General Methods for Monitoring Convergence of Iterative Simulations

Stephen Brooks and Andrew Gelman[†]

Abstract

We generalise the method proposed by Gelman and Rubin (1992a) for monitoring the convergence of iterative simulations by comparing between and within variances of multiple chains, in order to obtain a family of tests for convergence. We review methods of inference from simulations in order to develop convergence-monitoring summaries that are relevant for the purposes for which the simulations are used. We recommend applying a battery of tests for mixing based on the comparison of inferences from individual sequences and from the mixture of sequences. Finally, we discuss multivariate analogues, for assessing convergence of several parameters simultaneously.

Keywords: Markov chain Monte Carlo, Convergence Diagnosis, Inference

[†]Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, England and Department of Statistics, Columbia University, New York, NY 10027, USA. We thank Gareth Roberts, Xiao-Li Meng, Phillip Price, three anonymous referees and an Associate Editor for helpful comments. We also thank the U.S. National Science Foundation for grant DMS-9404305 and Young Investigator Award DMS-9457824, the U.K. Engineering and Physical Sciences Research Council, and the Research Council of Katholieke Universiteit Leuven for fellowship F/96/9.

1 Introduction and Background

Markov chain Monte Carlo (MCMC) algorithms have made a significant impact on the range of problems to which Bayesian analyses can be applied; see Gilks *et al* (1996). The method involves simulating from a complex and generally multivariate target distribution, $p(\theta)$, indirectly, by generating a Markov chain with the target density as its stationary density. Generally, we run $m \geq 1$ sequences of simulations, each of length n, $(\theta_{j1}, \theta_{j2}, \ldots, \theta_{jn})$, for $j = 1, \ldots, m$. (In this notation, each θ_{jt} is a vector.) If $m > 1$, the m sequences are usually, but not always, simulated independently.

One might like to analytically compute or estimate a convergence rate and then take sufficient iterations for any particular desired accuracy but this is not possible in general (Tierney, 1994). It is possible to find bounds on the rates of convergence for special classes of Markov chains (e.g., Rosenthal, 1995), but in practice it is difficult to apply these results effectively in the context of MCMC.

In the absence of any general techniques for *a priori* prediction of run lengths, it is necessary to carry out some form of statistical analysis in order to assess convergence. These procedures, which are called convergence diagnostics, fall into two general categories: those based solely on the output values (θ_{jt}) of the simulation, and those that also use additional information about the target density; the methods described in this paper are all of the first category. See Brooks and Roberts (1996) and Cowles and Carlin (1996) for reviews of commonly used techniques of convergence assessment.

Gelman and Rubin (1992a,b) point out that, in many problems, lack of convergence can be easily determined from multiple independent sequences but cannot be diagnosed using simulation output from any single sequence. They propose a method using multiple replications of the chain to decide whether or not stationarity has been achieved within the second half of each of the sample paths. The idea behind this is an implicit assumption that convergence will have been achieved within the first half of the sample paths, and the validity of this assumption is essentially being tested by the diagnostic. This method's popularity may be largely due to its implementational simplicity and the fact that generic code is widely available to implement the method. In addition, even when this method is not formally used, something like it is often done implicitly or informally.

For example, Green *et al* (1994) write, in implementing a Gibbs sampler, "In the present study we used ... one long chain. Experimentation with different starting values convinced us that the chain was converging and covering the entire posterior distribution." As is often the case in statistics, it generally useful to study the formal procedures that lie behind an informal method, in this case for monitoring the mixing of multiple sequences.

In this paper, we generalise the method of Gelman and Rubin (1992a) by (1) adding graphical methods for tracking the approach to convergence, (2) generalising the scale reduction factor to track measures of scale other than the variance, and (3) extending to multivariate summaries. Before presenting these extensions, we summarise the motivation underlying and the details of the original method.

1.1 Inference given Convergence

Operationally, effective convergence of Markov chain simulation has been reached when inferences for quantities of interest do not depend on the starting point of the simulations. This suggests monitoring convergence by comparing inferences made from several independently-sampled sequences with different starting points. Before considering methods of *comparing* inferences, we briefly discuss the standard method for *constructing* inferences under the assumption that convergence has been approximately reached. It is standard practice to discard observations within an initial transient phase. Most methods for inference are then based on the assumption that what remains can be treated as if the starting points had been drawn from the target distribution (for an exception, see Liu and Rubin, 1996).

Inference for random variables (e.g., posterior quantities in a Bayesian analysis), for which we use the notation $\psi = \psi(\theta)$, is straightforward: a $(1-\alpha)$ central posterior interval for a scalar random variable ψ under the target distribution is approximated by the $\alpha/2$ and $1-\alpha/2$ quantiles of the empirical simulation results, ψ_{it}, from all sequences together (if $m > 1$) and discarding burn-in iterations. This is, of course, also the standard procedure for obtaining intervals from direct (non-iterative) simulation, and is generally recommended no matter how convergence is monitored; for example, Gelman and Rubin (1992a) use an approximate Student-t posterior distribution to estimate convergence but then recommend using the empirical intervals once approximate convergence has been reached. When additional information about the target distribution is used, inference can be made much more precise, as in the "Rao-Blackwellization" procedure of Gelfand and Smith (1990), which uses the analytic form of the conditional target distribution of ψ given the rest of θ (see also Tanner and Wong, 1987). However, even these procedures are ultimately applied by averaging over the iterations of all the simulated sequences, after discarding burn-in.

This suggests that "convergence" can be quantified in terms of the properties of the empirical interval, as compared to the true 95% interval from the target distribution, which would be attained in the limit as $n \to \infty$.

1.2 Monitoring Convergence: The Original Method

As background, we present the method of Gelman and Rubin (1992a) using our general perspective of comparison of inferences. The method presupposes that m chains have been simulated in parallel, each with different starting points which are overdispersed with respect to the target distribution. A number of methods have been proposed for generating initial values for MCMC samplers. Gelman and Rubin (1992a) propose using a simple mode-finding algorithm to locate regions of high density and sampling from a mixture of t-distributions located at these modes to generate suitable starting values. An alternative approach is presented by Applegate *et al* (1990) who use the method of simulated annealing in this same context, see also Jennison (1993). Having obtained suitable starting points, the chains are then run for $2n$ iterations, of which the first n are discarded to avoid the burn-in period.

Given any individual sequence, and if approximate convergence has been reached, an assumption is made that inferences about any quantity of interest is made by computing the sample mean and variance from the simulated draws. Thus, the m chains yield m possible inferences; to answer the question of whether these inferences are similar enough to indicate approximate convergence, Gelman and Rubin (1992a) suggest comparing these to the inference made by mixing together the mn draws from all the sequences. Consider a scalar summary—that is, a random variable—ψ, that has mean μ and variance σ^2 under the target distribution, and suppose that we have some unbiased estimator $\widehat{\mu}$ for μ. Letting ψ_{jt} denote the t^{th} of the n iterations of ψ in chain j, we take $\widehat{\mu} = \bar{\psi}_{..}$, and calculate the between-sequence variance B/n, and the within-sequence variance W, defined by

$$\begin{aligned} B/n &= \frac{1}{m-1}\sum_{j=1}^{m}(\bar{\psi}_{j.} - \bar{\psi}_{..})^2 \\ W &= \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{t=1}^{n}(\psi_{jt} - \bar{\psi}_{j.})^2. \end{aligned}$$

Note that the ANOVA assumption of pooled within variances can be made here because, under convergence, the m within-chain variances will indeed be equal.

Having observed these estimates, we can estimate σ^2 by a weighted average of B and W,

$$\hat{\sigma}^2_+ = \frac{n-1}{n}W + \frac{B}{n},$$

which would be an unbiased estimate of the true variance σ^2, if the starting points of the sequences were drawn from the target distribution, but overestimates σ^2 if the starting distribution is appropriately overdispersed. (As we shall see in Section 2.1, if over-dispersion does not hold, $\hat{\sigma}^2_+$ can be too low, which can lead to falsely diagnosing convergence.) Accounting for the sampling variability of the estimator $\hat{\mu}$ yields a pooled posterior variance estimate of $\widehat{V} = \hat{\sigma}^2_+ + B/(mn)$.

The comparison of pooled and within-chain inferences is expressed as a variance ratio,

$$R = \frac{\widehat{V}}{\sigma^2},$$

which is called the scale reduction factor, or SRF. (Strictly speaking, the term "scale reduction factor" applies to $\sqrt{R}$; for convenience, we will mostly work with variance reduction factors in this paper.) Because the denominator of R is not itself known, it must be estimated from the data; we can gain an overestimate of R by (under)estimating σ^2 by W. Thus, we (over)estimate R by

$$\widehat{R} = \frac{\widehat{V}}{W} = \frac{m+1}{m}\frac{\hat{\sigma}^2_+}{W} - \frac{n-1}{mn}, \tag{1}$$

which is called the *potential* scale reduction factor, or PSRF, and can be interpreted as a convergence diagnostic as follows. If $\widehat{R}$ is large, this suggests that either estimate of the variance $\hat{\sigma}^2$ can be further decreased by more simulations, or that further simulation will increase W, since the simulated sequences have not yet made a full tour of the target distribution. Alternatively, if the PSRF is close to 1, we can conclude that each of the m sets of n simulated observations is close to the target distribution.

