FAST MCMC COMPUTATIONS FOR THE ESTIMATION OF SPARSE PROCESSES FROM NOISY OBSERVATIONS

Manuel Davy and Jérôme Idier

IRCCyN/CNRS, 1 rue de la Noë, BP92101, 44321 Nantes cedex 3, France
email : {davy,idier}@irccyn.ec-nantes.fr

ABSTRACT

This paper presents a fast MCMC algorithm specially designed for high dimensional models with block structure. Such models are often met in Bayesian inference, such as spectral estimation, harmonic analysis, blind deconvolution or signal classification. Our algorithm generates samples distributed according to a posterior distribution. We show that sampling the amplitudes together with the remaining model parameters leads to quicker computations than sampling from the marginal posterior, where amplitudes have been integrated out. Simulation results demonstrate the soundness of this approach for high dimensional models.

1. INTRODUCTION

A number of Bayesian inference problems in Signal Processing can be cast in the form of the following Gaussian model:

$$\mathbf{y} \;=\; \mathbf{D}(\boldsymbol{\theta})\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y} = [y(1), \ldots, y(N)]^{\mathsf{T}}$ is the vector of observed data, $\mathbf{D}(\boldsymbol{\theta})$ is a $N \times R$ matrix of basis functions (stored in columns) with possible r-dimensional parameters $\boldsymbol{\theta}$. The basis functions amplitudes are $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_R]^{\mathsf{T}}$ and the additive noise $\boldsymbol{\epsilon} = [\epsilon(1), \ldots, \epsilon(N)]^{\mathsf{T}}$ is zero-mean Gaussian, white with variance v_ϵ. The overall objective is the estimation of the parameters $\boldsymbol{\theta}$ as well as the amplitudes $\boldsymbol{\beta}$ and the noise variance v_ϵ. Example problems are spectral estimation [1], harmonic analysis of music [2], blind deconvolution [3, 4] and signal classification [5]. Bayes parameter estimation generally requires the computation of high dimensional intractable integrals, and a good solution consists of implementing Monte Carlo Markov Chain (MCMC) methods [6]. An important remark is that, in the above examples, the parameter $\boldsymbol{\theta}$ and amplitudes $\boldsymbol{\beta}$ can be decomposed into K blocks with similar structure,

$$\begin{array}{rcl} \boldsymbol{\theta} & = & [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K] \\ \boldsymbol{\beta} & = & [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K] \end{array} \quad \text{with } \mathbf{y} = \sum_{k=1}^{K} \mathbf{D}(\boldsymbol{\theta}_k)\boldsymbol{\beta}_k + \boldsymbol{\epsilon}$$

where the size of $\boldsymbol{\theta}_k$ (resp. of $\boldsymbol{\beta}_k$) is r_k with $r_1 + \ldots + r_K = r$ (resp. $R_k$with $R_1 + \ldots + R_K = R$). Example blocks are single sinusoids in spectral analysis [1] or individual notes in harmonic analysis [2]. In this paper, we exploit this specific structure to design a fast MCMC algorithm.

When selecting a Gaussian amplitude prior and an inverse Gamma noise variance prior, one can marginalize $\boldsymbol{\beta}$ and v_ϵ in the posterior probability distribution function (pdf) $p(\boldsymbol{\beta}, v_\epsilon, \boldsymbol{\theta}|\mathbf{y})$, and the inference can be based on $p(\boldsymbol{\theta}|\mathbf{y})$ only. The conditional pdfs $p(v_\epsilon|\boldsymbol{\theta}, \mathbf{y})$ and $p(\boldsymbol{\beta}|v_\epsilon, \boldsymbol{\theta}, \mathbf{y})$ are also available and can be used for inference as well. In practice, this results in decoupling the estimation of v_ϵ and $\boldsymbol{\beta}$ from that of $\boldsymbol{\theta}$. This can be directly used to design an efficient MCMC algorithm [1–3]: a Markov Chain composed of samples $\{\widetilde{\boldsymbol{\theta}}^{(l)}\}_{l=1,\ldots,L}$ with invariant distribution $p(\boldsymbol{\theta}|\mathbf{y})$ is built, then samples $\{\widetilde{v}_\epsilon^{(l)}\}_{l=1,\ldots,L}$ and $\{\widetilde{\boldsymbol{\beta}}^{(l)}\}_{l=1,\ldots,L}$ are generated using the conditional densities. In the following, this approach is referred to as *hierarchical sampling* (HS).

