

A distance-based diagnostic for trans-dimensional Markov chains

S. A. Sisson and Y. Fan¹

Abstract

Over the last decade the use of trans-dimensional sampling algorithms has become endemic in the statistical literature. In spite of their application however, there are few reliable methods to assess whether the underlying Markov chains have reached their stationary distribution. In this article we present a distance-based method for the comparison of trans-dimensional Markov chain sample output for a broad class of models. The presented method requires no a priori assumptions regarding the nature of parameters across model space. In addition, the diagnostic will simultaneously assess deviations between and within chains. Illustration of the analysis of Markov chain sample-paths is presented in simulated examples and in two common modelling situations: a finite mixture analysis and a change-point problem.

Keywords: Change-point problems; Convergence assessment; Finite mixture models; Markov chain Monte Carlo; Reversible jump, Trans-dimensional Markov chain.

¹S. A. Sisson is lecturer, School of Mathematics, University of New South Wales, Sydney 2052, Australia. (Email: Scott.Sisson@unsw.edu.au). Y. Fan is lecturer, School of Mathematics, University of New South Wales, Sydney 2052, Australia. (Email: Y.Fan@unsw.edu.au).

1 Introduction

Since the mid-1990's a number of trans-dimensional frameworks have been proposed for the generation of sample realisations drawn from distributions that span multiple models of possibly varying dimensionality (Green 2003; Sisson 2005). These stochastic algorithms generate a Markov chain of dependent realisations with a desired stationary distribution. As with all iterative algorithms, one obvious prerequisite to inference is that the chain converges to its equilibrium state.

In fixed dimensional cases, various methods have been proposed for assessing convergence without the need to analyse simulation output. One approach is to design an algorithm to produce independent draws directly from the target distribution (Nummelin, 1978; Mykland et al., 1995; Johnson, 1998; Propp and Wilson, 1996). Another approach is to develop theoretical (analytic) results bounding the difference between the simulation and target distributions after some specified number of iterations (Rosenthal 1995; Cowles and Rosenthal 1998; Roberts and Tweedie, 1999, 2000). Each of the above methods have achieved only limited success and cannot (yet) be easily extended to the variable dimension setting. In the absence of such methods, chain convergence is assessed by analysing multiple sampler output (Cowles and Carlin 1996). One drawback of this approach is that these diagnostics invariably fail to detect a lack of convergence when parts of the target distribution are missed entirely by all replicate chains. To reduce this problem, Gelman and Rubin (1992) advocate using well dispersed starting values for each replicate chain.

Few methods for monitoring trans-dimensional sampler output are available (Brooks

and Giudici 2000; Castelloe and Zimmerman 2002; Brooks et al. 2003; Sisson (2005) provides a review.) These methods can be summarised into two groups. The first such method extends the ideas of Gelman and Rubin (1992) to the trans-dimensional case, by monitoring of functionals of parameters which retain their interpretations as the sampler moves between topologies. Specifically, Brooks and Giudici (2000) advocate a two-way ANOVA decomposition of the functional over replicate chains. Castelloe and Zimmerman (2002) extended this approach to a weighted MANOVA design, preventing rarely visited models dominating the statistic, and permitting the monitoring of several functionals. However, there is a difficulty in identifying adequate functionals. In the multi-model setting, even common parameters may change meaning between models (Berger and Pericchi 1996; Fan and Brooks 2000). Brooks and Giudici (2000) propose the deviance as a default functional in the absence of superior options.

The second method for analysing trans-dimensional sampler output relies on monitoring model indicators; implementing non-parametric tests to assess differences in model probabilities. However, requisite test assumptions may be problematic in the Markov chain setting. For example, the Kolmogorov-Smirnov and χ^2 tests proposed by Brooks et al. (2003) require independent sample outputs. One obvious problem with monitoring model indicators is when between-model convergence is achieved before within-model convergence.

In this article we develop a convergence diagnostic for trans-dimensional Markov chains where the model may be represented within a marked point process framework (Stephens 2000; Cappé et al. 2003). Popular modelling scenarios in this general

framework include finite mixture problems (Richardson and Green 1997), change-point analyses (Green 1995) and variable selection for regression models (George and McCulloch 1995). The diagnostic we develop ensures that chain performance is assessed over all model parameters across all models.

We take our motivation from techniques developed for the analysis of labelled spatial point patterns (Moller and Waagepetersen 2004; Diggle 2003; Cuzick and Edwards 1990; Besag and Diggle 1977). One descriptor of spatial homogeneity is based on point-to-nearest-event distances. For observed events $s = \{s_1, \dots, s_n\}$ and specified points $\mathcal{V} = \{v_1, \dots, v_V\}$ in a region $A \subseteq \mathbb{R}^d$, let y_i denote the distance from the i -th point (v_i) to the nearest event (s_j) in A with respect to some distance measure. The distribution function of y may be estimated as

$$\hat{G}_1(y) = V^{-1} \sum_{i=1}^V I(y_i \leq y), \quad (1)$$

where $I(\cdot)$ is the indicator function, this is a measure of the empty spaces in A . When A is bounded, an edge-corrected estimator is required. Monte Carlo tests are often performed by simulation of replicate functions $\hat{G}_i(y)$, $i = 2 \dots, g$, under the null hypothesis (e.g. complete spatial randomness), on the statistic

$$\omega_i = \int_0^\infty \left(\hat{G}_i(y) - \bar{G}_i(y) \right)^2 dy, \quad i = 1, \dots, g, \quad (2)$$

where

$$\bar{G}_i(y) = (g - 1)^{-1} \sum_{j \neq i} \hat{G}_j(y).$$

Such methods have had previous application in the post-processing of Markov chain simulation output. Celeux et al. (2000) developed a point-to-nearest-event loss

function for parameter estimation in the context of highly modal k -dimensional mixture distributions. In deriving the Bayes estimator, Celeux et al. (2000) considered each Markov chain realisation as spatial pattern of size k in component space. Sisson and Hurn (2004) later noted this loss function could be extended to the variable-dimension framework.

In Section 2, we develop the above spatial descriptors to identify a novel measure of sample-path deviation between independent chains in the trans-dimensional Markov chain setting. Aspects of the diagnostic are illustrated through simulation in Section 3. A theoretical example, and two popular real data settings: the analysis of finite mixtures and a change-point problem, are examined in detail and contrasted with existing trans-dimensional diagnostics in Section 4. We conclude with a discussion.

