Les Cahiers du GERAD ISSN: 0711–2440

On testing MCMC convergence in Bayesian clustering

M. Asgharian
V. Partovi Nia

G–2014–52

August 2014

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

On testing MCMC convergence in Bayesian clustering

Masoud Asgharian[a]

Vahid Partovi Nia[b]

[a] *Departement of Mathematics and Statistics, McGill University, Montréal (Québec) Canada, H3A 0B9*

[b] *GERAD & Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal (Québec) Canada, H3C 3A7*

masoud@math.mcgill.ca
vahid.partovinia@polymtl.ca

August 2014

Les Cahiers du GERAD

G–2014–52

Abstract: While there has been a surge of articles on convergence diagnostic tools for MCMC on continuous stationary distributions and ordinal state spaces, Bayesian clustering has spawned demands for tools designed specifically for nominal finite state spaces — grouping space. To fill this gap we propose a simple quantitative convergence criterion for MCMC algorithms run on nominal state spaces that has an intuitive interpretation which is a one-dimensional goodness-of-fit statistic. We study the asymptotic behaviour of the statistic and estimate its variance using the regenerative simulation. The convergence assessment is performed via a formal statistical significance test. We study the performance of the proposed criterion via simulation. We finally consider the particular application of clustering of genetic mutants of the flowering plant *Arabidopsis thaliana*.

Key Words: Bayesian clustering, convergence assessment, Gibbs sampling, Markov chain Monte Carlo algorithm, regenerative simulation, split-merge algorithm.

Acknowledgments: The authors would like to thank David Stephens for his comments on an earlier draft of this manuscript. Masoud Asgharian is supported by Natural Sciences and Engineering Research Council (NSERC) of Canada, Fonds de recherche du Québec, and Health Services Research and Development Service of the U.S. Department of Veterans Affairs grants. Vahid Partovi Nia is supported by the Swiss National Science Foundation Fellowship and NSERC Discovery Grant.

1 Introduction

Clustering may be described as the partitioning of data into homogeneous groups. Classical clustering techniques employ a measure of dissimilarity and optimize a criterion in order to determine the allocation of data to different groups (Hartigan, 1975). Modern approaches are based on probabilistic models where homogeneous groups of data follow the same distribution (Murua et al., 2008; Everitt et al., 2011). From a statistical modelling viewpoint, clustering may be regarded as fitting a mixture model with an unknown number of components. When the number of clusters is fixed, a maximum likelihood approach via the EM algorithm is commonly adopted to allocate observations. Asymptotic model selection criteria such as AIC or BIC are commonly used in practice to choose the number of clusters (Fraley and Raftery, 2002). Bayesian clustering is considered as an alternative to mixture modelling. In Bayesian clustering a prior distribution is assumed on parameters and on groupings (Heard et al., 2006).

The goal of hierarchical Bayesian clustering is to find the maximum *a posteriori* (MAP) allocation of data using agglomerative or divisive algorithms. Alternatively, approximate sampling from the posterior distribution of allocations is performed via MCMC algorithms (Liu, 2001; Robert and Casella, 2004). Often these samples are used for consensus clustering representation. However, the cardinality of the space of groupings renders MCMC sampling very challenging, even for a small number of observations. More precisely, let $s(T, C)$ denote the Stirling number of the second kind, i.e., the number of ways that T observations can be classified into C non-empty clusters. Then, the number of possible groupings of T observations is given by $B(T) = \sum_{C=1}^{T} s(T, C)$, called the Bell number. This number grows rapidly with T, but not faster than the factorial of T. Even for moderately sized data sets, the Bell number is enormous, for instance in our application with only $T = 14$ genetic mutants, $B(14) \approx 1.9 \times 10^8$, and for $T = 100, B(100) \approx 4.8 \times 10^{115}$.

We discuss Bayesian hierarchical clustering with an unknown number of clusters C, under the assumption that the marginal posterior distributions of groupings are available up to a normalizing constant, i.e., all model parameters can be integrated out. This assumption reduces the posterior state space to that of groupings, so an MCMC algorithm such as Metropolis-Hastings or Gibbs sampling may be implemented only on the space of groupings. In particular, unlike common Bayesian clustering approaches, there is no need to apply a trans-dimensional MCMC algorithm, such as the reversible jump MCMC algorithm of Green (1995) when the marginal densities have analytically closed forms. The reversible jump algorithm allows the MCMC sampler to traverse parameter subspaces of varying dimensionality. When all model parameters are integrated out, the parameter space is the space of groupings, and it is of fixed dimension with cardinality equal to the Bell number.

Whereas MCMC convergence diagnostic tools have been studied extensively for continuous stationary distributions (Cowles and Carlin, 1996), Bayesian clustering has spawned demands for tools designed specifically for nominal state spaces. To the best of our knowledge, the only quantitative technique well suited for this purpose is Zellner and Min (1995), which requires the availability of full conditional distributions, so it is not generally applicable, for instance, in the Metropolis-Hastings algorithm. Brooks et al. (2003) have proposed nonparameteric convergence diagnostic tests for Bayesian clustering based on Kolmogrov-Smirnov and Pearson statistics. They assume nearly independent data after subsampling. However, concerns exist in estimating the empirical average after subsampling (Geyer, 1992; MacEachern and Berliner, 1994). Furthermore, the asymptotic theory for the Kolomgrov-Smirov and the Pearson tests are established only for the i.i.d. sampling scheme.

We introduce a convergence diagnostic criterion that compares the empirical probability mass function (pmf) to the posterior pmf on groupings unlike the methods developed for ordinal variables that check the convergence implicitly in terms of the moments. Such implicit convergence assessments can be misleading, since they only test stability of some moments. As discussed in Section **??**, it is easy to construct examples where moments stabilize, possibly around a wrong value, long before the empirical pmf concentrates around the true posterior pmf of groupings. Our criterion is applicable to an irreducible, aperiodic Markov chain defined on a finite state space (Roberts and Rosenthal, 2004). This implies that the Markov chain can be the output of a more general class of MCMC sampling algorithms, including the Gibbs sampler.

In Bayesian clustering, the higher the posterior is for a grouping, the more that grouping is weighted in the consensus clustering diagram. So it is important to check the convergence of the chain at least on groupings with high posterior pmf. We test convergence to equilibrium through the ratio of the empirical pmf to the true pmf known up to a normalizing constant. Averaged over states under consideration, this gives an intuitive, variance-like, one-dimensional goodness-of-fit statistic. Using the Markov chain Central Limit Theorem (CLT) (Jones, 2004; Galin, 2004), we derive the asymptotic distribution of this variance statistic. Under the hypothesis of stationarity, we expect this statistic to be small, so we propose to reject the null hypothesis for large values. In practice, we estimate the asymptotic variance by regenerative simulation (Mykland et al. (1995); Hobert et al. (2002)).

Regenerative simulation is one of a number of specialized techniques, including overlapping batch means and spectral variance methods (Flegal and Jones, 2010), for consistent estimation of the variance of a statistic in a Markov chain. Jones et al. (2006) employ this technique for MCMC convergence diagnostics. They form an asymptotic confidence interval for the expectation of a given function of interest, e.g., the posterior mean, and propose to continue the MCMC simulation process until the length of the half-width interval falls below a pre-specified threshold. In this way, convergence of the empirical mean is assessed, and convergence to the equilibrium distribution is only implicitly tested in terms of the specified moment of the chain. As the moment of the chain is the target of convergence, their method is not applicable for nominal state spaces. In contrast, our significance test aims to assess lack of convergence to the equilibrium distribution directly.

This paper is organized as follows. Section 2 introduces Bayesian clustering and the challenges of MCMC sampling from the posterior distribution on groupings. In Section 3 we define our test statistic and derive its asymptotic distribution. We implement our convergence criterion for Gibbs and split-merge Metropolis-Hastings (MH) sampling algorithms in Section 4, with application to clustering genetic mutants of the flowering plant *Arabidopsis thaliana*.

