))$, with $f_{\boldsymbol{\theta}}$ specified as in (1.2). We assume that the prior $p(\boldsymbol{\theta})$ decomposes as $p(d)p(k)p_k(\boldsymbol{\xi}_k)$, where the support of $p(d)$ is $[0, 1/2 - t]$ for some $t > 0$, $p(k)$ is a geometric distribution, and, given k, $\xi_j(j+1)^{\alpha+1/2} \sim \mathcal{N}(0, 1)$ independently except from the fact that it is truncated on

$$\Xi_k(\alpha, L) = \{\boldsymbol{\xi}_k \in \mathbb{R}^{k+1}; \sum_{j=0}^{k} (j+1)^{2\alpha+1} \xi_j^2 \leq L\}$$

where L is a large positive constant.

Let P_o^n be the true distribution of $\boldsymbol{x}$, that is, a Gaussian distribution $N(0 \times \mathbf{1}, \boldsymbol{T}(f_o))$, where the true spectral density f_o admits a FEXP representation, as in (1.2), with parameters $d_o > 0$, $k_o = +\infty$, and $\xi_{0,o}, \xi_{1,o}, \ldots$ belonging to the Sobolev ball of radius L, $\Xi(\beta, L) = \{(\xi_0, \xi_1, \ldots); \sum_{j=0}^{\infty}(j+1)^{2\beta+1}\xi_j^2 \leq L\}$. Throughout the proof we denote by $|A|_{Fr} = tr[A^T A]^{1/2}$ the Froebinius norm of a matrix A and by $\|A\|^2 = \sup\{x^T A^T A x; \|x\| = 1\}$ its euclidian norm.

We now prove that, provided $0 \leq d_o \leq 1/2 - t$, $\beta \geq \alpha > 3/2$, and for L_o small enough (compared to L), one has:

$$\mathbb{E}^{\pi_n}\left[w^2(\boldsymbol{\theta})\right] = w_o^2\left\{1 + o_p(1)\right\}, \quad w_o^2 = o_p(e^{u_n(\log n)^{3/2}})$$

for any sequence (u_n) such that $u_n \to +\infty$, and where w (resp. w_o) is used as a short-hand for w_{Corr} (resp. w_{Corr}^o) in the rest of the proof. One may obtain the same type of results for $\mathbb{E}^{\pi_n}[w(\boldsymbol{\theta})]$ using the same calculations.

One has:

$$\begin{aligned}\mathbb{E}^{\pi_n}\left[w^2(\boldsymbol{\theta})\right] &= \frac{\int w^2(\boldsymbol{\theta}) \exp\left\{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{T}(1/4\pi^2 f_{\boldsymbol{\theta}})\boldsymbol{x} - l_{n,o}\right\} |\mathbf{T}(f_{\boldsymbol{\theta}})|^{-1/2} p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int \exp\left\{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{T}(1/4\pi^2 f_{\boldsymbol{\theta}})\boldsymbol{x} - l_{n,o}\right\} |\mathbf{T}(f_{\boldsymbol{\theta}})|^{-1/2} p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \\ &\triangleq \frac{N_n}{D_n}\end{aligned}$$

where $l_{n,o} = -\left\{\boldsymbol{x}^T\mathbf{T}(f_o)^{-1}\boldsymbol{x} + \log\det\boldsymbol{T}(f_o)\right\}/2$.

Let $\epsilon_n^2 = (n/\log n)^{-\beta/(2\beta+1)}$. The idea of the proof is to first show that

$$\mathbb{E}^{\pi_n}\left[\frac{w^2(\boldsymbol{\theta})}{w_o^2}\right] = \mathbb{E}^{\pi_n}\left[\frac{w^2(\boldsymbol{\theta})}{w_o^2}\mathbb{1}_{S_{n,1}}(\boldsymbol{\theta})\mathbb{1}_{k\leq k_{n,1}}\right] + o_p(1) \tag{5.1}$$

where $S_{n,1} = \{\boldsymbol{\theta}; l(f_o, f_{\boldsymbol{\theta}}) \leq \epsilon_n\}$, $k_{n,1} = k_0\epsilon_n^{-1/\alpha}$ for some $k_0 > 0$, and $l(f, f')$ is the L^2 distance between spectral log-densities, $l(f, f') = \int_{-\pi}^{\pi}\{\log(f/f')\}^2$. Let $S_{n,1,k} = \{\boldsymbol{\theta}_k; \boldsymbol{\theta} = (k, \boldsymbol{\theta}_k) \in S_{n,1}\}$. In a second step, we show that

$$\sup_{\boldsymbol{\theta}\in S_{n,1,k}} |\log w(\boldsymbol{\theta}) - \log w_o| = o_p(1) \tag{5.2}$$

uniformly on $k \leq k_{n,1}$, which would then conclude the proof of Theorem 1.