The above HS approach is motivated by minimizing the variance of integral estimates: the smaller the parameter space, the more accurate the estimation. More precisely, the quadratic error made by estimators based on MCMC samples typically decreases at rate α/L. A smaller value is expected for α_{HS} since amplitudes have been integrated out in the HS approach. Yet, we argue that for high dimensional problems (large R) where $\mathbf{D}(\boldsymbol{\theta})$ and $\boldsymbol{\beta}$ can be decomposed into K blocks of same structure, it is more efficient to adopt a *Joint Sampling* (JS) approach, that generates samples $\{\widetilde{\boldsymbol{\beta}}^{(l)}, \widetilde{\boldsymbol{\theta}}^{(l)}\}_{l=1,\ldots,L}$ block-wise, in a sense that will be made clearer in the following. Actually, JS iterations have a much lower cost than HS iterations, and thus more samples can be generated within the same computation time. Thus, even if α_{JS} is larger than α_{HS}, the joint approach is more accurate (as confirmed by simulations in Section 5).

This paper is organized as follows. In Section 2, we briefly present the example of harmonic model and the estimation objectives. In Section 3, our fast MCMC algorithm is described. Section 4 is devoted to discussions and simulation results are presented in Section 5. Conclusions are proposed in Section 6.

2. BAYESIAN MODEL AND ESTIMATION

This section specifies the Bayesian model and introduces the example of harmonic analysis.

2.1. An example: Harmonic Models

In harmonic analysis, the matrix $\mathbf{D}(\boldsymbol{\theta})$ is composed of so-called *Gabor* basis functions of the form (for $t = 1, \ldots, N$):

$$\begin{array}{rcl} g_{k,m,i}(t) & = & \phi(t - i\Delta_t)\cos(\omega_{k,m}\, t\,/\omega_s) \\ h_{k,m,i}(t) & = & \phi(t - i\Delta_t)\sin(\omega_{k,m}\, t\,/\omega_s) \end{array} \tag{2}$$

where $\Delta_t \;=\; (N-1)/I$ for $i = 0, \ldots, I$, $m = 1, \ldots, M_k$ and $k = 1, \ldots, K$. In eq. (2), the windowing function ϕ has, e.g., Gauss or Hamming shape, see [2]. The overall model is actually hierarchical on two levels. At the highest level, the *overall level*, $\mathbf{y}$ is composed of K notes played by one or several instruments. At the middle level, the *notes level*, each note $k = 1, \ldots, K$ is composed of M_k partials, that is, sine waves with frequencies $\omega_{k,m}$,

$m = 1, \ldots, M_k$ roughly related by $\omega_{k,m} \approx m\omega_{k,1}$. At the lowest level, the *partials level*, each sine wave has a time-varying amplitude which is written as a linear combination of time shifted windows $\phi(t - i\Delta_t)$. The initial phase is determined by the amplitudes of both $g_{k,m,i}$ and $h_{k,m,i}$. Clearly, this model can be written block-wise as follows

$$\mathbf{y} \;=\; \sum_{k=1}^{K} \mathbf{y}_k + \boldsymbol{\epsilon} \text{ with } \mathbf{y}_k \;=\; \sum_{m=1}^{M_k} \mathbf{y}_{k,m}$$

where $\mathbf{y}_k$, $k = 1, \ldots, K$ are individual notes and $\mathbf{y}_{k,m} = [y_{k,m}(1), \ldots, y_{k,m}(N)]^\mathsf{T}$ are individual partials, with

$$y_{k,m}(t) \;=\; \sum_{i=0}^{I} \left(a_{k,m,i} g_{k,m,i}(t) + b_{k,m,i} h_{k,m,i}(t)\right)$$

for $t = 1, \ldots, N$. The amplitudes $a_{k,m,i}$ and $b_{k,m,i}$ $(i = 0, \ldots, I)$ are the elements of $\boldsymbol{\beta}$ corresponding to $y_{k,m}$. The parameters of this model are K and $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K]$ where $\boldsymbol{\theta}_k = [\omega_{k,1}, \ldots, \omega_{k,M_k}]$ (with size $r_k = M_k$) for $k = 1, \ldots, K$. Block size is $R_k = 2(I+1)M_k$, where the parameter I is generally assumed known, as its influence is not critical. Note that $\boldsymbol{\beta}$ and $\mathbf{D}(\boldsymbol{\theta})$ can have very high dimension (about 1000 to 5000) in standard harmonic estimation problems.