2 Bayesian clustering

In Bayesian clustering, each observation has a corresponding unknown grouping parameter which assigns it to a specific cluster. Let $\mathbf{y} = \{y_t\}_{t=1}^T$ represent the observations and $\mathbf{c} = \{c_t\}_{t=1}^T$ the unknown grouping parameters called labels, i.e., $c_t = c \in \{1, \dots, C\}$ if y_t is allocated to cluster c. In order to impose uniqueness in cluster labeling, we assume that the grouping parameters are in increasing order. The first observation, y_1, always has label 1; the second observation has label 1 if it belongs to the same group as y_1; otherwise, it has label 2, and so forth. Furthermore, we assume that there are no empty clusters. The likelihood function is given by

$$\pi(\mathbf{y} \mid \theta, \mathbf{c}) = \prod_{c=1}^{C} \prod_{\{t; c_t = c\}} \pi(y_t \mid \theta, \mathbf{c}),$$

where θ is the unknown model parameter, possibly multi-dimensional. We assume that, conditional on $\mathbf{c}$ and the model parameters, the observations are independent within and across clusters called partition model (Hartigan, 1990). Since the goal is to estimate the grouping parameter $\mathbf{c}$, the ideal scenario involves fitting a model with closed-form marginal posterior distributions (Heller and Ghahramani, 2005; Heard et al., 2006). In other words, the model parameters are integrated out with respect to their prior distribution given $\mathbf{c}$:

$$\pi(\mathbf{y} \mid \mathbf{c}) = \int \left\{ \prod_{c=1}^{C} \prod_{\{t; c_t = c\}} \pi(y_t \mid \theta, \mathbf{c}) \right\} \pi(\theta \mid \mathbf{c}) d\theta. \tag{1}$$

The state space of interest is that of all possible allocations under the posterior distribution $\pi(\mathbf{c} \mid \mathbf{y}) \propto \pi(\mathbf{y} \mid \mathbf{c})\pi(\mathbf{c})$, where $\pi(\mathbf{c})$ is the prior distribution on allocations. The Rao-Blackwellization of Equation (1) reduces the variance of MCMC-based estimators and facilitates the exploration of $\pi(\mathbf{c} \mid \mathbf{y})$ by the MCMC algorithm. Current literature offers several choices for the prior distribution $\pi(\mathbf{c} \mid \mathbf{y})$ (McCullagh and Yang, 2006; Heard et al., 2006; Booth et al., 2008). We assess the sensitivity of the posterior distribution to the choice of $\pi(\mathbf{c})$ by introducing a clustering prior sensitivity parameter ξ; denote the posterior distribution by

$\pi_\xi(\mathbf{c} \mid \mathbf{y}) \propto \pi(\mathbf{y} \mid \mathbf{c}) \{\pi(\mathbf{c})\}^\xi$, $0 \leq \xi \leq 1$. The sensitivity parameter ξ ranges from 0 to 1, defining a class of distributions with priors evolving from the uniform distribution to the prior of interest $\pi(\mathbf{c})$.

3 Convergence criterion

Let $\{X_t\}_{t\geq 1}$ be an irreducible, aperiodic Markov chain with discrete state space S_M of cardinality M. Following the Bayesian clustering context, X_t represents a grouping. The value of X_t is an integer that hypothetically refers to a distinct grouping. Therefore a state and a grouping are interchangeable words, with cardinality M being equal to the Bell number.

Let $\mathbf{P} = \{\mathrm{P}_{i,j}\}_{i,j=1}^M$ denote the transition probability matrix. By the Ergodic theorem (Meyn and Tweedie, 1993), there exists a unique stationary distribution $\mathbf{\Pi} = \{\Pi_i, i \in S_M\}'$, such that $\mathbf{P\Pi} = \mathbf{\Pi}$, satisfying $\Pi_j = \lim_{k\to\infty} \mathrm{P}_{i,j}^{(k)}, \forall i, j \in S_M$, where $\mathrm{P}_{i,j}^{(k)}$ is the transition probability from state i to state j in k steps.

In practice, we have a finite length ergodic Markov chain $X = \{X_t, t = 1, \ldots, n\}$ that may not visit all possible states M, i.e. visits only $m \leq \min(n, M)$ distinct states. The chain X_t is the result of an MCMC sampling algorithm run on groupings, such as split-merge Metropolis-Hastings (MH) or Gibbs sampling over n iterations. In the sequel, we assume that m is constant, and we comment on relaxing this assumption in Section 6. In other words, we assume that the region of the state space not visited after n iterations has negligible posterior probability. We later provide a lower bound of the number of iterations n to observe high mass states. (Section 3.3) Denote the state space of this Markov chain by $S_{m,n}$. Furthermore, suppose that Π_i is known only up to a normalizing constant Z; this suffices to implement the MH algorithm. Denote

$$\Pi_i = \frac{\pi_i}{Z} \quad \forall i \in S_M,$$

where $Z = \sum_{i=1}^M \pi_i > 0$. We assume that the state space S_M is prohibitively large that enumerating all states to compute the normalizing constant is computationally infeasible.

In the sequel, we define the test statistic V_n and derive its asymptotic distribution. The asymptotic variance can be consistently estimated using regenerative simulation under moment conditions that are relatively easy to verify.

The method of regenerative simulation identifies random times at which the Markov chain probabilistically restarts itself, by constructing a split chain $\tilde{X} = \{(\tilde{X}_1, \delta_1), (\tilde{X}_2, \delta_2), \ldots\}$ on space $S_M \times \{0, 1\}$, such that if $\delta_i = 1$, then $i + 1$ is a regeneration time. The construction of $\tilde{X}$ is based on the following minorisation condition: find a function $h : S_M \mapsto [0, 1]$ for which $\mathbb{E}_\Pi h = \sum_{i\in S_M} h(i)\Pi_i > 0$ and a probability measure Q such that, for all $x \in S_M$ and all measurable sets A,

$$\mathrm{P}(x, A) \geq h(x)Q(A). \tag{2}$$

Since S_M is countable, Equation (2) is satisfied by $h(x) = \mathbb{I}(x = i)$, for a fixed state $i \in S_M$ where $\mathbb{I}$ is the indicator function, and $Q(\cdot) = \mathrm{P}(i, \cdot)$. Generating the split chain is simple: $\tilde{X}_t = X_t$ and $\delta_t = \mathbb{I}(X_t = i)$ (Hobert et al. 2002). Assuming that X is initialized with $X_1 \sim \mathrm{P}(i, \cdot)$, then the chain probabilistically restarts itself at times $\tau_0 = 1, \tau_1 > \tau_0, \ldots$ defined by $\tau_{r+1} = \min\{t > \tau_r : \delta_{t-1} = 1\}, r \geq 0$, i.e., these correspond to the events $X_{\tau_{r+1}-1} = i$ when the chain returns to i. Constructing the split chain does not assume that the Markov chain is stationary; the only requirement is that $X_1 \sim Q$. Let $n_r = \tau_r - \tau_{r-1}, r \geq 1$ denote the length of the rth regeneration tour, and let $R(n)$ be the total number of tours in a chain of length n. For simplicity of notation, suppress the dependence on the number of iterations n and denote $R(n)$ by R. Since the Markov chain is aperiodic, it follows that $R \to \infty$ as $n \to \infty$.

Let g be a real-valued, Π-integrable function on S_M. The Ergodic Theorem implies that

$$\bar{g}_{\tau_R} = \frac{1}{\tau_R - 1} \sum_{t=1}^{\tau_R - 1} g(X_t) \to \mathbb{E}_\Pi g = \sum_{i\in S_M} g(i)\Pi_i$$

with probability 1 as $R \to \infty$. Note that τ_R is the start of the $(R+1)$st regeneration tour, hence the limits of the summation. In the sequel, the subscript Π indicates that the distribution in question is the stationary distribution Π. Furthermore, Hobert et al. (2002) show that, under the assumption that the minorisation condition holds, if X is geometrically ergodic and $\mathbb{E}_\Pi |g|^{2+\epsilon} < \infty$ for some $\epsilon > 0$, then the following CLT result is true:

$$\sqrt{R}\,(\bar{g}_{\tau_R} - \mathbb{E}_\Pi g) \xrightarrow{D} \text{Normal}_1(0, \sigma_g^2) \quad \text{as } R \to \infty, \tag{3}$$

where $\sigma_g^2 < \infty$, and Normal_d denotes the d-variate normal distribution. Moreover, a consistent estimator of σ_g^2 exists (Jones et al., 2006). Hobert et al. (2002) explain that the asymptotic variance σ_g^2 is related to the familiar Markov chain CLT result of Chan and Geyer (1994)

$$\sqrt{\tau_R - 1}\,(\bar{g}_{\tau_R} - \mathbb{E}_\Pi g) \xrightarrow{D} \text{Normal}_1(0, \gamma_g^2) \quad \text{as } \tau_R \to \infty, \tag{4}$$

where $\gamma_g^2 = \text{var}_\Pi\{g(X_1)\} + 2\sum_{k=2}^{\infty} \text{cov}_\Pi\{g(X_1), g(X_k)\} < \infty$, by $\sigma_g^2 = \gamma_g^2 \mathbb{E}_\Pi h$. In our case, this gives $\sigma_g^2 = \gamma_g^2 \Pi_i$.

3.1 Test statistic

The empirical estimator

$$\hat{\pi}_i = (\tau_R - 1)^{-1} \sum_{t=1}^{\tau_R - 1} \mathbb{I}(X_t = i),$$

is consistent as $R \to \infty$ by the Ergodic theorem. Recall that Π_i is the posterior mass at state i (often unobservable) and π_i is the posterior mass up to the normalizing constant Z; recall that π_i is used in calculation of the acceptance probability of the Metropolis-Hasting sampler, so it is available. As a result, for R large, we expect the ratio $f_i = \hat{\pi}_i/\pi_i$ to be approximately equal to Z^{-1}, for all $i \in S_{m,n}$. The intuition is that, under equilibrium, the ratios f_i are approximately constant; hence, we suggest using their variance to assess convergence to stationarity. Define the variance test statistic

$$V_n = \frac{R}{m} \sum_{i \in S_{m,n}} \left(f_i - \overline{f}\right)^2, \tag{5}$$

where $\overline{f} = m^{-1} \sum_{j \in S_{m,n}} f_j$.

