MCMC Convergence Diagnostics in Action

Chantal Guihenneuc–Jouyaux
URA CNRS 1323, Université Paris 5, Paris, France,
Sonia Knight
QUT, Brisbane, Australia,
Kerrie L. Mengersen
QUT, Brisbane, Australia,
Sylvia Richardson
INSERM, Villejuif, France,
Christian P. Robert
UPRES-A CNRS 6085, Université de Rouen and CREST–INSEE, Paris, France

Abstract Existing convergence diagnostics for Markov chain Monte Carlo (MCMC) algorithms are presented as a phylogeny according to three consecutive convergence requirements encountered in practice: exploration (is the chain now sampling from the stationary distribution and is subsampling required?), stationarity (is the chain adequately moving around the state space?) and estimation (can the required parameters be estimated with adequate precision?). The performance of the different diagnostics in Bayesian analysis is then illustrated and evaluated through four complex applications: mixtures and hidden Markov models, quantitative trait loci (QTL) estimation, stochastic volatility models, and HIV progression models, based on earlier and present experiments.

Key Words: Markov chain Monte Carlo, asymptotics, diagnostics, stationary distribution, ergodicity, renewal theory, CLT, parallel chains, mixing, subsampling, `CODA`, hidden Markov chains, stochastic volatility.

1 Introduction

Markov chain Monte Carlo (MCMC) methods have become a standard statistical tool for the analysis of complex models. A number of explanations may be given for this. First, their theoretical foundation is sound, especially compared with many competing numerical solutions, as detailed by Besag *et al.* (1995) and Robert and Casella (1999), among many others. Second, their relative ease of implementation is appealing. This is obvious when considering the range of problems to which it has been applied and the variety of backgrounds of those practitioners and researchers who are adopting the tool. Third, it is a very open technique, in the sense that, although the common algorithms (Gibbs, Metropolis–Hastings) are by far the most well-known, more efficient alternatives are the subject of active research and implementation. Examples include simulated annealing (Jennison, 1993; Neal, 1998), simulated tempering (Geyer and Thompson, 1995), Langevin algorithms (Roberts and Tweedie, 1995), acceleration and perfect sampling (Propp and Wilson, 1996; Murdoch and Green, 1998).

As with all approximation methods, however, valid use of MCMC requires rigorous assessment of the pertinence of the results. Although the simulated chains do theoretically converge to the required distribution, this must be confirmed in practice. Convergence properties based on Harris recurrence (Tierney, 1994; Chan and Geyer, 1994; Athreya *et al.*, 1996), the Central Limit theorem (Tierney, 1994; Robert, 1995; Roberts and Tweedie, 1996; Roberts and Rosenthal, 1998) geometric ergodicity (Schervish and Carlin, 1992; Mengersen and Tweedie, 1996; Polson, 1996; Roberts and Tweedie, 1996) are somehow reassuring but their premises are often difficult to verify in practice and/or their bounds are impractically conservative. As a consequence, a body of literature has grown on *diagnostics*, which here we define as tools for assessing convergence "on line" (as one or several parallel chains are running), which are applicable to a wide range of problems, and which are (perhaps relatively) easily implemented.

Classification of this diagnostic toolkit has taken a number of forms. Cowles and Carlin (1996) provide a brief description of each method and summarise them according to graphical or quantitative nature, use of single or multiple chains, employment of the univariate or full joint distribution, focus on bias and/or variance, applicability, ease of use and theoretical basis. Brooks and Roberts (1999) focus on categorisation of methods by their theoretical basis and make comparisons based on use of the full joint density, multiple chains, applicable samplers, computational expense, interpretability and restrictions. Robert and Casella (1999) consider three increasingly stringent types of convergence (convergence of the chain to the stationary distribution, convergence of the empirical average to the appropriate distribution, convergence of a sample to *iid*-ness) and classify methods according to these. In Section 3, we provide brief details of an alternative classification proposed by Mengersen, Robert and Guihenneuc–Jouyaux (1999) which is based on three consecutive stages. First, one may wish to confirm that the MCMC chain is indeed sampling from the stationary distribution. Second, it is important to ascertain that the chain is adequately exploring the support of this distribution. Third, adequate precision of estimation of parameters of interest must be realised. This tripartite phylogeny is then further subdivided according to general approach.

The purpose of this paper, however, is to get closer to the practitioner's point of view, by focusing on the application of these tools, their interpretation and the relative contribution that they make to the overall assessment of convergence. Although they fully serve their purpose, namely to illustrate and compare the various diagnostics in a compact presentation, the examples used in the reviews of Robert (1995), and Cowles and Carlin (1996) might be considered as academic, in the sense that they cover cases where the call to MCMC algorithms is not the only solution or at least where the convergence issues are rarely at stake. As in Brooks (1998), we choose here to present more realistic examples which require careful and complex modelling (including latent variables) and which expose some of the difficulties of implementing the various diagnostics. The drawback of this choice, obviously, is that we cannot cover all possible applications of MCMC algorithms, since the selected examples present idiosyncrasies which do not extend to other models and highlight specific defects of the different diagnostics which are not necessarily damning in different settings. These examples are, nonetheless, representative of a large area of complex statistical models, often called *latent variable models* or *missing data models*, for which the sole feasible approach is to use MCMC methods. The four detailed examples, described in Section 2, are the result of active research problems considered by the authors in the last years. These comprise mixtures and hidden Markov models, estimation of gene order and quantitative trait loci (QTL), stochastic volatility, and HIV progression models. They provide the context in which to illustrate a selection of the diagnostics described in Section 3. Section 4 then provides some cohesion about analysis and consequent conclusions for each example, and some observations about the diagnostic toolkit as a whole.

2 Examples

2.1 Example 1: Mixtures and hidden Markov chains

As a first example we consider mixtures of distributions,

$$\tilde{f}(x) = \sum_{j=1}^{k} p_j \, f(x|\xi_j) \; , \tag{1}$$

where $p_1 + \ldots + p_k = 1$, and the $f(x|\xi_j)$'s are from exponential families. We refer to Diebolt and Robert (1994), Mengersen and Robert (1996), Robert (1996) or Gruet *et al.* (1999) for details on the MCMC implementation of Bayesian inference in these models. Let us only mention two important points: first, Bayesian noninformative modeling is possible on reparameterised versions of (1), as shown by Mengersen and Robert (1996), Roeder and Wasserman (1997) and Robert and Titterington (1998). For instance, a three component normal mixture can be rewritten as

$$p\mathcal{N}(\mu, \tau^2) + (1-p)q\mathcal{N}(\mu + \tau\theta, \tau^2\sigma^2) + (1-p)(1-q)\mathcal{N}(\mu + \tau\theta + \tau\sigma\varepsilon, \tau^2\sigma^2\omega^2) .$$

Second, mixtures are a special case of *latent variable models*, where the latent variables are the indicator variables $z_i \in \{1, \ldots, k\}$ such that

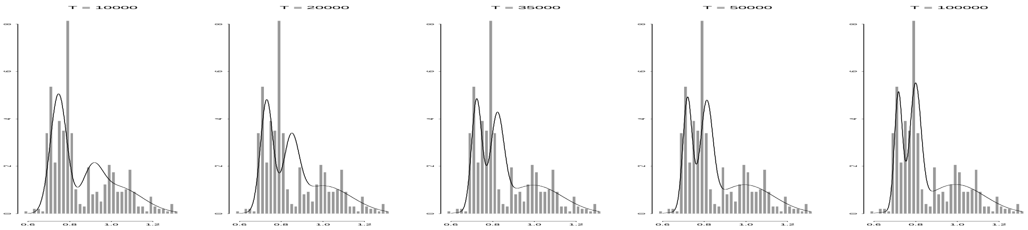
$$z_i \sim \mathcal{M}_k(1; p_1, \ldots, p_k) \qquad \text{and} \qquad x_i|z_i \sim f(x|\xi_{z_i}).$$

The corresponding MCMC algorithms then take advantage of this structure to work on the completed model, where a Gibbs sampler can most often be implemented, although this is not necessarily the best possible choice.

Convergence assessment in this setup is, to say the least, an open problem because of the complex structure of the posterior distribution. Indeed, when developing the likelihood as a sum of k^n exponential-type terms, it is easy to see that the posterior distribution also looks like a weighted sum of (perhaps standard conjugate) distributions and is thus most likely multimodal. Moreover, the complexity of the posterior associated with (1) increases dramatically with k, with significant modes around k' component submodels ($k' < k$). A standard MCMC algorithm is then unlikely to explore the whole range of the posterior modes and usually stays in the neighbourhood of one of the major modes, with rare jumps between modes. While most of the modes are minor and possibly redundant, they cannot be excluded from the start. Moreover, the invariance of (1) under permutation of the indices implies that the posterior distribution associated with a symmetric prior possesses $k!$ modes of equal height. This symmetric feature of the posterior should be observed in the Markov chain but most MCMC implementations have difficulties to move from one mode to another.

The complexity of the stationary distribution $f(\theta)$ and the possible call to component-wise Metropolis–Hastings steps prohibits the use of the methods involving a detailed study of the transition kernel like distance estimation (see Mengersen *et al.*, 1999). Moreover, the data augmentation structure also prevents the call to techniques based on renewal theory, since the renewal probabilities are decreasing exponentially quickly with the sample size. On the reverse, Figure 1, which represents a sequence of 3 component normal mixture densities based on the estimated parameters for different iteration sizes, constitutes a very rudimentary convergence assessment since the underlying parameters have not at all converged (see Figure 4).

Figure 1: Markov Chain Monte Carlo estimates of a 3 component normal mixture at different iterations, based on the dataset of Izenman and Sommer (1988). (*Source:* Robert and Mengersen, 1998.)



More complex mixture structures will also be considered in this example, with similar difficulties in the convergence assessment. For instance, a normal hidden Markov chain introduces a Markovian dependence between the latent variables

$$P(z_t = u|z_{t-1} = v) = p_{vu} , \qquad x_t|z_t \sim \mathcal{N}(\mu_{z_t}, \sigma^2) ,$$

thus marginally remaining a mixture model. As shown for instance in Robert and Titterington (1998), this model faces the same difficulties as mixture models in terms of Bayesian inference and MCMC convergence assessment and it requires the development of specific diagnostics. The following examples are particular cases of mixture and hidden Markov models, respectively.

2.2 Example 2: Mixture models for QTL localisation

The mixture models described in the previous section provide a natural expression of biological characteristics in genetics. For example, the expression in its progeny of a sire's genotype Qq is a mixture of genotypes QQ, Qq and qq, with weights depending on the allele frequency p_Q, the dams' genotypes and other genetic information. A recent focus of such models is the construction of gene marker maps for parentage testing, disease resistance diagnosis and *quantitative trait loci* (QTL) detection. Mapping is made possible through the collection of marker information, which is genotype information observed at known locations (loci) along the genome. Inter-gene (or inter-marker) distances are proportional to recombination rates where a recombinant event is the observed result of a cross-over between two chromosomes. Estimation of gene order and detection of the QTL can both be undertaken in a Bayesian framework using MCMC. This is particularly useful for complex situations such as a half-sib design which is characterized by a large number of families, each family consisting of a large number of nuclear families with a single parent in common, but with all dam genotype information missing. Three key advantages of this approach over competing methods is the estimation of dam allele frequencies, the interpretation of the latent variables as allocation of progeny to the different genotype classes, and the direct interpretation of different "models" as different gene orders or QTL positions.

George *et al.* (1998, 1999) provide more details on the genetic and statistical models and the computational method appropriate to these analyses. For the QTL localisation problem (George *et al.*, 1999), the statistical representation of the genetic model is given by $Y_{ji} = \mu_{Q_{ji}} + e_{ji}$, where for offspring i in family j, Y_{ji} is the trait value, $Q_{ji} \in \{QQ, Qq, qq\}$ is the QTL genotype, $\mu_{Q_{ji}}$ is the expected trait value given QTL genotype Q_{ji} and e_{ji} is the random deviation from $\mu_{Q_{ji}}$.

For s half-sib families, the likelihood of the quantitative trait data and the marker data becomes a mixture of four within family likelihoods, conditioned on the recombination rates $\underline{\theta}$, the marker allele frequency in the dams $\underline{p}$, the locus order α_l ($QM_1M_2, .., M_N$ for $l = 1$; $...M_{l-1}QM_{l+1}$ for $l = 2, .., N$; $M_1M_2...M_NQ$ for $l = N + 1$), the residual variance σ_R^2, p_Q, $\underline{\mu}$,

$$\prod_{j=1}^{s} \{p_Q^2 L_{QQ}(\underline{Y}_j \mid p_Q, \mu_1, \mu_2, \sigma_R^2) + p_Q(1-p_Q)L_{Qq}(\underline{Y}_j, \underline{m}_j, m_j \mid \underline{\theta}, \underline{p}, p_Q, \underline{\mu}, \sigma_R^2, \alpha_l) +$$
$$(1-p_Q)p_Q L_{qQ}(\underline{Y}_j, \underline{m}_j, m_j \mid \underline{\theta}, \underline{p}, p_Q, \underline{\mu}, \sigma_R^2, \alpha_l) + (1-p_Q)^2 L_{qq}(\underline{Y}_j \mid p_Q, \mu_2, \mu_3, \sigma_R^2)\} \ .$$

where $\underline{Y}_j$ and m_j are the quantitative trait data and marker data for offspring of the jth sire, respectively, and m_j is marker data for the jth sire. Marker data for the sire includes (assumed) information on how the genotypes are oriented on a pair of chromosomes (phase), while the orientation is unknown for offspring.

The single family likelihoods for sires with QTL genotypes QQ, qq and $x \in \{Qq, qQ\}$ are given by

$$L_{QQ}(\underline{Y}_j \mid p_Q, \mu_1, \mu_2, \sigma_R^2) \propto \prod_{i=1}^{n_j} \{p_Q \phi(Y_{ji} \mid \mu_1, \sigma_R^2) + (1-p_Q)\phi(Y_{ji} \mid \mu_2, \sigma_R^2)\} \ ,$$

$$L_{qq}(\underline{Y}_j \mid p_Q, \mu_2, \mu_3, \sigma_R^2) \propto \prod_{i=1}^{n_j} \{p_Q \phi(Y_{ji} \mid \mu_2, \sigma_R^2) + (1-p_Q)\phi(Y_{ji} \mid \mu_3, \sigma_R^2)\} \ ,$$

where the product is over the n_j offsprings, and

$$L_x(\underline{Y}_j, \underline{m}_j, m_j \mid \underline{\theta}, \underline{p}, p_Q, \underline{\mu}, \sigma_R^2, \alpha_l) \propto \prod_{i=1}^{n_j} \sum_{k=1}^{3} p(m_{ji}, Q_{ji} = k \mid m_j, \alpha_l, Q_j = x) \phi(Y_{ji} \mid \mu_k, \sigma_R^2) \,,$$

where the sum is over the three QTL classes for the n_j offsprings, Q_j is the sire's QTL genotype, and $p(m_{ji}, k \mid m_j, \alpha_l, Q_j)$ is the conditional probability of an offspring genotype m_{ji} and a QTL class k (ϕ is the Gaussian density).

The four steps in the MCMC sampling procedure comprised allocation of sires to QTL genotypes, of offspring to QTL genotypes, the allocations being denoted by $\underline{z}$, generation of new values for the genetic parameters and of a new locus ordering. In the spirit of this paper, we focus on the parameters $\underline{\theta}$, p_Q, $\underline{\mu}$, $\underline{z}$ and $\underline{\alpha}$. We use two simulation designs considered in George *et al.* (1998): Design 1 with a large QTL, 20 sires and 200 offspring per sire, and Design 2 with a small QTL, 5 sires and 200 offspring per sire. The large QTL had the same expected trait value, μ_{ji}^G, but a smaller residual variance σ_R^2. The first design exhibits quite rapid convergence whereas the second requires much more careful implementation and a much longer run time. In this study, four markers, with a QTL between the 3rd and 4th markers, were specified.

Assuming independence of recombination rates along the chromosome, the prior for $\underline{\theta}$ was $p(\underline{\theta} \mid \alpha_\ell) = \prod_{i=1}^N p(\theta_{i,i+1} \mid \alpha_\ell)$. For the recombination rate between markers, a Gaussian prior centered around the published value $\hat{\theta}_{i,i+1}$ and with specified standard deviation sd, say, was placed on θ. The Gaussian density was truncated at $\hat{\theta}_{i,i+1} \pm 2 \times sd$. Since no information was available on the position of the QTL, a uniform density over $[0, K]$ was placed on the recombination rate, with K the upper bound of the allowable support for θ. This careful choice of prior, based on genetic information, considerably improved convergence. Although stronger priors could be proposed for the allele frequency for the QTL, the following results are based on a flat prior. Convergence of p_Q, and hence stability of the allocation vector $\underline{z}$, was improved by the use of a specialized proposal density: p_Q was sampled from a hybrid of two proposal densities, the uniform and a normal density centered around 0.5. The window size h of the uniform, the variance of the normal density and the proportion with which the uniform and normal are sampled were calibrated by the acceptance rate.

