Stat Comput (2006) 16:355–362
DOI 10.1007/s11222-006-9391-y

Bayesian model learning based on a parallel MCMC strategy

Jukka Corander · Mats Gyllenberg · Timo Koski

Received: July 2005 / Accepted: June 2006

Abstract We introduce a novel Markov chain Monte Carlo algorithm for estimation of posterior probabilities over discrete model spaces. Our learning approach is applicable to families of models for which the marginal likelihood can be analytically calculated, either exactly or approximately, given any fixed structure. It is argued that for certain model neighborhood structures, the ordinary reversible Metropolis-Hastings algorithm does not yield an appropriate solution to the estimation problem. Therefore, we develop an alternative, non-reversible algorithm which can avoid the scaling effect of the neighborhood. To efficiently explore a model space, a finite number of interacting parallel stochastic processes is utilized. Our interaction scheme enables exploration of several local neighborhoods of a model space simultaneously, while it prevents the absorption of any particular process to a relatively inferior state. We illustrate the advantages of our method by an application to a classification model. In particular, we use an extensive bacterial database and compare our results with results obtained by different methods for the same data.

Keywords Bayesian analysis · Markov chain Monte Carlo · Model learning · Parallel search

J. Corander (✉) · M. Gyllenberg · T. Koski
Rolf Nevanlinna Institute, Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68, FIN-00014, Finland
e-mail: jukka.corander@helsinki.fi

T. Koski
Department of Mathematics, University of Linköping, S-58183 Linköping, Sweden

1 Introduction

A common problem in Bayesian modelling is the learning of structural layers of probability models, such as the number of components in mixture distributions, edge sets of graphical models, choice of predictor variables in regression, and so on. Such learning problems are often formulated by separating the structural layer of a model from the quantitative layer, which refers to the actual values of the model parameters. Markov chain Monte Carlo (MCMC) algorithms have gained a considerable popularity for Bayesian model learning. In particular, for structural learning where a suitable parametric dimension of a model is unknown *a priori*, the reversible jump MCMC algorithm introduced by Green (1995), has been extensively applied. A recent survey of Markov chain methodology suitable for variable-dimensional modelling is given by Sisson (2005).

Despite of the theoretically solid basis of the MCMC based approach, complex modelling applications may pose challenges which lead to practically unacceptable precision level in the numerical solutions derived by the algorithms. For instance, estimates of posterior probabilities for models may fluctuate considerably, or the most relevant model structures can even remain undetected. Several strategies have been developed to enhance the performance of the MCMC for structural learning, e.g. single-chain Metropolis-coupled MCMC (Geyer and Thompson, 1995), graphical monitoring of MCMC model composition (Giudici and Castelo, 2003), proposal improvement (Brooks et al., 2003), parallel coupled MCMC (Altekar et al., 2004).

Generally, when analytical integration can be incorporated in MCMC, it often increases the numerical tractability of the learning problem. This feature can be exploited for a wide variety of learning tasks, such as edge determination for graphical models (Giudici and Castelo, 2003), discovery of gene

regulatory binding motifs (Jensen et al., 2004), and unsupervised classification of molecular marker data (Corander et al., 2004). For instance, for complex data sets, the classification model of Corander et al. (2004) can contain several millions of quantitative parameters. Learning of such models using the reversible jump MCMC type algorithms, where the quantitative parameters are sequentially updated, is numerically intractable.

We consider here a computational strategy for Bayesian structural learning of models, where the marginal likelihood with respect to a prior measure for the quantitative parameters can be calculated analytically, for an arbitrary value of the structural layer. It is further assumed that the structural layer has a finite set of distinct values, referred to as the model space. Applicability of our strategy is further widened if also analytical approximations to the marginal likelihood are exploited, in a similar fashion as in the methods discussed in Sisson (2005).

It is generally recognized that MCMC methods are a preferred alternative to direct optimization algorithms, due to the tendency of the latter to be trapped at local maxima. However, for many structural learning problems, the target distribution is highly multi-modal, and even the MCMC algorithms may experience difficulties in identifying the most relevant areas of the model space. Also, the neighborhood structure of the model space may yield an unexpected obstacle for convergence of the MCMC method in practice. To avoid such problems, Corander et al. (2006) developed a parallel approach specific to unsupervised classification on a heuristic basis. Here, we establish consistency of an analogous approach to the general model learning problem for discrete spaces.

The structure of the paper is as follows. In the next section we formulate the Bayesian structural model learning problem. In Section 3 we establish the consistency of a novel non-reversible MCMC estimation algorithm. Numerical examples concerned with unsupervised classification are considered in Section 4 to illustrate the preferable features of our method. Some concluding remarks and possibilities for future research are given in the final section.

