ELSEVIER

Signal Processing 81 (2001) 19–37

SIGNAL PROCESSING

www.elsevier.nl/locate/sigpro

Model selection by MCMC computation

C. Andrieu[a,1,*], P.M. Djurić[b,2], A. Doucet[a,3]

[a]Engineering Department, Cambridge University, Trumpington Street, Cambridge CB2 1PZ, UK
[b]Department of Electrical and Computer Engineering, Stony Brook State University of New York, NY, USA

Received 11 May 1999; received in revised form 20 January 2000

Abstract

MCMC sampling is a methodology that is becoming increasingly important in statistical signal processing. It has been of particular importance to the Bayesian-based approaches to signal processing since it extends significantly the range of problems that they can address. MCMC techniques generate samples from desired distributions by embedding them as limiting distributions of Markov chains. There are many ways of categorizing MCMC methods, but the simplest one is to classify them in one of two groups: the first is used in estimation problems where the unknowns are typically parameters of a model, which is assumed to have generated the observed data; the second is employed in more general scenarios where the unknowns are not only model parameters, but models as well. In this paper, we address the MCMC methods from the second group, which allow for generation of samples from probability distributions defined on unions of disjoint spaces of different dimensions. More specifically, we show why sampling from such distributions is a nontrivial task. It will be demonstrated that these methods genuinely unify the operations of detection and estimation and thereby provide great potential for various important applications. The focus is mainly on the reversible jump MCMC (Green, Biometrika 82 (1995) 711), but other approaches are also discussed. Details of implementation of the reversible jump MCMC are provided for two examples. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Bayesian model selection; Markov chain Monte Carlo methods; Reversible jump MCMC

1. Introduction

In many signal processing problems the unknowns of interest are the model that generates the observed data and its parameters. In a practical setting, there are more than one candidate models for the data, and the main objective is to choose the best of them according to a predefined criterion. The models have their own parameters, which in general may not be related at all, and often they, too, have to be estimated. A coherent approach of finding the best model is based on the Bayesian methodology. Although the theory of the procedure is fairly simple, its practical implementation is plagued with several nontrivial difficulties. One of them is the multidimensional integrations needed to obtain the marginal posterior probabilities of the models.

The integration problem has most commonly been alleviated by applying asymptotic approximations and invoking the Laplace's method for integration. Quite often, however, the so obtained

[*] Corresponding author.

[1] Supported by AT&T Laboratories, Cambridge, UK.

[2] Supported by the National Science Foundation under Award No. MIP-956743.

[3] Supported by EPSRC, UK.

PII: S0165-1684(00)00188-2

Nomenclature

$\mathbb{R}$	set of real numbers
$\mathbb{R}^+$	set of positive real numbers
$\bigotimes_{n=1}^{k}\mathscr{A}_n$	set with elements $(a_1, a_2, \ldots, a_n)$ where $a_j \in \mathscr{A}_j$ for $j = 1, \ldots, n$
$[\boldsymbol{A}]_{i,j}$	ith row, jth column of matrix $\boldsymbol{A}$
$\vert\boldsymbol{A}\vert$	determinant of matrix $\boldsymbol{A}$
$\boldsymbol{A}^{\mathrm{T}}$	matrix $\boldsymbol{A}$ transposed
	$\mathbf{z} \triangleq (z_1, \ldots, z_{j-1}, z_j, z_{j+1}, \ldots, z_k)^{\mathrm{T}}, \boldsymbol{z}_{-j} \triangleq (z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_k)^{\mathrm{T}}$.
$\mathbf{0}_{n\times p}$	null matrix of dimension $n \times p$
$\boldsymbol{I}_n$	identity matrix of dimension $n \times n$
$\mathbb{1}_E(\boldsymbol{z})$	indicator function of the set E(1 if $\boldsymbol{z} \in E$, 0 otherwise).
$\delta_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{z})$	delta Dirac measure such that $\int_A \delta_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{z}) = 1$ if $\boldsymbol{x} \in A$ and 0 otherwise
$\lfloor z \rfloor$	highest integer strictly less than z
$\boldsymbol{z} \sim \pi(\boldsymbol{z})$	$\boldsymbol{z}$ is distributed according to $\pi(\boldsymbol{z})$
$\boldsymbol{z}\vert\boldsymbol{y} \sim \pi(\boldsymbol{z})$	the conditional distribution of $\boldsymbol{z}$ given $\boldsymbol{y}$ is $\pi(\boldsymbol{z})$

Probability distribution	$\mathscr{F}$	$f_{\mathscr{F}}(\cdot)$
Inverse Gamma	$\mathscr{IG}(\alpha,\beta)$	$(\beta^{\alpha}/\Gamma(\alpha))z^{-\alpha-1}\exp(-\beta/z)\mathbb{1}_{[0,+\infty)}(z)$
Gamma	$\mathscr{G}a(\alpha,\beta)$	$(\beta^{\alpha}/\Gamma(\alpha))z^{\alpha-1}\exp(-\beta z)\mathbb{1}_{[0,+\infty)}(z)$
Gaussian	$\mathscr{N}(\boldsymbol{m},\Sigma)$	$\vert 2\pi\Sigma\vert^{-1/2}\exp(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{m})^{T}\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{m}))$
Uniform	$\mathscr{U}_A$	$[\int_A \mathrm{d}\boldsymbol{z}]^{-1}\mathbb{1}_A(\boldsymbol{z})$

approximations may be relatively poor, which altogether lead to incorrect results. Recently, very interesting alternatives have emerged, most of them based on the Markov chain Monte Carlo (MCMC) sampling technique. Some of them can be viewed as generalized MCMC schemes where the random walks within the parameter spaces of the standard MCMC computations are extended to include walks that allow for jumps between parameter spaces of different dimensions.

MCMC sampling was initially proposed as a method for drawing samples from nonconventional densities, and so it was restricted to problems where the densities had fixed dimensionalities. For this reason, in the 1980s and early 1990s most of the efforts in the field were focused on applications related to estimation. Good reviews of the method can be found in [20,21,23,36], and in signal processing context, in [2,31]. Additional references on the subject can be found in [13].

In 1994, Grenander and Miller proposed a method with the same objective, to draw samples from a desired (target) distribution, but with an important difference in that the dimensionality of the parameter space was not fixed [26]. In other words, the target distribution was defined over the joint sample space of the models and their parameters. With this, they generalized the sampling-based methodology in an important way, which in fact allows for simultaneous detection and estimation of signals. Their approach was based on jump-diffusion dynamics with the essential features that at random times the process jumps between parameter spaces corresponding to different models and of different dimensions, and once within a model of fixed dimensionality, it follows a Langevin diffusion [26]. Earlier work had also been done in the field based on the construction of a continuous-time Markov birth–death process as described by Preston [33]. Ripley first applied this idea to the simulation of point processes [35].

In signal processing, the theory of random sampling based on jump-diffusion processes has found interesting applications. Besides in target tracking applications, it has been used to resolve problems related to estimation of directions of arrival of signals from an unknown number of moving signal sources [30]. This work has been extended to a more general setting to include automated tracking and recognition of moving targets by fusion of multi-sensor data, specifically, narrow sensor array outputs and high-resolution images [30]. Additional contributions in the same application area have been reported in [28,29,40].

In 1995, Green showed that the MCMC methodology can be generalized in order to cope with problems where the data generating models are also unknown [27]. His approach represents a flexible framework for construction of reversible Markov chain samplers, where moves between different parameter spaces are also allowed. The Markov chain, thus, not only moves within parameter spaces that correspond to particular models, but it also jumps between them, their difference in dimensionality notwithstanding. A nice feature of the approach is that it is a generalization of the standard MCMC theory.

Since the appearance of [27], there have been many efforts to apply the methodology to solving various model selection problems. In signal processing some topics that have been examined are the joint detection and estimation of harmonic signals [1,16], model order determination of autoregressive processes [43], restoration of distorted autoregressive signals [44], variable selection [14], parametric modeling and estimation of time-varying spectra by chirps with Gaussian envelopes [15], and unsupervised image segmentation [5,10].

Also in 1995, Carlin and Chib proposed a third approach to model selection [8]. Their idea is based on the use of conventional Markov chains and the concept of a supermodel. Namely, the supermodel is defined over a composite parameter space, which in fact is the product space of all model parameters. The implementation of their method requires the use of pseudopriors, that is, linking densities with no physical meaning, and standard MCMC methodology.

In this paper, the reversible jump MCMC and Carlin and Chib's algorithms are described in detail, and a brief exposition on the jump-diffusion methodology is provided. MCMC solutions to several typical signal processing problems are also given. They include the detection of harmonics embedded in noise and deconvolution of impulsive processes. In particular, the paper is organized as follows: in Section 2 we introduce the problem and give three illustrative examples. In Section 3 we provide a description of MCMC samplers for model selection. In Section 4 we present the applications of these samplers. Readers only interested in practical implementation of the algorithms can skip the sections and subsections marked with a ◆. The other sections and subsections are more methodological and aim at giving a detailed description of the construction of MCMC samplers for model selection, pointing out the arising "theoretical" problems.

In this paper we assume that the reader is familiar with Bayesian statistics and MCMC algorithms for fixed model problems, i.e. the data augmentation algorithm, Gibbs sampler and Metropolis-Hastings (MH) algorithm. Newcomers should refer to [19] in this special issue, and reviews such as [23,36,38,42].

2. Problem formulation

We assume that data $\boldsymbol{y}_{1:T}$ can be described by a model that belongs to a family of models $(\mathscr{M}_n)_{n=1,\ldots,N}$ which might be defined on different spaces and where N may be infinite. We therefore consider a parameter space consisting of a union of possibly heterogeneous subspaces. We present here three example that are of importance in signal processing: detection of the number of sinusoids in noise, choice of modelling between autoregressive and harmonic process and Bernoulli Gaussian estimation [17].

Example 1 (*Nested models*). We want to model the data $\boldsymbol{y}_{1:T}$ with one of the following models:

$$\begin{aligned}\mathscr{M}_0\colon\quad & y_t = w_{0,t}, \quad k=0,\\ \mathscr{M}_k\colon\quad & y_t = \sum_{j=1}^{k} (a_{c_{j,k}} \cos[\omega_{j,k} t] + a_{s_{j,k}} \sin[\omega_{j,k} t])\\ & \qquad + w_{k,t}, \quad k \geqslant 1,\end{aligned} \tag{1}$$

where for a given k, $w_{k,t} \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma_k^2)$. The model $\mathcal{M}_k$ describes the data in terms of k sinusoids in white Gaussian noise. The unknown parameters for $\mathcal{M}_k$ are $\boldsymbol{\theta}_k = (a_{c_{1,k}}, a_{s_{1,k}}, \omega_{1,k}, \ldots, a_{c_{k,k}}, a_{s_{k,k}}, \omega_{k,k}, \sigma_k^2)$. Bayesian inference is performed on the parameter space $\boldsymbol{\Theta} = \bigcup_{n=0}^{k_{\max}} \{n\} \times \boldsymbol{\Theta}_n$ where $\boldsymbol{\Theta}_n = (\mathbb{R}^2 \times (0,\pi))^n \times \mathbb{R}^+$, i.e. if $k = n$, the unknown parameters $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_n$. The space $\boldsymbol{\Theta}$ is a union of disjoint spaces. In this case, one says that the models are nested as $\boldsymbol{\Theta}_{n+1} = \mathbb{R}^2 \times (0,\pi) \times \boldsymbol{\Theta}_n$ for $n = 0, \ldots, k_{\max}$.

Example 2 (*Models of different natures*). The data $\boldsymbol{y}_{1:T}$ can be best represented by one of the following two models:

- $\mathcal{M}_1$, an AR process of order k_{AR}, k_{AR} being fixed, excited by white Gaussian noise $v_t \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma_{k_{\text{AR}}}^2)$,

$$y_t = \sum_{i=1}^{k_{\text{AR}}} a_i y_{t-i} + v_t. \tag{2}$$

- $\mathcal{M}_2$, $k_{\sin}$ sinusoids, $k_{\sin}$ being fixed, embedded in a white Gaussian noise sequence $w_t \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma_{k_{\sin}}^2)$, i.e.

$$y_t = \sum_{j=1}^{k_{\sin}} (a_{c_j} \cos[\omega_j t] + a_{s_j} \sin[\omega_j t]) + w_t. \tag{3}$$

Consequently the parameter space for models 1 and 2 is $\boldsymbol{\Theta} = \{1\} \times \boldsymbol{\Theta}_1 \cup \{2\} \times \boldsymbol{\Theta}_2$, so it is also a union of disjoint spaces. Either the data represent an AR process of order k_{AR} with unknown parameters $(\boldsymbol{a}_{k_{\text{AR}}}, \sigma_{k_{\text{AR}}}^2)$ in $\boldsymbol{\Theta}_1 = \mathbb{R}^{k_{\text{AR}}} \times \mathbb{R}^+$, or they are $k_{\sin}$ sinusoids in noise with unknown parameters $(a_{c_1}, a_{s_1}, \omega_1, \ldots, a_{c_{k_{\sin}}}, a_{s_{k_{\sin}}}, \omega_{k_{\sin}}, \sigma_k^2)$ in $\boldsymbol{\Theta}_2 = (\mathbb{R}^2 \times (0,\pi))^{k_{\sin}} \times \mathbb{R}^+$.