The reversible jump sampler was characterized by two tasks: proposing a move between alternate locus orderings that induces good mixing and determining the probability of acceptance of a particular move. Forward and backward moves were constructed in which the QTL pivoted around a randomly selected marker, based on the *Haldane map distance.* Equal weighting was given to the different orderings, so $p(\alpha_\ell) \sim 1/(N+1)$ where $N+1$ is the number of possible locus orderings.

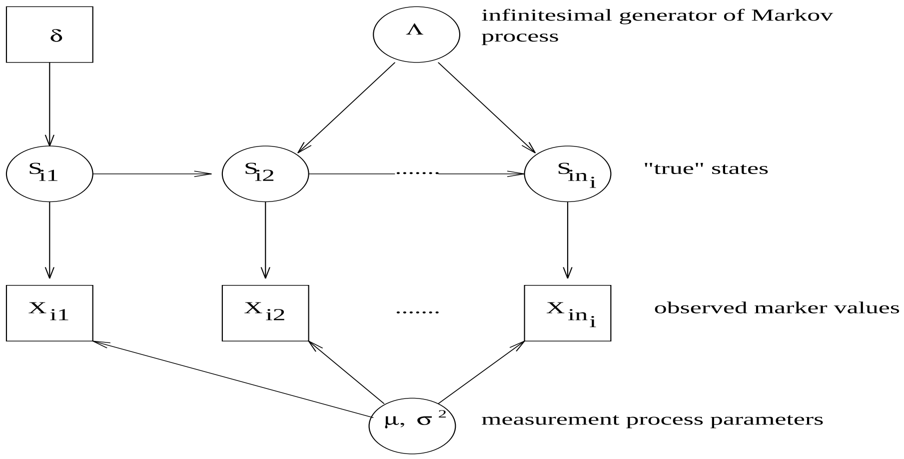
Loci orderings with low posterior probabilities (eg < 0.1) were not considered in convergence assessment due to the insufficient lengths of their parameter chains and an obvious lack of importance in the QTL mapping problem. Parallel chains were run for the Exploration phase, but since the MCMC runs were computationally demanding, at the completion of this stage only one run was randomly selected to continue.

2.3 Example 3: Markov model for HIV progression

Markov chain models of the natural history of HIV play an important part in AIDS modeling, for the description of HIV progression, the estimation of transition rates and the evaluation of the therapy effects on progression. The modelling is done through a special type of hidden Markov structure, as in §2.1, with the difference that the underlying model operates in continuous time. The different states of the Markov chain then correspond to a discretisation of continuous markers (*CD4 cell count*), which are highly variable due to within-individual effects and measurement errors.

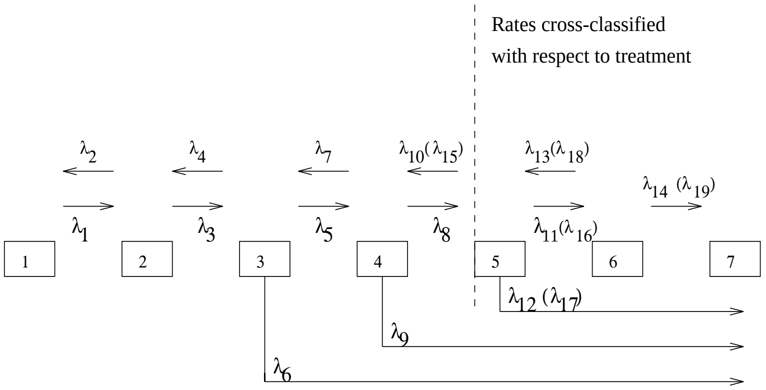
Guihenneuc–Jouyaux *et al.* (1998) use for instance a Bayesian hierarchical model, represented by Figure 2, whose first level is a disease process, that is a Markov model on the unobserved true states, and whose

Figure 2: Directed acyclic graph representation of the Bayesian hierarchical model of HIV progression. (*Source:* Guihenneuc–Jouyaux *et al.*, 1998.)



second level is the measurement process, linking the true states with the marker values. The authors apply this representation to the "San Francisco Men Health Study" dataset: This database consists of 430 patients monitored roughly every 6 months from mid 1984 through September 1992 and it contains 3833 follow-up times. At each follow-up time, the therapy is registered and a patient is *treated* at time t if he has received AZT and/or Pentamidine before t. Guihenneuc–Jouyaux *et al.* (1998) make the additional assumption of the persistence of the treatment effect but alternatives also apply. Only the patients with CD4 cell count under 350 can have therapy. For the disease process, they assume an underlying time homogeneous Markov process with transient stages 1-6 (corresponding to the disease progression and thus based on the CD4 cell count) and absorbing stage 7 (corresponding to AIDS and established on clinical symptoms). The model allows back flows and direct passages to AIDS (stage 7) from stages 3–5, as represented in Figure 3. There is a possibility of therapy at stage 5 and above and thus two kinds of transition rates to be estimated, corresponding to treated and untreated follow-ups. The resulting model has 19 transition rates (noted by λ_1 to λ_{19}).

Figure 3: Markov model of disease progression. (*Source:* Guihenneuc–Jouyaux *et al.*, 1998.)



The measured log(CD4)'s, X_{ij}, given the true stage S_{ij}, are independent Gaussian with unknown mean $\mu_{S_{ij}}$ and variance $\sigma^2_{S_{ij}}$. The prior distribution on the exponentials of the means, $m_i = \exp(\mu_i)$, is derived from order statistics on $[50, 1100]$. More precisely, if $U_1, \ldots, U_{20}$ are 20 i.i.d. replications from a uniform distribution on $[50, 1100]$ and if $U_{[1]}, \ldots, U_{[20]}$ denote the corresponding *decreasing* ordered variables, then m_i is distributed as $U_{[4*(i-1)]}$ for $i > 1$ (the first mean is equal to 1100). The choices of the sample size of 20 and of the spacing of 4 between the indexes of $U_{[\cdot]}$ ensure a mean spacing $\mathbb{E}[m_i - m_{i+1}]$ equal to 200 CD4 (and thus $\mathbb{E}[m_2] = 900, \mathbb{E}[m_3] = 700, \mathbb{E}[m_4] = 500, \mathbb{E}[m_5] = 300, \mathbb{E}[m_6] = 100$), corresponding to a classical discretisation in the literature. The full conditional distribution of m_i given the other means is then a Beta distribution, $\mathcal{B}e(4,4)$ on $[m_{i+1}, m_{i-1}]$. There are four variances: one for the first stage (var1), since it corresponds to an open ended interval of values of CD4, one for stages from 2 up to 4 (var2), where no treatment is given, and two for the stages 5 and 6, one corresponding to the treated patient (vartT) and one to the untreated patients (varNT). The parameter set includes in addition the unobserved stages of the disease process for the 3833 follow-up times.

Given the complexity of the model, due to both the presence of the latent variables S_{ij} and to exponentials of matrices, the only possible implementation is via a MCMC algorithm, which is a Gibbs sampler in Guihenneuc–Jouyaux *et al.* (1998). Since the full conditional distribution of the transition rates in the Gibbs sampling involves the exponential of the infinitesimal generator matrix Λ, i.e. products of terms $\exp\{\Lambda t_{ij}\}$, where t_{ij} is the interval between follow-ups $j-1$ and j for patient i, the actual implementation requires an additional Metropolis–Hastings step and does not lead to a closed form kernel. The complexity of this distribution and the simulation of a large number of latent variables lead to huge running times which prohibit the use of parallel runs. The diagnostics studied in Section 4 correspond to a single run of 10,000 iterations.

2.4 Example 4: Stochastic Volatility

Stochastic volatility models are quite popular in Finance where they describe series with sudden changes in their magnitude (see, e.g., Kim, Shephard and Chib, 1998). They rely on a latent linear process (y_t^*), the so-called *volatility*, which governs the variance of the observables $(t = 1, \ldots, T)$:

$$\begin{cases} y_t^* = \alpha + \beta y_{t-1}^* + \sigma \epsilon_{t-1}^* \ , \\ y_t = e^{y_t^*/2} \epsilon_t \ , \end{cases} \tag{2}$$

where the ϵ_t's and ϵ_t^* are iid $\mathcal{N}(0,1)$ rv's, and $y_0^* \sim \mathcal{N}(\alpha, \sigma^2)$. The (observed) likelihood $L(\alpha, \beta, \sigma | y_0, \ldots, y_T)$ is then implicit, since it implies integrating out the latent variables y_t^*. On the contrary, the completed likelihood, which corresponds to the full model (2), is available in closed form, as

$$\sigma^{-T+1} \exp - \left\{ (y_0^* - \alpha)^2 + \sum_{t=1}^{T} (y_t^* - \alpha - \beta y_{t-1}^*)^2 \right\} / 2\sigma^2 \exp - \sum_{t=0}^{T} \left\{ y_t^2 e^{-y_t^*} + y_t^* \right\} / 2 \ .$$

The simulation of the y_t^*'s can be done by conditional Metropolis–Hastings steps, based on an approximation of

$$f(y_t^* | y_{t-1}^*, y_{t+1}^*, y^t, \theta) \propto \exp \left\{ -\frac{(y_t^* - \alpha - \beta y_{t-1}^*)^2}{2\sigma^2} - \frac{(y_{t+1}^* - \alpha - \beta y_t^*)^2}{2\sigma^2} - y_t^*/2 - y_t^2 e^{-y_t^*}/2 \right\} .$$

For instance, $y_t^*/2 + y_t^2 e^{-y_t^*}/2$ can be approximated by $(y_t^* - \log(y_t^2))^2/2$ as in Billio *et al.* (1998). The proposal distribution is then a Gaussian

$$\mathcal{N}\left(\frac{[\alpha + \beta y_{t-1}^* + \beta(y_{t+1}^* - \alpha)]\sigma^{-2} + \log(y_t^2)/2}{(1+\beta^2)\sigma^{-2} + 1/2}, \left[(1+\beta^2)\sigma^{-2} + 1/2\right]^{-1} \right) ,$$

with the appropriate Metropolis–Hastings acceptance probability. As discussed in Kim *et al.* (1998), several alternatives are available in the literature, some of which are superior to this solution (see also Jacquier *et al.*, 1994, or Shephard and Pitt, 1997). The point, however, is in analysing the performances of different diagnostics on this algorithm, rather than proposing a more efficient solution.

This model is illustrated by two simulated samples. First, a sample of size 500 was obtained using $\alpha = 0$, $\beta = 0.9$ and $\sigma = 0.5$. This parameterisation induces a very stable chain, as will be indicated by the diagnostics employed in Section 4. A second sample of size 250 observations was then obtained using $\alpha = 0$, $\beta = 0.7$ and $\sigma = 0.1$. In this case, mixing is very slow.

3 A Phylogeny of Diagnostics

Given a simulated chain $\theta^{(t)}, t = 1, , ..T$ (or more generally, several chains generated in parallel), the purpose of most convergence diagnostics is to assess or control the proximity of the distribution of $\theta^{(t)}$ and of the target (stationary) distribution f and/or of the accuracy of an approximation of a parameter of interest $\mathbb{E}_f[h(\theta)]$, for instance by the empirical average

$$S_T = \frac{1}{T} \sum_{t=1}^{T} h(\theta^{(t)}). \tag{3}$$

As discussed in the Introduction, different philogenies can be developed for these convergence diagnostics. The categorisation proposed by Mengersen *et al.* (1999) employs three main stages (Exploration, Stationarity, Estimation) and develops subcategories based on methodological approaches. This is briefly described in Table 1, with details of methods, motivations, the corresponding authors and references provided in the above paper. The motivation behind the distinction between the *Exploration* and *Stationarity* stages is that the fact that the chain is stationary, that is, that $\theta^{(t)} \sim f$, is not a sufficient condition for good convergence properties, in the sense that the chain may be very slowly mixing, as in the case of multiple mode posterior distributions. Similarly, good mixing properties do not necessarily lead to acceptable estimators (3), for instance when the support of h is concentrated, compared with the support of f.

Given the goal of this paper, which is to illustrate the use of the different diagnostic tools in specific setups, we will not develop a general comparison of the available diagnostics. Criticisms and limitations of each tool can be found in Mergersen *et al.* (1998) as well as on the companion Web site `http://www.ensae.fr/crest/statistique/robert/McDiag`, which aims at keeping an updated vision of the available diagnostics.

As commonly acknowledged and as is evident from the above phylogeny, no single diagnostic in the toolkit addresses all stages of convergence or all problems. Moreover, different diagnostics will give conflicting confirmation of convergence, possibly only agreeing for the more simple applications. As a result, it is good practice to employ a variety of tools, with the admission that some expert arbitration is often needed to solve disputes between diagnostic outputs or even to calibrate some of the methods. Of course, conservatism is recommended, in that it is generally preferable to err by running the chain for longer than necessary in order to satisfy all diagnostics, even though relatively extreme stopping rules that may be advocated by some methods must be weighed accordingly. Indeed, problem-specific examination of the goals of these methods and the validity of the assumptions on which they are based may also be revealing. (See, for example, the caution advised by Brooks and Roberts (1997) when focus is on quantile estimation.)

4 Diagnostic Methods Illustrated

We now examine the performances of the items of the above toolkit in the setups of the various examples introduced in Section 2, even though some of the methods in the toolkit are not considered here. Preference is given to those which are relatively new and to representatives of the various cells of the phylogeny of Section 3. Further detail about these and other methods is provided in part by Mengersen *et al.* (1999) and the other reviews described in the Introduction. Some of the methods are also available in the software package CODA (Best *et al.*, 1995) designed specifically for MCMC diagnostics.

Table 1: Tripartite Categorisation of MCMC diagnostics, following Mergersen *et al.* (1998).

	Exploration: *Is* $(\theta^{(t)})$ *sampling from* f*?*
Graphical Approaches	raw plots, cumulative averages, histograms and Box plots; graphical and quantitative assessment of CUSUMs
Distance estimation	comparison of f_t and f via $\|\|f_t - f\|\|$ (L1, L2), $\mathrm{Var}(f_t/f)$; use of approximation $\hat{f}_t$ based on the transition kernel
Lower Bounds	discretisation of $(\theta^{(t)})$ into a two-state Markov chain; Spearman test; bounds on the Total Variation distance
	Stationarity: *Does* $(\theta^{(t)})$ *cover the support of* f*?*
Graphical Approaches	raw plots and associated graphs (single and parallel runs); plots based on renewal theory using discretised chains
Lower Bounds	based on time series spectral methods or on coupling theory
Stopping Rules	comparison of variance within and between parallel chains; nonparametric (eg, Kolmogorov-Smirnov or Kuiper) tests
Control Variates	based on Riemann sums, lower bounds on the probability mass of the MCMC sample, or divergence assessment
	Estimation: *Can* $(\theta^{(t)})$ *adequately estimate* $\mathbb{E}_f[h(\theta)]$*?*
Stopping rules	tests based on spectral estimation, taking $(\theta^{(t)})$ as a time series; comparison of Rao-Blackwellised, importance sampling and Riemann estimators
Variance Assessment	comparison of variances within and between parallel chains; estimation based on renewal theory
Asymptotic Normality	appeal to the CLT to construct parametric and nonparametric tests and plots

4.1 Graphical Methods

Trace (raw) plots, cumulative plots, autocorrelation and crosscorrelation plots, histograms and Box plots have (rightly) been a favourite among diagnostic tools. While they provide a fast and easily interpretable impression of convergence, they are nonetheless prone to misdiagnosis. Their use in conjunction with other tools with stronger theoretical background is thus highly recommended. On the reverse, note that that most diagnostics can lead to graphical implementations, as will be seen in the following paragraphs.