2 Bayesian predictive modelling

Let $\mathbf{x}$ denote a generic data set, for which we consider a finite set $\{p(\cdot \mid \theta_\delta, \delta \in \Delta\}$ of probability models. Here δ is taken as the structural layer of a probability model, while $\theta_\delta \in \Theta_\delta$ represents the quantitative layer, taking values on a space dependent on the specific structure. The Bayesian approach (see, e.g. Schervish, 1995) to learning plausibilities of the elements of Δ on the basis of the observed $\mathbf{x}$, specifies formally the predictive, or marginal data distribution as the mixture

$$p(\mathbf{x}) = \sum_{\delta\in\Delta} p(\delta)p(\mathbf{x} \mid \delta) = \sum_{\delta\in\Delta} p(\delta) \int_{\Theta_\delta} p(\mathbf{x} \mid \theta_\delta) d\mu(\theta_\delta), \quad (1)$$

where $\mu(\theta_\delta)$ is a prior probability measure on the Borel space (Θ_δ, τ) where τ is a sigma-algebra, and $p(\mathbf{x} \mid \theta_\delta)$ is the conditional data distribution, or the likelihood, given θ_δ. The value of $p(\delta)$ can be interpreted as the prior predictive weight, or the prior probability of the structural layer δ, such that $\sum_{\delta\in\Delta} p(\delta) = 1$.

In structural model learning one is typically interested in the posterior probabilities of $\delta \in \Delta$, as measures of the model plausibility to explain the information carried by the data. These are defined according to

$$p(\delta \mid \mathbf{x}) = \frac{p(\delta)p(\mathbf{x} \mid \delta)}{\sum_{\delta\in\Delta} p(\delta)p(\mathbf{x} \mid \delta)}. \quad (2)$$

The standard MCMC approach (see Sisson, 2005 or Robert and Casella, 2005) to estimating (2), or to identification of the model structure corresponding to the maximal ability to explain the data (given by $\arg\max_{\delta\in\Delta} p(\delta \mid \mathbf{x})$), is to construct a Markov chain in Δ, with the time homogeneous distribution corresponding to (2). From a realization $\{\delta_t, t = 0, 1, \ldots\}$ of such a chain, the posterior probabilities can be consistently estimated as

$$p_n(\delta \mid \mathbf{x}) = n^{-1} \sum_{t=1}^{n} I(\delta_t = \delta), \quad (3)$$

such that $p_n(\delta \mid \mathbf{x}) \to p(\delta \mid \mathbf{x})$, as $n \to \infty$. Similarly, the value of δ maximimizing (2), can be identified from the realization $\{\delta_t, t = 0, 1, \ldots\}$. Although the estimator (3) can be straightforwardly applied to even multiple independent realizations of a Markov chain, it is subject to certain numerical deficiencies from the practical point of view. Namely, as it is based on the relative frequencies of visits to specific models structures, estimated posterior probabilities may fluctuate largely due to a large size of the model space Δ and problems with the mixing of the chain. Also, when several independent realizations are used to calculate the estimates, they cannot be consistently weighted with respect to their representativeness of the posterior distribution. Therefore, any chains that have explored only relatively inferior areas of the model space, will be given too much weight in (3) in practice, although a consistent estimate is obtained when n tends to infinity.

One of the most commonly used MCMC algorithms for structural learning is the Metropolis-Hastings (MH) algorithm (see Tierney, 1994; Carlin and Chib, 1995; Chib and

Greenberg, 1995), which is defined through the following transition kernel, governing the probability of transition from the current state δ_t to a proposal state δ^* as:

$$\min\left(1, \frac{p(\delta^*)p(\mathbf{x}\mid\delta^*)}{p(\delta_t)p(\mathbf{x}\mid\delta_t)}\frac{q(\delta_t\mid\delta^*)}{q(\delta^*\mid\delta_t)}\right), \tag{4}$$

where $q(\delta^*\mid\delta_t)$ is the probability of choosing state δ^* as the candidate for the next state at δ_t, and $q(\delta_t\mid\delta^*)$ is the probability of restoration of the current state. When the proposal mechanism is deliberately chosen, the algorithm can be used to generate an aperiodic, irreducible, and reversible Markov chain, whose time homogeneous distribution equals the sought posterior distribution (2).

Given its generality, the MH algorithm has been applied to an extremely wide range of model learning situations. However, the success of the algorithm for a particular application is largely dependent of the possibilities to design effective proposal distributions. For instance, the neighborhood structure of the proposal space may create barriers for the convergence of the generated Markov chain in practice. The MH algorithm introduced in Corander et al. (2004) for Bayesian unsupervised classification, has the property that for certain move types (merge and split), typically either

$$\frac{q(\delta_t\mid\delta^*)}{q(\delta^*\mid\delta_t)}\to 0, \quad \text{or} \quad \frac{q(\delta_t\mid\delta^*)}{q(\delta^*\mid\delta_t)}\to 1, \tag{5}$$

when the amount of observations in the data $\mathbf{x}$ increases. Consequently, the transitions in the chain realization may become mainly determined by the proposal ratio, while the models' abilities to predict the data structure are almost ignored. In general, for any proposal distribution where the neighborhood structure of the space Δ implies strongly asymmetric proposal ratios, the ordinary MH algorithm may fail to produce a reasonable approximation to the posterior. Such transition mechanisms can easily arise for model dimension changing MCMC samplers, where the number of sample paths leading away from a state, is to a large extent distinct from the number of paths leading to the same state.

The ordinary MH algorithm may fail to produce a reliable approximation to the posterior also due to large size of the model space Δ. Then, a single chain can easily remain in a subspace of Δ representing relatively inferior predictive ability of models. Such problems were early recognized for various applications, and several modifications have been introduced in the literature, see Sisson (2005). From the theoretical point of view, the challenge of utilizing simultaneous multiple stochastic search processes, is to allow them interact while preserving the consistency properties.