2 Method

2.1 The diagnostic

We now adapt the above methods to develop a convergence diagnostic for a trans-dimensional Markov chain, when the model is formulated in a point process setting.

Proposition 2.1 *Denote the target density by*

$$\pi = \sum_{k \in \mathcal{K}} \rho_k \pi_k, \quad \mathcal{K} = \{1, 2, \dots\},$$

where π_k denotes the normalised density of the model with $k \in \mathcal{K}$ components, and $\rho_k \in [0, 1]$ the weight accorded to model k , $\sum \rho_k = 1$. Let parameter specification $\xi = \{\xi_1, \dots, \xi_k\}$ from model π_k be considered as k events residing in component space, $A \subseteq$

\mathbb{R}^d . Then the ξ_i follow an inhomogeneous point process on A with locally integrable and finite intensity function, $\mu_k : A \rightarrow [0, \infty)$, where $\int_B \mu_k(\beta) d\beta < \infty$ for all bounded $B \subseteq A$. Accordingly the intensity function arising through the density π is the mixture $\mu = \sum_{k \in \mathcal{K}} \rho_k \mu_k$.

That is, the target density π determines a corresponding inhomogeneous intensity function, μ , defined on the d -dimensional component space, A . We are able to estimate aspects of μ based on the spatial point process descriptor (1) using sample output from a trans-dimensional Markov chain.

Proposition 2.2 *Consider a Markov chain with stationary distribution π and sample path $\xi^{(1)}, \dots, \xi^{(N)}$. Each sample realisation of varying length may be expressed as $\xi^{(r)} = \{\xi_1^{(r)}, \dots, \xi_{K^r}^{(r)}\}$, where $\xi_j^{(r)} \in A \subseteq \mathbb{R}^d$, for $r = 1, \dots, N, j = 1, \dots, K^r$. Let $\mathcal{V} \subseteq A$ denote a subset of component space. Then for each $v \in \mathcal{V}$, define*

$$x_r = \text{distance from } v \text{ to the nearest component in } \xi^{(r)}.$$

The distribution function of x , $F(x; v)$, is naturally estimated as

$$\hat{F}(x; v) = N^{-1} \sum_{r=1}^N I(x_r \leq x). \quad (3)$$

The function $F(x; v)$ has a subtly different interpretation to that of $G_1(y)$ estimated by (1). In the manner of Celeux et al. (2000), the distribution $F(x; v)$ is estimated from N replicated spatial patterns (i.e. Markov chain realisations) as the distribution of the distance from v to the nearest component of the distribution π . Compare this to (1) where $G_1(y)$, estimated from $V = |\mathcal{V}|$ points, characterises the degree of aggregation or regularity in a *single* spatial pattern as observed from *all* points in \mathcal{V} .

Hence we have reduced the N variable dimension samples to a family of univariate distribution functions, each of which describes a local area of component space A . For an underlying intensity μ defined on A and implied by π , the function $F(x; v_i)$ will in general be distinct from $F(x; v_j)$ when observed from different reference points, $v_i \neq v_j, v_i, v_j \in \mathcal{V}$. If the cardinality of \mathcal{V} is sufficiently large, the essential characteristics of μ (and hence π) are captured by the functions $F(x; v), v \in \mathcal{V}$. Accordingly any discrepancies between two independent sampler implementations will be apparent in the estimates of the $F(x; v)$. Monitoring the differences between these distributions will provide information on the sampling properties of replicated chains, permitting an assessment of chain convergence.

2.2 Monitoring Convergence

Suppose that we have $C \geq 2$ independent trans-dimensional sample outputs of replicate chain implementations, S_1, \dots, S_C , and estimates of the functions $F^{S_c}(x; v)$ (given by (3)), for $v \in \mathcal{V}$ and $c = 1, \dots, C$. There are several ways to compare replicate sampler output in this setting. Here we consider the following:

a) Pairwise comparisons

Define

$$u^{c_1 c_2}(v) = \int_0^\infty |F^{S_{c_1}}(x; v) - F^{S_{c_2}}(x; v)|^p dx$$

using L^p distance with $p \in \mathbb{R}^+$, as the discrepancy between chains S_{c_1} and S_{c_2} as observed from reference point $v \in \mathcal{V} \subseteq A$. This integral may be approximated via a standard one-dimensional numerical integration over some range $[0, x_Z]$.

As the intensity, μ , is assumed to be non-homogeneous under π , we contrast each pair of chains from all possible reference points to obtain:

$$\begin{aligned} u^{c_1 c_2} &= \int_{\mathcal{V}} u^{c_1 c_2}(v) dv = \int_{\mathcal{V}} \int_0^\infty |F^{S_{c_1}}(x; v) - F^{S_{c_2}}(x; v)|^p dx dv \quad (4) \\ &\approx V^{-1} \sum_{v \in \mathcal{V}} \int_0^\infty |F^{S_{c_1}}(x; v) - F^{S_{c_2}}(x; v)|^p dx. \end{aligned}$$

In practice, to approximate the above integral over \mathcal{V} we adopt a finite set $\mathcal{V} = \{v_1, \dots, v_V\}$. Note that from (4), when $\mathcal{V} = A$, then $u^{c_1 c_2} = 0$ if and only if the chains S_{c_1} and S_{c_2} consist of identical realisations. If desired, an overall distance measure for $C \geq 3$ chains can be obtained by averaging over all pairwise distances:

$$u = \binom{C}{2}^{-1} \sum_{c_1=1}^C \sum_{c_2 \neq c_1} u^{c_1 c_2}. \quad (5)$$

We may then plot the quantities $u^{c_1 c_2}$ and u and monitor the discrepancies between sample outputs towards the value zero.

b) Monte Carlo test

As an alternative to pairwise comparisons, we may compare each chain, S_c , to the rest in analogy with the Monte Carlo hypothesis tests as described in Section 1. Following (2), for $c = 1, \dots, C$, we define

$$w^c(v) = \int_0^\infty |F^{S_c}(x; v) - \bar{F}^c(x; v)|^p dx,$$

where

$$\bar{F}^c(x; v) = (C - 1)^{-1} \sum_{k \neq c} F^{S_k}(x; v)$$

is calculated using all chain replicates except S_c . Then, integrating over \mathcal{V} gives

$$w^c = \int_{\mathcal{V}} w^c(v) dv \approx V^{-1} \sum_{v \in \mathcal{V}} \int_0^\infty |F^{S_c}(x; v) - \bar{F}^c(x; v)|^p dx \quad (6)$$

which may be evaluated as before. In this manner, if C is large, w^c may be interpreted through a Monte Carlo test of equal sampling distributions of all chains. That is, if we suppose chains $S_k, k \neq c$, have converged, then the above permits the test that the chain S_c has also converged. In practice however, the number of replicated chains, C , will rarely be large and the assumption that all chains $S_k, k \neq c$ have converged will be false. However the statistic remains conceptually useful, particularly for the identification of problem chains. If all chain replicates have converged then we expect w^c to approach zero for all chains. If the value of w^c is consistently larger for one chain than the rest, then we may discard this chain and replace it with new simulations.