A particular graphical method has been devised by Yu (1995) and Yu and Mykland (1998), in connection with the convergence to $\mathbb{E}_f[h(\theta)]$ of the empirical average (3) for a given function h, through cumulative sums (CUSUM)

$$C_T^i = \sum_{t=1}^{i} [h(\theta^{(t)}) - S_T], \qquad i = 1, \cdots, T \tag{4}$$

where S_T is defined in (3). The authors graphically evaluate the mixing speed of the chain or assess the correlation between the $\theta^{(t)}$s on the CUSUM: a highly irregular graph concentrated around 0 indicates good mixing properties. The output is sometimes compared with a CUSUM graph based on an *iid* normal sequence. A quantitative measurement of the "hairiness" of the CUSUM plots, D_T, is

$$D_t = \frac{1}{t}\sum_{t=1}^{t} d_i$$

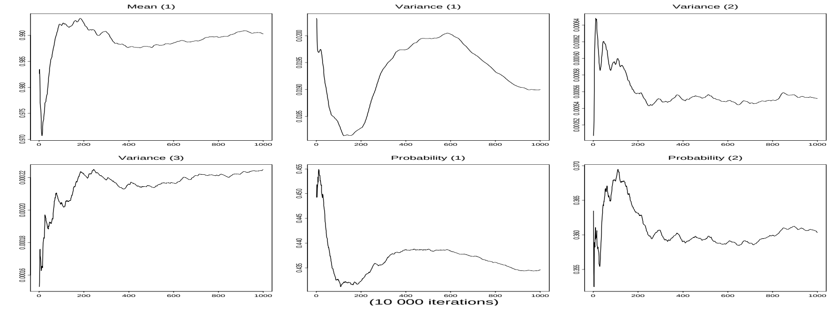
where d_i is 1 if the CUSUM plot has a local minimum or maximum and 0 otherwise. A D_t of zero therefore indicates a totally smooth plot and a value of one indicates maximum "hairiness". Confidence intervals can be calculated assuming the sequence d_i is *iid* and distributed symetrically about the mean,

$$\frac{1}{2} + Z_{\alpha/2}\sqrt{\frac{1}{4t}}.$$

Chains with good mixing stabilise within, say, 95% confidence limits. Brooks (1998) extends this approach by constructing stopping rules based on rank statistics. More advanced evaluations are possible, as shown by Heidelberger and Welch (1983) (see §4.5).

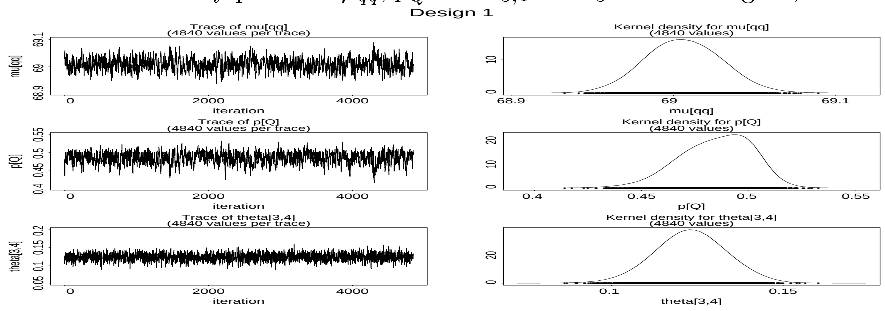
Example 1 Simple assessments of convergence for the mixture model problem do not work well. For instance, plots of the cumulative averages of the parameters may fail to settle (Figure 4), even though the corresponding estimates of the densities are stable (Figure 1). This phenomenon exhibited in Robert and Mengersen (1998) can be blamed on the weak identifiability of mixtures where different parameters may produce almost identical distributions. This phenomenon does not appear for hidden Markov models, though.

Figure 4: Evolution of the estimates of some parameters of the 3 component mixture for the dataset of Izenman and Sommer (1988). (*Source:* Robert and Mengersen, 1998.)



Example 2 The problem of misleading graphical diagnostics described in the previous example did not surface in the QTL analysis. For the "well behaved" Design 1, trace and kernel density plots for selected parameters over iterations of the saved chain visiting the true model (α_3) are shown in Figure 5. Similar plots for the "poorly behaved" Design 2, are shown in Figure 6. Both designs use a thinning interval of 10 to overcome storage problems and substantial autocorrelation to lag 20. The corresponding cumulative averages were completely stable after 100 and 5000 iterations under Design 1 and 2 respectively.

Figure 5: Trace and kernel density plots for μ_{qq}, p_Q and $\theta_{3,4}$ for α_3 under Design 1, obtained by `CODA`.



Figures 7 and 8 illustrate the typical relationship between the autocorrelations and the cusum path of all parameters. The larger the autocorrelations, the wider the range and the "stickier" the path of the cusums C_T (solid line) relative to the cusum path of an *iid* random variate (dotted line). Plots of C_T and D_T shown on the left of Figure 8 were derived under Design 2 and illustrate poor mixing. This was observed for all parameters in both designs, but of more concern for a greater number of parameters under the second design. Tighter cusum plots and improved D_T statistics were achieved only after thinning the chain (right of Figure 8). The amount of thinning was chosen as the most substantial autocorrelation over all parameters, using visual assessment; 10 and 15 were used for Designs 1 and 2 respectively. Once correlation in the chains is broken in this manner, convergence for all parameters is reached after 1,900 iterations under Design 1. These chain lengths apply after imposing burnins recommended by the methods of §4.5. The slowest parameters to converge under Design 1 for this method are μ_{qq}, $\theta_{2,3}$, $\theta_{3,4}$, $p_{3,1}$ and $p_{4,2}$. For Design 2, the D_T statistics for σ_R, $\theta_{2,3}$ and $\theta_{3,4}$ fall just short of the confidence limits, although their cusum plots look fine. While this was remedied by increasing the thinning parameter to 20, μ_{QQ} now had not converged. Brooks (1998) describes an extension of this statistic which estimates the burnin and thinning parameters, rather than estimating them *a priori*.

Example 3 Figure 9 illustrates the evolution of the simulated parameters via raw plots. As discussed above, this type of graph is difficult to interpret. In the present setup, it does not exhibit strong convergence problems, even though some transition rates are markedly slower than others. This is in contrast with complementary diagnostics as discussed in the following subsections. The autocorrelations of the parameter chains, given in Figure 10, are reasonable over $10,000$ iterations.

Example 4 As indicated in §2.4, the first simulated example has good mixing properties, as backed by the CODA output graph (not represented here). On the opposite, the second sample leads to a very slow mixing for the same MCMC algorithm. Figure 11 illustrates this feature with very irregular (raw) sample paths for the three parameters and, correlatively, heavy tailed estimations of the posterior densities. This is a case where the raw plots are sufficient to signal the poor properties of the MCMC algorithm but, as seen below, every indicator turns red in this case.

4.1.1 Graphical Methods for Convergence of Model Choice

Choice of models may be considered as a variable in an MCMC analysis. To date, diagnostics for convergence of these special chains have received little attention. There is a number of issues associated with such chains that deserve special attention. First, while by design the chain is made to traverse across the entire (relatively small) discrete model space, it is not imperative for it to "settle down" into a particular model, so it is not always clear what is meat by "convergence". Second, although very long chains may be run, the number of visits to a particular state may be quite small, especially after thinning and burnin if these are imposed, providing comparatively little basis for diagnostics of each state. This might be exacerbated if the MCMC algorithm is designed to allow for stabilisation within a state before collection of values, thus sacrificing stability over the state space of models for stability of variables within states. Third, the very nature of small discrete spaces precludes the use of many of the available diagnostics. (See Brooks and Giudici, 1998, for proposals in this direction.)

In Example 1, such chains may arise through the assessment of the number of components of a mixture model, via reversible jump; see Richardson and Green (1995) for this problem in detail. As illustrations of possible diagnostics to this situation we consider here the QTL problem.

Example 2 QTL localisation as parametrised by George et al (1999) has the additional problem of monitoring the choice of locus order $\underline{\alpha}$. For a four marker-one QTL problem, five locus orders ($\alpha_1 = QMMMM, \ldots, \alpha_5 = MMMMQ$) are possible. Under a reversible jump algorithm, the posterior probabilities for the candidate locus orders were respectively (0.000, 0.002, 0.998, 0.000, 0.000) under Design 1 and (0.157, 0.248, 0.312, 0.141, 0.142) under Design 2. After thinning, only 4,840 iterations from a chain of 50,000 in Design 1 and 28,120 from 900,000 in Design 2 are available to monitor the convergence of parameters within locus order α_3. Although both designs identified the true gene order with highest posterior probability, assessment of stationarity across the model space was in many respects more difficult for Design 2. Under Design 1, a particular locus order was very quickly chosen and, apart from a few flyspots on the trace plot, it remained predominant for most of the simulation. Under Design 2, the chain kept jumping

Figure 6: Trace and kernel density plots for μ_{Qq}, p_Q and σ_R for α_3, under Design 2, obtained by CODA.

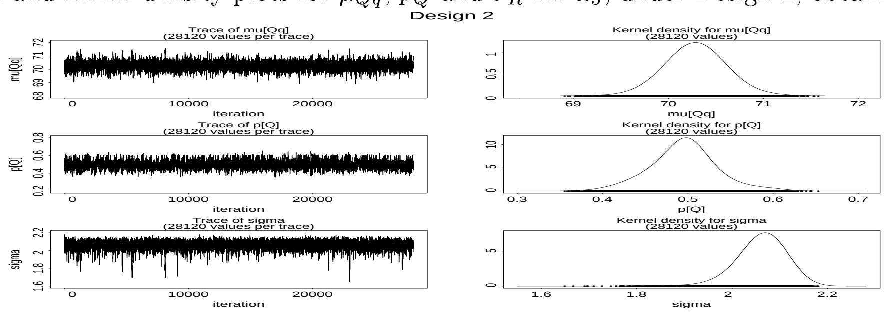


Figure 7: Autocorrelation plot for σ under Design 2.

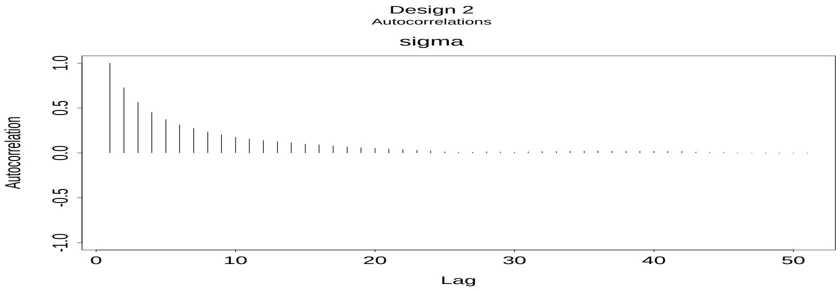


Figure 8: CUSUM plots for σ under Design 2 (solid lines) and an *iid* random variate (dotted lines) before and after thinning.

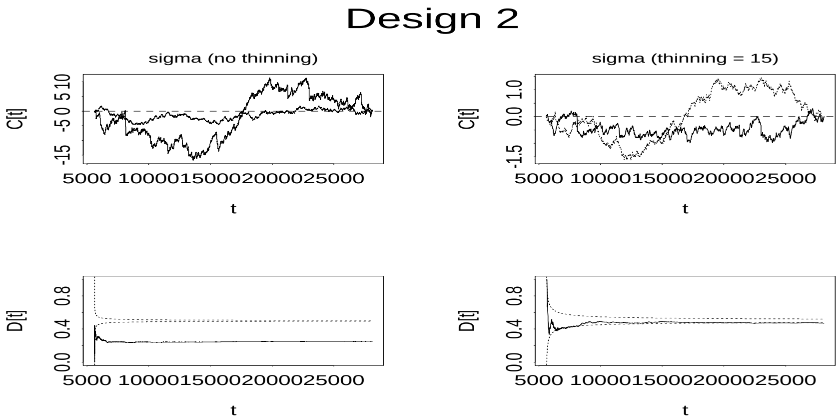


Figure 9: Plots of the simulation output for the parameters of interest in the HIV model, based on 10,000 iterations, obtained by `CODA`. (*Source:* Guihenneuc–Jouyaux *et al.*, 1998.)

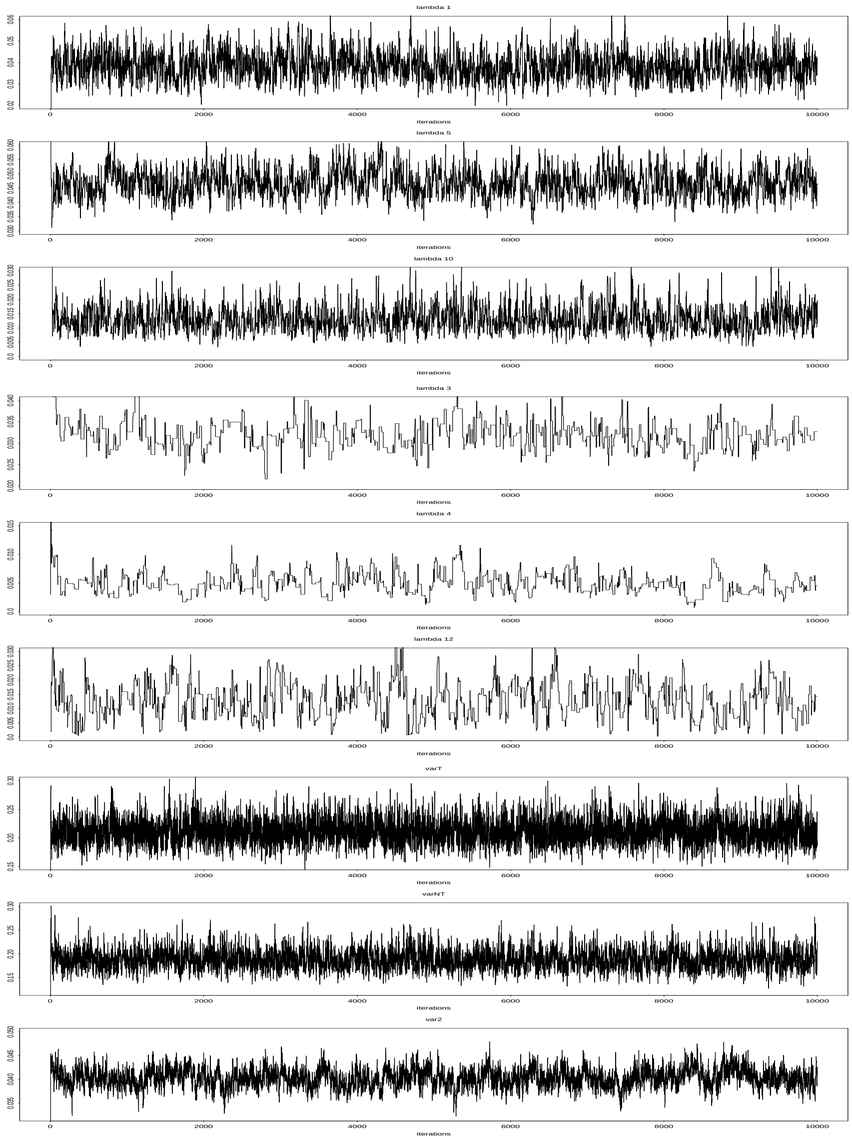


Figure 10: Autocorrelations for the parameters of interest in the HIV model, based on 10,000 iterations, obtained by CODA. (*Source:* Guihenneuc–Jouyaux *et al.*, 1998.)

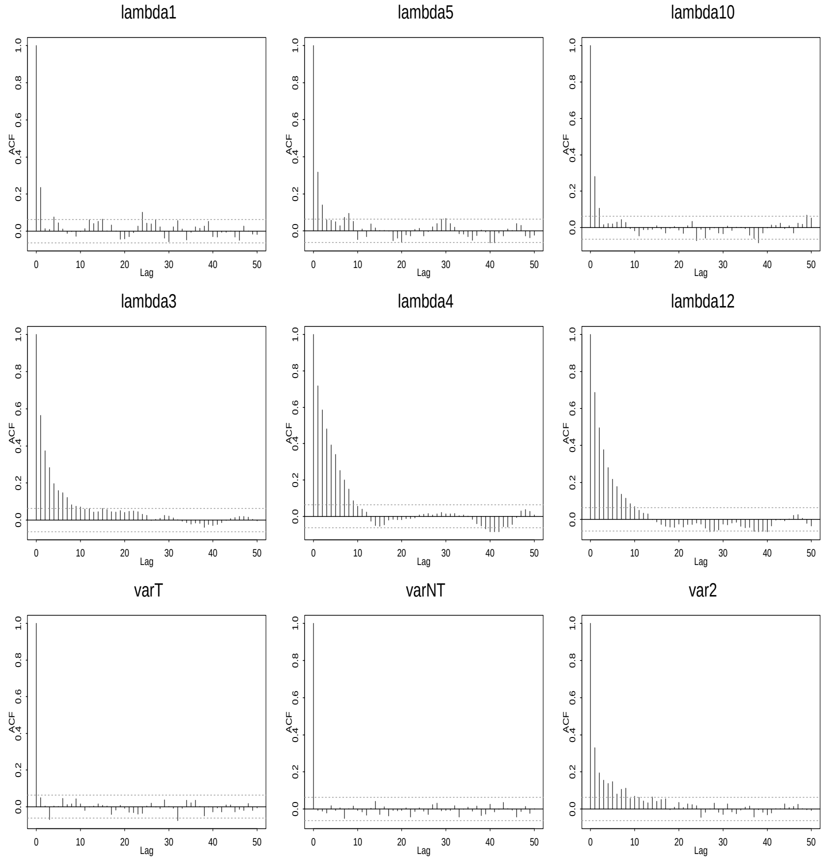


Figure 11: Output of the Metropolis–Hastings algorithm for the stochastic volatility model (25) and density estimation, based on 100,000 iterations, for $\alpha = 0$, $\beta = .7$ and $\sigma = .1$, obtained by CODA.