We now prove (5.1). As proven in Rousseau et al. (2012) and Kruijer and Rousseau (2011) (for the true posterior, but the proof for the approximate posterior π_n follows the same lines), there exists $c > 0$ such that

$$P_o^n\left[D_n \geq e^{-cn\epsilon_n^2}\right] = o(1). \tag{5.3}$$

Indeed, let $k_n = \lfloor k_0(n/\log n)^{1/(2\beta+1)}\rfloor$ with $k_0 > 0$ and

$$\mathcal{B}_n = \left\{\boldsymbol{\theta};\ k = k_n, d_o \leq d \leq d_o + n^{-a}, |\xi_j - \xi_{j,o}| \leq n^{-a}, j = 1, \cdots k_n\right\},$$

then it is easy to see that, for some $c_1 > 0$,

$$\mathbb{P}^{p(\boldsymbol{\theta})}[\mathcal{B}_n] \geq e^{-c_1 k_n \log n},$$

where $\mathbb{P}^{p(\boldsymbol{\theta})}$ denotes the prior probability. As in Rousseau et al. (2012) and Kruijer and Rousseau (2011), let $\Omega_n(\boldsymbol{\theta}) = \{\boldsymbol{x}; \tilde{l}_n(\boldsymbol{\theta}) - l_{n,o} \geq -n\epsilon_n^2\}$, with

$$\tilde{l}_n(\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{x}^T\boldsymbol{T}(1/4\pi^2 f_{\boldsymbol{\theta}})\boldsymbol{x} - \frac{1}{2}\log\det\left[\boldsymbol{T}(f_{\boldsymbol{\theta}})\right]$$

then we have that for all $\boldsymbol{\theta} \in \mathcal{B}_n$,

$$P_o^n \left[\Omega_n(\boldsymbol{\theta})^c\right] = o(1) \tag{5.4}$$

if a is chosen large enough. Indeed let $A(\boldsymbol{\theta}) = \boldsymbol{T}^{1/2}(f_o)\boldsymbol{T}(1/4\pi^2 f_{\boldsymbol{\theta}})\boldsymbol{T}^{1/2}(f_o) - \boldsymbol{I}_n$ and $B(\boldsymbol{\theta}) = \boldsymbol{T}^{1/2}(f_o)\boldsymbol{T}(f_{\boldsymbol{\theta}})^{-1}\boldsymbol{T}^{1/2}(f_o) - \boldsymbol{I}_n$. Using Lemma 2.4 in the supplement of Kruijer and Rousseau (2011), for all $\boldsymbol{\theta} \in \mathcal{B}_n$, we have that $\mathrm{tr}\left[\boldsymbol{T}^{-1/2}(f_o)\boldsymbol{T}(f_{\boldsymbol{\theta}})\boldsymbol{T}^{-1/2}(f_o)\right] \leq 2n$ for n large enough so that all eigenvalues of $\boldsymbol{T}^{1/2}(f_o)\boldsymbol{T}(f_{\boldsymbol{\theta}})^{-1}\boldsymbol{T}^{1/2}(f_o)$ are bounded from below by $1/2$ and

$$\log|\boldsymbol{I}_n + B(\boldsymbol{\theta})| = \mathrm{tr}\left[B(\boldsymbol{\theta})\right] - \frac{1}{2}\mathrm{tr}\left[\left((\boldsymbol{I}_n + \tau B(\boldsymbol{\theta}))^{-1}B(\boldsymbol{\theta})\right)^2\right] \geq \mathrm{tr}\left[B(\boldsymbol{\theta})\right] - 2\mathrm{tr}\left[B(\boldsymbol{\theta})^2\right].$$

Moreover using Lemma 2.4 in the supplement of Kruijer and Rousseau (2011), we have that $\mathrm{tr}\left[B(\boldsymbol{\theta})\right] = \mathrm{tr}\left[A(\boldsymbol{\theta})\right] + O(n^\epsilon)$ for all $\epsilon > 0$ so that, since for n large enough and $\boldsymbol{\theta} \in \mathcal{B}_n$, $2\mathrm{tr}\left[B(\boldsymbol{\theta})^2\right] + n^\epsilon \leq n\epsilon_n^2$,

$$P_o^n \left[\Omega_n(\boldsymbol{\theta})^c\right] \leq \mathbb{P}_{\boldsymbol{z}\sim\mathcal{N}(0,\boldsymbol{I}_n)}\left[\boldsymbol{z}^T A(\boldsymbol{\theta})\boldsymbol{z} - \mathrm{tr}[A(\boldsymbol{\theta})] \geq n\epsilon_n^2\right].$$

One also has:

$$|A(\boldsymbol{\theta})|_{Fr}^2 = \frac{n}{2\pi}\int_{-\pi}^{\pi}\left(\frac{f_o(\lambda)}{f_{\boldsymbol{\theta}}(\lambda)} - 1\right)d\lambda + \text{Error},$$

where the Error is controlled by Lemma 2.6 in the supplement of Kruijer and Rousseau (2011), since we have

$$|A(\boldsymbol{\theta})|_{Fr}^2 = \mathrm{tr}\left[\left(\boldsymbol{T}(f_{\boldsymbol{\theta}}^{-1})\boldsymbol{T}(f_o - f_{\boldsymbol{\theta}})\right)^2\right]$$

and $f_o = f_{\boldsymbol{\theta}} b$ where (taking $\xi_j = 0$ for $j > k$)

$$b(\lambda) = (2 - 2\cos\lambda)^{d-d_o}\exp\left\{\sum_{j=0}^{\infty}(\xi_{o,j} - \xi_j)\cos j\lambda\right\}$$

and we note that, for $\boldsymbol{\theta} \in \mathcal{B}_n$, the function $\exp\left\{\sum_{j=0}^{k}\xi_j\cos j\lambda\right\}$ is Lipschitz with constant $O(k_n^{(3/2-\beta)_+}) = O(1)$, for a is large enough (see Lemma 3.1 in the supplement of Kruijer and Rousseau (2011)). Therefore we obtain

