Medical Infobahn for Europe
A. Hasman et al. (Eds.)
IOS Press, 2000

Convergence Diagnosis for Gibbs Sampling Output

Ulrich Mansmann

Department of Medical Biometry, University of Heidelberg, Heidelberg, Germany

Abstract: Gibbs sampling is a technique to calculate a complex posterior distribution as steady state measure of a Markov chain. The fundamental problem of inference from Markov chain simulation is that there will always be areas of the target distribution that have not been covered by the finite chain. Deciding when to stop the chain in order to have reached enough coverage of the support of the target distribution is an important matter. Techniques based on one long single chain and on multiple chains are discussed in the framework of a linear mixed effects model. The diagnostics used do not provide a consistent view on the convergence. Practical consequences on the estimates are shown.

1. Introduction: Markov chain Monte Carlo (MCMC) methods are powerful tools and easy to apply. Thus, there are risks of serious errors: *inappropriate modelling* (the assumed model is not realistic from a substantive point of view or does not fit the data), *errors in calculation or programming* (the simulated stationary distribution is not the desired target distribution), and *slow or no convergence* (the chain remains for many iterations in a region influenced by the starting values, or the chain moves between two equilibrium states). This article discusses possible measures against the third kind of error considering a linear mixed effects model fitted to longitudinal data from a study on *the antiatherosclerotic effect of Allium sativum* [1].

Slow convergence is a problem with deterministic algorithms as well. For example, a sound use of the EM algorithm requires a careful check of convergence. Consequences of a naive convergence consideration for the EM algorithm in the context of finite normal mixtures are discussed in [2]. In MCMC, the practical task in monitoring convergence is to estimate how much the inference based on the Markov chain differs from the desired target distribution. This is related to the problem of detecting if a Markov chain has forgotten its starting point.

Techniques for a *single long chain* based on spectral methods [3], and on eigenvalue bounds [4] as well as a technique for *multiple chains* based on variance analytic methods [5] will be discussed. An excellent review on the theoretical backgrounds of convergence assessment is given in [6]. Furthermore, it is of interest to study autocorrelations for each variable in each chain. High autocorrelations within chains indicate slow mixing and usually slow convergence.

2. The model: The model considered is given as a *directed graph*. The graph allows a formal language to represent the full joint distribution of all quantities considered as a simple factorisation in terms of conditional distributions. Figure 1 shows the model graph for the problem under consideration. For more information on the model see [7].

The posterior distribution of interest is of high dimension. Generally, the monitoring of convergence is performed on one-dimensional margins of the full distribution. Methods for the full-dimensional measure do exist which are computationally expensive . Generally, it is recommended to monitor all fixed effect nodes (founder nodes with no parents: b1, b2, alpha0, sigma1, tau). It may also be of interest to monitor internal nodes. To shorten the discussion, the monitoring of convergence will be restricted to the key parameter *b2* which quantifies treatment-time interaction and parameter *sigma1* which describes the random

effect. Because, in hierachical models, it is important to look at a plot of the random effect variance. If the starting value of this parameter is too close to zero, the Gibbs sampler can get stuck for a long time close to the starting values.

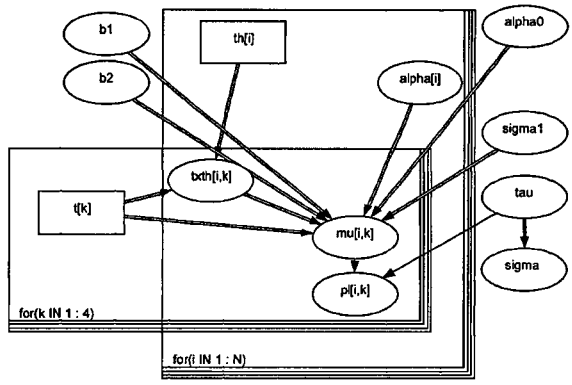

Figure 1: Model graph

3. Convergence analysis based on a single chain: The chain will be started with an extreme choice of starting values: list(alpha0 = 0, b1=-400, b2=1,000, tau = 1,000, sigma1 = 0.1). This strategy is used to get an impression of how influential starting points may be in the future history of the chain. Figure 2 shows the first 100 iterations of sampler for *sigma1* and *b2*. The chain seem to reach a stable state after 30 iterations. One may conclude on heuristic grounds that after 1,000 iteration the chain has forgotten its initial values and has reached the target distribution and additional 1,000 iterations may be sufficient to perform the estimations. This will now be inspected more systematically by three tools of convergence analysis for a single chain: The *Gewecke*, the *Heidelberger-Welch*, and the *Raftery-Lewis* diagnostic [6].