To resolve the problem with the MH based classification algorithm of Corander et al. (2004, 2006) introduced heuristically a non-reversible MH classification algorithm. The relatively superior properties of the approach were demonstrated for extensive molecular data sets. In the next section we study formally the properties of an analogous algorithm for the general Bayesian model learning problem, and establish its consistency properties.

3 A non-reversible learning algorithm

Let $q(\cdot\mid\delta)$ denote a generic fixed distribution that assigns probabilities on Δ, conditional on δ. Although in a typical implementation of MCMC there are several types of proposal distributions, these are all assumed here to be fixed, and it is sufficient to consider them jointly by using the generic notation. The distribution $q(\cdot\mid\delta)$ is assumed to be such that the non-reversible Markov chain with state space Δ, governed by the transition kernel

$$\min\left(1, \frac{p(\delta^*)p(\mathbf{x}\mid\delta^*)}{p(\delta_t)p(\mathbf{x}\mid\delta_t)}\right), \tag{6}$$

is aperiodic and irreducible, where δ^* refers to a value generated from $q(\cdot\mid\delta_t)$. Since the state space Δ of the chain is finite, it follows (e.g., Häggström, 2002) that (6) defines also a positive recurrent Markov chain. The stationary distribution of such a chain is in general unknown, and not equal to the posterior (2). However, it is still possible to construct a consistent estimate of (2) from a realization $\{\delta_t, t = 0, 1, \ldots\}$. Interestingly, non-reversible Markov chain samplers have been earlier shown to have advantageous convergence properties (Diaconis et al., 2000) for certain types of discrete distributions.

For an arbitrary n, let $\Delta_n \subseteq \Delta$ denote the subspace of Δ that has been visited by the process $\{\delta_t, t = 0, 1, \ldots, n\}$, i.e. $\delta \in \Delta_n$ if $\delta = \delta_t$, for any $t = 1, 2, \ldots, n$. The following lemma establishes posterior estimates that can be constructed from $\{\delta_t, t = 0, 1, \ldots, n\}$.

Lemma 1. *Let $\{\delta_t, t = 0, 1, \ldots, n\}$ be a Markov chain defined according to* (6)*. Then, the estimate of posterior probability mass for all $\delta \in \Delta$,*

$$p_n(\delta\mid\mathbf{x}) = \frac{p(\mathbf{x}\mid\delta)}{\sum_{\delta\in\Delta_n} p(\mathbf{x}\mid\delta)},$$

and the estimate of posterior mode,

$$\arg\max_{\delta\in\Delta_n} p(\delta\mid\mathbf{x})$$

are consistent, i.e.

$$p_n(\delta\mid\mathbf{x}) \to p(\delta\mid\mathbf{x}),$$

$$\arg\max_{\delta\in\Delta_n} p(\delta\mid\mathbf{x}) \to \arg\max_{\delta\in\Delta} p(\delta\mid\mathbf{x}),$$

a.s. as $n\to\infty$.

Proof: The lemma follows from the standard properties of positive recurrent Markov chains, see Isaacson and Madsen (1976), or Häggström (2002), as each state of the finite state space will be visited with certainty as n tends to infinity. Notice that the general theory ensures the existence of the stationary distribution under the stated conditions. □

In the non-reversible chain the proposal distribution $q(\cdot \mid \delta)$ will have a key role, as for the reversible chains. However, an advantage of using the non-reversible solution is that the potential asymmetry of the proposal ratio $q(\delta_t \mid \delta^*)/q(\delta^* \mid \delta_t)$ does not affect the behavior of the algorithm, since only the predictive ability of models $p(\mathbf{x} \mid \delta)$ enters the transition kernel. A further advantage is that the explicit proposal probabilities need not be calculated in the practical implementation of the algorithm, e.g., reduces the computation time considerably for certain applications. For the non-reversible case it is possible to utilize any fixed proposal distribution, such that the conditions of aperiodicity and irreducibility are satisfied. Then, Lemma 1 applies, and a consistent estimate of the posterior may be be constructed from a realization of any number of chains having the time homogeneous distribution according to (6).

From the practical perspective, the aperiodicity and irreducibility conditions are analogously required even in the reversible case, and therefore, the same basic rules apply to the design of the proposal distribution. For reversible chains it is necessary to design proposal distributions for which the actual proposal probabilities of arbitrary states can be analytically calculated in the implementation, since they appear in the transition probability matrix. On the contrary, the non-reversible chain can utilize any fixed proposal distribution for which the proposal probabilities can be theoretically shown to satisfy the given conditions, even if the actual probabilities of arbitrary state transitions cannot be calculated analytically.

The non-reversible MH algorithm provides considerable flexibility to the proposal design, and e.g. solves the problem (5) related to the neighborhood structure of Δ. Nevertheless, as in the standard reversible MCMC case, a single chain solution may be insufficient for complex applications to provide a reliable approach to the identification of representative areas of Δ. Therefore, we define below a more efficient multiple chain solution for which the consistency of the estimates $p_n(\delta \mid \mathbf{x})$ and $\arg\max_{\delta \in \Delta_n} p(\delta \mid \mathbf{x})$ can still be validated.