From our perspective of monitoring convergence based on the behaviour of potential inferences, Gelman and Rubin's approach applies to all those situations where inferences are summarised by posterior means and variances, even if the posterior distributions are not themselves believed to be normally distributed. Other measures of convergence may be more appropriate if one is planning to use distributional summaries other than the first two moments.

1.3 Correcting for Sampling Variability in the Variance Estimates

One can refine the PSRF to account for sampling variability in the variance estimates, following the method of Fisher (1935) based on independent samples from the normal distribution. The key parameter for this adjustment is the estimated degrees of freedom, d, for a Student-t approximation to the posterior inference based upon the simulations. The degrees of freedom can be estimated by the method of moments: $d \approx 2\widehat{V}/\widehat{Var}(\widehat{V})$.

Gelman and Rubin (1992a) incorrectly adopt the correction factor $d/(d-2)$. This represents the ratios of the variances of the t_d and the normal distributions. The use of this incorrect factor has led to a number of problems, in that the corrected SRF (CSRF) can be infinite or even negative in the cases where convergence is so slow that $d < 2$. In order to correctly account for the sampling variability, the correction factor $(d+3)/(d+1)$ should be used. The correct factor, $(d+3)/(d+1)$, removes the problems associated with that proposed in the original paper.

We can motivate this correction factor as follows. Given an estimator $\widehat{\psi}$, with associated sampling variance s^2 estimated on d degrees of freedom, the position is not the same as if the variance were known exactly. Our estimate of the variance is itself subject to variability and allowance for this is made by adopting the t-distribution rather than the normal. Thus, it is incorrect to assume that the information supplied by $\widehat{\psi}$ is $1/s^2$, as though the estimate were known to follow a normal distribution with this variance. Instead, we need to account for not only the estimate s^2, but also the degrees of freedom upon which it is based. Thus, the correction factor $(d+3)/(d+1)$ comes from the evaluation of Fisher's information for the t_d distribution, correcting our estimator $\widehat{V}$ for the degrees of freedom on which it is based, see Fisher (1935).

Thus, we define

$$\widehat{R}_c = \frac{d+3}{d+1}\widehat{R} = \frac{d+3}{d+1}\frac{\widehat{V}}{W}.$$

This correction is minor, because at convergence d tends to be large.

2 An Iterated Graphical Approach

Gelman and Rubin (1992a) originally propose running the chains for $2n$ iterations and basing the calculation of $\widehat{R}_c$ upon the final n of these, for each parameter of interest. They suggest that if $\widehat{R}_c$ is close to 1 for each parameter, then the latter half-sequences can be said to have converged. However, if $\widehat{R}_c$ is much greater than 1 for any parameter of interest, this suggests that one should either continue the simulation or assume that there is some facet of the model which is causing slow convergence.

The method of monitoring convergence by examining a ratio, $\widehat{R}_c$, ignores some information in the simulations. In particular, at convergence, the following three conditions should hold:

1. The mixture-of-sequences variance, V, should stabilise as a function of n. (Before convergence, we expect $\hat{\sigma}^2_+$ to decrease with n, only increasing if the sequences explore a new area of parameter space, which would imply that the original sequences were not overdispersed for the particular scalar summary being monitored.)

2. The within-sequence variance, W, should stabilise as a function of n. (Before convergence, we expect W to be less than V.)

3. As previously discussed, $\widehat{R}_c$ should approach 1.

Monitoring $\widehat{R}_c$ alone considers only the third of these conditions.

An alternative, graphical approach to monitoring convergence is to divide the m sequences into batches of length b. We can then calculate $V(k)$, $W(k)$, and $\widehat{R}_c(k)$ based upon the latter half of the observations of a sequence of length $2kb$, for $k = 1, ..., n/b$, for some suitably large n. Note that we still choose to discard the first half of each sample, but argue that this choice is reasonable in terms of computational efficiency. If we discard a greater proportion of the samples then this would be wasteful, since the diagnostic would be based upon only a small fraction of the samples available. However, if we choose to discard less than half of each sample, then the sample of observations used at each iteration would change too little, and the $\widehat{R}$ value would remain high for too long, due to its continued dependence upon samples within the burn-in. Thus, the diagnostic would be wasteful in the sense that it would diagnose convergence later than necessary. Discarding half of the available sample at each iteration provides a compromise which attempts to maximise overall efficiency.

The iterative diagnostic procedure can be done either as observations are generated, or post-simulation as with the original method. In this way, we construct sequences that we can plot against k (or, equivalently $2kb$) in order to monitor the chains as they converge. In order to choose a suitable value of b, we note that a small value of b increases computational expense, whereas a large value provides little extra information in terms of convergence monitoring. In practice, experience suggests that taking $b \approx n/20$ provides a reasonable output, whilst minimising computational overheads.

In addition to the plot of $\widehat{R}_c(k)$ against k, another useful diagnostic is obtained by plotting the two scale factors, $V^{1/2}(k)$ and $W^{1/2}(k)$, as a function of k together on the same plot. (We use the $\frac{1}{2}$ power so that the factors are on a more directly interpretable scale.) Approximate convergence is not attained until both lines stabilise (conditions 1 and 2), and they stabilise at the same value (condition 3). We illustrate with an example.

2.1 Example: Weibull Regression in Censored Survival Analysis

Grieve (1987) provides data that measure photocarcinogenicity or survival times for four groups of mice subjected to different treatments. The survival times are assumed to follow a Weibull distribution, so that the likelihood is given by $\prod_i \rho e^{\beta' z_i} t_i^{\rho-1} \exp(-e^{\beta' z_i} t_i^{\rho})$, where t_i denotes the failure or censure time of an individual, $\rho > 0$ is the shape parameter of the Weibull distribution, β is a vector of unknown parameters and the z_i denote covariate vectors assigning each observation to one particular treatment group.

Following Dellaportas and Smith (1993), we assume vague normal prior distributions for the β_i parameters and a similarly vague gamma prior distribution for ρ, and we use the Gibbs sampler to fit the above model to Grieve's data. Given overdispersed starting points for five replications of the sampler, the iterated PSRF values for ρ are given in Figure 1a.

Figure 1 here

We can see that had we only run 1000 iterations of our sampler, we might easily have mis-diagnosed convergence, even with multiple sequences, having only calculated a single $\widehat{R}_c$ value. (In fact, after 1000 iterations, the $\widehat{R}_c$ values for the other parameters in the model—the four elements of β—are below 1.06 also.) Mis-diagnosing convergence may lead to both biased and falsely-precise inferences about the parameters. For example, based on 1000 iterations the 90% empirical interval for ρ of $[2.94, 4.34]$, has only 79.8% coverage probability in the target distribution (as estimated from iterates 2001–3000 of all five replications).

However, Figure 1 only checks the third of the three conditions necessary for convergence. We would not be misled in this way if we looked at the plots of $V^{1/2}$ and $W^{1/2}$, as displayed in Figure 1b. In the crucial period from iterations 800 to 1300, when $\widehat{R}_c$ goes rapidly below 1.1 (misleadingly suggesting convergence), neither V nor W has stabilised—they both appear to be increasing in this region. This suggests that neither of the first two convergence conditions have yet been satisfied. It is not until beyond around 2500 iterations that both plots begin to level out at a constant value, satisfying all three conditions above for convergence. The behaviour of V and W as a function of time suggests that the starting points were not sufficiently overdispersed. In this case, all five replications have explored only a single portion of the parameter space. At iteration 1300, one of the chains "discovers" the remaining portion, which is subsequently explored by the others by iteration 2500. By continuing the chain beyond 5000 iterations, we can be reasonably certain that convergence was achieved after about 2500 iterations.

Thus, the iterated, graphical method presents a qualitative element to the diagnostic, which gives the user a better "feel" for what is really happening with the chains, in a similar way to the cusum plot of Yu (1995). However, given a large number of parameters, the univariate graphical approach soon becomes impractical, which suggests that a numerical approach for checking the stability of V and W would be desirable.

3 General Univariate Comparisons

A limitation of the original convergence diagnostic is the assumption of normality of the marginal distribution of each scalar quantity, ψ. Normality is assumed explicitly when using the correction factor $(d+3)/(d+1)$ and, more importantly, implicitly when comparing the mixing of sequences by monitoring means and variances. The assumption of normality can be made more reasonable by appropriate transformations such as logs and logits, as Gelman and Rubin (1992a) suggest, but it is still unappealing, especially considering that MCMC methods are often used for highly non-normal, and even multimodal, densities. Here, we define a family of potential scale reduction factors, each of which has an interpretation as a measure of mixing and with the property that $\widehat{R} \to 1$ as convergence is approached, but which avoid the assumption of normality.