c) **Potential scale reduction factor (PSRF)**

The diagnostic of Gelman and Rubin (1992) is a standard fixed-dimensional convergence tool, and as such is easily interpretable by many practitioners. If we denote $x_{rc}, r = 1, \dots, N, c = 1, \dots, C$ (the distance from reference point v to the nearest model component at iteration r in the c^{th} chain) as the scalar estimand to be monitored, this diagnostic may be implemented directly for each $v \in \mathcal{V}$. Specifically this involves monitoring the corresponding PSRFv:

$$\hat{R}_v = \left(\frac{\frac{N-1}{N}W_v + \frac{1}{N}B_v}{W_v} \right)^{1/2} \quad (7)$$

where B_v and W_v denote the between and within sequence variances, given by

$$B_v = N(C-1)^{-1} \sum_{c=1}^C (\bar{x}_{.c} - \bar{x}_{..})^2, \quad \text{where} \quad \bar{x}_{.c} = N^{-1} \sum_{r=1}^N x_{rc}, \quad \bar{x}_{..} = C^{-1} \sum_{c=1}^C \bar{x}_{.c}$$

$$W_v = C^{-1} \sum_{c=1}^C s_c^2, \quad \text{where} \quad s_c^2 = (N-1)^{-1} \sum_{r=1}^N (x_{rc} - \bar{x}_{.c})^2.$$

This diagnostic suggests possible convergence to the target distribution if \hat{R}_v declines to 1 as $N \rightarrow \infty$, for all $v \in \mathcal{V}$. If \hat{R}_v is high for any reference point v , then proceeding with further simulations may improve inference.

For sequences of length N , Gelman and Rubin (1992) hypothesised convergence of all chains after the first $N/2$ iterations, and accordingly evaluated this on the basis of chain iterations $N/2 + 1, \dots, N$. An incremental assessment on the performance of chains of length $2 \leq N_0 \leq N$ is provided by evaluating the diagnostic over successively larger portions of the full sample-path. We adopt this procedure here.

2.3 Reference point determination

The similarity of sample-paths under replicate chain implementations is assessed on component space, A , with respect to reference points $v \in \mathcal{V} \subseteq A$. If the intensity function μ admitted by π is complex, then a sufficient number and range of v are required to detect any discrepancy between samples: a point located near to one particular component of a model will miss any discrepancies in other components.

For densities π consisting of a few well defined components, only a few, well-chosen reference points may be needed. However, such a specification of \mathcal{V} is unlikely to be practical. Except where implementing a regular grid is feasible (Diggle and Matérn 1981, for example), we adopt the approach of randomly selecting equal numbers of reference points from the full sample-path of each replicate chain (Celeux et al. 2000). Their number, determined by the perceived complexity of the (common) intensity μ , should increase in proportion to the total number of model components and their

proximity in A . We demonstrate in Section 3.2, via a simple non-trivial example, that the number of reference points need not be large to obtain accurate results. As a general guide, we advocate starting with around 100 reference points, and evaluate their sufficiency by re-evaluating the diagnostic on progressively enlarged sets.

3 Simulation Studies

We now present two simulated studies. The first evaluates the proposed diagnostic’s ability to assess between-model convergence. The second study empirically evaluates the number of reference points needed to obtain stable diagnostic results.

3.1 Example 1: Convergence in model probability

One important issue in assessing the performance of trans-dimensional Markov chains is evaluating the convergence of model probabilities. A wide range of univariate diagnostics could be used to this effect (Brooks et al. 2003, for example) although convergence in model probabilities alone is not a sufficient criterion for within model convergence. A desirable feature of any variable-dimensional convergence diagnostic is that it would naturally incorporate a measure of this kind.

We consider a simple one-dimensional mixture, under two “chain” replicates S_1 and S_2 , and monitor the statistics $u^{12} = w^1 = w^2$ with L^1 distance ($p = 1$). Define

$$X_{nq} = \{x_1, \dots, x_{nq}\}, \quad Y_{n(1-q)} = \{(y_1, z_1), \dots, (y_{n(1-q)}, z_{n(1-q)})\}$$

to be a sequence of one and two component random vectors of length nq and $n(1 - q)$ respectively. Hence the pair $Z_q = \{X_{nq}, Y_{n(1-q)}\}$ constitutes a set of n realisations

from a mixture distribution with a probability of $q \in [0, 1]$ of observing a realisation with one component. Suppose that, for fixed n , the “true” probability of the one component model is given by q^* , and a random sample from the full distribution by Z_{q^*} . We may then evaluate the discrepancy between Z_{q^*} and Z_q through u^{12} for a sequence of values of q . If u^{12} adequately discriminates between mis-proportioned model probabilities, we would expect the case $q = q^*$ to yield the smallest discrepancy.

Specifically, with $n = 1000$, consider the case where $x_i, y_j, z_j \sim N(0, 1)$ independently for $i = 1, \dots, nq, j = 1, \dots, n(1 - q)$, and where $\mathcal{V} = \{v_i = \Phi^{-1}(\frac{i-1/2}{20}) : i = 1, \dots, 20\}$ are quantiles of the standard Normal density. Figure 1 (a) displays u^{12} evaluated between Z_{q^*} and Z_q for a range of values of $q \in [0, 1]$ in the three cases defined by $q^* = 0.2, 0.5, 0.8$. In each instance u^{12} is minimised at the true model proportions. In the case of $q^* = 0.5$, Figure 1 (b, c) illustrates the disparity between estimates of $F(x; v_1)$ determined from Z_{q^*} and Z_q under the extreme cases of exact ($q = 0.5$) and strongly mismatched ($q = 0.9$) model proportions respectively. From these images it is clear why mis-allocation of model probabilities is easily detected.

