DISCOVERING HIDDEN STRUCTURES IN MOLECULAR DATA USING A BAYESIAN PARTITION MODEL APPROACH

Pekka Marttinen

Academic dissertation

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium XIV, the Main Building (Fabianinkatu 33), on September 26th, 2008, at 12 noon.

Department of Mathematics and Statistics
Faculty of Science
University of Helsinki

HELSINKI 2008

Supervisor Prof. Jukka Corander
Department of Mathematics
Åbo Akademi University
Finland

Reviewers Prof. Esa Läärä
Department of Mathematical Sciences
University of Oulu
Finland

Doc. Aki Vehtari
Department of Biomedical Engineering and
Computational Science
Helsinki University of Technology
Finland

Opponent Prof. Carlo Berzuini
Medical Research Council Biostatistics Unit
Institute of Public Health
Cambridge, UK

ISBN 978-952-92-4359-4 (Paperback)
ISBN 978-952-10-4922-4 (PDF)
http://ethesis.helsinki.fi
Yliopistopaino
Helsinki 2008

List of articles included

This thesis is based on the following five original articles which are referred to in the text by Roman numerals:

I J. Corander and P. Marttinen. Bayesian identification of admixture events using multi-locus molecular markers. *Molecular Ecology*, 15:2833–2843, 2006.

II J. Corander, P. Marttinen, and S. Mäntyniemi. Bayesian identification of stock mixtures from molecular marker data. *Fishery Bulletin*, 104:550–558, 2006.

III P. Marttinen, J. Corander, P. Törönen, and L. Holm. Bayesian search of functionally divergent protein subgroups and their function specific residues. *Bioinformatics*, 22:2466–2474, 2006.

IV P. Marttinen, J. Tang, B. De Baets, P. Dawyndt, and J. Corander. Bayesian clustering of fuzzy feature vectors using a quasi-likelihood approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008. http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.53

V P. Marttinen, A. Baldwin, W. P. Hanage, C. Dowson, E. Mahenthiralingam, and J. Corander. Bayesian modeling of recombination events in bacterial populations. 2008. Submitted.

Author's contributions in articles I-V

I,II PM and JC had jointly the main responsibility in designing the models and algorithms used in these articles, PM had the main responsibility in implementation and testing, and PM took part in writing the articles while JC had the main responsibility in writing the articles.

III The modeling approach was developed jointly by all authors. PM had the main responsibility in implementation, PM and PT carried out the testing of the methods, and PM had the main responsibility in writing the article.

IV,V PM contributed the main part in all aspects of the articles.

Contents

1 Introduction

One of the central characteristics of contemporary biosciences is the accumulation of data into rapidly growing databases. Such data often are in the form of vectors of observations made from the items under investigation. Our main focus is on data reflecting the molecular level features, examples including genetic data, either in the form of genetic markers or DNA sequences, and amino acid sequence data from proteins. For recently published examples of such data, see e.g. the comprehensive genetic marker data of human populations by Jakobsson et al. (2008), or the protein database maintained by The Uniprot Consortium (2008). Because these kinds of data are immediately linked to the genetic composition of individuals present in data, they provide a firm stepping stone to answering questions related to the evolutionary histories of the individuals. A fundamental task is to identify groups of individuals in a data set, who are related from an evolutionary perspective, and to quantify the extent and the nature of these relationships. In this work, we consider these kinds of questions, and our goal is in developing novel statistical methods which would further promote the understanding of the investigated phenomena.

The Bayesian approach (see e.g. Bernardo and Smith, 1994) has recently been recognized as a suitable framework in bioinformatics and related fields (e.g. Beaumont and Rannala, 2004; Wilkinson, 2007) for tackling questions about the statistical properties of data. The flexibility of the Bayesian approach stems from ability to combine information from several sources within a single statistical model, while being able to coherently take the uncertain nature of such information into account. The sources of information comprise typically expert knowledge about the phenomenon under investigation and the actual observed data. Model development within the Bayesian approach is initiated by encoding the available expert knowledge into prior probability distributions for models and the related parameters. The prior distributions are updated by the information embodied in the observed data into posterior distributions through the use of the Bayes' formula. Although the essential elements of the Bayesian approach have been formulated decades ago, widespread utilization of the Bayesian approach has been restricted by difficulties related to the computation. Advances in computational techniques (e.g. Gilks et al., 1996) as well as the increased efficiency of computers since the early 1990's have been followed by a surge of applications successfully solved by the Bayesian methods. Still, Bayesian computation may require fairly extensive computational power especially in situations where learning of high-dimensional models for large-scale data sets is required.