An alternative interpretation of the $\widehat{R}$ diagnostic is as a (squared) ratio of interval lengths, rather than as a variance ratio. This interpretation leads to an alternative implementation that can be used whenever sequences are long enough. Suppose we seek a $100(1-\alpha)\%$ interval for some parameter, ψ. Then we construct $\widehat{R}$, based upon the final n of the $2n$ iterations, as follows.

1. From each individual chain, take the empirical $100(1-\alpha)\%$ interval, i.e., the $100\frac{\alpha}{2}\%$ and

the $100(1-\frac{\alpha}{2})\%$ points of the n simulation draws. Thus, we form m within-sequence interval length estimates.

2. From the entire set of mn observations, gained from all chains, calculate the empirical $100(1-\alpha)\%$ interval, gaining a total-sequence interval length estimate.

3. Evaluate $\widehat{R}$ defined as

$$\widehat{R}_{\text{interval}} = \frac{\text{length of total-sequence interval}}{\text{mean length of the within-sequence intervals}}.$$

This method is considerably simpler than the original method. It is very easy to implement and does not even require second moments, let alone normality. $\widehat{R}_{\text{interval}}$ is still a PSRF, but based upon empirical interval lengths as a measure of information, rather than variance estimates. Like the normal-based diagnostic, $\widehat{R}_{\text{interval}}$ has the property of approaching 1 as the chains converge. Note that we do have the constraint that n be large enough to determine the empirical within-sequence intervals, but in practice, this does not generally seem to be a problem.

However, this is not the only possible interval-based approach. An alternative diagnostic can be defined based on coverage probability, as follows.

1. From each individual chain, take the end-points of the empirical $100(1-\alpha)\%$ interval to get a "within-chain" interval estimate.

2. For each of the within-chain intervals, estimate the empirical coverage probability of that interval with reference to the combination of all m sequences combined; i.e., calculate the proportion of observations from all m sequences combined, which lie within that interval.

3. Compute the average empirical coverage probability of the m intervals. At convergence, this will equal the nominal coverage probability, $1-\alpha$.

Non interval-based alternatives are also possible. For example, if higher order moments can be assumed to exist, and are of interest, one can replace the numerator and denominator in (1) by an empirical estimate of the central s^{th} order moments calculated from all sequences together and the mean s^{th} order moment calculated from each individual sequence, respectively. Thus, we define

$$\widehat{R}_s = \frac{\frac{1}{mn-1}\sum_{j=1}^{m}\sum_{t=1}^{n}|\psi_{jt}-\bar{\psi}_{..}|^s}{\frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{t=1}^{n}|\psi_{jt}-\bar{\psi}_{j.}|^s}, \tag{2}$$

for any value of s which seems reasonable within the context of the problem at hand. Corresponding graphical displays plot the numerator and denominator of $\widehat{R}_{\text{interval}}$, or $\widehat{R}_s$, together for sequences of length $2kb$, $k=1,\ldots,n/b$ as described in Section 2. These plots may then clarify the reasons for non-convergence.

The second-order measure, $\widehat{R}_2$, is clearly similar to R defined in Section 1.2; some algebraic manipulation yields

$$\widehat{R}_2 = \frac{(m-1)n}{mn-1}\frac{\hat{\sigma}_+^2}{W} + \frac{n-1}{mn-1} \tag{3}$$

(compare to (1)), and for moderately large n (so that $\frac{n-1}{n} \approx 1$),

$$\widehat{R}_2 \approx \frac{m-1}{m+1}\widehat{R} + \frac{2}{m+1}. \tag{4}$$

For example, if $m = 5$ and $\frac{\sigma_+^2}{W} = 1.2$, then $\widehat{R} = 1.24$ and $\widehat{R}_2 = 1.16$.

In defining these alternative measures, we ignore the correction factor $(d+3)/(d+1)$. This is appealing because we no longer assume normality but, of course, also means that we ignore some of the sampling variability in our estimate. In general, this correction factor has negligible effect upon the decision as to when convergence is achieved since firstly, if $\widehat{R}_c < 1.1$, say, then we should have enough observations so that its sampling variability is negligible; and secondly, in those cases where the effect is non-negligible, the chains are far from convergence and the ratio $\widehat{R} = \widehat{V}/W$ will be large anyway. However, if this is a concern, the correction factor $(d+3)/(d+1)$ can be computed (using the implicit normal model) and multiplied into the alternative measures; this is similar to the semi-parametric bootstrap idea of using a model to appropriately scale the variance of an empirical distribution (Efron and Tibshirani, 1993).

3.1 Example: Weibull Regression in Censored Survival Analysis (Revisited)

Here we use the above methods to assess the convergence of the shape parameter in the example of Section 2.1.

Figure 2 here

Figure 2 displays the numerator and denominator of the interval based convergence measure $\widehat{R}_{\text{interval}}$ and the alternative moments estimators $\widehat{R}_s$ for $s = 3$ and 4. Comparing to Figure 1, we can see that all of the moment-based methods behave similarly to $\widehat{R}_c$, with the numerator and denominator nearly converging around 1000 iterations and then subsequently rising again between 1500 and 2000 iterations. However, for the interval-based method, $\widehat{R}_{\text{interval}}$ decreases monotonically as the simulation proceeds, and this is a general characteristic of this particular measure. Thus, plotting the interval-based statistic, as with $\widehat{V}$ and W, may provide security against falsely diagnosing convergence too early, as discussed in Section 2.1.

3.2 Example: Truncated Cauchy Distribution

As a contrast, we apply the interval-based and normal-theory methods to the p-dimensional Cauchy distribution. We use a Metropolis algorithm with a Cauchy jumping kernel with scale of $2.4/\sqrt{p}$ (see Gelman *et al*, 1995), and with the target distribution truncated to lie within a sphere of radius $100\sqrt{p}$, to ensure eventual convergence. We performed a simulation study for the case $p = 5$, with $m = 10$ sequences, each of length $2n = 2000$ and, following Gelman *et al* (1995), starting points drawn from an overdispersed Cauchy distribution (the target distribution scaled by a factor of 4). We repeated the simulation 100 times, each time computing the potential scale reduction factors, $\widehat{R}_c^{1/2}$ and $\widehat{R}_{\text{interval}}$ (based on central 80% intervals). We also computed the *eventual* scale reduction factor: the length of the empirical central 80% interval from the last halves of the simulated sequences mixed together, divided by the true central 80% interval from the target distribution. For simplicity, we consider the convergence of just one parameter—the first element of the p-dimensional vector parameter.

In this example, which is far from normal, we were surprised to find that the normal-based convergence diagnostic performs much better than the interval-based measure. One way to see this is to classify the simulations by whether the potential scale reduction factor is greater or less than 1.2, which is sometimes used as a guideline for "approximate convergence." Of the 100 simulations, 54 had achieved $\widehat{R}_c^{1/2} < 1.2$ at 2000 iterations and 46 had not. Of the 54 cases with this apparent "convergence," $\widehat{R}_c^{1/2}$ ranged from 1.03 to 1.19, and the eventual scale reduction factors, based on the true target distribution, ranged from 0.6 to 1.6, with only 8 cases having eventual scale reduction factors greater than 1.2. Of the 46 simulations in which "convergence" had apparently not been reached, $\widehat{R}_c^{1/2}$ ranged from 1.2 to 2.1, and the eventual scale reduction factors ranged from 0.6 to 4.5, with 26 cases having eventual scale reduction factors greater than 1.2. This performance is not bad for the purposes of monitoring convergence—when $\widehat{R}_c^{1/2} < 1.2$, we can be pretty sure that convergence has been reached, but when $\widehat{R}_c^{1/2} > 1.2$, we can say little. (If we were to adopt the more stringent standard of $\widehat{R}_c^{1/2} < 1.1$, we would find that 21 out of the 100 simulations reached convergence, and, reassuringly, 20 of these had eventual scale reduction factors below 1.1.)

In contrast, $\widehat{R}_{\text{interval}}$ is often wildly optimistic in this example. All 100 of our simulations reported $\widehat{R}_{\text{interval}} < 1.2$, even though, as noted above, in 34 of the cases, the eventual scale reduction was greater than 1.2, with a high of 4.5. The corrected version, $[(d+3)/(d+1)]^{1/2}\widehat{R}_{\text{interval}}$, is not much better: 93 of the simulations yield a value below 1.2, and 28 of these have eventual scale reductions greater than 1.2, with a high of 3.1.

Why is this happening? This is an example in which a large number of simulation draws—even if obtained independently, rather than via MCMC—are needed to accurately estimate the mean, variance, and 80% interval of the target distribution. The 80% intervals from each individual sequence are themselves far more variable than from the mixture of all 10 sequences. The sampling distribution of interval widths is skewed in the positive direction, so that the denominator of $\widehat{R}_{\text{interval}}$ tends to be larger than its numerator. In fact, in 98 of the 100 simulations above, $\widehat{R}_{\text{interval}}$ was less than 1, and in 67 of the simulations, $((d+3)/(d+1))^{1/2}\widehat{R}_{\text{interval}}$ was less than 1. In general, this suggests that if the interval method is used, it is not necessarily appropriate to combine the widths of individual sequences by simple averaging.