3.2 Example 2: Simulation of reference points

We examine the effect of the number of reference points required for a robust inference in a simple example. Consider the following three models:

$$\mathbf{M1:} N(1, 5); \quad \mathbf{M2:} N(-5, 2), N(5, 1); \quad \mathbf{M3:} N(-5, 2), N(5, 1), N(6, 1).$$

M3 differs from M2 by an additional nearby component, a situation which commonly arises in practice. We construct three artificial “chains” of length 10,000. Two con-

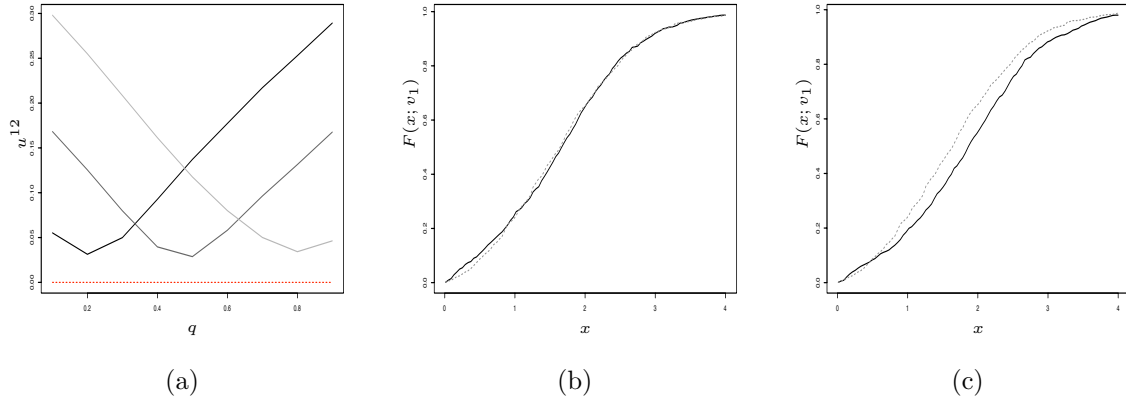


Figure 1: (a) Estimates of u^{12} for each $q^* = 0.2, 0.5, 0.8$; (b) Estimates of $F(x; v_1)$ for Z_{q^*} (dashed line) and Z_q (solid line) for $q = q^* = 0.5$; (c) Same as (b) but with $q^* = 0.5, q = 0.9$.

sist of realisations from M1 and M2 in equal proportions. The third is constructed with proportions 0.45, 0.5 and 0.05 respectively for M1–M3, with all M3 realisations occurring in the first 2,000 iterations. Accordingly there is a model probability and an extra component mismatch between chain 3 and the rest. Non-convergence should be detected up to 4,000 iterations, after which chain homogeneity should not be rejected. Reference points were sampled uniformly over all chains realisations. For samples with more than one component, we then randomly select one of these. Diagnostic values were again calculated using L^1 distance.

Figure 2 (a–c) respectively illustrates the diagnostic plots of u , $\max\{w^c : c = 1, 2, 3\}$ and $\max\{\hat{R}_v : v \in \mathcal{V}\}$ for increasing numbers of reference points. Means and maxima of the statistics were considered to reduce the complexity of the plots. Numbers of reference points considered were $|\mathcal{V}| = 1$ (lightest), 3, 9, 30, 60, 120, 240 (darkest). All monitored statistics exhibit strong similarity, and rapidly converge

after only a modest increase in the cardinality of \mathcal{V} . The case of only a single reference point is clearly the most variable. Beyond 4,000 iterations, all chains can be considered to have effectively converged. The figure shows that the PSRFv's are effectively one after this point, and the pairwise and Monte Carlo comparisons remain approximately constant and within 0.1 of zero, for the remainder of the chain. Note that Figures 2 (a,b) do not level off at zero as the sample-paths consist of non-identical realisations.

In general, the complexity of the problem and the total number of components are determining factors in the optimum number of reference points required. We postulate that around 100 reference points are prudent for most applications, although this should be verified in each case. We note that given the additive forms of (4) and (6), and the independent calculations of (7), extra reference points may be added sequentially until diagnostic stability is achieved.

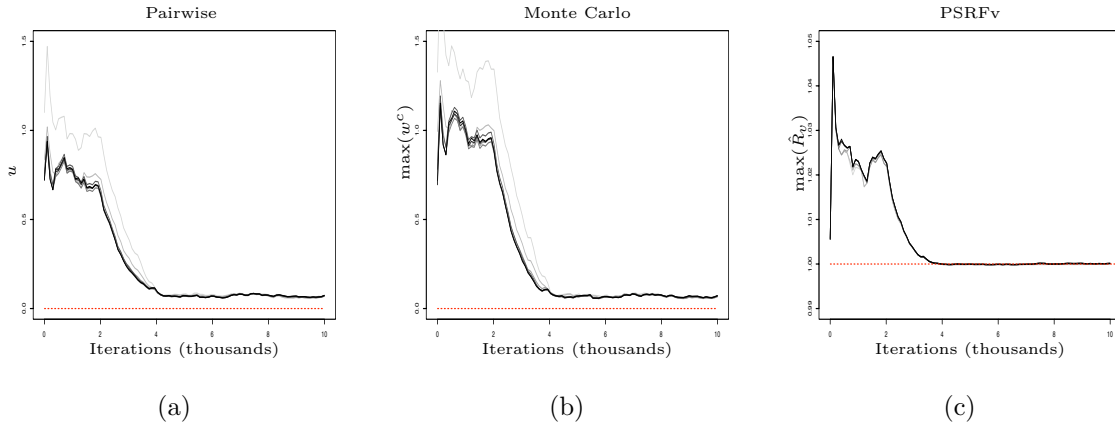


Figure 2: (a) mean of pairwise comparisons; (b) maximum of Monte Carlo tests; (c) maximum PSRFv. For varying numbers $|\mathcal{V}| = 1$ (lightest), 3, 9, 30, 60, 120, 240 (darkest).

4 Examples

We now illustrate the proposed trans-dimensional diagnostics and compare them to existing methods by consideration of both a theoretical example and two real data analyses. The analyses evaluate the well known mixture of Normals sampler of Richardson and Green (1997), which has been previously assessed by Brooks and Giudici (2000), and provide a re-examination of a change-point model analysis implemented before the development of any trans-dimensional convergence diagnostics.

