Bayesian unsupervised classification framework based on stochastic partitions of data and a parallel search strategy

Jukka Corander · Mats Gyllenberg · Timo Koski

Received: 20 June 2007 / Revised: 13 June 2008 / Accepted: 13 February 2009 /
Published online: 4 March 2009
© Springer-Verlag 2009

Abstract Advantages of statistical model-based unsupervised classification over heuristic alternatives have been widely demonstrated in the scientific literature. However, the existing model-based approaches are often both conceptually and numerically instable for large and complex data sets. Here we consider a Bayesian model-based method for unsupervised classification of discrete valued vectors, that has certain advantages over standard solutions based on latent class models. Our theoretical formulation defines a posterior probability measure on the space of classification solutions corresponding to stochastic partitions of observed data. To efficiently explore the classification space we use a parallel search strategy based on non-reversible stochastic processes. A decision-theoretic approach is utilized to formalize the inferential process in the context of unsupervised classification. Both real and simulated data sets are used for the illustration of the discussed methods.

Keywords Bayesian classification · Markov chain Monte Carlo · Statistical learning · Stochastic optimization

Mathematics Subject Classification (2000) 62H30 · 68T10

J. Corander (✉)

Department of Mathematics, Åbo Akademi University, Turku 20500, Finland
e-mail: jukka.corander@abo.fi

M. Gyllenberg

Department of Mathematics and Statistics, Rolf Nevanlinna Institute,
University of Helsinki, P. O. Box 68, Helsinki 00014, Finland

T. Koski

Department of Mathematics, Royal Institute of Technology,
100 44 Stockholm, Sweden

1 Introduction

Model-based classification is a useful tool for pattern exploration in a wide range of applications, ranging from pure data mining to allocation of observations to pre-defined classes. An *unsupervised classification*, also termed *clustering*, of observed data may be understood as a scientific learning task where a putative hidden structure underlying the observations is to be discovered. Although heuristic numerical methods for classification have been widely exploited for decades already, there has been a gradually increasing interest in the model-based approach as the computational restrictions have become less significant. For a broad survey of model-based strategies to unsupervised classification, see Bock (1996). In particular, in many of the papers reviewed by Bock (1996), advantages of a model-based approach over the heuristic alternatives have been clearly demonstrated.

The so far most exploited model-based strategy to unsupervised classification is theoretically formulated in terms of the so called latent class model (see, e.g. Duda et al. 2000), for which practical implementations are typically based on the standard EM- or Gibbs sampler-algorithm (Robert and Casella 1999). To some extent, the reversible jump Metropolis–Hastings algorithm (see, Green 1995; Robert and Casella 1999; Sisson 2005) has also been used in this context to fit Bayesian unsupervised classification models. However, for large and complex data sets, the latent class modelling framework using standard Monte Carlo computation may yield highly unstable results due to intrinsic difficulties in the class identifiability and computational tractability.

Here we consider the general case of clustering items represented by discrete-valued feature vectors. A predictive Bayesian approach is described by utilizing generalizations of the de Finetti type representation results under a model which generates stochastic partitions of data based on an random urn allocation scheme. Such a model structure was earlier considered in a molecular biological context by Corander et al. (2004, 2007). The stochastic partition framework enables a parallel search strategy based on non-reversible Markov chains, where distinct search processes can learn from each other in an unsupervised manner. To improve upon our previous work on this modelling strategy, we treat, in particular, formally the question of how inferences should be made concerning cluster contents, using a decision-theoretic approach.

The structure of this article is as follows. In Sect. 2 we treat the unsupervised classification problem using a predictive Bayesian approach. Statistical MCMC-based learning with parallel stochastic processes is considered in the section thereafter. A decision-theoretic formulation of the problem of making inferences about the cluster structures is developed in Sect. 4. Numerical illustrations for real and simulated data sets are given in Sect. 5, and some remarks are provided in the final section.

2 Model-based unsupervised predictive classification

Consider a set N of n items for which we have available finite feature vectors $\mathbf{x}_i$, such that each element x_{ij} in $\mathbf{x}_i$ belongs to a discrete alphabet, $x_{ij} \in \mathcal{X}_j = \{1, \dots, r_j\}$, $r_j \geq 1$, $j = 1, \dots, d$. The feature space is thus represented by the Cartesian product

$$\mathbf{x}_i \in \mathcal{X} = \prod_{j=1}^d \mathcal{X}_j, \quad (1)$$

and we use $\mathbf{x}^{(N)} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ to denote the set of the n feature vectors. When the value of at least one single element x_{ij} is unknown, the observations $\mathbf{x}^{(N)}$ contain missing data. Under the commonly utilized assumption of complete randomness in the patterns of presence of missing data, the missing observations are coherently handled in the predictive unsupervised classification model described in detail below.

A common probabilistic formulation of the unsupervised classification problem is obtained through a latent class model (Duda et al. 2000) with a fixed number k of classes, which capture distinct patterns in the data. Such patterns can be thought emerging when sampling data vectors from a distribution having substantial local probability mass concentrations scattered over the space $\mathcal{X}$. The rationale behind the statistical inference based on a latent class formulation is that heterogeneous data sets cannot be adequately represented using a single distribution from a parametric family. By allowing k to increase, the model is in principle able to capture in a probabilistic manner even subtle characteristics of an underlying structured population of items. As a consequence, ability to predict features for future data is improved, given that the population has a sufficiently stable representation. However, in this respect it is also of importance to avoid an overly complex model, which typically results in poor predictive performance.

Gyllenberg et al. (1997), and Gyllenberg and Koski (2001, 2002) have considered Bayesian classification of binary vectors using a latent class formulation combined with the principle of minimizing stochastic complexity (see Rissanen 1987, 1995). This strategy is asymptotically equivalent to maximization of the posterior probability of a classification, however, the latter formulation provides a more natural means for application of stochastic estimation methods, such as Markov chain Monte Carlo (MCMC). Also, another benefit of the Bayesian formulation is the possibility of deriving intuitive measures of classification uncertainty in addition to the optimal estimates under a specified loss function, which will be discussed in Sect. 4.

To formulate the unsupervised classification problem using a framework distinct from the latent class model, one can use an approach based on stochastic partitions of N (Corander et al. 2004, 2007). Let $S = (s_1, \dots, s_k)$, ($1 \leq k \leq n$) be a partition of N , i.e., a collection of subsets or *classes* of N , such that $\bigcup_{c=1}^k s_c = N$, $s_c \cap s_{c^*} = \emptyset$, for all $c, c^* = 1, \dots, k$, $c \neq c^*$. We let $\mathcal{S}$ denote the space of such partitions. The uncertainty related to the grouping of the items prior to actually observing their features can be described by introducing the probability measure

$$p(S) \geq 0, \quad S \in \mathcal{S}, \quad \sum_{S \in \mathcal{S}} p(S) = 1, \quad (2)$$

which is a prior distribution over the space of possible classification structures. A default impartial option for the choice of prior $p(S)$ is the uniform distribution in $\mathcal{S}$:

$$p(S) = |\mathcal{S}|^{-1}, \quad S \in \mathcal{S}, \quad (3)$$

which arises under a particular random urn model, see [Corander et al. \(2007\)](#). This is an example of random generation of combinatorial structures as discussed by [Van Cutsem \(1996\)](#).