Incidentally, the convergence diagnostic based on empirical coverage probabilities performs well in this example. In our 100 simulations, the average empirical coverage probabilities (ECP's) of the nominal 80% intervals ranged from 0.62 to 0.79, and they were highly predictive of the eventual (true) scale reduction factors: all of the simulations for which the ECP was less than 0.70 had eventual SRF's greater than 1.2; of the simulations for which the ECP was between 0.70 and 0.75, 58% had eventual SRF's greater than 1.2; and of the simulations for which the ECP was between 0.75 and 0.80 (i.e., quite close to the nominal coverage probability of 0.80), only 3% had eventual SRF's greater than 1.2.

Despite the problems with the interval-based PSRF, we emphasise that our methods have not "failed" in this example, because we would recommend using the alternative measures in addition to the traditional $\widehat{R}_c$ and not accepting convergence until all the measures are satisfied.

Thus, we might plot the original $\widehat{R}_c$ statistic, together with the corresponding numerator and denominator ($\widehat{V}$ and W) and the $\widehat{R}_{interval}$ diagnostic. The $\widehat{R}_c$ provides us with a scale for the diagnostic, which is interpretable in terms of the utility of continuing to run the chain, whereas the other plots, provide some reassurance against falsely diagnosing convergence too early, ie; the $\widehat{V}$ and W plots should both settle down and to the same location before convergence is diagnosed. The idea is to use a wide variety of diagnostics so that if all appear to suggest that convergence has been achieved, then the user can have some confidence in that conclusion. Having discussed these univariate extensions to the original method, we now move on to discuss an extension of the original method to the case where we are interested not in scalar functionals, but more general multivariate functionals of the chain.

4 Multivariate Extensions

There are a number of multivariate approaches available to us. When we are interested in the estimation of a vector parameter ψ based upon observations $\psi_{jt}^{(i)}$ denoting the ith element of the parameter vector in chain j at time t, the direct analogue of the univariate approach in higher dimensions is to estimate the posterior variance-covariance matrix by

$$\widehat{V} = \frac{n-1}{n} W + \left(1 + \frac{1}{m}\right) B/n,$$

where

$$W = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{t=1}^{n} (\psi_{jt} - \bar{\psi}_{j\cdot})(\psi_{jt} - \bar{\psi}_{j\cdot})'$$

and

$$B/n = \frac{1}{m-1} \sum_{j=1}^{m} (\bar{\psi}_{j\cdot} - \bar{\psi}_{\cdot\cdot})(\bar{\psi}_{j\cdot} - \bar{\psi}_{\cdot\cdot})',$$

denote the (p-dimensional) within and between-sequence covariance matrix estimates of the p-variate functional ψ, respectively. Thus, we can monitor both $\widehat{V}$ and W, determining convergence when any rotationally-invariant distance measure between the two matrices indicates that they are "sufficiently" close.

Before we go on to discuss scalar measure of distance between the $\widehat{V}$ and W matrices, we should point out that similar extensions can also be provided to the alternative diagnostics proposed in Section 3. In particular, the interval-based diagnostic in the univariate case becomes a comparison of volumes of total and within-chain convex hulls. Thus, for each chain we calculate the (hyper-)volume within 95%, say, of the points in the sample and compare the mean of these with the volume from 95% of the observations from all samples together. However, such a method is likely to be computationally expensive in even moderate dimensions, so we have not adopted this approach here.

4.1 Scalar Distance Measures

We would like to summarise the distance between $\widehat{V}$ and W with a scalar measure which should approach 1 (from above) as convergence is achieved, given suitably over-dispersed starting points. Many such measures can be constructed; we will work with the maximum root statistic, which is the maximum SRF of any linear projection of ψ, given by

$$\widehat{R}^p = \max_a \frac{a'\widehat{V}a}{a'Wa}, \tag{5}$$

which we call the *Multivariate PSRF* (or MPSRF) and which we can compute via the following lemmas.

Lemma 4.1 *For two non-singular, positive definite and symmetric matrices M and N,*

$$\max_a \frac{a'Ma}{a'Na} = \lambda,$$

where λ is the largest eigenvalue of the symmetric, positive definite matrix $N^{-1}M$.

Proof: See Mardia *et al* (1979), Section A.9.2.

Lemma 4.2

$$\widehat{R}^p = \frac{n-1}{n} + \left(\frac{m+1}{m}\right)\lambda_1,$$

where λ_1 is the largest eigenvalue of the symmetric, positive definite matrix $W^{-1}B/n$.

Proof:

$$\begin{aligned}
\widehat{R}^p &= \max_a \frac{a'\widehat{V}a}{a'Wa} \\
&= \max_a \frac{a'\left[\frac{n-1}{n}W + B\left(\frac{m+1}{mn}\right)\right]a}{a'Wa} \\
&= \frac{n-1}{n} + \left(\frac{m+1}{m}\right)\max_a \frac{a'Ba/n}{a'Wa} \\
&= \frac{n-1}{n} + \left(\frac{m+1}{m}\right)\lambda_1, \quad \text{by Lemma 4.1.}
\end{aligned}$$

□

Clearly, under the assumption of equal means between sequences, $\lambda_1 \to 0$ and so $\widehat{R}^p$ tends to 1 for reasonably large n.

Note that these measures are incalculable in the case where W is singular. If both W and B are singular, then this suggests that the underlying problem is ill-posed, for example two or more parameters may be very highly correlated. However, if only W is singular, then this suggests that the problem lies with the sampler, for example one or more variables have failed to move within the current set of iterations. Thus, in the case where $\widehat{R}^p$ is incalculable, due to the singularity of W, we still gain valuable information by examining the determinants of both the B and W matrices. Another instance in which the bound may be incalculable is if the matrix is too large to invert or decompose. This is essentially a constraint imposed by the hardware available and we have yet to try a model where this has been a problem.

In order to monitor the three conditions for convergence described in Section 2, we might also monitor plots of the determinants of, say, the pooled and within chain covariance matrices, which should stabilise as a function of n. Thus, the multidimensional approach involves monitoring one (or more) of the measures above, together with the determinants say, of the pooled and within chain covariance matrices.

4.2 Bounding the Multivariate Scale Reduction Factors

One reason for choosing $\widehat{R}^p$ among all the possible multivariate generalisations of the SRF is that it may be used as an approximate upper bound to the maximum of the univariate $\widehat{R}$ statistics over all p variables.

Lemma 4.3 *For $k = 1, \ldots, p$, let $\widehat{R}(k)$ be the univariate PSRF defined in (1), applied to the simulations of the k^{th} element of θ. If $\widehat{R}^{max}$ denotes the maximum of the $\widehat{R}(k)$ values (over $k = 1, ..., p$) then,*

$$\widehat{R}^{max} \leq \widehat{R}^p,$$

where $\widehat{R}^p$ is the MPSRF defined in (5), applied to the vector of parameters, θ.

Proof: Let I_k denote a vector of zeroes, with kth entry replaced by 1. Then

$$\begin{aligned}
\widehat{R}^p &= \max_a \frac{a'\widehat{V}a}{a'Wa} \\
&\geq \max_k \frac{I_k'\widehat{V}I_k}{I_k'WI_k} \\
&= \max_k \widehat{R}(k) \\
&= \widehat{R}_{max}
\end{aligned}$$

$\square$

Thus, the multivariate measure $\widehat{R}^p$ is an upper bound to the largest of the original measures over each of the elements of θ.

4.3 Examples

4.3.1 Example: Weibull Regression in Censored Survival Analysis (Revisited)

Here, we return to the example of Section 2.1 and examine the performance of the multivariate measures for the vector of five parameters corresponding to the shape parameter ρ, together with the 4-dimensional parameter vector, β.

Figure 3 here

Figure 3a plots the maximum of the individual PSRF's over the five parameters (for $m = 5$), together with the bound, $\widehat{R}^5$. Similarly, Figure 3b plots the determinants of the pooled and within-chain covariance matrices. Note the similarity in shape between the MPSRF and $\widehat{R}^{max}$ plots, and also with the univariate moment-based plots discussed earlier. Like the PSRF's, the MPSRF plot also possesses a second peak after about 2000 iterations.

In this case we can clearly see how the MPSRF bounds above the $\widehat{R}^{max}$ diagnostic, with $\widehat{R}^5 \approx 1.5 \times \widehat{R}^{max}$ within iterations 1–3000. Obviously, this bound improves as the simulations continue, since the MPSRF itself also converges to 1. Figure 3a illustrates that monitoring the MPSRF may well cause us to diagnose convergence later than the univariate PSRF plots. This difference is essentially a measure of the lack of convergence attributed to the interaction between the scalar functionals that we have chosen to study.

4.3.2 Example: Bivariate Normal Model with a Non-Identified Parameter

Let us consider an example in which we have observations y whose values depend upon two parameters θ and ϕ:

$$y_i \sim N(\theta_i + \phi_i, 1).$$

Clearly, θ and ϕ are not identified by the likelihood, but may be separated via their prior distribution, $p(\theta, \phi)$.