Definition 1. Parallel interacting processes. Let $\{\delta_{tj}, t = 0, 1, \ldots; j = 1, \ldots, m\}$ and $\{Z_t, t = 0, 1, \ldots\}$ be $m + 1$ stochastic processes defined as follows:

(1) Define a sequence of strictly decreasing probabilities $\{\alpha_t, t = 1, 2, \ldots\}$, such that $\alpha_t > \alpha_{t+1}$, and $\alpha_t \to 0$ as $t \to \infty$.

(2) Define the stochastic process $\{Z_t, t = 0, 1, \ldots\}$ as $Z_0 = 0$, and $P(Z_t = 1) = \alpha_t$, $P(Z_t = 0) = 1 - \alpha_t$, independently for $t = 1, 2, \ldots$.

(3) Let δ_{0j}, $j = 1, \ldots, m$, be arbitrary initial states of $\{\delta_{tj}, t = 0, 1, \ldots; j = 1, \ldots, m\}$. Given a realization $\{Z_t, t = 0, 1, \ldots\}$, the transition mechanism of the processes $\{\delta_{tj}, t > 0; j = 1, \ldots, m\}$ depends on values of Z_t according to the following.

(4) For each t, such that $Z_t = 0$, transition from δ_{tj} to the next state $\delta_{(t+1)j}$ is determined according to the probability (6), for $j = 1, \ldots, m$.

(5) For each t, such that $Z_t = 1$, transition from δ_{tj} to the next state $\delta_{(t+1)j}$ is determined according to the following distribution over the space $Q_t^{(\Delta)} = \{\delta_{tj}, j = 1, \ldots, m\}$ of candidate states:

$$P_t\left(\delta_{(t+1)j} = \delta_{tj}\right) = \frac{p(\delta_{tj})p(\mathbf{x} \mid \delta_{tj})}{\sum_{j=1}^{m} p(\delta_{tj})p(\mathbf{x} \mid \delta_{tj})}, \tag{7}$$

independently for $j = 1, \ldots, m$.

The m processes$\{\delta_{tj}, t = 0, 1, \ldots; j = 1, \ldots, m\}$ defined above are not time homogeneous Markov chains. However, as $t \to \infty$, their transition probabilities converge to those of the time homogeneous Markov chain defined in (6). The interaction strategy exploited here, is reminiscent of those Bayesian particle filtering methods (e.g., Doucet et al., 2000), where particles are allowed to cluster according to a stochastic process.

The parallel interacting processes are defined to yield an efficient, yet consistent scheme for exchange of information between them. Even if consistency of posterior estimates can be shown for any strictly decreasing sequence of probabilities $\{\alpha_t, t = 1, 2, \ldots\}$, for practical implementation it is necessary to be more specific about them. In a classification application, Corander et al. (2006) utilized a logarithmic rate of decrease with

$$\alpha_t = 0, t = 1; \alpha_t = \frac{1}{q \log t}, \quad t = 2, 3, \ldots$$

where $q \geq 1$ can be chosen suitably, for instance $q \in [5, 10]$. As the probabilities α_t govern the possibility for the processes to interact, a reasonable choice of q allows chains to interact feasibly often and prevents too rapid isolation as t increases. The logarithmic rate of decrease is similar to the cooling schedule providing consistency and optimality properties for simulated annealing, see e.g. Gidas (1985). The multiplicative constant can be given a value from quite a wide range, though one needs to acknowledge that it is computationally inefficient to let the chains mix all the time. The general idea is to let the chains to explore independently of each other several directions in the model space for a reasonable amount of

time, and then investigate which of them have found plausible areas. The exact lengths of the intervals between interaction times are not relevant, but if they are too long, then a lot of computational operations may have been unnecessarily wasted on chains staying at clearly inferior states.

The role of the proposal distribution used in the last step of Definition 1 is to let the conditional posterior probabilities (7) to guide the interaction events. Given the above definition, the processes have a tendency to coalesce towards the states which are associated with higher marginal likelihoods. Also, when multiple model structures with roughly equal marginal likelihoods are present, the probabilities (7) lead to a more dispersed proposal distribution.

The following theorem establishes the fact that we can construct consistent estimates of the posterior mode and the individual posterior probabilities from a realization of the m interacting processes.

Theorem 1. *The estimates given in Lemma 1 remain consistent for the parallel interacting processes specified in Definition 1.*

The proof will be given in Appendix. Firstly, we investigate a time homogeneous Markov chain with the state space Δ, determined by the transition probability matrix P according to (6). Secondly, we establish the behavior of the interacting processes between the interaction times, as $t \to \infty$.

4 Numerical illustration

In this section we apply our parallel learning method to a Bayesian classification model, analyzing a large and well studied bacterial database and compare our results with results obtained by other methods. From a statistical perspective, this application corresponds to the fitting of a large scale multinomial mixture model, where both the number of components and the component specific probabilities for the observable attributes are unknown parameters. On the other hand, from the biological perspective the classes of the mixture model represent underlying biologically relevant groups in the bacterial population.

It widely recognized that mixture models with an unknown number of components represent generally a tremendous challenge for statistical inference. Also, the particular suitability of the Bayesian paradigm has been rightfully acknowledged in this context.

