Bayesian noise model selection and system identification using Chib's approximation based on the Metropolis-Hastings sampler

Andrei Bărbos, Audrey Giremus, Jean-François Giovannelli

UMR 5218 - IMS - Laboratoire de l'Intégration du Matériau au Système

351 Cours de la Libération, 33405 Talence cedex, France

{andrei-cristian.barbos, audrey.giremus, jean-francois.giovannelli}@ims-bordeaux.fr

Résumé − Dans cette comunication, le problème de la sélection du modèle de bruit qui corrompt des données mesurées a été étudié. Les données étant obtenues en sortie d'un système d'ordre deux, ce problème est couplé a un problème d'identification. Dans le cadre bayésien, une solution jointe a été proposée. La sélection de modèle repose sur l'évidence dont le calcul est délicat. L'approche utilisée pour calculer l'évidence est fondée sur les méthodes MCMC mais ne repose pas sur l'estimateur de la moyenne harmonique qui est connu pour présenter des problèmes de convergence. Elle met à profit la condition de réversibilité de l'algorithme de Metropolis-Hastings. Les performances de la solution ont été évaluées avec des résultats satisfaisants.

Abstract − In this paper the problem of model selection has been applied to the identification of the model for the noise that affects the measured data. Performing the identification of a second order system has been considered as a side problem, in the view that the output of such a system is what we are measuring. A joint solution for the two distinct problems has been proposed in the context of the Bayesian statistical modelling. The main problem was with the approximation of the model evidence the solving of which required the use of numerical methods. However, our approach is not based on the well-known estimator of the harmonic mean which can exhibit bad convergence properties. As an alternative, the proposed estimation method takes advantage of the reversibility property of the Metropolis sub-kernel. The performances of the proposed solution have been assessed with encouraging results.

1 Introduction

The main problem addressed in this paper is performing model selection in the Bayesian context. The objective is to determine the appropriate noise model for noise corrupted observed data. This problem is coupled with a system identification one, more specifically with the identification of a second order system.

The problem of model selection has a long history, dating as back as Jeffreys and continuing into present time with the works of [1–4,7–9] and the references therein, just to name a few. Despite being under scrutiny for such a long time, there still isn't a definite solution to the problem. The issue that most often arises is the presence of intractable integrals in the posterior distribution of the models. The workaround usually involves numerical sampling methods, most often in the class of Markov Chain Monte Carlo (MCMC), in order to be able to sample from the posterior. Regarding the numerical sampling methods, there are two distinct approaches: within-model and across-model sampling. The within-model approach consists in running a Markov chain for each of the considered models whereas the across-model approach runs just one Markov chain for all the models, which is able to make jumps also across the model space and not only within the parameter space. Unfortunately, it is not an easy task to design good jumping rules so as to allow a thorough exploration of both the model and parameter spaces. As such, a within-model approach was considered in our case.

The joint problem of noise model selection and system identification has been recently tackled in [6], where a different approach based on the Laplace-Metropolis approximation was used to approximate the model evidence.

Section 2 deals with the Bayesian modelling of the problem under study, *i.e.* specifying the likelihood, prior distribution and obtaining the posterior. Section 3 introduces the evidence approximation method, section 4 presents the numerical results, while section 5 concludes the paper.

This study has been carried out with financial support from the French State, managed by the French National Research Agency (ANR) in the frame of the "Investments for the future" Programme IdEx Bordeaux - CPU (ANR-10-IDEX-03-02)

2 Bayesian noise model selection

The Bayesian statistical modelling process, as indicated in [10], comprises three steps: model formulation, model estimation and model selection. The first step involves specifying the likelihood and the prior distributions, the second step involves applying Bayes rule to determine the posterior law while the third step is all about determining the most appropriate model given the available data.

2.1 System and noise models

As the emphasis is on the noise model selection part, we have chosen not to complicate the problem at hand by selecting an intricate model for the system, thus we have chosen to perform the identification of a second order system. It is performed by analysing the step response of the system, considering that the unit step signal is applied at time $t = 0$. The model for the step response of the system is defined in the time domain and it is given by:

$$r(t) = G_0 \left(1 - \cos(2\pi f_0 \Delta t) \exp\left(-\frac{\Delta t}{\tau_0}\right)\right) u(\Delta t), \quad (1)$$

where G_0 is the gain, $\Delta t = t - t_0$ with $t_0 > 0$ represents the propagation delay, f_0 is the ripple frequency and τ_0 is a time constant controlling the decay of the ripples. The vector $\boldsymbol{\theta}^s = [G_0, t_0, \tau_0, f_0]$ groups together all the parameters.