The conditionals are easy to compute, so we may use the Gibbs sampler in order to explore the posterior distribution. However, if the prior distribution is relatively diffuse, θ and ϕ will be highly (negatively) correlated in the posterior distribution, so that convergence may be very slow. In order to speed convergence, we take a transformation of variables from (θ, ϕ) to (θ, η), where $\eta_i = \theta_i + \phi_i$. Now, since the likelihood informs about η_i directly, overall sampler convergence is improved as discussed in Gelfand et al. (1995).

Since only the η_i are properly identified by the likelihood, and this is the main parameter of interest, it is common to use the PSRF to assess the convergence of the η_i parameters alone. However, Carlin and Louis (1996, p203, 356 and 361-2) show how, given vague but proper prior distributions for θ and ϕ, monitoring the η_i sequences can mistakenly diagnose convergence too early. Note that this problem would not occur if the recommended strategy of Gelman and Rubin (1992a) is followed, which is to simultaneously monitor the convergence of *all* the parameters in a model. However, in higher dimensions, monitoring all parameters may be expensive and it is tempting to monitor only those parameters which are of interest, though this is not a practice recommended by the authors.

Figure 4 here

For example if we have only a single observation $y = 0$ and take prior distributions

$$p(\theta) \sim N(\mu_\theta, \sigma_\theta^2) \qquad p(\phi) \sim N(\mu_\phi, \sigma_\phi^2),$$

then Figure 4 provides the Gibbs sampler output from five replications for the θ, ϕ and η sequences, with $\mu_\theta = \mu_\phi = 50$ and $\sigma_\theta^2 = \sigma_\phi^2 = 100$. Even by eye, it is clear that the θ and ϕ sequences cannot have converged before the 200th iteration, though the PSRF suggests that the η sequence converges immediately. In fact this is not the case, and not until both the θ and ϕ sequences converge, does the η sequence converge to the correct location. This contradicts the widely held belief that one need only monitor the parameters of interest in order to assess their convergence.

Figure 5 here

Figure 5 provides plots of the MPSRF based upon the two parameters θ and η and also the PSRF sequences for the two parameters θ and η individually. Running the Gibbs sampler for 1000 iterations and calculating the PSRF's at intervals of 20 iterations, the calculation of the MPSRF sequence took approximately the same time to compute as the two PSRF sequences combined. However, it is clear that the MPSRF sequence reflects the fact that convergence has not been achieved. Note also, the similarity in shape and scale between the MPSRF sequence and the PSRF sequence for θ.

Obviously, as the number of y-observations increases we would need to monitor more and more PSRF sequences, which quickly becomes prohibitive. However, the MPSRF summarises the PSRF sequences in a single sequence with comparable computational expense. The interpretational advantage of a single plot over a number of such plots is obvious.

4.3.3 Example: Inference for a Hierarchical Pharmacokinetic Model

For a more applied example, we consider a hierarchical nonlinear pharmacokinetics model to which a Gibbs-Metropolis algorithm was applied, see Gelman *et al* (1996) for details. Five parallel chains were run for 100,000 iterations each, with starting points drawn from an overdispersed version of the posterior distribution estimated from an earlier version of the model. To save computer memory, only every 20th iteration was saved, a fact that need not concern us here. The model has 141 parameters (17 anatomical and physiological parameters on each of 6 individual subjects, 17 population means and variances, and 5 measurement variances) and, after 100,000 iterations, approximate convergence was reached, in the sense that $\widehat{R}_c < 1.2$ for all parameters.

Figure 6 here

Figure 6 displays the plots of $\widehat{R}_c$ for all the model parameters. After an initial shaky period, these generally settle down quite well, and the plots show the convergence behaviour over time for all the parameters at once. (By comparison, it would be much harder to view the convergence by looking at plots of the simulation sequences themselves.) Alternatively, the multivariate scale reduction factor (plotted in Figure 7) is displayed as a single plot. It is clear from this plot that the chains are at least near convergence towards the end of the sequence, but the fact that the plot only really settles down towards the end of the run perhaps suggests that the chains should be run a little longer, to be sure. Note that all of the univariate PSRF plots indicate approximate convergence, which is supported by the MPSRF plot.

Figure 7 here

It should be noted that whilst the scale for the MPSRF plot in Figure 7 is considerably larger than that for the univariate PSRF measures, this is not believed to be a result of the large number of parameters being considered. In fact, repeating the MPSRF calculation for three sets of 141 independent normal samples of size 1000, the MPSRF value takes values no larger than 1.3. This suggests that the MPSRF plots should be considered to be on a similar scale to the univariate PSRF plots and that, in general, large values of the MPSRF statistic suggests real problems in the convergence of functions of the parameters other than the individual parameters themselves. Thus, the large scale for the MPSRF plot in Figure 7 suggests that, at the very least, if we are interested in quantities besides the 141 parameters individually, then it might also be sensible to examine their PSRF's.

5 Discussion

The aim of this paper has been to suggest a number of alternative implementations of the original convergence diagnostic proposed by Gelman and Rubin (1992a). Clearly an iterative approach introduces a qualitative element to the original method, providing us with information on *how* the chains are converging. Such an approach can be applied to all the methods we discuss in this paper.

The original approach can be immediately generalised to other summaries beyond second moments, thus losing the restrictive normality assumption while retaining the desirable properties of the method (most notably, that $\widehat{R}$ converges to 1 as the simulations mix). In many problems, it would be most appropriate to apply a battery of convergence tests based on the sorts of inferences that are of interest in the ultimate output of the simulations. However, the interval-based approach can be "over-optimistic," particularly in the case of skewed distributions, as shown in the example of the Cauchy distribution. Once a simulation is judged to be well-mixed (as measured by the PSRF's), it makes sense to examine the plots of their numerators and denominators, to verify that the "between" variation is decreasing and the "within" variation is increasing as the simulations converge. For higher dimensional problems, the MPSRF has a distinct advantage in terms of its interpretability, summarising the individual PSRF's and providing a bound on them as the chains converge. Verification of the MPSRF may also be obtained by plotting the relevant determinants.

We have discussed how each of the extensions of the original method provide safeguards against false diagnosis of convergence under different situations. However, all methods based solely upon sampler output can be fooled (see Asmussen *et al*, 1992), and multiple-chain-based diagnostics, whilst safer than single-chain-based diagnostics, can still be highly dependent upon the starting points of the simulations.

Having said that, we believe that, in practice, it may be useful to combine a number of the alternative approaches proposed in this paper in diagnosing convergence. For example, for high dimensional problems, the MPSRF may be calculated, together with the PSRF's for each of the parameters of interest, which might also be combined with plots of the determinants of the covariance matrices and estimated variances, respectively. Note the distinction between this "cocktail" of diagnostics and simply plotting the PSRF's for parameters of interest alone. In combining a number of approaches, the strengths of each of the approaches may be combined, and some reliability may be placed on the conclusions thus drawn.

References

Applegate, D., R. Kannan and N. G. Polson (1990), Random Polynomial Time Algorithms for Sampling from Joint Distributions. Technical report, Carneggie-Mellon University, Tech. Rep. No 500

Asmussen, S., P. W. Glynn and H. Thorisson (1992), Stationarity Detection in the Initial Transient Problem. *ACM Transactions on Modelling and Computer Simulation* **2**, 130–157

Brooks, S. P. and G. O. Roberts (1996), Diagnosing Convergence of Markov Chain Monte Carlo Algorithms. Technical report, University of Cambridge

Carlin, B. P. and T. A. Louis (1996), *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall

Cowles, M. K. and B. P. Carlin (1996), Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association* **91**, 883–904

Dellaportas, P. and A. F. M. Smith (1993), Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling. *Applied Statistics* **42**, 443–460

Efron, B. and R. J. Tibshirani (1993), *An Introduction to the Bootstrap*. Chapman and Hall: New York

Fisher, R. A. (1935), *The Design of Experiments*. Oliver and Boyd, Edinburgh

Gelfand, A. E., S. K. Sahu and B. P. Carlin (1995), Efficient Parametrizations for Normal Linear Mixed Models. *Biometrika* **82**, 479–488

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

Gelman, A., F. Y. Bois and J. Jiang (1996), Physiological Pharmacokinetic Analysis using Population Modelling and Informative Prior Distributions. *Journal of the American Statistical Association* **91**, 1400–1412

Gelman, A., G. O. Roberts and W. R. Gilks (1995), Efficient Metropolis Jumping Rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.), *Bayesian Statistics 5*, New York: Oxford University Press

Gelman, A. and D. Rubin (1992a), Inference from Iterative Simulation using Multiple Sequences. *Statistical Science* **7**, 457–511

Gelman, A. and D. B. Rubin (1992b), A Single Series from the Gibbs Sampler Provides a False Sense of Security. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.), *Bayesian Statistics 4*, pp. 625–631, New York: Oxford University Press

Gilks, W. R., S. Richardson and D. J. Spiegelhalter (1996), *Markov Chain Monte Carlo in Practice*. Chapman and Hall

Green, E. J., F. A. Roesch, A. F. M. Smith and W. E. Strawderman (1994), Bayesian Estimation for the Three Parameter Weibull Distribution with Tree Diameter Data. *Biometrics* **50**, 254–269

Grieve, A. P. (1987), Applications of Bayesian Software: Two Examples. *Statistician* **36**, 283–288

Jennison, C. (1993), Discussion to Smith and Roberts (1993). *Journal of the Royal Statistical Society, Series B* **55**, 54–56

Liu, C. and D. B. Rubin (1996), Markov-Normal Analysis of Iterative Simulations Before Their Convergence. *Journal of Econometrics* To Appear.