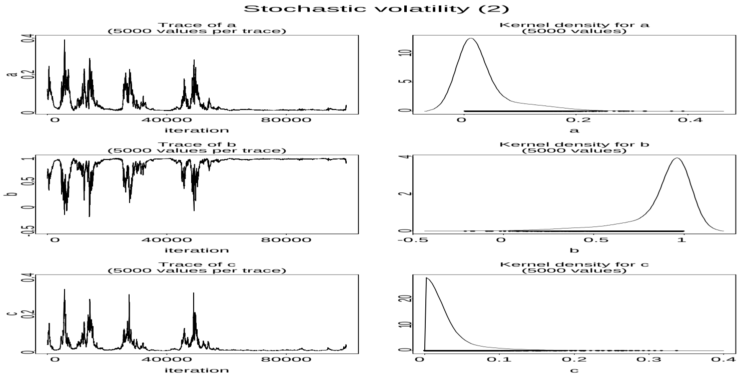
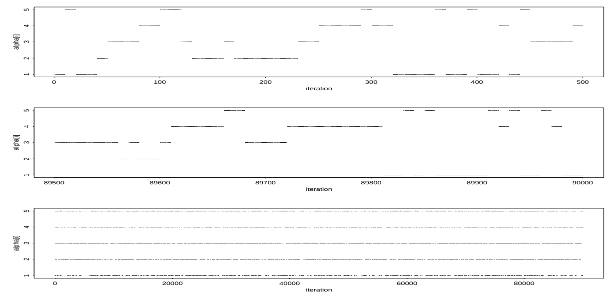


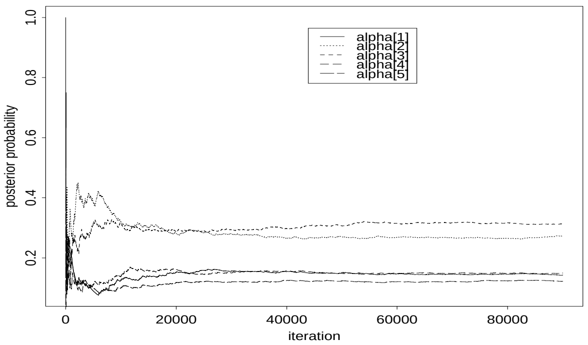
Figure 12: Trace plots of the locus order α_l, $l = 1, .., 5$, under Design 2, for the first and last 500 iterations and all 90,000 iterations, obtained by `CODA`.



between models. Figure 12 plots a trace of the locus orders, generated under the reverse jump sampler, for this latter design, based on the last 22,500 iterations.

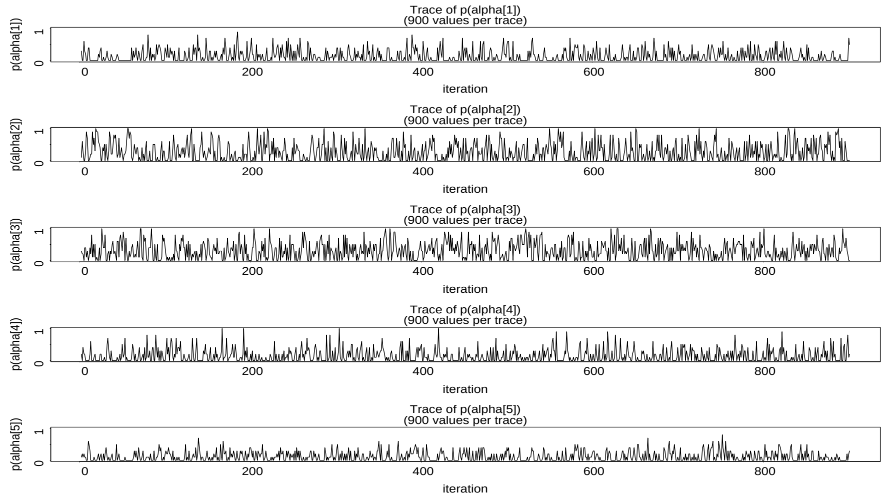
Alternative ways of monitoring convergence of $\underline{\alpha}$ under Design 2 were explored. For example, a transition probability matrix was constructed based on the full chain. A positive indication of stationarity was observed by comparing equivalent matrices based on the first and last 10% of the chain. Although a natural statistic to monitor is the posterior probability of a particular locus order, $\hat{p}(\alpha_i)$, diagnostics on this may be misleading since the increasing number of cycles necessitates decreasing variance in the statistic. The trace plots in Figure 13 indicate the degree of nonstationarity for Design 2 using this statistic.

Figure 13: Trace plots of the posterior probabilities of each locus order, $\hat{p}(\alpha_i)$ under Design 2, obtained by `CODA`.



One might instead construct bounds for the trace plot; these may be chosen as upper and lower percentiles of a large number of evolutions of $\hat{p}(\alpha_i)$ based on randomisations of the chain. The above statistic fails against this comparison. An alternative is to block the chain into cycles of 100 iterations and monitor the proportion of times, $p(\alpha_i)$, that the chain is in model α_i within these blocks. Trace plots for this modified chain are given in Figure 14. While these "flypaper" plots are difficult to interpret, diagnostics described in the following sections (using the logit transformation when appropriate) acknowledge much greater stationarity. Perhaps the appearance of convergence is expected, however, with the highly variable, almost random choice of locus order from the small discrete space. Even so, not all diagnostics (see §4.9) will firmly support or deny convergence with a chain of only 900 bins. The D_T statistics for example (not shown here) are poor indicators since the confidence limits are too wide for practical interest. The value of the binning approach has a limitation, however, since it is dependent on the size of the block and the length of the chain. Ironically, a very long chain was required in order to obtain sufficient values of the sub-chain on which to perform some of the diagnostics.

Figure 14: Trace plot of $p(\alpha_i)$ under Design 2, computed for blocks of 100 cycles, obtained by `CODA`..



4.2 Rao-Blackwellisation

Rao-Blackwellisation was introduced by Gelfand and Smith (1990) as a way of improving on the standard estimates (3) by conditioning. For instance, in a Gibbs setting, when $\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)})$, and $h(\theta) = h(\theta_1)$, S_T can be replaced by

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[h(\theta_1)|\theta_2^{(t)}] \ .$$

A similar construction can be derived for general Metropolis–Hastings algorithms, by conditioning on the simulated variables and integrating out the uniforms, as shown by Casella and Robert (1996). An additional use of Rao-Blackwellisation is to estimate marginal posterior distributions as

$$\frac{1}{T}\sum_{t=1}^{T} \pi_1(\theta_1|\theta_2^{(t)}) \tag{5}$$

when the full conditional distribution π_1 is available. This can be used in the Riemann control variate technique of §4.3.

Example 1 Although Rao-Blackwellisation can be implemented for the mixture model problem, the comparison of usual estimates and Rao-Blackwell corrections brings very little information as whether convergence has been achieved since both estimates are indistinguishable from the start (see Robert and Mengersen, 1998, for illustrations). This is a most common occurrence in low dimensional problems and suggests the use of more powerful alternatives to (3), as in the next paragraph, which make use of Rao-Blackwellisation in a non-parametric way, that is through (5).

Example 2 The Rao-Blackwellised estimate of S_T for μ_i, for which we know the full conditional, is

$$\frac{1}{T}\sum_{t=1}^{t=T} \frac{N^{(t)}\bar{Y}_i^{(t)}}{N^{(t)}+\frac{1}{k}}$$

where $N^{(t)}$ is the number of offspring (over all families) allocated genotype i in the t^{th} iteration, $\bar{Y}_i^{(t)}$ is the average of their phenotypes and k is large. Plots of these (not shown here) support the conclusions reached for Example 1, namely that the estimates are indistinguishable from the unmodified cumulative averages. In one sense this is to be expected since the unmodified estimates are so stable; in another sense it is heartening that in such a complex application of mixture models this phenomenon is observed.

4.3 Riemann Control Variates

Philippe (1997a,b) bases her convergence diagnostics on Riemann sums, i.e. on the analytic representation of integrals as

$$\int h(\theta)f(\theta)d\theta = \lim_T \sum_{t=1}^{T} (\theta^{[t]} - \theta^{[t-1]})h(\theta^{[t]})f(\theta^{[t]}),$$

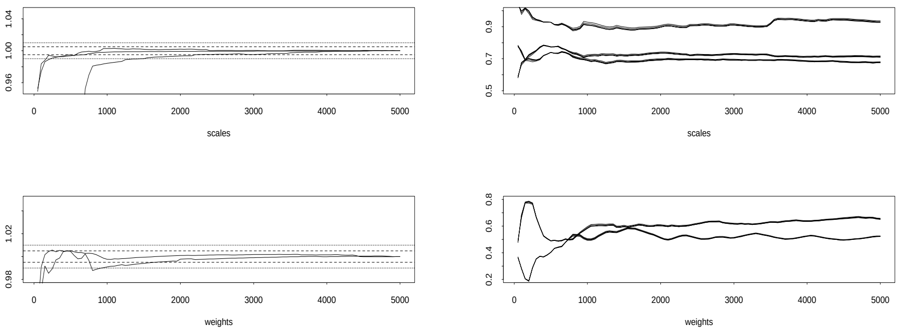
where the $\theta^{[t]}$'s are the order statistics for a sample from f or from another distribution, dependent or not. This approach performs well mainly in dimension 1, although in higher dimensions p the method can be applied to integrals depending only on a univariate marginal of f. Philippe and Robert (1998) develop an extension through a Rao-Blackwellised estimate of the marginals based on the conditional densities. When these are available there are in addition p control variates for each function of interest h. In particular, for $h(\theta) = 1$, these variates can be monitored until they all fall within a reasonable distance of 1.

Example 1 Figure 15 exhibits the result of a Riemann sum diagnostic for a 3 component exponential mixture,

$$q_0\mathcal{E}xp(\lambda_0) + (1-q_0)q_1\mathcal{E}xp(\lambda_0\tau_1) + (1-q_0)(1-q_1)\mathcal{E}xp(\lambda_0\tau_1\tau_2) .$$

The graphs on the left describe the convergence to 1 of the control variates for λ_0, $\lambda_1 = \lambda_0\tau$, (top) and q_0, q_1 (bottom). It thus takes under 2500 iterations to reach the 1% error band. The graphs on the right give the convergence curves of the different Rao-Blackwellised Riemann sum estimates of the five parameters of the mode, with a strong similarly between these. Note however, that the estimators have not stabilised, as shown by the jump in one scale curve, shortly before 4000 iterations. (See Philippe and Robert, 1998, for details.)

Figure 15: *Left:* Riemann control variates for the different parameters of a 3 component exponential mixture, with 1% and 5% error bounds (250 observations). *Right:* Convergence of the Riemann estimators for the different parameters. (*Source:* Gruet *et al.*, 1999.)



Example 2 Riemann sums are possible in this example, and their application is anticipated to be very useful as a comparative measure with unmodified and Rao-Blackwellised averages. Its application was almost impractical here due to the complexity of the model, the length of the chains and the manipulation required

of the chains, although the latter might have been avoided had the necessary elements been captured during the MCMC analysis. This diagnostic may hence become more accessible with the availability of appropriate software.

4.4 Diagnosis through Discretisation

Discretised chains afford explicit analysis of convergence. Raftery and Lewis (1992) reduce $\{\theta^{(t)}, t = 1, , ..T\}$ to a (pseudo) two-state Markov chain of the form $z^{(t)} = \mathbb{I}_{\theta^{(t)} \leq \underline{\theta}}$, where $\underline{\theta}$ is arbitrary. The authors then derive some approximations of the warming time t_0 and of the convergence time T from the transition matrix of $Z^{(t)}$,

$$\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

(see also Brooks and Roberts, 1997).

Guihenneuc–Jouyaux and Robert (1998) propose a discretisation based on several disjoint small sets A_i $(i = 1, .., k)$, such that $s_i(\theta) = \mathbb{I}_{\theta \in A_i}$ satisfies a corresponding regeneration equation, $P(\theta^{(t+1)} \in B|\theta^{(t)}) \geq s_i(\theta^{(t)})\nu_i(B)$, where ν_i is a probability measure and s_i a positive function. Subsampling is then undertaken at renewal times τ_n $(n > 1)$, which are defined as the successive instants when the Markov chain enters one of these small sets with splitting. Guihenneuc–Jouyaux and Robert examine the mixing rate of $(\xi^{(n)})$ through the difference in the number of visits N_j to a given state j from two different starting points u and v, based on the *divergence*

$$\mathrm{div}_j(u, v) = \lim_{n\to\infty} \mathbb{E}_u[N_j(n)] - \mathbb{E}_v[N_j(n)]\,. \tag{6}$$

The divergence estimates can be plotted, along with confidence intervals, for various sets of states j, u, v, derived from only two independent parallel chains, until all of these estimates stabilize. Unfortunately, renewal methods are difficult to implement in missing data setups such as those found in the examples of this paper.

Example 1 In the analysis of a normal hidden Markov mixture, Raftery and Lewis' assessment of convergence is somewhat ambivalent, suggesting between $5,000$ and $200,000$ iterations to reach convergence. This is, indeed, expected by the nature of the model and is supported by the contradiction in estimates given by other methods, as described below.

Example 2 Raftery and Lewis' diagnostic applied to within model convergence suggests nonconvergence after 5,000 iterations for parameters displaying substantial autocorrelations at lag 10 (lag 20 before thinning), namely p_Q, μ_{Qq} and μ_{qq} under Design 1 and σ_R, p_Q, μ_{QQ}, μ_{Qq}, μ_{qq}, $\theta_{2,3}$ and $\theta_{3,4}$ under Design 2. The worst offenders of nonconvergence are p_Q under Design 1 (14,000 iterations) and μ_{QQ} and σ_R under Design 2 (34,000 iterations), which agree with other diagnostics (see the comparison in Tables 2-4), although Raftery and Lewis' recommended chain lengths of 14,000 and 34,000 respectively are much longer. On the other hand, burnin periods of at most 35 seem cavalier, especially in light of the results of §4.5. The chain for the (blocked) probability of being in model α_l is shorter than the minimum of 4,000 blocks (or 4,000,000 iterations in the original chain) needed to assess convergence using Raftery and Lewis' method.

Example 3 Raftery and Lewis' evaluation of convergence for this problem suggests between 10 and 200 iterations for the warm-up time and between 4000 and $110,000$ iterations for convergence time.

Example 4 The diagnostic of Raftery and Lewis gives at most $20,000$ as the total number of iterations required for convergence of the first stochastic model. For the second model, however, this requirement is over a million. This is in line with expectations but the actual cycle times are quite different from those proposed by Heidelberger and Welch (1983), for example (see the following subsection).

