Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 35 (2014) 1474 – 1481

Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

Bayesian estimations with isotropic and anisotropic matrices for a multinomial logit model

Nozomi Matsumoto[a], Takeshi Kurosawa[b,*]

[a] *Department of Mathematical Information Science, Graduate School of Science, Tokyo University of Science, 1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan*

[b] *Department of Mathematical Information Science, Faculty of Science, Tokyo University of Science, 1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan*

Abstract

This paper studies a Metropolis-Hastings (MH) algorithm of unknown parameters for a multinomial logit model. The MH algorithm which is one of the Bayesian estimation requires prior and proposal distributions. A selection of the prior and proposal distributions is an important issue of the Bayesian estimation. However, there is no a decisive approach for the determination of prior and proposal distributions. A posterior distribution is obtained from two distributions. The MH algorithm generates samples from the posterior distribution of the unknown parameters. Unless we give appropriate distributions, it leads to an inappropriate posterior distribution. In this paper, we discuss differences in the behaviors of autocorrelation functions in a selection of the proposal distributions.

Peer-review under responsibility of KES International.

Keywords: Bayesian estimation, isotropic, anisotropic, proposal distribution, autocorrelation functions, Metropolis-Hasting algorithm

1. Introduction

For an estimation of a multinomial logit model, the likelihood estimation is widely used because the maximum likelihood estimators of unknown parameters are consistent and asymptotically normal distributed and the uniqueness of the optimal solution of the maximum likelihood estimation is ensured under weak conditions for the multinomial logit model. On the other hand, the Bayesian estimation obtains distributions of the unknown parameters which are treated with random variables. A concept of the Bayesian estimation is completely different from the one of the likelihood estimation. Since the Bayesian estimation is based on flexible assumptions, it is widely used for estimations of statistical models, not limited to the estimation of the multinomial logit model.

One of the representative methods of the Bayesian estimations are Markov chain Monte Carlo methods [1] such as Gibbs sampling[2], Metropolis method[3], and Metropolis-Hastings (MH) method[4]. The Gibbs sampling cannot be

* Corresponding author. Tel.: +81-3-5228-8202 ; fax: +81-3-3260-4293.
E-mail address: tkuro@rs.kagu.tus.ac.jp

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.211

directly applied to the estimation of the multinomial logit model because the posterior distribution for their unknown parameters does not have a full conditional distribution. The MH algorithm is an effective method which approximates to the posterior distribution, regardless of the forms of statistical models and prior distributions. In the Bayesian estimation, we require two important distributions to generate a Markov chain: 1. A *prior distribution*, 2. A *proposal distribution*. Using two distributions with observed data, we get a *posterior distribution* of unknown parameters. In this paper, we discuss an appropriate selection of the proposal distribution on a MH algorithm for a multinomial logit model. In section 2, we introduce briefly a MH algorithm and their diagnostics. In the next section, we give a case study and proceed to a conclusion of this study.

2. Metropolis-Hastings algorithm

The following procedure is the Metropolis-Hastings algorithm (MH). There are a lot of literatures and books relating to the MH algorithm (See E.g.[5]).

The Metropolis-Hastings algorithm

Given a current value $\boldsymbol{\theta}^{(s)}$,

1. Generate $\boldsymbol{\theta}^*$ from $J(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(s)}, \boldsymbol{y})$;
2. Compute the acceptance ratio
$$r = \frac{p(\boldsymbol{\theta}^*)p(\boldsymbol{y} \mid \boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(s)})p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(s)})} \times \frac{J(\boldsymbol{\theta}^{(s)} \mid \boldsymbol{\theta}^*, \boldsymbol{y})}{J(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(s)}, \boldsymbol{y})};$$
3. Set $\boldsymbol{\theta}^{(s+1)} = \begin{cases} \boldsymbol{\theta}^* & \text{if } u < r \\ \boldsymbol{\theta}^{(s)} & \text{otherwise} \end{cases}$, where u is a sample from uniform distribution $U(0, 1)$.

In the above algorithm, J is a *proposal distribution* and $P(\boldsymbol{y} \mid \boldsymbol{\theta})$ is a *posterior distribution*. The proposal distribution generates candidates for the parameter values. Sometimes the value is rejected following a value of an acceptance ratio r. Although a choice of the prior distribution is an important issue for the Basysian etimation, we mainly focus on proposal distribution in this paper.