The example material consists of 5313 strains of bacteria belonging to the family *Enterobacteriaceae*. Each strain is characterized by $d = 47$ binary characters describing the occurrence (1) or nonoccurrence (0) of certain biochemical reactions (Farmer et al., 1985). The strains were classified by Farmer et al. (1985) into 104 nomen species or corresponding biogroups representing 27 genera. The by far most frequent species was *Escherichia coli* comprising 1708 strains, that is, almost one third of the material.

We have implemented the non-reversible MH algorithm into the BAPS software introduced by Corander et al. (2004), available at `http://www.rni.helsinki.fi/jic/bapspage.html`. Corander et al. (2006) utilized in their classification approach a stochastic partition $S = (s_1, \ldots, s_k)\,(1 \leq k \leq n)$ of the observed n items in a data set. The model space Δ is in this case the space of partitions of a finite integer n which correspond to all possible classification solutions for the data. Under a uniform prior distribution on Δ, and certain forms of generalized exchangeability, the posterior of the partitions equals

$$p(S \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid S)}{\sum_{S \in \Delta} p(\mathbf{x} \mid S)}, \tag{8}$$

where the marginal likelihood is determined as

$$p(\mathbf{x} \mid S) = \prod_{c=1}^{k} \prod_{j=1}^{d} \frac{\Gamma\left(\sum_{l=0}^{1} \lambda_{cjl}\right)}{\Gamma\left(\sum_{l=0}^{1} \lambda_{cjl} + n_{cjl}\right)} \prod_{l=0}^{1} \frac{\Gamma(\lambda_{cjl} + n_{cjl})}{\Gamma(\lambda_{cjl})}, \tag{9}$$

and $\Gamma(\cdot)$ is the gamma function. The hyperparameters of (9) can for this application be given the well-known reference form

$$\lambda_{cjl} = 1/2, l = 1, \ldots, r_j; j = 1, \ldots, d; c = 1, \ldots, k, \tag{10}$$

corresponding to the Jeffrey's prior.

To search for the posterior optimal classification Corander et al. (2006) considered MH transition kernel defined as

$$\min\left(1, \frac{p(\mathbf{x} \mid S^*)}{p(\mathbf{x} \mid S)}\right), \tag{11}$$

where S is the current classification and S^* a proposal classification in a non-reversible Markov chain with the state space Δ. The fixed proposal mechanism uses the following four different possibilities:

- With probability 1/2, merge two randomly chosen classes s_c, s_{c*}.
- With probability 1/2, split a randomly chosen class s_c into two new classes, with cardinalities uniformly distributed between 1 and $|s_c| - 1$ and with elements randomly chosen from s_c.
- Move an arbitrary individual from a randomly chosen class $s_c, |s_c| > 1$, into another randomly chosen class s_{c^*}.
- Choose one individual randomly from each of two randomly chosen classes s_c and s_{c^*}, and exchange them between the classes.

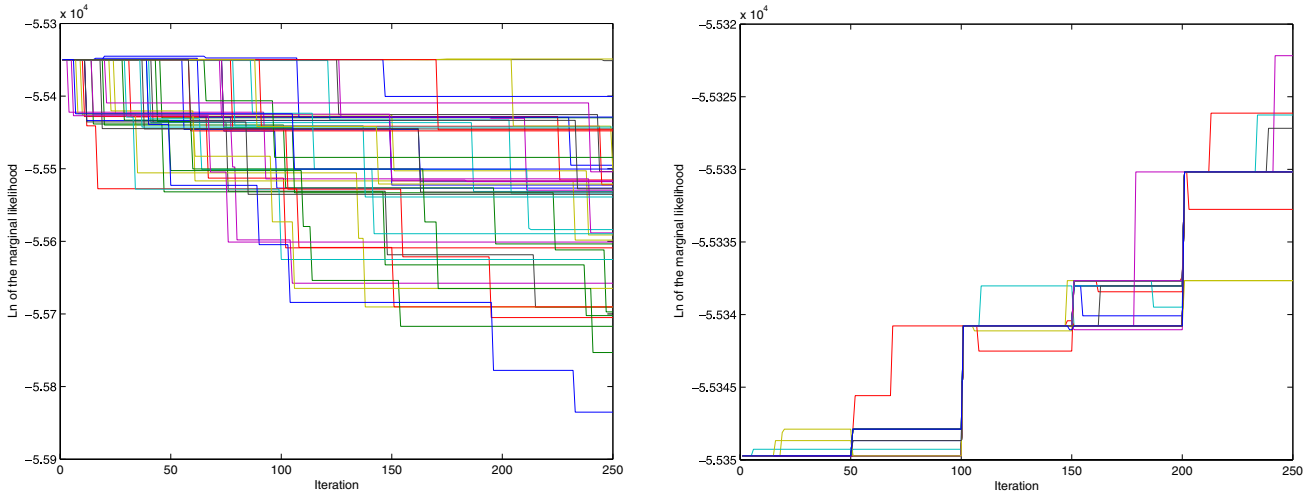

Fig. 1 Predictive abilities ($\log_e p(\mathbf{x} \mid S)$, vertical axis) of the Bayesian classification model for the *Enterobacteriaceae* data, over 250 iterations of 50 processes all started at the same initial configuration, produced with the reversible (left) and non-reversible (right) MH algorithms

The crucial conditions for the Lemma 1 and Theorem 1, are the irreducibility and aperiodicity of the Markov chain defined in (11). The validity of these conditions for the above type proposal follows from the lemma below.