Mardia, K. V., J. T. Kent and J. M. Bibby (1979), *Multivariate Analysis*. Academic Press, London

Rosenthal, J. S. (1995), Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. *Journal of the American Statistical Association* **90**, 558–566

Tanner, M. and W. Wong (1987), The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association* **82**, 528–550

Tierney, L. (1994), Markov Chains for Exploring Posterior Distributions. *Annals of Statistics* **22**, 1701–1762

Yu, B. (1995), Discussion to Besag *et al* (1995). *Statistical Science* **10**, 3–66

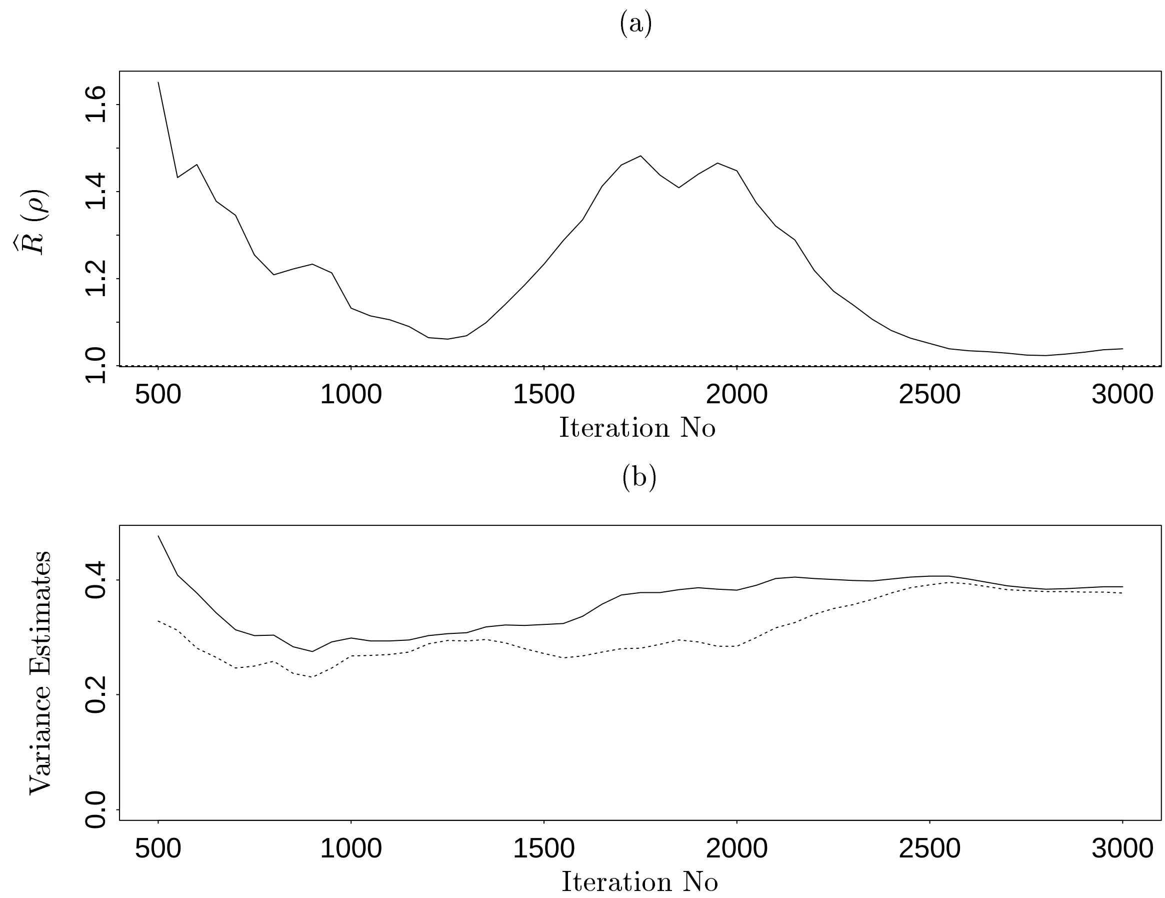

Figure 1: (a) Iterative PSRF Plot for ρ in the Weibull example (from $m = 5$ parallel sequences and with $b = 50$). After 1000 or 1200 iterations, the PSRF could misleadingly suggest that convergence has been obtained. (b) Plots of $\widehat{V}^{1/2}$ (solid) and $W^{1/2}$ (dotted) for this example. Not until about 2500 iterations are both lines stable and close to each other.

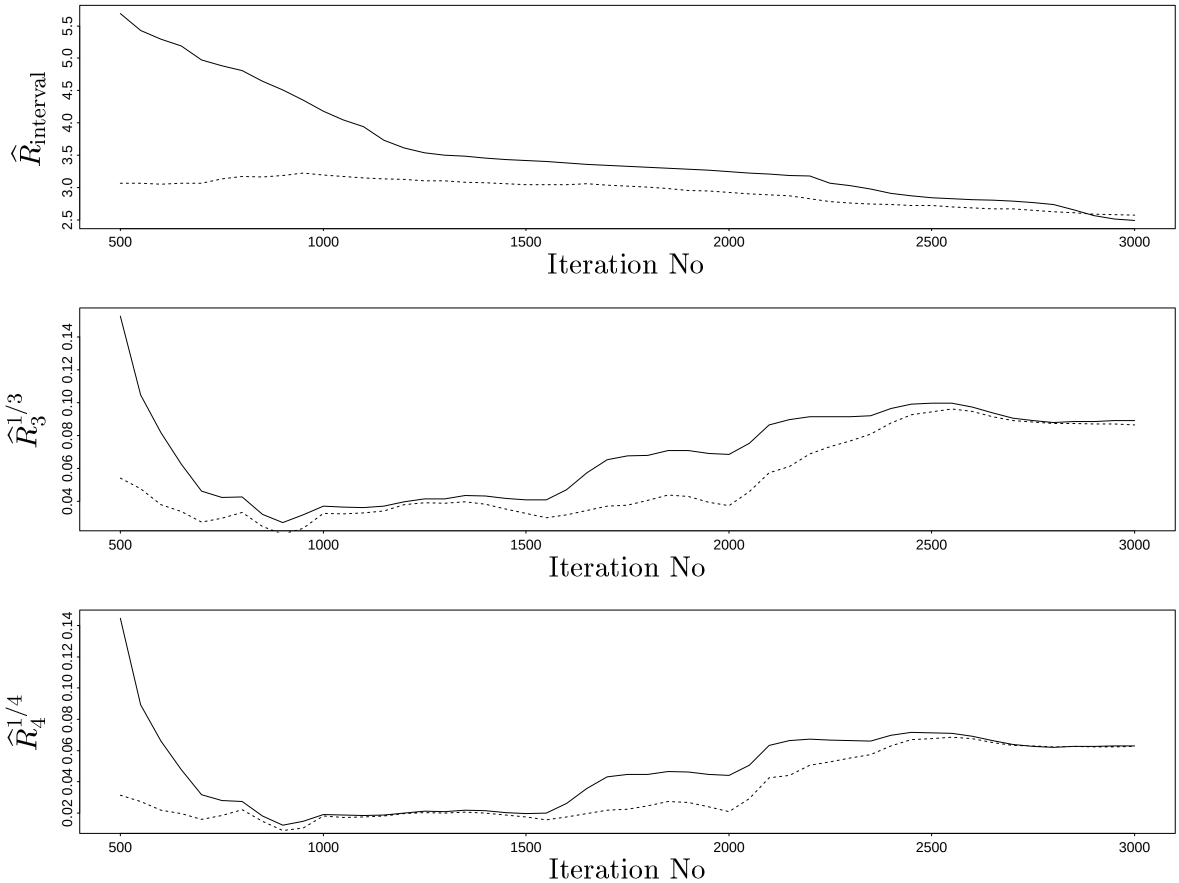

Figure 2: Numerators (solid lines) and denominators (dotted lines) for several univariate convergence measures for the shape parameter ρ, in the Weibull example: (a) Interval-based measure, (b) $\widehat{R}_3^{1/3}$, and (c) $\widehat{R}_4^{1/4}$. For each plot, the solid line represents the scale from the mixture of the m simulated sequences, and the dotted line represents the average scale of the m individual sequences. Like Figure 1b, these plots indicate that approximate convergence has not been reached before about 2500 iterations.