Remark 3.1 *For R large, we have the approximation*

$$V_n \approx \frac{R}{m} \sum_{i \in S_{m,n}} (f_i - Z^{-1})^2 = \frac{R}{m} \frac{1}{Z^2} \sum_{i \in S_{m,n}} \frac{(O_i - E_i)^2}{E_i^2},$$

where $O_i = n\hat{\pi}_i$ is the observed number of visits to state i, and $E_i = n\Pi_i$ is the expected number of visits under the stationary distribution. Therefore, V_n somehow mimics the Pearson goodness-of-fit statistic, but with denominator E_i^2. This means that distances corresponding to states that have low probability Π_i are more heavily weighted.

Remark 3.2 *The stabilizing coefficient R/m is required to avoid obtaining an asymptotically degenerate distribution for V_n as $R \to \infty$. The numerator R comes from the stabilizing rate for $\hat{\pi}_i$ of order $R^{\frac{1}{2}}$ in Equation (3), and the denominator m gives an intuitive interpretation of V_n as the variance of the ratios f_i.*

First, we simplify the expression of V_n as follows.

$$\begin{aligned} V_n &= \frac{R}{m} \sum_{i \in S_{m,n}} \left(f_i - \frac{1}{m} f_i - \frac{1}{m} \sum_{\substack{j \in S_{m,n} \\ j \neq i}} f_j \right)^2 \\ &= \frac{R}{m} \sum_{i \in S_{m,n}} \left(\mathbf{a}_i' \mathbf{f}\right)^2 \\ &= \frac{R}{m} (\mathbf{A}\mathbf{f})'(\mathbf{A}\mathbf{f}), \end{aligned}$$

with the following notation: $\mathbf{f} = (f_1, \ldots, f_m)'$, for $i \in S_{m,n}$, $\mathbf{a}_i = \mathbf{e}_i - m^{-1}\mathbf{1}_m$, where $\mathbf{e}_1, \ldots, \mathbf{e}_m$ are the standard basis vectors of $\mathbb{R}^m$ and $\mathbf{1}_m$ is a column vector of 1s of length m, and $\mathbf{A}$ is the $m \times m$ symmetric matrix with ith row equal to $\mathbf{a}_i'$ and $\mathbf{A}'$ is the transpose of $\mathbf{A}$. Moreover, we notice that for all $i \in S_{m,n}$,

$$\begin{aligned}\left(\mathbf{f} - Z^{-1}\mathbf{1}_m\right)' \mathbf{a}_i &= \left(1 - \frac{1}{m}\right) \frac{\hat{\pi}_i - \Pi_i}{\pi_i} - \frac{1}{m} \sum_{\substack{j \in S_{m,n} \\ j \neq i}} \frac{\hat{\pi}_j - \Pi_j}{\pi_j} \\ &= f_i - \overline{f}.\end{aligned}$$

This result gives the useful representation $V_n = (\mathbf{C}\mathbf{w}_n)'(\mathbf{C}\mathbf{w}_n)$, where $\mathbf{w}_n = (w_{1,n}, \ldots, w_{m,n})'$, with $w_{i,n} = \sqrt{R}(\hat{\pi}_i - \Pi_i)$, and $\mathbf{C}$ is the $m \times m$ symmetric matrix defined by

$$\mathbf{C} = \mathbf{A} \times \mathrm{diag}\left\{(\sqrt{m}\pi_i)^{-1}\right\}.$$

Theorem 3.1 presents the asymptotic distribution of V_n as $R \to \infty$. See the Appendix for proof.

Theorem 3.1 *For an irreducible, aperiodic, discrete state space Markov chain with equilibrium distribution* Π*,*

$$\mathbf{C}\mathbf{w}_n \xrightarrow{D} \mathrm{Normal}_m(\mathbf{0}, \mathbf{C}\Sigma\mathbf{C}') \quad \text{as } R \to \infty,$$

for m fixed, where $\mathbf{\Sigma}$ is the asymptotic variance-covariance matrix of $\mathbf{w}_n$. Consequently,

$$V_n = (\mathbf{C}\mathbf{w}_n)'(\mathbf{C}\mathbf{w}_n) \xrightarrow{D} \sum_{i=1}^{m} \lambda_i Z_i^2,$$

where $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of the matrix $\mathbf{C}\Sigma\mathbf{C}'$, and $Z_i \overset{\text{iid}}{\sim} \mathrm{Normal}_1(0,1)$, $i = 1, \ldots, m$.

3.2 Implementation

In consensus clustering high mass states play a major role in the depiction of the consensus graph, so we restrict our attention to k high mass states. These states are, approximately, the highest posterior states. Below we describe regenerative sampling. This method helps to estimate the variance-covariance matrix $\mathbf{\Sigma}$ consistently. As a note of caution, the performance of regenerative simulation suffers when the state space explored is large. Therefore, we suggest that all the remaining states be merged and renamed as the new state $k+1$. We propose a hypothesis test for convergence at confidence level $(1-\alpha)$, for a fixed α. We assume that the number of regeneration tours R is large enough for the distribution of V_n to be well approximated by that of $\sum_{i=1}^{k+1} \lambda_i Z_i^2$. The eigenvalues λ_i are unknown, but they can be estimated by $\hat{\lambda}_i$, the eigenvalues of the estimator $\hat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$. Each entry in $\hat{\mathbf{\Sigma}}$ is a consistent estimator obtained by regenerative simulation (Mykland et al. (1995); Hobert et al. (2002)). Details appear in the Appendix.

1. Set $t = n$.
2. Run the MCMC algorithm for t iterations, and let i be the most frequently visited state. Split the chain into R regeneration tours defined by return visits to state i.
3. Compute the statistic V_t, the eigenvalue estimates $\hat{\lambda}_i$, $i = 1, \ldots, m$, and the p-value p_t; see the Appendix for details.

   If $p_t \leq \alpha$, reject the null hypothesis, continue for further n iterations, i.e. set $t = t + n$ and return to Step 2.
4. If $p_t > \alpha$, there is no evidence against the null hypothesis that the chain is in equilibrium by iteration t.

We consider the Lugannani and Rice saddlepoint approximation and its extension (Wood et al., 1993) to estimate the cumulative distribution function of $\sum_{i=1}^{k+1} \lambda_i Z_i^2$ and approximate the corresponding p-value.

The computational complexity of implementing this convergence criterion is dominated by the cost of computing $\hat{\mathbf{\Sigma}}$, of order $O(kR + k^2)$. See the Appendix for suggestions on reducing the computation time. For

k small, the method is rather fast and the suggested convergence criterion can be used as a stopping rule. When a small value of k is chosen, more states are merged as the $k+1$th state so the convergence assessment becomes less precise. We suggest assurance of convergence at least for $k=1$ as a minimum requirement for using the consensus clustering diagram. This happens in our example for the MH algorithm.

3.3 Minimum burn-in

Intuitively a chain with a large number of iterations n has a higher chance of visiting states with small posterior probability Π_i. A formal result on how to choose a minimum sample size for observing a state with a given probability can, however, be of practical concern in many applications. Theorem 3.2 below provides a result of this kind under equilibrium. In other words, it provides a lower bound for burn-in sample. Suppose E_i is the event of observing state i at least once in a reversible chain during a run of n iterations with P_{ij} being its one step transition probability.

Theorem 3.2 *Let $\epsilon > 0$ and $\xi > 0$ be given. Suppose $\Pi_i > \xi$ is the probability of state i at equilibrium. Then, for a reversible chain, $P(E_i) > 1-\epsilon$ if*

$$n > \frac{\log \epsilon}{\log\left\{(1-\frac{\xi}{1-\xi})(1-P_{i,i})\right\}} \tag{6}$$

See Appendix for the proof.

Theorem 3.2 is a guide for providing a lower bound for the number of iterations required to visit the state whose probability is greater than a given threshold. As expected, this lower bound depends on the probability that the chain remains at the state i which, in turn, reflects how good the mixing is. This fact was discussed by Peskun (1973) who showed that a chain with a smaller transition matrix trace has better mixing. Equation (6) suggests the better the mixing is, the less the sample size required for visiting a state with a given probability. Given that $P_{i,i}$ is unknown in many applications, it should be estimated from a preliminary sample. Under the best scenario, i.e. $P_{i,i}=0$, we have $n > \frac{\log \epsilon}{\log\left(1-\frac{\xi}{1-\xi}\right)}$. If for instance we wish the chance of observing a state whose posterior probability is greater than 0.001 be larger than 0.9999, the chain must be run at least for $n > 9196$ iterations. Note that this result would not apply for non-reversible chains such as the ordinary Gibbs sampler.

4 Case study

Messerli et al. (2007) study the metabolic pattern of 14 genetic mutants of *Arabidopsis thaliana* from measurements of 43 metabolites (mostly sugars, sugar alcohols, amino acids and organic acids), obtained by the method of gas chromatography mass spectrometry. Figure 1 presents the data, where mutants are represented by integer labels, and four replicates are available for each mutant; exceptionally, for mutant 1, only three replicates exist. These genetic mutants are described as follows: two mutants are defective in starch biosynthesis (13,14); four are defective in starch degradation (9–12); a comparative mutant accumulates starch as a pleiotropic effect (8); four are uncharacterized starch-excess mutants (4–7); and finally, three are wild-type plants (1–3). The goal is to a perform metabolomic characterization of these mutants via clustering.