To do statistical learning with respect to the putative partitions in S , given the observed feature data $\mathbf{x}^{(N)}$, it is necessary to derive an expression for the predictive likelihood of $\mathbf{x}^{(N)}$ conditional on any particular partition S . Let the predictive likelihood be denoted by $p(\mathbf{x}^{(N)}|S)$. By Bayes' theorem, the posterior of S equals then

$$p(S|\mathbf{x}^{(N)}) = \frac{p(\mathbf{x}^{(N)}|S)p(S)}{\sum_{S \in \mathcal{S}} p(\mathbf{x}^{(N)}|S)p(S)}, \quad (4)$$

which provides the basis for statistical learning under the Bayesian paradigm. However, derivation of a formal inference rule still requires the specification of a loss function in addition to the posterior probabilities, which will be considered in detail in Sect. 4. In the remainder of this section we show how a concrete form of the predictive likelihood $p(\mathbf{x}^{(N)}|S)$ arises from generalizations of the de Finetti representation theorem (see [Bernardo and Smith 1994](#); [Schervish 1995](#)).

Let $\mathbf{x}^{(s)}$ denote the observed feature vectors for any non-empty subset $s \subseteq N$ of the n items, and let cardinalities of sets be in general denoted by $|\cdot|$. We assume that the observed sequences of features $\mathbf{x}^{(s)}$ are *unrestrictedly infinitely exchangeable* (see Definitions 4.2, 4.3, and 4.13, and Propositions 4.2 and 4.18 in [Bernardo and Smith 1994](#)). This assumption is equivalent with the property that, if we combine any permutation of the item values $x_{1j}, \dots, x_{|s|j}$, for a fixed feature j , with arbitrary corresponding permutations over the remaining features $j^* = 1, \dots, d$, $j^* \neq j$, the same predictive probability mass function for $\mathbf{x}^{(s)}$ is obtained. Furthermore, we obtain a characterization of the sequences $x_{1j}, \dots, x_{|s|j}$ in terms of the sufficient statistics n_{jl}^s , $l = 1, \dots, r_j$, where n_{jl}^s represents the number of copies of value l for feature j observed among the items in s .

The unique probabilistic characterization of $\mathbf{x}^{(s)}$ under unrestricted infinite exchangeability equals

$$p(\mathbf{x}^{(s)}) = \int_{\Theta} \prod_{j=1}^d \prod_{l=1}^{r_j} p_{jl}^{n_{jl}^s} dQ(\theta), \quad (5)$$

where p_{jl} can be interpreted as the limit of the relative frequency of observing the value l for feature j among items sampled from an underlying class, i.e.

$$p_{jl} = \lim_{|s| \rightarrow \infty} \frac{n_{jl}^s}{|s|}. \quad (6)$$

Thus, in the context of a traditional latent class formulation this quantity is a class-conditional probability of an outcome for a feature. The symbol $\theta = (p_{11}, \dots, p_{1r_1}, \dots, p_{d1}, \dots, p_{dr_d})$ in (5) refers jointly to p_{jl} over different values and features, and $Q(\theta)$ is a probability measure over the space Θ , representing our prior beliefs about the limits of the relative frequencies in (6). The probability mass function in (5) defines

a predictive probability model for the observed items in s . It is important to notice that the above limit is taken relative to a fixed classification S . This, however, has very restricted practical consequences, as the only role of (6) is the operational interpretation provided for the quantities in (5) (see the discussion of exchangeability in [Bernardo and Smith 1994](#)).

By extending the unrestricted exchangeability assumption to hold over the classes $s_1, \dots, s_k$, we obtain the joint probabilistic characterization for $\mathbf{x}^{(N)} = \{\mathbf{x}^{(s_1)}, \dots, \mathbf{x}^{(s_k)}\}$ as

$$p(\mathbf{x}^{(N)}|S) = \int_{\Theta} \prod_{c=1}^k \prod_{j=1}^d \prod_{l=1}^{r_j} p_{cjl}^{n_{cjl}} dQ(\theta|S), \quad (7)$$

where the sufficient statistic n_{cjl} now represents the number of copies of value l for feature j observed among the items in s_c , and p_{cjl} , θ , and $Q(\theta|S)$, are defined analogously to (5) and (6).

The concrete probabilistic implication of the utilized exchangeability assumptions is that all observed features are considered conditionally independent given the stochastic partition S . The sufficient statistics n_{cjl} emerging under any particular value of S reflect how missing observations are explicitly handled by the predictive model in a statistically sensible manner. Consider a partition S with k classes. Then, the statistical uncertainty regarding the allocation of any particular item i to any of the k classes increases as a function of the number of missing observations in $\mathbf{x}_i$ due to the form of (7), where the values of the sufficient statistics n_{cjl} are changing to a lesser extent when the item is moved between the classes. This can also be seen from the formula (9) derived below. In the most extreme case, when only missing observations are present for an item, the sufficient statistics do not change at all when re-allocating the item, and obviously, the likelihood cannot be informative about the classification.

The extension of the unrestricted exchangeability assumption implies that the observed values of features within an arbitrary class provide no information about the values observed in any other class. The probability mass function in (7) defines a predictive probability model for the observed items in N . Our assumptions also lead to a more explicit form of the prior beliefs $Q(\theta|S)$, i.e., the product Dirichlet distribution

$$Q(\theta|S) \propto \prod_{c=1}^k \prod_{j=1}^d \prod_{l=1}^{r_j} p_{cjl}^{\lambda_{cjl}-1}, \quad (8)$$

where the hyperparameter $\lambda_{cjl} > 0$, for all index values (see [Zabell 1982](#)). Under this result, the so called sufficientness postulate, the explicit predictive probability mass function for the items equals

$$p(\mathbf{x}^{(N)}|S) = \prod_{c=1}^k \prod_{j=1}^d \frac{\Gamma(\sum_{l=1}^{r_j} \lambda_{cjl})}{\Gamma(\sum_{l=1}^{r_j} \lambda_{cjl} + n_{cjl})} \prod_{l=1}^{r_j} \frac{\Gamma(\lambda_{cjl} + n_{cjl})}{\Gamma(\lambda_{cjl})}, \quad (9)$$

where $\Gamma(\cdot)$ is the gamma function. To specify the hyperparameters, we utilize the well-known reference distribution derived originally by [Perks \(1947\)](#), which determines the hyperparameter values as

$$\lambda_{cjl} = r_j^{-1}, \quad l = 1, \dots, r_j; \quad j = 1, \dots, d; \quad c = 1, \dots, k. \quad (10)$$

3 A Bayesian learning algorithm

The Bayesian probability-based specification of the plausibility of each classification facilitates the construction of a reliable learning algorithm, since completely heuristic search strategies can then be avoided. Using an analogous classification model for molecular marker data, [Corander et al. \(2004\)](#) introduced an MCMC algorithm which provide a consistent estimate of (4). The algorithm utilizes an approach of simulating several independent parallel Markov chains, which appears to perform reliably even for fairly large data sets. However, the computational effort required may become in practice prohibitive for complex data sets. By introducing a dependence between the chains, and by modifying the transition mechanisms of the individual chains, it is possible to obtain considerably more efficient statistical learning algorithms. Such an MCMC strategy is utilized here, and it belongs to the class of general non-reversible Metropolis-Hastings algorithms for parallel Bayesian model learning proven to be consistent by [Corander et al. \(2006\)](#).