The task of identifying evolutionary relationships among sampled individuals is closely related to the task of classification, in which the individuals are divided into classes (or clusters) corresponding to divergent groups present in data. Depending on the amount of available prior information about the contents of the classes, the classification task can be considered either as unsupervised (no information) or supervised (training data available for all possible classes), with semi-supervised classification falling in between these two extremes. Further-

more, in unsupervised classification, the number of clusters may be unknown *a priori*. The articles I-V included in this work contain examples of all these types of classification tasks. A traditional way of addressing a classification task from the Bayesian perspective is to utilize latent class mixture models (Duda, 2001), in which the individuals are given labels specifying the classes to which they belong. We adopt a somewhat different approach where the structure of data is specified by a partition, i.e. a set of non-empty classes such that each individual belongs to exactly one of the classes. A distinction to the latent class model approach is that in the partition model approach the classes are completely unlabelled and defined only in terms of the individuals they contain. Consequently, the partition model formulation avoids the label-switching problem related to latent class models, where difficulties in simulation-based Bayesian inference may arise due to permuted class labels. More importantly, the partition model approach enables the utilization of efficient computational strategies, which is of utmost importance in unsupervised classification of large-scale data sets. In the articles I-V, generalizations and modifications of the basic partition model (see Section 2) are used to obtain answers to various biologically relevant questions about the structure of the considered data sets.

This summary part is structured as follows. In the two subsections of Section 2, the Bayesian approach for unsupervised classification (i.e. clustering) utilizing a partition model is formally defined, using the normative framework for Bayesian analysis discussed in Bernardo and Smith (1994). The first subsection formulates the clustering problem in terms of a Bayesian model learning framework, and discusses the specification of prior beliefs for the set of possible partitions. The predictive distribution for a partition model is derived in the second subsection. The derivation is done by using the basic concepts of Bayesian analysis (exchangeability, for example), and the goal is to characterize the fundamental assumptions concerning the data, from which the used form of the predictive distribution follows. The derivation is shown only for vectors of binary observations, which is a special case of the more general model derived in Corander et al. (2007). This simplifying constraint is adopted because it already enables the illustration of the most essential underlying assumptions while maintaining the notational clarity. The development in Section 2 (especially the second subsection) is more rigorous than what has been possible, or even appropriate, when deriving the models in the attached articles I-V. The more rigorous treatment is given here in order to provide a better understanding about the assumptions inherent in the models utilized in the attached articles. Section 3 discusses the computational strategies adopted for solving the statistical learning tasks. Especially, we review the used stochastic search strategies, and compare them with more standard Bayesian computation based on Markov chain Monte Carlo. Section 3 also includes results from a novel simulation experiment, which demonstrate the benefits obtained through our approach. Section 4 provides brief summaries of articles I-V included in this thesis.

2 Partition model for structured data

Bayesian model learning for unsupervised classification

In this subsection, we derive a Bayesian framework for the identification of the structure of a data set $\mathbf{x}$ consisting of N data items

$$\mathbf{x} = \{\mathbf{x}^{(r)}\}_{r=1}^{N}.$$

By structure we refer to a collection of subsets of the items, such that the data for each subset can adequately be modeled using a single family of distributions labeled by a finite-dimensional parameter. As the focus of our work is on modeling molecular data, the above data structure will in the sequel be referred to as *population structure*, with $\mathbf{x}$ representing a sample from some underlying population of interest. Here, we do not specify further the actual form of the data items $\mathbf{x}^{(r)}$. In the next subsection, we will consider a special case where each $\mathbf{x}^{(r)}$ is a vector of binary attributes.

The population structure for the data items will in the sequel be represented by a partition $S = \{s_1, \ldots, s_k\}$, where s_i are disjoint, nonempty subsets of $\{1, \ldots, N\}$, whose union is $\{1, \ldots, N\}$. To proceed with the Bayesian framework for model-selection, we need to specify a prior predictive distribution $p(\mathbf{x}|S)$, which represents the probability of data conditional on any specific model S. We also need to specify a prior distribution $p(S)$ which represents our beliefs about the plausibilities of different models before observing the data. Together, the predictive distributions $p(\mathbf{x}|S)$ and the prior distribution $p(S)$ define an overall belief model for the data:

$$p(\mathbf{x}) = \sum_{S \in \mathcal{S}} p(\mathbf{x}|S)p(S),$$

where $\mathcal{S}$ is the set of all possible partitions for the data items. The posterior distribution of S, given the observed data $\mathbf{x}$, is then obtained through Bayes' rule:

$$p(S|\mathbf{x}) = \frac{p(\mathbf{x}|S)p(S)}{\sum_{S \in \mathcal{S}} p(\mathbf{x}|S)p(S)}. \tag{1}$$

The derivation of the predictive distributions $p(\mathbf{x}|S)$ in the special case of binary data vectors will be considered in the next subsection. Here, we will concentrate on the possible forms of prior distributions $p(S)$ for the partitions.