2.2. Posterior pdf

The likelihood function related to eq. (1) is

$$p(\mathbf{y}|\boldsymbol{\beta}, v_\epsilon, \boldsymbol{\theta}) \;=\; [2\pi v_\epsilon]^{-N/2} \exp -\frac{1}{2v_\epsilon} ||\mathbf{y} - \mathbf{D}(\boldsymbol{\theta})\boldsymbol{\beta}||^2$$

Consider the prior structure $p(\boldsymbol{\beta}, v_\epsilon, \boldsymbol{\theta}) \;=\; p(\boldsymbol{\beta}|v_\epsilon, \boldsymbol{\theta}) p(v_\epsilon) p(\boldsymbol{\theta})$ where $p(\boldsymbol{\beta}|v_\epsilon, \boldsymbol{\theta})$ is zero-mean Gaussian $\mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}/v_\epsilon)$ where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is block-diagonal, the size of each block is R_k. The prior $p(v_\epsilon)$ is the inverse Gamma distribution $\mathcal{IG}(\nu_0/2, \gamma_0/2)$ (see [1–3]). The marginalized posterior is

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta})\, |\mathbf{S}(\boldsymbol{\theta})|^{1/2} \left[\frac{1}{2}\left(\mathbf{y}^\mathsf{T}\mathbf{P}(\boldsymbol{\theta})\mathbf{y} + \gamma_0\right)\right]^{-\frac{N+\nu_0}{2}} \qquad (3)$$

where $\mathbf{P}(\boldsymbol{\theta}) = \mathbf{I}_N - \mathbf{D}(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta})\mathbf{D}(\boldsymbol{\theta})^\mathsf{T}$ is a $N \times N$-dimensional matrix, $\mathbf{I}_N$ denotes the N-dimensional identity matrix, $|\cdot|$ denotes matrix determinant and $\mathbf{S}(\boldsymbol{\theta}) \;=\; [\mathbf{D}(\boldsymbol{\theta})^\mathsf{T}\mathbf{D}(\boldsymbol{\theta}) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}]^{-1}$ is a $R \times R$-dimensional matrix . The conditional posterior densities are given by

$$p(v_\epsilon|\boldsymbol{\theta}, \mathbf{y}) \;=\; \mathcal{IG}\left((N+\nu_0)/2, (\mathbf{y}^\mathsf{T}\mathbf{P}(\boldsymbol{\theta})\mathbf{y} + \gamma_0)/2\right) \quad (4)$$

$$p(\boldsymbol{\beta}|v_\epsilon, \boldsymbol{\theta}, \mathbf{y}) \;=\; \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{v_\epsilon}\mathbf{S}(\boldsymbol{\theta})\right) \qquad (5)$$

where $\boldsymbol{\mu} = \mathbf{S}(\boldsymbol{\theta})\mathbf{D}(\boldsymbol{\theta})^\mathsf{T}\mathbf{y}$.

In order to manipulate blocks of amplitudes/parameters, we introduce the notation with subscript $-k$ to denote all the components but component number k. Using these notations, each block can be written $\mathbf{y}_k \;=\; \mathbf{D}(\boldsymbol{\theta}_k)\boldsymbol{\beta}_k$ where $\mathbf{D}(\boldsymbol{\theta}_k)$ is a $N \times R_k$ matrix. Denote $\mathbf{e}_k = \mathbf{y} - \mathbf{D}(\boldsymbol{\theta}_{-k})\boldsymbol{\beta}_{-k}$. Then, $p(\boldsymbol{\theta}_k|\boldsymbol{\beta}_{-k}, v_\epsilon, \boldsymbol{\theta}_{-k}, \mathbf{y}) = p(\boldsymbol{\theta}_k|\mathbf{e}_k)$ is given in eq. (3), with

$$\mathbf{P}(\boldsymbol{\theta}_k) \;=\; \mathbf{I}_N - \mathbf{D}(\boldsymbol{\theta}_k)\mathbf{S}(\boldsymbol{\theta}_k)\mathbf{D}(\boldsymbol{\theta}_k)^\mathsf{T}$$

instead of $\mathbf{P}(\boldsymbol{\theta})$ and $\mathbf{e}_k$ instead of $\mathbf{y}$. Note that the dimension of $\mathbf{S}(\boldsymbol{\theta}_k)$ is $R_k \leq R$. Similarly, $p(\boldsymbol{\beta}_k|\boldsymbol{\beta}_{-k}, v_\epsilon, \boldsymbol{\theta}, \mathbf{y}) = p(\boldsymbol{\beta}_k|v_\epsilon, \boldsymbol{\theta}_k, \mathbf{e}_k)$ has the same structure as in eq. (5), with $\boldsymbol{\mu}_k = \mathbf{S}(\boldsymbol{\theta}_k)\mathbf{D}(\boldsymbol{\theta}_k)^\mathsf{T}\mathbf{e}_k$ instead of $\boldsymbol{\mu}$ and $\mathbf{S}(\boldsymbol{\theta}_k)$ instead of $\mathbf{S}(\boldsymbol{\theta})$.