For general properties of MCMC methods and for theoretical aspects of Markov chains, we refer to [Robert and Casella \(1999\)](#), and [Isaacson and Madsen \(1976\)](#), respectively. [Corander et al. \(2004\)](#) considered a Metropolis-Hastings algorithm, which simulates a Markov chain having the time homogeneous distribution equal to (4). This algorithm is defined by the transition kernel, which determines the probability of a transition from a current classification S to a new proposal classification S^* , as

$$\min \left(1, \frac{p(\mathbf{x}^{(N)}|S^*) p(S^*) q(S|S^*)}{p(\mathbf{x}^{(N)}|S) p(S) q(S^*|S)} \right), \quad (11)$$

where $p(\mathbf{x}^{(N)}|\cdot)$ is defined according to (9), $p(S)$ is the prior, $q(S^*|S)$ is the probability of choosing classification S^* as the new candidate when in S , and $q(S|S^*)$ is the probability of restoration of the current classification S . The proposal mechanism to derive S^* from S considered by [Corander et al. \(2004\)](#) was constructed from the following four different possibilities: split/merge of existing classes, re-allocation/exchange of items among existing classes. These simple transition mechanisms are analogous to those used by [Dawson and Belkhir \(2001\)](#), and define an aperiodic and irreducible finite Markov chain. The transition mechanisms are also similar to those generally used in the reversible jump MCMC algorithm introduced by [Green \(1995\)](#) for similar models, see also [Sisson \(2005\)](#).

[Aarts and Korst \(1989\)](#) showed that MCMC algorithms such as the Metropolis-Hastings algorithm defined above, are an efficient tool for solving combinatorial optimization problems like the one defined by the attempt to identify the partition maximizing the posterior probability (4). Since the state space $\mathcal{S}$ of the chain is finite,

it follows (e.g. Häggström 2002) that (11) defines also a positive recurrent Markov chain. Thus, for a realization of the chain $\{S_t, t = 0, 1, \dots\}$, the standard convergence result holds as

$$\lim_{T \rightarrow \infty} p_T \left(S | \mathbf{x}^{(N)} \right) = p \left(S | \mathbf{x}^{(N)} \right), \quad (12)$$

where

$$p_T \left(S | \mathbf{x}^{(N)} \right) = T^{-1} \sum_{t=0}^T I(S_t = S) \quad (13)$$

is the relative frequency of occurrence of state S . However, convergence of the chain may be slow in reality for large sets of items, and therefore, an alternative estimate is preferable

$$p_n^* \left(S | \mathbf{x}^{(N)} \right) = \frac{p \left(\mathbf{x}^{(N)} | S \right) p(S)}{\sum_{S \in \mathcal{S}^*} p \left(\mathbf{x}^{(N)} | S \right) p(S)}, \quad (14)$$

where $\mathcal{S}^*$ is the set of distinct values of S visited in m independent realizations of the Markov chain $\{S_{tj}, t = 0, 1, \dots; j = 1, \dots, m\}$.

To improve the learning algorithm we will use the two generic modifications to the Metropolis–Hastings algorithm introduced in Corander et al. (2006), namely, non-reversibility and interaction events among the chains over time. This is achieved by considering a positive recurrent non-reversible Markov chain $\{S_t, t = 0, 1, \dots\}$ with the transition kernel

$$\min \left(1, \frac{p(\mathbf{x}^{(N)} | S^*) p(S^*)}{p(\mathbf{x}^{(N)} | S) p(S)} \right), \quad (15)$$

where the predictive probabilities are defined as in (11), and where the proposal classifications are generated according to the above described mechanism. The following algorithm embeds the non-reversible transition kernel (15) into parallel process with interaction events to estimate the posterior probabilities (4).

Definition 1 A Bayesian parallel stochastic search algorithm

Let $\{S_{tj}, t = 0, 1, \dots; j = 1, \dots, m\}$ and $\{Z_t, t = 0, 1, \dots\}$ be $m + 1$ stochastic processes defined as follows:

- (1) Let $\alpha_1 = 0$, and define the sequence of probabilities $\{\alpha_t, t = 2, 3, \dots\}$ according to

$$\alpha_t = \frac{1}{q \log t},$$

where $q \geq 1$ can be chosen suitably, for instance $q \in [5, 10]$.

- (2) $Z_0 = 0$, and $p(Z_t = 1) = \alpha_t$, $p(Z_t = 0) = 1 - \alpha_t$, independently for $t = 1, 2, \dots$.
- (3) Let S_{0j} , $j = 1, \dots, m$, be arbitrary initial states of $\{S_{tj}, t = 0, 1, \dots; j = 1, \dots, m\}$. For each t , such that $Z_t = 0$, transition to the next state $S_{(t+1)j}$ is determined according to the (15) using proposal values generated as in (11), independently for $j = 1, \dots, m$.
- (4) For each $t = 0, 1, \dots$, define the distribution

$$p_t(S_{tj}) = \frac{p(\mathbf{x}^{(N)}|S_{tj}) p(S_{tj})}{\sum_{j=1}^m p(\mathbf{x}^{(N)}|S_{tj}) p(S_{tj})}, \quad (16)$$

over the space of the current states $\{S_{t1}, S_{t2}, \dots, S_{tm}\}$. For each $t = 0, 1, \dots$ such that $Z_t = 1$, the transition to the next state is determined according to distribution (16), such that the next state for each chain is sampled from this distribution, independently for $j = 1, \dots, m$.

The m processes in $\mathcal{S}$ according to the above definition are not time homogeneous Markov chains, however, as $t \rightarrow \infty$, their transition probabilities become increasingly similar to those defined for the Markov chain having the (unknown) stationary distribution determined by the proposal distribution in (11) and the acceptance probability in (15). The m processes behave analogously to independent time homogeneous Markov chains between any two consecutive times, where $Z_t = 1$ equals unity. As the interaction probability α_t approaches zero when $t \rightarrow \infty$, the length of time intervals between interactions increases towards infinity.

Forced by the conditional probabilities (16), the processes tend to coalesce towards the states of $\mathcal{S}$ among $\{S_{t1}, S_{t2}, \dots, S_{tm}\}$, which are associated with better predictive ability for the observed items. Also, in cases where multiple classifications with roughly equally good predictive values of $\mathcal{S}$ are present, the form of (16) effectively prevents the chains from coalescing into a single value.