Assume that there are S samples of the Markov chain except those in the burn-in period. The following diagnostics are used in this study.

1. Trace plot of the Markov chain
   Check whether the chain illustrated along the time series moves stable without depending on an initial value.
2. The autocorrelation function (cf. E.g,[5])
   A *sample autocorrelation function* computes correlation among the values of the chain. The lag-t autocorrelation function estimates the correlation between elements of the sequence that are t steps apart. The *effective sample size* also helps to measure the autocorrelations of the chain.
3. Geweke's diagnostic[6]
   The approach which tests the difference between the means of the values in the early sequence A and in the later sequence B. Based on[6], we put the sample size p_A and p_B in two groups for $p_A = 0.1S$ and $p_B = 0.5S$, respectively.
4. Gelman-Rubin's diagnostic[7]
   Gelman-Rubin's diagnostic $\hat{R}$ gets a look at variances of multiple chains starting from different values for each of the parameters. It regards that the chain converges if the values of $\hat{R}$ are all less than 1.1 or 1.2 in practice.

Although we used all the diagnostics in our simulation, the results of trace plot and Gelman-Rubin's are not omitted in this paper.

3. Simulation studies

We conduct a simulation study for a Metropolis-Heistings (MH) algorithm. We use weather data at the Japan Meteorological Agency in Yokohama (From January first, 2009 to December 31th, 2011)[8]. Table 1 shows the variables and their corresponding parameters. Note that the quantitative variables are normalized.

Table 1. Variables corresponding to their parameters.

	variable	parameter	quantitative
Weather in the next morning (sunny, cloudy, rainy)	Y	–	–
Atmospheric pressure	X_1	β_1	Yes
Maximum temperature	X_2	β_2	Yes
Hours of sunshine	X_3	β_3	Yes
Whether rainy or not in the today afternoon	X_4	β_4	No
Constant (sunny)	–	β_5	No
Constant (cloudy)	–	β_6	No

Suppose that the statistical model with these data $(X_1, \ldots, X_4) = (x_{1n}, \ldots, x_{4n})$ $(n = 1, \ldots, 1095)$ is the multinomial logit model as follows:

$$P(Y_n = i) = \frac{e^{z_{in}}}{\sum_{j=1}^{3} e^{z_{jn}}} \quad (n = 1, \ldots, 1095), \tag{1}$$

where $z_{1n} = \beta_1 x_{1n} + \beta_2 x_{2n} + \beta_5, z_{2n} = \beta_3 x_{3n} + \beta_6, z_{3n} = \beta_4 x_{4n}$. In (1), the values 1,2, and 3 of Y_n correspond to "sunny", "cloudy", and "rainy", respectively. We estimate the unknown coefficient parameters $\beta_1, \ldots, \beta_6$. These parameters are noted as a vector of the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_6)^{\mathrm{T}}$. We describe the settings to run the MH algorithm as follows.

Based on a random walk chain algorithm, a proposal value $\boldsymbol{\beta}^*$ given a current value $\boldsymbol{\beta}^{(s)}$ in s-th iteration is generated as follows:

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}^{(s)} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \mathrm{diag}({\delta_1}^2, \ldots, {\delta_6}^2), \tag{2}$$

where $\mathbf{0}$ is the 6-dimensional vector. In this paper, we proceed with our estimation using the covariance matrix without the covariances as in (2). The MH algorithm consisting of $S = 55,000$ iterations with the initial value $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ is conducted. To discuss the differences in proposal distributions, we suppose on ahead that the *noninformative prior distribution* is

$$\boldsymbol{\beta} \sim N(\mathbf{0}, 1000000\boldsymbol{I}), \tag{3}$$

where $\boldsymbol{I}$ is the identity matrix. The samples of the chain in the first $5,000$ iterations are discarded as those in the burn-in period and samples in the remaining $50,000$ iterations are used for the estimation.

We seek an appropriate covariance matrix $\boldsymbol{\Sigma}$ of the random walk chain algorithm in Section 3.1 and 3.2 under the noninformative prior (3). We treat isotropic and anisotropic covariance matrices in Section 3.1 and Section 3.2, respectively. This simulation study is performed to compare the behaviors of the isotropic covariance matrix and the anisotropic covariance matrix.

