An empirical Bayes approach for joint Bayesian model selection and estimation of sinusoids via reversible jump MCMC

Alireza Roodaki, Julien Bect, Gilles Fleury

AN EMPIRICAL BAYES APPROACH FOR JOINT BAYESIAN MODEL SELECTION AND ESTIMATION OF SINUSOIDS VIA REVERSIBLE JUMP MCMC

Alireza Roodaki, Julien Bect, and Gilles Fleury

E3S — SUPELEC Systems Sciences
Dept. of Signal Processing and Electronic Systems, SUPELEC, Gif-sur-Yvette, France.
Email: {alireza.roodaki, julien.bect, gilles.fleury}@supelec.fr

ABSTRACT

This paper addresses the sensitivity of the algorithm proposed by Andrieu and Doucet (IEEE Trans. Signal Process., 47(10), 1999), for the joint Bayesian model selection and estimation of sinusoids in white Gaussian noise, to the values of a certain hyperparameter claimed to be weakly influential in the original paper. A deeper study of this issue reveals indeed that the value of this hyperparameter (the scale parameter of the expected signal-to-noise ratio) has a significant influence on 1) the mixing rate of the Markov chain and 2) the posterior distribution of the number of components. As a possible workaround for this problem, we investigate an Empirical Bayes approach to select an appropriate value for this hyperparameter in a data-driven way. Marginal likelihood maximization is performed by means of an importance sampling based Monte Carlo EM (MCEM) algorithm. Numerical experiments illustrate that the sampler equipped with this MCEM procedure provides satisfactory performances in moderate to high SNR situations.

1. INTRODUCTION

In this paper, we address the problem of detection and estimation of sinusoids in white Gaussian noise, assuming that the number of component is unknown. A fully Bayesian algorithm, based on the Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) technique [8, 9], has been proposed for this problem in [1]. Similar algorithms have also been used for other applications such as polyphonic signal analysis [3], array signal processing [12], and nuclear emission spectra analysis [10]. However, to the best of our knowledge, the sensitivity of the algorithm to the value of its hyperparameters has never been clearly discussed.

Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)^t$ be a vector of N observations of an observed signal. We consider the finite family of embedded models $\{\mathscr{M}_k,\ 0 \leq k \leq k_{\max}\}$, where $\mathscr{M}_k$ assumes that $\mathbf{y}$ can be written as a linear combination of k sinusoids observed in white Gaussian noise. Let $\boldsymbol{\omega}_k = (\omega_{1,k}, \ldots, \omega_{k,k})$ be the vector of radial frequencies in model $\mathscr{M}_k$, and let $\mathbf{D}_k$ be the corresponding $N \times 2k$ design matrix defined by

$$\mathbf{D}_k(i+1, 2j-1) \triangleq \cos(\omega_{j,k} i), \quad \mathbf{D}_k(i+1, 2j) \triangleq \sin(\omega_{j,k} i)$$

for $i = 0, \ldots, N-1$ and $j = 1, \ldots, k$. Then the observed signal $\mathbf{y}$ follows under $\mathscr{M}_k$ a normal linear regression model:

$$\mathbf{y} = \mathbf{D}_k . \mathbf{a}_k + \mathbf{n},$$

where $\mathbf{n}$ is a white Gaussian noise with variance σ^2. The unknown parameters are assumed to be the number of components k and $\boldsymbol{\theta}_k = \{\mathbf{a}_k, \boldsymbol{\omega}_k, \sigma^2\}$.

Assuming that no (or little) information is available about the vector of amplitudes $\mathbf{a}_k$, the conditionally conjugate g-prior is usually recommended as a default prior in the Bayesian variable selection literature [14, 21]. Under this prior, the distribution of $\mathbf{a}_k$ conditionally to σ^2, k and $\boldsymbol{\omega}_k$ is Gaussian with $\sigma^2/g\,(\mathbf{D}_k^t \mathbf{D}_k)^{-1}$ as its covariance matrix, where g is a positive parameter. Following [1], a zero-mean g-prior for $\mathbf{a}_k$ will be used in this paper. Our results, however, are likely to remain relevant for any covariance matrix of the form $\sigma^2/g\,\Sigma_k$ (with Σ_k possibly depending on k and $\boldsymbol{\omega}_k$).