Lemma 2. *Let $q(\cdot \mid \delta)$ be a generic proposal distribution on Δ, which has non-zero entropy on Δ. Assume $p(\delta) > 0$, for all $\delta \in \Delta$. Then the finite Markov chain defined by (6) is irreducible and aperiodic.*

Proof: We prove first irreducibility. Assume that the posterior is non-negative, $p(\delta \mid \mathbf{x}) > 0$, $\delta \in \Delta$. Hence, there is a positive probability to propose a sequence of state transitions leading from any $\delta \in \Delta$ to any other state $\delta^* \in \Delta$. Second, we show the aperiodicity. If an irreducible Markov chain is periodic, then all diagonal elements in its transition matrix are zero. Thus, if the Markov chain defined by the current algorithm is periodic, there exists no $\delta \in \Delta$, such that the acceptance probability is smaller than unity for some proposed candidate state. But this will happen if and only if the posterior ratio (6) is always larger than one on Δ, which is clearly impossible. Therefore, there will always be at least one non-zero element on the diagonal of the transition matrix, and the finite Markov chain cannot be periodic. □

For the elementary facts on Markov chains used in the above lemma we refer to Isaacson and Madsen (1976).

One of the main motivations behind the non-reversible classification algorithm introduced by Corander et al. (2006) was the inappropriate behavior of the reversible algorithm of Corander et al. (2004) observed for large data sets. In Fig. 1, we illustrate the advantages of a non-reversible approach, by a comparison of the behavior of the reversible (left) and non-reversible (right) MH algorithms for the above Bayesian classification model when applied to the *Enterobacteriaceae* data. Both panels of the figure show 250 iterations of 50 processes all started at the same initial configuration of the classification model, the vertical axis corresponding to the logarithm of $p(\mathbf{x} \mid S)$. It can be seen that most of the chains produced with the reversible MH method, tend towards worse solutions, whereas the non-reversible processes behave in the opposite way. This situation provides an example of the impact a scaling effect of the neighborhood of a particular model $\delta \in \Delta$ has to the convergence of a reversible Markov chain.

The interaction strategy we have utilized, enables exploration of several local neighborhoods of the model space simultaneously, while it does not allow any particular search process to be absorbed by a relatively inferior state (as compared to the other $m - 1$ processes) in the long run. Conversely, when several modes with roughly equal local masses exist in the posterior distrbution, the interaction scheme tends to allow these to be searched, without collapsing the search into a single neighborhood. In addition, the classification problem studied in our example illustrates the advantage of considering a non-reversible MH algorithm, when the reversible algorithm leads to practically unacceptable solution in model learning.

In a series of papers Gyllenberg et al. (1997, 1998, 1999a,b,c) and Gyllenberg et al. (1999) have studied Farmer's *Enterobacteriaceae* and classified it by the method of minimizing stochastic complexity (SC), Bayesian predictive identification and cumulative classification. The classification with the least stochastic complexity (SC = 21.2920) contained 63 classes and is described in detail in Gyllenberg et al. (1999c).

In contrast to the SC-minimizing classification, the MH-classification obtained by the method described in the present paper contained 129 classes and had an SC-value of 21.9968. However, the two classifications share a common

structure. Generally speaking, the larger classes in the SC-minimizing classification were split into two or more classes in the MH-classification. The MH-classification contained 61 pure classes (classes containing only one nomen species), compared with only 15 for the SC-minimizing classification. In both classifications one third of these were pure *E. coli* classes (20 and 5, respectively). Of the pure MH-classes 14 were perfect (containing all the strains of the nomen species) and 7 almost perfect (lacking a few strains). The corresponding figures for the SC-minimizing classification were 5 and 1. In addition, the SC-minimizing classification had 10 classes containing almost all the strains in a nomen species plus a few odd strains. Such classes did not appear in the MH-classification.

To summarize, the classification obtained by the MH-algorithm confirms the established taxonomy of *Enterobacteriaceae*, even to a higher degree than the SC-minimizing classification. Even the deviations from Farmer's et al. (1985) taxonomy make sense. *E. coli* is the most abundant of the *Enterobacteriaceae* species and it is very heterogeneous. It is therefore microbiologically natural to subdivide it into several classes. On the other hand, some of the MH-classes, as well as SC-classes, collected several species from the same genus. There is nothing remarkable with that. The notions of bacterial *species* and *genera* are man made and an objective definition of these concepts is still lacking. There is therefore no reason to assume that a mathematical algorithm should produce classes corresponding to species.