The specification of the probabilities α_t governing the possibility for the processes to interact, has been deliberately chosen to allow chains to interact feasibly often and to prevent too rapid isolation of them. The logarithmic rate of decrease in the probabilities is similar to the cooling schedule required for consistency and optimality for simulated annealing, see e.g. Gidas (1985). However, the exact value of q in the definition of α_t is not of considerable importance in practice, as long as it is within a reasonable range, such that the chains neither interact too seldom or too often. In the former case, a considerable amount of simulation effort is wasted if many chains are trapped to local modes of the posterior, whereas some have identified areas associated with much higher probabilities, as interaction would then enable chains to escape from the local modes. On the other hand, it is not sensible to attempt to let the chains interact all the time, as this would in general tend not to markedly change the conditional posterior distribution of the states of the process at any given point in time.

As shown in Corander et al. (2006) for the general case of Bayesian model learning with a finite model space, (14) is also a consistent estimate of $p(\mathcal{S}|\mathbf{x}^{(N)})$ under the interacting parallel processes as defined above. The advantages of using this estimate

rather than (13) are the relative stability of the sum $\sum_{S \in \mathcal{S}^*} p(\mathbf{x}|S)p(S)$, and that results from independent realizations of the chain can be joined in a meaningful way, which is not possible for (13). In the relative frequency based estimation, some chains may become stuck in regions of $\mathcal{S}$ associated with low values of $p(\mathbf{x}^{(N)}|S)p(S)$, and obtain thereby too much weight in the estimate (13).

4 Decision rules for inferring unsupervised classifications

4.1 Decision rules

In our earlier works where the random partition models for classifying molecular biological data were investigated, a primary candidate for the optimal classification was simply suggested to be the partition maximizing the posterior probability

$$\hat{S} = \arg \max_{S \in \mathcal{S}} p(S|\mathbf{x}^{(N)}). \quad (17)$$

Here we consider the issue of inferring partitions more formally by using the general Bayesian decision-theoretic framework (see, e.g. [Bernardo and Smith 1994](#); [Schervish 1995](#)).

Let ω denote generally the value of an estimator of a partition of data, provided by a decision rule $\delta(\mathbf{x}^{(N)})$. Further, let $L(S, \omega) \geq 0$ be a loss function specifying the loss incurred by the use of ω when S would be the true generating partition. The posterior risk associated with a decision rule equals

$$r_{\delta(\mathbf{x}^{(N)})} = \sum_{S \in \mathcal{S}} L(S, \omega) p(S|\mathbf{x}^{(N)}), \quad (18)$$

and the optimal partition to be reported on the basis of the information in $\mathbf{x}^{(N)}$ is then

$$\omega_{opt} = \arg \min_{\omega \in \mathcal{S}} r_{\delta(\mathbf{x}^{(N)})}. \quad (19)$$

Analogously to general estimation problems, the metric implied by the loss function plays a central role in the inference. However, the choice of an appropriate metric is more complicated in the context of partitions due to discreteness of the space $\mathcal{S}$. A variety of different putative metrics are considered in the next subsection.

Apart from the most simplistic loss functions $L(S, \omega)$, it is challenging to derive an estimator minimizing the risk (18), as no analytical approach can be directly utilized. Theorem 2 below shows how a posterior-optimal estimate can be obtained using the previously described MCMC approach.

Let $\mathcal{S}_m = \{S_j, j = 1, \dots, m\}$ be a sample of m distinct partitions from the posterior distribution (4) obtained using the parallel interacting search processes according to Definition 1. The corresponding estimates of the posterior probabilities of the

partitions equal

$$\hat{p}(S_j | \mathbf{x}^{(N)}) = \frac{p(\mathbf{x}^{(N)} | S_j) p(S_j)}{\sum_{S_j \in \mathcal{S}_m} p(\mathbf{x}^{(N)} | S_j) p(S_j)}. \quad (20)$$

Let $\hat{r}_{\delta(\mathbf{x}^{(N)})} = \sum_{S_j \in \mathcal{S}_m} L(S_j, \omega) \hat{p}(S_j | \mathbf{x}^{(N)})$ be the empirical posterior risk associated with the estimate ω . Define the probability distribution

$$p(\omega) \propto \exp(-\hat{r}_{\delta(\mathbf{x}^{(N)})}), \quad \omega \in \mathcal{S}, \quad (21)$$

and define a partition estimator minimizing the empirical risk according to

$$\hat{\omega}_{opt}^{(t)} = \arg \min_{\omega \in \mathcal{S}_t} \hat{r}_{\delta(\mathbf{x}^{(N)})}, \quad (22)$$

where $\mathcal{S}_t \subseteq \mathcal{S}$ is the space of partitions visited by a parallel stochastic process otherwise analogous to Definition 1, except that the terms $p(\mathbf{x}^{(N)} | S)p(S)$ in the transition kernels are replaced by $\exp(-\hat{r}_{\delta(\mathbf{x}^{(N)})})$ for the corresponding partitions. Thus, the distribution (21) reflects the behavior of the risk function of a partition estimator, and it can be used for identification of the posterior optimal partition as stated in the following theorem.

Theorem 2 (Identification of the posterior optimal partition using stochastic optimization). *Let the global estimate minimizing the empirical posterior risk be denoted as $\hat{\omega}_{opt} = \arg \min_{\omega \in \mathcal{S}} \hat{r}_{\delta(\mathbf{x}^{(N)})}$. Then, $\hat{\omega}_{opt}^{(t)} \xrightarrow{a.s.} \hat{\omega}_{opt}$, as $t \rightarrow \infty$, and $\hat{\omega}_{opt}^{(t)} \xrightarrow{a.s.} \omega_{opt}$, as both $t \rightarrow \infty$ and $\mathcal{S}_m \rightarrow \mathcal{S}$.*

Proof The theorem is a direct consequence of the general convergence result for non-reversible interacting processes stated in Corander et al. (2006). For a modified proof of this convergence result with slightly weaker conditions, see also Corander et al. (2008). $\square$

The above theorem shows that a consistent estimate of the empirical posterior risk minimizing partition can be obtained by a stochastic parallel search otherwise analogous to that used to estimate the posterior probabilities, except that the kernel function of the distribution is replaced by (21). Hence, the mode of the distribution corresponds to the partition minimizing the empirical risk. Moreover, as the posterior sample size increases, the empirical risk function converges to the true posterior risk function.