4.5 Spectral Analysis

Heidelberger and Welch (1983) and Geweke (1992), among others, consider $(h(\theta^{(t)}))$ as a time series and derive diagnostics from the corresponding spectral density $\psi_h(w)$ which gives the asymptotic variance $\psi_h^2(0)$ of the empirical average (3). Resorting to batching, Geweke (1992) takes the first T_A observations and the last T_B observations and computes the partial expectations

$$\delta_A = \frac{1}{T_A} \sum_{t=1}^{T_A} h(\theta^{(t)}), \qquad \delta_B = \frac{1}{T_B} \sum_{t=T-T_B+1}^{T} h(\theta^{(t)}) \tag{7}$$

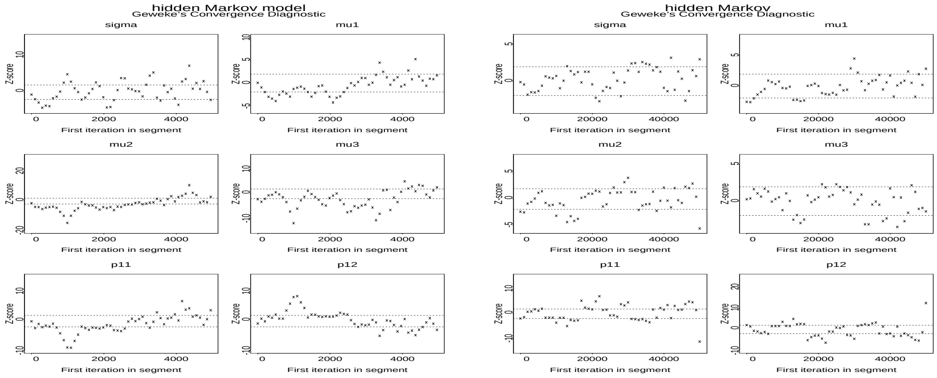
and the corresponding estimates σ_A^2 and σ_B^2 of $\psi_h^2(0)$. The convergence diagnostic is based on the fact that the standardized difference between δ_A and δ_B is (asymptotically in T) a standard normal variable. Heidelberger and Welsh (1983) use

$$B_T(S) = \frac{S_{[Ts]} - [Ts]\bar{\theta}}{(T\hat{\psi}_h^2(0))^{1/2}} , \qquad 0 \leq s \leq 1 ,$$

where $S_t = \theta^{(1)} + \ldots + \theta^{(t)}$ and $\bar{\theta} = S_T/T$, and $\hat{\psi}_h^2(0)$ is an estimate of $\psi_h^2(0)$, is approximately a Brownian bridge. (Note the connection with the CUSUM technique of §4.1.) Both these methods may be implemented through `CODA`

Example 1 In the analysis of a normal hidden Markov chain the inference afforded by these two methods is quite contradictory despite the foundational similarity of their approaches. Geweke's method gives an acceptance of stationarity for most parameters after $5,000$ iterations (Figure 16), but Heidelberger and Welch's rejects convergence of most parameter chains for the same number of iterations. Moreover, Geweke's estimate is the minimum of that supported by Raftery and Lewis's diagnostic. This difference may be due to the cruder estimation of the variance in Geweke's approach.

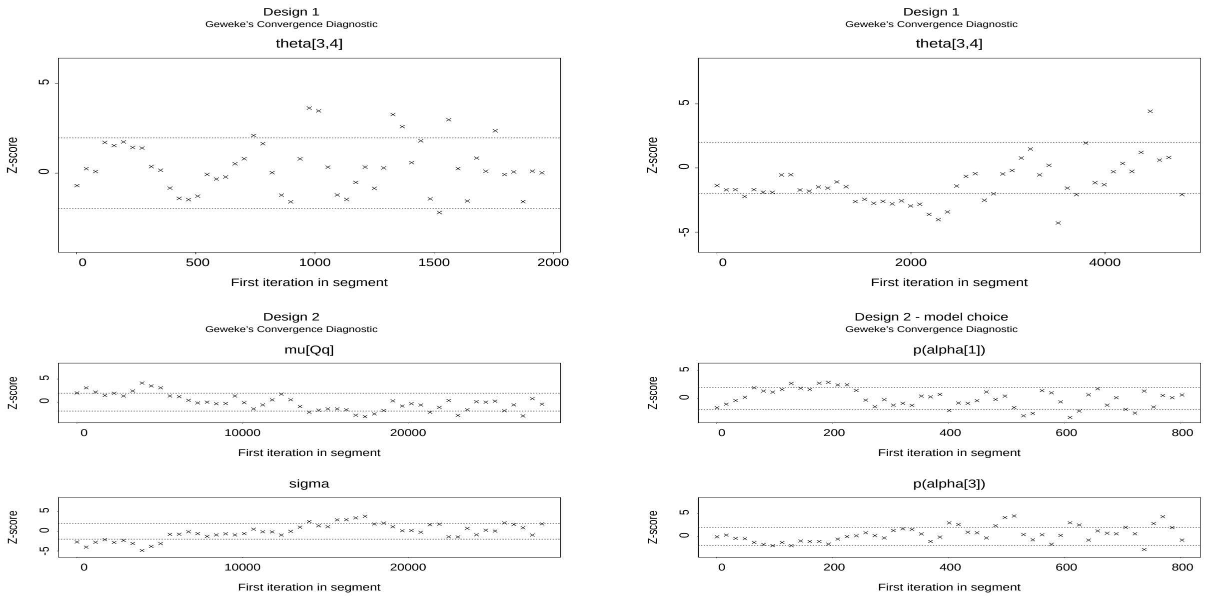
Figure 16: Geweke's diagnostic for a hidden Markov model, obtained by `CODA`, for 5000 and $50,000$ iterations.



Example 2 When the chain of 4,800 iterations under Design 1 is analysed, both Heidelberger and Welch's (1983) and Geweke's (1992) methods agree on convergence for all parameters, with a burnin of 2,400 necessary for $\theta_{3,4}$. These methods also agree when as few as 2,000 iterations are analysed, but conclude no burnin is necessary for any parameters. Similar contradictions also hold for both methods under Design 2. When fewer than 5,000 iterations are analysed, all parameters converge and burnins are only indicated for $p_{1,2}$, which shows no concern under any other diagnostics. Yet when 28,000 iterations are analysed, all parameters are passed with a burnin of 5,600 for σ_R and 2,800 for μ_{QQ}. Figure 17 plots Geweke's statistic for various parameters and chain lengths. Although these methods may be acknowledging stationarity too early, Gelman

and Rubin's method in §4.6 is not as reassuring. Based on this diagnostic, burnin lengths of 2,400 and 5,600 were used for assessing estimation properties under Designs 1 and 2 in §4.1, 4.8 and 4.9. The probability of being in locus order α_i under Design 2 passed Geweke's and Heidelberger and Welch's methods for 900 blocks.

Figure 17: Geweke's diagnostic for various parameters and chain lengths in the QTL examples, obtained by `CODA`.



Example 3 The two methods of Geweke and Heidelberger and Welch agree for the HIV analysis. Geweke's diagnostic (Figure 18) indicates that $10,000$ iterations are sufficient, and Heidelberger and Welch's diagnostic also gives a positive signal for the 19 parameters of interest after $10,000$ iterations.

Example 4 For the first stochastic volatility analysis, both Geweke's and Heidelberger and Welch's criteria pass β and σ after 5000 iterations.

4.6 Within and Between Chain Variance Comparison

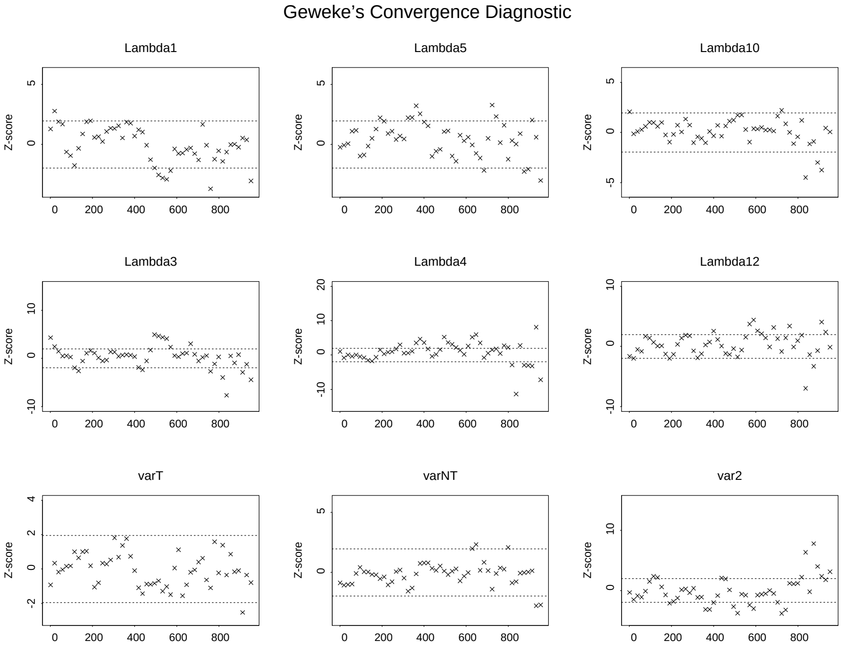
Gelman and Rubin (1992) derive a starting distribution μ related to the modes of f. With the possible addition of an importance sampling step (SIR), they then generate M chains $(\theta_m^{(t)})$ $(1 \le m \le M)$. For any function of interest h, a stopping rule is derived from the difference between *within* and *between* variance estimators,

$$B_T = \frac{1}{M} \sum_{m=1}^{M} (\overline{\xi}_m - \overline{\xi})^2 \ , \quad W_T = \frac{1}{M} \sum_{m=1}^{M} s_m^2 = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{T} \sum_{t=1}^{T} (\xi_m^{(t)} - \overline{\xi}_m)^2 \ ,$$

with $\xi_m^{(t)} = h(\theta_m^{(t)})$, $\overline{\xi}_m = (\xi_m^{(1)} + \ldots + \xi_m^{(T)})/T$, and $\overline{\xi} = (\overline{\xi}_1 + \ldots + \overline{\xi}_M)/M$. The posterior variance of ξ, estimated by

$$\hat{\sigma}_T^2 = \frac{T-1}{T} W_T + B_T \ ,$$

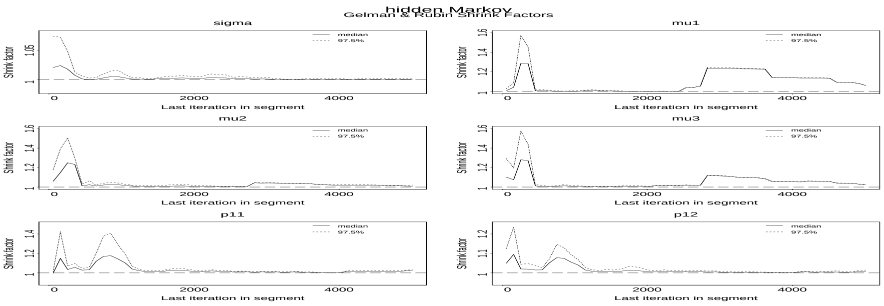
Figure 18: Geweke's diagnostic plot for the parameters of interest in the HIV model, based on $10,000$ iterations, obtained by `CODA`. (*Source:* Guihenneuc–Jouyaux *et al.*, 1998.)



may be compared with W_T, which is asymptotically equivalent, graphically or through the use of a t test. (See also Brooks and Gelman, 1998, for multivariate extensions.)

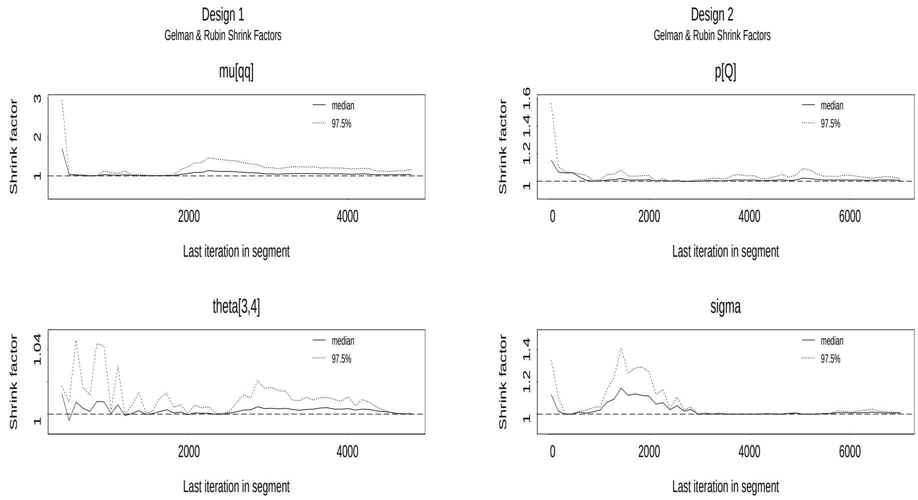
Example 1 Gelman and Rubin's output for the analysis of a normal hidden Markov chain requires close to $50,000$ iterations in the case of the first mean, μ_1 (Figure 19), which is an order of magnitude larger than that required by Geweke's and at the upper end of that recommended by Raftery and Lewis'. Note the lack of monotonicity of the graph around 2500 iterations for μ_1, μ_2 and μ_3, which is a frequent occurrence for this method.

Figure 19: Gelman and Rubin's (1992) shrink factors for an hidden Markov model, obtained by `CODA`.



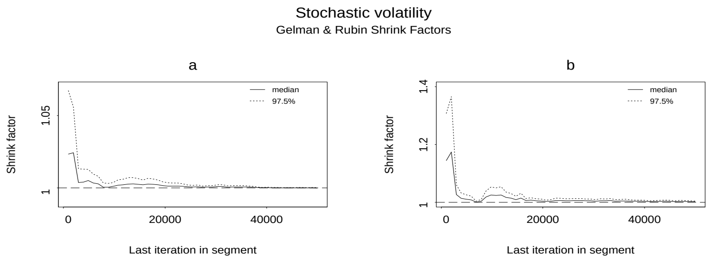
Example 2 Gelman and Rubin (1993) shrink factors are plotted in Figure 20. Most parameters settle down by 4,000 iterations under Design 1, although μ_{QQ}, μ_{qq} and especially θ_1 (not shown) do not settle, indicating a chain length of 4,800 is insufficient. With only 7,020 iterations available for the second chain in Design 2, it is not unexpected that μ_{qq} and p_Q among other parameters have not stabilised.

Figure 20: Gelman and Rubin's (1992) shrink factors for the QTL examples, obtained by `CODA`.



Example 4 As anticipated, for the first simulation under the stochastic volatility model, the diagnostics are almost all unanimous in their conclusion of convergence after about ten thousand iterations, as shown by Figure 21. For the second simulation, the diagnostics pointed to the lack of stability and high correlations.

Figure 21: Gelman and Rubin's (1992) shrink factors of α *(a)* and β *(b)* for a stochastic volatility model (25), with $\alpha = 0$, $\beta = .9$ and $\sigma = .5$, obtained by `CODA`.



4.7 Allocation Maps

A graphical assessment of the stability of a discrete valued parameter $(z^{(t)})$ over the course of the simulation $t = 1, .., T$, may be undertaken via an allocation map proposed by Robert (1997). For example in the latent variable expression of a mixture model (see Section 2), $z_i^{(t)} = j$ indicates allocation of the ith observation to the jth component of the mixture at cycle t. The map then depicts the values of j as different "colours" (eg, black, white, grey) and plots $z_i^{(t)}$ for all (i, j). This representation can also be extended to continuous setups with missing data, as in Example 4. In this case, the levels of gray correspond to a discretization of the latent variable.

Example 1 In the analysis of the same data as in Figure 1, using a three component mixture model, Figure 22 provides a description (or *map*) of the evolution of the allocations for the 485 points of the sample, each component being represented by a different grey level. The left map corresponds to 5000 iterations of the Gibbs sampler, with a random start and a large precision $\zeta^{-1} = 1$. The sudden modification on the upper part of the image shows that stationarity has not been achieved. On the contrary, the stable bands of the image on the right are indicative of a strong stability. The corresponding mixture estimates on top of both images confirm this assessment. This graphical check is not always informative, though, as shown by Figure 23, which represents allocation maps for various numbers of iterations in the exponential example of Gruet *et al.* (1999). Exponential mixtures seem so weakly identifiable that most observations cannot afford a stable allocation structure and the allocation map cannot be used as a convergence diagnostic. In general, there are difficulties with the interpretation of such maps, whose stability depends on the choice of the prior distribution (see Robert, 1997, vs. Richardson and Green, 1997).

Figure 22: Allocation map for the dataset of Izenman and Sommer (1988) and a 3 component normal mixture model. The shades correspond to the three components *(white is for* 1*, grey for* 3 *and dark for* 2*)*. (*Source:* Robert, 1997.)