$$\text{Error} = O(\log n\|b - 1\|_\infty) = O(\log n),$$

and $|A(\boldsymbol{\theta})|_{Fr}^2 = O(n^{1-a} + \log n) = o(n^\tau n\epsilon_n^2)$, for some $\tau > 0$, which combined with Lemma 1.3 in the supplement of Kruijer and Rousseau (2011), leads to, for some $c > 0$,

$$P_o^n \left[\Omega_n(\boldsymbol{\theta})^c\right] \leq e^{-cn^\tau}.$$

We now turn to the second part of the proof. Let $S_{n,j,k} = \{\boldsymbol{\theta}_k; l(f_o, f_{(k,\boldsymbol{\theta}_k)}) \in ((j-1)\epsilon_n, j\epsilon_n)\}$, for $j = 1, \cdots, J_n$ where $J_n = O(\epsilon_n^{-1})$, and let

$$\begin{aligned} N_{n,j,k} &= w_o\int_{S_{n,j,k}} w(\boldsymbol{\theta})^2\exp\left\{\tilde{l}_n(\boldsymbol{\theta}) - l_{o,n}\right\}p(\boldsymbol{\theta})\,d\boldsymbol{\theta} \\ &= w_o\int_{S_{n,j,k}}\frac{\exp\left\{l_n(\boldsymbol{\theta}) - l_{o,n}\right\}}{w(\boldsymbol{\theta})}p(\boldsymbol{\theta})\,d\boldsymbol{\theta}. \end{aligned}$$

We prove that uniformly in $J_0 \leq j \leq J_n$ and $k \in \mathbb{N}$, $N_{n,j,k} = o_p(e^{-cn\epsilon_n^2})$, for some $J_0 > 0$. To do so we bound

$$P_o^n \left[\sup_{\boldsymbol{\theta} \in S_{n,j,k}} |\log(w_o/w(\boldsymbol{\theta}))| > v_{n,j,k} \right]$$

for $j \geq 2$ and for a properly chosen $v_{n,j,k}$. Let $k_{n,j} = k_0(j\epsilon_n)^{-1/\alpha}$, $J_{n,1} = J_1 n^{-\alpha\epsilon}(\log n)^{\alpha}\epsilon_n^{-1}$, with $J_1 > 0$. Define for $C > 0$ (large enough) and $\tau > 0$ (small enough)

(5.5)

$$v_{n,j,k} = k^{3/2} n^{\epsilon} j\epsilon_n, \quad \text{if } k \leq k_{n,j}, \quad j \geq 1$$

$$v_{n,j,k} = (C \log n(k \wedge n^{(1+\epsilon)/(\alpha-1/2)} k^{-1/(\alpha-1/2)}) + k^{\tau}) k_{n,j}^{-\alpha+1/2}, \quad \text{if } k > k_{n,j}.$$

Note that if ϵ is small, $Ck \log n = o(n^{\epsilon})$ for some $k \geq k_{n,j}$ only if $(j\epsilon_n)^{-1/\alpha} \lesssim n^{\epsilon}(\log n)^{-1}$, i.e. if $j \gtrsim n^{-\alpha\epsilon}(\log n)^{\alpha}\epsilon_n^{-1} := J_{n,1}$. We write $-\log w(\boldsymbol{\theta}) + \log(\xi_o) = z' A(\boldsymbol{\theta}) z/2$ with $\boldsymbol{z} = \boldsymbol{T}^{-1/2}(f_o)\boldsymbol{x} \sim \mathcal{N}(0, I_n)$, and, under P_o, and we bound successively

$$\sup_{\boldsymbol{\theta} \in S_{n,j,k}} \left(\boldsymbol{z}^T [A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)] \boldsymbol{z}^T + \mathrm{tr}[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)] \right)$$

and

$$\sup_d \left| \boldsymbol{z}^T [A_o - A(\boldsymbol{\theta}_d)] \boldsymbol{z} + \mathrm{tr}[A_0 - A(\bar{\boldsymbol{\theta}}_d)] \right|$$

with $\bar{\boldsymbol{\theta}}_d = (d, k, \bar{\boldsymbol{\xi}}_{d,k})$ and $\bar{\boldsymbol{\xi}}_{d,k} = \mathrm{argmin}_{\boldsymbol{\xi} \in \mathbb{R}^{k+1}} l(f_0, f_{d,k,\boldsymbol{\xi}})$, see Kruijer and Rousseau (2011). We now study

$$\sup_{\boldsymbol{\theta} \in S_{n,j,k}} \left(z' [A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)] z + \mathrm{tr}[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)] \right).$$