Apart from the optimal classification structure, in some applications it may be of interest to do also other types of inference. From the sample $\mathcal{S}_m$ of the posterior distribution over the space of partitions, various interesting uncertainty characterizations related to the classification may be derived. For instance, the marginal probability of the number of classes supported by the data, is represented by the estimate

$$\hat{p}(|S| = k | \mathbf{x}^{(N)}) = \sum_{S \in \mathcal{S}_m : |S|=k} p(S | \mathbf{x}^{(N)}) p(S). \quad (23)$$

Similarly, the conditional posterior uncertainty about the allocation of any single item, given the empirical risk minimizing partition $\hat{\omega}_{opt}^{(t)}$, may be characterized using the probabilities defined as follows. Let $S_{-i} = (s_1, \dots, s_k)$ be a partition of $N \setminus \{i\}$, and let $S_{-i} \otimes_c \{i\} = (s_1, \dots, s_c \cup \{i\}, \dots, s_k)$ be the partition where the class s_c of S_{-i} is augmented with i . The conditional probabilities

$$\frac{p(S_{-i} \otimes_c \{i\} | \mathbf{x}^{(N)})}{\sum_{c=1}^k p(S_{-i} \otimes_c \{i\} | \mathbf{x}^{(N)})}, \quad (24)$$

then reflect how strongly the data assigns a particular item to a particular class in the local neighborhood of a partition.

4.2 Loss functions and partition metrics

Perhaps the simplest possible loss function for the estimation of partitions is the zero-one loss, which results in the estimate in (17). Formally, this loss function can be defined as

$$L(S, \omega) = \begin{cases} -1, & \text{if } \omega = S \\ 0, & \text{if } \omega \neq S \end{cases}, \quad (25)$$

from which it directly follows that the estimate minimizing the posterior risk (18) equals the posterior mode (17), as the risk of the estimate ω equals $-p(\omega | \mathbf{x}^{(N)})$. The zero-one loss function can be considered plausible in situations where the posterior distribution (4) is highly concentrated on a single partition over the space $\mathcal{S}$. Then, attempts to use a synthesis of many partitions yield a negligible effect due to the form of the weights in the risk expression (18). On the other hand, when the posterior mass is more evenly distributed over a subspace of $\mathcal{S}$, other loss functions may be expected to have more reasonable statistical properties.

One of the most elementary metrics in the partition space is the normalized Hamming distance

$$D_1(S, S^*) = 1 - \frac{A(S, S^*)}{\binom{n}{2}}, \quad (26)$$

where $A(S, S^*)$ is the number of pairs of items for which the two partitions agree, i.e. an agreement is present for a pair of items when either they are in the same class, or in different classes, both in S and S^* . Usually, $1 - D_1(S, S^*)$ is referred to as the Rand index (Rand 1971). Ramifications of the basic Rand index were discussed in Hubert and Arabie (1985), as it was noticed that the Rand index easily attains elevated values towards unity when n is large. The adjusted Rand index introduced by Hubert and Arabie (1985) is more attractive in this respect, as its expected value for two random partitions of a set of items approaches zero. Using the notation by

Hubert and Arabie (1985), the adjusted Rand index is defined as

$$AR(S, S^*) = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} \right] - \sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} / \binom{n}{2}}, \quad (27)$$

where n_{ij} is the size of the intersection of the i th class in S and j th class in S^* , and further, $n_{i\cdot}$ and $n_{\cdot j}$ are the total sizes of the corresponding classes, respectively. A distance measure may then be defined as $D_2(S, S^*) = 1 - AR(S, S^*)$.

Yet another alternative metric is provided by considering the Shannon metric for partitions, which is defined in terms of the entropy. Let the entropy $H(S)$ of a partition S be

$$H(S) = - \sum_{c=1}^k p_c \log_2 p_c, \quad (28)$$

where $p_c = |s_c|/n$. Furthermore, let the conditional entropy $H(S^*|S)$ be defined as

$$H(S^*|S) = - \sum_{c=1}^k p_c \sum_{c^*=1}^{k^*} p_{c^*|c} \log_2 p_{c^*|c}, \quad (29)$$

where $p_{c^*|c} = |s_{c^*} \cap s_c|/|s_c|$, and c^* is the index for the k^* classes in S^* . The Shannon metric is now given as

$$D_3(S, S^*) = H(S^*|S) + H(S|S^*). \quad (30)$$

Any of the above metrics, or any other metrics as well, can be used as a loss function $L(S, \omega)$ to determine the optimal partition, which minimizes the weighted distance to the other partitions in $\mathcal{S}$. Thus, for such loss functions the posterior risk minimizing estimate can be interpreted as a “mean partition” in a certain sense. In Sect. 5 we illustrate using a simulation study the relative behavior of the above three metrics and the approximations discussed below.

4.3 Approximations to decision rules for inferring partitions

The zero-one loss function $L(S, \omega)$ in (25) is numerically particularly attractive, as the corresponding optimal estimate may be easily identified given the posterior sample $\mathcal{S}_m$. However, it may be anticipated that more elaborate loss functions based on suitable metrics in the partition space provide more satisfactory inferences in the ordinary statistical sense. Since a statistically consistent minimization of the empirical posterior risk function requires considerable computational effort, it is of interest to also investigate approximate and numerically fast strategies to obtain estimates that reflect the overall similarity of partitions.

One such strategy was informally suggested in Dawson and Belkhir (2001) and Corander et al. (2004) in the molecular biological unsupervised classification context.

In their strategy, the complete linkage clustering algorithm (e.g. [Mardia et al. 1979](#)) is utilized to process the information from a set of samples from the posterior distribution over the space of partitions. Using the posterior sample, [Dawson and Belkhir \(2001\)](#) calculate the distance between any two items as one minus the relative number of sampled partitions where the two items are assigned to the same class, because they use an ordinary reversible Metropolis–Hastings algorithm for the estimation. Instead, [Corander et al. \(2004\)](#) calculate the same distance as one minus the posterior probability of co-assignment using the analytically calculated posterior weights. This latter distance is obtained by first denoting the event that two items $\{i, j\}$ are assigned to the same class as $i \sim j$, for which the marginal posterior probability then equals

$$\hat{p}(i \sim j | \mathbf{x}^{(N)}) = \sum_{S \in \mathcal{S}_m: i \sim j} p(S | \mathbf{x}^{(N)}) p(S). \quad (31)$$

Both [Dawson and Belkhir \(2001\)](#) and [Corander et al. \(2004\)](#) considered the representation of the marginal posterior probabilities of co-assignment to provide a convenient tool for post-processing automatically the information in the posterior sample. Their choice can in fact be motivated by decision-theoretic arguments, as shown below.

Instead of defining the global loss related to a partition estimate ω as in (18), consider a related loss function, which determines the loss incurred by the use of ω for a particular pair of items, say $\{i, j\}$, such that

$$L_{\{i,j\}}(S, \omega) = \begin{cases} -1, & \text{if } i \sim j \text{ in } \omega \text{ and } i \not\sim j \text{ in } S \\ 0, & \text{otherwise} \end{cases}. \quad (32)$$

It then follows that the empirical risk associated with $L_{\{i,j\}}(S, \omega)$ equals the marginal posterior probabilities of co-assignment

$$\hat{r}_{\delta(\mathbf{x}^{(N)})}^{[i,j]} = \sum_{S_j \in \mathcal{S}_m} L_{\{i,j\}}(S_j, \omega) \hat{p}(S_j | \mathbf{x}^{(N)}) = -\hat{p}(i \sim j | \mathbf{x}^{(N)}). \quad (33)$$

Let $\mathbf{D}$ be a distance matrix for the n items, where the element (i, j) is determined by

$$\mathbf{D}_{ij} = 1 - \hat{p}(i \sim j | \mathbf{x}^{(N)}). \quad (34)$$

Using this particular distance matrix, it is possible to obtain a certain risk-minimizing estimate with a deterministic algorithm (complete linkage clustering algorithm), as shown by the following result. Let $\mathcal{T}$ be an ultrametric tree (e.g. [Mardia et al. 1979](#)) for a set of n items based on the distance matrix $\mathbf{D}$. Further, let $\mathcal{T}(d)$ denote the partition of N defined by $\mathcal{T}$ at the distance level $d \in [0, 1]$.

