2011 8th International Multi-Conference on Systems, Signals & Devices

MCMC FOR PARAMETERS ESTIMATION BY BAYESIAN APPROACH

H. Ait Saadi

Saad Dahlab University
SET Laboratory
Département de l'électronique
BP 270, Blida

F. Ykhlef, A. Guessoum

Saad Dahlab University
LATSI laboratory
Département de l'électronique
BP 270, Blida

ABSTRACT

This article discusses the parameter estimation for dynamic system by a Bayesian approach associated with Markov Chain Monte Carlo methods (MCMC). The MCMC methods are powerful for approximating complex integrals, simulating joint distributions, and the estimation of marginal posterior distributions, or posterior means. The Metropolis-Hastings algorithm has been widely used in Bayesian inference to approximate posterior densities. Calibrating the proposal distribution is one of the main issues of MCMC simulation in order to accelerate the convergence.

Index Terms— MCMC, Metropolis-Hastings, Bayesian approach, MMSE, dynamic system, parameters estimation

1. INTRODUCTION

The Systems identification or looking for models from experimental data is a major concern in most disciplines. It is usually described as a collection of methods or techniques to determine mathematical models able to reproduce as authentically as possible the performance of a physical, chemical, biological or economic system. Once the model structure is chosen or given, the problem of system identification is reduced to a problem of parameter estimation. The methods of estimation in statistical parametric identification are usually based on minimization the prediction error and the method of maximum likelihood. These techniques often require a sufficiently large number of data [1, 2].

The Bayesian approach is particularly efficient in the presence of short data measurements in contrast to other methods. It's taking into account both the information provided by the observations and knowledge available to the experimenter. The approach is presented as a method of estimating random models and requires the calculation of estimators from a posteriori probability distribution generally very complicated. To solve a variety of "unsolvable" problems in Bayesian inference (integration, normalization, marginalization, expectation, optimization), we use the Markov Chain Monte Carlo approach (MCMC). Many applications can be envisaged with MCMC; some of them are introduced in [4, 5, 6, 7, 8, 9].

Markov chain simulation is essentially a general technique based on generating samples from proposal distributions and then correcting (acceptance or rejection) those samples to approximate a target posterior distribution. Here we must know both the target $p(.)$, (up to a normalizing constant), and the proposal, $q(x)$. The most powerful and efficient MCMC methods are: the Metropolis-Hastings and Gibbs sampler and their variants.

In [3], for dynamic systems estimation and closed loop controller, a very good study have been presented for posterior computation via MCMC. However the authors underline some questions about the MCMC convergence acceleration, the proposal density choice and the estimation of the noise variance.

In this paper we want to compute the marginal posterior density function of a subset of parameters by using MCMC, and thus the MAP estimation of these parameters. The variance of the noise is added as nuisance parameter and the proposal distribution is calibrated to ensure a fast convergence of MCMC algorithm with low rejection rate.

2. BACKGROUND

Let consider a linear time-invariant model as described by [1]:

$$M(\theta): \quad y(t) = G(q,\theta)u(t) + H(q,\theta)e(t). \tag{1}$$

The structure of the general model is :

$$A(q,\theta)y(t) = \frac{B(q,\theta)}{F(q,\theta)}u(t) + \frac{C(q,\theta)}{D(q,\theta)}e(t) \tag{2}$$

with :

$$A(q,\theta) = 1 + a_1 q^{-1} + \cdots a_{n_a} q^{-n_a} \tag{3}$$
$$B(q,\theta) = b_1 q^{-1} + \cdots b_{n_b} q^{-n_b} \tag{4}$$
$$F(q,\theta) = 1 + f_1 q^{-1} + \cdots f_{n_f} q^{-n_f} \tag{5}$$
$$C(q,\theta) = 1 + c_1 q^{-1} + \cdots c_{n_c} q^{-n_c} \tag{6}$$
$$D(q,\theta) = 1 + d_1 q^{-1} + \cdots d_{n_d} q^{-n_d} \tag{7}$$

The parameter vector to be determined is :

$$\theta = [a_1, b_1, c_1, d_1, f_1 \ldots a_{n_a}, b_{n_b}, c_{n_c}, d_{n_d}, f_{n_f}, \sigma_e^2] \tag{8}$$