The parameter $\delta^2 = 1/g$, called the Expected SNR (ESNR), controls the expected size of the amplitudes. Owing to its influence on the performance of the algorithm, and assuming again that no (or little) information is available, the hyperparameter δ^2 is given in [1] a conjugate inverse gamma prior with parameters α_{δ^2} and β_{δ^2}, that we denote by $\mathscr{IG}(\alpha_{\delta^2}, \beta_{\delta^2})$. Such a hierarchical Bayes approach is usually hoped to increase the robustness of the statistical analysis; see [18, Section 10.2] for more information. The first parameter is set to $\alpha_{\delta^2} = 2$, in order to have an heavy-tailed "weakly informative" prior (with infinite variance). It is claimed in [1, Section V.D] that the value of β_{δ^2} has a weak influence on the performance of the algorithm.

The contribution of this paper, which can be seen as a continuation of [1], is twofold. First, on the basis of extensive numerical experiments, we argue that the value of β_{δ^2} can have a strong influence on 1) the mixing rate of the Markov chain and 2) the posterior distribution of the number of components. Second, instead of using a fixed value for the hyperparameter β_{δ^2}, we investigate the capability of an Empirical Bayes (EB) approach to estimate it from the data, in the spirit of the approach used in [2, 6] to estimate δ^2. More precisely, since the marginal likelihood of β_{δ^2} is not available in closed form, we implement an Importance Sampling (IS) based Monte Carlo Expectation Maximization (MCEM) algorithm [13, 20] to maximize it numerically.

The paper is outlined as follows. Section 2 recalls the hierarchical Bayesian model and the RJ-MCMC sampler proposed in [1]. Section 3 discusses the influence of β_{δ^2} on both the mixing rate of the Markov chain and the posterior distribution of the number k of components. Section 4 explains the fundamentals of the MCEM algorithm, which is used for estimating β_{δ^2}. Section 5 presents the results of our numerical experiments and discusses the pros and cons of the Empirical Bayes approach in estimating β_{δ^2}. Finally, Section 6 concludes the paper and gives directions for future work.

2. BAYESIAN FRAMEWORK

This section describes the prior distribution and the RJ-MCMC sampler considered in this paper, following [1] unless explicitly stated otherwise.

2.1 Prior distributions

The joint prior distribution of the unknown parameters is chosen to have the following hierarchical structure:

$$\begin{aligned} p\left(k, \boldsymbol{\theta}_k, \delta^2\right) \;=\; & p\left(\mathbf{a}_k \mid k, \boldsymbol{\omega}_k, \sigma^2, \delta^2\right)\, p\left(\boldsymbol{\omega}_k \mid k\right) \\ & \times p(k)\; p\left(\sigma^2\right)\; p\left(\delta^2\right). \end{aligned} \tag{1}$$

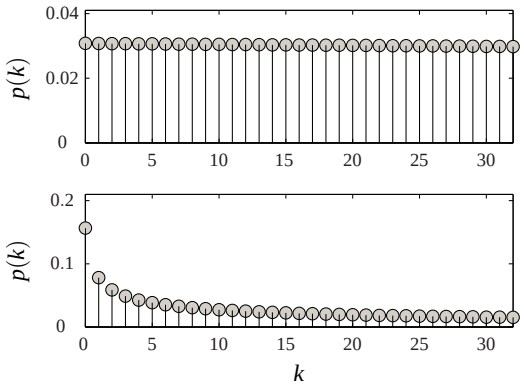

Figure 1: Truncated negative binomial prior on k corresponding to $\alpha_\Lambda = 1.0$ (upper plot) and $\alpha_\Lambda = 0.5$ (lower plot), with $k_{\max} = 32$ and $\beta_\Lambda = 0.001$.

The conditional distribution of $\mathbf{a}_k$ is the g-prior distribution already described in the introduction. Conditional on k, the components of $\boldsymbol{\omega}_k$ are independent and identically distributed, with a uniform distribution on $(0,\pi)$. The noise variance σ^2 is endowed with Jeffrey's improper prior, i.e. $p(\sigma^2) \propto 1/\sigma^2$, where the symbol $\propto$ denotes proportionality.

The prior distribution of k is defined in [1] in two steps, following once again the hierarchical Bayes philosophy. First, k is given a Poisson distribution with mean Λ, truncated to $\{0,1,\ldots,k_{\max}\}$. Then, to increase the robustness of the inference in a context of weak prior information on k, the hyperparameter Λ is given a conjugate Gamma prior, with shape parameter $\alpha_\Lambda \approx \frac{1}{2}$ and scale parameter $\beta_\Lambda \approx 0$. This is equivalent to using for k a (truncated) negative binomial prior[1] that puts a strong emphasis on small values. In this paper, we set $\alpha_\Lambda = 1$ in order to have an almost flat prior for k over $\{0,\ldots,k_{\max}\}$; see Figure 1 for a comparison of the two prior distributions.