Note that $n^{-1}|A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)|_{Fr}^2 = n^{-1}\mathrm{tr}\left[(A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d))^2\right]$ converges towards 0, so we only need control the approximation error, which we split into the approximation error of $\mathrm{tr}\left[(\boldsymbol{T}(f_o)\boldsymbol{T}(f_{\boldsymbol{\theta}}^{-1} - f_{\bar{\boldsymbol{\theta}}_d}^{-1}))^2\right]$ and of $\mathrm{tr}\left[(\boldsymbol{T}(f_o)(\boldsymbol{T}(f_{\boldsymbol{\theta}})^{-1} - \boldsymbol{T}(f_{\bar{\boldsymbol{\theta}}_d})^{-1})^2)\right]$. We now consider the first term. Note that

$$f_{\boldsymbol{\theta}}^{-1} - f_{\bar{\boldsymbol{\theta}}_d}^{-1} = f_{\bar{\boldsymbol{\theta}}_d}^{-1}(\exp[\sum_{j=0}^{k} (\xi_j - (\bar{\xi}_{d,k})_j)\cos(j\lambda)] - 1) := f_{\bar{\boldsymbol{\theta}}_d}^{-1} b_{\boldsymbol{\theta}}(\lambda)$$

where

$$\sup_{\lambda \in [-\pi,\pi]} |b_{\boldsymbol{\theta}}(\lambda)| \leq \sum_{j=0}^{k} |\xi_j - (\bar{\xi}_{d,k})_j| \leq \sqrt{k}\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}_{d,k}\| \lesssim \sqrt{k} j\epsilon_n.$$

Note that if $k \geq k_{n,j}$ then $k^{1/2-\alpha} \leq \sqrt{k} j\epsilon_n$ and we bound instead

$$\sup_x |b_{\boldsymbol{\theta}}(x)| \leq \sum_{j=0}^{k} |\xi_j - (\bar{\xi}_{d,k})_j| \leq \sqrt{k_{n,j}}\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}_{d,k}\| + k_{n,j}^{-\alpha+1/2}.$$

Lemma 2.1 of the supplement of Kruijer and Rousseau (2011) implies that the approximation error of the first term is bounded by $O((\sum_{j=1}^{k} |\xi_j - (\bar{\xi}_{d,k})_j| n^{\epsilon})^2) = O((n^{\epsilon'}\sqrt{k} j\epsilon_n)^2)$ for all $\epsilon' > 0$ and all $k \leq k_{n,j}$ and is bounded by $O((k_{n,j}^{-\alpha+1/2} n^{\epsilon'})^2)$ for all $k \geq k_{n,j}$ and all $\epsilon' > 0$. From Lemma 2.4 of the supplement of Kruijer and

Rousseau (2011), the same bounds apply to the second term. Finally we obtain that

$$|A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)|_{Fr} = O(n^{\epsilon}\sqrt{k}j\epsilon_n) \quad \text{if } k \leq k_{n,j}$$
$$|A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)|_{Fr} = O(n^{\epsilon}k_{n,j}^{-\alpha+1/2}) \quad \text{if } k > k_{n,j}.$$

This implies that for all $\xi \in S_{n,j,k}$:

- If $k \leq k_{n,j}$, since $v_{n,j,k}(\sqrt{k}j\epsilon_n)^{-1} = kn^{\epsilon}$,
$$\mathbb{P}\left(\boldsymbol{z}^T[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)]\boldsymbol{z} + \mathrm{tr}[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)] > v_{n,j,k}\right) \leq e^{-ckn^{\epsilon}},$$
- If $k > k_{n,j}$, since $v_{n,j,k}k_{n,j}^{\alpha-1/2} \geq n^{\epsilon}$,
$$Pr\left(\boldsymbol{z}^T[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)]\boldsymbol{z} + \mathrm{tr}[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)] > v_{n,j,k}\right) \leq e^{-cv_{n,j,k}k_{n,j}^{\alpha-1/2}}.$$

Moreover a Taylor expansion of $A(\xi')$ around $A(\boldsymbol{\theta})$ implies that $\forall \delta > 0$ and for all ξ, ξ'

(5.6)
$$|\boldsymbol{z}^T[A(\boldsymbol{\theta}) - A(\boldsymbol{\theta}')]\boldsymbol{z} + \mathrm{tr}[A(\boldsymbol{\theta}) - A(\boldsymbol{\theta}')]| \lesssim (\boldsymbol{z}^T\boldsymbol{z} + n)n^{2(d-d_o)_+ + \delta}[\sum_{j=0}^{k}|\xi_j - \xi_j'| + |d - d'|].$$

Without loss of generality we can restrict ourselves on $\Omega_n = \{\boldsymbol{z}^T\boldsymbol{z} \leq 2n\}$, since $\mathbb{P}[\Omega_n^c] = o(e^{-cn})$ for some positive c.

- If $k \leq k_{n,j}$ and $j \leq J_{n,1}$, if $\|\boldsymbol{\xi} - \boldsymbol{\xi'}\| \leq n^{-1-\epsilon}k^{-1/2}$ and $|d - d'| \leq n^{-1-\epsilon}$,
$$|z'[A(\boldsymbol{\theta}) - A(\boldsymbol{\theta'})]z + tr[A(\boldsymbol{\theta}) - A(\boldsymbol{\theta'})]| = o(1), \quad \text{uniformly.}$$

If $k \leq k_{n,j}$ and $j \geq J_{n,1}$, then
$$|z'[A(\boldsymbol{\theta}) - A(\boldsymbol{\theta'})]z + tr[A(\boldsymbol{\theta}) - A(\boldsymbol{\theta'})]| = o(n\epsilon_n^2 j^2), \quad \text{uniformly.}$$

Let $E_{n,j,k}$ be the covering number of $S_{n,j,k}$ by balls satisfying the above constraint then
$$E_{n,j,k} \leq e^{C'k\log n}$$