2.3. Estimation objectives

In the harmonic example above, the estimation objective consists of computing MMAP estimates of the number of notes $\widehat{K}$ and the numbers of partials $\widehat{\mathbf{M}}$. The frequencies are estimated via the following MMSE integral

$$\widehat{\boldsymbol{\omega}} \;=\; \int \boldsymbol{\omega} p(\boldsymbol{\omega}, K, \mathbf{M}|\mathbf{y}) \delta_{\widehat{K}, \widehat{\mathbf{M}}}(K, \mathbf{M})\, d\boldsymbol{\omega} dK\, d\mathbf{M}$$

which cannot be computed analytically ($\delta_u(v)$ denotes the Dirac delta function, i.e. $\delta_u(v) = 0$ whenever $v \neq u$). In such situations, Monte Carlo integration is a convenient approach [1–3, 6], and $\widehat{\boldsymbol{\omega}} \approx \sum_{l=1}^{L} \widetilde{\boldsymbol{\omega}}^{(l)}/L$ where $\{\widetilde{\boldsymbol{\omega}}^{(l)}\}_{l \in \mathcal{L}}$ is part of the joint sample $\{\widetilde{\boldsymbol{\beta}}^{(l)}, \widetilde{v}_\epsilon^{(l)}, \widetilde{\boldsymbol{\omega}}^{(l)}, \widetilde{K}^{(l)}, \widetilde{\mathbf{M}}^{(l)}\}_{l=1,\ldots,L}$ distributed according to $p(\boldsymbol{\beta}, v_\epsilon, \boldsymbol{\omega}, K, \mathbf{M}|\mathbf{y})$, where $\mathcal{L}$ is the set of indices l that have the MMAP values for K and $\mathbf{M}$, i.e., such that $(\widetilde{K}^{(l)}, \widetilde{\mathbf{M}}^{(l)}) = (\widehat{K}, \widehat{\mathbf{M}})$ (the MMAP estimate of, e.g., K is the most represented value among $\widetilde{K}^{(1)}, \ldots, \widetilde{K}^{(L)}$.

3. MCMC ALGORITHM BASED ON JOINT SAMPLING

In this section, we present a fast Metropolis-within-Gibbs algorithm aimed at generating samples $\{\widetilde{\boldsymbol{\theta}}^{(l)}\}_{l=1,\ldots,L}$. For the sake of clarity, the number of blocks K is assumed known here, however, this algorithm can be extended to unknown K so as to include reversible jumps. In order to alleviate notations, we omit the notations $\widetilde{\cdot}^{(l)}$ and $\cdot^\star$ whenever it is clear from the context that we deal with Markov chain samples and with candidates.

Algorithm 1: Joint sampling MCMC algorithm

- Initialization
  - Sample $\widetilde{\boldsymbol{\theta}}^{(1)}$ according to its prior distribution
  - Sample $\widetilde{\sigma}_v^{2(1)}$ from the pdf in eq. (4)
  - Sample the amplitudes $\widetilde{\boldsymbol{\beta}}^{(1)}$ from the pdf in eq. (5)
  - Set $l \leftarrow 2$
- Iterations While $l < L$, do

  `Step 1.1:` For $k = 1, \ldots, K$, update block #k as follows
  - sample $\boldsymbol{\theta}_k^\star$ according to the proposal distribution $q_1(\boldsymbol{\theta}_k^\star|\widetilde{\boldsymbol{\theta}}_{-k}^{(l,l-1)}, \widetilde{\boldsymbol{\theta}}_k^{(l-1)}, \widetilde{\boldsymbol{\beta}}_{-k}^{(l,l-1)}, \mathbf{y})$ and perform a MH test w.r.t

$$p(\boldsymbol{\theta}_k|\widetilde{\boldsymbol{\beta}}_{-k}^{(l,l-1)}, \widetilde{v}_\epsilon^{(l-1)}, \widetilde{\boldsymbol{\theta}}_{-k}^{(l,l-1)}, \mathbf{y}) \propto |\mathbf{S}(\boldsymbol{\theta}_k)|^{1/2} \left[\frac{1}{2}\left(\mathbf{e}_k^\mathsf{T}\mathbf{P}(\boldsymbol{\theta}_k)\mathbf{e}_k + \gamma_0\right)\right]^{\frac{N+\nu_0}{2}} p(\boldsymbol{\theta})$$

    which has the structure of the pdf in eq. (3), see Subsection 2.2. This yields $\widetilde{\boldsymbol{\theta}}_k^{(l)}$.
  - sample the amplitudes $\widetilde{\boldsymbol{\beta}}_k^{(l)}$ from

$$p(\boldsymbol{\beta}_k|\widetilde{\boldsymbol{\beta}}_{-k}^{(l,l-1)}, \widetilde{v}_\epsilon^{(l-1)}, \widetilde{\boldsymbol{\theta}}_k^{(l)}, \widetilde{\boldsymbol{\theta}}_{-k}^{(l,l-1)}, \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{S}(\boldsymbol{\theta}_k)/\widetilde{v}_\epsilon^{(l-1)}),$$