4.1 Data modelling

We fit the following hierarchical Bayesian model, suitable for high-dimensional, small-sample data sets (Partovi Nia and Davison, 2012). Given the data allocation vector $\mathbf{c}$,

$$\begin{aligned} y_{vctr} - \mu &\overset{\text{iid}}{\sim} \text{Normal}_1(\gamma_{vc}\theta_{vc} + \eta_{vct}, \sigma^2) \\ \gamma_{vc} &\overset{\text{iid}}{\sim} \text{Bernoulli}(p) \\ \theta_{vc} &\overset{\text{iid}}{\sim} \text{Normal}_1(0, \sigma_\theta^2) \\ \eta_{vct} &\overset{\text{iid}}{\sim} \text{Normal}_1(0, \sigma_\eta^2), \end{aligned} \tag{7}$$

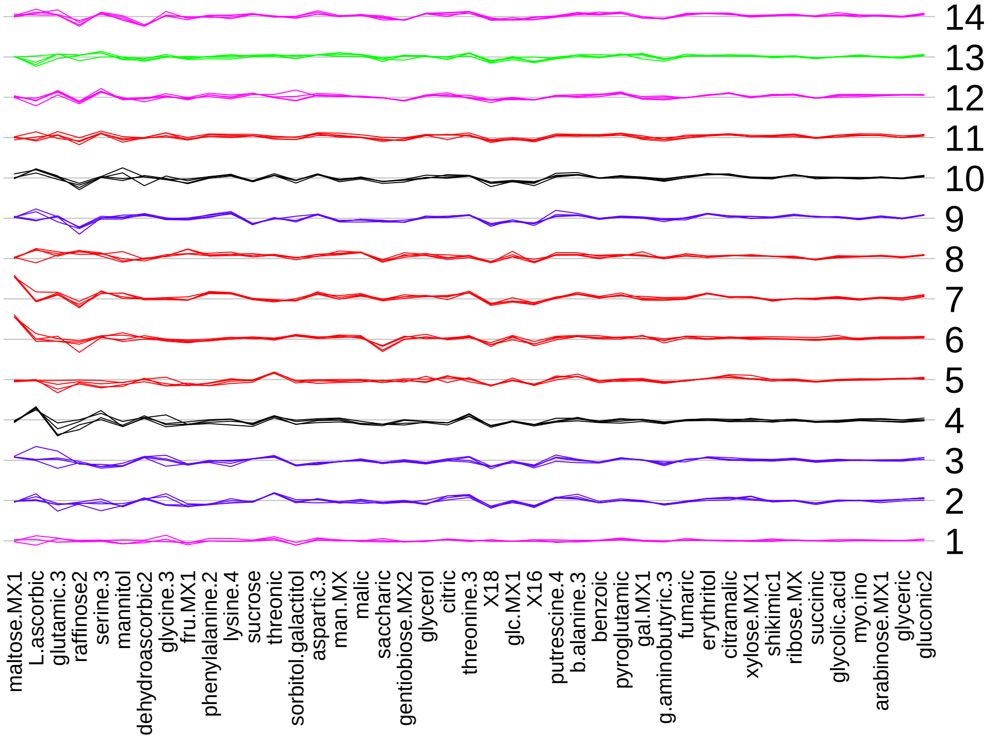

Figure 1: Plot of the log spectra (solid lines) of the metabolite data. Different colors indicate the category of mutant: black for those defective in starch biosynthesis, red for those defective in starch degradation, green for the comparative plant, blue for the uncharacterized mutants, and magenta for the wild types.

where Bernoulli(p) denotes the Bernoulli distribution with success probability p. The subscripts $v = 1, \ldots, V$, $c = 1, \ldots, C$, $t = 1, \ldots, T_c$, $r = 1, \ldots, R_{ct}$ denote, respectively, variable, cluster, mutant in cluster, and replicate, where V is the number of variables, C is the number of clusters, T_c is the number of mutants in cluster c, and R_{ct} is the number of replicates of mutant t in cluster c. The Bernoulli variable γ_{vc} controls the appearance of the clustering mean θ_{vc} to adjust for noise variables. The continuous parameter η_{vct} is added to account for the between-mutant error in cluster c. The model parameters σ^2 and σ^2_η are the between-replicate and between-mutant variance components, respectively, while σ^2_θ is the variance of the disappearing random mean component θ_{vc}.

Under the model specification in (7), the model parameters η_{vct}, θ_{vc}, and γ_{vc} can be integrated out, a marginal likelihood mixture of two Normal distributions for each replicate

$$p\text{Normal}_1(\mu, \sigma^2 + \sigma^2_\eta + \sigma^2_\theta) + (1-p)\text{Normal}_1(\mu, \sigma^2 + \sigma^2_\eta).$$

Hyperparameters $\mu, \sigma^2_\eta, \sigma^2_\theta, \sigma^2$ and p are estimated using the empirical Bayes approach. The estimated parameters and their asymptotic standard errors are $\mu = 0.083(0.028), \sigma^2 = 0.159(0.005), \sigma^2_\theta = 5.100(2.721), \sigma^2_\eta = 0.373(0.033)$, and $p = 0.034(0.019)$.

Following Heard et al. (2006), we assume that the assignment of mutants to clusters is exchangeable. So it suffices to specify a prior on T_c's, the number of observations in cluster c ($c = 1, \ldots, C$), and on the total number of clusters C, where $\sum_{c=1}^{C} T_c = T$ is the total number of mutants to be clustered:

$$\pi(\mathbf{c}) = \Pr(T_1, \ldots, T_C \mid C) \Pr(C),$$

and a uniform discrete prior is suggested for the total number of clusters,

$$\Pr(C = c) = 1/T, \quad c = 1, \ldots, T,$$

and the uniform multinomial-Dirichlet distribution is placed on the cluster totals given the number of clusters. This yields the following prior

$$\pi(\mathbf{c}) \propto \frac{(C-1)!T_1!\dots T_C!}{T(T+C-1)!}.$$

Figure 6 (left panel) presents the Bayesian hierarchical agglomerative clustering dendrogram, built using the posterior distribution of the data allocation vector as the similarity measure (Partovi Nia and Davison, 2012). The dendrogram suggests the presence of five clusters in our data set. As the agglomerative method may result in a poor approximation of the MAP clustering, we also explore the space of partitions by MCMC sampling (Rasmussen, 2000). We consider the following two algorithms: (i) a reversible Gibbs sampler, i.e., a Gibbs sampler that updates cluster labels in a random order; and (ii) a split-merge Metropolis-Hastings sampler. The split-merge sampler (Jain and Neal, 2004) uses Metropolis-Hastings updates to explore the space of cluster allocations via split and merge moves with restricted Gibbs scans embedded within. The advantage is that groups of observations can be updated at one time, and, if the proposed move is supported by the data, then it is likely to be accepted, see Jain and Neal (2004) and Jain and Neal (2007) for more discussion. In contrast, the Gibbs sampling algorithm with incremental updates is prone to becoming trapped in local modes of the posterior distribution. We have implemented the split-merge sampler with five local Gibbs scans and run the two samplers for $n = 5 \times 10^4$ iterations. The results presented are for sensitivity parameter $\xi = 0.5$; other choices of ξ gave similar results, suggesting that the prior $\pi(\mathbf{c})$ does not play a major role in the analysis.

To understand the performance of our proposed model better and make sure that our criterion truly detects convergence/lack of convergence, we have calculated the posterior for all $1.9 \times 10^8 = B(14)$ possible groupings for our case study. This formidable task becomes, of course, rather infeasible as the number of subjects to cluster increases. The computation 1.9×10^8 possible groupings posterior took two weeks on a Linux UBUNTU 12.04 LTS machine with 16 GB RAM, run on 8 parallel processors. This calculation provided us the true normalizing constant and assured us that both samplers visited the true MAP groupings.

4.2 Convergence comparison with moment-based method

The moment-based method is appropriate for ordinal state spaces while clustering is concerned with nominal states (groupings). For the sake of comparison we may therefore extract a binary Markov chain from the MCMC samples of groupings by considering a specific grouping. In another words, the finite state-space chain $X_t = 1$ if the chain is in a specific grouping or $X_t = 0$ otherwise. Then the convergence of the chain is tested with respect to this grouping. We suggest that the test be run on the most frequently visited state (grouping). This state is a proxy for the highest mass posterior state for a good sampler and hence it plays a key role in making inference and even sometimes is the objective of an MCMC run. Therefore, before starting the testing procedure, the maximum observed grouping must be found in the grouping chain, then a binary chain must be constructed and fed into the convergence testing procedure.

The maximum a posteriori grouping estimation using the Gibbs sampler and using the MH sampler yield the same grouping. Note that the MAP grouping found by both MCMC samplers is different to the grouping found by the agglomeration method, see Figure 6. Both samplers suggest the following three clusters $\{6,7\}$, $\{1,8\}$, $\{2,4,5,13,3,9,10,11,12,14\}$ as the maximum a posteriori grouping, with the estimated probability $\hat{\pi}_i = 0.43$ using the Gibbs sampler and $\hat{\pi}_i = 0.34$ using the MH sampler. This difference in the estimated probabilities motivated us to run our convergence test to see if both chains have converged. As the chain is binary, we may use the convergence method of Jones et al. (2006) developed for ordinal state spaces. Jones et al. (2006) recommend simulating the Markov chain until the length of the half-width confidence interval falls below a specified threshold ϵ.