4.1 Theoretical Example

Consider a point process model specification, where one of the unknown parameters enters the likelihood only through some even power; e.g. a mixture of Normals with parameter components (ψ, μ, σ) respectively denoting the component weight, mean and standard deviation, and component density $f(x) = (2\pi\sigma^2)^{-1/2} \exp\{-(x - \mu)^2/(2\sigma^2)\}$. The density π is then mirrored about the origin for all positive and negative combinations of σ in each component. Suppose we implement a fast-mixing reversible-jump sampler with split/merge between-model transitions. Suppose also that likely values of σ are significantly distant from zero, so that a split/merge step will never produce a change in sign. It follows that unless the sampler is instructed to consider negative in addition to positive values, the starting location of the chain will determine the region of parameter space in which the sampler will remain indefinitely. Now consider two chain implementations S_1 and S_2 with initial values of $(\psi, \mu, \sigma) = (\psi_0, \mu_0, \sigma_0)$ and $(\psi, \mu, \sigma) = (\psi_0, \mu_0, -\sigma_0)$ respectively, and with the same random number seed. The

two realised sample-paths will be identical, except that the signs of σ will be opposite.

How would the various trans-dimensional diagnostics evaluate the performance of S_1 and S_2 ? As the model probabilities are identical, Brooks et al. (2003) would conclude that both chains have converged immediately from initialisation. Both Brooks and Giudici (2000) and Castelloe and Zimmerman (2002) require monitoring a statistic that retains interpretation over all models, and the former propose the deviance as a default choice in the absence of preferred candidates. In this example, likelihood evaluations (and hence the deviance) will be identical for both chains, and accordingly immediate convergence is the only possible inference. Similar conclusions will be reached if the chains have different random seeds. When nearest-component distances from points $v \in \mathcal{V}$ taken from *both* chains are evaluated, the difference between components (ψ, μ, σ) and $(\psi, \mu, -\sigma)$ is apparent, resulting in large differences in the estimated $F(x; v)$. Accordingly a verdict of non-convergence is obtained.

4.2 Analysis of finite mixtures

A k -component finite mixture analysis assumes an underlying density of the form

$$f(y) = \sum_{j=1}^k \psi_j f_j(y|\theta_j)$$

for observed data $y = (y_1, \dots, y_n)$, where ψ_j is the weight of the j^{th} component. The distribution f_j , with parameter vector θ_j , is usually taken to be common across all components. Richardson and Green (1997) developed a reversible jump algorithm for the case where $f(y)$ constitutes a mixture of Normal densities under weak prior information. Here we consider this model and sampler implemented for the publicly

available “Enzyme” dataset. This sampler and dataset was also analysed by Brooks and Giudici (2000) in the particular context of sampler convergence assessment. Denoting a model configuration by $\xi = \{\xi_1, \dots, \xi_k\}$, for a finite mixture of Normals we would have $\xi_j = (\psi_j, \mu_j, \sigma_j^2) \in A, j = 1, \dots, k$, where $A = [0, 1] \times \mathbb{R} \times \mathbb{R}^+$, with μ_j and σ_j^2 representing the mean and variance of the j -th component. We implement five chain replications, S_1, \dots, S_5 , each of length 400,000. For the set of reference points, $v \in \mathcal{V}$, we randomly select twenty components from the full sample-path of each chain, providing $|\mathcal{V}| = 100$ in total.

Initially, we examine the performance of existing methods. Figure 3 (a,b) illustrates the diagnostic of Brooks et al. (2003) which provides a test of between-chain homogeneity with respect to model probabilities. These Kolmogorov-Smirnov (all pairwise comparisons) and χ^2 (all chains simultaneously) tests require independent realisations. Based on the estimated convergence rate (Brooks et al. 2003) we retain every 400th iteration to obtain approximate independence. If the statistics stay above the critical value, homogeneity cannot be rejected. Figure 3 (c) illustrates the two multivariate PSRF’s of Castelloe and Zimmerman (2002) using the deviance as the default statistic to monitor. The first (solid line) shows the ratio of between- and within- chain variation; hence a value of 1 indicates an absence of a chain effect. The second shows the ratio of within-model variation and the within-model, within-chain variation. A value of 1 indicates absence of a chain, model and chain-model interaction.

The Kolmogorov-Smirnov statistic cannot reject immediate convergence, with all pairwise chain comparisons well above the critical value of 0.05. The χ^2 statistic

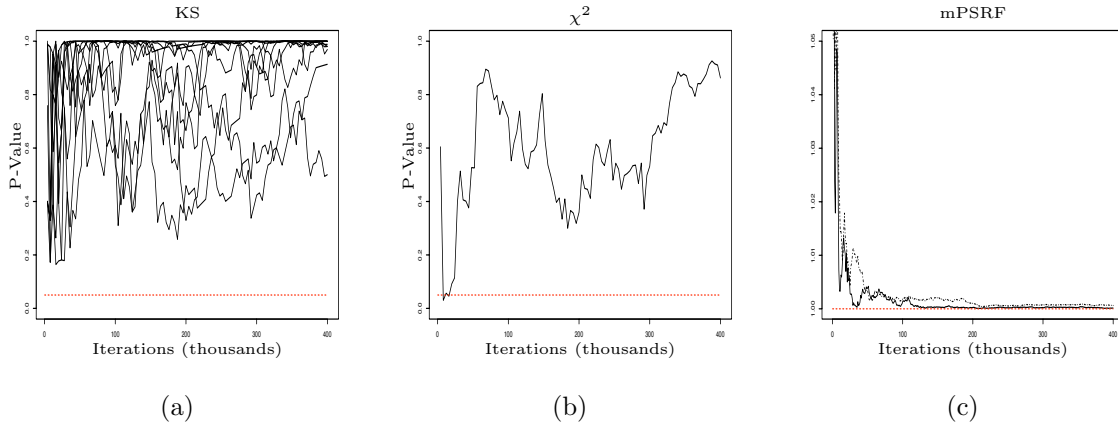


Figure 3: Convergence assessment for Enzyme data. (a,b) Kolmogorov-Smirnov and χ^2 tests of Brooks et al. (2003). (c) mPSRF's of Castelloe & Zimmerman (2002)

cannot reject convergence after the first 10,000 iterations. The mPSRF's of Castelloe and Zimmerman (2002) approach one rapidly, suggesting a lack of chain, model or interaction effect: both are within 2 decimal places (d.p.) of one beyond 43,000 iterations and within 3 d.p. beyond 166,000 iterations. This latter benchmark is supported by the independent analysis of Brooks and Giudici (2000) who demonstrate evidence for convergence after around 150,000 iterations, although they conclude that their chain lengths of only 200,000 iterations were too short for certainty.