Theorem 3 (Identification of a posterior optimal partition using a deterministic algorithm). *Let $\mathcal{T}$ be the ultrametric tree obtained with the complete linkage clustering algorithm from $\mathbf{D}$ and let $\alpha \in [0, 1]$. Then, a partition estimate ω for which the empirical risk $\hat{r}_{\delta(\mathbf{x}^{(N)})}^{[i,j]}$ does not exceed $-\alpha$ for any pair of items $\{i, j\}$, is given by the partition $\mathcal{T}(1 - \alpha)$.*

Proof The theorem follows from the fundamental property of the complete linkage algorithm (Mardia et al. 1979), combined with the definition of the pairwise loss function. Namely, in the ultrametric tree $\mathcal{T}$ obtained with the complete linkage algorithm, lineages for a subset $s_c \subseteq N$ of items cannot coalesce at distance level d , unless all $\binom{s_c}{2}$ pairs of items have distances smaller than or equal to d . $\square$

However, the above approach still necessitates the use of a subjective judgement regarding the level of distance at which the ultrametric tree is cut, to obtain a partition estimate. To obtain a partition estimate minimizing the posterior risk $\hat{r}_{\delta(\mathbf{x}^{(N)})}$ without such a judgement, it is possible to utilize an ultrametric tree in an automated fashion by a numerical optimization procedure. This can be defined as the decision rule

$$\hat{\omega}_{opt} = \arg \min_{d \in [0, 1]: \omega = \mathcal{T}(d)} \hat{r}_{\delta(\mathbf{x}^{(N)})}, \quad (35)$$

which identifies a distance level d yielding from an ultrametric tree $\mathcal{T}$ a partition minimizing the posterior weighted average distances over $\mathcal{S}_m$. Such a deterministic decision-rule could also be used in combination with any hierarchical clustering algorithm, for instance single linkage or average linkage algorithm, in addition to the earlier mentioned complete linkage algorithm. Also, fast heuristic non-hierarchical clustering algorithms could be used for this purpose, by defining $\hat{r}_{\delta(\mathbf{x}^{(N)})}$ to be the objective function to be minimized.

As a summary of the decision-theoretic approach to inferring partitions, it can be concluded that this statistical problem involves two main components. The first component is the choice of an appropriate metric in the partition space and the second one is the choice of the computational approach to identify the partition minimizing the empirical risk function $\hat{r}_{\delta(\mathbf{x}^{(N)})}$ defined in Sect. 4.1. The approach based on stochastic computation to minimize the empirical risk is theoretically supported by its consistency property. However, as this method is computationally very demanding for large sets of items, heuristic approximations, such as those discussed above, are attractive alternative candidates for obtaining partition estimates minimizing an empirical risk function. On the other hand, it would be necessary to investigate their statistical performance in detail to make any general recommendations. In the next section we illustrate the behavior of the discussed approximations under the various metrics considered earlier, and compare their relative performance to the risk minimization based on stochastic optimization.

5 Illustrations of the Bayesian classification learning approach

5.1 Simulated example

In this section we first illustrate how various decision rules and their approximations behave when applied to a simulated data set, for which the underlying generating model is known. To specify a generating model structure, we simulate independent and identically distributed vectors of probabilities $\mathbf{p}_c$, with the elements $p_{cj} \sim \text{Uniform}(0, 1)$, for $c = 1, \dots, 10$; $j = 1, \dots, 50$. These probabilities are taken to characterize 10

Table 1 Distances between the generating partition and the partition inferred by combining metric D with the method M

$D \backslash M$	M_1	M_2	M_3	M_4	M_5
D_1	0.155	0.225	0.795	0.199	0.230
D_2	0.155	0.225	0.795	0.199	0.230
D_3	0.205	0.303	0.795	0.180	0.284

The metrics are as defined in Sect. 4.2. Method M_1 refers to the stochastic estimate considered in Theorem 2, $M_2 - M_4$ refer to deterministic estimates based on (35) and single, average and complete linkage hierarchical clustering algorithms, respectively, and M_5 refers to deterministic estimates based on (35) and Ward's hierarchical clustering method. All reported distances are defined according to the metric D_2 to ensure comparability of the values

distinct sources of data, each determined by its 50-dimensional vector of probabilities for binary attributes. Given the generated probabilities, 10 binary vectors of length 50 are independently sampled from each source, leading to 100 observed items in total. The information content of such a data set is deemed substantial, but not large enough to yield negligible uncertainty with respect to the posterior uncertainty concerning the structure of the unsupervised classification.

The parallel MCMC estimation procedure was applied to the data using 2000 iterations of 100 simultaneous processes, with the stochastic intensity of interaction determined by $q = 10$, which results on average in interaction events approximately after every [40,75] iterations of the algorithm in the considered range. For the stochastic partition model, we used here and in the subsequent real data analyses the same default choices of priors, i.e. $p(S) = |S|^{-1}$ and $p_{cj} \sim \text{Beta}(1/2, 1/2)$, $c = 1, \dots, k$; $j = 1, \dots, d$, which were considered in Sect. 2.

After the posterior sample $\mathcal{S}_m$ was extracted from the realization of the parallel processes, both the stochastic and deterministic approaches to obtaining the risk-minimizing partition estimate were used. The stochastic estimate was obtained by using exactly the same amount of simulation effort as originally used to create the sample $\mathcal{S}_m$. In Table 1, we have collected values of the distance between the generating partition and the various putative estimates discussed in the previous sections. These values can be compared with the distance (based on the metric D_2) between the mode partition of the sample $\mathcal{S}_m$ and the generating partition, which equals 0.172.

It is seen that for these simulated data, only the stochastic estimate based on either the metric D_1 or D_2 is closer to the generating partition than the mode estimate. This illustrates that an elaborate risk function may indeed yield more accurate inferences compared to the mode estimate, when the data are not very conclusive about how the item set should be classified. In fact, one would anticipate that this relationship between the two types of estimates holds quite generally when the posterior distribution is not very peaked in the partition space. Moreover, out of the deterministic approximate decision rules, only the one utilizing the complete linkage hierarchical clustering algorithm has relatively comparable accuracy with the estimates based on posterior mode and stochastic minimization of risk.