2.2 Sampling structure

The hierarchical structure and prior distributions just described make it possible to integrate parameters $\mathbf{a}_k$ and σ^2 out of the posterior distribution analytically. This *marginalization* step [17] yields the following marginal posterior distribution:

$$p\left(k,\boldsymbol{\omega}_k,\delta^2,\Lambda\,|\,\mathbf{y}\right) \propto \left(\mathbf{y}^t\mathbf{P}_k\mathbf{y}\right)^{-N/2} \frac{\Lambda^k\pi^{-k}}{k!\,(\delta^2+1)^k} \times p(\delta^2)\,p(\Lambda)\,\mathbb{1}_{(0,\pi)^k}(\boldsymbol{\omega}_k)\,, \tag{2}$$

with

$$\mathbf{P}_k \;=\; \mathbf{I}_N - \frac{\delta^2}{1+\delta^2}\,\mathbf{D}_k\left(\mathbf{D}_k^t\mathbf{D}_k\right)^{-1}\mathbf{D}_k^t$$

when $k \geq 1$ and $\mathbf{P}_0 = \mathbf{I}_N$.

The joint posterior distribution (2) is the target distribution of the RJ-MCMC sampler. In the following, different steps for sampling from the target distribution are briefly described. For more detailed expressions please refer to [1, 8].

The RJ-MCMC sampler, that leaves the target density (2) invariant, consists of a Metropolis-Hastings (MH) move for updating the value of k and $\boldsymbol{\omega}_k$, followed by a sequence of Gibbs moves to update δ^2 and Λ. (The conditional distribution of δ^2 given k, $\boldsymbol{\omega}_k$, Λ and $\mathbf{y}$ is sampled from by first *demarginalizing* [17] σ^2 and $\mathbf{a}_k$ and then sampling from the full conditional distribution.)

Since the problem under consideration is trans-dimensional, the proposal distribution for the MH move updating k and $\boldsymbol{\omega}_k$ is in fact a mixture of proposal distributions performing within-model moves (updating radial frequencies without changing k) and between-models moves ("birth" and "death" moves, which respectively add and remove components). Except for a modification described below, the moves implemented in our sampler are the same as in [1].

2.3 Correction of the birth ratio in [1]

In the birth move proposed in [1], and also used in this paper, the insertion of a new sinusoid is proposed as follows: first a new radial frequency is sampled from the uniform distribution on $(0,\pi)$ and, then, it is inserted at a random location[2] among the existing ones. According the theory of RJ-MCMC samplers [8] and using the same proportion of birth and death moves as in [1], the move is accepted with probability $\alpha_{\text{birth}} = \min\{1, r_{\text{birth}}\}$, where

$$r_{\text{birth}} = \left(\frac{\mathbf{y}^t\mathbf{P}_{k+1}\mathbf{y}}{\mathbf{y}^t\mathbf{P}_k\mathbf{y}}\right)^{-N/2} \frac{1}{1+\delta^2}\,. \tag{3}$$

One should note that the birth ratio computed in [1] differs from (3) by a $1/(k+1)$ factor. A similar mistake in computing RJ-MCMC ratios has been reported in the field of genetics [11]. Note that this additional factor is equivalent to using a different prior distribution over k. A detailed justification of (3) will be provided in a forthcoming paper.

3. SENSITIVITY OF THE ALGORITHM TO β_{δ^2}

This section first reviews related work concerning the role of δ^2 in the Bayesian variable selection literature, and then proceeds to describing the role of β_{δ^2} in the present problem.

3.1 Review of related work in Bayesian variable selection

It has been highlighted in the variable selection literature that the parameter δ^2, which controls the expected relative size of the amplitudes with respect to σ, implicitly defines a "dimensionality penalty" from the model selection point of view [2, 6]. Indeed, considering that $p(k)$ is approximately constant for $k \in [0, k_{\max}]$, we have

$$\log p\left(k,\boldsymbol{\omega}_k\,|\,\mathbf{y},\delta^2\right) \;\approx\; -\frac{N}{2}\log\left(\mathbf{y}^t\mathbf{P}_k\mathbf{y}\right) - F\cdot k + C, \tag{4}$$

where $F = \log\left(\pi\left(1+\delta^2\right)\right)$ and C is a constant which does not depend on k and $\boldsymbol{\omega}_k$. F can be interpreted as a dimensionality penalty, which penalizes complex models. Thus, δ^2 plays the role of a regularization parameter, "large" values of which favor sparse signal representations at the expense of detection sensitivity. Conversely, "small" values of δ^2 typically lead to the selection of overfitting models (i.e., in terms of detection performance, false positives).