Figure 3 presents the length of the half-width confidence interval for the rth posterior moment for $r = 1$, computed every 100 iterations; this length is compared to threshold $\epsilon = 0.05$ confirming convergence of both chains. Figure 2 confirms convergence of the Gibbs sampler after about 8,000 iterations, but never accepts convergence of the MH sampler. This result conflicts with that of Jones et al. (2006), as depicted in Figure 3,

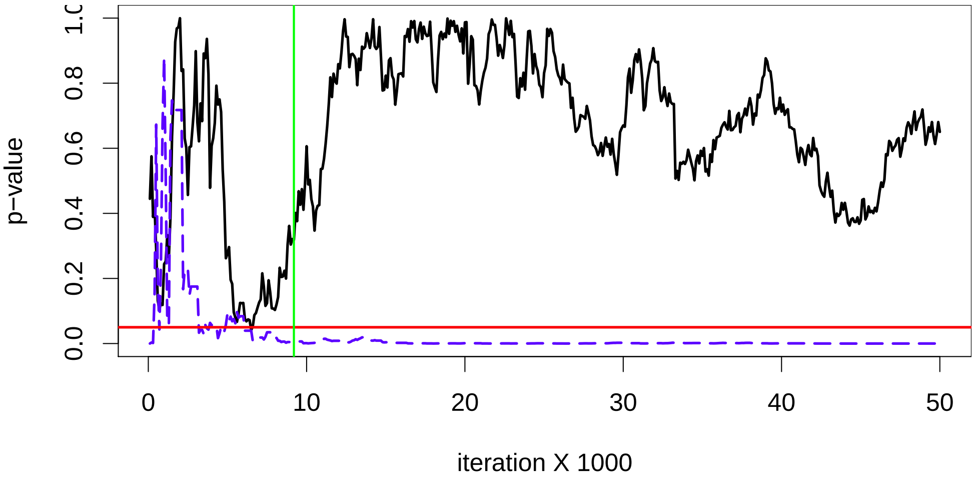

Figure 2: Convergence of our suggested method when a binary Markov chain is extracted from the MCMC samples of groupings for data of Figure 1. The solid line is the Gibbs sampler and the dashed line is the split-merge Metropolis-Hastings sampler. Plot of p-value of V_n versus the number of iterations n. A horizontal line is drawn at 0.05 as the threshold for the p-values. The samples are converged if the curves fall *above* the threshold. The vertical line represents the minimum sample size of observing a state with posterior larger than 0.001 with probability 0.9999, derived using Theorem 3.2, see also Figure 3.

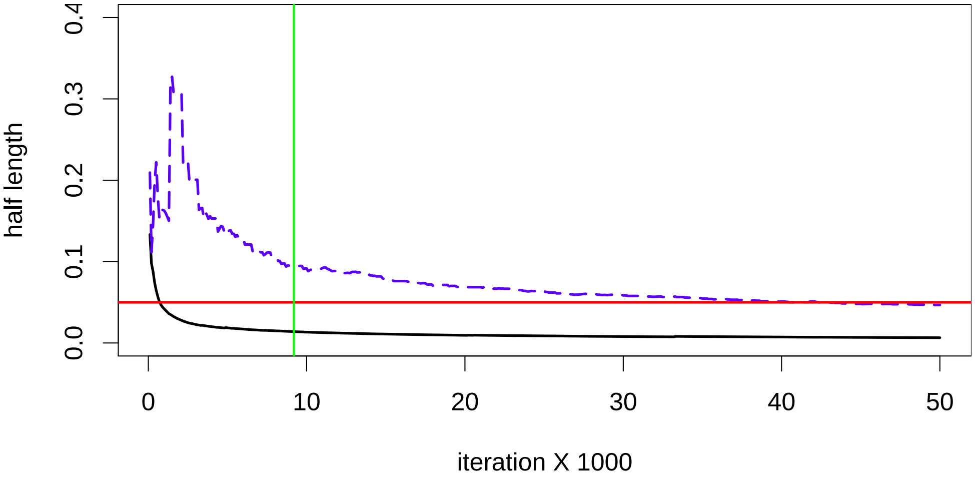

Figure 3: Convergence methods of Jones et al. (2006) when a binary Markov chain is extracted from the MCMC samples of groupings for data of Figure 1. The solid line is the Gibbs sampler and the dashed line is the split-merge Metropolis-Hastings sampler. The vertical axis is the half length of confidence interval ($\alpha = 0.05$) for the mean of the two chains. The horizontal line shows the threshold value at $\varepsilon = 0.05$. The samples are converged if the curves fall *below* the threshold. The vertical line represents the minimum sample size of observing a state with posterior larger than 0.001 with probability 0.9999, derived using Theorem 3.2, see also Figure 2.

accepts convergence for both chains. The main reason for this conflicting result is that our criterion uses the extra available information (the posterior being known up to a constant), but Jones et al. (2006) only check the moment stability (here the empirical probability) of a specified state. We, in addition, check whether this stabilized empirical probability matches the posterior. Figures 2 suggests that even for grouping only 14 observations, a chain with 50,000 iterations may not satisfy the minimum convergence property for the MH sampler.

The conflict of convergence between the Gibbs and the MH sampler using our criterion motivated us to study our test on a binary chain in more details. We consider two different situations: (i) binary chains having different mixing and (ii) binary chains having the same mixing, but one chain being compared to a slightly incorrect posterior. Wang (1981) introduces the Markov-Bernoulli chain with stationary distribution Bernoulli(p) and dependence parameter $\rho \in (0, 1)$, where the correlation between trials that are k steps apart equals ρ^k. The autocorrelation ρ^k reflects the mixing, the smaller the ρ, the better the mixing will be.

We generate two Markov-Bernoulli chains of length $100,000$ with $p = 0.43$ and $\rho = 0.1$, or $\rho = 0.9$. Figure 4 shows the p-value computed every 200 iterations against the threshold of 0.05. The criterion confirms that the chain with a better mixing, $\rho = 0.1$, converges to the stationary distribution faster than the chain with the lower mixing $\rho = 0.9$. For the former chain, the p-value is almost always above the threshold, which indicates that there is no evidence to reject the null hypothesis of stationarity. In comparison, for the latter chain, after 1,000 iterations, the p-value drops below the threshold of 0.05, and remains around this value until the end of the run. As both chains have the same stationary distribution we expect that the p-value for the chain with the lower mixing will rise eventually, but this phenomenon requires a longer run.

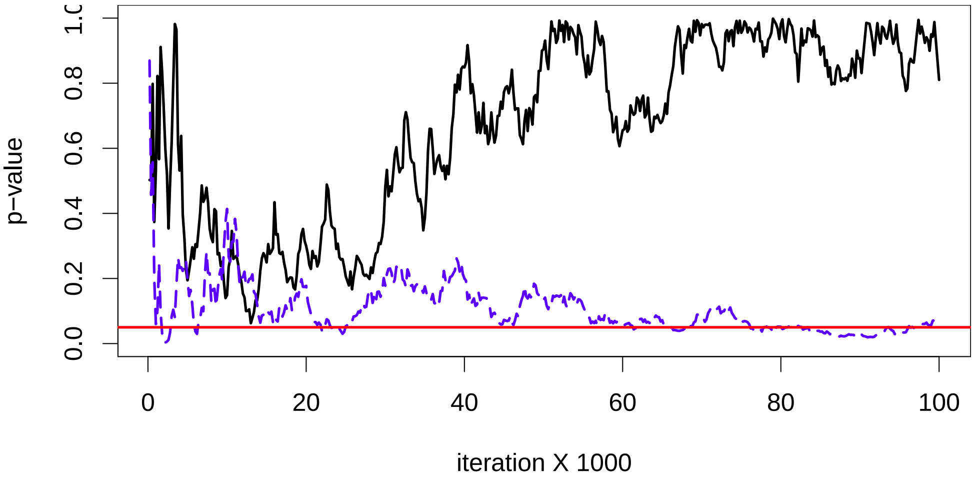

Figure 4: Plot of p-value of V_n versus the number of iterations n number for a binary Markov chain simulated by the algorithm of Wang (1981). All chains have the same stationary distribution Bernoulli($p = 0.43$).

When a multi-state chain is converted to a binary chain, the low mixing property may affect the observed proportion, $\hat{\pi}_i$, of the state under consideration in a small run. This happens simply because a multi-state chain with low mixing is more likely to become trapped in a low mass region and therefore requires a longer run to match the observed proportion with the posterior.