    see eq. (5) and Subsection 2.2.
- `Step 1.2:` Sample $\widetilde{v}_\epsilon^{(l)}$ from $p(v_\epsilon|\widetilde{\boldsymbol{\beta}}^{(l)}, \widetilde{\boldsymbol{\theta}}^{(l)}, \mathbf{y})$, given by

$$\mathcal{IG}\left((N+\nu_0)/2,\, (||\mathbf{e}_0||^2 + \gamma_0)/2\right),$$

  which has the same structure as the pdf in eq. (4), where $\mathbf{e}_0 = \mathbf{y} - \mathbf{D}(\widetilde{\boldsymbol{\theta}}^{(l)})\widetilde{\boldsymbol{\beta}}^{(l)}$.
- Set $l \leftarrow l + 1$,

In the above algorithm, MH stands for "Metropolis -Hastings" and we used the notation $\widetilde{\boldsymbol{\beta}}_{-k}^{(l,l-1)} = [\widetilde{\boldsymbol{\beta}}_1^{(l)}, \ldots, \widetilde{\boldsymbol{\beta}}_{k-1}^{(l)}, \widetilde{\boldsymbol{\beta}}_{k+1}^{(l-1)}, \ldots, \widetilde{\boldsymbol{\beta}}_K^{(l-1)}]$, and a similar notation for $\widetilde{\boldsymbol{\theta}}_{-k}^{(l,l-1)}$. Overall, Step 1.1 has the structure of a "one-block-at-a-time" Metropolis-Hastings sampler operating on blocks $(\boldsymbol{\beta}_k, \boldsymbol{\theta}_k)$. Each block $(\boldsymbol{\beta}_k, \boldsymbol{\theta}_k)$ is generated conditionally on the remaining blocks $(\boldsymbol{\beta}_{-k}, \boldsymbol{\theta}_{-k})$ using marginal/conditional sampling: first, $\boldsymbol{\theta}_k$ is sampled from the marginal posterior $p(\boldsymbol{\theta}_k|\boldsymbol{\beta}_{-k}, \boldsymbol{\theta}_{-k}, \ldots)$, second $\boldsymbol{\beta}_k$ is generated from the conditional posterior $p(\boldsymbol{\beta}_k|\boldsymbol{\theta}_k, \boldsymbol{\beta}_{-k}, \boldsymbol{\theta}_{-k}, \ldots)$. The overall structure is that of a Gibbs sampler, which iterates on $\boldsymbol{\theta}$, v_ϵ and $\boldsymbol{\beta}$. Convergence issues of such MCMC algorithms have been extensively discussed in a number of works (see [1, 3]), and it is not addressed again here.

4. DISCUSSION

In this section, we compare the computation complexity of Algorithm 1.1 and that of the algorithms proposed in [1–3].

4.1. Standard Hierarchical Sampling (HS) algorithm

For the sake of clarity, Algorithm 2 below recalls the structure of HS algorithms proposed in [1–3, 5].

Algorithm 2: Standard HS MCMC for sparse processes

- Initialization: see Algorithm 1
- Iterations While $l < L$, do

  Step 2.1: For $k = 1, \ldots, K$, update block #k as follows
  - sample $\boldsymbol{\theta}_k^\star$ according to the proposal distribution $q_2(\boldsymbol{\theta}_k^\star|\widetilde{\boldsymbol{\theta}}^{(l,l-1)}, \mathbf{y})$ and form $\boldsymbol{\theta}^\star = [\ldots, \widetilde{\boldsymbol{\theta}}_{k-1}^{(l)}, \boldsymbol{\theta}_k^\star, \widetilde{\boldsymbol{\theta}}_{k+1}^{(l-1)}, \ldots]$. Perform a MH test w.r.t $p(\boldsymbol{\theta}|\mathbf{y})$, see eq. (3). This yields $\widetilde{\boldsymbol{\theta}}_k^{(l)}$.

  Step 2.2: Sample $\widetilde{v}_\epsilon^{(l)}$ from $p(v_\epsilon|\widetilde{\boldsymbol{\theta}}^{(l)}, \mathbf{y})$ given in eq. (4)

  Step 2.3: Sample $\widetilde{\boldsymbol{\beta}}^{(l)}$ from $p(\boldsymbol{\beta}|\widetilde{v}_\epsilon^{(l)}, \widetilde{\boldsymbol{\theta}}^{(l)}, \mathbf{y})$, see eq. (5)

  Set $l \leftarrow l + 1$.

Remark: In the case of strong correlations among the sampled parameters, Algorithm 2 is expected to perform better than Algorithm 1: It is generally more efficient to sample *jointly* parameters with strong correlations. In the examples presented in this paper, parameters are not strongly correlated.

