A VARIATIONAL CONTROL VARIABLE FOR ASSESSING THE CONVERGENCE OF THE GIBBS SAMPLER

Chuanhai Liu, Jun Liu, Donald B. Rubin
Department of Statistics, Harvard University, Cambridge, MA 02138

KEY WORDS: Markov chain, mutual information.

1. Introduction

The Gibbs sampler is an iterative simulation scheme for generating samples that converge to draws from a target distribution $\pi(X)$ of random variable X. An iterative simulation scheme is used because direct simulation from the target distribution can not be easily implemented. To define the iterative Gibbs sampler, partition the components of X as $X = (x_1, \ldots, x_d)$ where x_i is g_i-dimensional, and thus X is a $g = \sum_{i=1}^{d} g_i$ dimensional random vector. The Gibbs sampler is easy to implement when the set of d conditional distributions,

$$\pi(x_i \mid X_{[-i]}) \quad i = 1, 2, \ldots, d, \tag{1}$$

where $X_{[-i]}$ denotes $\{x_j,\ j \neq i\}$, are easy to draw from. The basic idea of the scheme is to construct a Markov chain with the target $\pi(X)$ as its equilibrium distribution. The chain is initiated by a draw from some starting density $p_0(X)$ (or a fixed point), then each variate x_i is visited and updated by a sample drawn from the conditional distribution $\pi(x_i \mid X_{[-i]})$. For example, the most widely used visiting scheme is a systematic one that visits each variate in turn. Detailed descriptions are found in Geman and Geman (1984), Gelfand and Smith (1990), and a variety of other recent references. Under certain regularity conditions, as long as each variate is visited infinitely often, the distribution of X defined in such a manner will converge to $\pi(X)$, and the rate is usually geometric (Geman and Geman 1984; Tanner and Wong 1987; Schervich and Carlin 1991; Liu, Wong and Kong 1991).

However, despite the wealth of theoretical results on the rate of convergence given in previously mentioned papers, no practically applicable criteria for stopping the the sequence of iterations appear there. In practice, Tanner and Wong (1987), Gelfand and Smith (1990) proposed some useful exploratory techniques to monitor the convergence visually by plotting marginal densities, and by observing the stability of certain summary variables like the means or variances of components of X. However, their techniques are questionable for use with high dimensional problems, which are common in application. An interesting example in Gelman and Rubin (1991a) indicates that we can be seriously misled by such exploratory techniques.

One applicable proposal for monitoring convergence is using independent multiple chains. Specifically, m Markov chains $\{X^{(j,t)}, j = 1, 2, \ldots, m;\ t = 0, 1, 2, \ldots\}$ are independently simulated with starting points $X^{(j,0)} \sim p_0(X)$, $j = 1, 2, \ldots, m$, where it is required that the starting distribution $p_0(X)$ is overdispersed relative to the target one, and that such "overdispersion" is preserved from iteration to iteration. Let $p_t(X)$ be the distribution of $X^{(j,t)}$, where for the same t, the parallel samples are i.i.d. Gelman and Rubin (1991b), by using multiple chains, propose a method based on a comparison, for each scalar function of X about which inferences are desired, of the "within" sample variance for each parallel chain and the "between" sample variance of different chains. It provides a "conservative" inference for the distribution of these components and gives a factor by which this distribution might become sharper if the simulations were continued indefinitely.

Here, we define a scalar global control variable based on samples drawn, and thereby reduce the assessment of the convergence of a high dimensional Gibbs sampling procedure to that of a one dimensional random variable. The constructed control variable incorporates information on the overall performance of convergence of the full joint distribution of X so that the judgement made based on our method is more encompassing than one based on subcomponents. The proposed control variable can be combined with the use of Gelman and Rubin's method to provide a stopping criteria. In Section 2, the control variable is constructed and justified by simple theoretical arguments; Section 3 provides an example to illustrate our method; Section 4 concludes with a brief discussion.