Example 3 (*Variable selection: Bernoulli–Gauss (BG) problem*). We assume that the underlying process x_t is an AR process of known dimension k_{AR} excited by a BG sequence. The process of interest x_t is observed in an additive white Gaussian noise w_t and we observe y_t. More precisely,

$$x_t = \sum_{i=1}^{k_{\text{AR}}} a_i x_{t-i} + v_t, \tag{4}$$

$$y_t = x_t + w_t, \tag{5}$$

where $w_t \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$, and v_t is an iid sequence which takes the value 0 with probability $(1 - \lambda)$ or is drawn from a Gaussian distribution of variance $\sigma_0^2 > 0$ with probability λ, i.e.[4]

$$\pi(\mathrm{d}v_t) = \lambda \mathcal{N}(0,\sigma_0^2)\,\mathrm{d}v_t + (1 - \lambda)\,\delta_0(\mathrm{d}v_t). \tag{6}$$

From an algorithmic point of view, it is convenient to introduce a missing Bernoulli sequence $\boldsymbol{r}_{1:T}$, such that

$$\begin{aligned} \pi(\mathrm{d}v_t | (r_t = 0)) &= \delta_0(\mathrm{d}v_t), \\ \pi(\mathrm{d}v_t | (r_t = 1)) &= \mathcal{N}(0,\sigma_0^2)\,\mathrm{d}v_t. \end{aligned} \tag{7}$$

In this model, $(a_1, \ldots, a_{k_{\text{AR}}})$, σ^2, λ and σ_0^2 are assumed known, and the sequence $\boldsymbol{r}_{1:T}$ is unknown. Clearly, there are 2^T possible sequences $\boldsymbol{r}_{1:T} \in \{0,1\}^T$ and we order them according to $\boldsymbol{r}_{1:T}(n), n = 0, \ldots, 2^T - 1$, such that $n = \sum_{t=0}^{T-1} 2^t r_t(n)$. Here the objective is to perform Bayesian inference of the unknowns $(\boldsymbol{r}_{1:T}, \boldsymbol{v}_{1:T})$ on the space

$$\boldsymbol{\Theta} = \bigcup_{n=0}^{2^T - 1} \{n\} \times \bigotimes_{t=1,\ldots,T} (\{0\} \times \{0\})^{1 - r_t(n)} \cup (\{1\} \times \mathbb{R})^{r_t(n)} \tag{8}$$

with the notational convention that $(A)^1 = A$ and $(A)^0 = \emptyset$ for a set A. Hence for each sequence the unknown parameters $(\boldsymbol{r}_{1:T}, \boldsymbol{v}_{1:T})$ lie in a different subspace. If $r_t = 0$ then $v_t = 0$ and thus $(r_t, v_t) \in \{0\} \times \{0\}$. If $r_t = 1$ then $v_t \sim \mathcal{N}(0,\sigma_0^2)$ and thus $(r_t, v_t) \in \{1\} \times \mathbb{R}$. This definition might seem tedious, but it clearly points out that the prior and posterior probability distributions of $(\boldsymbol{r}_{1:T}, \boldsymbol{v}_{1:T})$ are defined on 2^T distinct subspaces.

In the general case the probability distribution will be defined on a space of the form $\boldsymbol{\Theta} \triangleq \bigcup_{n=1}^{N} \{n\} \times \boldsymbol{\Theta}_n$ and can be written for any $A \in \mathscr{B}(\boldsymbol{\Theta})$

$$\begin{aligned} \int_A \pi(k,\mathrm{d}\boldsymbol{\theta}) &= \Pr((k,\mathrm{d}\boldsymbol{\theta}) \in A) \\ &= \int_A \sum_{n=1}^{N} \pi_n(n,\mathrm{d}\boldsymbol{\theta}_n) \mathbb{I}_{\{n\} \times \boldsymbol{\Theta}_n}(k,\boldsymbol{\theta}), \end{aligned} \tag{9}$$

[4] We adopt the standard notation $\pi(\mathrm{d}\boldsymbol{\theta})$ for the probability of a small set around $\boldsymbol{\theta}$ and we will denote with $\pi(\boldsymbol{\theta})$ the associated density function, defined with respect to a proper measure.

where

$$\mathbb{I}_{\{n\}\times\boldsymbol{\Theta}_n}(k,\boldsymbol{\theta}) = \begin{cases} 1, & (k,\boldsymbol{\theta})\in\{n\}\times\boldsymbol{\Theta}_n, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Note that $(k,\boldsymbol{\theta})$ is in one of the spaces $\{n\}\times\boldsymbol{\Theta}_n$, and the probability of k equal to n and $\boldsymbol{\theta}$ being in an infinitesimal set centered around $\boldsymbol{\theta}_n$ is $\pi_n(n,\mathrm{d}\boldsymbol{\theta}_n)$. Thus, the probability that $\boldsymbol{\theta}$ is in the subspace $\boldsymbol{\Theta}_n$ is

$$\pi_n(n) = \int_{\boldsymbol{\Theta}_n} \pi_n(n,\mathrm{d}\boldsymbol{\theta}_n). \tag{11}$$

Conditional upon n, the parameters $\boldsymbol{\theta}$ are distributed according to the probability distribution $\int_A \pi_n(\mathrm{d}\boldsymbol{\theta}_n|n) = \int_A \pi_n(n,\mathrm{d}\boldsymbol{\theta}_n)/\pi_n(n)$.

Remark 1. Actually, one could write

$$\int_A \pi(k,\mathrm{d}\boldsymbol{\theta}) = \int_A \sum_{n=1}^{N} \pi_n(n,\mathrm{d}\boldsymbol{\theta})\mathbb{I}_{\{n\}\times\boldsymbol{\Theta}_n}(k,\boldsymbol{\theta}),$$

that is, drop the subscript n from $\boldsymbol{\theta}$ when $\boldsymbol{\theta}\in\boldsymbol{\Theta}_n$. We add this index to improve readability.

Again, the number N of possible models can either be finite or infinite; in either case the following condition must naturally hold:

$$\sum_{n=1}^{N} \pi_n(n) = \sum_{n=1}^{N} \int_{\boldsymbol{\Theta}_n} \pi_n(n,\mathrm{d}\boldsymbol{\theta}_n) = 1. \tag{12}$$

In order to perform Bayesian model selection, one is interested in evaluating quantities such as the model probabilities $\pi_n(n)$. There is usually no closed-form analytical expression for such quantities, so one has to resort to numerical methods. In this scenario, MCMC has proved to be an efficient means for solving this kind of integration problems [27,34]. Our aim is now to define and construct MCMC algorithms in order to obtain samples from the distribution $\pi(k,\mathrm{d}\boldsymbol{\theta})$. A straightforward solution would be to run N independent fixed model Markov chains (where N must be finite), one for each distribution, subsequently compare the quantities $\pi_n(n)$, and finally choose among the different models. Such a strategy, however, would not take advantage of possible relations between parameters from different spaces. An MCMC sampler that could take advantage of such relations would be of great interest, particularly in the case of nested models (see Example 1).

3. MCMC algorithms for model selection

In this section, we describe how to build MCMC algorithms for model selection, i.e. how to construct ergodic Markov chains admitting $\pi(k,\mathrm{d}\boldsymbol{\theta})$ as their invariant distribution. In the case of model selection, the main difficulty for the Markov chain is to be able to jump from one subspace $\boldsymbol{\Theta}_n$ to another subspace $\boldsymbol{\Theta}_m$. Green [27] has developed a general methodology that addresses this problem. Our aim is to present this methodology, progressively, starting in Section 3.1 with a very simple case where $N = 2$, first when there is no measure theoretic problem (see Section 3.1.2) and then when this kind of problem arises (see Section 3.1.3). Then in Section 3.2 we show how to extend the algorithm for $N \geqslant 2$. In Section 3.3, we describe Carlin and Chib's algorithm and give a brief exposition of the jump-diffusion methodology.

3.1. The case $N = 2$

3.1.1. Goals

We want to sample from the distribution

$$\int_A \pi(k,\mathrm{d}\boldsymbol{\theta}) = \int_A \pi_1(1,\mathrm{d}\boldsymbol{\theta}_1)\mathbb{I}_{\{1\}\times\boldsymbol{\Theta}_1}(k,\boldsymbol{\theta}) + \int_A \pi_2(2,\mathrm{d}\boldsymbol{\theta}_2)\mathbb{I}_{\{2\}\times\boldsymbol{\Theta}_2}(k,\boldsymbol{\theta}) \tag{13}$$

defined on $\boldsymbol{\Theta} \triangleq \{1\}\times\boldsymbol{\Theta}_1 \cup \{2\}\times\boldsymbol{\Theta}_2$ where $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ may be disjoint. To this aim, we need to construct a Markov transition kernel $K(k^*,\mathrm{d}\boldsymbol{\theta}^*\,|\,k,\boldsymbol{\theta})$ admitting $\pi(k,\mathrm{d}\boldsymbol{\theta})$ as invariant distribution

$$\int_A \int_{\boldsymbol{\Theta}} K(k^*,\mathrm{d}\boldsymbol{\theta}^*\,|\,k,\boldsymbol{\theta})\pi(k,\mathrm{d}\boldsymbol{\theta}) = \int_A \pi(k^*,\mathrm{d}\boldsymbol{\theta}^*), \tag{14}$$

i.e. if the current state of the Markov chain is distributed according to $\pi(\cdot)$ then, after one iteration via the transition kernel $K(\cdot\,|\,\cdot)$, the new state is also distributed according to $\pi(\cdot)$. $\int_A K(k^*,\mathrm{d}\boldsymbol{\theta}^*\,|\,k,\boldsymbol{\theta})$ is the probability of being in a set $A\in\mathscr{B}(\boldsymbol{\Theta})$ when starting from $(k,\boldsymbol{\theta})\in\boldsymbol{\Theta}$.

A sufficient condition for a transition kernel $K(k^*,\mathrm{d}\boldsymbol{\theta}^*\,|\,k,\boldsymbol{\theta})$ to admit $\pi(k,\mathrm{d}\boldsymbol{\theta})$ as invariant distribution is the detailed balance, or reversibility, condition [23]:

$$\int_A \pi(k,\mathrm{d}\boldsymbol{\theta})\int_B K(k^*,\mathrm{d}\boldsymbol{\theta}^* \mid k,\boldsymbol{\theta}) = \int_B \pi(k^*,\mathrm{d}\boldsymbol{\theta}^*)\int_A K(k,\mathrm{d}\boldsymbol{\theta} \mid k^*,\boldsymbol{\theta}^*), \tag{15}$$

i.e. the probability of being in any set A and jumping into any set B is equal to the probability of being in a set B and jumping into any set A when the chain is in its stationary regime. It is trivial to show that Eq. (15) implies Eq. (14).

When the current state of the Markov chain is $(k,\boldsymbol{\theta})$, the principle of the MH algorithm is to propose a candidate $\boldsymbol{\theta}^*$ in $\boldsymbol{\Theta}_1 \cup \boldsymbol{\Theta}_2$ according to a proposal distribution $q(\mathrm{d}\boldsymbol{\theta}^* \mid \boldsymbol{\theta})$.[5] Then this candidate is accepted or rejected according to an acceptance probability $\alpha((k,\boldsymbol{\theta}),(k^*,\boldsymbol{\theta}^*)) = \min\{1, r((k,\boldsymbol{\theta}),(k^*,\boldsymbol{\theta}^*))\}$, which ensures reversibility of the transition kernel with respect to $\pi(\cdot)$.

By analogy with the classical (fixed dimension) case, a satisfactory expression for $\alpha((k,\boldsymbol{\theta}) \in A, (k^*,\boldsymbol{\theta}^*) \in B)$ is the following:

$$r((k,\boldsymbol{\theta}) \in A,(k^*,\boldsymbol{\theta}^*) \in B) = \frac{\int_B \pi(k^*,\mathrm{d}\boldsymbol{\theta}^*)}{\int_A \pi(k,\mathrm{d}\boldsymbol{\theta})}\frac{\int_A q(\mathrm{d}\boldsymbol{\theta} \mid \boldsymbol{\theta}^*)}{\int_B q(\mathrm{d}\boldsymbol{\theta}^* \mid \boldsymbol{\theta})}. \tag{16}$$

This expression is the probability version of the standard MH acceptance ratio [27]. To be of practical interest, as one wants to work with points of the space $\boldsymbol{\Theta}$ and not subsets, it requires considering the limit of this ratio when the sets A and B collapse around the points $(k^*,\boldsymbol{\theta}^*)$ and $(k,\boldsymbol{\theta})$. This ratio is not necessarily defined. The existence and evaluation of the limit of the ratio in (16) are the main difficulties in generalizing the MCMC method. These are the main topics discussed in the following subsections.

3.1.2. Jumping between $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$: simple case◆

When the current state of the Markov chain is $(n,\boldsymbol{\theta}_n)$ two events can occur: either the chain stays in $\{n\} \times \boldsymbol{\Theta}_n$, or it moves to the other subspace. The design of MCMC updates within a subspace is based on standard methods such as the Gibbs sampler or the MH algorithm. So, for the sake of clarity, in this subsection we discuss only the case where we propose moves from one subspace to the other, i.e. $n \neq m$. Then,[6]

$$q(\mathrm{d}\boldsymbol{\theta}^* \mid \boldsymbol{\theta}) = q_{1,2}(\mathrm{d}\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1)\mathbb{I}_{\boldsymbol{\Theta}_2 \times \boldsymbol{\Theta}_1}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) + q_{2,1}(\mathrm{d}\boldsymbol{\theta}_1^* \mid \boldsymbol{\theta}_2)\mathbb{I}_{\boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2}(\boldsymbol{\theta}^*,\boldsymbol{\theta}), \tag{17}$$

i.e., when $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1$, a value $\boldsymbol{\theta}^* = \boldsymbol{\theta}_2^* \in \boldsymbol{\Theta}_2$ is proposed according to the distribution $q_{1,2}(\mathrm{d}\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1)$ and, when $\boldsymbol{\theta} = \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_2$, a value $\boldsymbol{\theta}^* = \boldsymbol{\theta}_1^* \in \boldsymbol{\Theta}_1$ is proposed according to the distribution $q_{2,1}(\mathrm{d}\boldsymbol{\theta}_1^* \mid \boldsymbol{\theta}_2)$.