In the works included in this thesis (I, II, III, IV, V), the most basic form of prior representation for the partitions is given by

$$p(S) = C * I\left(|S| \leq K\right), \tag{2}$$

where C is a constant, and $I\left(|S| \leq K\right)$ is an indicator function which equals unity, if the partition S has at most K clusters, and zero otherwise. K denotes here a user-specified upper bound for the number of clusters, considered to be suitable for the problem at hand. Thus, (2) assigns equal probability to all

partitions which have at most K clusters. The prior (2) is a slightly constrained version of a more general prior

$$p(S) = C, \tag{3}$$

which assigns equal probability to all possible partitions S. A generating process for partitions with the distribution (3) has been described in Stam (1983), and consists of throwing N balls one by one into u urns randomly, such that each urn has equal probability. If the number of urns u is itself assigned a distribution

$$p(u) = |\mathcal{S}|^{-1} e^{-1} \frac{u^N}{u!}, \ u = 1, 2, \ldots,$$

the contents of the non-empty clusters will define a partition, whose distribution is given by (3). Thus, partitions with distribution (2) can be generated by utilizing a similar urn construction as presented above, but by keeping only those models which have at most K clusters.

In our work, the upper bound K in (2) was originally introduced for implementational purposes, and in practice it is specified to be large enough to avoid any constraints which would hide the underlying data structure, unless it is especially of interest to consider only some particular values for the number of clusters, as e.g. in Corander et al. (2008b). The prior (2) can be seen as an agreeable representation of ignorance about the correct structure. Indeed, under this prior, the posterior distribution (1) simplifies to

$$p(S|\mathbf{x}) = \frac{p(\mathbf{x}|S)}{\sum_{S \in \mathcal{S}, |S| \leq K} p(\mathbf{x}|S)},$$

from which it can be seen that the posterior inferences are determined by $p(\mathbf{x}|S)$, the abilities of the different models to describe data. This is a favorable characteristic especially when considering some simple hypotheses concerning the structure of a model S. For example, we may consider the origin of some specific data item r by keeping the population structure otherwise fixed, and calculate the probabilities for the clusters as origins for the item r. By using the prior (2), all the putative clusters will be considered *a priori* equally likely as origins for the data item, and the resulting distribution will be determined by how well the observations in the different clusters are compatible with the observations for the data item r. On the other hand, although the prior (2) is seemingly uninformative, it in fact embodies fairly strong information about the number of clusters. This is caused by the fact that the number of possible partitions with a specific number k of clusters, given by the Stirling number of the second kind (see e.g. Abramowitz and Stegun, 1965, p. 824), depends strongly on k. For example, for a data set of nine items, there are 7770 ways to partition the data into four subsets, and only one possible partition consisting of one subset. Therefore, because all models are considered equally likely, four clusters would in this case be considered 7770 times more likely *a priori* than one cluster. Although the influence of a prior diminishes as the amount of information in $\mathbf{x}^{(r)}$ increases, the use of the prior (2) as such may consequently

lead to solutions with unreasonably many small clusters when the data is only sparsely informative, as illustrated in I and Corander et al. (2008a).

There are several ways to improve the basic version (2) of the prior distribution when only sparsely informative data are available. Dirichlet process mixture models (Antoniak, 1974) provide a completely unrelated alternative which essentially differs from the above-derived framework by specifying a different prior distribution $p(S)$ for the partitions. In this prior, the probability of assigning a data item r into a cluster depends linearly on the number of items already assigned to the cluster, with also a small probability of introducing a novel cluster. This probability further depends on a user-specified model parameter. Counts of clusters of different sizes observed in a partition S constitute then a sufficient statistic, for which a distribution in a closed form is given by the Ewens sampling formula (Ewens, 1972; Antoniak 1974). Such a prior avoids the problem of introducing many small clusters, because the *a priori* expected number of clusters will depend logaritmically on the number of items in the data set (Antoniak, 1974). On the other hand, for the comparison of simple hypotheses, for example which of two possible partitions is more likely, the prior based on Dirichlet processes may be less sensible if the two partitions have different numbers of variable-sized clusters. In such a situation one of the partitions may be strongly preferred *a priori*, while the uniform prior (2) would give equal prior probabilities to the alternatives. Thus, the problem with the uniform prior (2) becomes reversed with the Dirichlet process prior. While a uniform prior in the model space (2) brings about strong information about the number of clusters, the more sensible distribution for the number of clusters obtained by Dirichlet processes effectuates strong preferences within the model space. Both these phenomena emanate from the discrepancy in the number of ways to partition a data set into different numbers of clusters.

In some of our works, we have improved the basic prior (2) by utilizing additional information concerning the geographical sampling locations of the data items. In particular, we may utilize a prior which is uniform over all possible partitions which are compatible with the sampling information, by assigning non-zero probabilities only to partitions in which the data items sampled from the same location belong to the same cluster. Such priors can be justified by biological knowledge of the behavior of the species under investigation, which may state for example that individuals belonging to different demographic units (family, species, etc.) are very unlikely to be collected at the same location. In I and II we show that such priors considerably increase the power to correctly detect clusters, and, from the practical point of view, remove the undesired behavior of the prior (2) when the data are only sparsely informative. A more detailed utilization of spatial information taking into account the actual coordinates of the sampling locations has been considered, e.g., in Corander et al. (2008a).