4.2. Algorithmic complexity

In Algorithms 1 and 2, the most computationally intensive part consists of the computation of $\mathbf{y}^\mathsf{T}\mathbf{P}(\theta)\mathbf{y}$, which requires the inversion of $\mathbf{S}^{-1} = \mathbf{D}^\mathsf{T}\mathbf{D} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$, with dimension R_k in Algorithm 1, and with dimension R in Algorithm 2, where $R_k \leq R$.

Efficient implementations only require one computation of the target distribution for each MH step, as the value of the posterior, for the last accepted sample, can be stored and reused. Then, each iteration of Algorithm 2 requires $K + 1$ computation of $\mathbf{S}(\boldsymbol{\theta})$ with size R whereas each iteration of Algorithm 1 requires $2K$ computations of $\mathbf{S}(\boldsymbol{\theta}_k)$ with size R_k. In addition, Step 1.1 requires K sampling from the multivariate normal pdf $\mathcal{N}(\boldsymbol{\mu}_k, \mathbf{S}(\boldsymbol{\theta}_k)/\widetilde{v}_\epsilon^{(l-1)})$ with dimension R_k, which usually requires one Choleski decomposition. This decomposition can be avoided, however. Before explaining this point, we introduce an efficient implementation for the computation of $\mathbf{y}^\mathsf{T}\mathbf{P}(\theta)\mathbf{y}$ required in Algorithms 1 and 2 (see Algorithm 3).

Algorithm 3: Computation of $\mathbf{y}^\mathsf{T}\mathbf{P}(\boldsymbol{\theta})\mathbf{y}$ with maximum cost of each step

- Compute $\mathbf{D}^\mathsf{T}\mathbf{D} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$ $\rightarrow \mathrm{O}(R^2N)$
- Compute Choleski factors $\mathbf{C}\mathbf{C}^\mathsf{T} = \mathbf{D}^\mathsf{T}\mathbf{D} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$ $\rightarrow \mathrm{O}(R^3)$
- Compute $\mathbf{u} = \mathbf{D}(\boldsymbol{\theta})^\mathsf{T}\mathbf{y}$ $\rightarrow \mathrm{O}(RN)$
- Solve the triangular system $\mathbf{C}\mathbf{v} = \mathbf{u}$ for $\mathbf{v}$ $\rightarrow \mathrm{O}(R^2)$
- Compute $\mathbf{y}^\mathsf{T}\mathbf{P}(\theta)\mathbf{y} = \mathbf{y}^\mathsf{T}\mathbf{y} - \mathbf{v}^\mathsf{T}\mathbf{v}$ $\rightarrow \mathrm{O}(N)$

The computation of $|\mathbf{S}(\boldsymbol{\theta})|^{-1}$ is straightforward since $|\mathbf{S}(\boldsymbol{\theta})| = |\mathbf{C}|^{-2}$ where $\mathbf{C}$ is triangular. As can be seen, the overall maximal complexity of Algorithm 3 is $\mathrm{O}(R^3 + NR^2)$. In many applications, however, matrix $\mathbf{D}$ is either sparse (harmonic analysis) or with specific structure that can be used to form directly $\mathbf{D}^\mathsf{T}\mathbf{D}$ (spectral analysis) and the actual complexity is much smaller than $\mathrm{O}(NR^2)$. Finally, the overall complexity of Algorithm 3 is $\mathrm{O}(R^3)$. Algorithm 3 is used in both Algorithm 1 (with $\boldsymbol{\theta}_k$ instead of $\boldsymbol{\theta}$ and R_k instead of R) and Algorithm 2. In Algorithm 1, the Choleski factor $\mathbf{C}$ can be reused in the multivariate Gaussian sampling, which is performed as follows.

Algorithm 4: Multivariate Gaussian Sampling with maximum cost of each step

- Sample a i.i.d. vector $\mathbf{w}$ according to $\mathcal{N}(0, 1)$
- Solve the linear triangle system $\mathbf{C}\boldsymbol{\mu}_0 = \mathbf{w}$ $\rightarrow \mathrm{O}(R^2)$
- Solve the linear system $\mathbf{C}\boldsymbol{\mu}' = \mathbf{D}(\boldsymbol{\theta})^\mathsf{T}\mathbf{y}$ $\rightarrow \mathrm{O}(R^2)$
- Solve the linear system $\mathbf{C}^\mathsf{T}\boldsymbol{\mu} = \boldsymbol{\mu}'$ $\rightarrow \mathrm{O}(R^2)$
- Compute $\boldsymbol{\beta} = \boldsymbol{\mu}_0 + \boldsymbol{\mu}$