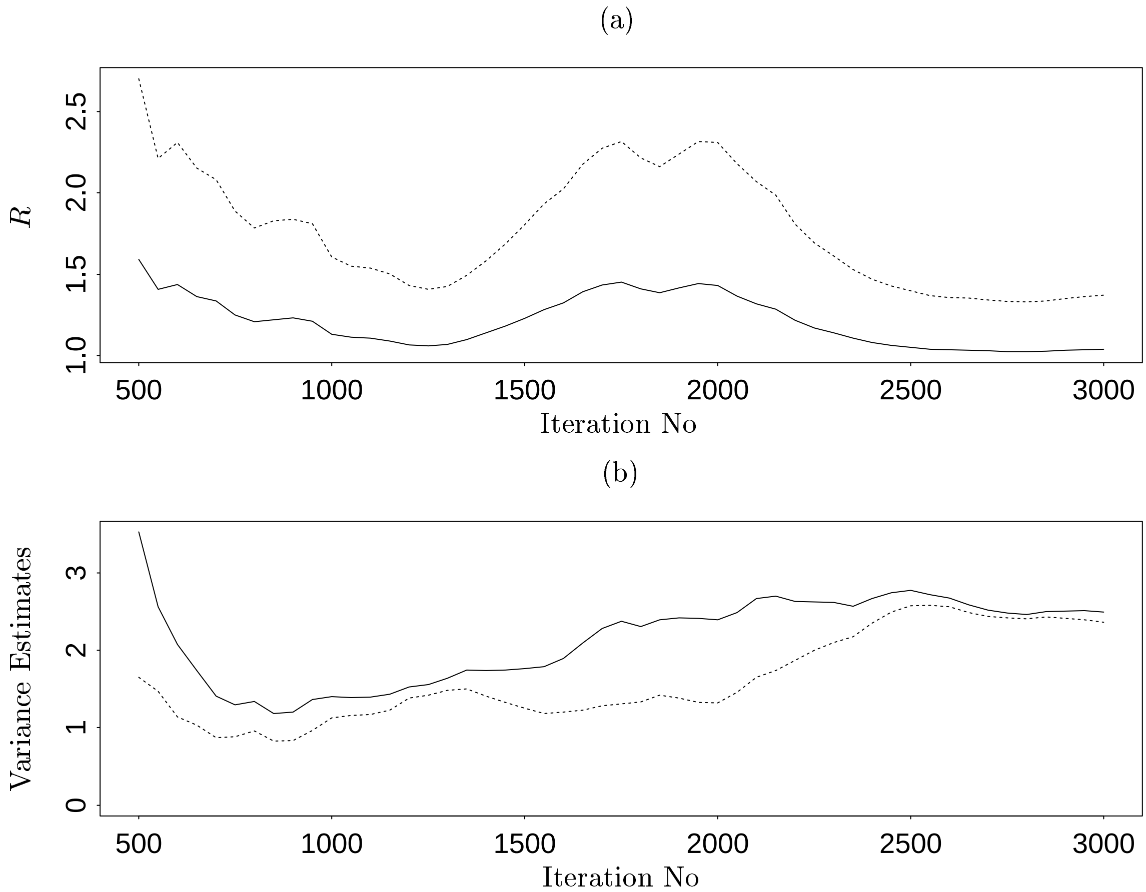

Figure 3: (a) Plots of maximum of the individual PSRF's ($\widehat{R}^{max}$, solid) together with bound based upon $\widehat{R}^5$ (dotted). (b) Plot of the the determinants ($\times 10^6$) of the pooled (solid) and within-chain (dotted) covariance matrices.

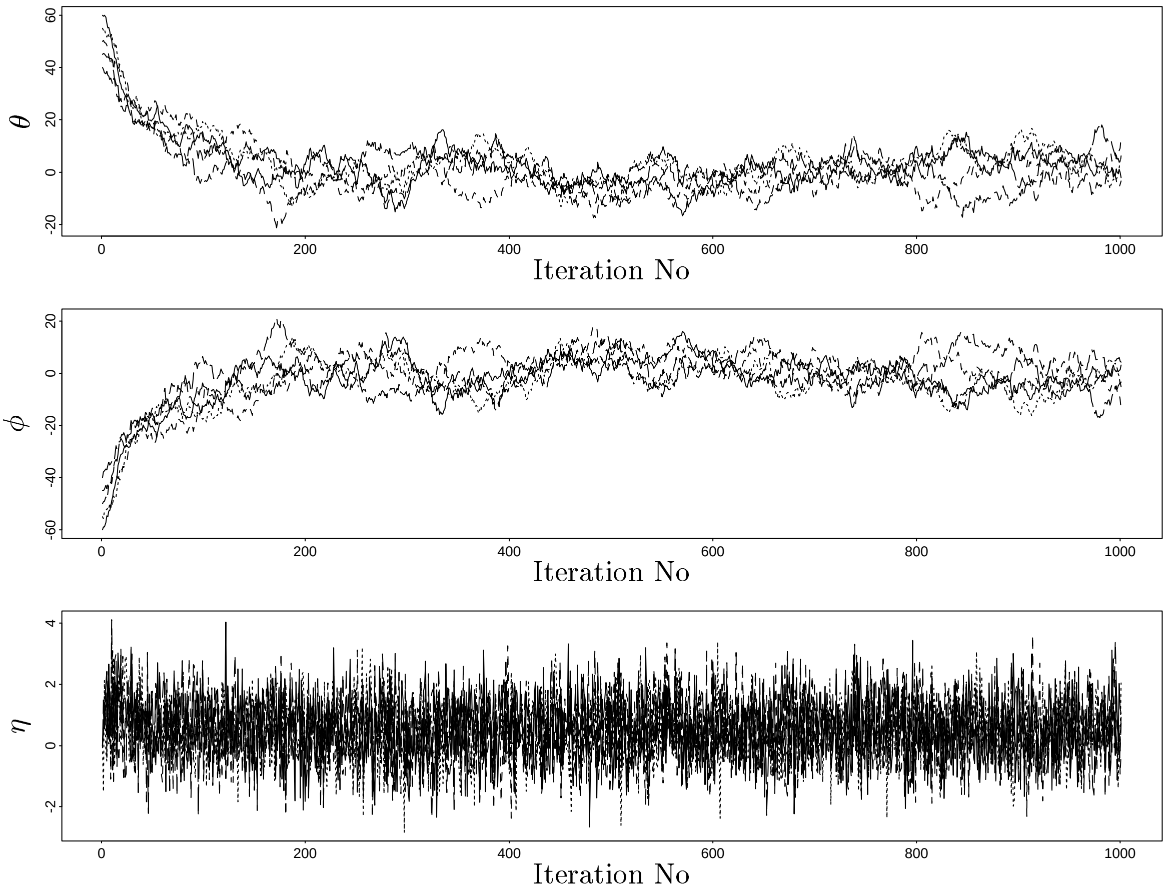

Figure 4: Raw parameter values for the bivariate normal example.