3.1. Determination of the appropriate isotropic covariance matrix

We find an appropriate covariance matrix $\boldsymbol{\Sigma}$ in (2) when it is restricted to be an *isotropic covariance matrix* (homoscedasticity). The isotropic covariance matrix is proportional to the identity matrix

$$\boldsymbol{\Sigma} = \delta^2 \boldsymbol{I}. \tag{4}$$

Namely, we assume that $\delta_1^2 = \cdots = \delta_6^2$ in (2). We simulate with the following three different variances δ^2 in (4) :

$$1.\ \delta^2 = 0.1^2, \qquad 2.\ \delta^2 = 0.003^2, \qquad 3.\ \delta^2 = 0.0001^2.$$

In fact, although some variances were applied to the MH algorithm in our simulation, we pick up only three cases. In this simulation, we use the noninformative prior distribution (3) for a prior distribution $p(\boldsymbol{\beta})$ to find an appropriate covariance matrix without depending on the prior distribution.

Table 2. Comparison with the posterior means (std. deviation).

	β_1	β_2	β_3	β_4	β_5	β_6
$\delta^2 = 0.003^2$	−0.2858	−0.4395	−0.3077	2.7403	1.9782	1.8432
	(0.0701)	(0.0695)	(0.0670)	(0.1902)	(0.1359)	(0.1364)
$\delta^2 = 0.0001^2$	−0.2858	−0.4395	−0.3077	2.7403	1.9782	1.8432
	(0.0730)	(0.0765)	(0.0674)	(0.1822)	(0.1195)	(0.1176)
$\delta^2 = 0.1^2$	−0.2847	−0.4367	−0.3076	2.7157	1.9513	1.8189
	(0.0606)	(0.0724)	(0.0675)	(0.1931)	(0.1530)	(0.1515)

Three different proposal variances are evaluated with multiple convergence diagnostics. We see in Table 2 that the posterior means and its standard deviations in three different variances δ^2 are nearly equal. We cannot observe at first glance the differences among the three cases in the simulation from the view point of the posterior means. In that sense, the MH algorithm enables us to obtain the estimates of the unknown parameters without depending on a choice of the proposal distribution. However, the choice of the proposal distribution depends on the convergence of the chain. Moreover, the purpose of the Bayesian estimation is to get a posterior distribution without dependency on a proposal distribution, not a point estimation. Figure 1 shows the values of the autocorrelation functions. Here, we pick up only the result relating to β_1.

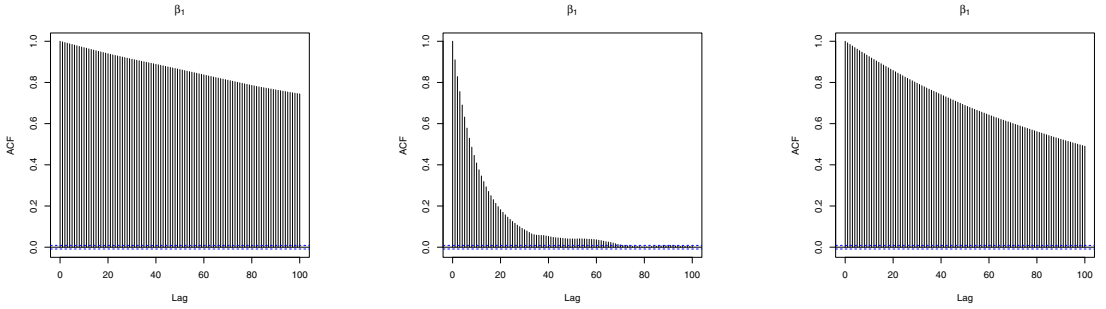

Fig. 1. The values of the autocorrelation functions in β_1.

Obviously, the behaviors of the autocorrelation functions are different. The high values of the autocorrelation functions mean that the values in the chain are dependent. It is an unwelcome fact, since we want to get independent samples from the chain. Although the MH algorithm intrinsically yields a dependency chain, it is desirable to get the more independent chain. In that sense, the middle variance is the most appropriate in the view point of the autocorrelation functions. The chain given in the large variance $\delta^2 = 0.1^2$ has high autocorrelations, which implies that the chain does not accept for a new proposal value. Hence the chain goes sideways for a long period. Similarly, the chain given in the small proposal variance $\delta^2 = 0.0001^2$ has also high autocorrelations. The small variance with $\delta^2 = 0.0001^2$ means that $\boldsymbol{\beta}^*$ is very close to $\boldsymbol{\beta}^{(s)}$. It leads to the high correlation between the present value and the proposal value. The following table shows the acceptance ratios of the each variance.