When comparing the plausibilities of the estimates it should be kept in mind that the values in Table 1 are based on the metric D_2 , which favors in principle the combination

$D_2 \setminus M$ for any method M . However, given that the number of generated items is only moderate, the adjustment in the Rand index is rather negligible in magnitude, and consequently, the two metrics D_1 , D_2 yield here identical partition estimates.

For the simulated data set with 100 items and 50 binary features, running 2000 iterations of 100 parallel interacting processes took ~ 5 min on a standard laptop computer equipped with a 1.66 GHz dual core processor. The increase in computation time induced by the interaction step, compared to independent parallel non-reversible processes, is of rather negligible magnitude, as the steps between the interaction events largely determine the amount of computational effort required by the algorithm.

5.2 Two applications from biology

We now illustrate our partition learning approach by considering unsupervised classification of two real biological data sets representing challenging applications. In both cases it is of central importance to learn the number of homogeneous groups present in the data on the basis of the observed features for each sampled individual. Given the extensive sizes of the data sets, we restrict our illustration here to the identification of the posterior optimal classifications with respect to (17). Also, for these data the expected gain from using minimization of the empirical risk function is rather small, due to the relatively high concentration of the estimated posterior distribution at the mode partition. Running the parallel MCMC algorithm over 1,000 iterations with 100 interacting processes took ~ 1 h for the first biological data set. As the second data set contains a considerably larger number of units and has a highly complex structure, it is also computationally much more challenging. A single run of the MCMC algorithm over 400,000 iterations with 50 parallel processes took ~ 10 days on a desktop PC equipped with a 2.8 GHz processor. To ensure the stability of the mode partition estimate, the MCMC algorithm was run twice using the same setup (#iterations, #chains).

The first data set comprises 1056 individuals from a worldwide sample of 52 human populations, who have been genotyped for 377 microsatellite marker loci in the genome. A microsatellite, also called simple sequence repeat, short tandem repeat or variable number tandem repeat, is a short (2–5 base) motif that is repeated multiple times and is flanked by unique DNA. Differences among individuals are generally detected as a length difference in the amplified product. In the human data, di-, tri- and tetranucleotide loci distributed over all 22 autosomal chromosomes were characterized for the sampled individuals. The same data set has been previously investigated by Corander et al. (2004) using reversible independent stochastic processes and by Rosenberg et al. (2002) using a Gibbs sampler algorithm and a latent class mixture model.

The actual data set consists for each individual of two observations on 377 discrete-valued features, for which the number of possible values varies between 4 and 32. The two observations correspond to the two alleles present in a diploid organism at any particular locus over the genome. The amount of missing data is 3.8%, and the missing observations are fairly uniformly distributed across the loci. Notice that the missing data are handled by the partition model as discussed in Sect. 2. Further

Fig. 1 Log-traces of 100 dependent non-reversible processes for the human data analyzed earlier in [Corander et al. \(2004\)](#) using independent reversible Markov chains

details about the data are available in the supplementary material of [Rosenberg et al. \(2002\)](#).

An important biological aspect in the present setting is characterization of the molecular variation observed in a sample from a target population. The biological hypothesis corresponding to any particular classification S according to the predictive model is equivalent to the statement that the individuals allocated in the same class represent a sample from a random mating unit of the investigated population, whereas any collection of classes does not constitute a random mating unit. In this sense a population is genetically structured when $k > 1$, meaning that the measured genomic characteristics differ on average systematically between individuals representing distinct ancestral sources. For the human population, this construct reflects the evolutionary history in terms of geographical colonization patterns over the world.

[Corander et al. \(2004\)](#) used 100 independent parallel MCMC chains to estimate the Bayesian classification of the same data set at sample population level. Although these were sufficient to produce a stable posterior estimate, it can be seen from ([Corander et al. 2004; Fig. 2](#)), that majority of the 100 chains had over 5000 iterations stayed on a considerably lower level of predictive value (9) and did not contribute in practice to the search of the posterior mode.

In [Fig. 1](#), we show logarithms of the traces of (9) produced by our interacting algorithm. The advantage of the interaction step is clearly visible, as the processes have rapidly coalesced towards values of S representing better predictive ability. Here, all processes have found the posterior mode after roughly 600 iterations, whereas in the analysis of [Corander et al. \(2004\)](#), first MCMC chain found the same mode after 1,000 iterations, and the majority of the chains did not manage to find the mode within 10,000 iterations. Our algorithm was thus able to identify the same posterior mode estimate of S ($k = 7$) as found in [Corander et al. \(2004\)](#), however, by utilizing only a fraction of computational resources.

The second unsupervised classification example utilizes data from Gyllenberg et al. (1997) who investigated 5313 strains of bacteria belonging to the family *Enterobacteriaceae*, representing 104 species or biogroups. The data consisted of 47 binary characters corresponding to biochemical reactions for each specimen (strain, isolate). The biological aspect reflected by the unsupervised classification approach is the taxonomic structure of the *Enterobacteriaceae* population. The phenotypic features of the bacterial strains measured in terms of the biochemical reactions tend to vary between different species and different ecological niches colonized by bacteria, which provides the basis for identifying distinct groups in the data.

For this data set, the algorithm of Corander et al. (2004) completely fails to produce a meaningful classification, due to the large class sizes of the underlying structure, which strongly affect the transition probabilities in (11). Thus, this example illustrates the deficiency of the ordinary reversible Metropolis-Hastings algorithm to handle transitions in the parameter space properly when the complexity of the modelled data increases.

Both runs of the interacting MCMC algorithm produced a classification consisting of 129 classes, where only some (<5) individual bacteria were differently allocated. Compared to the results of Gyllenberg et al. (1997), our classes represented generally subclasses of their classes. Such splits were often present with respect to the species boundaries, which makes microbiological sense. The classification result is visualized in Fig. 2, which clearly shows that the model and the learning algorithm are able to distinguish sensible patterns with high resolution even in a data set with highly complex structure. As a comparison, the data are also shown in Fig. 3 using a random permutation of the strains, which shows that no patterns are easily discernible in the data as such.

6 Discussion

In the present work we have considered a generic unsupervised classification problem in terms of a Bayesian predictive formulation. Despite of the biological nature of our real examples, the classification framework is valid for any application where the features are discrete-valued in the same fashion as here, and where the theoretical assumptions are considered as plausible approximations. It is worthwhile to notice that the Bayesian predictive approach is not based on the concept of a “true” statistical model, but simply intends to maximize the predictive ability to describe features observed in non-deterministic data.

Our parallel interacting MCMC algorithm is implemented in BAPS 2.1 software, available at <http://www.abo.fi/fak/mnf/mate/jc/software/baps.html>. Despite of the molecular biological setting where the earlier version of the software was introduced, it is possible to use BAPS 2.1 for unsupervised classification of generic items related with observations from multiple discrete alphabets.