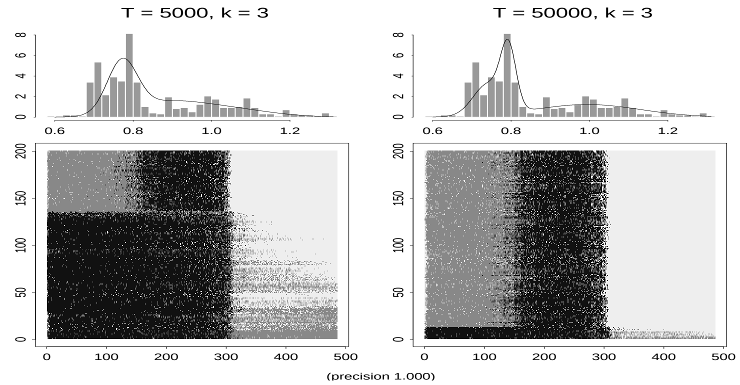
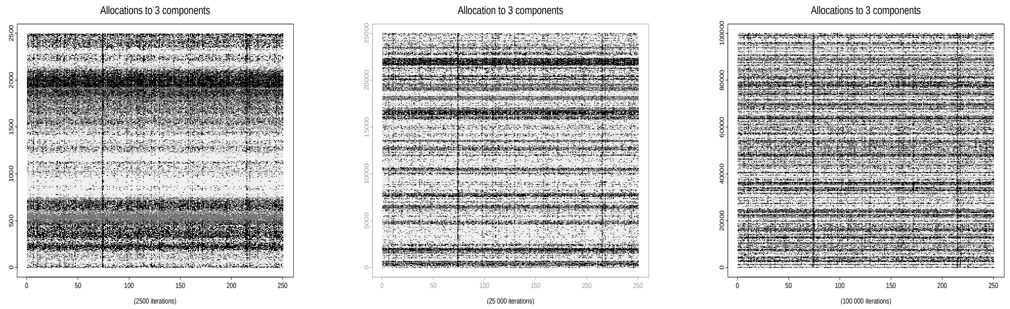


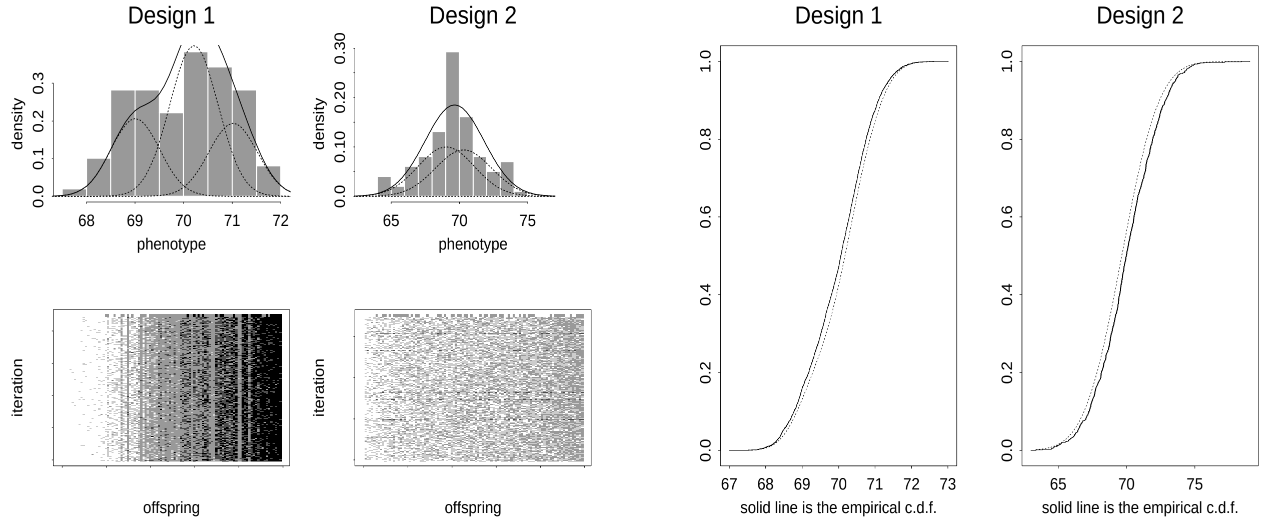
Figure 23: Allocation maps for a 3 component exponential mixture. (*Source:* Gruet *et al.*, 1999.)



Example 2 Figure 24 contains the allocation maps for a random sample of 100 (of the 200) offspring for the first family in both designs. Offspring are plotted in order of their phenotypes. Design 2 had more trouble allocating offspring to genotypes because of the larger variance associated with the model. Note also, the first sire in this design is almost always allocated a homozygous genotype and hence can only give rise to offspring of genotype Qq or qq. Stability of the allocation does not change over the length of the chains, and so density plots of the mixture look the same (for both designs) whether 1000 or all iterations were used. The empirical and estimated cumulative density functions are also a close fit.

Example 3 Figure 25 shows another use of the allocation maps in the follow-up study. It presents the successive states of several patients along the MCMC iterations and points out patients with special features, namely those with very little variability in the allocations, like Patient 15, or those with a very unstable

Figure 24: Density plots, allocation maps and cumulative density plots for the QTL examples. The shades correspond to the three QTL genotypes *(white is for qq, grey for Qq and dark for QQ)*.



pattern, like Patients 1 and 10, or yet patients with a stable progression in the disease states, like Patient 18. These maps provide here a practical tool to understand better the origin of the within individual variability. In fact, unstable patterns lead to select a subgroup of patients who at first progress to higher states and after come back to states of less gravity (like Patients 1 and 10). Moreover, the existence of such trajectories motivates the introduction of back flows between states in the Markov process.

Example 4 For the first simulation under the stochastic volatility model, Figure 26 reproduces the allocation maps. As with the other diagnostics, there is a show of strong stability in the successive values of the y_t^*'s for this case. The allocation map for the second simulation is given in Figure 27. This is also informative about the poor mixing anticipated for this case, with wide areas of constant values for the simulated y_t^*'s.

4.8 Assessment of Asymptotic Normality

The Central Limit Theorem provides the asymptotic distribution for the standardized version of (1). The conditions under which this holds, however, are often difficult to confirm (see, e.g., Roberts and Rosenthal, 1998) and the consequent estimates are difficult to compute. Some practical convergence diagnostics are nonetheless based on manageable versions of the Central Limit Theorem.

Using a discretisation modelled on that of Raftery and Lewis for a collection of sets A, Chauveau and Diebolt (1997) proposed a Shapiro-Wilk normality test on parallel chains. A general alternative developed by Robert *et al.* (1999) involves subsampling a single chain at increasingly distant times t_k, the latter being generated by a Poisson distribution. If N_T denotes the number of sampling times before T and with estimates of $\mathbb{E}^f[h(\theta)]$ and $\mathrm{var}^f(\mathrm{h}(\theta))$, $\widehat{\mu}_T$ and $\widehat{V}_T$, taken from the whole chain, the normalized sum

$$S_T = \sum_{k=1}^{N_T} (h(x^{(t_k)}) - \widehat{\mu}_T) \Big/ \sqrt{N_T\, \widehat{V}_T}$$

is asymptotically $\mathcal{N}(0,1)$. Usual tests of normality such as Kolmogorov-Smirnov or Shapiro-Wilk, as well as normality plots, may thus be applied.

Figure 25: Allocation maps for nine selected patients in the HIV model. The first axis is indexed by the successive follow-up times (hence the different scales) and the second axis by the MCMC iterations. (Different patients correspond to different gray scales.)

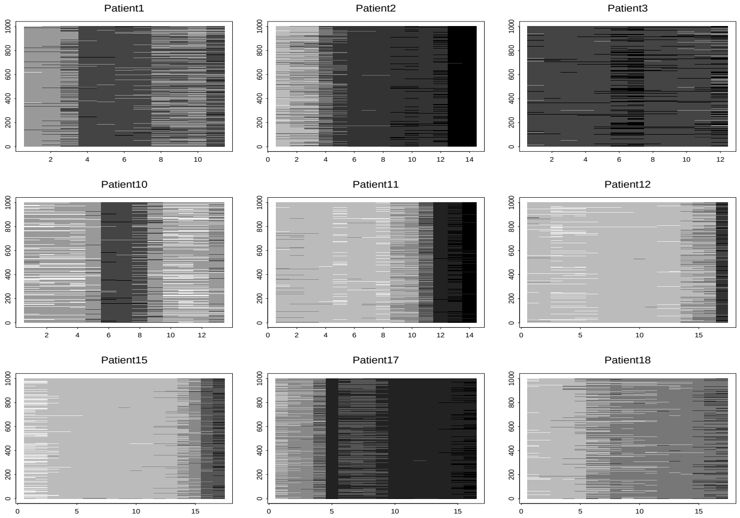


Figure 26: Allocation map (*top*) and average vs. true allocation *(bottom)* for the y_t^*'s in the case of the stochastic volatility model (25), with $\alpha = 0$, $\beta = .9$ and $\sigma = .5$. (The true allocations are represented by dashes.)

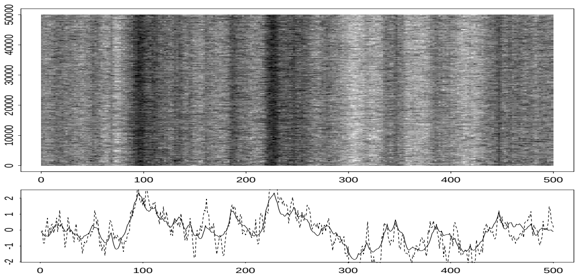
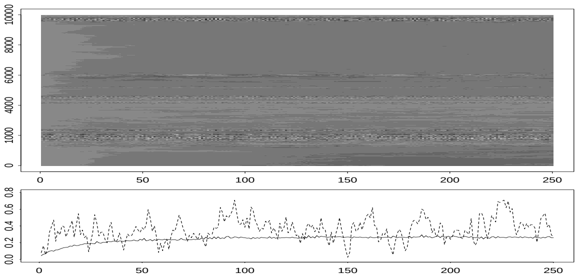
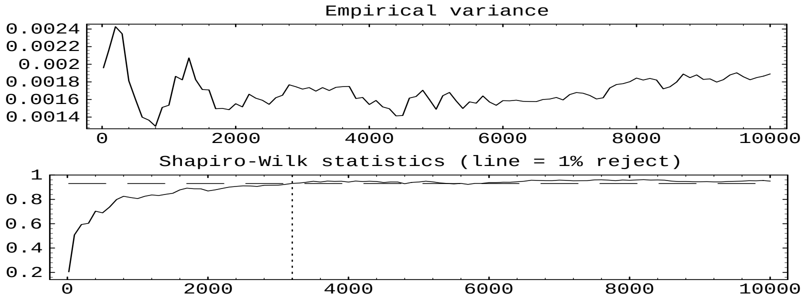


Figure 27: Allocation map (*top*) and average vs. true allocation *(bottom)* for y_t^*'s in the case of the stochastic volatility model, with $\alpha = 0$, $\beta = .7$ and $\sigma = .1$.



Example 1 Figure 28 illustrates the performance of the method of Chauveau and Diebolt (1997) in the case of a two component normal mixture, $p\mathcal{N}(\mu_1, \sigma_1^2) + (1-p)\mathcal{N}(\mu_2, \sigma_2^2)$ when the set A in the indicator function $\mathbb{I}_A(\theta^{(t)})$ is chosen in the tail of the posterior distribution of σ_2. For a proper conjugate prior, the 1% level is reached after approximately 3,500 iterations. The control panel *(top)* also estimates the asymptotic variance of the standardized average, which can be compared with an algebraic estimate as in Chauveau *et al.* (1998). Convergence occurs much more quickly for this criterion, but the discretisation effect must also be taken into account.

Figure 28: CLT control of the convergence of the posterior variance of σ_2. (*Source:* Chauveau and Diebolt, 1997.)



The allocation chain $(z^{(t)})$ can be processed through the diagnostic developed by Robert *et al.* (1999). The control device is built as follows: for each observation, the chain $(z_j^{(t)})$ $(j = 1, \dots, n)$ is independently subsampled (in j), while the estimates $\widehat{\mu}_{jT}$ and $\widehat{V}_{jT}$ are derived from the whole chain. If N_j denotes the number of subsampled $z_j^{(t)}$'s at time T, the normalized versions of the average latent variables $(j = 1, \dots,)$

$$\zeta_j = N_j^{-1/2} \widehat{V}_j^{-1/2} \sum_{k=1}^{N_j} (z_j^{(t_{jk})} - \widehat{\mu}_j) \; . \tag{8}$$

should be approximately i.i.d. $\mathcal{N}(0,1)$ for T large enough, thus providing a convergence control on line. Figure 29 represents a graphical evaluation of the normality of these samples for successive numbers of iterations, by plotting the sample histogram and the normal fit, as well as the normality adequation plot of Ghosh (1996), along with the Kolmogorov-Smirnov p-value. (The graphical normality assessment of Ghosh (1996) is based on the properties of the third derivative of the logarithm of the empirical moment generating function, called T_3-function, in the normal case.) The convergence assessment derived from the asymptotic normality of the ζ_j's thus occurs faster than with the other criteria. This is related to the high exchangeability between components observed in exponential mixtures.

Global control spreadsheets can be constructed, as in Figure 30, which represents a first evaluation of convergence for a normal hidden Markov chain. Besides the allocation map, this graph compares the average allocation over the iterations with the conditional expected allocation based on the estimated parameters, $\mathbb{E}[z_j | x_1, \dots, x_n, \hat{\mu}^T, \hat{\sigma}^T, \hat{\mathbf{P}}]$, more to assess the fit of a mixture (or hidden Markov model) than to check the stationarity of the MCMC chain. The cumulated averages are quite stable, compared with the regular mixture case (see Robert and Mengersen, 1998).

Example 2 The method of Robert *et al* (1999) was used to calculate the normalised sums (after burnins recommended in §4.5) for 100 subsamples of each parameter. Subsamples were generated from a Poisson distribution with a mean of 10 for all parameters, or 3 in the case of $\underline{\alpha}$. The Kolmogorov-Smirnov p-values are consistently low for p_Q in Design 1 and σ_R in Design 2, so that when the minimum p-value (solid line) over all parameters is plotted, even if these individual p-values (dotted lines) rise over 0.05 another p-value plunges to give a pessimistic view (left hand panels of Figure 31). Even though all p-values are (briefly)

Figure 29: Convergence control by normality assessments for the sample from a 3 component exponential mixture: the graphs in inserts are Ghosh's (1996) normality adequation plots and KL gives the significance level for Kolmogorov-Smirnov p-value. (*Source:* Gruet *et al.*, 1999.)

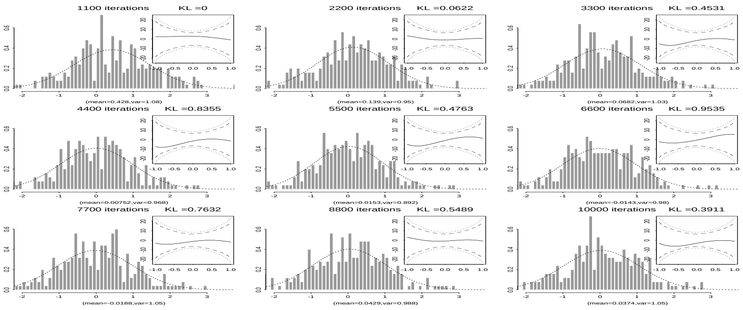
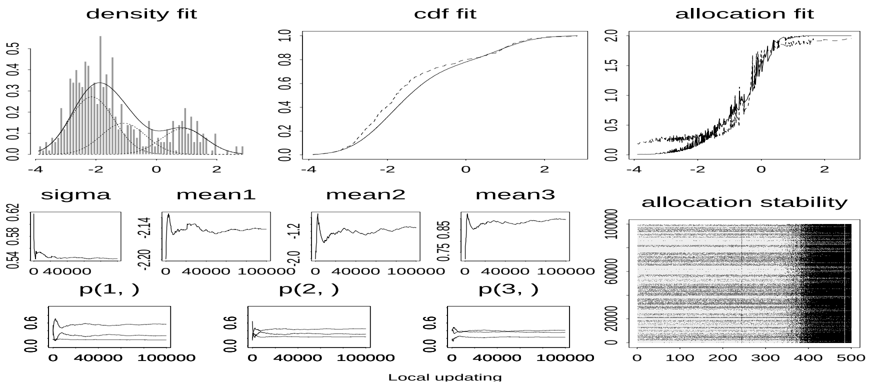
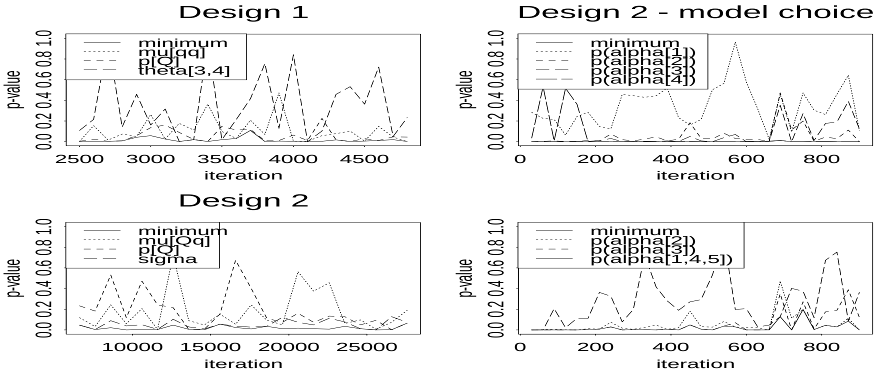


Figure 30: Control spreadsheet for a hidden Markov model estimation, based on 500 observations. (*Source:* Robert *et al.*, 1999.)



acceptible after 500 iterations for Design 1 and 10,000 iterations for Design 2, longer chains seem necessary to estimate parameters with desired precision. When monitoring $\underline{\alpha}$, p-values for $p(\alpha_2)$ and $p(\alpha_3)$ are acceptible after 700 iterations. By summing the probabilities of choosing one of the remaining locus orders (all with posterior probabilites less than 0.16), the minimum p-value is also acceptable, as shown in the right hand panels of Figure 31.