What is actually measured at the output of the system is a sampled and noise corrupted version of the step response, where we have assumed an additive model for the noise. As such, the model for the measurements is:

$$x_n = r_n + w_n, \quad (2)$$

where $r_n = r(nT_s)$ with T_s the sampling period and w_n is the noise component. All N measurements have been collected in the vector $\boldsymbol{x} = [x_1, \ldots, x_N]^T$.

The interest towards the noise model selection problem stems from the fact that the parameter estimation procedure is tightly connected to the chosen noise model. As such under the wrong hypothesis, the estimated values tend to be off with respect to the true ones. Table 1 presents the expressions for each of the three chosen models. All three models have a simple parametrisation with only one parameter, γ_k, and they are of zero mean[1]. Each model is assigned a parameter vector $\boldsymbol{\theta}_k = [\boldsymbol{\theta}^s, \gamma_k]$, $k = \{1, 2, 3\}$, which we seek to estimate.

The general expression for the likelihood function is obtained by making the assumption of i.i.d. measurements and is given by equation (3). Replacing the f_k term with either one of the expressions found in table 1 yields the likelihood function for the respective model.

$$f(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M} = k) = \prod_{n=1}^{N} f_k\left[(x_n - r_n(\boldsymbol{\theta}^s))|\gamma_k\right] \quad (3)$$

[1] location parameter for the Cauchy distribution

TAB. 1: Noise distributions

Gauss	$f_1(\omega|\gamma_1) = (2\pi)^{-1/2}\ \gamma_1^{1/2} \exp\left(-\gamma_1\omega^2/2\right)$
Laplace	$f_2(\omega|\gamma_2) = 2^{-1}\ \gamma_2^{1/2} \exp\left(-\gamma_2^{1/2}|\omega|\right)$
Cauchy	$f_3(\omega|\gamma_3) = \pi^{-1}\gamma_3^{1/2}\left[1 + \gamma_3\omega^2\right]^{-1}$

2.2 Prior distributions

In the Bayesian modelling, the prior distribution plays an important role as it is the tool to inject available initial information into the inference problem. In our case however, we have chosen to use *uninformative* uniform prior distributions for the parameters so as to provide a solution as general as possible,

$$\pi(\theta_{k,l}|\mathcal{M} = k) = \mathcal{U}(\theta_{k,l}) \quad (4)$$

where $\theta_{k,l}$, $l = 1, \ldots, L\,(= 5)$ is the l-th component of the parameter vector $\boldsymbol{\theta}_k$. Moreover, for a given parameter $\theta_{k,l}$, we have used the same prior for all models.

Under the assumption of the parameters being independent among each other, the prior law factorises as:

$$\pi(\boldsymbol{\theta}_k|\mathcal{M} = k) = \prod_{l=1}^{L} \pi(\theta_{k,l}|\mathcal{M} = k). \quad (5)$$

2.3 Bayesian model selection

Two key components in performing model selection are the posterior distribution and the *model evidence*. For a given model k, $k \in \{1, \ldots, K\}$, the posterior writes:

$$\mathcal{P}(\mathcal{M} = k|\boldsymbol{x}) = \frac{\mathcal{P}(\mathcal{M} = k) f(\boldsymbol{x}|\mathcal{M} = k)}{\sum_{j=1}^{K} \mathcal{P}(\mathcal{M} = j) f(\boldsymbol{x}|\mathcal{M} = j)}, \quad (6)$$

where the term

$$f(\boldsymbol{x}|\mathcal{M} = k) = \int f(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M} = k)\, \pi(\boldsymbol{\theta}_k|\mathcal{M} = k)\ \mathrm{d}\boldsymbol{\theta}_k \quad (7)$$

represents the aforementioned model evidence and the term $\mathcal{P}(\mathcal{M} = k)$ represents the model prior probability. In order not to introduce prior bias towards a given model, we have considered the models to be *a priori* equiprobable.

The noise model selection procedure consists in selecting the model with the highest posterior probability, which is an optimal choice as it minimises the zero/one risk [6].

$$\widehat{\mathcal{M}} = \arg\max_k \mathcal{P}(\mathcal{M} = k|\boldsymbol{x}) \quad (8)$$

We have not managed to elude the common problem of having to deal with intractable integrals as in our case we cannot evaluate the integral found in equation (7). The following sections discusses the solution employed in order to overcome this problem.