In the middle variance $\delta^2 = 0.003^2$, $\boldsymbol{\beta}^*$ is accepted as the value of $\boldsymbol{\beta}^{(s+1)}$ for 38 % of the iterations. It lies in an appropriate range of a MH algorithm. The Geweke's diagnostics are shown Table 4 below. Some Geweke's

Table 3. Acceptance rate in $\delta^2 = 0.1^2$, $\delta^2 = 0.003^2$, and $\delta^2 = 0.0001^2$.

	$\delta^2 = 0.1^2$	$\delta^2 = 0.003^2$	$\delta^2 = 0.0001^2$
Percentage of the acceptance	0.2%	38%	86%

diagnostics are not calculated in the small variance. Under the significance level for $\alpha = 0.05$, the null hypothesis of that there is no difference between the mean in the early sequence and in the later sequence is rejected. In the small and the large variances, the chains do not converge from the view point of Geweke's diagnostic.

Table 4. Geweke's diagnostics

	β_1	β_2	β_3	β_4	β_5	β_6
$\delta^2 = 0.0001^2$	−3.2654	−2.7750	−2.3006	−	−	3.2986
$\delta^2 = 0.003^2$	−1.2334	−0.7610	−0.3377	−0.8853	−0.9211	−1.1105
$\delta^2 = 0.1^2$	3.2910	3.7490	1.8180	−0.8100	1.8380	3.9800

From the above discussion, we may conclude that the appropriate variance is $\delta^2 = 0.003$ and the autocorrelation functions of the parameter β_1 is appropriate. However, the autocorrelation functions of the other parameters are not good (See Figure 2). This fact is discussed in the next subsection.

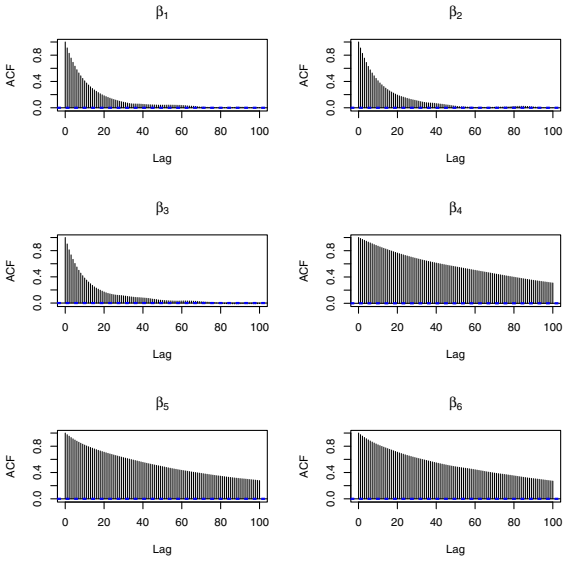

Fig. 2. Plots of the autocorrelation functions when $p(\boldsymbol{\beta})$ is the noninformative prior distribution and $\boldsymbol{\Sigma}$ is isotropic.

3.2. Isotropic matrix

In Section 3.1, we found the appropriate isotropic covariance matrix. However, we never conclude that this isotropic covariance matrix leads to an effective estimation for all the parameters when the covariance matrix is restricted to be isotropic. Obviously the autocorrelations with respect to β_4, β_5, and β_6 take high values in comparison with the results of β_1, β_2 and β_3 (See Figure 2). We note that β_4, β_5, and β_6 correspond to the parameters with the

qualitative variables. However, the acceptance ratio of the chain is in the acceptable range. Hence we use ${\delta_1}^2, \ldots, {\delta_6}^2$ so that the covariance matrix (2) is not restricted to be isotropic in this section to get the chains with low autocorrelations. The matrix is called an *anisotropic covariance matrix* (heteroscedasticity). Meanwhile, suppose that anisotropic variances ${\delta_1}^2, \ldots, {\delta_6}^2$ are given by

$$ {\delta_1}^2 = {\delta_2}^2 = {\delta_3}^2 = 0.003^2, \quad {\delta_4}^2 = {\delta_5}^2 = {\delta_6}^2 = 0.03^2. \tag{5} $$