5 Discussion

The general non-reversible algorithm we have introduced here has some important advantages over the standard reversible MH approach to exploring posterior probabilities of models under the current setting. Firstly, the lack of a proposal probability ratio in the transition kernel leads to more efficient computation and avoids eventual paradoxical behavior induced by the reversibility condition when proposal ratios are strongly asymmetric. Secondly, the posterior estimates based on Lemma 1 are considerably more stable than the corresponding estimates based on the relative frequency of visits to a model. Thirdly, the non-reversible algorithm enables a parallel implementation where chains may exchange information about the posterior distribution in a consistent manner. Finally, the lack of a proposal probability ratio in the transition kernel provides a considerable freedom in the design of proposal mechanisms. This opens also the possibility of using intelligent search operators instead of the random type proposals exploited in our classification example. For instance, the proposal probability of a particular model structure may be defined to be proportional to its suitability according some effective heuristic exploratory tools.

Some earlier MCMC methods have exploited analytical approximations to the marginal likelihood $p(\mathbf{x} \mid \delta)$ within the posterior sampling scheme (see Sisson, 2005). Applicability of our strategy is clearly further widened if similar approximations are used under the non-reversible MH algorithm. One of the crucial challenges in modelling based on MCMC computation is the assessment of the convergence towards the target distribution. For large scale applications it would be extremely helpful if the algorithms had some intrinsic intelligence, e.g., in the sense that they could autonomously decide whether a particular region of a model space is worth further exploration. Similarly, in a parallel system the available computational resources could be autonomously re-allocated to concentrate more on learning some specific parts of a model. A highly valuable feature of such a self-monitoring system would be the possibility to assess whether continued computation is expected to provide further gains in the data prediction with a reasonable probability, or should the process be terminated. Whereas our non-reversible approach provides quite accessible means for the first two objectives, the final aspect is more left open for further research.

Appendix

Proof of Theorem 1: Let $\{Y_t\}$ be a Markov chain with arbitrary initial distribution and P (6) as transition matrix. Then, (Defs. II.1.6,II.1.8, and Lemma III.2.1, in Isaacson and Madsen, 1976) the probability of ever visiting any state j from any state i equals unity. Further, let in the sequel $\Delta_t \subseteq \Delta$ denote generally the subpace of Δ that has been visited by time t. Define now the first time when $\{Y_t\}$ has visited every state,

$$T_\delta = \min\{t \mid \Delta_t = \Delta, Y_0 = \delta\},$$

and the probability

$$g_\delta^{(n)} = P(T_\delta = n).$$

Clearly, we have $\sum_{n=1}^{\infty} g_\delta^{(n)} = 1$, for all $\delta \in \Delta$. Then, for any $\varepsilon > 0$, and any $\delta \in \Delta$ exists an $n_0^\delta(\varepsilon)$ such that

$$\sum_{n=1}^{n_0^\delta(\varepsilon)} g_\delta^{(n)} > 1 - \varepsilon,$$

that is $P(T_\delta = n_0^\delta(\varepsilon)) > 1 - \varepsilon$. We set $n_0(\varepsilon) = \max_{\delta \in \Delta}(n_0^\delta(\varepsilon))$.

Let $\{X_t\}$ be any of the m parallel interacting processes specified in Definition 1. We examine here the behavior of this single process, as it is sufficient to establish the consistency properties. Let $i = 1, 2, \ldots$ and define $Z_{t_i} = 1$,

$t_{i+1} = \min\{t > t_i \mid Z_{t_{i+1}} = 1\}$. Let $\tau_i = t_{i+1} - t_i - 1$, i.e. τ_i is the number of transitions of $\{X_t\}$ that are distributed according P in $t = t_i, t_i + 1, \ldots, t_{i+1}$. We may now establish the following behavior of τ_i:

$$\begin{aligned}
\mathbb{P}(\tau_i \geq n_0(\varepsilon)) &= 1 - \mathbb{P}(\tau_i < n_0(\varepsilon)) \\
&= 1 - \sum_{t=0}^{n_0(\varepsilon)-1} \mathbb{P}(\tau_i = t) \\
&= 1 - \sum_{t=0}^{n_0(\varepsilon)-1} \left(\alpha_{t_i+t+1} \prod_{s=t_i+1}^{t_i+t} (1 - \alpha_s) \right) \\
&\geq 1 - \sum_{t=0}^{n_0(\varepsilon)-1} \alpha_{t_i+t+1} \\
&\geq 1 - n_0(\varepsilon) \max\{\alpha_t \mid t = t_i + 1, \ldots, t_i + n_0(\varepsilon)\}.
\end{aligned}$$

Given the previous results, for any $\varepsilon > 0$, there exists an $i > i_0$, such that $\mathbb{P}(\tau_i \geq n_0(\varepsilon)) > 1 - \varepsilon$. Further, for any $\varepsilon > 0$, and i sufficiently large,

$$\begin{aligned}
&\mathbb{P}((\Delta_{t_{i+1}} \backslash \Delta_{t_i}) = \Delta) \\
&= \sum_{\delta \in Q_{t_i}^{(\Delta)}} \mathbb{P}((\Delta_{t_{i+1}} \backslash \Delta_{t_i}) = \Delta, X_{t_i+1} = \delta) \\
&= \sum_{\delta \in Q_{t_i}^{(\Delta)}} \mathbb{P}((\Delta_{t_{i+1}} \backslash \Delta_{t_i}) = \Delta \mid X_{t_i+1} = \delta) \mathbb{P}(X_{t_i+1} = \delta) \\
&\geq \sum_{\delta \in Q_{t_i}^{(\Delta)}} \mathbb{P}(\tau_i \geq n_0(\varepsilon)) \mathbb{P}((\Delta_{t_{i+1}} \backslash \Delta_{t_i}) = \Delta \mid X_{t_i+1} = \delta) \\
&\quad \mathbb{P}(X_{t_i+1} = \delta) \\
&\geq (1 - \varepsilon)^2 \sum_{\delta \in Q_{t_i}^{(\Delta)}} \mathbb{P}(X_{t_i+1} = \delta) \\
&\geq (1 - \varepsilon)^2.
\end{aligned}$$