978-1-4577-0411-6/11/$26.00 ©2011 IEEE

The structures FIR, ARX, ARMAX, Box-Jenkins or OE, are particular cases from this model. A one-step-ahead predictor of (2), is given by :

$$\hat{y}(t|t-1,\theta) = H^{-1}(q,\theta)G(q,\theta)u(t)+ \left[1 - H^{-1}(q,\theta)\right] y(t) \quad (9)$$

where:

$$G(q,\theta) = \frac{B(q,\theta)}{A(q,\theta)F(q,\theta)} \quad (10)$$

$$H(q,\theta) = \frac{C(q,\theta)}{A(q,\theta)D(q,\theta)} \quad (11)$$

The prediction error is :

$$\begin{aligned} y(t) - \hat{y}(t|t-1,\theta) &= -H^{-1}(q,\theta)G(q,\theta)u(t) \\ &\quad + H^{-1}(q,\theta)y(t) \\ &= e(t,\theta) \end{aligned} \quad (12)$$

The estimation of θ is based on the prediction-error identification-method (PEM), generally with quadratic norm:

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \frac{1}{2}\left[y(t) - \hat{y}(t|t-1,\theta)\right]^2 \quad (13)$$

where $Z^N = [y(1), u(1), y(2), u(2), \ldots y(N), u(N)]$.

The Likelihood function of N-independent measurements or observations data $Y_N \triangleq \{y(1), \ldots, y(N)\}$ is given by:

$$p(Y_N\,\theta) = \prod_{t=1}^{N} p_e(y(t) - \hat{y}(\,t|\,t-1,\theta)) \quad (14)$$

where $p(\cdot)$ is noise probability density function.

Under some regularity conditions of [2], the maximum of likelihood estimator (MLE) is asymptotically Gaussian with:

$$\hat{\theta}^{MV}_{N(N\to\infty)} \sim \mathcal{N}(\theta_0, I_Y^{-1}(q_0)) \quad (15)$$

where I_Y^{-1} is the Fisher information matrix and θ_0 is the unknown true parameters vector such as :

$$I_Y(\theta_0) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\ln p(Y_N\,|\theta\,)\right]\Bigg|_{\theta=\theta_0} \quad (16)$$

The MLE is asymptotically efficient and unbiased (It's generally biased for finite N).

3. BAYESIAN ESTIMATION

We now depart from the classical approach to statistical estimation in which the parameter θ of interest is assumed to be a deterministic but unknown constant. Instead, we assume that θ is a random variable whose particular realization we must estimate. The Bayesian approach, when applicable, can therefore improve the estimation accuracy. In [10], an overview of all the principal estimation approaches is given. In Bayesian estimation the most popular formulations are: maximum a posteriori (MAP), maximum likelihood (ML), minimum variance (MV) or equivalently minimum mean-squared error (MMSE). In determining the MAP or MMSE estimators we first require the posterior distribution $p(\theta|Y_N)$ which represents the sum of the knowledge available. We can use Bayes' rule to determine it as:

$$p(\,\theta|\,Y_N) = \frac{p(Y_N|\,\theta)p(\theta)}{p(Y_N)} \quad (17)$$

$$= \frac{p(Y_N|\,\theta)p(\theta)}{\int p(Y_N|\,\theta)p(\theta)d\theta} \quad (18)$$

where $p(\theta)$ is the prior density (the knowledge possessed before measurement), $p(Y_N|\theta)$ is called the likelihood done by (more likely to be true) and $p(Y_N)$ is called the evidence (scales the posterior to assure its integral is unity).

The MAP estimation is the value of θ, that maximizes the posterior density:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|Y_N) \quad (19)$$

The MMSE estimation of one component θ^i(the optimal estimator in terms of minimizing the Bayesian MSE) is the mean of the posterior density :

$$\hat{\theta}^i_{MMSE} = E(\theta^i|Y_N) = \int_{\theta_i} \theta^i p(\theta^i|Y_N)\; d\theta^i \quad (20)$$

The posterior density distribution of one component from θ(marginal posterior) is obtained by calculating the following integral:

$$p(\theta^i|Y_N) = \int p(\theta|Y_N)d\theta^1 \ldots d\theta^{i-1}d\theta^{i+1} \ldots d\theta^d \quad (21)$$

The calculation of this integral in most real applications is complicated and involves numerical difficulties and cannot be solved analytically. It is therefore the same for the term of equation (13), which is often difficult to assess and the problem is usually circumvented by working with the function $\log p(\theta\,|Y_N)$. However, in this case the marginal densities cannot be calculated. This is the reason why Bayesian inference remained an elegant theoretical construction with little practical application until the introduction of efficient approximations techniques and in particular of Markov chain Monte Carlo (MCMC) simulation. There are two basic types of MCMC algorithms, the Metropolis algorithm and the Gibbs sampler, have been widely used in diverse fields.