In Figure 4 (a–c) we illustrate the plots of $u^{c_j c_k}$ given by (4), w^{c_k} given by (6) and \hat{R}_v given by (7). Here all three statistics tell a similar story: up to around 100,000 iterations variation between chains is plainly still reducing; beyond 300,000 iterations, differences between the chains appear to have stabilised. The intervening iterations mark a gradual transition between these two states. On closer inspection, Figure 4 (b) indicates that beyond 100,000 iterations two chains (solid and dotted

lines) are visibly different to the rest, which appear difficult to tell apart. We have to wait 100,000 more iterations before one chain (solid line) joins the rest, and a further 100,000 before all chains appear indistinguishable.

Taken together, the plots in Figure 4 suggest that there is evidence for between-chain homogeneity for three of the chains after 100,000 iterations, and similarly for all five chains after 300,000 iterations. Our more conservative estimate of convergence compared to those of the diagnostics in Figure 3 reflects the simultaneous monitoring of all model parameters over 100 ($v \in \mathcal{V}$) distributions, compared with just the model probability in Figure 3 (a,b), and the combined monitoring of model probability together with a single statistic; the deviance in Figure 3 (c).

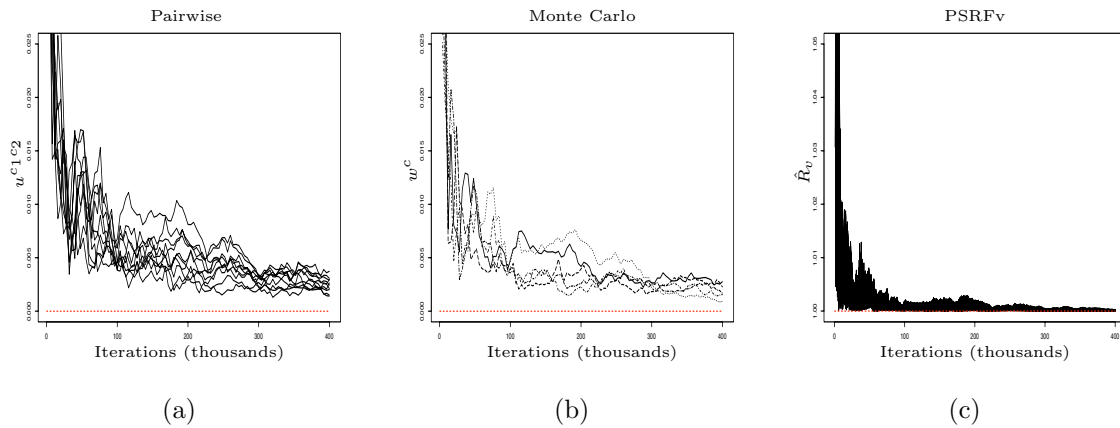


Figure 4: Convergence assessment for Enzyme data, using (a) pairwise comparisons $u^{c_j c_k}$; (b) Monte Carlo tests w^{c_k} ; (c) PSRFv's, R_{v_j} , for $j, k = 1, \dots, 5$.

4.3 A Bayesian change-point analysis

We now analyse sampler performance for a typical change-point problem. Models of this type may typically be expressed in a point process framework. Specifically we consider the modelling of prehistoric corbelled domes (late Minoan tholos data collected from Achladia on Crete (Cavanagh and Laxton 1982)), considered from a Bayesian perspective by Fan and Brooks (2000). This analysis was conducted prior to the availability of any suitable convergence diagnostics. Paired data arise in the form (d_i, r_i) , representing the distance, d_i , below the apex of the tomb at which the corresponding radius, r_i , are thought to approximately follow a log-linear model between a series of change-points. Thus the model is formulated as

$$\log(r_i) = \begin{cases} \log(\alpha_j) + \beta_j \log(d_i + \delta_j) + \epsilon_i & \gamma_{j-1} \leq d_i < \gamma_j \\ \log(c) + \epsilon_i & \gamma_J \leq d_i \end{cases},$$

with change-points $0 = \gamma_0 \leq \dots \leq \gamma_k$, subject to the continuity constraints

$$\alpha_j(\gamma_j + \delta_j)^{\beta_j} = \alpha_{j+1}(\gamma_j + \delta_{j+1})^{\beta_{j+1}}, \quad \alpha_k(\gamma_k + \delta_k)^{\beta_k} = c,$$

for $j = 1, \dots, k-1$, with $\epsilon_i \sim N(0, \sigma^2)$, and where the number of change-points, k , is also unknown. A full model specification may be realised as $\xi = \{\xi_1, \dots, \xi_k\}$ with $\xi_j = (\alpha_j, \beta_j, \delta_j, \gamma_j, \sigma_j^2)$, where σ_j^2 is set to equal to the error variance σ^2 . For the first component of each realisation, the change-point parameter γ_1 is always set to zero. Fan and Brooks (2000) implemented their sampler for 15 million iterations, and retained the final 7.5 million (thinned to 40,000) for the final analysis. We now determine whether this choice was justified. We implement five chains, S_1, \dots, S_5 , each of length 15 million and retain every 375th realisation to yield 40,000 iterations

in total. We form \mathcal{V} by randomly sampling 20 components from each chain.

Figure 5 illustrates the diagnostics of Brooks et al. (2003) and Castellote and Zimmerman (2002). Plots (a, b) were obtained by a further sub-sampling of every 30 iterations to attain approximate independence. Both Kolmogorov-Smirnov and χ^2 statistics are above the critical value of 0.05 after about 1.5 million iterations. Accordingly, based on model probabilities alone, we are unable to reject the hypothesis of between-chain homogeneity beyond this point in the simulation. Figure 5 (c) displays the mPSRF's of Castellote and Zimmerman (2002) again implemented using the deviance as the monitored statistic, Both plots converge rapidly to one; being within 2 d.p. of one after 800,000 iterations and within 3 d.p. after 3.1 million. Accordingly there is again a lack of evidence for chain, model or an interaction effect.