Let us assume that the current state of the Markov chain at iteration i of our MCMC sampler is $(k^{(i)},\boldsymbol{\theta}_{k^{(i)}}^{(i)}) = (n,\boldsymbol{\theta}_n) \in \{n\} \times \boldsymbol{\Theta}_n$. We propose to jump to the state $(m,\boldsymbol{\theta}_m^*) \in \{m\} \times \boldsymbol{\Theta}_m$ using the proposal distribution $q_{n,m}\ (\mathrm{d}\boldsymbol{\theta}_m^* \mid \boldsymbol{\theta}_n)$. With probability $\alpha((n,\boldsymbol{\theta}_n),(m,\boldsymbol{\theta}_m^*))$ we set $(k^{(i+1)},\boldsymbol{\theta}_{k^{(i+1)}}^{(i+1)}) = (m,\boldsymbol{\theta}_m^*)$, otherwise $(k^{(i+1)},\boldsymbol{\theta}_{k^{(i+1)}}^{(i+1)}) = (k^{(i)},\boldsymbol{\theta}_{k^{(i)}}^{(i)})$ where the acceptance ratio is the limit of the ratio when the sets $\mathrm{d}\boldsymbol{\theta}_n$ and $\mathrm{d}\boldsymbol{\theta}_m^*$ collapse around $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_m^*$

$$r((n,\boldsymbol{\theta}_n),(m,\boldsymbol{\theta}_m^*)) = \frac{\pi_m(m,\mathrm{d}\boldsymbol{\theta}_m^*)q_{m,n}(\mathrm{d}\boldsymbol{\theta}_n \mid \boldsymbol{\theta}_m^*)}{\pi_n(n,\mathrm{d}\boldsymbol{\theta}_n)q_{n,m}(\mathrm{d}\boldsymbol{\theta}_m^* \mid \boldsymbol{\theta}_n)} \tag{18}$$

according to Eq. (16). The acceptance probability is $\alpha((n,\boldsymbol{\theta}_n),(m,\boldsymbol{\theta}_m^*)) = \min\{1, r((n,\boldsymbol{\theta}_n),(m,\boldsymbol{\theta}_m^*))\}$.

If we assume that for $(a,b) \in \{1,2\}$

$$\begin{aligned} \pi_a(a,\mathrm{d}\boldsymbol{\theta}_a) &= \pi_a(a,\boldsymbol{\theta}_a)\mu_a(\mathrm{d}\boldsymbol{\theta}_a), \\ q_{a,b}(\mathrm{d}\boldsymbol{\theta}_b \mid \boldsymbol{\theta}_a) &= q_{a,b}(\boldsymbol{\theta}_b \mid \boldsymbol{\theta}_a)\mu_b(\mathrm{d}\boldsymbol{\theta}_b), \end{aligned} \tag{19}$$

i.e. $\pi_a(a,\mathrm{d}\boldsymbol{\theta}_a)$ and $q_{a,b}(\mathrm{d}\boldsymbol{\theta}_b \mid \boldsymbol{\theta}_a)$ admit, respectively, a density with respect to $\mu_a(\mathrm{d}\boldsymbol{\theta}_a)$ and $\mu_b(\mathrm{d}\boldsymbol{\theta}_b)$, then

$$\begin{aligned} &\frac{\pi_m(m,\mathrm{d}\boldsymbol{\theta}_m^*)q_{m,n}(\mathrm{d}\boldsymbol{\theta}_n \mid \boldsymbol{\theta}_m^*)}{\pi_n(n,\mathrm{d}\boldsymbol{\theta}_n)q_{n,m}(\mathrm{d}\boldsymbol{\theta}_m^* \mid \boldsymbol{\theta}_n)} \\ &\quad = \frac{\pi_m(m,\boldsymbol{\theta}_m^*)\mu_m(\mathrm{d}\boldsymbol{\theta}_m^*)q_{m,n}(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}_m^*)\mu_n(\mathrm{d}\boldsymbol{\theta}_n)}{\pi_n(n,\boldsymbol{\theta}_n)\mu_n(\mathrm{d}\boldsymbol{\theta}_n)q_{n,m}(\boldsymbol{\theta}_m^* \mid \boldsymbol{\theta}_n)\mu_m(\mathrm{d}\boldsymbol{\theta}_m^*)} \\ &\quad = \frac{\pi_m(m,\boldsymbol{\theta}_m^*)q_{m,n}(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}_m^*)}{\pi_n(n,\boldsymbol{\theta}_n)q_{n,m}(\boldsymbol{\theta}_m^* \mid \boldsymbol{\theta}_n)}, \end{aligned} \tag{20}$$

[5] To simplify the notation, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ define implicitly the index of the subspaces to which they belong, i.e. k and k^*.

[6] From now on we do not mention any set A and integral of the type $\int_A p(\mathrm{d}\boldsymbol{\theta})$ and simply use the simpler notation $p(\mathrm{d}\boldsymbol{\theta})$.

i.e. the ratio of Eq. (18) exists and can be computed straightforwardly using the densities. This is for example the case when $\mu_n(\mathrm{d}\boldsymbol{\theta}_n)$ and $\mu_m(\mathrm{d}\boldsymbol{\theta}_m)$ are Lebesgue measures. The assumptions we made to ensure the existence and compute the ratio of probability measures might be restrictive in some applications. In the next subsection, we consider a more general case.

Remark 2. In the case of a standard MH algorithm where $\boldsymbol{\Theta} \subset \mathbb{R}^d$, a very common case is when $\pi(\mathrm{d}\boldsymbol{\theta})$, $q(\mathrm{d}\boldsymbol{\theta} \mid \boldsymbol{\theta}^*)$ and $q(\mathrm{d}\boldsymbol{\theta}^* \mid \boldsymbol{\theta})$ admit densities $\pi(\boldsymbol{\theta})$, $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*)$ and $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta})$ with respect to the Lebesgue measure $\mathrm{d}\boldsymbol{\theta}$. Then the acceptance ratio takes the conventional form

$$r(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta})}. \tag{21}$$

3.1.3. Jumping between $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$: complicated case◆

We start with an example. Let us consider the case where the two sets are nested, i.e. for example $\boldsymbol{\Theta}_2 = \boldsymbol{\Theta}_1 \times \boldsymbol{\Psi}_{1,2}$ and one wants to jump from $\boldsymbol{\Theta}_1$ to $\boldsymbol{\Theta}_2$. In many cases one wishes to link the current state $\boldsymbol{\theta}_1$ and the candidate state $\boldsymbol{\theta}_2^*$ by exploiting that the current state in $\boldsymbol{\Theta}_1$ is $\boldsymbol{\theta}_1$ deterministically. For instance, one could add an extra component $\boldsymbol{\varphi}_{1,2} \in \boldsymbol{\Psi}_{1,2}$ to $\boldsymbol{\theta}_1$ to propose $\boldsymbol{\theta}_2^*$ as follows:

$$\boldsymbol{\theta}_2^* = (\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}). \tag{22}$$

This is reasonable in the case of two nested models: one wants to keep as much information as possible when moving from parameter space $\boldsymbol{\Theta}_1$ to $\boldsymbol{\Theta}_2$ and vice versa (note that otherwise one could have used two different Markov chains, one within each parameter space). We address this task in Section 4.1.1 in the context of sinusoids corrupted by noise. Let us assume that we are in the subspace with k sinusoids and we want to jump to a subspace with $k+1$ sinusoids. Then, a sensible move is to propose a new frequency $\boldsymbol{\varphi}_{1,2}$ according to a proposal distribution while keeping the current values of the other frequencies. The reverse move, i.e. the move to jump from $\boldsymbol{\Theta}_2$ to $\boldsymbol{\Theta}_1$, is automatically defined and consists of removing the appropriate components of $\boldsymbol{\theta}_2^*$. The acceptance ratio of the move from $\boldsymbol{\Theta}_1$ to $\boldsymbol{\Theta}_2$ is still equal to

$$r((1,\boldsymbol{\theta}_1),(2,\boldsymbol{\theta}_2^*)) = \frac{\pi_2(2,\mathrm{d}\boldsymbol{\theta}_2^*)q_{2,1}(\mathrm{d}\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^*)}{\pi_1(1,\mathrm{d}\boldsymbol{\theta}_1)q_{1,2}(\mathrm{d}\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1)}. \tag{23}$$

Reparametrizing in terms of $(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2})$, one obtains for the numerator

$$\begin{aligned}&\pi_2(2,\mathrm{d}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}))q_{2,1}(\mathrm{d}\boldsymbol{\theta}_1 \mid (\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}))\\ &\quad= \pi_2(2,\mathrm{d}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}))\delta_{\boldsymbol{\theta}_1}(\mathrm{d}\boldsymbol{\theta}_1)\end{aligned} \tag{24}$$

while the denominator becomes

$$\begin{aligned}&\pi_1(1,\mathrm{d}\boldsymbol{\theta}_1)q_{1,2}(\mathrm{d}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}) \mid \boldsymbol{\theta}_1)\\ &\quad= \pi_1(1,\mathrm{d}\boldsymbol{\theta}_1)q_{1,2}(\mathrm{d}\boldsymbol{\varphi}_{1,2} \mid \boldsymbol{\theta}_1)q_{1,2}(\mathrm{d}\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_1)\\ &\quad= \pi_1(1,\mathrm{d}\boldsymbol{\theta}_1)q_{1,2}(\mathrm{d}\boldsymbol{\varphi}_{1,2} \mid \boldsymbol{\theta}_1)\delta_{\boldsymbol{\theta}_1}(\mathrm{d}\boldsymbol{\theta}_1)\end{aligned} \tag{25}$$

as only $\boldsymbol{\varphi}_{1,2}$ is sampled, whereas $\boldsymbol{\theta}_1$ is kept fixed.

If we now assume that

$$\begin{aligned}&\pi_a(a,\mathrm{d}\boldsymbol{\theta}_a) = \pi_a(a,\boldsymbol{\theta}_a)\mu_a(\mathrm{d}\boldsymbol{\theta}_a),\\ &q_{a,b}(\mathrm{d}\boldsymbol{\varphi}_{a,b} \mid \boldsymbol{\theta}_a) = q_{a,b}(\boldsymbol{\varphi}_{a,b} \mid \boldsymbol{\theta}_a)\bar{\mu}_a(\mathrm{d}\boldsymbol{\varphi}_{a,b}),\end{aligned} \tag{26}$$

then the acceptance ratio satisfies

$$\begin{aligned}&r((1,\boldsymbol{\theta}_1),(2,\boldsymbol{\theta}_2^*))\\ &\quad= \frac{\pi_2(2,(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}))\mu_2(\mathrm{d}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}))}{\pi_1(1,\boldsymbol{\theta}_1)q_{1,2}(\boldsymbol{\varphi}_{1,2} \mid \boldsymbol{\theta}_1)\mu_1(\mathrm{d}\boldsymbol{\theta}_1)\bar{\mu}_1(\mathrm{d}\boldsymbol{\varphi}_{1,2})}\end{aligned} \tag{27}$$

which requires the existence and the evaluation of the limit of the ratio of measures

$$\frac{\mu_2(\mathrm{d}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}))}{\mu_1(\mathrm{d}\boldsymbol{\theta}_1)\bar{\mu}_1(\mathrm{d}\boldsymbol{\varphi}_{1,2})}. \tag{28}$$

In numerous cases, μ_2, μ_1 and $\bar{\mu}_1$ are the Lebesgue measures on the sets $\boldsymbol{\Theta}_2$, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Psi}_{1,2}$, thus this ratio is equal to 1.

Finally, one can extend this idea to a more general case where there exists a deterministic invertible relationship $f_{1,2}(\cdot)$ between $\boldsymbol{\Theta}_2 \times \boldsymbol{\Psi}_{2,1}$ and $\boldsymbol{\Theta}_1 \times \boldsymbol{\Psi}_{1,2}$ of the form

$$\begin{pmatrix}\boldsymbol{\theta}_2^*\\ \boldsymbol{\varphi}_{2,1}^*\end{pmatrix} = \begin{pmatrix}f_{1,2}^{\boldsymbol{\theta}}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2})\\ f_{1,2}^{\boldsymbol{\varphi}}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2})\end{pmatrix}. \tag{29}$$

Assuming that

$$\pi_1(1,\mathrm{d}\boldsymbol{\theta}_1) = \pi_1(1,\boldsymbol{\theta}_1)\mu_1(\mathrm{d}\boldsymbol{\theta}_1),$$

$$q_{1,2}(\mathrm{d}\boldsymbol{\varphi}_{1,2} \mid \boldsymbol{\theta}_1) = q_{1,2}(\boldsymbol{\varphi}_{1,2} \mid \boldsymbol{\theta}_1)\bar{\mu}_1(\mathrm{d}\boldsymbol{\varphi}_{1,2}),$$

$$\pi_2(2,\mathrm{d}f^{\boldsymbol{\theta}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2})) = \pi_2(2,f^{\boldsymbol{\theta}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}))\mu_2(\mathrm{d}f^{\boldsymbol{\theta}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2})),$$

$$q_{2,1}(\mathrm{d}f^{\boldsymbol{\varphi}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2})\,|\,f^{\boldsymbol{\theta}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2})) = q_{2,1}(f^{\boldsymbol{\varphi}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2})\,|\,f^{\boldsymbol{\theta}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2})) \times \bar{\mu}_2(\mathrm{d}f^{\boldsymbol{\varphi}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2})), \tag{30}$$

it follows that

$$r((1,\boldsymbol{\theta}_1),(2,\boldsymbol{\theta}^*_2)) = \frac{\pi_2(2,f^{\boldsymbol{\theta}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}))q_{2,1}(f^{\boldsymbol{\varphi}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2})\,|\,f^{\boldsymbol{\theta}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}))}{\pi_1(1,\boldsymbol{\theta}_1)q_{1,2}(\boldsymbol{\varphi}_{1,2}\,|\,\boldsymbol{\theta}_1)} \times \frac{\mu_2(\mathrm{d}f^{\boldsymbol{\theta}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}))\bar{\mu}_2(\mathrm{d}f^{\boldsymbol{\varphi}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}))}{\mu_1(\mathrm{d}\boldsymbol{\theta}_1)\bar{\mu}_1(\mathrm{d}\boldsymbol{\varphi}_{1,2})}. \tag{31}$$

Again, in numerous cases, μ_2, μ_1, $\bar{\mu}_1$ and $\bar{\mu}_2$ are Lebesgue measures and therefore this ratio limit satisfies

$$\frac{\mu_2(\mathrm{d}f^{\boldsymbol{\theta}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}))\bar{\mu}_2(\mathrm{d}f^{\boldsymbol{\varphi}}_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2}))}{\mu_1(\mathrm{d}\boldsymbol{\theta}_1)\bar{\mu}_1(\mathrm{d}\boldsymbol{\varphi}_{1,2})} = \mathscr{J}_{f_{1,2}} \triangleq \left|\det\frac{\partial f_{1,2}(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2})}{\partial(\boldsymbol{\theta}_1,\boldsymbol{\varphi}_{1,2})}\right|, \tag{32}$$

where $\mathscr{J}_{f_{1,2}}$ is the Jacobian of the transformation $f_{1,2}(\cdot)$.