2. Assessing the convergence by a variational control variable

At iteration t of the Gibbs sampler, we construct a set of $m(m-1)$ values of the control variable U, each using two different parallel chains i and j,

$$U^{(i,j,t)} = \frac{\pi(X^{(j,t)})}{\pi(X^{(i,t)})} \cdot \frac{T(X^{(i,t)} \mid X^{(j,t-1)})}{T(X^{(j,t)} \mid X^{(j,t-1)})} \qquad (2)$$

where the ratio of target densities at $X^{(j,t)}$ and $X^{(i,t)}$ can be obtained from knowing only the conditional densities (Besag 1974),

$$\frac{\pi(X)}{\pi(Y)} = \prod_{p=1}^{d} \frac{\pi(x_p \mid x_1, \ldots, x_{p-1}, y_{p+1}, \ldots, y_d)}{\pi(y_p \mid x_1, \ldots, x_{p-1}, y_{p+1}, \ldots, y_d)},$$

and $T(X \mid Y)$ represents the conditional probability density for the transition from state Y to state X in one step of the Gibbs sampler scheme such that

$$p_n(X) = \int T(X \mid Y) p_{n-1}(Y) dY.$$

Typically when a systematic visiting scheme for the variables is used,

$$T(X \mid Y) = \prod_{p=1}^{d} \pi(x_p | x_1, \ldots, x_{p-1}, y_{p+1}, \ldots, y_d).$$

In the case of $d = 2$, (i.e., that special case of data augmentation, Tanner and Wong 1987), (2) has a simpler form

$$U^{(i,j,t)} = \frac{\pi(x_2^{(j,t-1)} \mid x_1^{(i,t)})}{\pi(x_2^{(j,t-1)} \mid x_1^{(j,t)})}. \qquad (3)$$

When the transition function has the same support space as the target distribution for the Markov chain, which is true for most of applications of the Gibbs sampler, $U^{(i,j,t)}$ in (2) has the following property which reveals information on convergence:

$$E_0(U^{(i,j,t)}) = Var_\pi(\frac{p_t(X)}{\pi(X)}) + 1. \qquad (4)$$

The expectation on the left hand side of (4) is taken under the initial distribution and the variance on the right hand side is taken under the target distribution. Therefore, the expectation of $U^{(i,j,t)}$ is an unbiased estimates of a relative distance between distribution p_t and the target distribution π. For fixed t, the $m(m-1)$ values of and $i \neq j, i, j = 1, 2, \ldots, m$, $U^{(i,j,t)}$ are identically distributed with cumulative distribution function $F^{(t)}(u)$.

Proof of (4): Since $(X^{(i,t-1)}, X^{(i,t)})$ and $(X^{(j,t-1)}, X^{(j,t)})$ are independent for $i \neq j$, we have

$$\begin{aligned}
& E\left\{\frac{\pi(X^{(j,t)})T(X^{(i,t)} \mid X^{(j,t-1)})}{\pi(X^{(i,t)})T(X^{(j,t)} \mid X^{(j,t-1)})}\right\} \\
= & \iiint \frac{\pi(y)T(x|z)}{\pi(x)T(y|z)} p_t(x)T(y|z)p_{t-1}(z)dxdydz \\
= & \iint \frac{p_t(x)T(x|z)p_{n-1}(z)}{\pi(x)} dzdx = \int \frac{p_t^2(x)}{\pi(x)} dx \\
= & 1 + \int (\frac{p_t(x)}{\pi(x)} - 1)^2 \pi(x)dx. \qquad \square
\end{aligned}$$

From the above argument one may be happy to have such a wonderful thing as the control variable U by estimating the expectation of which one can tell the distance between the two distributions. It is indeed wonderful when the variance of U is tolerable, which includes the cases where distributions have lower bounds larger than zero, and some other cases. However in many real examples, the variance of U can be exceedingly large, including possibly infinity. Thus a direct use of the sample mean and variance of U is usually avoided when one is not sure about its variance level. Taking logarithm of U is not a bad idea. It is observed that

$$\begin{aligned}
& E_0\{\log(U^{(i,j,t)})\} \\
= & E_0\left[\log\{\frac{T(X^{(i,t)} \mid X^{(j,t-1)})}{T(X^{(j,t)} \mid X^{(j,t-1)})}\}\right] \\
= & -2I(X^{(t)}, X^{(t-1)}) \qquad (5)
\end{aligned}$$

in which $I(X, Y)$ for any pair of random variables is a measure of the dependence called *symmetrized mutual information.* It is generally defined as

$$\begin{aligned}
I(X,Y) &= \frac{1}{2}\left[E_J\left(\log \frac{f(X,Y)}{f_X(X)f_Y(Y)}\right)\right. \\
&+ \left. E_I\left(\log \frac{f_X(X)f_Y(Y)}{f(X,Y)}\right)\right] \geq 0,
\end{aligned}$$

where $f_X(x)$ and $f_Y(y)$ are marginal distributions of X and Y respectively, "E_J" indicates the expectation taken under the joint distribution $f(x, y)$, and "E_I" the expectation taken under independent measure $f_X(x)f_Y(y)$. The second equality of (5) can be justified by that if (y, z) are consecutive draws from one simulation chain, and x is from another independent chain, then