We compare the chains of the anisotropic with those of the isotropic covariance matrix using the MH algorithm. We omit here a discussion of the process to obtain (5). Suppose that the noninformative prior distribution (3) is used for the prior distribution $p(\boldsymbol{\beta})$ for similar reasons in the case of the isotropic matrix. Given the above settings, the simulated results are shown in Figure 3.

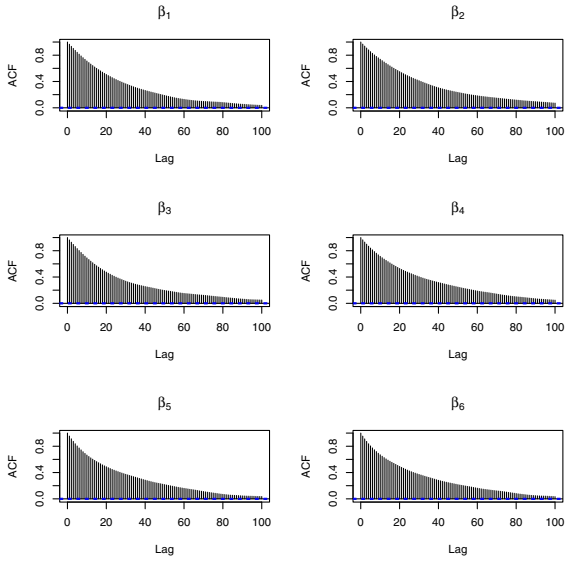

Fig. 3. Plots of the autocorrelation functions when $p(\boldsymbol{\beta})$ is the noninformative prior distribution and $\boldsymbol{\Sigma}$ is anisotropic.

Clearly, the chains β_4, β_5, and β_6 with the anisotropic matrix take the lower autocorrelations than those in the case of the isotropic one (Compare Figure 2 with Figure 3). Although we have the results of the balanced estimation, β_1, β_2, and β_3 have higher autocorrelations than the results of being restricted to be isotropic. The anisotropic matrix enables us to obtain the lower autocorrelations for some specific parameters. However, we loose the effectiveness of the chains for the specific parameters β_1, β_2, and β_3. Although the use of the proposal distribution with an anisotropic covariance matrix yields a balanced estimation, it does not always lead to an estimation with lower autocorrelations for all the parameters.

3.3. Further discussion (informative prior)

Until the previous subsections, we applied the noninformative prior distribution (3) to get an appropriate proposal distribution without depending on the influence of the prior distribution. Hence we apply a informative prior instead of the the noninformative prior distribution (3) for the further discussion. Here we assume that "informative" is

$$ \boldsymbol{\beta} \sim N(\mathbf{0}, 10\boldsymbol{I}). $$

We proceed with our estimation using the informative prior distribution.

Figure 4 shows the results of the autocorrelation functions under the informative prior with the isotropic covariance matrix. It is similar to the result under the noninformative prior with the isotropic matrix (Compare with Figure 2).

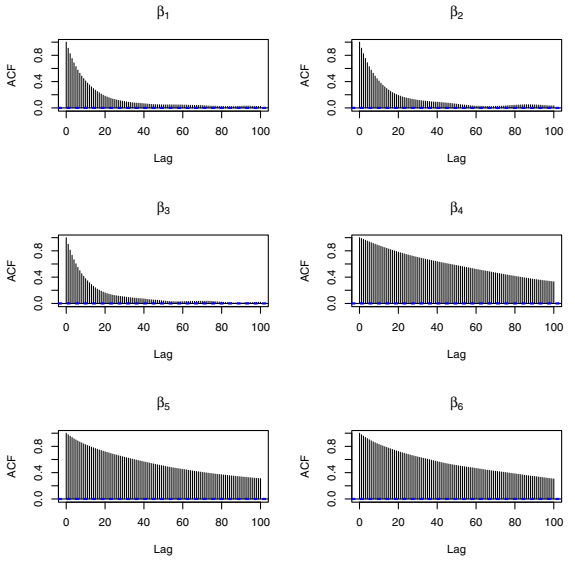

Fig. 4. Plots of the autocorrelation functions when $p(\boldsymbol{\beta})$ is the informative prior distribution and $\boldsymbol{\Sigma}$ is isotropic.