Remark 3. The algorithm described in Section 3.1.2 is a special case of this last framework, when $\boldsymbol{\varphi}_{1,2} = \boldsymbol{\theta}_2$, $\boldsymbol{\varphi}_{2,1} = \boldsymbol{\theta}_1$ and $f_{1,2}(\cdot)$ is such that

$$(\boldsymbol{\theta}_2\ \ \boldsymbol{\theta}_1) = f_{1,2}(\boldsymbol{\theta}_1\ \ \boldsymbol{\theta}_2). \tag{33}$$

Remark 4. The theoretical framework described by Green [27] is more general, as he shows that a sufficient condition for the acceptance ratio to be defined and evaluated is to be able to find a symmetric dominating measure on the probability distributions $\pi_m(m,\mathrm{d}\boldsymbol{\theta}_m)q_{m,n}(\mathrm{d}\boldsymbol{\varphi}_{m,n}\,|\,\boldsymbol{\theta}_m)$. However the practical framework presented by Green is the same as the one discussed here.

3.1.4. Practical implementation

In this section, we summarize reversible jump MCMC algorithms to jump between spaces for two models defined on $\boldsymbol{\Theta}_n$ and $\boldsymbol{\Theta}_m$ from an algorithmic point of view. For the sake of simplicity all the distributions are assumed to have a density with respect to the Lebesgue measure. The algorithm requires the definition of a deterministic invertible mapping between $\boldsymbol{\Theta}_m \times \boldsymbol{\Psi}_{m,n}$ and $\boldsymbol{\Theta}_n \times \boldsymbol{\Psi}_{n,m}$

$$(\boldsymbol{\theta}^*_m,\boldsymbol{\varphi}^*_{m,n}) = f_{n,m}(\boldsymbol{\theta}_n,\boldsymbol{\varphi}_{n,m}) \tag{34}$$

and for the reverse move $(\boldsymbol{\theta}^*_n,\boldsymbol{\varphi}^*_{n,m}) = f_{m,n}(\boldsymbol{\theta}_m,\boldsymbol{\varphi}_{m,n})$ where

$$f_{m,n}\circ f_{n,m}(\boldsymbol{\theta}_n,\boldsymbol{\varphi}_{n,m}) = (\boldsymbol{\theta}_n,\boldsymbol{\varphi}_{n,m}).$$

The fact that the mapping is invertible implies that there is dimension matching between $\boldsymbol{\Theta}_m \times \boldsymbol{\Psi}_{m,n}$ and $\boldsymbol{\Theta}_n \times \boldsymbol{\Psi}_{n,m}$, i.e. $\dim(\boldsymbol{\theta}^*_m) + \dim(\boldsymbol{\varphi}^*_{m,n}) = \dim(\boldsymbol{\theta}_n) + \dim(\boldsymbol{\varphi}_{n,m})$. Two proposal densities for $\boldsymbol{\varphi}_{n,m}$ and $\boldsymbol{\varphi}^*_{m,n}$, respectively $q_{n,m}(\boldsymbol{\varphi}_{n,m}\,|\,\boldsymbol{\theta}_n)$ and $q_{m,n}(\boldsymbol{\varphi}^*_{m,n}\,|\,\boldsymbol{\theta}_m)$, have to be defined. Then, if the current state is $(n,\boldsymbol{\theta}_n)$, the move of jumping from $\boldsymbol{\Theta}_n$ to $\boldsymbol{\Theta}_m$ is

Move from $\boldsymbol{\Theta}_n$ to $\boldsymbol{\Theta}_m$

(1) Sample $\boldsymbol{\varphi}_{n,m} \sim q_{n,m}(\boldsymbol{\varphi}_{n,m}\,|\,\boldsymbol{\theta}_n)$ and perform the invertible transformation $(\boldsymbol{\theta}^*_m,\boldsymbol{\varphi}^*_{m,n}) = f_{n,m}(\boldsymbol{\theta}_n,\boldsymbol{\varphi}_{n,m})$.

(2) Accept the move with probability

$$\alpha((n,\boldsymbol{\theta}_n),(m,\boldsymbol{\theta}^*_m)) = \min\{1, r((n,\boldsymbol{\theta}_n),(m,\boldsymbol{\theta}^*_m))\} \tag{35}$$

otherwise stay at $(n,\boldsymbol{\theta}_n)$.

If the current state is $(\boldsymbol{\theta}_m,\boldsymbol{\varphi}_m)$, the associated reverse move from $\boldsymbol{\Theta}_m$ to $\boldsymbol{\Theta}_n$ is

Move from $\boldsymbol{\Theta}_m$ to $\boldsymbol{\Theta}_n$

(1) Sample $\boldsymbol{\varphi}_{m,n} \sim q_{m,n}(\boldsymbol{\varphi}_{m,n}\,|\,\boldsymbol{\theta}_m)$ and perform the invertible transformation $(\boldsymbol{\theta}^*_n,\boldsymbol{\varphi}^*_{n,m}) = f_{m,n}(\boldsymbol{\theta}_m,\boldsymbol{\varphi}_{m,n})$.

(2) Accept the move with probability

$$\alpha((m,\boldsymbol{\theta}_m),(n,\boldsymbol{\theta}^*_n)) = \min\{1, r^{-1}((n,\boldsymbol{\theta}^*_n),(m,\boldsymbol{\theta}_m))\} \tag{36}$$

otherwise stay at $(m,\boldsymbol{\theta}_m)$.

where

$$r((a,\boldsymbol{\theta}_a),(b,\boldsymbol{\theta}^*_b)) = \frac{\pi_b(b,\boldsymbol{\theta}^*_b)q_{b,a}(\boldsymbol{\varphi}^*_{b,a}\,|\,\boldsymbol{\theta}^*_b)}{\pi_a(a,\boldsymbol{\theta}_a)q_{a,b}(\boldsymbol{\varphi}_{a,b}\,|\,\boldsymbol{\theta}_a)}\mathscr{J}_{f_{a,b}} \tag{37}$$

and $\mathscr{J}_{f_{a,b}}$ is the Jacobian of the transformation $f_{a,b}(\cdot)$

$$\mathscr{J}_{f_{a,b}} = \left|\det\frac{\partial f_{a,b}(\boldsymbol{\theta}_a,\boldsymbol{\varphi}_{a,b})}{\partial(\boldsymbol{\theta}_a,\boldsymbol{\varphi}_{a,b})}\right|. \tag{38}$$

Note that this algorithm is not guaranteed to produce, even asymptotically, samples from the correct distributions, as two extra properties, namely irreducibility and aperiodicity, of the Markov chain need to be checked. The reader interested in the precise definition of these notions should refer to [36,42].

Remark 5. To obtain the Jacobian in Eq. (37) we made the important assumption that all the distributions admit a density with respect to the Lebesgue measure. This assumption is justified in most applications. However, in more general cases, one should be careful and use expression (31).

In this subsection we have shown how to sample parameters from two distinct distributions and possibly take advantage of the relation between the parameters of the two distributions. We now present an extension to the more general case where $N \geqslant 2$ different distributions are involved.

3.2. Reversible jump MCMC for $N \geqslant 2$ models

3.2.1. Goals

We want now to sample from the distribution

$$\int_A \pi(k,\mathrm{d}\boldsymbol{\theta}) = \sum_{n=1}^{N} \int_A \pi_n(n,\mathrm{d}\boldsymbol{\theta}_n)\mathbb{I}_{\{n\}\times\boldsymbol{\Theta}_n}(k,\boldsymbol{\theta}) \tag{39}$$

defined on $\boldsymbol{\Theta} \triangleq \bigcup_{n=1}^{N} \{n\} \times \boldsymbol{\Theta}_n$. To this aim, we need to construct a Markov transition kernel $K(k^*,\mathrm{d}\boldsymbol{\theta}^* \mid k,\boldsymbol{\theta})$ admitting $\pi(k,\mathrm{d}\boldsymbol{\theta})$ as an invariant distribution. As in the previous section we assume that one can define a family $(f_{n,m}(\cdot,\cdot))_{(m,n)\in\{1,\dots,N\}^2}$ of invertible mappings between $\boldsymbol{\Theta}_n \times \boldsymbol{\Psi}_{n,m}$ and $\boldsymbol{\Theta}_m \times \boldsymbol{\Psi}_{m,n}$

$$(\boldsymbol{\theta}_m^*,\boldsymbol{\varphi}_{m,n}^*) = f_{n,m}(\boldsymbol{\theta}_n,\boldsymbol{\varphi}_{n,m}) \tag{40}$$

and that there exists a family of proposal distributions $q_{n,m}(\mathrm{d}\boldsymbol{\varphi}_{n,m} \mid n,\boldsymbol{\theta}_n)$ with $(m,n) \in \{1,\dots,N\}^2$ from which we know how to sample. In practice, we need an MCMC transition kernel which mixes moves within each subspace (as described previously) and moves between subspaces. So the proposal distribution can be written as

$$q(\mathrm{d}\boldsymbol{\varphi} \mid k,\boldsymbol{\theta}) = \sum_{n=1}^{N} \sum_{m=1}^{N} \rho_{n,m}(\boldsymbol{\theta}_n) q_{n,m}(\mathrm{d}\boldsymbol{\varphi}_{n,m} \mid n,\boldsymbol{\theta}_n)\mathbb{I}_{\boldsymbol{\Psi}_{n,m}\times\boldsymbol{\Theta}_n}(\boldsymbol{\varphi};\boldsymbol{\theta}), \tag{41}$$

where, for any $n \in \{1,\dots,N\}$, $0 \leqslant \rho_{n,m}(\boldsymbol{\theta}_n) \leqslant 1$ and $\sum_{m=1}^{N} \rho_{n,m}(\boldsymbol{\theta}_n) = 1$. Hence if $\boldsymbol{\theta} = \boldsymbol{\theta}_n \in \boldsymbol{\Theta}_n$ then, with probability $\rho_{nn}(\boldsymbol{\theta})$, $\boldsymbol{\varphi}_{n,n} \in \boldsymbol{\Psi}_{nn}$ is sampled according to $q_{nn}(\mathrm{d}\boldsymbol{\varphi}_{n,n} \mid \boldsymbol{\theta}_n)$ (classical MCMC move) and, with probability $\rho_{n,m}(\boldsymbol{\theta})$, $n \neq m$, $\boldsymbol{\varphi}_{n,m} \in \boldsymbol{\Psi}_{n,m}$ is sampled according to $q_{n,m}(\mathrm{d}\boldsymbol{\varphi}_{n,m} \mid \boldsymbol{\theta}_n)$ (reversible jump MCMC move). In many applications, $\rho_{n,m}(\boldsymbol{\theta}_n) = 0$ for most couples (n,m). Note that a possible extension would consist of proposing several moves from $\boldsymbol{\Theta}_n$ to $\boldsymbol{\Theta}_m$. Here we do not consider this case because it is a trivial extension of the above scheme.

With the assumption that all probability distributions admit a density with respect to the Lebesgue measure, the acceptance probability of a move from $\boldsymbol{\Theta}_n$ to $\boldsymbol{\Theta}_m$ satisfies

$$r((n,\boldsymbol{\theta}_n),(m,\boldsymbol{\theta}_m^*)) = \frac{\pi_m(m,\boldsymbol{\theta}_m^*)\rho_{m,n}(\boldsymbol{\theta}_m^*)q_{m,n}(\boldsymbol{\varphi}_{m,n}^* \mid \boldsymbol{\theta}_m^*)}{\pi_n(n,\boldsymbol{\theta}_n)\rho_{n,m}(\boldsymbol{\theta}_n)q_{n,m}(\boldsymbol{\varphi}_{n,m} \mid \boldsymbol{\theta}_n)} \times \mathcal{J}_{f_{n,m}}, \tag{42}$$

where $\mathcal{J}_{f_{n,m}}$ is the Jacobian, when only continuous variables are involved in the transformation, of the invertible mapping $f_{n,m}(\cdot,\cdot)$ between the spaces $\boldsymbol{\Theta}_n$ and $\boldsymbol{\Theta}_m$. One must not forget the terms $\rho_{n,m}(\boldsymbol{\theta}_n)$ and $\rho_{m,n}(\boldsymbol{\theta}_m^*)$ that are part of the proposal in the acceptance ratio. They did not appear in Eq. (37) as in that case N was equal to 2 and the only permitted moves were from space $\boldsymbol{\Theta}_n$ to space $\boldsymbol{\Theta}_m$ with $n \neq m$. Invariance of this kernel with respect to the distribution $\pi(k,\mathrm{d}\boldsymbol{\theta})$ is ensured by the detailed balance condition.