Predictive distribution of a partition model

In this subsection, our goal is to derive one particular form of the predictive model for the data $\mathbf{x}$ implied by the partition model S. Throughout the subsection, we only consider vectors of binary observations, and the presented results should be considered as special cases in this respect, although we do not state this explicitly in the sequel. Thus, the items in the data set can be expressed as

$$\mathbf{x}^{(r)} = \left(x_1^{(r)}, \ldots, x_J^{(r)}\right) \in \{0,1\}^J, \text{ for all } r = 1, \ldots, N.$$

If the value of an observation is equal to one, we will refer to it as a success. Otherwise, the observation will be termed a failure. A predictive model is a probability measure P which specifies the form of the joint distribution for the data $\mathbf{x}$. We will show that such a probabilistic description for the data can be obtained by utilizing fairly simplistic assumptions concerning the data. The following definition is central to our derivation:

Definition 1 *(Exchangeability). Random quantities $q_1, q_2, \ldots, q_w$ are said to be (finitely) exchangeable under a probability measure P when the joint belief distribution satisfies*

$$p(q_1, \ldots, q_w) = p\left(q_{\pi(1)}, \ldots, q_{\pi(w)}\right)$$

for any permutation π defined on the set $\{1, \ldots, w\}$. Furthermore, the sequence $(q_w)_{w=1}^{\infty}$ is said to be infinitely exchangeable if every finite subsequence is (finitely) exchangeable.

The definition of exchangeability captures the notion of symmetry among the observations. In particular, it implies that the ordering of the observations will not affect our inferences. Exchangeability in its various forms, as well as the representation theorems described below, is considered in detail in Bernardo and Smith (1994, Chapter 4).

Given the partition S, our data set $\mathbf{x}$ can be divided into groups according to the cluster label and the observed features. Let

$$\mathbf{x}_{ij} = \{x_j^{(r)} |\ r \in s_i\}$$

denote the observations for feature j in cluster s_i, $i = 1, \ldots, k$, $j = 1, \ldots, J$. If no further information about the data items is available, a plausible assumption is that the joint probability of the set $\mathbf{x}_{ij}$ is the same regardless of the order of observations in $\mathbf{x}_{ij}$. Formally, the observations in $\mathbf{x}_{ij}$ are considered as a part of an infinitely exchangeable sequence. A well-known representation theorem by de Finetti (1930) states that this determines the form of the predictive distribution for such observations. By using the following notation:

$$n_{ij1} = \sum_{r \in s_i} x_j^{(r)},$$

representing the number of successes among $\mathbf{x}_{ij}$, and

$$n_{ij0} = \mid s_i \mid - n_{ij1},$$

representing the number of failures among $\mathbf{x}_{ij}$, the predictive distribution can be written as

$$p(\mathbf{x}_{ij}) = \int_0^1 \theta_{ij}^{n_{ij1}} (1-\theta_{ij})^{n_{ij0}} dQ(\theta_{ij}), \tag{4}$$

where θ_{ij} is a random quantity, which can be interpreted as a limiting frequency of successes in the infinitely exchangeable sequence from which $\mathbf{x}_{ij}$ were obtained, and Q is a probability measure describing our beliefs about θ_{ij}. The form of (4) means that, assuming exchangeability, we can proceed as if the observations in $\mathbf{x}_{ij}$ are conditionally independent Bernoulli random quantities with *parameter* θ_{ij}. When the corresponding conditional probability $p(\mathbf{x}_{ij}|\theta_{ij})$ of the observed data $\mathbf{x}_{ij}$ is considered as a function of θ_{ij}, the resulting function is referred to in the conventional terminology as the *likelihood* function. Also, the distribution Q is referred to as the *prior* distribution of θ_{ij}.

Whereas the observations in a single cluster for feature j, $\mathbf{x}_{ij}$, can be plausibly considered exchangeable, there is in general no reason to treat symmetrically the observations with respect to different clusters or features. To derive the predictive model for such a situation, we will need to consider a set of infinitely exchangeable sequences. The following definition provides the required means:

Definition 2 *(Unrestricted exchangeability). Sequences of random quantities,* $\mathbf{q}_w = (q_{w1}, q_{w2}, \ldots)$, $w = 1, \ldots, m$, *are said to be unrestrictedly exchangeable if each sequence is infinitely exchangeable and, in addition, for all* $n_w \leq N_w$, $w = 1, \ldots, m$,

$$p(\mathbf{q}_1(n_1), \ldots, \mathbf{q}_m(n_m)|y_1(N_1), \ldots, y_m(N_m)) = \prod_{w=1}^{m} p(\mathbf{q}_w(n_w)|y_w(N_w)),$$

where $\mathbf{q}_w(n_w) = (q_{w1}, \ldots, q_{wn_w})$, *the vector of* n_w *first items in sequence* $\mathbf{q}_w$, *and* $y_w(N_w) = q_{w1} + q_{w2} + \ldots + q_{wN_w}$, $w = 1, \ldots, m$, *the sum of the entries in* $\mathbf{q}_w(N_w)$.