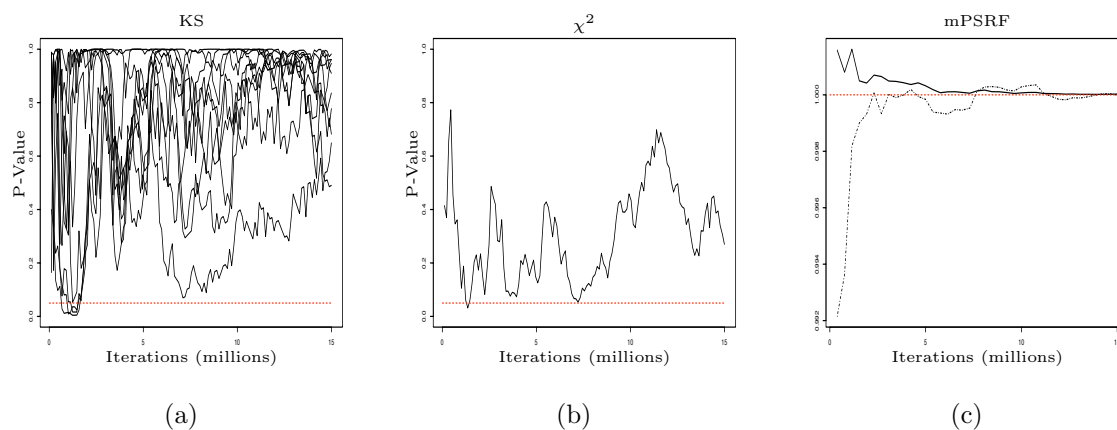


Figure 5: Convergence assessment for tombs data. (a,b) Kolmogorov-Smirnov and χ^2 tests of Brooks et al. (2003). (c) mPSRF's of Castellote & Zimmerman (2002).

An examination of diagnostics based on point-to-nearest-event distances, however, reveals a less than perfect accord among all chains (Figure 6). The plot of pairwise

comparisons, $u^{c_j c_k}$, indicates that the immediate differences between chains have been overcome by the 5 million iteration mark. However, there is a clear grouping of the statistics into those which contrast chain S_3 (grey lines) and those which do not. This divergence is most obvious in Figure 6 (b), where chain S_3 (dash-dot) visibly exhibits the largest discrepancy among all chains, from the 6 million iteration mark onwards. While this discrepancy is in decline towards the end of the simulation, it is by no means clear that chain homogeneity has been achieved at this point. Examination of the PSRFv's (Figure 6 (c)) reveals clear non-homogeneity between chains. Both before the 4 million iteration mark, and at intervals beyond the 7.5 million iteration mark three reference points are consistently indicating a lack of between-chain convergence.

Further investigation reveals that the parameter indicating the location of the first change-point, γ_2 , has not converged in chain S_3 . Figure 6 (d) shows density estimates of γ_2 from all five chains, estimated for the three component model with posterior model probability of approximately 0.18. Chain S_3 (black line) has a very different density estimate in the left tail. Revealingly, the γ parameter of the three reference points which exhibited marked deviations away from one in Figure 6 (c), are located in the left tail of the distribution as indicated on the graph.

In terms of the original work by Fan and Brooks (2000), given the lack of any trans-dimensional diagnostic tools at the time of their analysis, the authors were reasonably confident their chains had converged after 7.5 millions iterations. Based on the above, excluding chain S_3 , we might be reasonably confident that the chains had reached a sufficient degree of homogeneity to support a thesis of convergence at this point. However, as demonstrated, without the ability to identify, monitor and

possibly exclude any problem chains, an even longer burn-in may have been advisable.

5 Discussion

Perhaps the major difficulty with convergence diagnostics based solely on sample-path replicates is that an inability to distinguish between chains is a necessary but insufficient criterion to establish convergence. Only via reference to the analytic form of the posterior may claims of sufficient criteria be made (Fan et al. 2005, for example), and even then these should be well justified. This problem may be alleviated by increasing the number of chain replicates and using a dispersed set of starting points (Gelman and Rubin 1992), thereby reducing the probability that all chains have converged to the same “local” model. Descriptive measures of chain performance may be used to infer the likelihood that the chain has converged, but in addition offer summary descriptions of other aspects of chain performance such as mixing, chain efficiency and the relative performance of different samplers. Hence, output-based diagnostics remain the easiest and fastest way of monitoring MCMC chain performance, and therefore provide a valuable tool for the practitioner.

In this article we have introduced a new method for the assessment of trans-dimensional Markov chain samplers, based upon the distributions of point-to-nearest-event distances, when the underlying model may be formulated in a point process framework. The diagnostic describes the difference in intensity function estimates on component space between chain replicates, which naturally incorporates an assessment of between-model mixing as the distributions $F(x; v)$ are derived from the full

parameter vector. Estimates of the distributions $F(x; v)$ may then be contrasted in a number of ways, including via the well-known Gelman and Rubin diagnostic. The proposed diagnostic has the advantage that it oversteps many difficulties faced by existing methods, and it is straight-forward to implement. Similarly to Brooks and Giudici (2000) and Castellote and Zimmerman (2002), but unlike Brooks et al. (2003), this diagnostic may equally be applied in the fixed-dimensional setting.

One element that has not been discussed is that of the distance metric implied by x in Proposition 2.2. Here we have adopted Euclidean distance on the component space, A , however issues such as the relative scaling of parameters suggest consideration should be given to whether this metric is sensible. For example, in their distance-based loss function for the analysis of finite mixtures, Celeux et al. (2000) examined Euclidean distance of the transformed variables $\log(\psi/(1 - \psi)), \mu, \log \sqrt{\sigma^2}$ for Normal components in order that their random walk sampler would not be hindered by constraints upon the support. Similarly, Sisson and Hurn (2004) found that a distance metric of $d(v, \xi) = \min_j \|v - \xi_j\| / (1 + \|v - \xi_j\|)$ (where $\|\cdot\|$ denotes Euclidean distance), was effective in their analysis which benefited from a measure sensitive to local changes. In general, some consideration should be given to these issues in any implementation of sample-path comparisons.

Software for implementing this diagnostic tool is freely available at

<http://www.maths.unsw.edu.au/~scott>

Acknowledgements

SAS and YF are supported by the Australian Research Council through the Discovery Grant scheme (DP0664970), and YF by the Faculty of Science, UNSW. The authors wish to thank Professor S. P. Brooks, Professor P. J. Green and Juha Vierinen for the use of their software to implement various simulations in this article. Two of these are publicly available at <http://www.maths.bris.ac.uk/~peter>; <http://juha.vierinen.net/files/convergenceMan.html>.