Figure 31: Kolmogorov-Smirnov p-values for normalised sums of selected parameters (dotted lines) and of the minimum over all parameters (solid lines) in the QTL examples.



Example 3 The above normality diagnostic applies for the HIV problem, based on the unobserved stages. These stages are simulated at each follow-up time conditionally on their neighbours, due to the Markov structure of the disease process. If $s_{i,j,t}$ is the unobserved stage of patient i at the j-th follow-up time ($j \geq 1$) and at iteration t, then the full conditional distribution of $s_{i,j,t} \in \{1,2,3,4,5,6\}$ is discrete, based on the exponential of the infinitesimal generator matrix and on $(t_{ij}, t_{i(j+1)})$. The chains $s_{i,j,t}$ are subsampled at random times (for each couple (i,j)) and produce a sample of 3403 normalized sums. As noted in Robert *et al.* (1999), random subsampling does not eliminate the correlation between the $s_{i,j,t}$'s. As illustrated in Figure 32, a consequence of this residual correlation is an artificial overestimation of the mean and the variance of the statistic S_T, leading to an unstable normality adequation plot and significant Kolmogorov–Smirnov tests. We thus only consider the subsample corresponding to the last follow-up time for each patient, which contains 404 points. Figure 33 shows that the closeness to normality of the sample of the normalised S_T's is now somehow constant along the iterations, in the sense that the values of the Kolmogorov–Smirnov p-values vary uniformly over the 10,000 iterations.

Figure 32: Asymptotic normality assessment for the HIV model based on the 3403 follow-up times (same legend as Figure 29).

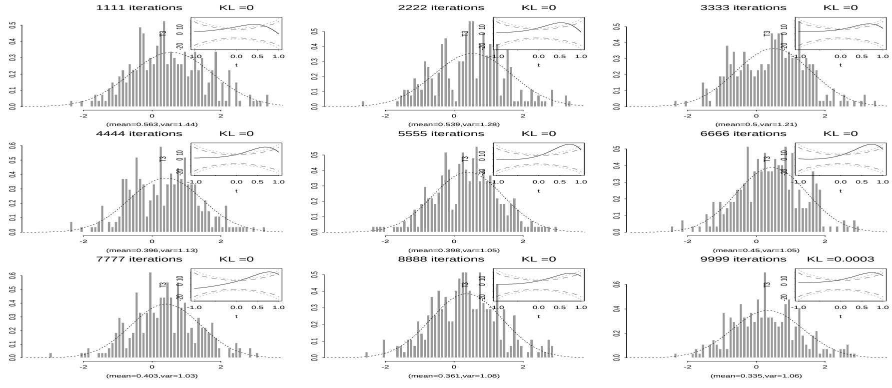
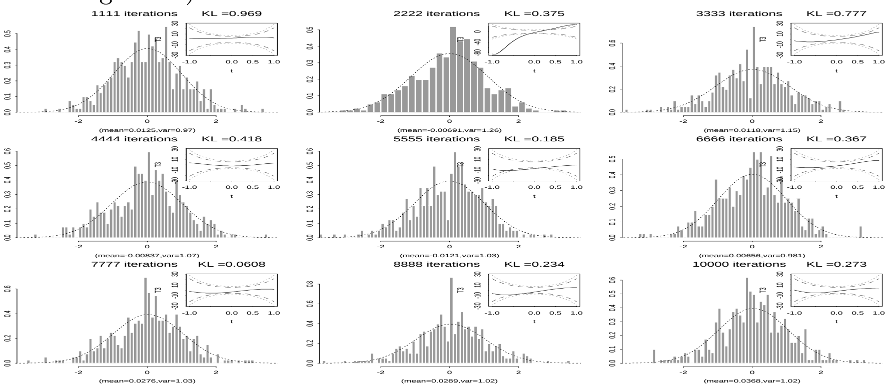
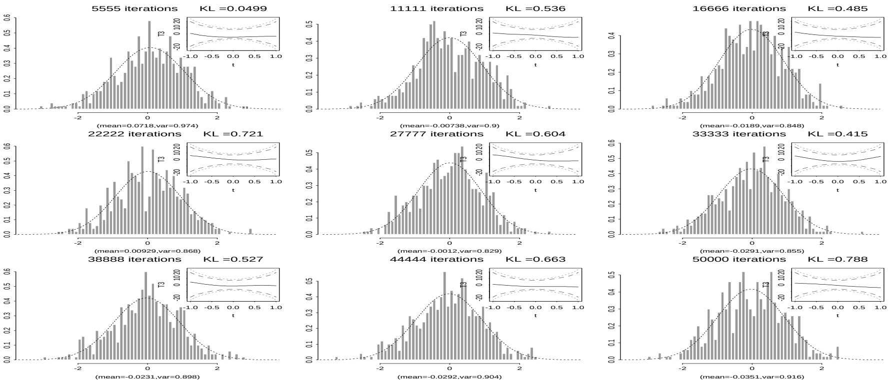


Figure 33: Asymptotic normality assessment for the HIV model based on the last follow-up time for each patient (same legend as Figure 29).



Example 4 For the volatility model, a similar assessment can be derived from the latent structure underlying the model, even though these are continuous, instead of discrete as in Example 1. Figure 34 presents the sequence of normality tests on the normalized y_t^*'s subsamples, as in §2.3.3, confirming the strong stability exhibited by other diagnostics (for the first simulated dataset).

Figure 34: Convergence control by normality assessment for the missing data sample for the stochastic volatility model (25), with $\alpha = 0$, $\beta = .9$ and $\sigma = .5$.



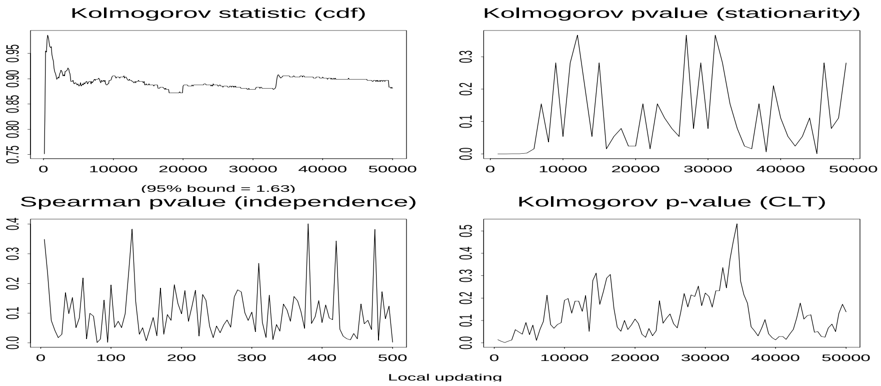
4.9 Nonparametric Tests

Robert *et al.* (1999) proposed a nonparametric stopping rule for exploring the support of the stationary distribution. Given M parallel chains $(\theta^{(m,t)})$ $(m = 1, \ldots, M)$, independently initialized from π_0, a Spearman test of independence between the starting value $(\theta_i^{(1,0)}, \ldots, \theta_i^{(M,0)})$ and $(\theta_i^{(1,t)}, \ldots, \theta_i^{(M,t)})$ is undertaken for every component i of θ, as in (7).

These authors also suggest Kolmogorov-Smirnov or Kuiper tests, for use in the Stationarity stage, based on a single chain. After introducing a batch size G leading to (quasi-) independent observations, comparison is made between the two halves of the MCMC sample, subsampled at intervals of G. For multi-dimensional chains, this can be computed on either the function of interest or on each component of the vector $\theta^{(t)}$. The p-values may be plotted against T or a stopping rule based on K can be derived since the asymptotic cdf of K is easily approximated. Improvements may also be made by taking into account the sequential nature of the test.

Example 1 In the case of a two component normal mixture, $p\mathcal{N}(\mu_1, \sigma_1^2) + (1-p)\mathcal{N}(\mu_2, \sigma_2^2)$, non-parametric tests can also be part of the control spreadsheets discussed in the previous subsection. This is illustrated in Figure 35, which plots the evolution of the various statistics/p-values against the iterations and show how several stopping rules easily follow. (The Kolmogorov statistic represented in Figure 35 is not relevant for convergence control, since it evaluates the fit of the estimated vs. the empirical cdf.) The Kolmogorov p-value for stationarity tests for equality of the distribution of both halves of the MCMC sample; the Spearman p-value evaluates independence from the starting value and the Kolmogorov p-value for the CLT assesses the normality of the sample (8), as above. These procedures thus lead to more conservative assessments of the necessary numbers of iterations, except for the Spearman p-value, which reaches acceptance levels after a few iterations.

Figure 35: Convergence indicators for the hidden Markov model estimation. The subsampling parameters for the CLT assessment are $\nu = 10$ and $d = 0$. (*Source:* Robert *et al.*, 1999.)



Example 2 The Kolmogorov-Smirnov p-values for comparing the two halves of the parameter chains (after burnin) under both designs were computed using a batch size of 10. Figure 36 shows a plot of the minimum p-value over all parameters for Design 1 is acceptable from the outset, while the corresponding plot for Design 2 flattens out to zero after only 5000 iterations. This second plot is a reflection on p_Q. A chain of 15,000 (after 5,000 burnin) not only exhibits large troughs in the nonparametric plots for μ_{Qq} and σ_R, but also fails the spectral diagnostics of §4.5. As with the previous section, a longer chain is necessary for estimating parameters under Design 2. The nonparametric test fails convergence for $p(\alpha_i)$, even when low probability models are summed, which is most likely due to the small sample size.

Example 4 Figure 37 shows a nonparametric test of stationarity for the first simulation of the stochastic model. Surprisingly, this provides more optimistic conclusions than those obtained from Gelman and Rubin's diagnostic, even though the gap in the p value between iteration $10,000$ and $15,000$ indicates that the chain is exploring another region in the support of the stationary distribution. (Note the similar feature in Figure 21 for β.)

5 Discussion and Conclusions

As "expected", no clearcut feature converges from the examination of these four examples, in the sense that (a) it is impossible to exclude one diagnostic, apart from those which do not apply, (b) the stopping rules give different values, sometimes with different orders of magnitude, (c) well-grounded methods like the CLT

Figure 36: Kolmogorov-Smirnov p-values for nonparametric tests of selected parameters (dotted lines) and of the minimum over all parameters (solid lines) in the QTL examples.

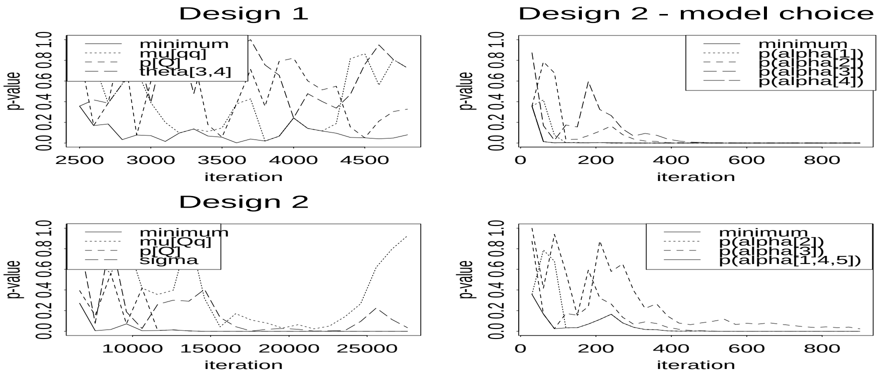
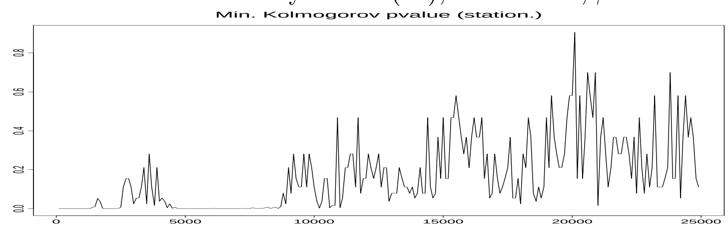


Figure 37: Successive values of the minimum p-value evaluating the stationarity of the sample by comparing both halves, in the case of the stochastic volatility model (25), with $\alpha = 0$, $\beta = .9$ and $\sigma = 0.5$.



test or the Riemann control variates are not necessarily conservative because they do not cover all aspects of MCMC convergence.

The mixture/hidden Markov example (Example 1) showed that elaborate convergence diagnostic spreadsheets could be devised, as in Robert *et al.* (1999), but that the complexity of the model was associated with difficult convergence assessments in the sense that estimated densities and allocation variables showed enough stability to "signal" convergence, while the parameters themselves remained sufficiently unstable to lead to rejection of convergence. The variety of convergence behaviours exhibited in this setup is reproduced by the number of proposed diagnostics, many of which have been originally devised for mixture convergence assessment. This reproduces the conclusion of Mengersen *et al.* (1999) about the pertinence of constructing specialized criteria for special problems and about the possible lack of adequacy of mainstream methods for treating complex models. Note, at last, that the estimation and simulation in mixture models remain a challenging area, when convergence can take different meanings depending on the identifiability imposed on the parameters. (For instance, a symmetric prior on the pairs (p_i, ξ_i) in (1) should lead to symmetric posteriors on the different components.)

The disparity in conclusions obtained under the different diagnostics is illustrated for Designs 1 and 2 (Example 2) in Tables 2 and 3 respectively. Here, comparison is made between recommendations for thinning, burnin and overall chain length from various assessments. While it is clear that overall substantive features are identified by a majority of diagnostics, there was some disagreement on the individual level (Table 4). In this example it is not enough to acknowledge convergence if most parameters behave. Firstly, even if all marginal densities converge we are not guaranteed convergence of the joint distribution. Secondly, most parameters are nuisance parameters and only important ones were slow to converge. The QTL example also illustrated three other issues. First, there is, of course, substantial interaction between diagnostics and refinement of the MCMC algorithm in order to improve convergence. Much of the stability that was

Diagnostic	Thinning[a] (k)	Burnin (t_0)	Chain length[b] (T)	Normality
§4.1 *Graphical Methods*	*Autocorrelations* $k = 15$ μ_{Qq}, μ_{qq}, p_Q $k = 5$ μ_{QQ}, σ_R $k = 3$ $\theta_{2,3}, \theta_{3,4}$	*Trace plots* $t_0 = 0$ for all		*Density Plots* $p_Q, \theta_{4,5}$ not Normal
§4.1 *Cusums*	when $k = 1$ performed badly[c] use $k = 10$ for all	use burnin from §4.5	$T = 4,300$ μ_{qq} $T = 4,000$ $\theta_{2,3}, \theta_{3,4}$ $T < 2,900$ rest	-
§4.2 *Rao-Blackwellisation*	-	-	same as cumulative averages for $\mu_{QQ}, \mu_{Qq}, \mu_{qq}$	-
§4.4 *Raftery & Lewis*	$k = 2$ p_Q	$t_0 = 16$ p_Q $t_0 = 8$ μ_{qq} $t_0 = 5$ $\mu_{Qq}, \theta_{2,3}, \theta_{3,4}$ $t_0 = 2:3$ rest	$T = 14,000$ p_Q $T = 9,000$ μ_{qq} $T = 6,000$ μ_{Qq} $T = 4:5,000$ rest	-
§4.5 *Heidelberger & Welch*	-	$t_0 = 2,400$ $\theta_{3,4}$ $t_0 = 0$ rest	$T = 4,800$ $\theta_{3,4}$ $T < 1,000$ rest	-
§4.5 *Geweke*	-	$t_0 = 2,400$ $\theta_{3,4}$ $t_0 < 500$ for all	$T = 4,800$ $\theta_{3,4}$ $T < 1,000$ rest	-
§4.6 *Gelman & Rubin*	-	$t_0 < T/2$	$T > 4,800$ $\mu_{QQ}, \mu_{qq}, p_{2,2}$ $T = 3:4,000$ $\mu_{Qq}, p_Q, \theta_{2,3}, \theta_{3,4}$ $T = 3,000$ $\theta_{4,5}, p_{2,1}, p_{4,1}$ $T < 2,000$ rest	-
§4.8 *CLT*	use $k = 10$ for all[d]	use burnin from §4.5	$p > 0.05$ for T<3,000 for all but frequently low for p_Q, μ_{qq}	-
§4.9 *Nonparametric*	use $k = 10$ for all[d]	use burnin from §4.5	$p > 0.05$ for T<3,000 for all but troughs at T=3,800 and T=4,600 for some	-

[a] k is further thinning after k=10 imposed for storage reasons.