The derivation of the partition-based unsupervised classification model through exchangeability provides a theoretical motivation to a naive Bayesian classifier, where features are assumed conditionally independent, as this corresponds to the product form of the predictive likelihoods (9). Therefore, the classification model investigated

Fig. 2 An image of the *Enterobacteriaceae* data (5313×47 binary matrix) sorted according to the posterior estimate (17). Every row represents the binary vector of observations for a single strain/isolate, such that red (green) color in a particular column corresponds to the value zero (one) in that element of the vector. The vectors are ordered with respect to the estimated classification, such that black horizontal lines indicate class boundaries and the classes are themselves ordered with respect to their size. The details in the lower part of the image are not well discernible due to the small sizes of the classes that are present there (color figure online)

here is basically subject to the same restrictions as a naive Bayes classifier, which, in fact, has been shown to perform surprisingly well even for dependent features, see e.g. the discussion in Hand and Yu (2001). It would also be possible to generalize the exchangeability assumptions, by introducing dependencies between the features inside the item classes in analogy with the supervised Bayesian classifiers based on graphical models, such as the ones described in Friedman et al. (1997) and Cerquides and De Mántaras (2005). However, such an approach induces an extreme increase in the computational complexity, as the learning of the graph structures is in general an NP-hard problem. When some suitable auxiliary prior information about the utilized features is available, it is possible to make use of less restrictive forms of exchangeability and introduce a Markovian dependence structure among the features inside the classes, as illustrated by Corander and Tang (2007) in a molecular biological context. Such an approach does then not require learning of the dependence structure itself, which results in computational complexity fairly similar to the currently considered models.

In the derivation of the unsupervised classification approach we have deliberately utilized well-known forms of reference priors, both for the class-specific parameters and for the partitions themselves. It is clear that the choices of priors will always have some impact on the inferences, in particular in a situation involving choices among

Fig. 3 An image of the *Enterobacteriaceae* data (5313×47 binary matrix) where the rows are randomly ordered. The color coding is equivalent to that used in Fig. 2 (color figure online)

multiple models of varying parametric dimensionality. However, in a general machine learning setting it is difficult to motivate any particular subjective choices of priors, unless some auxiliary information is available, on which the priors can be conditioned. Therefore, we consider the reference choices to be of practical value with respect to the applicability of the classification method. Also, it can be seen from the results of Gyllenberg et al. (1997), and Gyllenberg and Koski (2002), that the predictive likelihood approaches asymptotically (as the number of items in classes increases) the stochastic complexity of a classification under the reference priors.

The stochastic partition model has certain particularly attractive properties, both from the theoretical and practical perspective. For instance, compared to the traditional latent class model, the partition approach acknowledges directly the existing symmetry among classes of data, and does not thus lead to numerical obstacles provided by the artificial labels. Nevertheless, the numerical efficiency of the stochastic search considered here could still be improved to a considerable extent by designing more intelligent search operators instead of the proposals of random type. Such strategies were recently utilized by Corander et al. (2008) and Marttinen et al. (2006) to reduce the computational complexity of Bayesian stochastic learning for challenging models.

It is worth noticing that in the non-reversible MCMC algorithm it is possible to use any search operators that yield a stationary distribution for the corresponding Markov chain. Thus, the limitation of the standard reversible MCMC algorithms to proposal distributions for which the proposal probabilities are explicitly calculable, is avoided. However, in future we aim to investigate further generalizations of this

computational framework, by considering adaptive proposals leading to non-stationary Markov chains.

Acknowledgments The authors wish to thank two anonymous reviewers for the comments and suggestions that did lead to a significant improvement of the original manuscript. The work of JC was supported by the grant no. 121301 from Academy of Finland.

References

- Aarts EHL, Korst J (1989) Simulated annealing and Boltzmann machines. Wiley, Chichester
- Bernardo JM, Smith AFM (1994) Bayesian theory. Wiley, Chichester
- Bock H-H (1996) Probabilistic models in cluster analysis. *Comput Stat Data Anal* 23:5–28
- Cerquides J, De Mántaras RL (2005) TAN classifiers based on decomposable distributions. *Mach Learn* 59:323–354
- Corander J, Tang J (2007) Bayesian analysis of population structure based on linked molecular information. *Math Biosci* 205:19–31
- Corander J, Ekdahl M, Koski T (2008) Parallel interacting MCMC for learning of topologies of graphical models. *Data Mining Knowl Discovery* (in press)
- Corander J, Gyllenberg M, Koski T (2006) Bayesian model learning based on a parallel MCMC strategy. *Stat Comput* 16:355–362
- Corander J, Gyllenberg M, Koski T (2007) Random partition models and exchangeability for Bayesian identification of population structure. *Bull Math Biol* 69:797–815
- Corander J, Waldmann P, Marttinen P, Sillanpää MJ (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* 20:2363–2369
- Dawson KJ, Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet Res Camb* 78:59–77
- Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley, New York
- Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. *Mach Learn* 29:131–163
- Gidas B (1985) Nonstationary Markov chains and convergence of the annealing algorithm. *J Stat Phys* 39:73–130
- Green P (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82:711–732
- Gyllenberg HG, Gyllenberg M, Koski T, Lund T, Schindler J, Verlaan J (1997) Classification of *Enterobacteriaceae* by minimization of stochastic complexity. *Microbiology* 143:721–732
- Gyllenberg M, Koski T, Verlaan M (1997) Classification of binary vectors by stochastic complexity. *J Multivar Anal* 63:47–72
- Gyllenberg M, Koski T (2001) Probabilistic models for bacterial taxonomy. *Int Stat Rev* 69:249–276
- Gyllenberg M, Koski T (2002) Bayesian predictiveness, exchangeability and sufficientness in bacterial taxonomy. *Math Biosci* 177, 178:161–184
- Hand DJ, Yu K (2001) Idiot's Bayes: not so stupid after all. *Int Stat Rev* 69:385–398
- Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2:193–218
- Häggström O (2002) Finite Markov chains and algorithmic applications. Cambridge University Press, Cambridge
- Isaacson DL, Madsen RW (1976) Markov chains: theory and applications. Wiley, New York
- Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic Press, London
- Marttinen P, Corander J, Törönen P, Holm L (2006) Bayesian search of functionally divergent protein subgroups and their function specific residues. *Bioinformatics* 22:2466–2474
- Perks W. (1947) Some observations on inverse probability including a new indifference rule. *J Inst Actuaries* 73:285–334
- Rand WM (1971) Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc* 66:846–850
- Rissanen J (1987) Stochastic complexity. *J Roy Stat Soc B* 49:223–239
- Rissanen J (1995) Fisher information and stochastic complexity. *IEEE Trans Inf Theory* 42:40–47
- Robert CP, Casella G (1999) Monte Carlo statistical methods. Springer, New York
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldmann MW (2002) Genetic structure of human populations. *Science* 298:2381–2385

- Schervish MJ (1995) *Theory of statistics*. Springer, New York
- Sisson SA (2005) Transdimensional Markov chains: a decade of progress and future perspectives. *J Am Stat Assoc* 100:1077–1089
- Van Cutsem B (1996) Combinatorial structures and structures for classification. *Comput Stat Data Anal* 23:165–188
- Zabell SL (1982) W.E. Johnson's 'Sufficientness' postulate. *Ann Stat* 10:1091–1099