5 Clustering

For simplicity, we assign to each distinct visited grouping an integer label. This labelling is consistent in the sense that states visited by both samplers have identical integer labels. This time we target the top 10 most visited states for each sampler. For a more detailed analysis we calculated the true posterior order of each top visited grouping. In the ideal case both samplers hit the true ordering. Furthermore, in the ideal scenario the proportion of visiting each grouping $\hat{\pi}_i$ matches the true posterior probability π_i for all 10 top visited states. The Gibbs sampling and split-merge MH algorithms visit 697 and 268 distinct states, respectively (out of which 228 states are visited by both algorithms). This suggests that the Gibbs sampling algorithm explores the state space of data groupings more freely than the split-merge algorithm. Figure 5 (left panel) confirms that the Gibbs sampling has better convergence properties as π_i and $\hat{\pi}_i$ fall closer to the reference line $\pi_i = \hat{\pi}_i$. Figure 5 (right panel) shows that both samplers visit the true 5 highest posterior groupings, but afterwards the MH sampler is more likely to be trapped in low mass region and visit the groupings with smaller posterior probability. In the right panel of Figure 5 the ratio $f_i = \frac{\hat{\pi}_i}{\pi_i}$ is nearly constant for the Gibbs

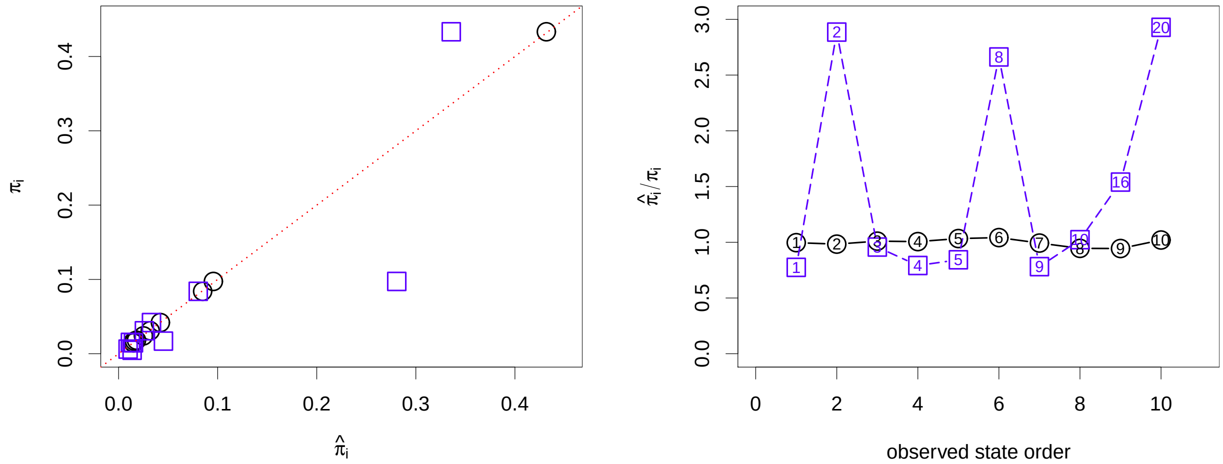

Figure 5: Left panel: plot of π_i versus $\hat{\pi}_i$ for the Gibbs sampler (circle) and the split-merge Metropolis-Hastings sampler (square). Right panel: the top observed state i versus $f_i = \hat{\pi}_i/\pi_i$: the Gibbs sampling (circles connected with solid lines) and split-merge MH algorithm (squares connected with dashed lines); large $\frac{\hat{\pi}_i}{\pi_i}$ ratios correspond to states that were visited more often than expected. The digit inside the circles and squares are the order according to the true posterior.

sampler. This is not the case for the MH sampler. Running the convergence criterion again targeting top 10 most visited states, yields a graph similar to Figure 2, accepting the convergence for the Gibbs sampler and rejecting the convergence for the MH sampler.

For the multi-class case we chose the top 10 states because this choice gave us estimated posterior probabilities > 0.001. The consistent estimation of variance covariance matrix $\mathbf{\Sigma}$ for a group of states using the regeneration method requires the states being observed more than once. An appropriate choice of the number of states can be found by computing the average time of regeneration. If the convergence is calculated each m iterations, the regeneration length should be at least $\frac{m}{2}$ giving on average two sequences to update the variance estimation. The average regeneration for both chains was around 70 iterations for top 10 states, and we updated p-values every 200 iterations.

Figure 2 suggests that the Gibbs sampler has converged, unlike MH sampler. Having computed the posterior for all possible groupings, we further studied convergence of both samplers graphically using Figure 5. Figure 5 also confirms the result of Figure 2. We have therefore used the result of the Gibbs sampler to cluster our data.

Figure 6 (right panel) displays a representative dendrogram based on the output of the Gibbs sampling algorithm; colours denote different MAP clusters. This dendrogram is reliable, since the chain is confirmed to have converged. This dendrogram was obtained by taking the most frequently visited allocation as the reference allocation, and constructing a dendrogram using observed agglomerations and divisions of this reference allocation. Figure 7 is the result of applying consensus clustering (Murua et al., 2008) to the output of the reversible Gibbs sampler. Again such a diagram can be drawn for the Gibbs sampler only as the convergence for the MH sampler is doubtful. In consensus clustering for each pair of observations, an indicator function is defined, taking value 1 if the observation belongs to the same cluster, and 0 otherwise. The values of these indicator functions are estimated by MCMC empirical averages, approximating the posterior probability that a pair of observations falls into the same cluster. Several threshold values are applied, giving different clustering possibilities associated to different uncertainties.

We used `R` (http://www.r-project.org/) for our computation; in particular, the contributed packages `bclust` (Partovi Nia and Davison, 2012) for clustering, `labeltodendro` (Partovi Nia and Stephens, 2010) and `ape` for visualization, and `partition` for generating all possible groupings of 14 mutants.

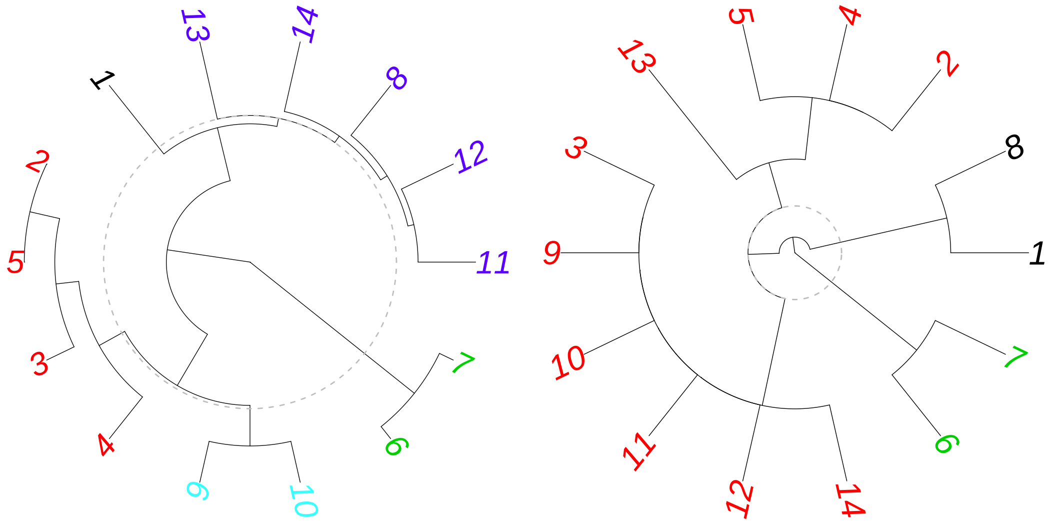

Figure 6: **Tree diagrams** *Left*: Dendogram resulting from agglomerative Bayesian clustering. *Right*: Dendrogram extracted from the reversible Gibbs sampler output. The dashed gray circle shows the MAP clustering cutting point. The colours highlights the MAP grouping found by cutting the tree with the dashed circle.

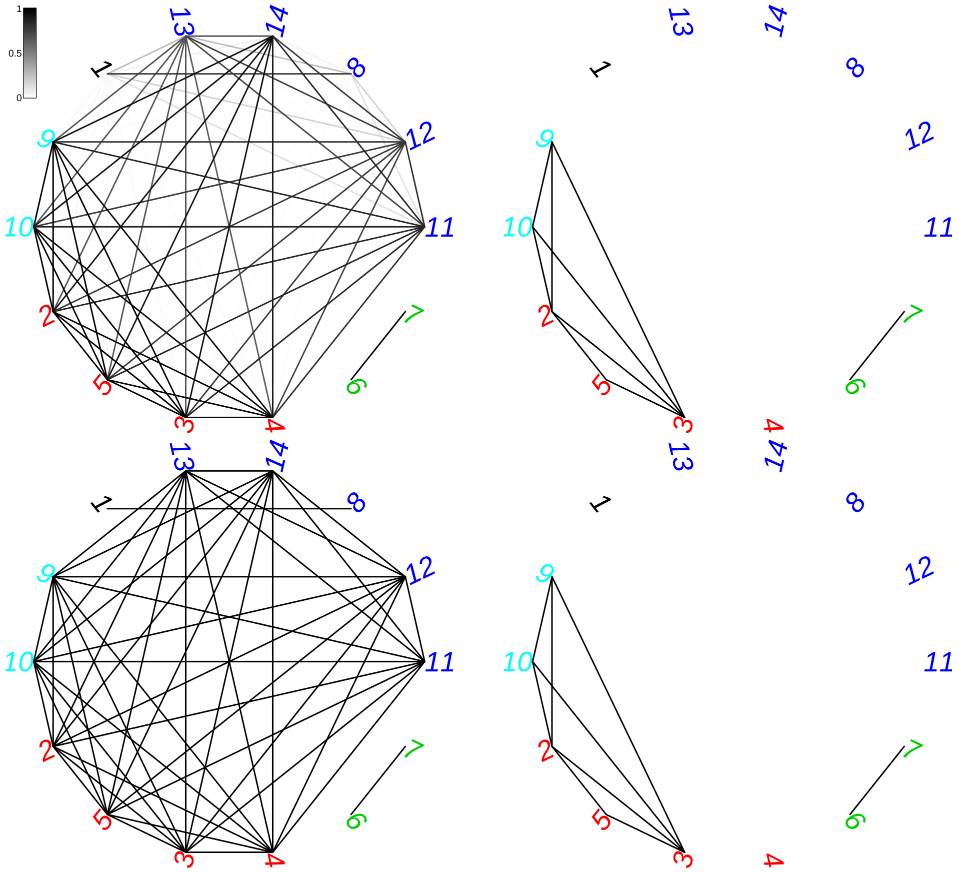

Figure 7: Consensus clustering diagrams. *Top left*: no threshold. *Top right*: threshold at 0.99. *Bottom left*: threshold at 0.9. *Bottom right*: threshold at 0.5. Labels are colored according to the optimal grouping found by the agglomerative algorithm; see Figure 6 (left panel).