- If $k > k_{n,j}$. First if $k > k_2 n\epsilon_n^2$, for some $k_2 > 0$ possibly large, then uniformly over $\sum_{j=1}^{k}|\xi_j - \xi_j'| \leq n^{-1-\epsilon}k$ and $|d - d'| \leq n^{-1-\epsilon}$,
$$|\boldsymbol{z}^T[A(\boldsymbol{\theta}) - A(\boldsymbol{\theta'})]\boldsymbol{z} + \mathrm{tr}[A(\boldsymbol{\theta}) - A(\boldsymbol{\theta'})]| = o(k)$$

The covering number of $S_{n,j,k}$ by the above constraints depends on k. Define $K_n(k)$ such that $K_n(k)^{-(\alpha-1/2)} = n^{-1-\epsilon}k$, i.e. $K_n(k) = n^{(1+\epsilon)/(\alpha-1/2)}k^{-1/(\alpha-1/2)}$. Then
$$E_{n,j,k} \leq \exp(C\log n(k \wedge K_n(k)))$$

Now if $k < k_2 n\epsilon_n^2$, then uniformly over $\sum_{j=1}^{k}|\xi_j - \xi_j'| \leq n^{-1-\epsilon}(nj^2\epsilon_n^2)$ and $|d - d'| \leq n^{-1-\epsilon}$,
$$|\boldsymbol{z}^T[A(\boldsymbol{\theta}) - A(\boldsymbol{\theta'})]\boldsymbol{z} + \mathrm{tr}[A(\boldsymbol{\theta}) - A(\boldsymbol{\theta'})]| = o(nj^2\epsilon_n^2)$$
and
$$E_{n,j,k} \leq \exp(Ck\log n).$$

A simple chaining argument in $S_{n,j,k}$ implies that for all $j \leq J_{n,1}$ and all $k \leq k_{n,j}$, there exists $M > 0$ such that

(5.7)
$$\mathbb{P}\left[\Omega_n \cap \{\sup_{\boldsymbol{\theta}\in S_{n,j,k}}\left(\boldsymbol{z}^T[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)]\boldsymbol{z} + \mathrm{tr}[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)]\right) > v_{n,j,k} + M\}\right] \leq e^{-ckn^{\epsilon}}$$

and if $j \geq J_{n,1}$, for any $\delta > 0$ and n large enough

(5.8)
$$\mathbb{P}\left[\Omega_n \cap \{\sup_{\boldsymbol{\theta} \in S_{n,j,k}} \left(\boldsymbol{z}^T[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)]\boldsymbol{z} + \mathrm{tr}[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)]\right) > v_{n,j,k} + \delta n j^2 \epsilon_n^2\}\right] \leq e^{-ckn^{\epsilon}}.$$

Note also that for all $k \leq k_{n,j}$, $v_{n,j,k} = k^{3/2} n^{\epsilon} j \epsilon_n \leq n^{\epsilon}(j\epsilon_n)^{1-3/(2\alpha)}$ so that whenever $\alpha > 3/2$ and $j \leq J_{n,1}$, $v_{n,j,k} = o(1)$ and if $j \geq J_{n,1}$, $v_{n,j,k} = o(nj^2\epsilon_n^2)$.

If $k > k_{n,j}$ and $k > k_2 n\epsilon_n^2$, for all j, and all $\delta > 0$

$$\mathbb{P}\left[\sup_{\boldsymbol{\theta} \in S_{n,j,k}} \left(\boldsymbol{z}^T[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)]\boldsymbol{z} + \mathrm{tr}[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)]\right) > v_{n,j,k} + \delta k\right] \leq e^{-ck^{\tau}}, \tag{5.9}$$

If $k < k_2 n\epsilon_n^2$, for all $\delta > 0$ and n large enough,

(5.10)
$$\mathbb{P}\left[\Omega_n \cap \{\sup_{\boldsymbol{\theta} \in S_{n,j,k}} \left(\boldsymbol{z}^T[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)]\boldsymbol{z} + \mathrm{tr}[A(\boldsymbol{\theta}) - A(\bar{\boldsymbol{\theta}}_d)]\right) > v_{n,j,k} + \delta n j^2 \epsilon_n^2\}\right] \leq e^{-c(k\log n + n^{\epsilon})},$$

and for all $k > k_{n,j}$, $v_{n,j,k} = o(k + nj^2\epsilon_n^2)$. We now study

$$\sup_d \left|\boldsymbol{z}^T[A(\boldsymbol{\theta}_o) - A(\bar{\boldsymbol{\theta}}_d)]\boldsymbol{z} + \mathrm{tr}[A(\boldsymbol{\theta}_o) - A(\bar{\boldsymbol{\theta}}_d)]\right|$$

We have

$$|A(\boldsymbol{\theta}_o) - A(\bar{\boldsymbol{\theta}}_d)|_{Fr}^2 = 0 + \mathrm{Error},$$

where Error is the approximation error of the trace by its limiting integral. We bound separately the approximation error of $\mathrm{tr}\left[(\boldsymbol{T}(f_o)\boldsymbol{T}(f_o^{-1} - f_{\bar{\boldsymbol{\theta}}_d}^{-1}))^2\right]$ and of $\mathrm{tr}\left[(\boldsymbol{T}(f_o)(\boldsymbol{T}(f_o)^{-1} - \boldsymbol{T}(f_{\bar{\boldsymbol{\theta}}_d})^{-1})^2)\right]$. We have, see Kruijer and Rousseau (2011),

$$f_{\bar{\boldsymbol{\theta}}_d} = f_o e^{(d-d_o)H_k - \Delta_{d_o,k}}, \quad H_k(x) = \sum_{j>k} \eta_j \cos(jx), \quad \eta_j = 2/j \quad \Delta_{d_o,k} = \sum_{j>k} \theta_{o,j}\cos(jx)$$

so that

$$f_{\bar{\boldsymbol{\theta}}_d} - f_o = f_o((d-d_o)H_k - \Delta_{d_o,k}) + O(((d-d_o)H_k - \Delta_{d_o,k})^2 x^{-2|d-d_o|})$$

