

ON THE JOINT BAYESIAN MODEL SELECTION AND ESTIMATION OF SINUSOIDS VIA REVERSIBLE JUMP MCMC IN LOW SNR SITUATIONS

Alireza Roodaki, Julien Bect and Gilles Fleury

Department of Signal Processing and Electronic Systems,
SUPELEC, Gif-sur-Yvette, France.

ABSTRACT

This paper addresses the behavior in low SNR situations of the algorithm proposed by Andrieu and Doucet (IEEE T. Signal Proces., 47(10), 1999) for the joint Bayesian model selection and estimation of sinusoids in Gaussian white noise. It is shown that the value of a certain hyperparameter, claimed to be weakly influential in the original paper, becomes in fact quite important in this context. This robustness issue is fixed by a suitable modification of the prior distribution, based on model selection considerations. Numerical experiments show that the resulting algorithm is more robust to the value of its hyperparameters.

Index Terms— Bayesian model selection; reversible jump MCMC; prior calibration; Bayesian sensitivity analysis; spectral analysis.

1. INTRODUCTION

Detection and separation of signals in low SNR conditions has many applications in various fields such as communication, radar and sonar—to name but a few. Moreover, sinusoids are one of the most common kind of signals used in these applications. The problem of joint detection and estimation of sinusoids in low SNR situations, assuming unknown number of components, is therefore of general importance.

A fully Bayesian algorithm based on Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) technique [1] for handling this problem, not specifically in low SNR situations, has been proposed in [2]. This algorithm, of course with appropriate modifications, has been used for other applications such as polyphonic signal analysis [3], array signal processing [4], and nuclear emission spectra analysis [5]. However, to the best of our knowledge, the behavior of this algorithm in low SNR situations has never been studied. To present the problem more explicitly, in the following we will introduce the notations used in the algorithm.

Let $\mathbf{y} = (y_1, y_2, \dots, y_N)^t$ be a vector of N independent observations. Based on the model \mathcal{M}_k (for $k = 0, 1, \dots, k_{\max}$), \mathbf{y} can be represented by summation of k sinusoids together with a white Gaussian noise. Defining the $N \times 2k$ matrix containing the sinusoids with different radial frequencies, \mathbf{D}_k , as below

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

for $i = 0, \dots, N-1$ and $j = 1, \dots, k$, one can write the normal linear regression model for the current problem

with k components:

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

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

As in many Bayesian model selection approaches for normal linear regression problem, the well-known conditionally conjugate g -prior [6, 7, 8], which provides tractable computations, has been assigned as a prior over the amplitudes in the model proposed in [2]. The g -prior is a zero mean multivariate normal distribution with $\sigma^2/g(\mathbf{D}_k^t \mathbf{D}_k)^{-1}$ as its covariance matrix. The variable called g controls the expected size of the amplitudes. This parameter has been substituted by δ^{-2} in [2] and δ^2 has been called the Expected SNR (ESNR).

Owing to the influence of the ESNR on the performance of the algorithm, particularly in the Bayesian model selection part, several approaches for setting or estimating it have been proposed in the variable selection literature; see [7, 8, 9] and references therein. To keep the Fully Bayesian spirit, a vague conjugate Inverse-Gamma (\mathcal{IG}) prior has been assigned over ESNR in [2], i.e. $p(\delta^2 | \alpha_{\delta^2}, \beta_{\delta^2}) = \mathcal{IG}(\cdot | \alpha_{\delta^2}, \beta_{\delta^2})$. Although it was mentioned that the performance of the proposed algorithm is not sensitive to the value of the scale parameter β_{δ^2} , our experiments have shown that this parameter becomes influential when dealing with low SNR signals.