In the Bayesian variable selection literature, many researchers have tried to either set an appropriate fixed value to δ^2 or estimate it using different approaches. In [4], several fixed values for δ^2 are compared in a model averaging framework, and $\delta^2 = \max\{N, p^2\}$ is recommended as a default ("benchmark") value, where p denotes the number of variables. Several approaches for the estimation of δ^2, both EB or fully Bayesian, have been proposed and compared

[1] Indeed, the marginal prior distribution of k is given by

$$p(k) = \frac{\Gamma(k+\alpha_\Lambda)}{\Gamma(\alpha_\Lambda)\,k!}\left(\frac{\beta_\Lambda}{\beta_\Lambda+1}\right)^{\alpha_\Lambda}\left(\frac{1}{\beta_\Lambda+1}\right)^{k},$$

which is a negative binomial distribution. See, e.g., [5, Section 2.7 and 17.2], where the negative binomial distribution is advocated as a robust alternative to the Poisson distribution.

[2] Note that the same ratio would be obtained if the radial frequency were sorted instead [16].

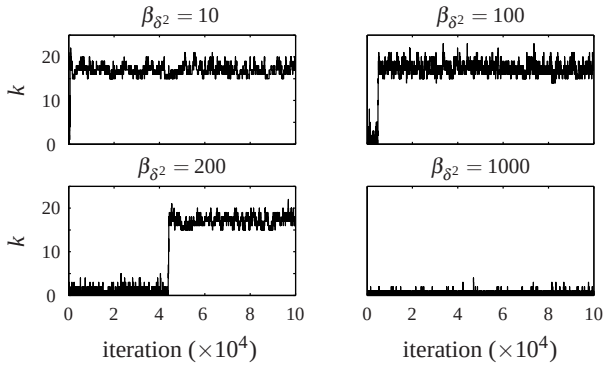

Figure 2: Mixing of the chain for different values of β_{δ^2}. The true model is $\mathscr{M}_{15}$, and the sampler is initialized in $\mathscr{M}_0$.

in [2, 6, 14]. It is concluded in [2] that the Maximum Marginal Likelihood (MML) approach is superior to the others (in terms of mean square error), but the conclusions of [14]—in a slightly different setting—suggest that some fully Bayesian approaches can perform just as well.

3.2 Role of β_{δ^2}

Our numerical experiments have revealed that the value of β_{δ^2} can have a significant influence on 1) the posterior distribution of the number of components and 2) the convergence rate of the Markov chain.

The former fact can be understood in light of Section 3.1 where the role of δ^2 as a dimensionality penalty has been highlighted. Indeed, since β_{δ^2} is a scale parameter for the prior distribution of δ^2, it can be expected that, probably to a lesser extent, β_{δ^2} should play a similar role. In other words, high values of β_{δ^2} are expected to favor sparse solutions, with a risk of omitting low SNR components, whereas low values of β_{δ^2} are expected to allow solutions with many components (high values of k). This point will be further discussed in Section 5 on the basis of numerical results.

Let us now discuss the influence of β_{δ^2} on the mixing of the sampler. We have found that large values of β_{δ^2} lead to a sampler that has severe mixing issues and often gets trapped in local modes of the target distribution. This issue is illustrated in Figure 2, which shows the mixing of the chain for different values of β_{δ^2} in a case where the true model is $\mathscr{M}_{15}$, the number of samples $N = 64$, and the sampler is initialized in $\mathscr{M}_0$. The mixing issue of the chain when $\beta_{\delta^2} > 100$ is highlighted in this figure, which causes the sampler to get stuck for many iterations at a local mode. In fact, when $\beta_{\delta^2} = 1000$ the sampler cannot escape from the local mode after 100k iterations. This convergence issue might similarly happen when the true signal is near null model and the sampler is initialized near full model. So, for large values of β_{δ^2}, the algorithm is sensitive to the initialized state. On the other hand, too small values of β_{δ^2} which corresponds to assuming low ESNR, would cause the algorithm to explore many regions of low probability of the space in low SNR situations which can be really computationally expensive and causes convergence problems.