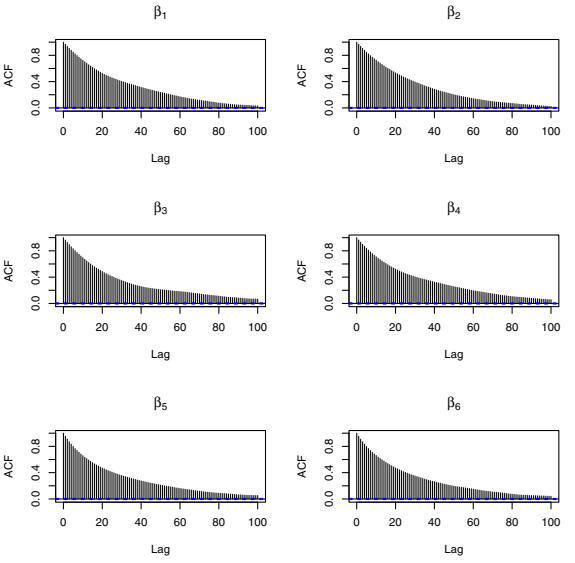

Fig. 5. Plots of the autocorrelation functions when $p(\boldsymbol{\beta})$ is the informative prior distribution and $\boldsymbol{\Sigma}$ is anisotropic.

The chains are also unbalanced. The autocorrelation functions of the chains in β_1, β_2, and β_3 keep low values. On the other hand, the ones in the other paramters keep high values. The use of the informative prior does not give a balanced estimation under the proposal distribution with the isotropic covariance matrix. Figure 5 shows the results of the informative prior with the anisotropic covariance matrix. The use of the anisotropic covariance matrix leads to the similar result under the noninformative prior with the anisotropic matrix (Compare Figure 3 and 5). Namely, the use of the anisotropic covariance matrix yields the effectiveness of the estimation for β_4, β_5, and β_6 and then we get a balanced estimation. Instead of this, the estimation looses the effectiveness for the specific parameters β_1, β_2, and β_3. This phenomenon is similar to the noninformative prior discussed in Section 3.2.

4. Conclusion

In this paper, we showed behaviors of a MH algorithm with a random walk algorithm. Under the noninformative prior distribution, we seeked the appropriate proposal distribution without depending on the prior distribution. When the covariance matrix was restricted to be isotropic, it seems that the plausible estimation is obtained because some diagnostics showed the convergence of the chain.

Next we considered the anisotropic covariance matrix in the random walk chain algorithm to reduce the values of the autocorrelation functions for the specific parameters. Although the anisotropic covariance matrix with the appropriate values for some specific parameters was set, it did not lead to the effective estimation for all the parameters. While the chains for some specific parameters had lower autocorrelations, the chains for the other parameters have higher autocorrelations. That is, the anisotropic covariance matrix yielded a balanced estimation, however, it did not lead to an estimation with lower autocorrelations for all the parameters. In other words, the anisotropic covariance matrix adjusts how effective to estimate for each of the parameters.

This study is limited to a selection of the proposal distribution and we mainly discuss the behaviors of the autocorrelation functions. It does not contain a selection of the prior distribution. We shall also give a concept of the selection of the prior distribution and show a simulation study with a balanced and efficient chain in the presentation of the conference.

References

1. Tierney, L.. Markov chains for exploring posterior distributions. *The Annals of Statistics* 1994;**22**:1701–1762.
2. Geman, S., Geman, D.. Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1984;**6**:721–741.
3. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* 1953;**21**:1087–1092.
4. Hastings, W.K.. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970;**57**:97–109.
5. Hoff, P.D.. *A First Course in Bayesian Statistical Methods*. New York: Springer Texts in Statistics; 2009.
6. Geweke, J.. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics* 1992; **4**:169–193.
7. Gelman, A.. Inference and monitoring convergence. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J., editors. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall/CRC; 1996, p. 131–143.
8. The Japan Meteorological Agency, . `http://www.jma.go.jp/jma/indexe.html`; 2013.