The definition captures the idea that, given the total number of successes in the first N_w observations from the wth sequence, only that total is relevant when predicting the outcome of any subset of size n_w of the N_w first observations. Furthermore, this total can not be used to predict the outcomes of $\mathbf{q}_{w'}(n_{w'})$ in any other sequence if we know the corresponding value of $y_{w'}(N_{w'})$. Notice, however, that in the absence of knowledge of $y_{w'}(N_{w'})$, the number of successes from the first sequence *might* be considered relevant for assessing the probabilities for the second sequence.

By assuming that the observations $\mathbf{x}_{ij}$ for different clusters and features, $i = 1, \ldots, k$, $j = 1, \ldots, J$, are unrestrictedly exchangeable, a generalization of the

basic representation theorem (de Finetti, 1938) yields the following predictive representation for the data:

$$p(\mathbf{x}) = \int_{[0,1]^m} \prod_{i=1}^{k} \prod_{j=1}^{J} \theta_{ij}^{n_{ij1}} (1-\theta_{ij})^{n_{ij0}} dQ(\theta), \tag{5}$$

where θ represents jointly all θ_{ij} and Q is a joint probability distribution for θ. Recall that θ_{ij} can be interpreted as the limiting relative frequency of successes in the sequence from which $\mathbf{x}_{ij}$ were obtained. We will adopt here the simplest kind of joint distribution for θ_{ij} by setting

$$dQ(\theta) = \prod_{i=1}^{k} \prod_{j=1}^{J} dQ(\theta_{ij}), \tag{6}$$

according to which the knowledge of the limiting frequency in a sequence corresponding to some cluster and feature does not change our beliefs about outcomes in any other sequence. Substituting (6) into (5) yields

$$p(\mathbf{x}) = \prod_{i=1}^{k} \prod_{j=1}^{J} \int_{[0,1]} \theta_{ij}^{n_{ij1}} (1-\theta_{ij})^{n_{ij0}} dQ(\theta_{ij}). \tag{7}$$

The exact form of the prior distribution $dQ(\theta_{ij})$ must be specified according to the problem at hand. We proceed by assuming that the probability of another success from the sequence from which $\mathbf{x}_{ij}$ have already been observed depends linearly on the number of observed successes n_{ij1} among $\mathbf{x}_{ij}$. It follows from this assumption (Zabell, 1982, Corollary 2.2) that (7) has the form

$$p(\mathbf{x}) = \prod_{i=1}^{k} \prod_{j=1}^{J} \frac{\Gamma\left(\sum_{l=0}^{1} \alpha_{ijl}\right)}{\Gamma\left(\sum_{l=0}^{1} (\alpha_{ijl} + n_{ijl})\right)} \prod_{l=0}^{1} \frac{\Gamma\left(\alpha_{ijl} + n_{ijl}\right)}{\Gamma(\alpha_{ijl})} \tag{8}$$

for some constants $\alpha_{ijl},\ i = 1, \ldots, k,\ j = 1, \ldots, J,\ l = 0, 1$. When all $\alpha_{ijl} > 0$, this coincides with the analytical form of the integral (7) arising from the specification

$$\theta_{ij} \sim Beta(\alpha_{ij1}, \alpha_{ij0}) \tag{9}$$

of the prior beliefs about $\theta_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, J$. The family of $Beta$ distributions (9) includes prominent choices for noninformative, or reference, prior beliefs about θ_{ij} (see e.g. Bernardo and Smith, 1994, Chapter 5.4). Such prior distributions can be used to express ignorance about the value of the parameter, or to claim objectivity, as may be aspired for example when making inferences of a scientific hypothesis. Example reference choices include the uniform prior density, $Beta(1, 1)$, following from the Bayes-Laplace postulate, according to which all outcomes should be considered equally probable (Bayes, 1763; Laplace, 1814), and Jeffreys prior, $Beta(1/2, 1/2)$, which one may regard as the first choice due to the property that the prior beliefs derived using Jeffreys rule are invariant under data transformations (Jeffreys, 1961; see also Perks, 1947).

3 Computational strategies

In Bayesian analysis, the information about a variable of interest is captured by the corresponding posterior distribution. A typical goal of Bayesian computation is to provide an estimate of this distribution. The estimated distribution allows one to compare relative plausibilities of various hypotheses concerning the data, and to summarize the results in terms of the mean or mode of the distribution, for example. In the current setting we are primarily interested in the posterior distribution of partition S. In general, to estimate a posterior distribution, one may need to evaluate integrals which are intractable in practice. However, even in cases where the statistical characteristics of the posterior distribution can not be obtained directly, it is usually possible to simulate samples from the distribution. Such samples provide an indirect way of approximating the distribution. Various methods have been developed for obtaining samples from the posterior distribution, e.g. direct sampling, importance sampling, rejection sampling, or Markov chain Monte Carlo sampling (see e.g. Gelman et al., 2004). Alternatively, it is sometimes possible to use numerical integration. Markov chain Monte Carlo (MCMC) methods, e.g. Gilks et al. (1996), Sisson (2005), are perhaps the most commonly used techniques for sampling from posterior distributions. The idea in MCMC is to generate a stochastic process whose stationary distribution is the posterior distribution from which one wishes to sample realizations. After an initial, so-called burn-in period, the values simulated from such a process can be considered as samples from the stationary distribution, i.e. the posterior distribution. If the space of possible states for the chain is discrete and finite, such as the space of all possible partitions $\mathcal{S}$, the probability of a single state S can be estimated as the number of time points in which the chain occupies the state, divided by the total number of observations simulated from the chain, i.e.