A possible solution to the mixing issue would be to use a combination of simulated annealing and MCMC sampler as is done, for example, in [7]. In the next section we follow a different path and use an EB approach to estimate β_{δ^2} from the data.

4. IMPORTANCE SAMPLING BASED MCEM ALGORITHM

Hierarchical models are commonly used in Bayesian model (or variable) selection problems. However, this hierarchy should stop at some point with all remaining parameters assumed fixed. Then, based on some prior beliefs, these parameters can be set. However, for some parameters which no information is provided beforehand, rather than setting them to a fixed value, the EB approach uses the observed data to estimate them. It avoids using arbitrary choices which may be at odds with the observed data.

In this method, one tries to estimate β_{δ^2} such that the marginal likelihood is maximized. In other words,

$$\hat{\beta}_{\delta^2} = \text{argmax}_{\beta_{\delta^2}}\, p(\mathbf{y}|\beta_{\delta^2}).$$

This is similar to MML method proposed in [6] for estimating δ^2. The maximum likelihood may be easier to compute when the data is augmented by a set of latent variables, $\mathbf{u}$ say. These latent variables, in our case, are $\{\boldsymbol{\omega}_k, k, \delta^2, \Lambda\}$. Then, one can use the EM algorithm that entails, at iteration $r+1$, an E-step for computing the expected log-likelihood

$$Q(\beta_{\delta^2}|\hat{\beta}^r_{\delta^2}) = E_{\hat{\beta}^{(r)}_{\delta^2}}\Big\{\ln p(\mathbf{y},\mathbf{u}|\beta_{\delta^2})|\mathbf{y}\Big\} \tag{5}$$

and, an M-step, for maximization of $Q(\beta_{\delta^2}|\hat{\beta}^r_{\delta^2})$ over β_{δ^2} in order to obtain the MLE of it, $\hat{\beta}^{r+1}_{\delta^2}$.

However, in our case, computing the E-step is not possible analytically. Therefore, here, we propose to use Monte Carlo approximation of (5), which is called MCEM [13, 15], by simulating samples from $p(\mathbf{u}|\mathbf{y},\hat{\beta}^r_{\delta^2})$. Moreover, the Monte Carlo estimation of (5) can be implemented in a more efficient way using the idea of Importance Sampling (IS). As is explained in [13, 15], in this framework, samples are just generated from $p(\mathbf{u}|\mathbf{y},\hat{\beta}^0_{\delta^2})$, where $\hat{\beta}^0_{\delta^2}$ is the initial value. Then, for m number of generated samples, the E-step can be written as

$$Q(\beta_{\delta^2}|\hat{\beta}^r_{\delta^2}) = \sum_{t=1}^{m} w_t \ln p(\mathbf{y},u_t|\beta_{\delta^2}) \Big/ \sum_{t=1}^{m} w_t \tag{6}$$

where

$$w_t = \frac{p(u_t|\mathbf{y},\beta^{(r)}_{\delta^2})}{p(u_t|\mathbf{y},\beta^{(0)}_{\delta^2})}$$

are the weights which in our case would simplify to

$$w_t = \left(\frac{\beta^{(r)}_{\delta^2}}{\beta^{(0)}_{\delta^2}}\right)^{\alpha_{\delta^2}} \exp\left(-\frac{\beta^{(r)}_{\delta^2} - \beta^{(0)}_{\delta^2}}{\delta^2_t}\right).$$

Since the RJ-MCMC sampler introduced in Section 2 can easily generate m samples from $p(\mathbf{u}|\mathbf{y},\hat{\beta}^0_{\delta^2})$, these samples can be used to perform the IS based MCEM procedure. So, in each MCEM iteration, a batch of m samples is generated from the RJ-MCMC sampler in order to compute (6). The computationally efficient point of this procedure is that once the IS based MCEM algorithm is stopped, the generated samples are not discarded. They can be used to generate the desired posterior distribution of the unknown parameters by using the importance weights.

However, one should note that this procedure is sensitive to the value of $\hat{\beta}^0_{\delta^2}$. In order to reduce the variations of w_t, it is proposed in [13] to run a few burn-in iterations using a simple MCEM method without importance reweighting.