6 Discussion

Bayesian clustering via MCMC methods is concerned with exploring the nominal state space of groupings. We proposed a convergence criterion for MCMC algorithms on nominal finite state spaces that is widely applicable to algorithms such as Gibbs sampling and Metropolis-Hastings. We implemented this criterion to MCMC sampling of groupings, under the assumption that the marginal posterior distribution of groupings is tractable. In particular, we defined a one-dimensional variance-like statistic and proposed an intuitive hypothesis test for lack of convergence to the stationarity distribution which is known up to a normalizing constant. Theorem 3.1 presents the asymptotic distribution of this statistic. We recommend monitoring qualitatively the p-value of the test that rejects the null hypothesis for large values of the variance statistic being computed over high mass regions or over the top visited states. Since more frequently visited states and high mass regions may differ, we also derive an approximate lower bound in Theorem 3.2 for the number of iterations required to visit a high mass state, under the stationarity assumption.

Estimation of the asymptotic variance-covariance matrix in the Markov chain CLT is a computationally intensive problem that has generated some interest in recent years. We use the method of regenerative simulation for this task, with computational and storage costs of order $O(k^2)$, where k is the number of states under consideration in an MCMC sampling algorithm. The performance of regenerative simulation can be poor when the state space explored is large. We can resolve this deficiency by merging low mass states.

In practice, we recommend computing the proposed test statistic over the top k most visited states and merging the remaining states as the state $k+1$. In Section 4 we compared the proposed convergence assessment criterion to that of Jones et al. (2006) on MCMC output from the Gibbs sampling and MH split-merge algorithms, and concluded that the Gibbs sampler converges much faster than the split-merge Metropolis-Hastings samples. Given that the Gibbs sampler explores a wider neighbourhood for a small number of clustering objects, this observation may be expected.

In Bayesian clustering via non-conjugate models (Jain and Neal, 2007; Kim et al., 2006; Tadesse et al., 2005) sampling from the posterior distribution is performed by the reversible jump algorithm (Richardson and Green, 1997). For these models, the marginal posterior distribution of partitions may not have a tractable form, and our convergence criterion would not apply. Our preliminary study indicates that it is possible to devise a similar convergence diagnostic tool utilizing the detailed balance condition. We, finally, note that in the derivation of the asymptotic distribution of the proposed criterion m, the number of visited states is considered fixed.

Appendix

Proof of Theorem 3.1

Since $S_{m,n}$ and S are asymptotically interchangeable, the proof is given in terms of the former state space. Irreducible and aperiodic Markov chains on finite state spaces are uniformly ergodic (Roberts and Rosenthal, 2004) hence the condition of geometric ergodicity of X is satisfied.

Split the chain X into $R = R(n)$ regeneration tours, where n is the length of X. Let $n_r = \tau_r - \tau_{r-1}$ denote the length of the rth tour; the average tour length is $\bar{n} = R^{-1}\sum_{r=1}^{R} n_r$. For $i \in S_{m,n}$, define

$$s_{r,i} = \sum_{k=\tau_{r-1}}^{\tau_r - 1} \mathbb{I}(X_k = i),$$

the number of visits to state i in the rth tour, $r = 1, \ldots, R$. The pairs $(n_r, s_{r,i})$, $r = 1, \ldots, R$ are independent and identically distributed, for fixed i. Similarly, for $i \neq j$, define

$$s_{r,ij} = \sum_{k=\tau_{r-1}}^{\tau_r - 1} \mathbb{I}(X_k \in \{i,j\}) = s_{r,i} + s_{r,j},$$

the number of visits to states i or j in the rth tour.

For $i \in S_{m,n}$, the CLT result in Equation (3) applies with $g_i(x) = \mathbb{I}(x = i)$ (where $\mathbb{E}_\Pi |g_i|^{2+\epsilon} = \Pi_i \ \forall \epsilon > 0$). As $R \to \infty$,

$$\begin{aligned} w_{i,n} &= \sqrt{R}\,(\hat{\pi}_i - \Pi_i) = \sqrt{R}\left\{\frac{1}{\tau_R - 1}\sum_{k=1}^{\tau_R - 1} g_i(X_k) - \mathbb{E}_\Pi g_i\right\} \\ &\xrightarrow{D} \text{Normal}_1(0, \sigma_{ii}), \end{aligned}$$

where the asymptotic variance has the following expression (Hobert et al. 2002)

$$\sigma_{ii} = \frac{\mathbb{E}_Q\left\{(s_{1,i} - n_1 \mathbb{E}_\Pi g_i)^2\right\}}{\left\{\mathbb{E}_Q(n_1)\right\}^2} = \frac{\mathbb{E}_Q\left\{(s_{1,i} - n_1 \Pi_i)^2\right\}}{\left\{\mathbb{E}_Q(n_1)\right\}^2}. \tag{8}$$

By the Cramér-Wold Device, $\mathbf{w}_n \xrightarrow{D} \text{Normal}_m(\mathbf{0}, \mathbf{\Sigma})$ as $R \to \infty$, where $\mathbf{0}$ is an m-dimensional column vector of zeros and $\mathbf{\Sigma} = \{\sigma_{ij}\}_{i,j=1}^m$ is an $m \times m$ variance-covariance matrix. The diagonal elements of $\mathbf{\Sigma}$ are given in Equation (8), and the off-diagonal elements are

$$\begin{aligned} \sigma_{ij} &= \lim_{R\to\infty} \text{cov}(w_{i,n}, w_{j,n}) \\ &= \frac{1}{2}\lim_{R\to\infty}\left[\text{var}(w_{i,n} + w_{j,n}) - \text{var}(w_{i,n}) - \text{var}(w_{j,n})\right] \\ &= \frac{1}{2}\Bigg[\lim_{R\to\infty} \text{var}\left\{\sqrt{R}\left(\frac{1}{\tau_R - 1}\sum_{k=1}^{\tau_R - 1} g_{i,j}(X_k) - \Pi_i - \Pi_j\right)\right\} \\ &\quad - \sigma_{ii} - \sigma_{jj}\Bigg], \end{aligned} \tag{9}$$

where the asymptotic variance is given by the Markov chain CLT with $g_{i,j}(x) = \mathbb{I}(x \in \{i, j\})$. In particular, as $R \to \infty$,

$$\sqrt{R}\left\{\frac{1}{\tau_R - 1}\sum_{k=1}^{\tau_R - 1} g_{i,j}(X_k) - \mathbb{E}_\Pi g_{i,j}\right\} \xrightarrow{D} \text{Normal}_1(0, \eta_{ij}),$$

where

$$\begin{aligned} \eta_{ij} &= \frac{\mathbb{E}_Q\left\{\left[s_{1,ij} - n_1 \mathbb{E}_\Pi g_{i,j}\right]^2\right\}}{\left\{\mathbb{E}_Q(n_1)\right\}^2} \\ &= \frac{\mathbb{E}_Q\left\{\left[s_{1,ij} - n_1(\Pi_i + \Pi_j)\right]^2\right\}}{\left\{\mathbb{E}_Q(n_1)\right\}^2}. \end{aligned}$$

So, $\sigma_{ij} = \frac{1}{2}[\eta_{ij} - \sigma_{ii} - \sigma_{jj}]$.

Since $\mathbf{w}_n \xrightarrow{D} \text{Normal}_m(\mathbf{0}, \mathbf{\Sigma})$ as $R \to \infty$, it follows that $\mathbf{C}\mathbf{w}_n \xrightarrow{D} \text{Normal}_m(\mathbf{0}, \mathbf{C}\mathbf{\Sigma}\mathbf{C}')$ as $R \to \infty$. By expressing the variance test statistic V_n as $V_n = (\mathbf{C}\mathbf{w}_n)'(\mathbf{C}\mathbf{w}_n)$, we conclude that $V_n \xrightarrow{D} \sum_{i=1}^m \lambda_i Z_i^2$ as $R \to \infty$, where $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of $\mathbf{C}\mathbf{\Sigma}\mathbf{C}'$ and $Z_1, \ldots, Z_m$ are independent standard normal variables (Chernoff and Lehmann, 1954, Lemma 1).