Thus, the probability that the interacting processes visit every possible state of Δ between the interaction times, converges to unity, as $\alpha_t \to 0$ and $t_i \to \infty$. This fact is then sufficient to establish the result stated in Theorem 1.

Acknowledgment The work of J.C. and T.K. was supported by the Centre of Population Genetic Analyses, University of Oulu, Finland (Academy of Finland, grant no. 53297).

References

Altekar G., Dwarkadas S., Huelsenbeck J.P., and Ronquist F. 2004. Parallel metropolis-coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20: 407–415.

Brooks S.P., Giudici P., and Roberts G.O. 2003. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. J. Roy. Statist. Soc. B 65: 3–39.

Carlin B.P. and Chib S. 1995. Bayesian model choice via Markov-chain Monte Carlo methods. J. Roy. Statist. Soc. B 57: 473–484.

Chib S. and Greenberg E. 1995. Understanding the Metropolis-Hastings algorithm. Amer. Statist. 49: 327–335.

Corander J., Gyllenberg M., and Koski T. 2006. Bayesian unsupervised classification framework based on stochastic partitions of data and a parallel search strategy. Submitted to J. Statist. Comput. Simulation.

Corander J., Waldmann P., Marttinen P., and Sillanpää M.J. 2004. BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics 20: 2363–2369.

Diaconis P., Holmes S., and Neal R.M. 2000. Analysis of a nonreversible Markov chain sampler. Ann. App. Prob. 10: 726–752.

Doucet A., Godsill S., and Andrieu C. 2000. On sequential Monte Carlo sampling methods for Bayesian filtering. Stat. Comput. 10: 197–208.

Farmer J.J., Davis B.R., and Hickmanbrenner F.W. 1985. Biochemical identification of new species and biogroups of Enterobacteriaceae isolated from clinical specimens. J. Clin. Microbiology 21: 46–76.

Geyer C.J. and Thompson E.A. 1995. Annealing Markov chain Monte Carlo with applications to ancestral inference. J. Amer. Stat. Assoc. 90: 909–920.

Gidas B. 1985. Nonstationary Markov chains and convergence of the annealing algorithm. J. Statist. Phys. 39: 73–130.

Giudici P. and Castelo R. 2003. Improving Markov chain Monte Carlo search for data mining. Machine Learning 50: 127–158.

Green P. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82: 711–732.

Gyllenberg H.G., Gyllenberg M., Koski T., Lund T., Schindler J., and Verlaan, M. 1997. Classification of Enterobacteriaceae by minimization of stochastic complexity. Microbiology 143: 721–732.

Gyllenberg H.G., Gyllenberg M., Koski T., and Lund T. 1998. Stochastic complexity as a taxonomic tool. Computer Methods and Programs in Biomedicine 56: 11–22.

Gyllenberg H.G., Gyllenberg M., Koski T., Lund T., and Schindler J. 1999a. An assessment of cumulative classification. Quantitative Microbiology 1: 7–27.

Gyllenberg H.G., Gyllenberg M., Koski T., Lund T., and Schindler J. 1999b. Enterobacteriaceae taxonomy approached by minimization of stochastic complexity. Quantitative Microbiology 1: 157–170.

Gyllenberg H.G., Gyllenberg M., Koski T., Lund T., Mannila H., and Meek C. 1999c. Singling out ill-fit items in a classification. Application to the taxonomy of Enterobacteriaceae. Archives of Control Sciences 9: 97–105.

Gyllenberg M., Koski T., Lund T., and Gyllenberg H.G. 1999. Bayesian predictive identification and cumulative classification of bacteria. Bulletin of Mathematical Biology 61: 85–111.

Häggström O. 2002. Finite Markov Chains and Algorithmic Applications. Cambridge, Cambridge University Press.

Isaacson D.L. and Madsen R.W. 1976. Markov Chains: Theory and Applications, New York, Wiley.

Jensen S.T., Liu S., Zhou Q., and Liu J.S. 2004. Computational discovery of gene regulatory binding motifs: A Bayesian perspective. Stat. Sci. 19: 188–204.

Laskey K.B. and Myers J.W. 2003. Population Markov chain Monte Carlo. Machine Learning 50: 175–196.

Robert C.P. and Casella G. 2005. Monte Carlo Statistical Methods. 2nd edition, New York, Springer.

Schervish M.J. 1995. Theory of Statistics. New York, Springer.

Sisson S.A. 2005. Transdimensional Markov chains: A decade of progress and future perspectives. J. Amer. Stat. Assoc. 100: 1077–1089.

Tierney L.M. 1994. Markov chains for exploring posterior distributions. Ann. Statist. 22: 1701–1728.