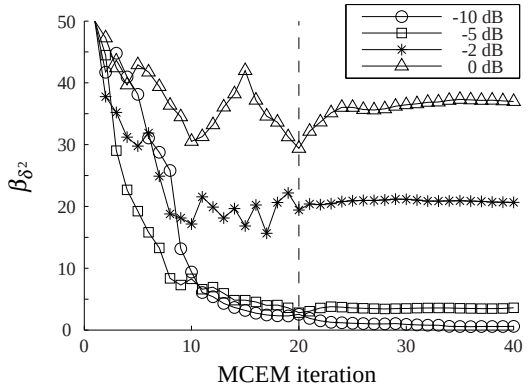

Figure 3: Estimated values of β_{δ^2} using the IS-based MCEM algorithm. The signal is generated under $\mathscr{M}_1$ with $N = 64$, $\omega_{1,1} = 0.2\pi$, for several values of the SNR (see legend). The vertical line indicates the burn-in period.

5. SIMULATION RESULTS AND DISCUSSION

In this section, we will investigate the capability of the IS based MCEM algorithm for assessing β_{δ^2} in different situations. Moreover, we will compare the performance of the sampler with several fixed values of β_{δ^2}. Simulations are performed on two different sample sizes $N = 64$ and $N = 256$ generated according to $\mathscr{M}_1$ with different SNRs. The SNR is defined as

$$\text{SNR} \triangleq \frac{\|\mathbf{D}_k \mathbf{a}_k\|^2}{N\sigma^2}.$$

The parameters of the single sinusoid are as follows: $\omega_{1,1} = 0.2\pi$, $-\arctan(a_{2,1}/a_{1,1}) = \pi/3$, and $a_{1,1}^2 + a_{2,1}^2 = 20$.

In the IS based MCEM algorithm, first, 20 burn-in iterations with $m = 100$ samples were carried out. Then, the 20 IS based MCEM procedure iterations with $m = 5000$ were performed to estimate β_{δ^2}. So, finally, in addition to an approximate estimate of β_{δ^2}, 100k samples from the RJ-MCMC sampler are obtained and can be used to produce the posterior distributions of the unknown parameters, of course by using the importance weights. Figure 3 shows the performance of the IS based MCEM algorithm in estimating the value of β_{δ^2} for different observed signals. This relation between the value of β_{δ^2} and SNR, that is illustrated in figure 3, is remarkably consistent with expectations. It is worthwhile to note that variation of the estimated values of β_{δ^2} is substantially reduced after the burn-in period, as it is shown in figure 3, which illustrates the convergence of the algorithm.

Table 1 presents the probabilities of $\arg\max p(k|\mathbf{y})$ in 100 realizations of the algorithms. In each realization, 100k samples were generated and the first 20k samples were discarded as the burn-in period. The results are presented for different fixed values of β_{δ^2} together with the results obtained by applying the IS based MCEM algorithm for estimating β_{δ^2}.

First, let us consider the case of fixed β_{δ^2}. From the results presented in Table 1, it can be concluded that the value of β_{δ^2} has a strong influence on the posterior distribution of the number of components. Indeed choice of β_{δ^2} would become more critical as the SNR decreases. Though the sampler produces reasonable results for a wide range of values of β_{δ^2}, i.e. $10 \leq \beta_{\delta^2} \leq 1000$, in high SNR situations (not shown here), the behavior of the sampler significantly varies by changing the value of this parameter in low SNR situations. For instance, when SNR $= -5$ dB, while the probability of detecting one component is almost the same for the mentioned interval, setting $\beta_{\delta^2} = 10$ provides a sampler which overestimates the number of components. On the other hand, larger values of β_{δ^2} leads to a sampler that underestimates the number of components. According to the obtained results, choosing a very small value for β_{δ^2}, one say, is not suitable. For the values of SNR < 0 dB, it makes convergence problems for the sampler by accepting most of proposed birth or death moves. More precisely, it leads to a sampler which explores all possible regions, even low probable ones, which would be really computationally expensive when $k_{\max}$ is large. However, one should note that for all simulations the samplers were initialized near null model, otherwise for values of $\beta_{\delta^2} > 100$ the results would definitely changed. In the case that $N = 256$, the sensitivity of the sampler to the choice of β_{δ^2} is less critical. This may be caused by the fact that the observed signal is more informative in this case. Finally, a fixed value of $\beta_{\delta^2} \in [50, 100]$ provides a sampler with more reasonable performance for most values of SNR.