[b] T includes t_0 for comparative purposes and may differ from the text.

[c] An extension of the cusum diagnostics used here (Brooks, 1998) does provide information on k and t_0.

[d] k is the Poisson sample mean for §4.8 and batch size for §4.9.

Table 2: Comparison of methods under different criteria under Design 1

observed in Section 4 was bought by careful attention to proposals and use of biologically acceptable priors where necessary. Second, stability of parameter estimation within a locus order is indicated even though convergence of $\underline{\alpha}$ was more difficult to determine with the same diagnostics. Third, although the choice of locus order could have been accomplished via a Metropolis–Hastings step, the complex change in the value and interpretation of parameters necessitated by a change in order made the reversible jump algorithm an attractive alternative.[1]

Although the diagnostics partially differed for the HIV example, no strong evidence of divergence was found (except for Raftery and Lewis', evaluation which may take extreme values). Moreover, the medical interpretation of the results (see Richardson and Guihenneuc–Jouyaux, 1996, and Guihenneuc–Jouyaux *et al.*, 1998) does not lead to contradictions with the AIDS literature.

The diagnostics unanimously identified relative stability in the first stochastic volatility model and lack of stability in the second. For instance, all diagnostics in CODA pointed out the poor performances of the MCMC algorithm for this dataset, which requires an alternative implementation. As for mixtures, the complexity of the algorithm excludes distance evaluations, while the latent variable structure prohibits the call to renewal diagnostics.

References

Athreya, K.B., Doss, H. & Sethuraman, J. (1996) On the Convergence of the Markov Chain Simulation Method. *Annals of Statistics* **24**, 69–100.

[1]For issues related with convergence diagnostic in variable dimension models, see also Brooks (1997).

Diagnostic	Thinning[a] (k)	Burnin (t_0)	Chain length[b] (T)	Normality
§4.1 *Graphical Methods*	*Autocorrelations* $k = 10:15$ μ_{QQ}, μ_{Qq} $\mu_{qq}, \sigma_R, p_Q, \theta_{2,3}, \theta_{3,4}$ $k = 1$ $\theta_{1,2}, \theta_{4,5}$ $k = 2:5$ for rest	*Trace plots* $t_0 = 0$ for all $t_0 = 0$ for all	*Cumulative averages*	*Density Plots* $\theta_{1,2}, \theta_{2,3}, \theta_{3,4}, \theta_{4,5}$ $p_Q, \theta_{4,5}$ not Normal
§4.1 *Cusums*	when $k = 1$ performed badly[c] use $k = 15$ for all	use burnin from §4.5	$T > 28,000$ $\sigma_R, \theta_{2,3}, \theta_{3,4}$, $\theta_{4,5}, p_{3,1}, p_{4,2}$ $T = 25,000$ $\mu_{qq}, p_{1,2}$ $T = 11,000$ $p_Q, p_{3,2}$ $T < 5,400$ rest	-
§4.2 *Rao-Blackwellisation*	-	-	same as cumulative averages for $\mu_{QQ}, \mu_{Qq}, \mu_{qq}$	-
§4.4 *Raftery & Lewis*	$k = 4:5$ μ_{QQ}, σ_R $k = 2:3$ μ_{Qq}, μ_{qq}, p_Q	$t_0 = 32:35$ μ_{QQ}, σ_R $t_0 = 12:15$ μ_{Qq}, μ_{qq}, p_Q $t_0 = 6:9$ $\theta_{2,3}, \theta_{3,4}$ $t_0 = 2:4$ rest	$T = 33:34,000$ μ_{QQ}, σ_R $T = 15:18,000$ μ_{Qq}, μ_{qq}, p_Q $T = 10,000$ $\theta_{2,3}, \theta_{3,4}$ $T = 4:5,000$ rest	-
§4.5 *Heidelberger & Welch*	-	$t_0 = 5,600$ σ_R $t_0 = 0$ rest	$T = 28,000$ $\sigma_R, \mu_{Qq}, \mu_{qq}$ $T < 10,000$ rest	-
§4.5 *Geweke*	-	$t_0 = 5,600$ σ_R $t_0 = 0$ rest	$T = 28,000$ $\sigma_R, \mu_{Qq}, \mu_{qq}$ $T < 10,000$ rest	-
§4.6 *Gelman & Rubin*	-	$t_0 < T/2$	Only 7,000 iterations collected for 2nd chain p_Q not converged at this point	-
§4.8 *CLT*	use $k = 10$ for all[d]	use burnin from §4.5	$p > 0.05$ for $T < 13,000$ for all but frequently low for p_Q, σ_R	-
§4.9 *Nonparametric*	use $k = 10$ for all[d]	use burnin from §4.5	$p > 0.05$ for $T < 10,000$ for all but flat thereafter for p_Q trough at T=20,000 for μ_{Qq}, p_Q, σ_R	-

[a] k is further thinning after k=10 imposed for storage reasons.

[b] T includes t_0 for comparative purposes.

[c] An extension of the cusum diagnostics used here (Brooks, 1998) does provide information on k and t_0.

[d] k is the Poisson sample mean for §4.8 and batch size for §4.9.

Table 3: Comparison of methods under different criteria under Design 2

Diagnostic	Design 2 σ_R	Design 1 $\theta_{3,4}$
§4.1 *Graphical Methods*	$k = 15$ (Fig. 7)	$k = 15$
	$t_0 = 0$ (Fig. 6)	$t_0 = 0$ (Fig. 5)
§4.1 *Cusums*	using $k = 15$, $t_0 = 5,600$	using $k = 10$, $t_0 = 2,200$
	$T > 28,000$ (Fig. 8)	$T > 4,000$
§4.4 *Raftery & Lewis*	$k = 4$, $t_0 = 32$, $T = 33,000$	$k = 1$, $t_0 = 5$, $T = 5,500$
§4.5 *Heidelberger & Welch*	$t_0 = 5,600$, $T = 28,000$	$t_0 = 2,400$, $T = 4,800$
§4.5 *Geweke*	$t_0 = 5,600$, $T = 28,000$ (Fig. 17)	$t_0 = 2,400$, $T = 4,800$ (Fig. 17)
§4.6 *Gelman & Rubin*	inconclusive (Fig. 20)	$t_0 < T/2$, $T = 4,000$ (Fig. 20)
§4.8 *CLT*	T? (Fig. 31)	$T < 3,000$ (Fig. 31)
§4.9 *Nonparametric*	T? (Fig. 36)	$T < 3,000$ (Fig. 36)

Table 4: Comparison of burnin(k), thinning(t_0) and chain length(T) for σ_R (Design 2) and $\theta_{3,4}$ (Design 1).

Besag, J., Green, E., Higdon, D. & Mengersen, K.L. (1995) Bayesian computation and stochastic systems (with discussion). *Statistical Science* **10**, 3–66.

Best, N., Cowles, M.K. & Vines, K. (1995) *CODA Manual Version 0.30.* MRC Biostatistics Unit, Cambridge.

Billio, M., Monfort, A., & Robert, C.P. (1998) The Simulated Likelihood Ratio method. Doc. travail CREST, Insee, Paris.

Brooks, S.P. (1997) Discussion to Richardson and Green (1997). *J. Royal Statist. Soc.* (Ser. B) **59**, 774–775.

Brooks, S.P. (1998) Quantitative Convergence Diagnosis for MCMC via CUSUMS. *Statistics and Computing* **8**, 267–274.

Brooks, S.P., Dellaportas, P. & Roberts, G.O. (1997) A Total Variation Method for Diagnosing Convergence of MCMC Algorithms. *J. Comput. Graph. Statist.* **6**, 251–265.

Brooks, S.P. & Gelman, A. (1998) Alternative Methods for Monitoring Convergence of Iterative Simulations. *J. Comput. Graph. Statist.* **7**, 434–455.

Brooks, S.P. & Giudici, P. (1998) Diagnosing convergence of reversible jump MCMC algorithms. Tech. report, University of Bristol.

Brooks, S.P. & Roberts, G.O. (1997) On Quantile Estimation and MCMC Convergence. Tech. report, University of Bristol.

Brooks, S.P. & Roberts, G.O. (1999) Diagnosing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing* (to appear).

Casella, G. & Robert, C.P. (1996) Rao-Blackwellisation of sampling schemes. *Biometrika* **83**, 81–94.

Chan, K.S. & Geyer, C.J. (1994) Discussion of "Markov chains for exploring posterior distribution". *Ann. Statis.* **22**, 1747–1758.

Chauveau, D. & Diebolt, J. (1997) MCMC convergence diagnosis via the Central Limit Theorem. Tech. report no. 22/97, U. Marne-la-Vallée.

Chauveau, D., Diebolt, J. & Robert, C.P. (1998) Control by the Central Limit theorem. In *Discretization Methods for MCMC Convergence Assessment*, C.P. Robert (Ed.), Chapter 5, 99–126. Lecture Notes in Statistics 135, Springer-Verlag, New York.

Cowles, M.K. & Carlin, B.P. (1996) Markov Chain Monte-Carlo convergence diagnostics: a comparative study. *J. Amer. Statist. Assoc.* **91**, 883–904.

Diebolt, J. & Robert, C.P. (1994) Estimation of finite mixture distributions by Bayesian sampling. *J. Royal Statist. Soc.* (Ser. B) **56**, 363–375.

Gelfand, A.E. & Smith, A.F.M. (1990) Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

Gelman, A. & Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**, 457–511.

George, A.W., Mengersen, K.L. & Davis, G.P. (1998) QTL detection for a half-sib design: a Bayesian approach. Technical Report, CiSSaIM, QUT.

George, A.W., Mengersen, K.L. & Davis, G.P. (1999) A Bayesian approach to gene ordering for a half-sib design. *Biometrics* (to appear).

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid et A.F.M. Smith (Eds.), 169–193. Oxford University Press, Oxford.

Geyer, C.J. & Thompson, E.A. (1995) Annealing Monte Carlo maximum likelihood with application to pedigree analysis. *J. Amer. Statist. Assoc.* **90**, 909-920.

Ghosh J. (1996) A new graphical tool to detect non normality. *J. Royal Statist. Soc.* (Ser. B) **58**, 691–702.

Gruet, M.A., Philippe, A., & Robert, C.P. (1999) MCMC control spreadsheets for exponential mixture estimation. *J. Comput. Graph. Statist.* (to appear).

Guihenneuc–Jouyaux, C., Richardson, S. & Lasserre, V. (1998) Convergence assessment in latent variable models: apllication to longitudinal modelling of a marker of HIV progression. In *Discretization Methods for MCMC Convergence Assessment*, C.P. Robert (ed.), Chapter 7, 147–160. Lecture Notes in Statistics 135, Springer-Verlag, New York.

Guihenneuc–Jouyaux, C. & Robert, C.P. (1998) Finite Markov chain convergence results and MCMC convergence assessments. *J. Amer. Statist. Assoc.* **93**(3), 1055–1067.

Heidelberger, P. & Welch, P.D. (1983) A spectral method for confidence interval generation and run-length control in simulation. *Commun. ACM* **24**, 233–245.

Izenman, A.J. & Sommer, C.J. (1988) Philatelic mixtures and multimodal densities. *J. Amer. Statist. Assoc.* **83**, 941–953.

Jacquier, E., Polson, N.G. & Rossi, P.E. (1994) Bayesian analysis of stochastic volatility models, *Journal of Business & Economic Statistics* **12**,, 371-417.

Jennison, C. (1993) Discussion on the meeting on the Gibbs sampler and other Markov Chain Monte Carlo methods. *J. Royal Statist. Soc.* (Ser. B) **55**, 54–56.

Kim, S., Shephard, N. & Chib, S. (1998) Stochastic volatility: likelihood inference and comparison with ARCH models. *Rev. Econom. Stud.* **65**, 361–393.

Mengersen, K. & Robert, C.P. (1996) Testing for mixtures: a Bayesian entropic approach. In *Bayesian Statistics 5*, J.O. Berger, J.M. Bernardo, A.P. Dawid, D.V. Lindley and A.F.M. Smith (Eds.), 255-276. Oxford University Press, London.

Mengersen, K.L., Robert, C.P. & Guihenneuc-Jouyaux, C. (1999) MCMC convergence diagnostics: a "reviewww". In *Bayesian Statistics 6*. J.O. Berger, J.M. Bernardo, A.P. Dawid, D.V. Lindley and A.F.M. Smith (Eds.). Oxford University Press, Oxford (to appear).

Mengersen, K.L. & Tweedie, R.L. (1996) Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24**, 101–121.

Murdoch, D.J. & Green, P.J. (1998) Exact Sampling from a Continuous State Space. *Scandinavian J. Statistics* **25**(3), 483–502.

Philippe, A. (1997a) Simulation output by Riemann sums, *J. Statist. Comput. Simul.* **59**(4), 295–314.

Philippe, A. (1997b) Importance sampling and Riemann Sums. *Prepub. IRMA Lille* **43**,VI.

Philippe, A. & Robert, C.P. (1998) Riemann sums for MCMC estimation and convergence monitoring. Doc. travail, CREST.

Polson, N.G. (1996) Convergence of Markov Chain Monte-Carlo Algorithms (with discussion). In *Bayesian Statistics 5*, J.M. Bernardo, J.O. Berger, A.P. Dawid et A.F.M. Smith (Eds.). Oxford University Press, Oxford, 297–323.

Propp, J.G. & Wilson, D.B. (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223–252.

Raftery, A.E. & Lewis, S. (1992) How many iterations in the Gibbs sampler? In *Bayesian Statistics 4*, J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith (Eds), 763–773. Oxford University Press, Oxford.

Richardson, S. & Green, P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Soc.* (Ser. B) **59**, 731–792.

Robert, C.P. (1995) Convergence control techniques for MCMC algorithms. *Statistical Science* **10**, 231–253.

Robert, C.P. (1996) Inference in mixture models. In *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson and D.J. Spiegelhalter (Eds.), 441–464. Chapman and Hall, London.

Robert, C.P. (1997) Discussion of Richardson and Green' paper. *J. Royal Statist. Soc.* (Ser. B) **59**, 758–764.

Robert, C.P. & Casella, G. (1999) *Monte Carlo Statistical Methods.* Springer-Verlag (to appear).

Robert, C.P. & Mengersen, K.L. (1998) Reparametrization issues in mixture estimation and their bearings on the Gibbs sampler. *Comput. Statist. Data Anal.* **29**, 325–343.

Robert, C.P., Rydén, T. & Titterington, D.M. (1999) Convergence controls for MCMC algorithms, with applications to hidden Markov chains. *J. Statist. Comput. Simul.* (to appear).

Robert, C.P. & Titterington, D.M. (1998) Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statistics and Computing* **8**(2), 145–158.

Roberts, G.O. & Rosenthal, J.S. (1998) Markov-chain Monte Carlo : some practical applications of theorical results. *The Canadian Journal of Statistics* **26**(1), 4–31.

Roberts, G.O. & Tweedie, R.L. (1996) Geometric convergence and Central Limit Theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83**, 95–110.

Roeder, K. & Wassserman, L. (1997) Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92**, 894–902.

Shephard, N. & Pitt, M.K. (1997) Likelihood analysis of non-Gaussian measurement time series, *Biometrika* **84**, 653–668.

Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701–1786.

Yu, B. (1995) Discussion to Besag *et al.* (1995). *Statist. Sci.* **10**, 54–58.

Yu, B. & Mykland, P. (1998) Looking at Markov Samplers Through Cusum Path Plots: A Simple Diagnostic Idea. *Statistics and Computing* **8**(3), 275–286.