The structure of this article is as follows. Section 2 briefly recalls the Bayesian algorithm proposed in [2]. Section 3 discusses first the “dimensionality penalty” induced by the hyperparameter δ^2 and then the effect of β_{δ^2} on the posterior distribution of k and δ^2 . Section 4 discusses solutions to the problem of choosing β_{δ^2} : since the usual data-driven approaches fail in low SNR situations, we propose to use a truncated Jeffrey prior instead. Section 5 presents numerical results that support the proposed method and discusses its sensitivity to the lower bound δ_{\min}^2 of the truncated prior. Finally, Section 6 concludes the article and addresses possible future works.

2. BAYESIAN FRAMEWORK

The full joint distribution of the observed signal and the unknown parameters, in the model proposed by [2], has the following hierarchical structure:

$$p(\mathbf{y}, k, \boldsymbol{\theta}_k, \delta^2) = p(\mathbf{y} | k, \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k | k, \delta^2) \times p(k) p(\delta^2). \quad (1)$$

2.1. Prior distributions

As proposed by [2], the prior over k is a Poisson distribution with mean Λ , truncated to $\{0, 1, \dots, k_{\max}\}$. Conditional on k , the ω_k 's are independent and identically distributed, with a uniform distribution on $(0, \pi)$. The noise variance σ^2 is endowed with Jeffrey's uninformative prior, i.e. $p(\sigma^2) \propto 1/\sigma^2$, where the symbol \propto denotes proportionality.

Furthermore, they have suggested to assign a conjugate $\mathcal{IG}(\alpha_{\delta^2}, \beta_{\delta^2})$ prior over ESNR and to set α_{δ^2} to two for having an infinite variance. However, as it can be seen in Figure 1, the posterior over δ^2 is severely sensitive to the value of β_{δ^2} .

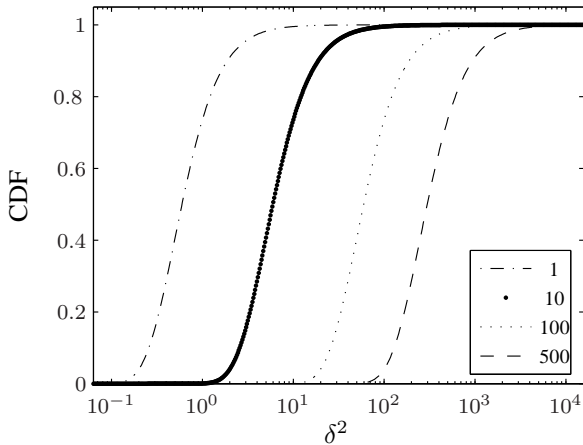


Figure 1: CDFs of priors over δ^2 for different values of β_{δ^2} .

The hyperparameter Λ has been assigned in [2] a Gamma prior, i.e. $p(\Lambda) = \mathcal{G}(\alpha_\Lambda, \beta_\Lambda)$, with $\alpha_\Lambda \approx \frac{1}{2}$ as a shape parameter and $\beta_\Lambda \approx 0$ as a scale parameter. This is equivalent to using a negative binomial prior over k that puts more emphasis on small values. In this paper, in order to have an almost flat prior over k , the parameter α_Λ is set to a value close to 1.

2.2. Sampling structure

Based on (1) and Bayes Theorem, after simply integrating \mathbf{a}_k and σ^2 out, the joint posterior distribution of k and ω_k , up to a normalizing constant, can be written as

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

with

$$\mathbf{P}_k = \mathbf{I}_N - \frac{\delta^2}{1 + \delta^2} \mathbf{D}_k (\mathbf{D}_k^t \mathbf{D}_k)^{-1} \mathbf{D}_k^t. \quad (3)$$

In the following, different steps for sampling from the above distribution are briefly described. For more detailed expressions, please refer to [1, 2].

The sampler consists of a Metropolis-Hastings (MH) move for the target density (2), which updates the values of k and ω_k , followed by a sequence of Gibbs moves to update δ^2 and Λ . The proposal kernel, in the MH step,

is a mixture of within-model moves, which update the radial frequencies without changing k , and between-models moves, which change the value of k by adding or removing a component (so-called birth/death move). The Gibbs move for δ^2 if performed by demarginalization of σ^2 and \mathbf{a}_k and then sampling from the ‘‘uncollapsed’’ posterior of δ^2 .