Turning to the results of the EB approach used here to automatically estimate the value of β_{δ^2} from the data, it can be seen from the table that the sampler equipped with the IS-based MCEM algorithm has a quite satisfactory behavior in moderate to high SNR situations (0 dB, -2 dB, and even -5 dB for $N = 256$). However, it is clear that the algorithm fails to select an appropriate value for β_{δ^2} in low SNR situations (-10 dB, and -5 dB for $N = 64$): the selected value is typically much too small, leading to severe overfitting. A similar behavior is observed in experiments under the null model $\mathscr{M}_0$ (not shown here).

In fact, based on Table 1, it seems that using $\beta_{\delta^2} = 50$ gives, in all the situations considered here, results that are similar to or better than the results of the EB approach. Additional experimental results under various configurations and sample sizes are required, however, to issue a general recommendation regarding the choice of an appropriate fixed value for β_{δ^2} (possibly depending on N) and, also, to confirm the capability of the EB approach to automatically select such a value in moderate to high SNR situations.

6. CONCLUSION

In this paper, first, the sensitivity of the RJ-MCMC algorithm proposed in [1] for detection and estimation of sinusoids to the hyperparameter β_{δ^2} has been investigated. Then, an IS-based MCEM algorithm has been used to estimate this parameter given the data, following an empirical Bayes (EB) approach. The IS-based MCEM method has proved able to automatically estimate an appropriate value for β_{δ^2} in moderate to high SNR situations.

The main limitation of the EB approach is that it cannot estimate a proper value for β_{δ^2} in very low SNR situations. This limitation was, however, predictable as in such cases the observed signal carries very little information about the parameter of interest. To overcome this limitation and avoid the problem of choosing a *scale* for $p(\delta^2)$, a truncated Jeffrey prior has been proposed in [19] and very promising results have been obtained.

As mentioned in Section 1, this model and RJ-MCMC sampler have also been used in other applications such as polyphonic signal analysis [3], array signal processing [12], and nuclear emission spectra analysis [10]. The contributions of this paper are likely to be useful in these applications as well.

REFERENCES

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

[2] W. Cui and E. I. George. Empirical Bayes vs. fully Bayes variable selection. *J. Stat. Plann. Inference*, 138(4):888–900, 2008.

[3] M. Davy, S. J. Godsill, and J. Idier. Bayesian analysis of polyphonic western tonal music. *J. Acoust. Soc. Am.*, 119:2498–2517, 2006.

N	β_{δ^2}	P_0	P_1	P_2	P_3	P_4
64	**1**	0.25	0.04	0.04	0.03	0.64
	10	0.64	0.13	0.05	0.02	0.16
	50	0.81	0.09	0.00	0.00	0.10
	100	0.87	0.11	0.00	0.01	0.01
	1000	0.97	0.02	0.01	0.00	0.00
	EB	0.05	0.04	0.02	0.06	0.83
256	**1**	0.01	0.05	0.16	0.18	0.60
	10	0.08	0.45	0.25	0.12	0.10
	50	0.18	0.76	0.04	0.02	0.00
	100	0.22	0.73	0.05	0.00	0.00
	256	0.35	0.63	0.02	0.00	0.00
	1000	0.48	0.51	0.01	0.00	0.00
	EB	0.00	0.22	0.16	0.12	0.50

N	β_{δ^2}	P_0	P_1	P_2	P_3	P_4
64	**1**	0.03	0.09	0.13	0.06	0.69
	10	0.09	0.56	0.12	0.07	0.16
	50	0.27	0.57	0.11	0.00	0.05
	100	0.31	0.60	0.08	0.00	0.01
	1000	0.54	0.45	0.01	0.00	0.00
	EB	0.01	0.22	0.25	0.12	0.42
256	**1**	0.00	0.71	0.22	0.05	0.02
	10	0.00	0.79	0.18	0.01	0.02
	50	0.00	0.92	0.06	0.00	0.02
	100	0.00	0.93	0.07	0.00	0.00
	256	0.00	0.99	0.00	0.01	0.00
	1000	0.00	0.99	0.01	0.00	0.00
	EB	0.00	0.92	0.05	0.02	0.01

N	β_{δ^2}	P_0	P_1	P_2	P_3	P_4
64	**1**	0.00	0.32	0.32	0.14	0.22
	10	0.00	0.68	0.23	0.07	0.02
	50	0.02	0.84	0.10	0.02	0.02
	100	0.01	0.93	0.04	0.01	0.01
	1000	0.02	0.97	0.01	0.00	0.00
	EB	0.00	0.69	0.22	0.04	0.05
256	**1**	0.00	0.89	0.10	0.01	0.00
	10	0.00	0.95	0.05	0.00	0.00
	50	0.00	0.95	0.04	0.00	0.01
	100	0.00	0.95	0.05	0.00	0.00
	256	0.00	1.00	0.00	0.00	0.00
	1000	0.00	1.00	0.00	0.00	0.00
	EB	0.00	0.94	0.04	0.02	0.00