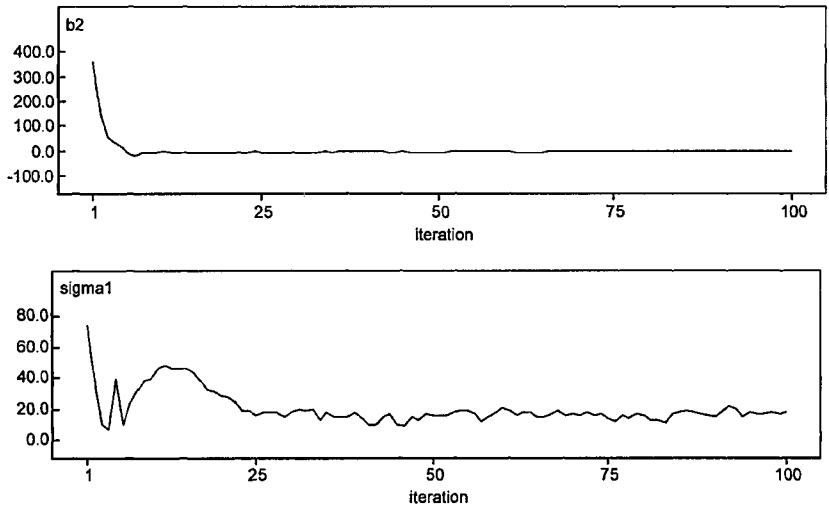

Figure 2: Behaviour of the Gibbs sampler starting from extreme points

The *Gewecke* diagnostic suggests that if a chain has converged by iteration n_0, then one should accept a *test of equal location* for two subsequences $\{x_i, i = n_0, ..., n_A\}$ and $\{x_i, i = n_B, ..., n_C\}$ with $n_0 < n_A << n_B < n_C$. The variability of the difference of the mean values of both subsequences is estimates form a consistent spectral density estimate. Under the null hypothesis of convergence the test statistic is standard normal distributed.

The *Heidelberger-Welch* diagnostic uses the assumption that in the steady state, the process observed is a weakly (covariance) stationary process. A new process is constructed from the standardized fluctuations of cumulative sums of the observed values around the mean value over the periode considered. The null hypothesis of stationarity implies that over a sufficient large period this process forms a Brownian Bridge. The *Cramer-von-Mises* statistic is used to judge convergence.

The *Raftery-Lewis* diagnostic quantifies how good a quantile of the target distribution can be estimated. A binary process is introduced (observed values below or above the quantile) which will well approximate a Markov chain when an appropriate subsequence is chosen. Once the subsequence is established, its approximate transition matrix can be estimated. The eigen-values of this matrix are easily calculated and used to estimate the geometric convergence rate of the chain which gives information on how long the chain has to run to approximate the target distribution with sufficient precision.

The z-score calculated from the *Gewecke* diagnostic (based on iterations 1,001 to 2,000) for parameter *b2* is 3.48 (*sigma1*: 2.83). The null hypothesis of stationarity is rejected for both parameters. The *Heidelberger-Welch* diagnostic accepts the null hypothesis of stationarity for *b2* as well as *sigma1*. The *Raftery-Lewis* diagnostic can not be applied because the 1000 iterations under study a not sufficient to perform the calculation.

Therefore, the chain is expanded to 15,000 iterations. Iteration 10,001 to 15,000 will be used for a second check. The z-score calculated from the *Gewecke* diagnostic for is 0.608 (*sigma1*: -0.897). The *Heidelberger-Welch* diagnostic also accepts the null hypothesis of stationarity for *b2* and *sigma1*. The *Raftery-Lewis* diagnostic states that for *b2* additional 3,120 iterations (for *sigma1* 5,504 additional iterations) to the 5,000 already used should be performed in order to estimate the 2.5^{th} percentile of the posterior distribution to a specified accuracy. Based on this result the sampler is run for additional 20.000 iterations and estimates of the parameters are based on the iterations 15,001 to 35,000. This time the *Raftery-Lewis* diagnostic calculates that 9,388 iterations are needed for *sigma1* and 11,332 for *b2*. With 15,000 iterations performed these requirements are met.

The *autocorrelation* analysis for *b2* gives 0.69 (Lag1), 0.17 (Lag 5), 0.023 (Lag 10) and 0.00021 (Lag 50). For *sigma1* one finds 0.57 (Lag1), 0.093 (Lag 5), 0.003 (Lag 10) and 0.015 (Lag 50). It may be necessary to increase the thinning interval to say , every 10^{th} or 50^{th} iteration, before calculating summary statistics and density estimates, in order to achieve a less highly correlated sample.