$$P(S|\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} I(S_t = S), \tag{10}$$

where T is the total number of simulated observations, S_t is the state of the chain at time point t and I is an indicator function, which equals unity if and only if S_t and S are the same.

The first version of our BAPS program, Corander et al. (2003), utilized a standard Metropolis-Hastings MCMC algorithm, which used random merge and split operators to traverse the partition space. In the second version, Corander et al. (2004), the algorithm was enhanced by simulating several parallel MCMC chains, and introducing new search operators which allowed moving single individuals between the clusters. However, an even more important innovation was that the probabilities of different partitions S were estimated from

$$P(S|\mathbf{x}) = \frac{P(\mathbf{x}|S)}{\sum_{S\in\mathcal{S}_T} P(\mathbf{x}|S)}, \tag{11}$$

where $\mathcal{S}_T$ is the set of all partitions visited during the first T iterations of the MCMC. Thus, as opposed to the previous version, the probabilities were

not based on the frequencies of visiting various partitions, as in (10), but on the relative marginal likelihoods which in our approach were available in analytical form for different models, as illustrated in the previous section in the special case of binary observations. Such an approach has several advantages over the frequency approach. Firstly, it is straightforward to combine the results from different chains. Secondly, the estimation is not impaired even if some chain gets stuck to a state of low predictive value, in which case the state is consequently assigned unreasonably high probability by the frequency-based approach. Thirdly, the empirical distribution calculated by (11) converges much faster than the one calculated from the relative frequencies (10), as the probabilities stabilize already when the models with high marginal likelihoods are visited just *once*. Finally, if (11) is used, it is not necessary to construct a Markov chain with a stationary distribution equal to the posterior distribution; rather, it is sufficient to use basically any available search method to find the most plausible models which are the most important for evaluating (11). This last notion was utilized in the third version (and subsequently) of the BAPS program (I), by employing a greedy stochastic optimization algorithm to find an estimate of S. To estimate the posterior distribution with such a strategy, the search algorithm is repeatedly executed several times, and a list of the models with the highest marginal likelihood values visited during the runs is maintained. Our experiments with both artificial and real data suggest that this strategy is both faster, and more likely to find the models which really have impact on the estimate of the posterior distribution. Analogous ideas utilizing stochastic optimization and model averaging have been used in the context of selection of variables in multivariate regression by Brown et al. (1999, 2002).

As opposed to a standard Metropolis-Hastings MCMC algorithm, which is based on random movement proposals in the search space, we can in the greedy optimization strategy suggest intelligent moves, which are more likely to lead to an improvement in the partition. This feature is especially crucial in splitting overly-sized clusters. Such clusters often correspond to a local optimum in the model space, the escaping from which is unlikely by moves of single individuals. Also, a random split, as in standard Metropolis-Hastings, would seldom be accepted. On the contrary, by using the distances calculated between the members of the cluster, and utilizing a fast clustering algorithm, we are able to suggest splits which have a decisively higher probability of improving the model. Use of similar intelligent search operators embedded in a non-reversible Metropolis-Hastings algorithm is discussed in Corander et al. (2006). Tu and Zhu (2002) give an example of utilizing intelligent (data-driven) steps within a reversible MCMC in the context of image segmentation. In their approach the image is partitioned at different scales prior to the actual MCMC algorithm, and the found partitions are used in the formulation of the proposal distributions for splitting a segment in a Metropolis-Hastings step.

In general, a greedy optimization as a means for Bayesian model learning can be criticized in two ways. First, it can be argued that there is no guarantee that such an algorithm would find the true global optimum, while the MCMC is theoretically guaranteed to find the optimal model when the length

of the simulated Markov chain approaches infinity. In a practical situation, in our opinion, the value of such a theoretical guarantee is rather negligible, since there is no guarantee that the convergence would happen in any reasonable time. The questionable behavior of standard MCMC especially when applied to large-scale model-learning problems has been documented in Jones et al. (2005), Laskey and Myers (2003), and Corander et al. (2006). Nevertheless, the stochastic greedy optimization can be, and should be, started several times from different initial assignments, in order to avoid the local optima, which may be reached in any single run. The second piece critical aspect related to the greedy optimization strategy is that MCMC can provide a better statistical characterization of the whole distribution, while a simple optimization algorithm gives only the mode of the distribution. However, by using (11) to combine the results from several runs of the optimization algorithm, a more honest representation of the posterior distribution is attainable. In I, II, and III we also employ Bayes factors (Kass and Raftery, 1995) in addition to the posterior distribution, and in V the marginal distributions of different variables, in order to characterize the uncertainty locally around the mode estimate of the distribution.