$$\begin{aligned}
& E_0\left[\log \frac{T(x \mid z)}{T(y \mid z)}\right] \\
= & \int \log\{T(x|z)\} p_t(x)p_{t-1}(z)\ dxdz
\end{aligned}$$

$$
\begin{aligned}
& -\int \log\{T(y|z)\}p_t(y,z)dydz \\
= \quad & -2I(X^{(t)}, X^{(t-1)})
\end{aligned}
$$

where $p_t(y,z) = T(y \mid z)p_{t-1}(z)$, is the joint distribution of $t-1$ and tth draws from a chain.

For reasonable density functions, the variance of $\log(U)$ is well bounded, which is a direct consequence of the fact that

$$\int \log^2\{f(x)\}f(x)dx < \infty$$

for those densities with tails not fatter than a Cauchy density. Typically, since $E(U)$ exists as we have shown, variance of $\log(U)$ is bounded. Our second example in Section 3 indicates the usefulness of this transformed statistic. Another approach might be the assessment of the convergence of the estimated $F^{(t)}(u)$ through Smirnov test statistic (Smirnov 1939) or by looking at the sequential plots of the sample c.d.f. (cumulative distribution function). Another way is to take logarithmic transformation of the variable U so as to result a more stable variance. A diagnosing procedure for assessing the convergence based on above facts is proposed as the following three steps.

First Calculate control samples $\{U^{(i,j,t)} : i \neq j,\ i,j = 1,\ldots,m\}$ or its subset at each iteration $t = 1,2,\ldots,2n$;

Second Plot the empirical distributions, i.e., sample cumulative distribution function of the variational control variable sequentially based on samples from from later half of the iterations, $t = n+1, n+2, \ldots, 2n$.

Third Use the idea of Gelman and Rubin (1991b) on the control variable U or its transformation to judge its distributional convergence. The simplest way is to compare the within and between variances of U in $m/2$ independent groups of paired series. More efficient methods use all $m(m-1)$ values of U and take in account the dependence structure among the U's that share sequences.

If there are any signs in above steps indicating the non-stationarity of the variational control variable, the Gibbs sampling iterations should not be terminated. It is also noted that in steps 2, samples from different iterations can be aggregated so as to improve the power of detecting although we lose the independence between samples by doing so.

3. Example: An Ising Model

As an illustration of our diagnosing procedure of the Gibbs sampler, a three dimensional $(N \times N \times N)$ Ising model for $N = 10$ is investigated. The model is used to describe a system of particles with each occupying one grid point of a $10 \times 10 \times 10$ cube. Let u or v denote the vertices of the 3-dim lattice, having form $u = (i,j,k), i,j,k, = 1,\ldots,N$. A random variable x_u, taking values either $+1$ or -1, is used to describe the spin of the particle located at position u in the cube. To be a bit more realistic than assuming independence between X's, a nearest-neighbor dependence structure proposed by the physicist Ising is employed. The full joint distribution of $\mathbf{X} = \{x_u;\ u \text{ are vertices}\}$ for the model can be written as

$$\pi(\mathbf{X} \mid \beta) \propto \exp\{\beta \sum_{\text{neighbors}} x_u x_v\} \tag{6}$$

The full conditional distribution of x_u is rather simple and neglected here. In the model β, taken as 0.25 in this investigation, is usually referred as a "temperature" parameter in statistical physics.

The value of the nearest-neighbor correlation function defined as

$$\rho(\beta) = E\{\frac{1}{3N^2(N-1)} \sum_{\text{neighbors}} x_u x_v\} \tag{7}$$

is of special physical interests. However, since no simple analytical function exists for our understanding, a simulation study is desirable. The Gibbs sampler can be employed because of the simple nearest-neighbor dependence structure of the field.