Therefore, from Lemma 2.6 in Kruijer and Rousseau (2011) in the supplement,

(5.11)
$$\begin{aligned}&\mathrm{error}\left(\mathrm{tr}\left[(\boldsymbol{T}(f_o)\boldsymbol{T}(f_o^{-1} - f_{\bar{\boldsymbol{\theta}}_d}^{-1}))^2\right]\right)\\ &= O((|d-d_o| + \|\Delta_{d_o,k}\|_{\infty})^2 \log n + (|d-d_o| + \|\Delta_{d_o,k}\|_{\infty}) l(f_o, f_{\bar{\boldsymbol{\theta}}_d})^{1/2} (\log n)^3).\end{aligned}$$

Similarly using Lemma 2.4 in the supplement of Kruijer and Rousseau (2011)

$$\mathrm{error}\left(\mathrm{tr}\left[(\boldsymbol{T}(f_o)(\boldsymbol{T}(f_o)^{-1} - \boldsymbol{T}(f_{\bar{\boldsymbol{\theta}}_d})^{-1})^2)\right]\right) = O((|d-d_o| + \|\Delta_{d_o,k}\|_{\infty})^2 n^{\epsilon})$$

for all $\epsilon > 0$. Let (k, d) be such that $\bar{\boldsymbol{\theta}}_d \in S_{n,j,k}$, then

(5.12)
$$|A(\xi_o) - A(\bar{\boldsymbol{\theta}}_d)|_{Fr} = O((|d-d_o| + \|\Delta_{d_o,k}\|_{\infty})n^{\epsilon}) = O(n^{\epsilon}(j\epsilon_n)^{(2\alpha-1)/(2\alpha)}), \quad \forall \epsilon > 0$$

see Lemma 3.1 in Kruijer and Rousseau (2011). We also have that for all $|d - d'| \leq n^{-2}$,

$$|\boldsymbol{z}^T[A(\bar{\boldsymbol{\theta}}'_d) - A(\bar{\boldsymbol{\theta}}_d)]\boldsymbol{z} - \mathrm{tr}[A(\bar{\boldsymbol{\theta}}'_d) - A(\bar{\boldsymbol{\theta}}_d)]| \leq (\boldsymbol{z}^T\boldsymbol{z} + n)n^{-1} = O_p(1)$$

uniformly, so that a simple chaining argument combined with (5.12) and Lemma 1.3 in the supplement of Kruijer and Rousseau (2011) implies that, for all $\epsilon > 0$

(5.13)
$$\mathbb{P}\left(\sup_d |\boldsymbol{z}^T[A(\bar{\boldsymbol{\theta}}_d) - A(\boldsymbol{\theta}_o)]\boldsymbol{z} - \mathrm{tr}[A(\bar{\boldsymbol{\theta}}_d) - A(\boldsymbol{\theta}_o)]| > n^{\epsilon}(j\epsilon_n)^{(2\alpha-1)/(2\alpha)}\right) \leq e^{-n^{\epsilon/2}}$$

Hence, combining (5.7), (5.9), (5.10) and (5.13) together with the fact that there exists $J_0, C_0 > 0$ such that for all $j \geq J_0$, setting $S_{n,j} = \{\boldsymbol{\theta}; (j-1)\epsilon_n \leq l(\boldsymbol{\theta}_o, \boldsymbol{\theta}) \leq j\epsilon_n\}$,

$$\int_{S_{n,j}} \exp\{l_n(\boldsymbol{\theta}) - l_n(\boldsymbol{\theta}_o)\}\, d\pi(\boldsymbol{\theta}) \leq e^{-C_0 j n \epsilon_n^2},$$

on a set having probability going to 1. Thus for all $\delta > 0$, there exists a set with P_o probability going to 1 such that

$$\begin{aligned}
&\sum_{j=J_0}^{J_n} \sum_{k\in\mathbb{N}} p(k) N_{n,j,k} \\
&\leq \sum_{j=1}^{J_n} \sum_{k=1}^{k_{n,1}} \sup_{(d,\bar{\boldsymbol{\xi}}_{d,k})\in S_{n,j,k}} \frac{w(\boldsymbol{\theta}_o)}{w(\bar{\boldsymbol{\theta}}_d)} \sup_{(d,\theta)\in S_{n,j,k}} \frac{w(\bar{\boldsymbol{\theta}}_d)}{w(\boldsymbol{\theta})} p(k) \int_{S_{n,j,k}} e^{l_n(\boldsymbol{\theta}) - l_n(\boldsymbol{\theta}_o)} dp(\boldsymbol{\theta}) \\
&\leq \sum_{j=J_0}^{J_n} e^{n^{\epsilon}(j\epsilon_n)^{(2\alpha-1)/(2\alpha)}} \sum_{k=1}^{k_{n,1}} e^{\delta(nj^2\epsilon_n^2 + k)} p(k) \int_{S_{n,j,k}} e^{l_n(\boldsymbol{\theta}) - l_n(\boldsymbol{\theta}_o)} dp(\boldsymbol{\theta}) \\
&\leq e^{-2cn\epsilon_n^2}
\end{aligned}$$