Except for a modification in the birth/death ratio, the moves implemented in our sampler are the same as in [2]. In the birth move, after proposing a new component by sampling its radial frequency from $U(0, \pi)$, it is randomly located among the previous components. Then, the move is accepted with probability $\alpha_{birth} = \min\{1, r_{birth}\}$, where

$$r_{birth} = \left(\frac{\mathbf{y}^t \mathbf{P}_{k+1} \mathbf{y}}{\mathbf{y}^t \mathbf{P}_k \mathbf{y}} \right)^{-N/2} \frac{1}{1 + \delta^2}. \quad (4)$$

One should note that the birth ratio (4) differs from the one reported in [2] by a multiplicative factor of $1/(k+1)$. A similar mistake for a similar algorithm has been found in the field of genetics [10]. Note that using the ratio given in [2] amounts to changing the prior distribution on k . This issue will be dealt with in greater detail in a forthcoming paper. In the meantime, the reader is referred to [11] for more information on the role of permutations and sorting in the computation of RJ-MCMC ratios.

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

In this section, the effect of β_{δ^2} on the performance of the algorithm in low SNR situations is discussed.

To better understand the importance of β_{δ^2} , the role of δ^2 will be discussed first, following the ideas introduced in [9, 12] to make a connection between Bayesian algorithms and model selection criteria. Let us assume, for the sake of simplicity, a flat prior over the number of components. Then, the log-posterior can be written as

$$\log p(k, \omega_k | \mathbf{y}, \delta^2) = -\frac{N}{2} \log(\mathbf{y}^t \mathbf{P}_k \mathbf{y}) - F \cdot k + C, \quad (5)$$

where $F = \log(\pi(1 + \delta^2))$ and C is a constant which does not depend on k and ω_k . F can be interpreted as a dimensionality penalty, which penalizes complex models. Therefore, large values of δ^2 , which result in large values of F , cause the algorithm to neglect small components with respect to the noise. Conversely, ‘‘small’’ values of δ^2 result in an algorithm which does not penalize enough ‘‘small’’ components and leads to overfitting.

In addition to—and partly because of—its role in the model selection properties of the algorithm, the value of δ^2 has a strong influence on the behavior of the resulting algorithm. For low values of δ^2 , the Markov chain has to visit much more often regions of the state space corresponding to high values of k , where the algorithmic complexity of running the chain is much higher. For high values of δ^2 , the posterior distribution has sharper peaks and valleys, which makes it much more difficult for the chain to explore, resulting in a slower convergence rate.

Turning to the role of β_{δ^2} , first, one should note that the \mathcal{IG} prior used in [2], although chosen to be weakly informative, is not really ‘‘vague’’ (see Figure 1). In fact, it

has a mode at $\beta_{\delta^2}/(\alpha_{\delta^2} + 1)$. By changing its scale parameter the behavior of the algorithm can be controlled just like changing the values of δ^2 itself, esp. in the low SNR situations where likelihood does not provide much information about δ^2 . Figure 2 displays the sensitivity of the posteriors of k and δ^2 to the hyperparameter β_{δ^2} in an experiment of signal detection under \mathcal{M}_1 with SNR = -1 dB, which is not very low. In this study, SNR is defined as $\|\mathbf{D}_k \mathbf{a}_k\|^2 / (N\sigma^2)$. It can be seen in this figure that the posterior of δ^2 is moving to the right by increasing the value of β_{δ^2} . Moreover, if one is interested in model selection based on the maximum of the posterior of the number of components, i.e. $\arg \max_{k \in \{0, \dots, k_{\max}\}} p(k | \mathbf{y})$, the selected models under $\beta_{\delta^2} = 1$, $\beta_{\delta^2} = 10$, and $\beta_{\delta^2} = 100$ would be \mathcal{M}_2 , \mathcal{M}_2 , and \mathcal{M}_1 , respectively. The differences in the results for Bayesian model averaging (not shown in this paper) are even more important.