3 Chib evidence approximation

The key component of the evidence approximation method proposed by Chib in [3] is being able to obtain samples from the posterior distribution. Moreover, the method itself is defined around the Metropolis-Hastings (MH) sampler. The first step in approximating the evidence is to recast the expression of the posterior distribution for the parameters by swapping the posterior with the evidence:

$$f\left(\boldsymbol{x}|\mathcal{M}=k\right)=\frac{f\left(\boldsymbol{x}|\boldsymbol{\theta}_k,\mathcal{M}=k\right)\pi\left(\boldsymbol{\theta}_k|\mathcal{M}=k\right)}{\pi\left(\boldsymbol{\theta}_k|\boldsymbol{x},\mathcal{M}=k\right)}. \quad (9)$$

As the posterior is not available in a closed form, the method then resorts to numerically approximating it.

The above equation remains valid no matter the value of $\boldsymbol{\theta}$, but as we will be performing a numerical approximation of the posterior, one wants the point in which the approximation is performed to lie in a region of high posterior density. This is so because more samples are available from such a particular region, thus ensuring a better accuracy of the approximation [2]. In our case we have used the posterior mean as the respective point.

To ease the numerical approximation of the posterior distribution, we had to decompose it as a product of $L=5$ conditional posterior laws[2], one law per parameter

$$\pi\left(\boldsymbol{\theta}_k^*|\boldsymbol{x},\mathcal{M}=k\right)=\prod_{l=1}^{L}\pi\left(\theta_{k,l}^*|\boldsymbol{x},\boldsymbol{\psi}_{k,l-1}^*\right) \quad (10)$$

where $\boldsymbol{\theta}_k^*=\left[\theta_{k,1}^*,\ldots,\theta_{k,L}^*\right]$ denotes the point at which we approximate the posterior and where for simplicity the notation $\boldsymbol{\psi}_{k,l-1}^*=[\theta_{k,1}^*,\ldots,\theta_{k,l-1}^*]$ was introduced.

Each conditional $\pi\left(\theta_l^*|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*\right)$ from equation (10) is approximated as the ratio of two expectations:

$$\pi\left(\theta_l^*|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*\right)=\frac{E_n\left\{\alpha\left(\theta_l,\theta_l^*|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*,\boldsymbol{\psi}_{l+1}\right)J_t\left(\theta_l,\theta_l^*|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*,\boldsymbol{\psi}_{l+1}\right)\right\}}{E_d\left\{\alpha\left(\theta_l^*,\theta_l|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*,\boldsymbol{\psi}_{l+1}\right)\right\}}, \quad (11)$$

where for simplicity the index k has been eliminated and the notation $\boldsymbol{\psi}_{l+1}=[\theta_{l+1},\ldots,\theta_L]$ has been introduced. The above formula was obtained starting from the reversibility property of the Metropolis-Hastings algorithm sub-kernel [3]. The numerator expectation is with respect to the distribution $\pi\left(\theta_l,\boldsymbol{\psi}_{l+1}|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*\right)$ whereas the denominator expectation is with respect to the distribution $\pi\left(\boldsymbol{\psi}_{l+1}|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*,\theta_l^*\right)\ J_t\left(\theta_l^*,\theta_l|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*,\boldsymbol{\psi}_{l+1}\right)$. The terms that intervene inside the expressions of the expectations, *i.e.* $\alpha\left(\circ,\circ|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*,\boldsymbol{\psi}_{l+1}\right)$ and $J_t\left(\theta_l,\theta_l^*|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*,\boldsymbol{\psi}_{l+1}\right)$, represent the acceptance ratio and the proposition law of the Metropolis-Hastings algorithm [5].

Both expectations in equation (11) are approximated using the Monte-Carlo integration technique. In the case of the numerator expectation, samples distributed according to the conditional posterior law $\pi\left(\theta_l,\boldsymbol{\psi}_{l+1}|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*\right)$ are required. To obtain these required samples we resorted to the use of the Metropolis-within-Gibbs (MwG) sampling approach, where the aforementioned conditional posterior is further decomposed into a series of conditional posteriors $\pi\left(\theta_j|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*,\theta_{-j}\right)$, where $j\in\{l,\ldots,L\}$ and $\theta_{-j}=[\theta_l,\ldots,\theta_{j-1},\theta_{j+1},\ldots,\theta_l]$. Under this approach, obtaining one sample from the above posterior requires sampling all the conditionals from the series whereas obtaining M samples requires repeating the process M times.

Approximating the denominator expectation requires samples distributed according to the product of distributions $\pi\left(\boldsymbol{\psi}_{l+1}|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*,\theta_l^*\right)\ J_t\left(\theta_l^*,\theta_l|\boldsymbol{x},\boldsymbol{\psi}_{l-1}^*,\boldsymbol{\psi}_{l+1}\right)$. By taking a closer look at its expression, one can notice that the conditional posterior distribution involved in it is nothing more than the distribution that must be sampled in order to approximate the numerator expectation for the subsequent term in the decomposition from (11). This recycling of samples is a noteworthy aspect of the model evidence approximation method as for estimating one term from the decomposition in equation (10) one must run the MwG sampler only once as opposed to twice if this wasn't the case.