3.2.2. Practical implementation

The main procedure of the algorithm is of the form

Reversible Jump MCMC algorithm

(1) Initialization: set $(k^{(0)},\boldsymbol{\theta}_{k^{(0)}}^{(0)}) \in \boldsymbol{\Theta}$, and $i = 1$.

(2) Iteration i.

- Sample k^* from the discrete distribution $(\rho_{k^{(i)}k^*}(\boldsymbol{\theta}_{k^{(i)}}))_{k^*=1,\dots,N}$.

- Apply the procedure **Move from $\boldsymbol{\Theta}_{k^{(i)}}$ to $\boldsymbol{\Theta}_{k^*}$** (with in this case $r((k,\boldsymbol{\theta}),(k^*,\boldsymbol{\theta}^*))$ **defined in** (42)).

where Move from Θ_n to Θ_m is described in Section 3.1.4.

3.3. Other approaches◆

3.3.1. Method of subspace extension and the algorithm of Carlin and Chib

One can, of course, generalize what has been presented in the previous subsection by extending the two subspaces $\boldsymbol{\Theta}_n$ and $\boldsymbol{\Theta}_m$. More precisely, we can introduce extended parameters $\bar{\boldsymbol{\theta}}_n \triangleq (\boldsymbol{\theta}_n,\boldsymbol{\varphi}_n)$ defined on extended sets $\bar{\boldsymbol{\Theta}}_n \triangleq \boldsymbol{\Theta}_n \times \boldsymbol{\Psi}_n$ and associated with probabilistic models $\bar{\pi}_n(n,\mathrm{d}\cdot)$ such that

$$\int_{\boldsymbol{\Psi}_n} \bar{\pi}_n(n,\mathrm{d}\bar{\boldsymbol{\theta}}_n) = \pi_n(n,\mathrm{d}\boldsymbol{\theta}_n) \tag{43}$$

which means that there exists a distribution $\bar{\pi}_n(\mathrm{d}\boldsymbol{\varphi}_n \mid n,\boldsymbol{\theta}_n)$ such that

$$\bar{\pi}_n(n,\mathrm{d}\bar{\boldsymbol{\theta}}_n) = \bar{\pi}_n(\mathrm{d}\boldsymbol{\varphi}_n \mid n,\boldsymbol{\theta}_n)\pi_n(n,\mathrm{d}\boldsymbol{\theta}_n). \tag{44}$$

Then, assuming that there exists a deterministic invertible mapping $f_{n,m}(\cdot)$ between $\bar{\boldsymbol{\Theta}}_n$ and $\bar{\boldsymbol{\Theta}}_m$, one can apply the same strategy as mentioned in the previous subsection.

A particular case of interest is the method proposed by Carlin and Chib [8], in which the following choice for family of spaces $(\boldsymbol{\Psi}_n)_{n=1,\dots,N}$ was made:

$$\boldsymbol{\Psi}_n = \boldsymbol{\Theta}_1 \times \dots \times \boldsymbol{\Theta}_{n-1} \times \boldsymbol{\Theta}_{n+1} \times \dots \times \boldsymbol{\Theta}_N. \tag{45}$$

This requires the definition of the following distributions:

$$\bar{\pi}_n(\mathrm{d}\boldsymbol{\varphi}_n \mid n,\boldsymbol{\theta}_n) \tag{46}$$

called "pseudopriors" as they do not have any meaning in the sense of statistical application, despite the fact that $\boldsymbol{\varphi}_n$ is composed of parameters similar to $\boldsymbol{\theta}_l$ for $l \neq n$. Note that the complete state space can then be written as

$$\begin{aligned}\boldsymbol{\Theta} &= \bigcup_{n=1}^{N} \{n\} \times \boldsymbol{\Theta}_1 \times \dots \times \boldsymbol{\Theta}_N \\ &= \{1,\dots,N\} \times \boldsymbol{\Theta}_1 \times \dots \times \boldsymbol{\Theta}_N \end{aligned} \tag{47}$$

and that N probabilistic models must be defined at the beginning of the procedure and, while it is carried out, this number cannot change.

The algorithm they propose is a Gibbs sampler on $\bar{\boldsymbol{\Theta}}$ which, contrary to Green's reversible jump MCMC, allows $\bar{\boldsymbol{\theta}}_n$ to be drawn first and then the new model m conditional upon $\bar{\boldsymbol{\theta}}_n$. A summary of the algorithm follows:

Carlin and Chib's algorithm

```
(1) Initialization
(2) Iteration $i$
    ● $\boldsymbol{\varphi}^{(i)}_{k^{(i-1)}} \sim \bar{\pi}_{k^{(i-1)}}(d\cdot \mid k^{(i-1)},\boldsymbol{\theta}^{(i-1)}_{k^{(i-1)}})$
    ● $\boldsymbol{\theta}^{(i)}_{k^{(i-1)}} \sim \bar{\pi}_{k^{(i-1)}}(d\cdot \mid k^{(i-1)},\boldsymbol{\varphi}^{(i)}_{k^{(i-1)}})$
(3) Draw the new index $k^{(i)}$

        $k^{(i)} \sim \bar{\pi}_k(\cdot \mid \boldsymbol{\theta}^{(i)}_k,\boldsymbol{\varphi}^{(i)}_k) \propto \bar{\pi}_k(\boldsymbol{\varphi}^{(i)}_k \mid k,\boldsymbol{\theta}^{(i)}_k)\pi_k(k,\boldsymbol{\theta}^{(i)}_k)$
                                                                        (48)
(4) Go to 2.
```

Note that Chib and Carlin make further assumptions that, with $\boldsymbol{\varphi}_k = (\varphi_{1,k},\dots,\varphi_{k-1,k},\varphi_{k+1,k},\dots,\varphi_N)$, the $\varphi_{m,k}$ are independent among themselves and of $\boldsymbol{\theta}_k$, conditional upon k. More precisely, they assume that

$$\bar{\pi}_k(k,\mathrm{d}(\boldsymbol{\theta}_k,\boldsymbol{\varphi}_k)) = \pi_k(k,\mathrm{d}\boldsymbol{\theta}_k)\prod_{\substack{m=1\\ m\neq k}}^{N} \bar{\pi}_k(\mathrm{d}\boldsymbol{\varphi}_{m,k} \mid k) \tag{49}$$

and that the pseudopriors do not depend on the current index, that is, for any k and distinct m and n such that $m \neq k$ and $n \neq k$ then

$$\bar{\pi}_m(\mathrm{d}\boldsymbol{\varphi}_{k,m} \mid m) = \bar{\pi}_n(\mathrm{d}\boldsymbol{\varphi}_{k,n} \mid n). \tag{50}$$

This algorithm has several drawbacks:

- N has to be finite,
- simulation of $\boldsymbol{\varphi}_m$ is required at each iteration, although they are neither used for estimation purposes nor for proposing the $\boldsymbol{\theta}_k$ in a "clever" way,
- pseudopriors must be carefully chosen in order for the exploration of the different indices to be efficient,
- it requires a Gibbs sampler, and hence the availability of full conditional distributions, which seriously limits the range of applications, even if Metropolized versions can be proposed.

Interesting remarks and comments on this algorithm can be found in [11,24], where connections between "Metropolized" Carlin and Chib type algorithm and reversible jump MCMC were independently noticed.

3.3.2. Jump-diffusion sampling

Much of the research on sampling from posterior distributions defined on $\boldsymbol{\Theta} \triangleq \bigcup_{n=1}^{N} \{n\} \times \boldsymbol{\Theta}_n$ was reinitiated by the work of Grenander and Miller published in 1994 [26]. The proposed method of inference is based on a random process which follows jump-diffusion dynamics, and whose samples are drawn from a posterior of the form (9).

The main idea in [26] is to construct a single posterior distribution over the union of all considered parameter spaces and then sample from it using a Markov process that has jump-diffusion dynamics. In particular, at random times the Markov process jumps from one of the parameter spaces to another, and in between the jumps, the process follows Langevin stochastic differential equations. Early work on simulations from a given probability density by using Langevin equations appears in [25].

The jump dynamics can be defined in various ways, and always must satisfy certain regularity and balance conditions as well as reversibility. The jump times are obtained from marginal jump intensities, which are computed from jump intensities chosen to satisfy a proper condition that ensure sampling from the desired posterior [26, Theorem 1,b]. Once the jump time is determined, a decision where to jump is made by using transition kernels, which are conditional probability densities of jumping from the current to a new parameter space. Two useful jump dynamics are the Gibbs and Metropolis–Hastings jump dynamics [32].

The diffusion process $\boldsymbol{\theta}^{(t)}$ within a fixed parameter space and between jump times satisfies the Langevin stochastic differential equation

$$\mathrm{d}\boldsymbol{\theta}_k^{(t)} = \frac{\mathrm{d}t}{2}\left(\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}_k}\log \pi_k(\boldsymbol{\theta}_k|k)\right)_{\boldsymbol{\theta}_k^{(t)}} + \mathrm{d}\boldsymbol{W}_k^{(t)}, \tag{51}$$

where $\boldsymbol{W}_k^{(t)}$ is a standard Brownian motion process whose dimension is the same as that of $\boldsymbol{\theta}_k^{(t)}$. In practice, (51) is approximated by a discrete-time version of it, and to be rigourous, a Metropolization of the algorithm is necessary to preserve the correct target distribution [37].

4. Examples

4.1. Analysis of sinusoids in noise

4.1.1. Data models

We do not motivate here the choice of the probabilistic model to perform model selection of sinusoids in Gaussian noise; the reader should refer to [1]. We just recall that the problem addressed is the selection of one of the following models:

$$\mathscr{M}_0\colon \quad y_t = w_{0,t}, \quad k = 0,$$

$$\mathscr{M}_k\colon \quad y_t = \sum_{j=1}^{k} (a_{c_{j,k}} \cos[\omega_{j,k} t] + a_{s_{j,k}} \sin[\omega_{j,k} t]) + w_{k,t}, \quad k \geqslant 1, \tag{52}$$

to represent the data for $t = 1, \ldots, T-1$, where $w_{k,t} \overset{\text{iid}}{\sim} \mathscr{N}(0,\sigma_k^2)$. In the present paper, dedicated to Bayesian computational methods and not Bayesian model selection, we use the priors introduced in [1], which allow for the amplitudes and the variance of the observation noise to be integrated out analytically, leading to the following posterior distribution for the number of sinusoids and their frequencies:

$$\pi_k(k,\mathrm{d}\boldsymbol{\omega}|\boldsymbol{y}_{1:T}) \propto \sum_{n=0}^{k_{\max}} (\boldsymbol{y}_{1:T}^{\mathrm{T}} \boldsymbol{P}_n \boldsymbol{y}_{1:T})^{-T/2} \times \frac{(\Lambda/((\delta^2+1)\pi))^n}{n!} \mathrm{d}\boldsymbol{\omega}_n \mathbb{I}_{\{n\}\times\boldsymbol{\Omega}_n}(k,\boldsymbol{\omega}). \tag{53}$$

The hyperparameters Λ and δ^2 can be interpreted as the mean number of expected sinusoids and the expected signal-to-noise ratio, respectively. A discussion of priors for model selection can be found in [3]. The matrix $\boldsymbol{P}_n$ is defined as

$$\boldsymbol{P}_n = \boldsymbol{I}_T - \boldsymbol{D}(\boldsymbol{\omega}_n)\boldsymbol{M}_n\boldsymbol{D}^{\mathrm{T}}(\boldsymbol{\omega}_n)$$

with

$$\boldsymbol{M}_n^{-1} = \frac{1+\delta^2}{\delta^2}\boldsymbol{D}^{\mathrm{T}}(\boldsymbol{\omega}_n)\boldsymbol{D}(\boldsymbol{\omega}_n) \tag{54}$$

and

$$[\boldsymbol{D}(\boldsymbol{\omega}_n)]_{i+1,2j-1} \triangleq \cos[\omega_{j,n} i] \quad (i = 1, \ldots, T-1,\ j = 1, \ldots, n),$$

$$[\boldsymbol{D}(\boldsymbol{\omega}_n)]_{i+1,2j} \triangleq \sin[\omega_{j,n} i] \quad (i = 1, \ldots, T-1,\ j = 1, \ldots, n). \tag{55}$$

The space where the probabilistic models are defined is $\boldsymbol{\Omega} \triangleq \bigcup_{n=0}^{k_{\max}} \{n\} \times \boldsymbol{\Omega}_n$ where $\boldsymbol{\Omega}_0 \triangleq \emptyset$,

$$\boldsymbol{\Omega}_n \triangleq \{\boldsymbol{\omega}_n ; \boldsymbol{\omega}_n \in (0,\pi)^n / \omega_{j_1,n} \neq \omega_{j_2,n} \text{ for } j_1 \neq j_2\} \quad (56)$$

and $k_{\max} \triangleq \lfloor (T-1)/2 \rfloor$. To perform model selection, one is interested in evaluating $\pi_k(k|\boldsymbol{y}_{1:T})$, for which one cannot obtain a closed-form expression.

4.1.2. Overview of the algorithm

For our problem, the following moves have been selected:

(1) Birth of a new sinusoid, i.e. proposing a new sinusoid with frequency ω^* at random on $(0,\pi)$.
(2) Death of an existing sinusoid, i.e. removing a sinusoid chosen randomly.
(3) Update of the parameters of all the sinusoids, when $k \neq 0$, and the variance of the observation noise.

The birth and death moves represent changes from k to $k+1$ and k to $k-1$, respectively. These moves are defined by heuristic considerations, the only condition to be fulfilled being to maintain the correct invariant distribution. A particular choice will only have influence on the convergence rate of the algorithm. Other moves may be proposed, but we have found that the ones suggested here lead to satisfactory results.