4. METROPOLIS-HASTINGS ALGORITHM

The Metropolis algorithm was introduced in its simplest form by Metropolis in 1953 and generalized by Hastings

in 1970 [8]. The algorithm is a procedure to simulate a sample of a univariate and multivariate distribution. We must know both the target $\pi = p(\theta|Y_N)$ (up to a normalizing constant), and the proposal density $q(x)$, a priori.

To ensure that the stationary $\pi(\cdot)$ can be achieved, whatever the starting point at which the string is initialized $\theta_{(0)}$, it is necessary that the chain is ergodic. This implies that the chain is π-irreductible and aperiodic.

Irreducibility means that there is a positive probability that the Markov chain can reach any non-empty set from all starting points [9].

$$q(\theta_{(n-1)}|\xi) > 0 \ \ \forall \theta, \xi$$

Aperiodicity ensures that the chain will not oscillate between different sets of states. These conditions are usually satisfied if the proposal distribution has a positive density on the same support as the target distribution. They can also be satisfied when the target distribution has a restricted support.

The samples are sequentially generated forming an ergodic Markov chain (irreducibility and aperiodicity) with invariant density equal to the required posterior density. Typically, in Markov chain simulation, samples are generated from the transition kernel or distribution.

$$p(\theta_{(k)} = \theta|\theta_{(k-1)}) \tag{22}$$

The *key*, however, is not really the chain itself, but the fact that the approximate distribution improves sequentially as it converges to the target posterior.

$$\lim_{t\to\infty} p(\theta_{(t)} = \theta|\theta_{(init)}) = p(\theta|Y_N) \quad \forall\theta_{(init)} \tag{23}$$

From the reversibility condition :

$$\pi(\xi)p(\theta|\xi) = \pi(\theta)p(\xi|\theta) \tag{24}$$

the invariance condition is satisfied :

$$\begin{aligned}\int \pi(\xi)p(\theta|\xi)d\xi &= \int \pi(\theta)p(\xi|\theta)d\xi \\ &= \pi(\theta)\int p(\xi|\theta)d\xi = \pi(\theta)\end{aligned}$$

The Hastings sampler is an accept-reject routine which induces a Markov chain by generating candidate random variates from a transition probability $q(\cdot)$. The algorithm is as follows:

1. Initialize $\theta_{(0)}$
2. In k iteration

   a) Generate a candidate sample $\xi_{(k)}$ from proposal:

   $$\xi_{(k)} \sim \mathrm{q}(./\theta_{(k-1)})$$

   b) Calculate the acceptance probability :

   $$\alpha(\xi_{(k)}|\theta_{(k-1)}) = min\left\{1, \frac{p(\xi_{(k)}))|\theta_{(k-1)}}{p(\theta_{(k-1)}|Y_N)} \times \frac{q(\theta_{(k-1)}|\xi_{(k)})}{q(\xi_{(k)}|\theta_{(k-1)})}\right\} \tag{25}$$

   c) Generate $u_k \sim \mathcal{U}_{[0\,1]}$ (uniform distribution):

   - If $u_k < \alpha$, accept the proposed $\xi_{(k)}$ and set $\theta_{(k)} = \xi_{(k)}$, with probability $\alpha(\xi_{(k)}|\theta_{(k-1)})$,
   - else reject with $\theta_{(k)} = \theta_{(k-1)}$

3. Increment k and return to step 2

The Metropolis-Hastings algorithm yields a Markov chain for which the reversibility condition holds, and consequently, if the chain is also ergodic, then it converges to the invariant distribution [3].

The speed with which the ergodic limit is reached depends on the choice of the instrumental distribution combined with a judicious choice of initial values. This distribution is available either analytically (a constant), is symmetrical (e.g, such that $q(x_{(n)}|y) = q(y|x_{(n)})$). Moreover, to improve the acceptance rate of the algorithm, it must be simulated quickly and must be chosen according to the distribution of interest. In particular, must be a good approximation of π and must cover all its support. Indeed, if the instrumental support of the distribution is too small, some areas of support of are not explored and thus there will be no simulated sample, however distribution instrumental support would create too much too wide releases, and slow convergence of the algorithm. In general, the rules of choice distributions are instrumental heuristics. In practice, different choices of instrumental delivery can define several variants of the Metropolis-Hastings algorithm.