It remains to show how to consistently estimate the entries in $\mathbf{\Sigma}$. Following Hobert et al. (2002), consistent estimators (as $R \to \infty$) of σ_{ii} and σ_{ij} are given by

$$\begin{aligned} \hat{\sigma}_{ii} &= \frac{\sum_{r=1}^R (s_{r,i} - \hat{\pi}_i n_r)^2}{R\bar{n}^{-2}} \\ \hat{\sigma}_{ij} &= \frac{1}{2}\left\{\frac{\sum_{r=1}^R (s_{r,ij} - \hat{\pi}_{ij} n_r)^2}{R\bar{n}^{-2}} - \hat{\sigma}_{ii} - \hat{\sigma}_{jj}\right\} \\ &= \frac{\sum_{r=1}^R (s_{r,i} - \hat{\pi}_i n_r)(s_{r,j} - \hat{\pi}_j n_r)}{R\bar{n}^{-2}}, \end{aligned}$$

respectively, where $\hat{\pi}_{ij} = \hat{\pi}_i + \hat{\pi}_j$. In practice, $\hat{\mathbf{\Sigma}}$ (the variance-covariance matrix with estimated entries) might not be positive semidefinite, so we find the nearest positive semidefinite matrix to $\hat{\mathbf{\Sigma}}$ in the Frobenius norm to within a given tolerance level (**?**). Then, let $\hat{\lambda}_1, \ldots, \hat{\lambda}_m$ be the eigenvalues of this approximation to $\hat{\mathbf{\Sigma}}$.

The computational complexity of computing $\hat{\mathbf{\Sigma}}$ is of the order $O(kR + k^2)$. If successive computations of $\hat{\mathbf{\Sigma}}$ (at iterations t_1 and t_2, $t_1 < t_2$) are based on the same splitting of the chain (i.e., using the same state x'), resulting in R_1 and R_2 tours, respectively, then for all states i visited up to time t_1, $s_{r,i}^{(t_2)} = s_{r,i}^{(t_1)}$ for $r \leq R_1$. In other words, values $s_{r,i}$ at time t_2 can be updated from values at time t_1, thus reducing computation time.

Proof of Theorem 3.2

$$\mathrm{P}(E_i) = \prod_{t=1}^{n} \mathrm{P}(X_t \neq i \mid X_{t-1} \neq i),$$

but under equilibrium

$$\mathrm{P}(E_i) = \{\mathrm{P}(X_2 \neq i \mid X_1 \neq i)\}^n.$$

On the other hand

$$\begin{aligned} \mathrm{P}(X_2 \neq i \mid X_1 \neq i) &= 1 - \mathrm{P}(X_2 = i \mid X_1 \neq i) \\ &= 1 - \mathrm{P}(X_1 \neq i \mid X_2 = i)\frac{\mathrm{P}(X_2 = i)}{\mathrm{P}(X_1 \neq i)} \end{aligned}$$

which equals

$$1 - \frac{1}{1 - \Pi_i} \sum_{j=1}^{M} \mathrm{P}(X_1 = j \mid X_2 = i)\mathrm{P}(X_2 = i) - \mathrm{P}(X_1 = i \mid X_2 = i)\mathrm{P}(X_2 = i),$$

and under reversibility

$$\begin{aligned} &= 1 - \frac{1}{1 - \Pi_i} \sum_{j=1}^{M} \mathrm{P}(X_2 = i \mid X_1 = j)\mathrm{P}(X_1 = j) - \mathrm{P}(X_1 = i \mid X_2 = i)\mathrm{P}(X_2 = i) \\ &= 1 - \frac{\Pi_i}{1 - \Pi_i}(1 - P_{i,i}). \end{aligned}$$

References

Booth, J.G., Casella, G. and Hobert, J.P. (2008) Clustering using objective functions and stochastic search. Journal of the Royal Statistical Society Series B 70(1), 119–139.

Brooks, S.P., Giudici, P. and Philippe, A. (2003) Nonparametric convergence assessment for MCMC model selection. Journal of Computational and Graphical Statistics 12(1), 1–22.

Chan, K.S. and Geyer, C.J. (1994) Comment on ‘Markov chains for exploring posterior distributions’ by L. Tierney. Annals of Statistics 22(4), 1747–1758.

Chernoff, H. and Lehmann, E.L. (1954) The use of maximum likelihood estimates in χ^2 tests for goodness of fit. The Annals of Mathematical Statistics 25(3), 579–586.

Cowles, M. and Carlin, B. (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association 91(434), 883–904.

Everitt, B., Landau, S., Leese, M. and Stahl, D. (2011) Cluster Analysis. New York: Wiley.

Flegal, J.M. and Jones, G.L. (2010) Batch means and spectral variance estimators in Markov chain Monte Carlo. The Annals of Statistics 38(2), 1034–1070.

Fraley, C. and Raftery, A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97(458), 611–631.

Galin, L.J. (2004) On the Markov chain Central Limit Theorem. Probability Surveys 1, 299–320.

Geyer, C.J. (1992) Practical Markov chain Monte Carlo. Statistical Science 7(4), 473–483.

Green, P.J. (1995) Reversible Jump MCMC computation and Bayesian model determination. Biometrika 82(4), 711–732.

Hartigan, J.A. (1975) Clustering Algorithms. New York: Wiley.

Hartigan, J.A. (1990) Partition models. Communications in Statistics, Theory and Methods 19, 2745–2756.

Heard, N.A., Holmes, C.C. and Stephens, D.A. (2006) A quantitative study of gene regulation involved in the immune response of *Anopheline* mosquitoes: An application of Bayesian hierarchical clustering of curves. Journal of the American Statistical Association 101(473), 18–29.

Heller, K.A. and Ghahramani, Z. (2005) Bayesian hierarchical clustering. In Proceedings of the 22nd international conference on Machine Learning, ICML '05, pp. 297–304. New York, USA: ACM.

Hobert, J.P., Jones, G.L., Presnell, B. and Rosenthal, J.S. (2002) On the applicability of regenerative simulation in Markov chain Monte Carlo. Biometrika 89(4), 731–743.

Jain, S. and Neal, R. (2007) Splitting and merging components of a nonconjugate Dirichlet process mixture model. Bayesian Analysis 2, 445–472.

Jain, S. and Neal, R.M. (2004) A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. Journal of Computational and Graphical Statistics 13(1), 158–182.

Jones, G.L. (2004) On the Markov chain central limit theorem. Probability Surveys 1, 229–320.

Jones, G.L., Haran, M., Caffo, B.S. and Neath, R. (2006) Fixed-width output analysis for Markov chain Monte Carlo. Journal of the American Statistical Association 101(476), 1537–1547.

Kim, S., Tadesse, M.G. and Vannucci, M. (2006) Variable selection in clustering via Dirichlet process mixture models. Biometrika 93(4), 877–893.

Liu, J.S. (2001) *Monte Carlo strategies in scientific computing*. New York: Springer.

MacEachern, S.N. and Berliner, L.M. (1994) Subsampling the Gibbs sampler. The American Statistician 48(3), 188–190.

McCullagh, P. and Yang, J. (2006) Stochastic classification models. In Proceedings of International Congress of Mathematicians, volume 3, pp. 669–686. European Mathematical Society.

Messerli, G., Partovi Nia, V., Trevisan, M., Kolbe, A., Schauer, N., Geigenberger, P., Chen, J., Davison, A.C., Fernie, A. R. and Zeeman, S.C. (2007) Rapid classification of phenotypic mutants of Arabidopsis via metabolite fingerprinting. Plant Physiology 143, 1481–1492.

Meyn, S.P. and Tweedie, R.L. (1993) Markov Chains and Stochastic Stability. London: Springer-Verlag.

Murua, A., Stanberry, L. and Stuetzle, W. (2008) On Potts model clustering, kernel k-means and density estimation. Journal of Computational and Graphical Statistics 17(3), 629–658.

Mykland, P., Tierney, L. and Yu, B. (1995) Regeneration in Markov chain samplers. Journal of the American Statistical Association 90, 233–241.

Partovi Nia, V. and Davison, A.C. (2012) High-dimensional Bayesian clustering with variable selection: The R package bclust. Journal of Statistical Software 47(5), 1–22.

Partovi Nia, V. and Stephens, D.A. (2010) Dendrogram representation of stochastic clustering in R. Unpublished Manuscript.

Peskun, P.H. (1973) Optimum Monte-Carlo sampling using Markov chains. Biometrika 60, 607–612.

Rasmussen, C.E. (2000) The infinite gaussian mixture model. In Advances in Neural Information Processing Systems, eds S. Solla, T. Leen and K.-R. Müller, volume 12, pp. 554–560. MIT Press.

Richardson, S. and Green, P. (1997) On Bayesian analysis of mixtures with an unknown number of components. Journal of the Royal Statistical Society. Series B 59(4), 731–792.

Robert, C.P. and Casella, G. (2004) Monte Carlo Statistical Methods. Second edition. New York: Springer.

Roberts, G.O. and Rosenthal, J. (2004) General state space Markov chains and MCMC algorithms. Probability Surveys 1, 20–71.

Tadesse, M.G., Sha, N. and Vannucci, M. (2005) Bayesian variable selection in clustering high-dimensional data. Journal of the American Statistical Association 100(470), 602–617.

Wang, Y. (1981) On the limit of the markov binomial distribution. Journal of Applied Probability 18, 937–942.

Wood, A. T.A., Booth, J.G. and Butler, R.W. (1993) Saddlepoint approximations to the CDF of some statistics with nonnormal limit distributions. Journal of the American Statistical Association 88(422), 680–686.

Zellner, A. and Min, C. (1995) Gibbs sampler convergence criteria. Journal of the American Statistical Association 90(431), 921–927.