N	β_{δ^2}	P_0	P_1	P_2	P_3	P_4
64	**1**	0.00	0.72	0.17	0.07	0.04
	10	0.00	0.86	0.08	0.05	0.01
	50	0.00	0.87	0.11	0.02	0.00
	100	0.00	0.95	0.05	0.00	0.00
	1000	0.00	0.98	0.02	0.00	0.00
	EB	0.00	0.88	0.09	0.02	0.01
256	**1**	0.00	0.91	0.09	0.00	0.00
	10	0.00	0.95	0.05	0.00	0.00
	50	0.00	0.98	0.02	0.00	0.00
	100	0.00	0.94	0.06	0.00	0.00
	256	0.00	0.98	0.02	0.00	0.00
	1000	0.00	1.00	0.00	0.00	0.00
	EB	0.00	0.98	0.02	0.00	0.00

Table 1: Probability of $\arg\max p(k|\mathbf{y}) = 0$, $\arg\max p(k|\mathbf{y}) = 1$, $\arg\max p(k|\mathbf{y}) = 2$, $\arg\max p(k|\mathbf{y}) = 3$, and $\arg\max p(k|\mathbf{y}) \geq 4$, are denoted, respectively, by P_0, P_1, P_2, P_3, and P_4. The value of the SNR is respectively -10 dB (top-left), -5 dB (top-right), -2 dB (bottom-left) and 0 dB (bottom-right). These probabilities have been estimated based on the output of 100 runs of the algorithm under $\mathscr{M}_1$ with two different sample sizes ($N = 64$ and $N = 256$). The length of the chain was set to 100k, with a burn-in period of 20k samples. Results are presented for several fixed values of β_{δ^2} and for the IS-based MCEM algorithm.

[4] C. Fernández, E. Ley, and M. Steel. Benchmark priors for Bayesian model averaging. *J. Econometrics*, 100(2):381–427, 2001.

[5] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis (second edition)*. Chapman & Hall / CRC, 2004.

[6] E. I. George and D. P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

[7] R. Gramacy, R. Samworth, and R. King. Importance tempering. *Stat. Comput.*, 20:1–7, 2010.

[8] P. J. Green. Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[9] P. J. Green. Trans-dimensional Markov chain Monte Carlo. In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 179–198. O.U.P., 2003.

[10] S. Gulam Razul, W. Fitzgerald, and C. Andrieu. Bayesian model selection and parameter estimation of nuclear emission spectra using RJMCMC. *Nucl. Instrum. Meth. A*, 497(2-3):492–510, 2003.

[11] J. Jannink and R. Fernando. On the Metropolis-Hastings acceptance probability to add or drop a quantitative trait locus in Markov chain Monte Carlo-based Bayesian analyses. *Genetics*, 166(1):641–643, 2004.

[12] J. R. Larocque and J. P. Reilly. Reversible jump MCMC for joint detection and estimation of sources in coloured noise. *IEEE Trans. Signal Process.*, 50:231–240, 2000.

[13] R. A. Levine and G. Casella. Implementations of the Monte Carlo EM algorithm. *J. Comput. Graph. Stat.*, pages 422–439, 2001.

[14] F. Liang, R. Paulo, G. Molina, M. Clyde, and J. Berger. Mixtures of g-priors for Bayesian variable selection. *J. Am. Stat. Assoc.*, 103(481):410–423, 2008.

[15] F. Quintana, J. Liu, and G. del Pino. Monte Carlo EM with importance reweighting and its applications in random effects models. *Comput. Stat. Data An.*, 29(4):429–444, 1999.

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

[17] C. Robert and G. Casella. *Monte Carlo Statistical Methods (second edition)*. Springer Verlag, 2004.

[18] C. P. Robert. *The Bayesian Choice (second edition)*. Springer, 2007.

[19] A. Roodaki, J. Bect, and G. Fleury. On the Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC in Low SNR Situations. In *Proc. 10th Int. Conf. on Information Science, Signal Processing and their Application (ISSPA), Kuala lumpur, Malaysia*, pages 5–8, 2010.

[20] G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Stat. Assoc.*, 85(411):699–704, 1990.

[21] A. Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. North-Holland/Elsevier, 1986.