Our Gibbs sampling scheme starts with $m = 40$ parallel chains in which the first 20 series start from the distribution (6) with $\beta = 0$, that is, $x_{i,j,k}$'s are independently drawn from Bernoulli(0.5); and the rest 20 start with all $x_{i,j,k} = 1$. At each iteration t of the Gibbs sampler, the sample means of ρ in (7) based on the first 20 series and the last 20 are computed respectively together with their estimated variances. Then two series of means of the parameter of interests with estimated variances, denoted by $\{\bar{\rho}_i^{(t)},\ s_i^{(t)} : t = n+1,\ldots,2n\}$ for $i = 1,2$, are obtained, and the mean series are plotted in Fig. 1. Here the latter half of the simulated sequences are used for inference. An adapted reduction coefficient is produced for this "two sample problem" with estimated variances for each data point $\bar{\rho}_i^{(t)}$ in the same spirit of Gelman and Rubin (1991):

$$\hat{R}_{2n} = [\frac{4n-2}{4n-1} + \frac{3B}{(4n-1)W}]^{1/2} \tag{8}$$

where $W = (1/2n)\sum[s_1^{(t)} + s_2^{(t)} + (n-1)(s_1^2 + s_2^2))$ with $4n-2$ degrees of freedom, s_1^2 and s_2^2 are the within sample variances of $\{\bar{\rho}_1^{(t)} : t = n+1, \ldots, 2n\}$ and $\{\bar{\rho}_2^{(t)} : t = n+1, \ldots, 2n\}$, respectively; $B/n = (\bar{\bar{\rho}}_1 - \bar{\bar{\rho}}_2)^2/2$ with 1 degree of freedom, $\bar{\bar{\rho}}_1$ and $\bar{\bar{\rho}}_2$ are the sample means of $\{\bar{\rho}_1^{(t)} : t = n+1, \ldots, 2n\}$ and $\{\bar{\rho}_2^{(t)} : t = n+1, \ldots, 2n\}$, respectively. A fact motivating the construction of such a "coefficient", under stationarity assumption, is that

$$E(\frac{(n-1)(s_1^2 + s_2^2)}{2n} + \frac{B}{n}) = \mathrm{var}(\bar{\rho}_i) = \frac{\mathrm{var}(\rho)}{400}.$$

Thus $\hat{R}_{2n}$ is typically larger than one when the current distribution is "overdispersed" relative to the target distribution, and will converge to one when the number of iteration increases, and thus provides an analytic indicator for convergence. Specially, $\frac{B}{2n}$ is used to take care of the variation of the estimated means. The interested readers are referred to Gelman and Rubin (1991) for detailed properties of this statistic. The calculated $\hat{R}$ for this example is shown in Fig. 3, which indicates that about 100 iterations are sufficient for doing inference on the nearest-neighbor correlation parameter ρ. This finding is in accordance with the plot of ρ in Fig. 1. However, are we sure that the whole distribution for $\mathbf{X}$ is enough close to the target distribution?

An analysis on the control variable U, more specially the mean and variance of $\log(U)$, reveals a surprising fact as being demonstrated in Fig. 2 and 3: at least 1000 iterations are needed to bring the joint distribution of $\{x_u, \text{all } u\}$ reasonably close to the target one. Our proposed method starts with computations of two groups of $\log(U)$'s for each iteration t, using the first 20 series and the last 20 series respectively so that each group is consists of 20×19 samples of $\log(U)$. Then two sample means of $\log(U)$ at iteration t for both groups are then calculated with their variances estimated by an estimator

$$\max\{\frac{s_U^{(t)}}{(m/2)(m/2-1)}[1 + (m/2-2)(\hat{r}_I + \hat{r}_{II})], 0\},$$

where $s_U^{(t)}$ is the sample variance in respective group; $\hat{r}_I$ is the sample estimate of correlation between $\log(U^{(i,j,t)})$ and $\log(U^{(i,k,t)})$, while $\hat{r}_{II}$ the estimate of that between $\log(U^{(i,j,t)})$ and $\log(U^{(k,j,t)})$) for $i \neq j \neq k$ within each group. This estimator for variances is valid because within each group there are only two types of dependence between $\log(U)$'s as being estimated by $\hat{r}_I$ and $\hat{r}_{II}$. For different i, j, k, l, $U^{(i,j,t)}$ is typically independent of $U^{(k,l,t)}$. Thus similar to the case of estimating nearest-neighbor correlation ρ, we obtain two independent series of sample means of $\log(U)$ from two groups at each iteration t with estimated variances. Two series of such sample means are plotted in Fig. 2, and the corresponding reduction coefficient $\hat{R}$ calculated by (8) is shown in Fig 3. Both plots strikingly demonstrate that substantially more iterations are needed to attain overall convergence of the joint distribution than those needed for inference on ρ.

Figure 1: The series of the sample means of the nearest-neighbor correlation coefficient ρ. The solid line is the sample means of the first 20 series drawn by the Gibbs sampler; the dotted line is the sample means of the last 20 series.