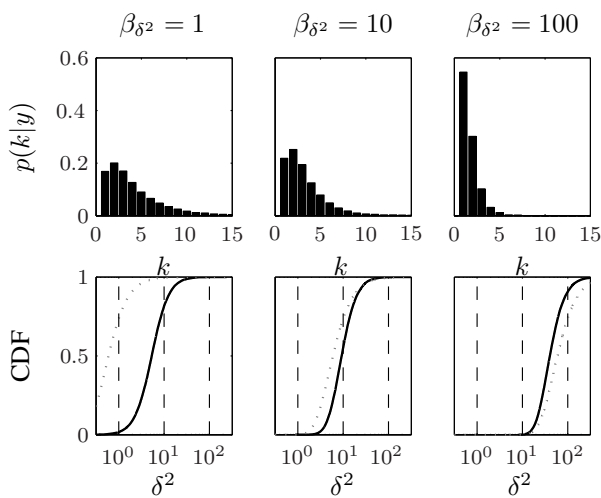


Figure 2: The posteriors of k and δ^2 under the experiment of signal detection with SNR = -1 dB and different values of β_{δ^2} . In the second row, the gray dotted lines show the prior and the black lines show the posterior of δ^2 . The length of the chain was set to 100k, with a burn-in period of 20k samples.

4. PROPOSED METHODS

In the following possible methods for either estimating a reasonable value for β_{δ^2} from the observed data or stabilizing the algorithm by modifying the prior are introduced.

4.1. Data-driven methods

In order to estimate a proper value for β_{δ^2} the first two approaches that may come to mind are the Fully Bayesian and the Empirical Bayes (EB) methods. The former one is constructed by assigning a vague conjugate Gamma prior over β_{δ^2} , that is, $\beta_{\delta^2} \sim \mathcal{G}(a, b)$. Then, one can update it by performing a Gibbs move with $\mathcal{G}(a + \alpha_{\delta^2}, b + \delta^{-2})$ as proposal distribution. On the other hand, the EB method is a data-driven approach in which the marginal likelihood of the parameter given the data, i.e. $p(\mathbf{y} | \beta_{\delta^2})$, is maximized. This idea has been used in [7, 9, 12] for estimating δ^2 . However, since in this problem, $p(\mathbf{y} | \beta_{\delta^2})$

does not exist in closed form, one should use Monte Carlo methods to estimate β_{δ^2} as in [13].

4.2. Using a truncated Jeffrey prior over δ^2

The idea of using an improper Jeffrey prior over ESNR, which provides a flat prior over the log (δ^2) in contrary to the current prior, has been mentioned in [2] but it is not used as $\delta^2 = 0$ would become an absorbing state of the Markov chain. Here, we propose to truncate the Jeffrey prior using a lower bound δ_{\min}^2 and an upper bound δ_{\max}^2 . The sensitivity of the algorithm to δ_{\max}^2 can be reduced by setting it to a large value, say 10000. However, choosing the value of the lower bound is less trivial, since it controls the minimal dimensionality penalty induced by the prior; a numerical sensitivity analysis will be carried out in the next section.

5. SIMULATION RESULTS AND DISCUSSION

In this section, we study the performance of the proposed solutions for reducing the sensitivity of the Bayesian algorithm to the prior over δ^2 . Simulations are carried out with the observed signal of length $N = 64$. In this paper, the problem of signal detection in low SNR situation is considered. The parameters of the single sinusoid are as follows: $\omega_{1,1} = 0.2\pi$, $-\arctan(a_{s1}/a_{c1}) = \pi/3$, and $a_{s1}^2 + a_{c1}^2 = 20$. The length of chain in all simulations was 100k, with a burn-in period of 20k samples.