Finally note that if there exists $\xi = (d, k, \theta) \in \cup_{j<J_0} S_{n,j,k}$, then $\bar{\theta}_{d,k} \in \cup_{j<J_0} S_{n,j,k}$ by definition of $\bar{\theta}_{d,k}$. Replacing ϵ_n by $J_0\epsilon_n$, with an abuse of notations, we write $S_{n,1,k} := \cup_{j\leq J_0} S_{n,j,k}$ for all k and we split the set k into $k \leq k_{n,1}$ and $k > k_{n,1}$.

$$\begin{aligned}
&\mathbb{E}^{\pi_n}\left[w^2(\boldsymbol{\theta}) \sum_{k=1}^{k_{n,1}} \mathbb{1}_{S_{n,1,k}}\right] \\
&\qquad \leq w_o^2 \sum_{k=1}^{k_{n,1}} \sup_{(d,\bar{\boldsymbol{\xi}}_{d,k})\in S_{n,1,k}} \frac{w^2(\bar{\boldsymbol{\theta}}_d)}{w_o^2} \sup_{(d,\boldsymbol{\xi})\in S_{n,1,k}} \frac{w^2(\boldsymbol{\theta})}{w^2(\bar{\boldsymbol{\theta}}_d)} p(k|x)\pi_n(S_{n,1,k}|k),
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}^{\pi_n}\left[w^2(\boldsymbol{\theta}) \sum_{k=1}^{k_{n,1}} \mathbb{1}_{S_{n,1,k}}\right] \\
&\qquad \geq w_o^2 \sum_{k=1}^{k_{n,1}} \inf_{(d,\bar{\boldsymbol{\xi}}_{d,k})\in S_{n,1,k}} \frac{w^2(\bar{\boldsymbol{\theta}}_d)}{w_o^2} \inf_{(d,\theta)\in S_{n,1,k}} \frac{w^2(\boldsymbol{\theta})}{w^2(\bar{\boldsymbol{\theta}}_d)} p(k|x)\pi_n(S_{n,1,k}|k),
\end{aligned}$$

we have using (5.13) associated with $j = 1$, with probability going to 1 uniformly in $(d, \bar{\boldsymbol{\xi}}_{d,k}) \in \cup_{k<k_{n,1}} S_{n,1,k}$ $w^2(\bar{\boldsymbol{\theta}}_d)/w_o^2 = 1 + o_p(1)$, and using (5.7) associated with $j = 1$, $w^2(\boldsymbol{\theta})/w^2(\bar{\boldsymbol{\theta}}_d) = 1 + o_p(1)$ uniformly in $(d, \theta) \in \cup_{k<k_{n,1}} S_{n,1,k}$. Let

$k > k_{n,1} = k_0 \epsilon_n^{-1/\alpha} \gtrsim (n/\log n)^{\frac{\beta/\alpha}{2\beta+1}} >> n\epsilon_n^2$, then

$$\mathbb{E}^{\pi_n}\left[w^2(\boldsymbol{\theta}) \sum_{k=k_{n,1}+1}^{\infty} \mathbb{1}_{S_{n,1,k}}\right] \leq \frac{1}{w_o D_n} \sum_{k=k_{n,1}+1}^{\infty} p(k) \int_{S_{n,1,k}} \frac{w(\boldsymbol{\theta})^{-1}}{w(\bar{\boldsymbol{\theta}}_d)^{-1}} \frac{w(\bar{\boldsymbol{\theta}}_d)^{-1}}{w(\boldsymbol{\theta}_o)^{-1}} \exp\left\{l_n(\boldsymbol{\theta}) - l_n(\boldsymbol{\theta}_o)\right\} p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

and on the set or x such that $D_n \geq e^{-cn\epsilon_n^2}$,

$$\begin{aligned} &\mathbb{E}^{\pi_n}\left[w^2(\boldsymbol{\theta}) \sum_{k=k_{n,1}+1}^{\infty} \mathbb{1}_{S_{n,1,k}}\right] \\ &\leq \frac{e^{cn\epsilon_n^2}}{w_o} \sum_{k=k_{n,1}+1}^{\infty} p(k) \sup_{(d,\theta)\in S_{n,1,k}} \frac{w(\boldsymbol{\theta})^{-1}}{w(\bar{\boldsymbol{\theta}}_d)^{-1}} \sup_{d:\bar{\boldsymbol{\theta}}_d \in S_{n,1,k}} \frac{w(\bar{\boldsymbol{\theta}}_d)^{-1}}{w(\boldsymbol{\theta}_o)^{-1}} \\ &\qquad \times \int_{S_{n,1,k}} \exp\left\{l_n(\boldsymbol{\theta}) - l_n(\boldsymbol{\theta}_o)\right\} d\pi(d,\theta) \\ &\leq \frac{e^{cn\epsilon_n^2}}{w_o} \sum_{k=k_{n,1}+1}^{\infty} p(k) e^{\epsilon k} \int_{S_{n,1,k}} \exp\left\{l_n(\boldsymbol{\theta}) - l_n(\boldsymbol{\theta}_o)\right\} p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \end{aligned} \tag{5.14}$$