4. Discussion

To be more conservative in judging the convergence of the distribution of the control variable U, a simple Smirnov statistic measuring the difference between $F^{(n)}(u)$ and $F^{(n')}(u)$ can be constructed by separating the m parallel simulation series into 2 groups with equal size. Thus at time $t = n$ and $t = n'$ we have estimated sample c.d.f. $F_1^{(n)}(u)$, $F_2^{(n)}(u)$, $F_1^{(n')}(u)$ and $F_2^{(n')}(u)$ using each group respectively. $F_1^{(n)}(u)$ and $F_2^{(n)}(u)$ are identically distributed, so

Figure 2: The solid line is the sample means of all the $\log(U)$'s created by the first 20 series drawn by the Gibbs sampler; the dotted line is the sample means of those created by the last 20 series.

Figure 3: The reduction coefficients for ρ (the thick line) and for $\log(U)$'s (the thin line).

are $F_1^{(n')}(u)$ and $F_2^{(n')}(u)$. $F_1^{(n)}(u)$ and $F_2^{(n')}(u)$ are independent, so are $F_1^{(n')}(u)$ and $F_2^{(n)}(u)$. The two Smirnov statistics

$$D_1(n, n') = \max \mid F_1^{(n)}(u) - F_2^{(n')}(u) \mid$$

and

$$D_2(n, n') = \max \mid F_2^{(n)}(u) - F_1^{(n')}(u) \mid$$

have the same distribution and when $F^{(n)}(u) = F^{(n')}(u) = F(u)$, the asymptotic distribution of $D_1(n, n')$ (or $D_2(n, n')$) can be found, for example, in Lehmann (1975). However one may lose power and is less efficient by doing so. Where possible the graph of $F_n(x)$ (the sample c.d.f.) should be plotted and inspected since this will reveal in a graphic way the nature of any substantial departure from the hypothesis (in our case, $F^{(n)}(u) = F^{(n')}(u)$).

It is not entirely clear how the quantitative difference between $p_n(X)$ and $\pi(X)$ is related to the distributional variation of $F^{(n)}(u)$. The sample mean of $U^{(i,j,n)}$ provides a quantitative measurement of the difference between $p_n(x)$ and $\pi(x)$. This suggests that examining the empirical distribution or the transformation of U are right ways to go. For example, one may wish to do logarithm transformation to the control variable U so as to make it more stable, and examine the distributional convergence of $\log(U)$. Whereas, the nice expectation formula (4) of U is substituted by the mutual information (5) for $\log(U)$. In this situation, a more sophisticated understanding is required.

Acknowledgement: We thank Professor Andrew Gelman at the University of California, Berkeley, for kindly reviewing our paper, which especially helped to clarify the conditions needed for (4).

References

[1] Besag, J. (1974), Spatial Interaction and the Statistical Analysis of Lattice systems (with discussion). *J. R. Statist. Soc.* **B, 36,** 192–236.

[2] Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith A. F. M. (1990), Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journal of the American Statistical Association,* **85**, 972–985.

[3] Gelfand A. E. and Smith A. F. M. (1990), Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association,* **85,** 398–409.

[4] Gelman A. and Rubin D. B. (1991a), A Single Series from the Gibbs Sampler Provides a False Sense of Security.

[5] Gelman A. and Rubin D. B. (1991b), Honest Inferences from Iterative Simulation. *Technical report, Dept. of Stat., Harvard University.*

[6] Gemam S. and Geman D. (1984), Stochastic Relaxation, Gibbs Distributions, and The Bayesian Restoration of Images. *IEEE Transactions on Pattern analysis and Machine Intelligence,* **6**, 721–741.

[7] Lehmann, E. L. (1975), Nonparametrics. *Hilden—Day Inc. McGraw—Hill International Book Company.*

[8] Liu, J. and Wong, W. H. and Kong, A. (1991), Correlation Structure and Convergence Rate of the Gibbs Sampler: Applications to the Comparisons of Estimators and Augmentation Schemes. *Submitted to Biometrika.*

[9] Schervich, M. and Carlin, B. (1991), On the Convergence of Successive Substitution Sampling. *Technical report, Dept. of Stat., Carnegie Mellon University.*

[10] Smirnov, N. V. (1939), On the Estimation of the Discrepancy between Empirical Curves of Distribution for two Independent Samples. *Bull. Univ. Moscow* **2**(2), 3–14.

[11] Tanner M. A. and Wong W. H. (1987), The Calculation of Posterior Distribution By Data Augmentation (with discussion). *Journal of the American Statistical Association,* **82**, 528–550.