Empirical illustration

In this subsection, we utilize the standard Gibbs sampling algorithm and a version of the greedy stochastic search algorithm to cluster a group of artificial data sets, in order to compare the performances of the two algorithms. The Gibbs sampling algorithm can be described as follows. Let the state of the Markov chain consist of c_r, $r = 1, \ldots, N$, the cluster labels for the data items, i.e. $c_r \in \{1, \ldots, K\}$ for all r, where K is the specified upper bound for the number of clusters. The cluster labels have no meaning beyond specifying which data items belong to the same cluster. Let c_{-r} denote the cluster labels of other data items except r. The conditional distribution for the cluster label c_r can be written as

$$P(c_r = c|c_{-r}, \mathbf{x}) \propto P(\mathbf{x}|c_r = c, c_{-r}) * P(c_r = c|c_{-r}), \tag{12}$$

where $P(\mathbf{x}|c_r = c, c_{-r})$ is the marginal likelihood of the data (8) conditional on the partition specified by the cluster labels, and $P(c_r = c|c_{-r})$ is the prior distribution for c_r, conditional on the cluster labels of the other data items. Because all partitions are here assumed *a priori* equally likely, it follows that

$$P(c_r = c|c_{-r}) = D, \tag{13}$$

where D is a constant. Thus, the probabilities are the same for all clusters which are represented in c_{-r}, and, in addition, one arbitrarily chosen empty cluster, while the probability of the rest of the empty clusters is zero (notice that assigning the data item to any of the empty clusters yields the same partition). Now, the Gibbs sampling algorithm simply consists of iterations, each of which updates in a random order the cluster labels for the data items, drawing the new values from the conditional distributions (12). Notice that the specified

algorithm can be considered as a standard algorithm for sampling from Dirichlet process mixture models (Neal, 1998, Algorithm 3), apart from the different prior specification for the cluster labels.

For the stochastic greedy optimization we use a basic version of the algorithm whose extensions have been used e.g. in I, II, III, and IV. In summary, the algorithm uses three types of operators to search the partition space. The first operator (Move) is analogous to one Gibbs sampling iteration, except that the data items are always moved to the cluster which yields the highest marginal likelihood improvement (can be zero in which case the item is not moved). The second operator (Merge) goes through all pairs of clusters, and merges the pair of clusters which leads to the highest increase in the marginal likelihood (if any pair yields an increase). The third operator (Split) uses the complete linkage algorithm (see e.g. Ripley, 1996) to identify putative subclusters, and splits a cluster according to the subdivision leading to the highest marginal likelihood value (if any splits are associated with higher values).

To illustrate how the presented stochastic greedy optimization and the Gibbs sampling MCMC strategies behave in practice, we employ a series of artificial data sets. We analyze each such data with both the described algorithms. The analyses are repeated twice, by starting both the algorithms from two different initial setups. In the first initial setup all data items are assigned into a single cluster. For the second setup, the complete linkage clustering algorithm is utilized to create $k_0 + 10$ initial clusters, where k_0 is the correct underlying number of clusters. The number of iterations in the MCMC is specified to be 1,000. We consider data sets of seven different degrees of complexity. The simplest data sets consist of 15 clusters, whereafter the number of clusters is increased up to 45 clusters using an equal spacing of 5 clusters. For each data set size, we generate five different data sets. The numbers of the data items are drawn independently for the different clusters from a uniform distribution over the set $\{5, 6, \ldots, 20\}$. Each data item is characterized by a binary vector $x^{(r)} \in \{0,1\}^J$, where we use value $J = 35$ for the length of the vector. We draw the frequency parameters θ_{ij} for different clusters i and features j independently from

$$\theta_{ij} \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right).$$

The value $J = 35$ was selected because in the preliminary experiments (the exact results not shown) we found it to be suitable for the comparison. With higher values (e.g. >50) both the algorithms were able to find the exactly correct partition in a short time, while with lower values (e.g. <20) the data was not informative enough for a proper learning of the partition, and consequently both the algorithms, especially the MCMC, ended up far away from the optimal model.

For reporting the results, we record for each run of the greedy algorithm the running time and the marginal likelihood of the model with the highest value. Because the total running time of the MCMC algorithm is largely determined by the pre-specified number of iterations, it provides an unsatisfactory basis for

the comparison of the efficiencies of the algorithms. A more reasonable estimate of the running time of the MCMC algorithm is obtained by recording the time to reach a model with an equal or higher marginal likelihood than the highest value found by the greedy algorithm. If the MCMC never reaches a model with such a value during the 1000 iterations, we record the time needed to reach the model with the highest value in that run. The average running times of the different algorithms are shown in Figure 1. To compare the goodness of the solutions found by the alternative algorithms, we select for both algorithms the model with the maximum marginal likelihood found in the two separate runs. We then use Bayes factors to compare these top-scoring models. The results of the comparison are shown in Figure 2.