for all $\epsilon > 0$, on a set of probability going to 1 using (5.9) and (5.10). A Markov inequality implies that

$$P_o^n\left[\sum_{k=k_{n,1}+1}^{\infty} p(k) e^{\epsilon k} \int_{S_{n,1,k}} \exp\left\{l_n(\boldsymbol{\theta}) - l_n(\boldsymbol{\theta}_o)\right\} d\pi(d,\theta) > u_n \sum_{k=k_{n,1}+1}^{\infty} p(k) e^{\epsilon k}\right]$$

is $O(1/u_n)$ for all u_n going to infinity, so that with probability going to 1,

$$\mathbb{E}^{\pi_n}\left[w^2(\boldsymbol{\theta}) \sum_{k=k_{n,1}+1}^{\infty} \mathbb{1}_{S_{n,1,k}}\right] \leq e^{-c_1 k_{n,1} + cn\epsilon_n^2} = o(1)$$

for some $c_1, c > 0$. We finally obtain that

$$\mathbb{E}^{\pi_n}\left[w^2(\boldsymbol{\theta})\right] = w^2(\boldsymbol{\theta}_o)(1 + o_p(1)) + w^{-1}(\boldsymbol{\theta}_o) e^{-Cn\epsilon_n^2} o_p(1),$$

for some $C > 0$. Simple computations show that for all $\epsilon > 0$,

$$P_o^n\left[|\log w(\boldsymbol{\theta}_o)| > n^{\epsilon}\right] = o(1),$$

we can also make precise the upper bound on $|\log w(\boldsymbol{\theta}_o)|$ by controlling better $|A(\boldsymbol{\theta}_o)|$, leading to a term of order $(\log n)^3$, which terminates the proof.

References

Adamchik, V. (2001). On the Barnes function. *Proceedings of the 2001 International Symposium on Symbolic and Algebraic Computation*, 101(474):15–20.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statist. Comput.*, 18(4):343–373.

Beran, J. (1993). Fitting long-memory models by generalized linear regression. *Biometrika*, 80(4):817.

Bertelli, S. and Caporin, M. (2002). A note on calculating autocovariances of long-memory processes. *J. Time Ser. Anal.*, 23(5):503–508.

Böttcher, A. and Silbermann, B. (1999). *Introduction to large truncated Toeplitz matrices*. Universitext. Springer-Verlag, New York.

Brockwell, P. and Davis, R. (2009). *Time series: theory and methods*. Springer Verlag, second ed. edition.

Chen, W., Hurvich, C., and Lu, Y. (2006). On the correlation matrix of the discrete Fourier transform and the fast solution of large Toeplitz systems for long-memory time series. *J. Am. Statist. Assoc.*, 101:812–821.

Dahlhaus, R. (1989). Efficient parameter estimation for self-similar processes. *Ann. Stat.*, 17(4):1749–1766.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Green, P. J. (2003). Trans-dimensional markov chain monte carlo. In Green, P., Lid Hjort, N., and Richardson, S., editors, *Markov chain Monte Carlo in practice*, pages 179–198. Oxford University Press.

Holan, S., McElroy, T., and Chakraborty, S. (2009). A Bayesian approach to estimating the long memory parameter. *Bayesian Analysis*, 4(1):159–190.

Hurvich, C., Moulines, E., and Soulier, P. (2002). The FEXP estimator for potentially non-stationary linear time series. *Stoch. Processes and their applications*, 97(2):307–340.

Jasra, A., Stephens, D., A. Doucet, A., and Tsagaris, T. (2011). Inference for Lévy driven stochastic volatility models via Sequential Monte Carlo. *Scand. J. of Statist.*, 38(1).

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputation and bayesian missing data problems. *J. Am. Statist. Assoc.*, 89:278–288.

Kruijer, W. and Rousseau, J. (2011). Bayesian semi-parametric estimation of the long-memory parameter under FEXP-priors. *ARrxiv 1202.4863*.

Leland, W., Taqqu, M., Willinger, W., and Wilson, D. (1994). On the self-similar nature of ethernet traffic. *IEEE/ACM Transactions on Networking*, 2(1):1–15.

Levinson, N. (1949). The Wiener RMS (root mean square) error criterion in filter design and prediction, appendix b of n. wiener, extrapolation, interpolation and smoothing of stationary time series.

Moulines, E. and Soulier, P. (2003). Semiparametric spectral estimation for fractional processes. In Doukhan, P., Oppenheim, G., and Taqqu, M., editors, *Theory and applications of long-range dependence*, pages 251–301. Birkhäuser Boston.

Neal, R. M. (2001). Annealed importance sampling. *Statist. Comput.*, 11:125–139.

Palma, W. (2007). *Long-memory time series: theory and methods*, volume 662. Wiley-Blackwell.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. R. Statist. Soc. B*, 59(4):731–792.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods, 2nd ed.* Springer-Verlag, New York.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statist. Science*, 16(4):351–367.

Rousseau, J., Chopin, N., and Liseo, B. (2012). Bayesian nonparametric estimation of the spectral density of a long or intermediate memory gaussian process. *Ann. Stat.*, (in press).

Schäfer, C. and Chopin, N. (2011). Sequential Monte Carlo on large binary sampling spaces. *Statist. Comput.*, (in press). 10.1007/s11222-011-9299-z.