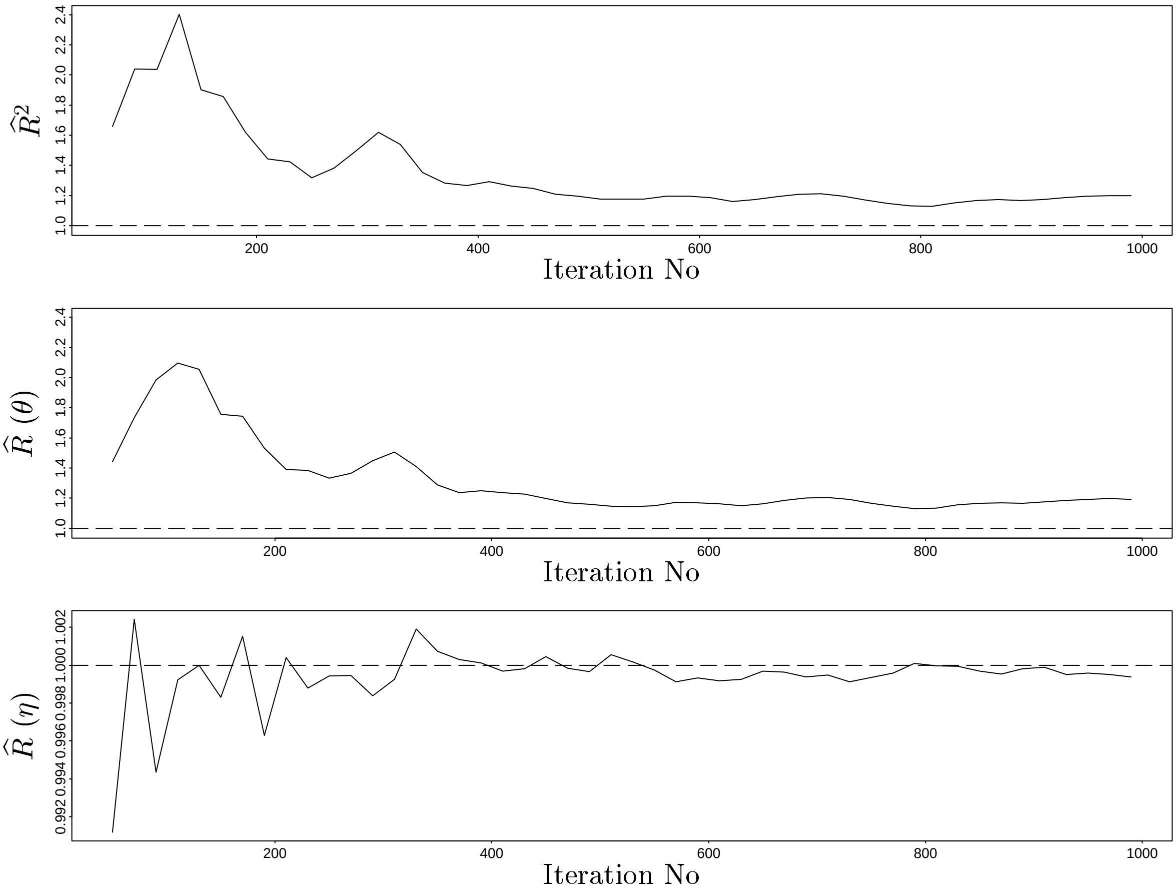

Figure 5: Diagnostic output sequences for the bivariate normal example. (a) MPSRF. (b) PSRF for θ. (c) PSRF for η.

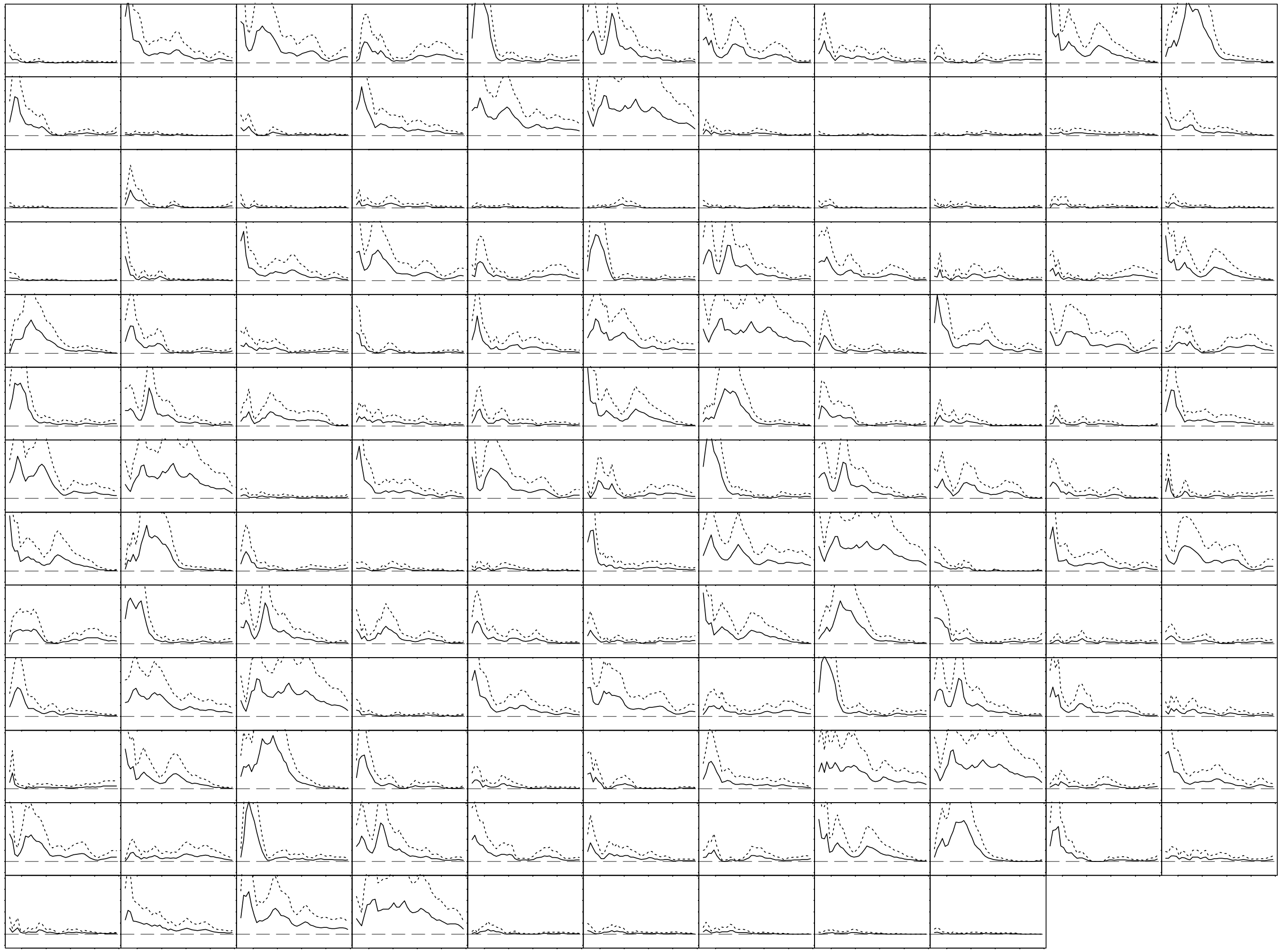

Figure 6: Plots of $\widehat{R}_c$ and 97.5% upper bounds for all the model parameters in the pharmacokinetic example, scaled so that they all range from 0.8 to 2.0 in the y-axis.

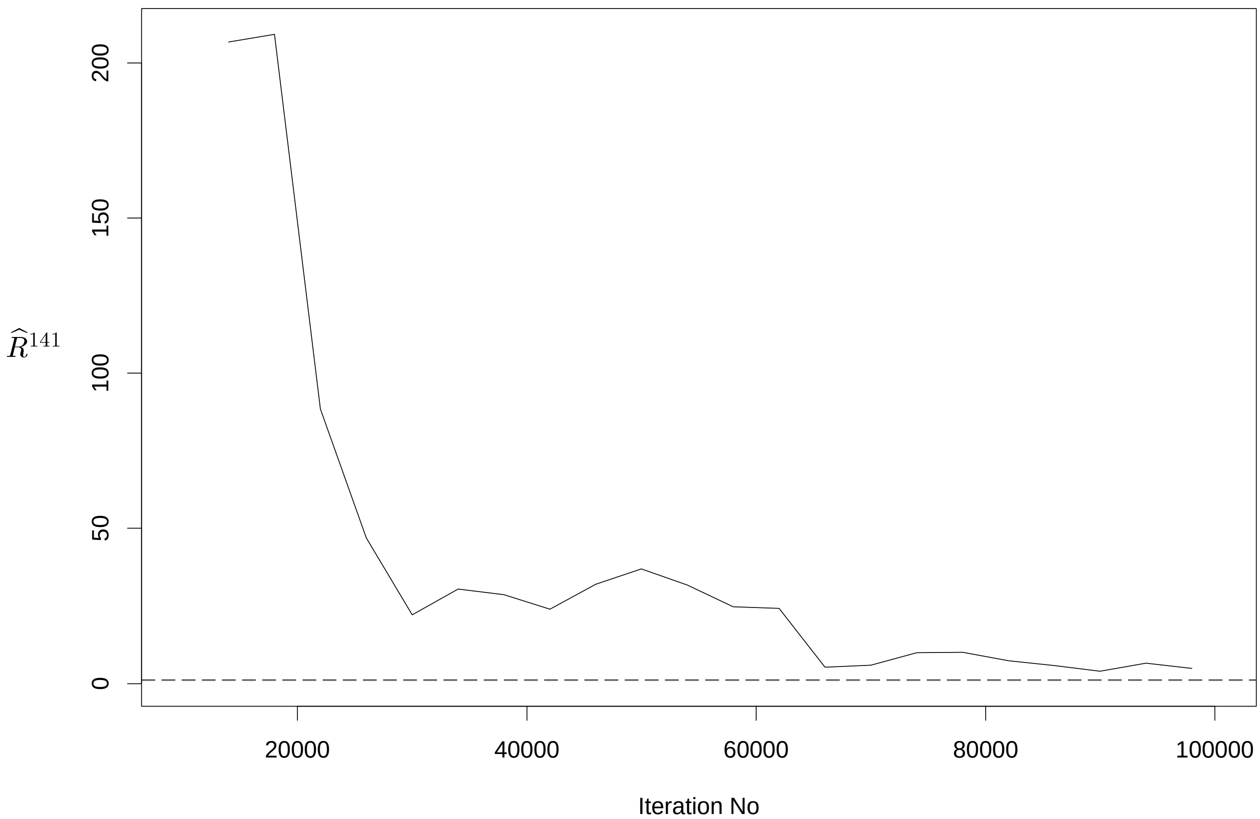

Figure 7: Plot of the MPSRF calculated from all 141 parameters in the pharmacokinetic example.