A special case of the Metropolis algorithm is random-walk Metropolis. It is defined by the relation $(q(\xi|\theta) = q(\theta - \xi))$. The candidate is generated as :

$$\xi_{(k)} = \theta_{(k-1)} + \varepsilon_k \tag{26}$$

Where the increment random variable (or random walk) ε is independent and identically distributed. A reasonable choice for this distribution is a symmetric Gaussian. The acceptance probability will be in this case:

$$\alpha = \left\{1, \frac{p(\xi_{(k)}|\theta_{(k-1)}, Y_{(N)})}{p(\theta_{(k-1)}, Y_{(N)})}\right\} \tag{27}$$

5. EXEMPLE OF SIMULATION WITH OE MODELE

We associate now MCMC with parameters estimation for an Output-Error model (OE). According the general structure (2), the likelihood function associated with an observed data record under independence assumption of random variables of the noise is given by :

$$\begin{aligned}p(Y_N|\theta) &= p(y(0)|\theta)\prod_{t=1}^{N} p(y(t)|Y(t-1),\theta) \\ &= p(y(0)|\theta)\prod_{t=1}^{N} p_e(y(t) - \hat{y}(t|t-1,\theta))\end{aligned} \tag{28}$$

where $p_e(.)$ represent the noise density function and $Y_N \triangleq \{y(1), y(2), \ldots, y(N)\}$. The posterior density $p(\theta|Y_N)$

calculated through the Bayes' rule is given by :

$$p(\theta|Y_N) = \frac{p(Y_N|\theta)p(\theta)}{p(Y_N)} = \frac{p(y(0)|\theta)p(\theta)}{\int p(Y_N|\theta)p(\theta)d\theta} \times \prod_{t=1}^{N} p_e(y(t) - \hat{y}(t|t-1,\theta)) \quad (29)$$

We take now the example of an Output-Error model (OE) defined by the flowing equation:

$$y(t) = (\frac{b_1 q^{-1} + b_2 q^{-2}}{1 + f_1 q^{-1}})u(t) + e(t) \quad (30)$$

For our simulation, we take: $F = [1 \;\; 0.8]$ and $B = [0 \;\; 0.5 \;\; 0.2]$, and $e(t)$ an independent and identically distributed (*i.i.d*) Gaussian noise with zero mean and variance σ_e^2.

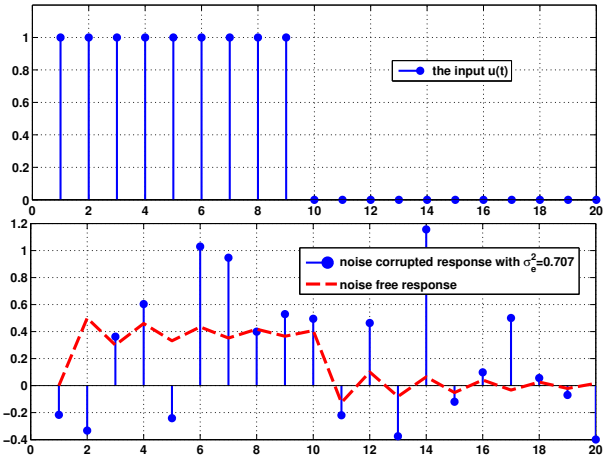

Figure 1. Example of System response of OE model, with $F = [1 \quad 0.8]$ and $B = [0 \quad 0.5 \quad 0.2]$.

The input $u(t)$ is a signal of 10 samples only, transiting from $1 \longrightarrow 0$ at $t = 5s$. The input and the response are shown by the Figure (1).

From a prior knowledge about the system and the parameters used $\theta = [f_1, b_1, b_2, \sigma_e^2]$, we know that $f_1 \in [-1, \quad 1]$ (for stability) et b_1 and b_2 are ≥ 0 (physical constraints). So we use a uniform prior density. According (26), the random walk method is adopted for the proposal density choice. The increment random variable is a gaussian perturbation $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. All the values of $\xi_{(k)}$ where f_1 is outside $[-1, +1]$ and b_1 or b_2 are < 0, will be rejected. The noise variance σ_e^2 is to be regarded as a nuisance parameter and we assign for it an $\mathcal{IG}(\alpha, \beta)$ prior PDF :

$$p(\sigma) \propto \frac{\beta^\alpha}{\Gamma(\alpha)}(\frac{1}{\sigma^2})^{\alpha+1} \exp(\frac{-\beta}{\sigma^2}) \; \mathbb{1}_{\mathbb{R}^+} \quad (31)$$

We assume that the nuisance parameter are independent of the desired parameters.