The resulting transition kernel of the simulated Markov chain is then a mixture of the different transition kernels associated with the moves described above. This means that at each iteration one of the candidate moves: birth, death or update is randomly chosen. The probabilities for choosing these moves are $\rho_{k,k+1}(\boldsymbol{\omega}_k)$, $\rho_{k,k-1}(\boldsymbol{\omega}_k)$ and $\rho_{k,k}(\boldsymbol{\omega}_k)$ respectively, where $\rho_{k,k+1}(\boldsymbol{\omega}_k) + \rho_{k,k-1}(\boldsymbol{\omega}_k) + \rho_{k,k}(\boldsymbol{\omega}_k) = 1$ for all $0 \leqslant k \leqslant k_{\max}$. The move is performed if the algorithm accepts it. For $k = 0$ the death move is impossible, so that $\rho_{0,-1}(\boldsymbol{\omega}_0) \triangleq 0$. For $k = k_{\max}$ the birth move is impossible and thus $\rho_{k_{\max},k_{\max}+1}(\boldsymbol{\omega}_{k_{\max}}) \triangleq 0$. Except in the cases described above, we take the following probabilities:

$$\begin{aligned} \rho_{k,k+1}(\boldsymbol{\omega}_k) &\triangleq c \min\left\{1, \frac{\pi_{k+1}(k+1)}{\pi_k(k)}\right\}, \\ \rho_{k+1,k}(\boldsymbol{\omega}_{k+1}) &\triangleq c \min\left\{1, \frac{\pi_k(k)}{\pi_{k+1}(k+1)}\right\} \end{aligned} \quad (57)$$

where $\pi_k(k) \propto (\Lambda^k/k!)\mathbb{I}_{\{0,\ldots,k_{\max}\}}$ is the prior probability of model $\mathscr{M}_k$ and c is a parameter which tunes the relative frequencies of dimension change and update moves. As pointed out in [27, p. 719], this choice ensures that

$$\rho_{k,k+1}(\boldsymbol{\omega}_k)\pi_k(k)[\rho_{k+1,k}(\boldsymbol{\omega}_{k+1})\pi_{k+1}(k+1)]^{-1} = 1 \quad (58)$$

which means that a MH algorithm on the sole dimension in the case of no observation would have 1 as acceptance probability. We take $c = 0.5$ and then $\rho_{k,k+1}(\boldsymbol{\omega}_k) + \rho_{k,k-1}(\boldsymbol{\omega}_k) \in [0.5,1]$ for all k [27]. For this algorithm, $\boldsymbol{\varphi}_{k,k+1} = \omega^*$ and $\boldsymbol{\omega}_{k+1} = f_{k,k+1}(\boldsymbol{\omega}_k, \omega^*)$ is just any concatenating function such as $\boldsymbol{\omega}_{k+1} = (\boldsymbol{\omega}_k, \omega^*)$.

One can then describe the main steps of the algorithm as follows:

Reversible Jump MCMC algorithm

```
(1) Initialization: set $(k^{(0)}, \boldsymbol{\theta}^{(0)}_{k^{(0)}}) \in \boldsymbol{\Theta}$.
(2) Iteration $i$.
    ● Sample $u \sim \mathscr{U}_{[0,1]}$.
    ● If $(u \leqslant \rho_{k^{(i)},k^{(i)}+1}(\boldsymbol{\omega}_{k^{(i)}}))$
        ○ then "birth" move.
        ○ else if
          $(u \leqslant \rho_{k^{(i)},k^{(i)}+1}(\boldsymbol{\omega}_{k^{(i)}}) + \rho_{k^{(i)},k^{(i)}-1}(\boldsymbol{\omega}_{k^{(i)}}))$
          then "death" move.
        ○ else update the parameters using
          a standard MH step.
        ○ End If.
(3) $i \leftarrow i + 1$ and go to 2.
```

We describe more precisely these different reversible jump moves below. In what follows, in order to simplify notation, we drop the superscript $\cdot^{(i)}$ from all variables at iteration i.

4.1.3. The birth and death moves

Suppose that the current state of the Markov chain is in $\{k\} \times \boldsymbol{\Theta}_k$, then

Birth move

- Propose a new frequency at random on $(0,\pi)$: $\omega^* \sim \mathscr{U}_{(0,\pi)}$ and set $\boldsymbol{\omega}_{k+1} = (\boldsymbol{\omega}_k, \omega^*)$.
- Evaluate $\alpha_{k,k+1}$, see (61), and sample $u \sim \mathscr{U}_{[0,1]}$.
- If $u \leqslant \alpha_{k,k+1}$ then the state of the Markov chain becomes $(k+1, \boldsymbol{\omega}_{k+1})$, else it remains at $(k, \boldsymbol{\omega}_k)$.

Assume that the current state of the Markov chain is in $\{k+1\} \times \boldsymbol{\Theta}_{k+1}$, then

Death move

- Choose a sinusoid at random among the $k+1$ existing sinusoids: $l \sim \mathcal{U}_{\{1,\dots,k+1\}}$.
- Evaluate $\alpha_{k+1,k}$, see (61), and sample $u \sim \mathcal{U}_{[0,1]}$.
- If $u \leqslant \alpha_{k+1,k}$ then the state of the Markov chain becomes $(k,\boldsymbol{\omega}_k)$, else it remains $(k+1,\boldsymbol{\omega}_{k+1})$.

The acceptance ratio for the proposed moves are deduced from expression (42)

$$\begin{aligned} & r((k,\boldsymbol{\omega}_k),(k+1,\boldsymbol{\omega}_{k+1})) \\ & \quad = \frac{\pi_{k+1}(k+1,\mathrm{d}\boldsymbol{\omega}_{k+1}|\boldsymbol{y}_{1:T})\rho_{k+1,k}(\boldsymbol{\omega}_{k+1})1/(k+1)}{\pi_k(k,\mathrm{d}\boldsymbol{\omega}_k|\boldsymbol{y}_{1:T})\rho_{k,k+1}(\boldsymbol{\omega}_k)\mathrm{d}\omega^*/\pi} \\ & \quad = \frac{\pi_{k+1}(k+1,\boldsymbol{\omega}_{k+1}|\boldsymbol{y}_{1:T})\rho_{k+1,k}(\boldsymbol{\omega}_{k+1})1/(k+1)}{\pi_k(k,\boldsymbol{\omega}_k|\boldsymbol{y}_{1:T})\rho_{k,k+1}(\boldsymbol{\omega}_k)1/\pi} \\ & \quad \times 1 \end{aligned} \tag{59}$$

which yields, after simplifications

$$\begin{aligned} & r((k,\boldsymbol{\omega}_k),(k+1,\boldsymbol{\omega}_{k+1})) \\ & \quad = \left(\frac{\boldsymbol{y}_{1:T}^{\mathrm{T}}\boldsymbol{P}_k\boldsymbol{y}_{1:T}}{\boldsymbol{y}_{1:T}^{\mathrm{T}}\boldsymbol{P}_{k+1}\boldsymbol{y}_{1:T}}\right)^{T/2} \frac{1}{(k+1)(1+\delta^2)}. \end{aligned} \tag{60}$$

Note that here the Jacobian is equal to one as $\mathrm{d}\boldsymbol{\omega}_{k+1}/(\mathrm{d}\boldsymbol{\omega}_k\,\mathrm{d}\omega^*) = 1$. Then the acceptance probabilities corresponding to the described moves are

$$\begin{aligned} & \alpha((k,\boldsymbol{\omega}_k),(k+1,\boldsymbol{\omega}_{k+1})) \\ & \quad = \min\{1, r((k,\boldsymbol{\omega}_k),(k+1,\boldsymbol{\omega}_{k+1}))\}, \\ & \alpha_{k+1,k}((k+1,\boldsymbol{\omega}_{k+1}),(k,\boldsymbol{\omega}_k)) \\ & \quad = \min\{1, r^{-1}((k,\boldsymbol{\omega}_k),(k+1,\boldsymbol{\omega}_{k+1}))\}. \end{aligned} \tag{61}$$

The update move consists of standard MH steps and is not detailed here, see [1] for details.

4.1.4. Merge and split moves?◆

In this subsection we illustrate the flexibility of the reversible jump methodology by considering sophisticated moves, which (similarly to [34]) we name split and merge moves. These moves are motivated by the following situation where the signal contains two sinusoids closely spaced in frequency. If a single high amplitude sinusoid has been created near the location of the two sinusoids, then the probability of removing this sinusoid so that it can be replaced by two smaller amplitude sinusoids can be low in practice. The split move will divide a sinusoid into two sinusoids in one step. The merge step will select two close sinusoids and replace them by one sinusoid. The proposal distributions for these steps have been selected to ensure that there is conservation of energy between the old and new configurations. Furthermore, we want our transformation to reflect the fact that we are more confident in the value of a sinusoid with high-energy than in a low-energy one. To simplify, we assume that the nuisance parameters, $\boldsymbol{a}_k,\sigma_k^2$ have not been integrated out, i.e. the space $\boldsymbol{\Theta}$ is as defined in Example 1 in Section 2. From a practical point of view it has been found that these sophisticated moves do not significantly improve the quality of the convergence of the algorithm towards the target distribution, meaning that the birth and death move are efficient enough. However, we think that they are of pedagogic interest as they illustrate the flexibility of the approach, and more importantly adaptations have proved to be useful for other types of regression problems for which ambiguities are more likely to occur [4].

Assume that there are $k+1$ sinusoids. Our proposal for the merge move begins by choosing at random a pair l of sinusoids which are adjacent in terms of their frequencies. To simplify notation, we will denote these two sinusoids as $(a_{c_1}, a_{s_1}, \omega_1)$ and $(a_{c_2}, a_{s_2}, \omega_2)$. One can merge these sinusoids, thus reducing $k+1$ by 1 and creating a new sinusoid (a_c^*, a_s^*, ω^*). The parameters of the proposed new sinusoid are obtained by first generating

$$u_0^* \sim \mathcal{U}_{(0,1)} \tag{62}$$

which will determine the fraction of energy attributed to the component $|a_c^*|\cos[\omega^* t]$. The signs of a_c^* and a_s^*, ε_c^* and ε_s^*, are drawn according to a uniform discrete probability distribution on $\{-1,1\}$, i.e.

$$\begin{aligned} & \Pr(\varepsilon_c^* = -1) = \Pr(\varepsilon_c^* = 1) = \tfrac{1}{2}, \\ & \Pr(\varepsilon_s^* = -1) = \Pr(\varepsilon_s^* = 1) = \tfrac{1}{2}. \end{aligned} \tag{63}$$

Consequently $\boldsymbol{\varphi}^*_{k+1,k} = (u^*_0, \varepsilon^*_c, \varepsilon^*_s)$. The transformation $f_{k+1,k}(\cdot,\cdot)$[7] is defined as

$$
\begin{aligned}
&a_c^{*2} + a_s^{*2} = a_{c_1}^2 + a_{s_1}^2 + a_{c_2}^2 + a_{s_2}^2,\\
&a_c^* = \varepsilon_c^* \sqrt{u_0^*(a_{c_1}^2 + a_{s_1}^2 + a_{c_2}^2 + a_{s_2}^2)},\\
&a_s^* = \varepsilon_s^* \sqrt{(1-u_0^*)(a_{c_1}^2 + a_{s_1}^2 + a_{c_2}^2 + a_{s_2}^2)},\\
&\omega^* = \frac{(a_{c_1}^2 + a_{s_1}^2)\omega_1 + (a_{c_2}^2 + a_{s_2}^2)\omega_2}{a_{c_1}^2 + a_{s_1}^2 + a_{c_2}^2 + a_{s_2}^2}.
\end{aligned}
\tag{64}
$$

Now, the reverse split move is largely determined. Assume that there are k sinusoids. Our proposal begins by choosing a sinusoid (a_c, a_s, ω) among the k existing ones with uniform probability. Then, this sinusoid is split into two components $(a_{c_1}, a_{s_1}, \omega_1)$ and $(a_{c_2}, a_{s_2}, \omega_2)$ with parameters conforming to Eq. (64). To generate these new parameters, we first generate an eight-dimensional random parameter vector $\boldsymbol{\varphi}^*_{k+1,k} = (u_1^*, u_2^*, u_3^*, u_4^*, \varepsilon^*_{c_1}, \varepsilon^*_{s_1}, \varepsilon^*_{c_2}, \varepsilon^*_{s_2})$ where

$$
u_1^* \sim \mathscr{U}_{(0,1)}, \quad u_2^* \sim \mathscr{U}_{(0,1)}, \quad u_3^* \sim \mathscr{U}_{(0,1)}, \quad u_4^* \sim \mathscr{U}_{(0,1)}
\tag{65}
$$

and $\varepsilon^*_{c_1}$, $\varepsilon^*_{s_1}$, $\varepsilon^*_{c_2}$, $\varepsilon^*_{s_2}$ are drawn according to a uniform discrete probability distribution on $\{-1,1\}$. Then the inverse $f_{k,k+1}(\cdot,\cdot)$ of the transformation is given by

$$
\begin{aligned}
&a_{c_1}^{*2} + a_{s_1}^{*2} = u_1^*(a_c^2 + a_s^2),\\
&a_{c_2}^{*2} + a_{s_2}^{*2} = (1-u_1^*)(a_c^2 + a_s^2),\\
&a_{c_1}^* = \varepsilon_{c_1}^* \sqrt{u_2^* u_1^*(a_c^2 + a_s^2)},\\
&a_{c_2}^* = \varepsilon_{c_2}^* \sqrt{u_3^*(1-u_1^*)(a_c^2 + a_s^2)},\\
&a_{s_1}^* = \varepsilon_{s_1}^* \sqrt{(1-u_2^*)u_1^*(a_c^2 + a_s^2)},\\
&a_{s_2}^* = \varepsilon_{s_2}^* \sqrt{(1-u_3^*)(1-u_1^*)(a_c^2 + a_s^2)},\\
&\omega_1^* = \omega - u_4^*\tilde{\sigma}\sqrt{\frac{a_{c_2}^{*2} + a_{s_2}^{*2}}{a}{}^{*2c_1} + a_{s_1}^{*2}},\\
&\omega_2^* = \omega + u_4^*\tilde{\sigma}\sqrt{\frac{a_{c_1}^{*2} + a_{s_1}^{*2}}{a_{c_2}^{*2} + a_{s_2}^{*2}}},
\end{aligned}
\tag{66}
$$