Consider the Harmonic model presented in Subsection 2.1, the size of each block (i.e., each note) k is $R_k = 2M_k(I+1)$. In standard harmonic estimation problems [2], one typically has $M_k \approx 20$, $I \approx 15$ and $R_k \approx 640$. With three notes, $R \approx 1920$ and the complexity of Algorithm 1 is one order of magnitude smaller than that of Algorithm 2. But one can do even better: each note is made of M_k sub-blocks with size $2(I+1) \approx 32$, and updating each of them one-at-a-time using Algorithm 1 yields even further improvements. The overall complexity of the latter approach is $\mathrm{O}\big((\sum_{k=1}^K M_k)[2(I+1)]^3\big)$ instead of $\mathrm{O}\big(([\sum_{k=1}^K M_k)2(I+1)]^3\big)$ in the hierarchical approach, that is, about 3600 smaller for the above example!

4.3. Proposal distributions

Aside computational complexity, our approach has another major advantage. The proposal q_1 in Algorithm 1 depends on the current amplitudes and parameters, and samples likely candidates. In particular, in spectral analysis [1], a stepwise approximation of the spectrum of $\mathbf{e}_k$, denoted $|\mathrm{TF}(\mathbf{e}_k)|^2$ (not available in Algorithm 2) can be used as q_1, rather than the Fourier transform of $\mathbf{y}$ used as q_2 in Algorithm 2, [1]. This yields an efficient proposal which does not sample already existing frequencies. Moreover, it becomes unnecessary to assume a g-prior structure $\Sigma_{\boldsymbol{\beta}} = \delta^2(\mathbf{D}^\mathsf{T}\mathbf{D})^{-1}$ (as proposed in [1]), since it is mainly aimed at avoiding 1) high dimensional determinant computation in eq. (3), and 2) superimposed frequencies. It is thus interesting to select $\Sigma_{\boldsymbol{\beta}} = v_{\boldsymbol{\beta}}\mathbf{I}_R$ and uniform frequencies prior, because in this case, the log posterior $p(\boldsymbol{\theta}_k|\widetilde{\boldsymbol{\beta}}_{-k}^{(l,l-1)}, \widetilde{v}_\epsilon^{(l-1)}, \widetilde{\boldsymbol{\theta}}_{-k}^{(l,l-1)}, \mathbf{y})$ can be straightforwardly computed from $|\mathrm{TF}(\mathbf{e}_k)|^2$, see [7, eq. (4)]. The example of spectral analysis shows that sampling the amplitudes block-wise enables dramatical reduction of the algorithm complexity.

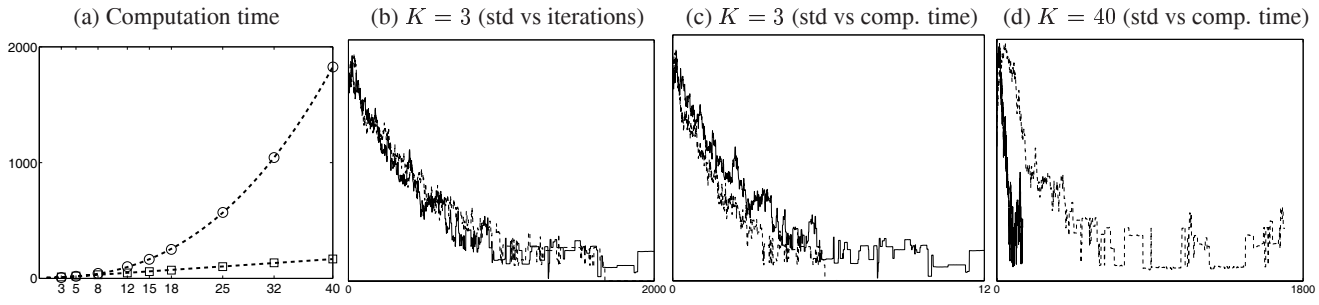

Fig. 1. Performance comparison of Algorithms 1 and 2. (a) Average computation time of Algorithm 1 (squares) and 2 (circles) as a function of the model order K, with $L = 2000$. In dashed lines, polynomial interpolation with order one (squares) and order 3 (circles). (b) Evolution of the standard deviation of the MMSE estimation error as a function of the iteration l for Algorithm 1 (solid lines) and 2 (dashed lines). (c) Same as (b), but plotted as a function of the average computation time. (d) Same as (c) with model order $K = 40$.

	Algorithm 1	Algorithm 2	True value
ω_1	0.0110	0.0110	0.0111
ω_2	0.0996	0.0992	0.0998
ω_3	0.1517	0.1516	0.1515
ω_4	0.1878	0.1878	0.1879
ω_5	0.2137	0.2142	0.2145
ω_6	0.2696	0.2689	0.2692
ω_7	0.4475	0.4473	0.4493
ω_8	0.4551	0.4551	0.4551

Table 1. Frequency estimation results using Algorithm 1 and Algorithm 2 in the case $K = 8$. For each algorithm, the values presented are MMSE estimates obtained by averaging samples from $l = 1900$ to $l = L = 2000$ over the 100 simulated Markov Chains.