The data-driven approaches estimate a reasonable value for the hyperparameter β_{δ^2} in high SNR situations but do not perform satisfactorily in low SNR situations. In fact, in these situations, our numerical experiments showed that β_{δ^2} is estimated to be very close to 0, which imposes too small δ^2 , using both methods. It has also been reported in [7] that the EB method tends to estimate δ^2 as 0 under the null model in a similar framework.

On the other hand, in the case of using a truncated Jeffrey prior over δ^2 , the value of δ_{\min}^2 determines the minimal dimensionality penalty. One should note that, a reasonable range of values for the lower bound is restricted, since having a high minimal penalty is not suitable. Moreover, setting δ_{\min}^2 to a large value might cause convergence issues. Thus, up to now, we have translated the problem of estimating a proper value for the hyperparameter β_{δ^2} to the problem of finding a reasonable value for δ_{\min}^2 . In the sequel, the sensitivity of the algorithm to the variations of this parameter is studied.

Figure 3 shows the posterior distributions for k and δ^2 for the same observed signal as Figure 2. As depicted in this figure, the sensitivity of the algorithm to the variations of δ_{\min}^2 is much less than that of β_{δ^2} . In fact no matter what the value of δ_{\min}^2 is, the model \mathcal{M}_1 would be selected based on the MAP of k . For further studying the sensitivity of the algorithm to the parameter δ_{\min}^2 , the probabilities of selected models based on $\arg \max_k p(k | \mathbf{y})$ in 100 realizations of the sampler for different values of SNR were estimated. Figure 4 shows the sensitivity of the algorithm to this parameter for the cases of SNR = -3 dB and SNR = -4 dB. In this figure, the algorithm was run with $\delta_{\min}^2 = 0.5$. The probabilities for other values of δ_{\min}^2 were

obtained using importance sampling. This method has already been used for the sensitivity analysis of Bayesian algorithms to their priors; see for instance [14]. It can be concluded from figure 4 that the probabilities are not very sensitive to the choice of δ_{\min}^2 . However, as the value of the lower bound increases, P_2 decreases while P_0 increases: this was predictable, as δ_{\min}^2 controls the minimal dimensionality penalty.

6. CONCLUSION

The main contribution of this paper has been to explain the lack of robustness, in low SNR situations, of the algorithm proposed in [2] and to propose solutions for fixing it. Simulation results showed that a truncated Jeffrey prior over δ^2 significantly improves the performance of the sampler in situations where the usual data-driven approaches (Empirical Bayes and Fully Bayes) fail. Sensitivity analyses, which are efficiently carried out using importance sampling, reveal that the resulting algorithm is rather robust to variations of the lower bound δ_{\min}^2 in a reasonable range. A natural direction for future work would be to propose a data-driven approach for the automatic selection of this threshold and to assess more systematically the performances of this algorithm.

References

- [1] P. J. Green, "Reversible jump MCMC computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.
- [2] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC," *IEEE T. Signal Proces.*, vol. 47, no. 10, pp. 2667–2676, 1999.
- [3] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acoust. Soc. Am.*, vol. 119, pp. 2498–2517, 2006.
- [4] J. R. Larocque and J. P. Reilly, "Reversible jump MCMC for joint detection and estimation of sources in coloured noise," *IEEE T. Signal Proces.*, vol. 50, pp. 231–240, 2000.
- [5] S. Gulam Razul, W. Fitzgerald, and C. Andrieu, "Bayesian model selection and parameter estimation of nuclear emission spectra using RJMCMC," *Nucl. Instrum. Meth. A*, vol. 497, no. 2-3, pp. 492–510, 2003.
- [6] A. Zellner, "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, (eds. P. K. Goel and A. Zellner), pp. 233–243, 1986.
- [7] F. Liang, R. Paulo, G. Molina, M. Clyde, and J. Berger, "Mixtures of g-priors for Bayesian variable selection," *J. Am. Stat. Assoc.*, vol. 103, no. 481, pp. 410–423, 2008.
- [8] C. Fernández, E. Ley, and M. Steel, "Benchmark priors for Bayesian model averaging," *J. Econometrics*, vol. 100, no. 2, pp. 381–427, 2001.
- [9] E. I. George and D. P. Foster, "Calibration and empirical Bayes variable selection," *Biometrika*, vol. 87, no. 4, pp. 731–747, 2000.
- [10] 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*, vol. 166, no. 1, pp. 641–643, 2004.
- [11] S. Richardson and P. J. Green, "On Bayesian analysis of mixtures with an unknown number of components," *J. Roy. Stat. Soc. B Met.*, vol. 59, no. 4, pp. 731–792, 1997.
- [12] W. Cui and E. I. George, "Empirical Bayes vs. fully Bayes variable selection," *J. Stat. Plan. Inference*, vol. 138, no. 4, pp. 888–900, 2008.
- [13] R. A. Levine and G. Casella, "Implementations of the Monte Carlo EM algorithm," *J. Comput. Graph. Stat.*, pp. 422–439, 2001.
- [14] J. Besag, P. Green, D. Higdon, and K. Mengersen, "Bayesian computation and stochastic systems (with discussion)," *Statistical Science*, vol. 10, no. 1, pp. 3–41, 1995.