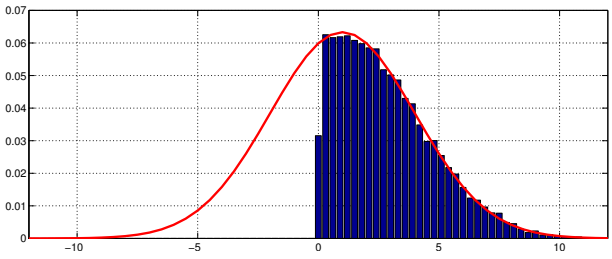

Figure 2. An example of positive normal distribution.

To accelerate the convergence of the Markov chain, we adopt some strategies :
First, the good calibration of the proposal distribution. In [11], an accept-reject algorithm has been proposed to simulate positive normal variables in the univariate case. This algorithm will increase the probability of acceptation (since b_1 and b_2 must be ≥ 0), and accelerate the convergence since it generates low reject, see the figure (2).
Second, the choice of σ_ε^2 is very important and must be done carefully. A wrong choice will compromise the convergence. In [12, 13, 14], the authors propose to adopt a combination of strategies aimed at evaluating and accelerating MCMC sampler convergence. According these recommendations, we propose to supervise the convergence by trying to have an acceptation rate between $(40\%-50\%)$. We test and repeat different parameters calibration for the 10^4 realizations only, till to have the good acceptance rate. Once the calibration is done, we compute for 5×10^4 realizations. The approximation of posterior densities is done by using histograms after a sufficiently long burn-in. The length of burn-in depends on the chain kernel and the rate of convergence, see [12, 16]. Heidelberger and Welch [15], suggest an iterative procedure, based upon repeated hypothesis tests of these statistics, in order to estimate the length of the burn-in. For simplicity, we monitor the convergence of the MCMC for several trials. The starting points for these chains are chosen to be widely dispersed in the target distribution (to prevent a convergence problem). The length of burn-in for our example it's around 1.5%. In order to shorten the length of burn-in, at the beginning, we take a relatively a large value of σ_ε^2, this value will be decreased by a half two times during the first 5000 iterations. The figure (3), depicts the first 5000 simulations of the chain.

For the simulation example described by (30), we take a noise variance of $\sigma_e^2 = 0.707$. The approximations of the marginal densities of the different parameters are shown in the figure (4). At the left, the posterior densities with 10^5 iterations and at the right with 5×10^4 iterations. The shapes of the densities are relatively the same. We can compare the maximum of these densities with the reel values. The PEM method minimizing the cost function given by (13), give us the following results: $f_1 = -0.89$ and $b_1 \simeq 0.4$ and $b_2 \simeq -0.32$. So, the "classical estimator" falls (f_1 and b_2) outside the known constraint interval. With bayesian estimation, $\hat{\theta}_{MAP} \simeq [0.8 \; 0.34 \; 0.14 \quad 0.68]$ and $\hat{\theta}_{MMSE} \simeq [0.376 \; 0.433 \; 0.327 \; 0.783]$. The MMSE

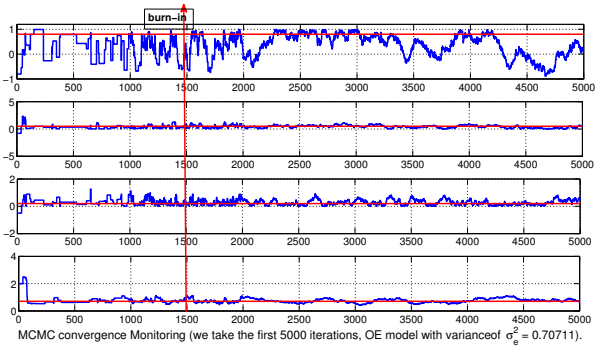

Figure 3. MCMC convergence monitoring of θ, we take the first 5000 iterations from Metropolis algorithm sampling, with $\sigma_e^2 = 0.707$.

estimator depends in general on the prior knowledge as well as the data. In the simulation example, we have a short data measurements and the MMSE estimator is "biased" towards the prior mean. We prefer the MAP estimator which is usually easier to determine since it involve only a maximization of the marginal posterior densities already computed by the MCMC.