5. SIMULATION RESULTS

In this section, we study the performance of Algorithms 1-2 in the case of spectral estimation (see [1] for a full presentation). The case of harmonic analysis is much more complicated, and simulation results can be found in [8].

In spectral estimation, $\boldsymbol{\theta}_k = \omega_k$ and $\mathbf{D}(\boldsymbol{\theta}_k) = [\mathbf{s}_k, \mathbf{c}_k]$ with $\mathbf{s}_k = [\ldots, \sin(\omega_k t), \ldots]^\mathsf{T}$ and $\mathbf{c}_k = [\ldots, \cos(\omega_k t), \ldots]^\mathsf{T}$. Block size is $R_k = 2$ and $r_k = 1$. Here, we assume the number of blocks (i.e., the number of frequencies) K known, though the algorithms could be straightforwardly extended to unknown K via reversible jumps. For the simulations, we create observations according to eq. (1) with $N = 500$ points and noise variance $v_\epsilon = 2$. The amplitudes, initial phases (denoted ψ_k) and frequencies in the simulated $\mathbf{y}$ are selected randomly i.i.d. as $\boldsymbol{\beta} \sim \mathcal{N}(1, 0.5)$, $\psi_k \sim \mathcal{U}([0, 2\pi])$ and $\omega_k \sim \mathcal{U}([0, \pi])$, where $\mathcal{U}[a, b]$ is the uniform distribution on $[a, b]$.

In simulations, the amplitudes prior is $\Sigma_{\boldsymbol{\beta}} = v_{\boldsymbol{\beta}} \mathbf{I}_R$ with $v_{\boldsymbol{\beta}} = 1$. For both Algorithm 1 and Algorithm 2, we simulated 100 Markov chains for the model orders $K = 3, 5, 8, 12, 15, 18, 25, 32, 40$ with $L = 2000$ samples. Results are given in Tab. 1 and plotted in Fig. 1. As can be seen, the computational cost of our algorithm increases *linearly* with K (as blocks size is $R_k = 2$, whatever K), whereas it increases in $O(K^3)$ for the standard algorithm. The estimation error (as a function of the iterations l) does not decrease significantly slower with our algorithm, whatever the model order. On the other hand, the estimation accuracy is much better with our method in terms of computation time. In our implementation, $\mathbf{D}^\mathsf{T}\mathbf{D}$ is formed directly, q_2 is as in [1] and q_1 is as in Subsection 4.3. The matlab files used in the simulations can be downloaded at `http://www.irccyn.ec-nantes.fr/~davy/fastMCMC_icassp.tar.gz`.

6. CONCLUSIONS

In this paper, we have introduced a fast MCMC algorithm for Bayesian inference. Typical applications are spectral estimation [1] or harmonic analysis of music [2], where dramatical reduction of algorithmic complexity has been demonstrated. This algorithm enables the extensive use of MCMC in applications that require high dimensional models, see e.g., [8].

7. REFERENCES

[1] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC," *IEEE Trans. Image Processing*, vol. 47, no. 10, pp. 456–463, Oct. 1999.

[2] M. Davy and S. Godsill, "Bayesian Harmonic Models for Musical Signal Analysis," in *Seventh Valencia International meeting Bayesian statistics 7*, Tenerife ,Spain, June 2002.

[3] C. Andrieu, E. Barat, and A. Doucet, "Bayesian Deconvolution of Noisy Filtered Point Processes," *IEEE Trans. Signal Processing*, vol. 49, no. 1, Jan. 2001.

[4] Q. Cheng, R. Chen, and T.-H. Li, "Simultaneous wavelet estimation and deconvolution of reflection seismic signals," *IEEE Trans. Geosci. Remote Sensing*, vol. 34, pp. 377–384, Mar. 1996.

[5] M. Davy, C. Doncarli, and J. Y. Tourneret, "Classification of Chirp Signals Using Hierarchical Bayesian Learning and MCMC Methods," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 377–388, Feb. 2002.

[6] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York: Springer, 2000.

[7] J.-F. Giovannelli, J. Idier, R. Boubertakh, and A. Herment, "Unsupervised frequency tracking beyond the Nyquist frequency using Markov chains," *IEEE Trans. Signal Processing*, vol. 50, no. 12, pp. 2906–2914, Dec. 2002.

[8] M. Davy, S. Godsill, and J. Idier, "Bayesian Estimation and Analysis of Harmonic Music," IRCCyN, CNRS, France, Tech. Rep., Dec. 2003.