Table 1
Parameters for the experiment

i	E_i	$-\arctan(a_{s_i}/a_{c_i})$	$\omega_i/2\pi$
1	20	0	0.2
2	20	$\pi/3$	$0.2 + 2/T$

where $\tilde{\sigma}$ is a predetermined constant. Eq. (66) is consistent with (64). Once ω_1 and ω_2 have been sampled, we must also verify that there is no other frequency located between ω_1 and ω_2. If there are such frequencies, then the move is rejected as the split/merge pair would not be reversible. The acceptance ratio of the merge move has the form, with $\boldsymbol{\theta}_k \triangleq (\boldsymbol{a}_k, \boldsymbol{\omega}_k, \sigma_k^2)$

$$
\begin{aligned}
&r((k+1, \boldsymbol{\theta}_{k+1}), (k, \boldsymbol{\theta}_k))\\
&\quad = \frac{\pi(k, \boldsymbol{\theta}_k \mid \boldsymbol{y}_{1:T})}{\pi(k+1, \boldsymbol{\theta}_{k+1} \mid \boldsymbol{y}_{1:T})} \frac{\rho_{k,k+1}(\boldsymbol{\theta}_k)k^{-1}\pi(u_1)\pi(u_2)}{\rho_{k+1,k}(\boldsymbol{\theta}_{k+1})k^{-1}}.\\
&\qquad \times \frac{\pi(u_3)\pi(u_4)\pi(\varepsilon_{c_1})\pi(\varepsilon_{s_1})\pi(\varepsilon_{c_2})\pi(\varepsilon_{s_2})}{\pi(u_0)\pi(\varepsilon_c)\pi(\varepsilon_s)} \mathscr{J}_{f_{k+1,k}}.
\end{aligned}
\tag{67}
$$

4.1.5. *Example of simulation*

We present here results for the case of two sinusoids ($N = 2$) with parameters given in Table 1, where $E_i = a_{s_i}^2 + a_{c_i}^2$. The number of observed samples was 64.

We do not discuss here the choice of δ^2 and Λ, which are estimated from the data (see [1]). We ran the algorithm for 20 000 iterations. Here we present estimators of different quantities of interest, aiming at performing model selection. In Figs. 1 and 2 we present the observed data and the posterior distribution of the models that allows us for example to choose the most probable model, and then conditional upon the knowledge of the "best" model to display the frequencies for the model with two sinusoids. The most probably frequencies for these two sinusoids can then be estimated from these histograms. Other estimators might be considered. For example if one is interested in estimating the original signal $\sum_{j=1}^{k}(a_{c_{j,k}}\cos[\omega_{j,k}t] + a_{s_{j,k}}\sin[\omega_{j,k}t])$ one might then consider the following estimator $\mathbb{E}(\sum_{j=1}^{k}(a_{c_{j,k}}\cos[\omega_{j,k}t] + a_{s_{j,k}}\sin[\omega_{j,k}t]) \mid \boldsymbol{y}_{1:T})$.

[7] If the birth/death move is also used, then one should introduce the notation $f^{[1]}_{k+1,k}(\cdot,\cdot), f^{[2]}_{k+1,k}(\cdot,\cdot)$ and $\rho^{[1]}_{k,k+1}, \rho^{[2]}_{k,k+1} \ldots$ but we do not, to simplify notation.

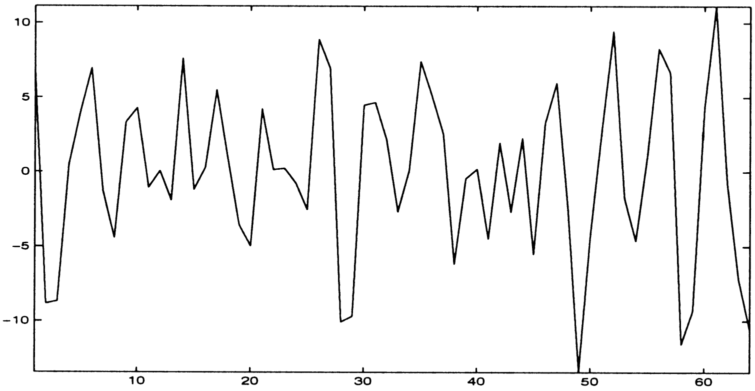

Fig. 1. Noisy observations.

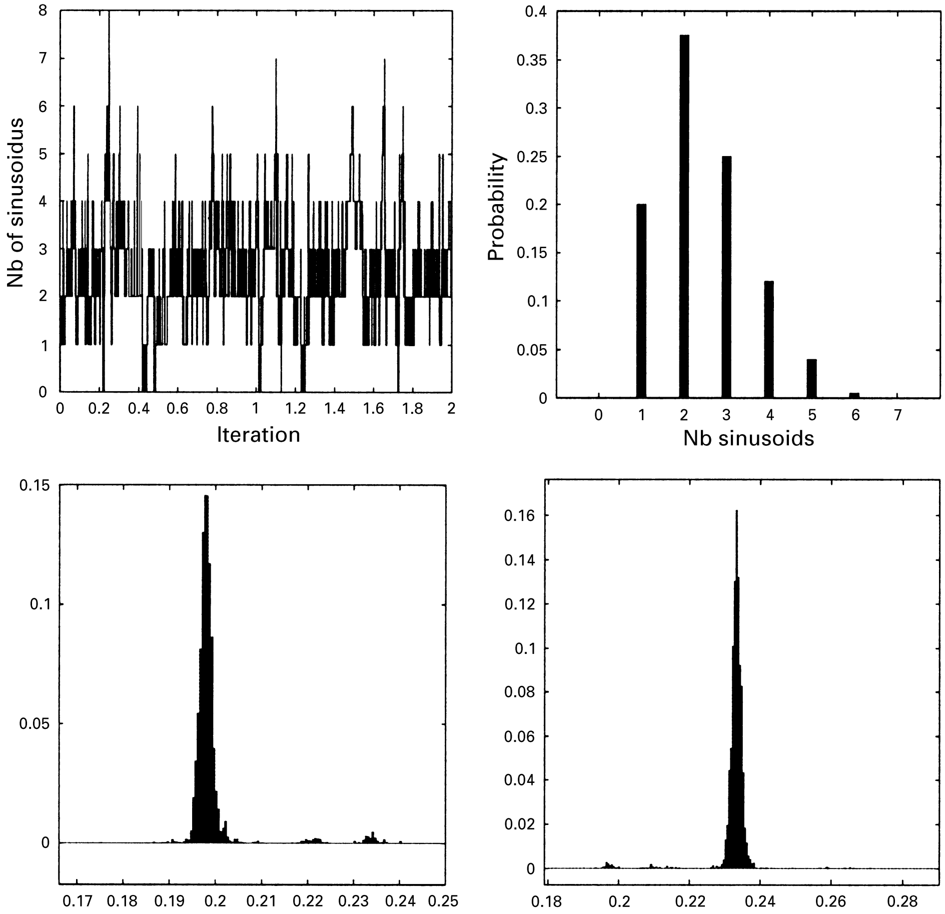

Fig. 2. Top: the component of the Markov chain corresponding to the dimension, and an estimate of $\pi(k\,|\,\boldsymbol{y}_{1:T})$. Bottom: estimation of $\pi(\omega/2\pi\,|\,\boldsymbol{y}_{1:T})$.

4.2. Bernoulli–Gauss deconvolution

4.2.1. Model of the data

In this model selection problem, we have to choose among the following data models:[8]

$$\mathscr{M}_{r_{1:T}}:\begin{cases} x_t = \begin{cases} \sum_{i=1}^{k_{\mathrm{AR}}} a_i x_{t-i} + v_t & \text{if } r_t = 1, \\ \sum_{i=1}^{k_{\mathrm{AR}}} a_i x_{t-i} & \text{if } r_t = 0, \end{cases} \\ y_t = x_t + w_t \end{cases} \tag{68}$$

and the space where the posterior distribution is defined is given by (8). We assume that $(x_0, \dots, x_{1-k_{\mathrm{AR}}}) = (0, \dots, 0)$. To simplify the presentation and focus on the model selection problem, we suppose that σ_v^2, $(a_i)_{i=1,\dots,k_{\mathrm{AR}}}$ and σ_w^2 are known. Note that these parameters could be estimated without any difficulty. We want to estimate the joint posterior distribution $\pi(\boldsymbol{r}_{1:T}, \mathrm{d}\boldsymbol{v}_{1:T} \,|\, \boldsymbol{y}_{1:T})$

$$\pi(\boldsymbol{r}_{1:T}, \mathrm{d}\boldsymbol{v}_{1:T} \,|\, \boldsymbol{y}_{1:T}) = \sum_{n=0}^{2^T-1} \pi_n(\boldsymbol{r}_{1:T}(n), \mathrm{d}\boldsymbol{v}_{1:T} \,|\, \boldsymbol{y}_{1:T}) \times \mathbb{I}_{\{n\}\times\boldsymbol{\Theta}_n}\left(\sum_{t=0}^{T-1} 2^t r_t(n), \boldsymbol{v}_{1:T}\right) \tag{69}$$

using MCMC, where $n = \sum_{t=1}^{T} 2^t r_t$, and $\boldsymbol{r}_{1:T}(n)$ denotes the corresponding sequence. More precisely,

$$\begin{aligned} &\pi_n(\boldsymbol{r}_{1:T}(n), \mathrm{d}\boldsymbol{v}_{1:T} \,|\, \boldsymbol{y}_{1:T}) \\ &\quad\propto \pi(\boldsymbol{y}_{1:T} \,|\, \boldsymbol{r}_{1:T}(n), \boldsymbol{v}_{1:T})\pi(\mathrm{d}\boldsymbol{v}_{1:T} \,|\, \boldsymbol{r}_{1:T}(n))\pi(\boldsymbol{r}_{1:T}(n)) \\ &\quad\propto \pi(\boldsymbol{y}_{1:T} \,|\, \boldsymbol{v}_{1:T})\pi(\mathrm{d}\boldsymbol{v}_{1:T} \,|\, \boldsymbol{r}_{1:T}(n))\pi(\boldsymbol{r}_{1:T}(n)) \end{aligned} \tag{70}$$

and $\pi(\mathrm{d}\boldsymbol{v}_{1:T} \,|\, \boldsymbol{r}_{1:T}(n)) = \prod_{t=1}^{T} \pi(\mathrm{d}v_t \,|\, r_t(n))$, where

$$\pi(\mathrm{d}v_t \,|\, r_t) = (1-\lambda)\delta_0(\mathrm{d}v_t)\mathbb{I}_{\{0\}}(r_t) + \frac{\lambda}{\sqrt{2\pi\sigma_0}} \exp\left(-\frac{v_t^2}{2\sigma_0^2}\right)\mathrm{d}v_t \mathbb{I}_{\{1\}}(r_t). \tag{71}$$

We present here an algorithm which stresses on the fact that the posterior distribution is defined on 2^T different subspaces $\boldsymbol{\Theta}_n$.

4.2.2. Algorithm

The algorithm we have chosen, consists of the following steps:

(1) Select t at random according to $\mathscr{U}_{\{1,\dots,T\}}$.

(2) If $r_t = 0$, then propose a new value of r_t, $r_t^* \sim q_t(r_t^* \,|\, \boldsymbol{r}_{1:T}, \boldsymbol{v}_{1:T}, \boldsymbol{y}_{1:T})$, so that $\rho_{k,k+2^t}(\boldsymbol{v}_{1:T}, \boldsymbol{r}_{1:T}) = q_t(1 \,|\, \boldsymbol{r}_{1:T}, \boldsymbol{v}_{1:T}, \boldsymbol{y}_{1:T})/T$ and $\rho_{k,k}(\boldsymbol{v}_{1:T}, \boldsymbol{r}_{1:T}) = q_t(0 \,|\, \boldsymbol{r}_{1:T}, \boldsymbol{v}_{1:T}, \boldsymbol{y}_{1:T})/T$.

(3) If $r_t = 1$, then propose a new value of r_t, $r_t^* \sim q_t(r_t^* \,|\, \boldsymbol{r}_{1:T}, \boldsymbol{v}_{1:T}, \boldsymbol{y}_{1:T})$, so that $\rho_{k,k-2^t}(\boldsymbol{v}_{1:T}, \boldsymbol{r}_{1:T}) = q_t(0 \,|\, \boldsymbol{r}_{1:T}, \boldsymbol{v}_{1:T}, \boldsymbol{y}_{1:T})/T$ and $\rho_{k,k}(\boldsymbol{v}_{1:T}, \boldsymbol{r}_{1:T}) = q_t(1 \,|\, \boldsymbol{r}_{1:T}, \boldsymbol{v}_{1:T}, \boldsymbol{y}_{1:T})/T$.

(4) Propose v_t^* to replace v_t, with $v_t^* \sim q_t(\mathrm{d}v_t^* \,|\, r_t^*, \boldsymbol{r}_{1:T}, \boldsymbol{v}_{1:T}, \boldsymbol{y}_{1:T})$ which admits $\mathbb{R}$ as support if $r_t^* = 0$ or $\{0\}$ if $r_t^* = 1$.