4 Simulation Results

One aspect worth mentioning with respect to the decomposition in equation (10) is the order in which it is performed. Initial experimental results indicated that the order in which the decomposition is carried out requires fine tuning for achieving better results. However, subsequent investigations indicated that this might not be case, or at least, not in the cases that were considered. For the presented results the order in which the decomposition was carried out was hand chosen and is the order in which the parameters are written in table 3.

In order to asses the performance of the presented model selection method, for each of the three considered noise types the method was run 100 times. At each run a number of 3000 posterior samples were drawn from which the first 900 samples were discarded as burn-in samples. The results for the noise model selection part are presented in table 2. As it can be observed the proposed method performed well, managing to select the correct model in all

Tab. 2: Confusion matrix

	$\widehat{M}=1$	$\widehat{M}=2$	$\widehat{M}=3$
$M=1$	100	0	0
$M=2$	0	100	0
$M=3$	0	0	100

[2] in general other decomposition strategies are possible, as for example performing a decomposition into groups of parameters

of the considered cases.

Table 3 introduces the prior interval for each parameter, the true values used in the experiments and presents the results for the system identification part. The estimated values for the parameters were consistent with the true values for all of the considered models. We have chosen to present the estimation results only for the Cauchy model as for it the estimated values exhibited the highest standard deviation. As it can be observed, the method performed also well in terms of estimating the values for the parameters.

Tab. 3: Estimation results for the case of Cauchy noise

	Prior	True	Estimated			PSD[3]
			Gauss	Laplace	Cauchy	Cauchy
G_0	$[1, 10]$	3	4.974	3.086	3.025	0.088
γ	$[0.1, 10]$	1	0.101	0.102	0.930	0.130
t_0	$[0, 0.5]$	0.2	0.491	0.205	0.198	0.004
f_0	$[1, 50]$	10	3.503	11.334	10.076	0.553
τ_0	$[0, 1]$	0.1	0.169	0.077	0.093	0.020

We have chosen to present also the estimated values for the other two models in the case when the true model is the Cauchy one. As it can be observed, the best results are obtained for the case of the Cauchy model whereas the worst are obtained for the case of the Gauss model. The results obtained for the Laplace model are rather close to the true values but they fall short of the ones obtained in the case of the Cauchy model. This can be explained by the fact that among the three distribution, the Gaussian one has the thinnest tails while the Cauchy has the thickest ones. As with the results, the Laplace distribution sits in between the previous two models. Care should be taken when analysing the estimation results for the γ_k parameter as it does not have the same meaning across all models. The fact that it is underestimated for the Gauss and Laplace models is the correct behaviour as for the case of the two models a smaller value accounts for a higher noise level which is the case considering the true model to be the Cauchy one.

5 Conclusions

This paper tackled the joint problem of noise model selection and system identification, where the system identification problem is nested inside the model selection one.

One of the biggest challenges in performing Bayesian model selection is computing the, most often intractable, model evidence. The solution to which we resorted, introduced by [3], makes use of a clever rewriting of the Bayes formula for the posterior law of the parameters such that the evidence is expressed as a function of the posterior law. As sampling the posterior law despite not having it in a complete form is possible, then the method proceeds with using these samples to approximate the posterior law, and in doing so to provide also an approximation for the marginal likelihood.

[3]Posterior Standard Deviation

References

[1] J. Berger and L. Pericchi. The intrinsic bayes factor for model selection and prediction. Technical report, Purdue University, Mars 1993.

[2] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

[3] S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

[4] H. Chipman, E. I. George, and R. E. McCulloch. The practical implementation of Bayesian model selection. In *Model selection*, pages 65–116. Institute of Mathematical Statistics, 2001.

[5] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Texts in Statistical Sciences. Chapman & Hall/CRC, Boca Raton, FL, 2004.

[6] J.-F. Giovannelli and A. Giremus. Bayesian noise model selection and system identification based on approximation of the evidence. In *IEEE Workshop on Statistical Signal Processing*, 2014.

[7] P. J. Green. Reversible jump Markov chain Monte Carlo computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

[8] R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, June 1995.

[9] M. Newton, A. Raftery, J. Satagopan, and P. Krivitsky. Approximate Bayesian inference with the weighted likelihood bootstrap. In *Bayesian Statistics 8*, pages 1–45. Oxford University Press, 2007.

[10] A. Tomohiro. *Bayesian Model Selection and Statistical Modeling*. Statistics: Textbooks and Monographs. CRC Press, Boca Raton, FL, 2010.