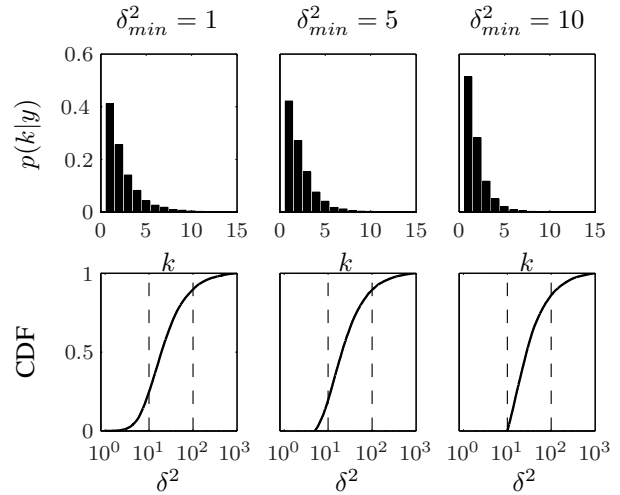


Figure 3: The posteriors of k and δ^2 under the experiment of signal detection with SNR = -1 dB and different values of δ_{\min}^2 .

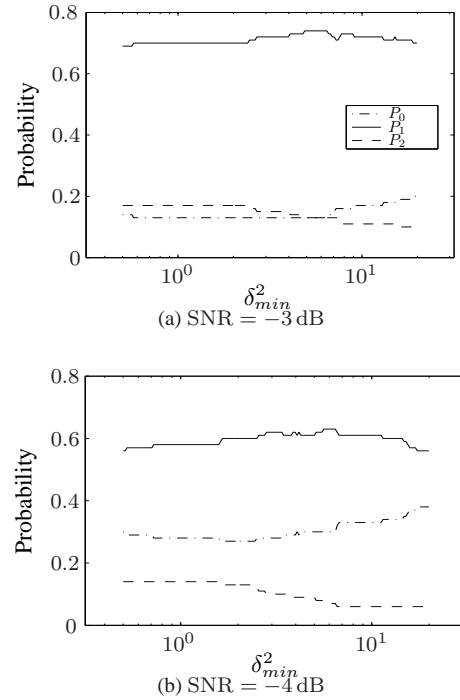


Figure 4: Probabilities of $\arg \max p(k | \mathbf{y}) = 0$, $\arg \max p(k | \mathbf{y}) = 1$, and $\arg \max p(k | \mathbf{y}) \geq 2$ are denoted, respectively, by P_0 , P_1 , and P_2 in 100 realization of the algorithm using $\delta_{\min}^2 = 0.5$. The probabilities for other values of δ_{\min}^2 , i.e. $\delta_{\min}^2 \in (0.5, 20]$, are estimated using the importance sampling method.