4. Convergence analysis based on multiple chains: The method of Gelman and Rubin [5] consists of analysing *m* independent sequences to form a distributional estimate of what is known about some random variable, given the observations simulated so far. It provides a basis for an estimate of how close the process is to convergence and, in particular, how much more simulations might improve the estimates. The method uses $m \geq 2$ independently simulated sequences of length *2n*, each beginning at different starting points which are over-dispersed with respect to the stationary distribution. The first n iterations are discarded. First, for any scalar functional of interest, the variance between the m sequence means are calculated. Second, the mean of the m within-sequence variances is determined. Now, rather than testing the null hypothesis that the algorithm has converged, a factor R is

given by which the estimated scale of the posterior distribution will shrink. An alternative method is proposed by Brooks and Gelman [8], which generalises the original method to consider more than one parameter.

The result using 4 chains which are over-dispersed with respect to the true posterior distribution shows that the factor R for parameter *sigma1* (as well as *b2*) reaches the value 1 after a little bit more than 400 iterations.

5. Consequences for the estimation of parameters: Table 1 summarises estimates of the SD of the mean estimates of *b2* and *sigma1* using different parts of the chain as well as thinning the chain.

Table 1: Influence of sampling strategies on the distribution of parameters

	b2			*sigma1*		
Iteration	Lag 1	Lag 10	Lag 50	Lag 1	Lag 10	Lag 50
1,001-2,000	1.88	1.82	2.19	2.57	2.66	3.09
10,001-15,000	1.73	1.70	1.66	2.76	2.81	3.01
15,001-30,000	1.73	1.78	1.75	2.69	2.61	2.60

This short view on the distribution of the parameters *b2* and *sigma1* shows that inference based on iterations 1,001-2,000 may produce a too narrow distribution. This may also happen for *sigma1* when estimation is based on iterations 10,001-15,000. The last line of Table 1 shows that iterations 15,001-30,000 can be used to gain a reliable result on the distributions of both parameters.

6. Discussion: The diagnostics used do not give a consistent view on the convergence of the sampler under study. The *Gelman-Rubin* diagnostic (which is offered in the WinBUGS software) assesses convergence after a short run in period (500 iterations). This and the *Heidelberger-Welch* method offer the most liberal view on the problem. A result which was not supported by the methods of *Raftery - Lewis* and *Geweke.*

As all statistical procedures, any convergence diagnostic technique cannot be guaranteed to successfully diagnose convergence. Each of the diagnostic methods has its own drawbacks and there is no one globally *best* method which works for all problems. The diagnostic methods discussed look at convergence of the chain itself, and do not try to provide a model for convergence. Notable exception of this is the method of *Raftery- Lewis.*

Because of the computational expenses, marginal and not full-dimensional diagnostics are offered in basic software like CODA [9]. The more rigorous and reliable methods, tend to be the most computationally expensive, and so the least well used. Whilst the existence of popular code is extremely useful, these programs are open to naive misuse and mis-interpretation. It is essential to have some understanding of the basic concepts behind any diagnostic that one chooses to use, and to understand its limitations in terms of the types of problems for which it may be applied.

References

[1] Koscielny J, Klüßendorf D, Latza R, Schmitt R, Radtke H, Siegel G, Kiesewetter H (1999) The antiatherosclerotic effect of Allium Sativum, Atherosclerosis, 144, 237-249.

[2] Friede T, Kieser M, Properties of an EM algortihm based procedure for blinded variance estimation (2000) XXth International Biometric Conference, Berkeley, California, July 2-7, 2000.

[3] Geyer CJ (1994) Practical Markov Chain Monte Carlo, Statistical Science, 7, 473-511

[4] Raftery AE, Lewis SM (1992) How many Iterations in the Gibbs sampler? In: Bernardo JM et al. (Eds) Bayesian Statistics 4.

[5] Gelman A, Rubin DB (1994) Inference from Iterative Simulations using Multiple Sequences, Statistical Science, 7, 457-511

[6] Brooks SP, Roberts GO (1996) Assessing Convergence of Markov Chain Monte Carlo Algorithms, Technical Report, Bristol

[7] Schuster E (2000) Modellwahl bei longitudinalen Daten – am Beispiel einer Studie zur Langzeitwirkung von Knoblauch auf Fettablagerungen in Arterien, MIE/GMDS2000, Hannover, 27.08-01.09.2000

[8] Brooks SP, Gelman A (1997) Alternative Methods for Monitoring Convergence of Iterative Simulations, Journal of Computational and Graphical Statistics, 6, 35-67

[9] Best N, Cowles K (1995) CODA Cambridge, MRC Biostatistics Unit