We now describe the algorithm more precisely, starting with the main procedure:

```
● Initialization
● Iteration $i$
    ○ Sample $t \sim \mathscr{U}_{\{1,\dots,T\}}$ and
      $r_t^* \sim q_t(r_t^* \,|\, \boldsymbol{r}_{1:T}^{(i-1)}, \boldsymbol{v}_{1:T}^{(i-1)}, \boldsymbol{y}_{1:T})$.
    ○ If $r_t^{(i-1)} = r_t^*$ Move from $\boldsymbol{\Theta}_k$ to $\boldsymbol{\Theta}_k$.
    ○ Else
            If $r_t^{(i-1)} < r_t^*$ then Move from $\boldsymbol{\Theta}_k$ to $\boldsymbol{\Theta}_{k+2^t}$
            Else Move from $\boldsymbol{\Theta}_k$ to $\boldsymbol{\Theta}_{k-2^t}$
```

where the procedures Move from $\boldsymbol{\Theta}_k$ to $\boldsymbol{\Theta}_{k+2^t}$ and Move from $\boldsymbol{\Theta}_k$ to $\boldsymbol{\Theta}_{k-2^t}$ are described below. The procedure Move from $\boldsymbol{\Theta}_k$ to $\boldsymbol{\Theta}_k$ corresponds to a standard MH step and is not described in details here.

Assume that the current model is k and that $(r_t^{(i-1)}, v_t^{(i-1)}) \in \{0\} \times \{0\}$ and $r_t^* = 1$, then,

[8] Note that we do not use the dangerous notation $x_t = \sum_{i=1}^{k_{\mathrm{AR}}} a_i x_{t-i} + r_t v_t$, which might suggest that when $r_t = 0$ then v_t can take any value.

Move from $\boldsymbol{\Theta}_k$ to $\boldsymbol{\Theta}_{k+2^t}$

- Propose a candidate v_t^* where $v_t^* \sim q_t(\mathrm{d}v_t^* \,|\, r_t^*, \boldsymbol{r}_{1:T}^{(i-1)}, \boldsymbol{v}_{1:T}^{(i-1)}, \boldsymbol{y}_{1:T})$.
- Set $(r_t^{(i)}, v_t^{(i)}) = (1, v_t^*)$ with probability $\min\{1,\alpha\}$ with

$$\alpha = \frac{\pi(\boldsymbol{r}_{-t}^{(i-1)}, \boldsymbol{v}_{-t}^{(i-1)}, 1, v_t^* \,|\, \boldsymbol{y}_{1:T}) q_t(0 \,|\, 1, v_t^*, \boldsymbol{r}_{-t}^{(i-1)}, \boldsymbol{v}_{-t}^{(i-1)} \boldsymbol{y}_{1:T})}{\pi(\boldsymbol{r}_{-t}^{(i-1)}, \boldsymbol{v}_{-t}^{(i-1)}, 0,0 \,|\, \boldsymbol{y}_{1:T}) q_t(v_t^* \,|\, 1, \boldsymbol{r}_{1:T}^{(i-1)}, \boldsymbol{v}_{1:T}^{(i-1)} \boldsymbol{y}_{1:T}) q_t(1 \,|\, \boldsymbol{r}_{1:T}^{(i-1)}, \boldsymbol{v}_{1:T}^{(i-1)} \boldsymbol{y}_{1:T})} \tag{72}$$

otherwise $(r_t^{(i)}, v_t^{(i)}) = (r_t^{(i-1)}, v_t^{(i-1)})$.

Similarly, assume that the current model is k and that $(r_t^{(i-1)}, v_t^{(i-1)}) \in \{1\} \times \mathbb{R}$ and $r_t^* = 0$, then,

Move from $\boldsymbol{\Theta}_k$ to $\boldsymbol{\Theta}_{k-2^t}$

- Propose the candidate $\{0\}$.
- Set $(r_t^{(i)}, v_t^{(i)}) = (0,0)$ with probability $\min\{1,\alpha\}$ with

$$\alpha = \frac{\pi(\boldsymbol{r}_{-t}^{(i-1)}, \boldsymbol{v}_{-t}^{(i-1)}, 0,0 \,|\, \boldsymbol{y}_{1:T}) q_t(v_t^{(i-1)} \,|\, 0,0, \boldsymbol{r}_{-t}^{(i-1)}, \boldsymbol{v}_{-t}^{(i-1)} \boldsymbol{y}_{1:T}) q_t(1 \,|\, 0,0, \boldsymbol{r}_{-t}^{(i-1)}, \boldsymbol{v}_{-t}^{(i-1)} \boldsymbol{y}_{1:T})}{\pi(\boldsymbol{r}_{1:T}^{(i-1)}, \boldsymbol{v}_{1:T}^{(i-1)} \,|\, \boldsymbol{y}_{1:T}) q_t(0 \,|\, \boldsymbol{r}_{1:T}^{(i-1)}, \boldsymbol{v}_{1:T}^{(i-1)} \boldsymbol{y}_{1:T})} \tag{73}$$

otherwise $(r_t^{(i)}, v_t^{(i)}) = (r_t^{(i-1)}, v_t^{(i-1)})$.

Note that in the special case where $q_t(r_t^* \,|\, \boldsymbol{r}_{1:T}, \boldsymbol{v}_{1:T}, \boldsymbol{y}_{1:T}) = p(r_t^* \,|\, \boldsymbol{r}_{-t}, \boldsymbol{y}_{1:T})$ and $q_t(\mathrm{d}v_t^* \,|\, \boldsymbol{r}_{1:T}, \boldsymbol{v}_{1:T}, \boldsymbol{y}_{1:T}) = p(\mathrm{d}v_t^* \,|\, \boldsymbol{r}_{1:T}, \boldsymbol{v}_{-t}, \boldsymbol{y}_{1:T})$ then the acceptance probabilities (72) and (73) are equal to 1. In practice, it is possible to evaluate the values $p(r_t^* \,|\, \boldsymbol{r}_{-t}, \boldsymbol{y}_{1:T})$ using a Kalman filter, and thus to sample from this discrete distribution. Sampling from $p(\mathrm{d}v_t^* \,|\, \boldsymbol{r}_{1:T}, \boldsymbol{v}_{-t}, \boldsymbol{y}_{1:T})$ is standard as it is either a Gaussian distribution when $r_t = 1$ or a delta Dirac mass on 0 when $r_t = 0$. It is very important here to observe that the algorithm boils down to a random scan Gibbs sampler [36] *due to the analytical properties of the model.* If it was impossible to sample exactly from $p(r_t^* \,|\, \boldsymbol{r}_{-t}, \boldsymbol{y}_{1:T})$ then this would not be the case, and the use of the reversible jump algorithm described above would then be unavoidable. To conclude, note that as shown in [17] it is possible to dramatically reduce the complexity of this algorithm by replacing this random scan Gibbs sampler by a deterministic scan Gibbs sampler, see [17] for details.

5. Conclusions

MCMC sampling is a powerful methodology for signal processing which has little been exploited in the signal processing community. With the recent advances of the theory of MCMC computations, this methodology has been generalized to allow for simultaneous selection of models and the estimation of their parameters. This has become possible once algorithms for sampling from target distributions defined over joint sample spaces of models and their parameters had been developed.

The main objective of this paper was to provide a summary of the theory and present examples of how one might apply it. Special care has been taken to pinpoint the subtleties of jumping from one parameter space to another, and in general, to show the construction of MCMC samplers in such scenarios. The focus in the paper was on the reversible jump MCMC algorithm because it is the most widely used of all existing methods; it is easy to use and is flexible and has nice properties. Many references have been cited, with the emphasis being given to articles with signal processing applications.

Acknowledgements

The authors would like to thank, J.F.G. de Freitas, J.R. Hopgood, E. Punskaya, P.T. Troughton and P.J. Walmsley, for their careful reading of an early version of this manuscript and helpful comments that have helped improving the present tutorial.

References

[1] C. Andrieu, A. Doucet, Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC, IEEE Trans. Signal Process. 47 (10) (October 1999) 2667–2676.

[2] C. Andrieu, A. Doucet, W.J. Fitzgerald, On Monte Carlo methods for Bayesian data analysis, in: A. Mees, R.L. Smith (Eds.), Nonlinear Dynamics and Statistics, Birkhäuser, 2000.

[3] C. Andrieu, A. Doucet, W.J. Fitzgerald, J.M. Peréz, Bayesian computational approaches to model selection, in: W.J. Fitzgerald, R.L. Smith, P.C. Young, A. Walden (Eds.), Nonlinear and Non Gaussian Signal Processing, Newton Institute Series, Cambridge University Press, Cambridge, 2000.

[4] C. Andrieu, J.F.G. de Freitas, A. Doucet, Robust full Bayesian learning for neural networks, IEEE Trans. Neural Networks, submitted for publication, available as Technical Report CUED/F-INFENG/TR. 343, University of Cambridge, UK, 1999.

[5] S.A. Barker, P.J.W. Rayner, Unsupervised image segmentation, Proceedings of ICASSP, Seattle, 1998.

[8] B.P. Carlin, S. Chib, Bayesian model choice via Markov chain Monte Carlo, J. Roy. Stat. Soc. B 57 (1995) 473–484.

[10] E. Clark, A. Quinn, A data driven Bayesian sampling scheme for unsupervised image segmentation, Proceedings of ICASSP, Vol. 6, Phoenix, 1999.

[11] P. Dellaportas, J.J. Forster, I. Ntzoufras, On Bayesian model and variable selection using MCMC, paper based upon a talk presented at the HSSS Workshop on Variable Dimension MCMC, New Forest, September 1997. Available from http://www.stat-athens.aueb.gr/~ptd/gvs.ps

[13] P.M. Djurić, Bayesian methods for signal processing, IEEE Signal Process. Mag. 15 (5) (1998) 26–28.

[14] P.M. Djurić, Variable selection by a reversible jump MCMC approach, Proceedings of EUSIPCO, Vol. 4, The Island of Rhodes, Greece, 1998, pp. 2013–2016.

[15] P.M. Djurić, S.J. Godsill, Parametric modeling and estimation of time-varying spectra, Proceedings of Asilomar, Asilomar, CA, 1998, pp. 292–296.

[16] P.M. Djurić, S.J. Godsill, W.J. Fitzgerald, P.J.W. Rayner, Detection and estimation of signals by reversible jump Markov chain Monte Carlo computations, Proceedings of ICASSP, Vol. 4, Seattle, 1998, pp. 2269–2272.

[17] A. Doucet, C. Andrieu, Iterative algorithms for optimal state estimation of jump Markov linear systems, Proceedings of ICASSP, Phoenix, Arizona, 1999.

[19] W.J. Fitzgerald, Signal processing applications of Markov chain Monte Carlo methods, Signal Processing, this issue.

[20] D. Gamerman, Markov Chain Monte Carlo, Chapman & Hall, London, 1997.

[21] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, Chapman & Hall, London, 1995.

[23] W.R. Gilks, S. Richardson, D.J. Spiegelhalter (Eds.), Markov Chain Monte Carlo in Practice, Chapman & Hall, London, 1996.

[24] S. Godsill, Some new relationships between MCMC model uncertainty methods, Technical Report CUED/F-INFENG/TR. 305, University of Cambridge, December 1997.

[25] U. Grenander, Tutorial in pattern theory, Division of Applied Mathematics, Brown University, Providence, RI, 1983.

[26] U. Grenander, M.I. Miller, Representation of knowledge in complex systems (with discussion), J. Roy. Stat. Soc. B 56 (1994) 549–603.

[27] P.J. Green, Reversible jump MCMC computation and Bayesian model determination, Biometrika 82 (1995) 711–732.

[28] A.D. Lanterman, M.I. Miller, D.L. Snyder, General Metropolis-Hastings jump-diffusions for automatic target recognition in infrared scenes, Opt. Eng. 36 (1997) 1123–1137.

[29] M.I. Miller, U. Grenander, J.A. O'Sullivan, D.L. Snyder, Automatic target recognition organized via jump-diffusion algorithms, IEEE Trans. Image Process. 6 (1997) 157–174.

[30] M.I. Miller, A. Srivastava, U. Grenander, Conditional-mean estimation via jump-diffusion processes in multiple target tracking/recognition, IEEE Trans. Signal Process. 43 (1995) 2678–2690.

[31] J.J.K. O'Ruanaidh, W.J. Fitzgerald, Numerical Bayesian Methods Applied to Signal Processing, Springer, Berlin, 1996.

[32] D.B. Philips, A.F.M. Smith, Bayesian model comparison via jump diffusions, in: Gilks, Richardson and Spiegelhalter, eds., Markov Chain Monte Carlo in Practice, Chapman & Hall, London, 1996, pp. 214–239.

[33] C.J. Preston, Spatial birth-and-death processes, Bull. Inst. Int. Stat. 39 (1976) 177–212.

[34] S. Richardson, P.J. Green, On Bayesian analysis of mixtures with unknown number of components, J. Roy. Stat. Soc. B 59 (4) (1997) 731–792.

[35] B.D. Ripley, Modelling spatial patterns (with discussion), J. Roy. Stat. Soc. B 39 (1977) 172–212.

[36] C.P. Robert, G. Casella, Monte Carlo Statistical Methods, Springer Series in Statistics, Springer, New York, 1999.

[37] G.O. Roberts, R.L. Tweedie, Exponential convergence of Langevin diffusions and their discrete approximations, Bernoulli 2 (1996) 341–364.

[38] A.F.M. Smith, G.O. Roberts, Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods, J. Roy. Stat. Soc. B 55 (1993) 3–23.

[40] D.D. Sworder, J.E. Boyd, Jump-diffusion in tracking/recognition, IEEE Trans. Signal Process. 46 (1998) 235–239.

[42] L. Tierney, Markov chains for exploring posterior distributions (with discussion), Ann. Stat. 22, 1994, 1701–1762.

[43] P.J. Troughton, S.J. Godsill, A reversible jump sampler for autoregressive time-series, Proceedings of ICASSP, Vol. 4, Seattle, 1998, pp. 2257–2260.

[44] P.J. Troughton, S.J. Godsill, MCMC methods for restoration of nonlinear distorted autoregressive signals, Proceedings of EUSIPCO, The Island of Rhodes, 1998, pp. 540–543.