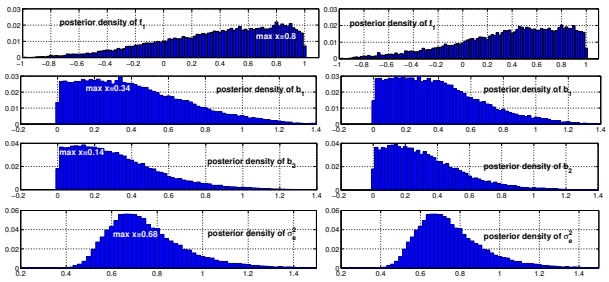

Figure 4. Histograms of the MCMC samples which represent the posterior marginal densities of parameters, OE model with $\sigma_e^2 = 0.707$, $F = [1, \quad 0.8]$ and $B = [0 \quad 0.5 \quad 0.2]$.

6. CONCLUSION

In absence of data measurements, the Bayesian approach can improve the estimation accuracy by incorporating the prior knowledge in the estimation. By using Markov Chain Monte Carlo methods (MCMC) methods, the marginal posterior densities can be calculated easily. The calibration of MCMC parameters for accelerating the convergence depends on the prior knowledge of the system studied. We can improve the performances according the considered application. The incorporation of the noise variance in the parameters to be estimated, affect the accuracy but remain acceptable. The MCMC methods can be explored more and more to profit of the Bayesian estimation advantages in many fields.

7. REFERENCES

[1] L. Ljung, *System identification theory for user*, Prentis-Hall, Englewood Cliffs, 1987.

[2] E. Walter, L. Pronzato, *identification de modèles paramétriques à partir de données expérimentales*,Masson, 1994.

[3] B. Ninness, S. Henrihsen and T. Brinsmead, "System identification via a computational Bayesian approach," in *Proc. Decision and Control Proceedings of the 41st IEEE Conference*, 2002, vol. 2, pp. 1820–1825.

[4] A. Mira, *MCMC Methods to Estimate Bayesian Parametric Models*, Elsevier BV, Handbook of Statistics, Vol. 25, 2005.

[5] W. Gilks, S, Richardson and D. Spiegelhalter, *Markov Chain Monte Carlo in practice*, Chapman and Hall, 1996.

[6] James V. Candy, *BAYESIAN SIGNAL PROCESSING: Classical, Modern, and Particle Filtering Methods*, WILEY, 2009.

[7] W. K. Ching and K. N. Michael, *Markov Chains: Models Algorithms and Applications*, Springer, 2006.

[8] W. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika,* vol. 57, pp. 97–109, 1970.

[9] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings Algorithm," *The American Statistician,* vol. 49, no. 4, pp. 327–335, Nov. 1995.

[10] S. M. Kay, Steven, *Fundamentals of statistical signal processing: estimation theory*, Prentice-Hall, 1993.

[11] V. Mazet, D. Brie and J. Idier, "Simulation of positive normal variables using several proposal distributions," *IEEE Statistical Signal Processing,* pp. 37–42, Jul. 2005.

[12] M. K. Cowles and B. P. Carlin, "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," in *Proc. Journal of the American Statistical Association*, 1996, , pp. 883–904.

[13] A. Gelman, G.O. Roberts and W.R. Gilks, *Efficient metropolis jumping rules*, J.M. Bernardo, J.O. Berger, A.P. Dawid et A.F.M. Smith Bayesian Statistics, Oxford University Press, pp. 599–608, 1996.

[14] G.O. Roberts, A. Gelman and W.R. Gilks, *Weak convergence and optimal scaling of random walk Metropolis algorithms*, The Annals of Applied Probability, vol. 7, no. 7, pp. 110–120, 1997

[15] P. Heidelberger, P. D. Welch, , *Simulation Run Length Control in the Presence of an Initial Transient*, Operations Research, no. 31, 1109–1144, 1983

[16] S.P. Brooks et G.O. Roberts, *Convergence assesment techniques for Markov chain Monte Carlo*, Statistics and Computing, vol. 7, no. 8, 319–335, 1998