References

- Berger, J. O. and L. R. Pericchi (1996). *Modeling and Prediction*, Chapter On the justification of default and intrinsic Bayes factors, pp. 276–293. Springer-Verlag.
- Besag, J. and P. J. Diggle (1977). Simple Monte Carlo tests for spatial patterns. *Appl. Stat.* 26, 327–333.
- Brooks, S. P. and P. Giudici (2000). MCMC convergence assessment via two-way ANOVA. *J. Comp. & Graph. Stat.* 9, 266–285.
- Brooks, S. P., P. Giudici, and A. Philippe (2003). On non-parametric convergence assessment for MCMC model selection. *J. Comp. & Graph. Stat.* 12, 1–22.
- Cappé, O., C. P. Robert, and T. Rydén (2003). Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers. *J. Roy. Stat. Soc. Ser. B* 65, 679–700.
- Castelloe, J. M. and D. L. Zimmerman (2002). Convergence assessment for re-

- versible jump MCMC samplers. Technical report, University of Iowa.
- Cavanagh, W. G. and R. R. Laxton (1982). Corbelling in the late minoan tholos tombs. *Annual of the British School at Athens* 77, 65–77.
- Celeux, G., M. Hurn, and C. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Stat. Assoc.* 95, 957–970.
- Cowles, M. K. and B. P. Carlin (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *J. Amer. Stat. Assoc.* 91, 883–904.
- Cowles, M. K. and J. S. Rosenthal (1998). A simulation approach to convergence rates for Markov chain Monte Carlo. *Stat. & Comp.* 8, 115–124.
- Cuzick, J. and R. Edwards (1990). Spatial clustering for inhomogeneous populations. *J. Roy. Stat. Soc. Ser. B* 52, 73–104.
- Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*, 2nd Ed. OUP.
- Diggle, P. J. and B. Matérn (1981). On sampling designs for the estimation of point-event nearest neighbour distances. *Scand. J. Stat.* 7, 80–84.
- Fan, Y. and S. P. Brooks (2000). Bayesian modelling of prehistoric corbelled domes. *The Statistician* 49, 339–354.
- Fan, Y., S. P. Brooks, and A. Gelman (2005). Output assessment for Monte Carlo simulations via the score statistic. *J. Comp. & Graph. Stat.*, To appear.
- Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–511.

- George, E. I. and R. E. McCulloch (1995). Stochastic search variable selection. In *Markov Chain Monte Carlo in Practice*. Chapman & Hall.
- Green, P. J. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika* 82, 711–732.
- Green, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems*, pp. 179–198. OUP.
- Johnson, V. E. (1998). A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *J. Amer. Stat. Assoc.*, 238–248.
- Moller, J. and R. P. Waagepetersen (2004). *Statistical Inference and Simulation for Spatial Point Processes*, Volume 100. Chapman & Hall/CRC Press.
- Mykland, P., L. Tierney, and B. Yu (1995). Regeneration in Markov chain samplers. *J. Amer. Stat. Assoc.* 90, 233–241.
- Nummelin, E. (1978). A splitting technique for Harris recurrent Markov chains. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 43, 309–318.
- Propp, J. G. and D. B. Wilson (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Rand. Struct. & Alg.* 9, 223–252.
- Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Stat. Soc. Ser. B* 59, 731–792.
- Roberts, G. O. and R. L. Tweedie (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stoch. Proc. & App.* 80, 211–229.
- Roberts, G. O. and R. L. Tweedie (2000). Rates of convergence of stochastically

- monotone and continuous time Markov models. *J. Appl. Prob.* 37, 359–373.
- Rosenthal, J. S. (1995). Minorization conditions and convergence rates for markov chain monte carlo. *J. Amer. Stat. Assoc.* 90, 558–566.
- Sisson, S. A. (2005). Trans-dimensional Markov chains: A decade of progress and future perspectives. *J. Amer. Stat. Assoc.* 100, 1077–1089.
- Sisson, S. A. and M. A. Hurn (2004). Bayesian point estimation of quantitative trait loci. *Biometrics* 60, 60–68.
- Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Stat.* 28, 40–74.

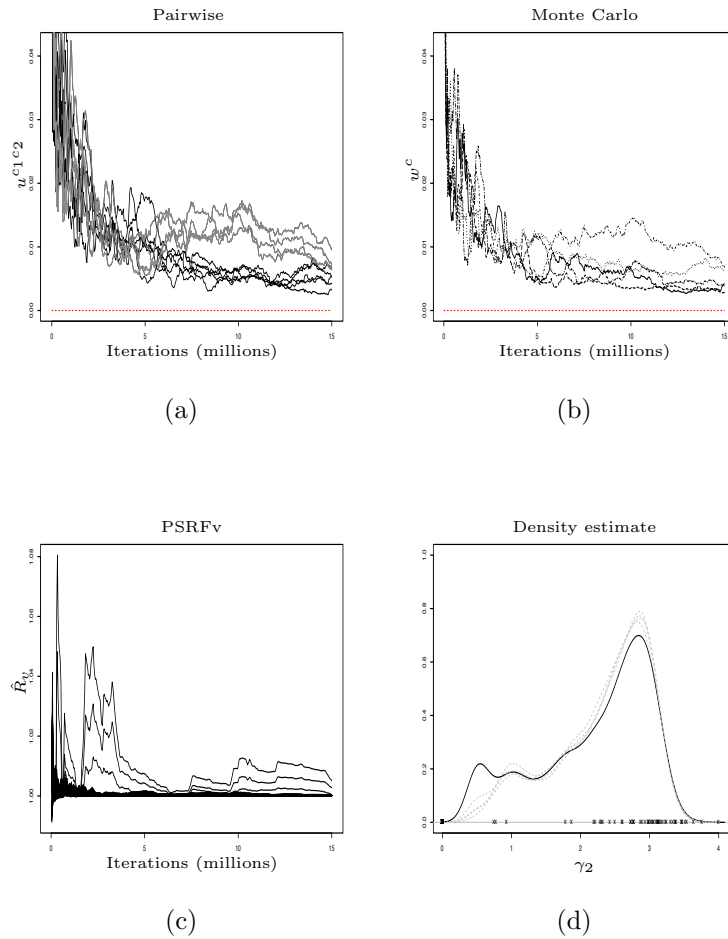


Figure 6: Convergence assessment for tombs data. (a) pairwise comparisons $u^{c_j c_k}$; (b) Monte Carlo tests w^{c_k} ; (c) PSRFv's, R_v , for $j, k = 1, \dots, 5$; (d) marginal density estimates of γ_2 (chain S_3 in black). Crosses indicate γ component of reference points.