As can be seen from Figure 1, the running time of the MCMC algorithm increases rapidly as the complexity of the data set increases. In contrast, the greedy algorithm is considerably faster for all the data set sizes. For example, the execution times are approximately 60 times longer for the MCMC algorithm started from $k_0 + 10$ as compared to the greedy search, when the most complex data sets are considered. A first look to Figure 1 would surprisingly suggest that the MCMC started from a single cluster is performing faster than the MCMC initialized with $k_0 + 10$ clusters. A closer inspection (exact results not given here) shows, however, that this is not the case. Recall that the time shown in Figure 1 is the time required to reach a model with an equal value to that found by the greedy algorithm or, if no such model is found, a time required to reach the model with the highest marginal likelihood value in that run. As can be seen from Figure 2, neither of the MCMC runs found a model with equal predictive value to that found by the greedy algorithm for the data sets with $k_0 = 45$. Moreover, for these data sets, the MCMC started with a single cluster never found a model of equal or higher value than the MCMC started with $k_0 + 10$ clusters. The surprisingly low running times (with $k_0 = 45$) in Figure 1 for the MCMC started with one cluster are therefore explained by the fact the algorithm got faster stuck to a local mode in the model space than the MCMC started with $k_0 + 10$ initial clusters.

These analyses serve to illustrate why we have chosen the stochastic greedy strategy. As the complexity of data increases, the greedy algorithm is able to find more plausible models than the MCMC, using only a fraction of time. Also, due to the fact that the optimal model with $k_0 = 45$, for example, was on average found around the 400th iteration by the MCMC, the actual running times for the complete MCMC runs (with 1,000 iterations) were about 2.5 times those shown in Figure 1. Furthermore, the artificial data sets analyzed here can be considered as fairly small with a relatively simple structure. For example, the *Neisseria* analysis (Tang et al., 2008) included 5175 strains of bacteria, each characterized by 3285 nucleotides. A single run of the greedy optimization algorithm implemented in the latest BAPS version took about 1 day. Because the details of the algorithm are somewhat different from the one used here, it is difficult to estimate exactly how long the MCMC analysis would take for such data. However, given that the MCMC runs took about 50 times longer even to find the top-scoring model with the three largest data set types in the performed

example analyses, even a cautious extrapolation would indicate that a running time of about two months for the Gibbs sampling MCMC would be required to reach models of equal probability to the greedy algorithm.

The results obtained in this section decisively illustrate the suboptimality of using standard MCMC in solving a large-scale unsupervised classification task. Although such behavior has recently been documented elsewhere (Corander et al., 2006; Daumé, 2007), most present-day Bayesian modelers still rely on MCMC in making their inferences about the population structure (Falush et al., 2003; Dawson and Belkhir, 2001). Furthermore, the used algorithms often utilize Gibbs sampling not only to sample the cluster labels, but also to sample the underlying parameters, which in our analysis were integrated out analytically, increasing the computational burden even further. Notice, however, that the considered efficient algorithmic solutions are facilitated by the ability to evaluate analytically the marginal likelihoods of different partitions. If this is not possible, then the Gibbs sampling of the parameters may be the only available solution. However, as illustrated in the articles I-V, the seemingly simple partition modeling framework presented in Section 2 can be generalized to various situations while maintaining the property of being analytically integrable. It is yet worth mentioning that the computational issues related to MCMC are not merely confronted when using the Gibbs sampling, but also with the standard Metropolis-Hastings sampling. For example, in the comparison of algorithms for sampling from Dirichlet process mixture models (Daumé, 2007), an algorithm which combines Gibbs sampling and Metropolis-Hastings is also considered, and the conclusion is that it is sometimes better than mere Gibbs, and sometimes worse, but not conclusively to either direction.

4 Summaries of the original articles

Article I, Bayesian identification of admixture events using multilocus molecular markers

By utilizing molecular marker data, Bayesian mixture models can be used to identify genetically divergent subgroups in data. Admixture models, on the other hand, attempt to distinguish the ancestral sources of the alleles observed in each individual. In contrary to the mixture models, in which individuals are assigned as a whole to different clusters, in admixture models individuals are allowed to be associated with multiple ancestral populations, representing sources of their observed alleles. In Article I, we discuss the problems related to the identifiability of simultaneously estimating the putative ancestral populations and admixture proportions for the individuals. As a solution, we suggest to divide the estimation of the admixture proportions into two phases. In the first phase, a mixture model is utilized to identify genetically diverged subgroups in the data. In the second phase, the clusters identified in the first phase are used as alternative source populations for the alleles observed in the individuals. To estimate the statistical significance related to the estimated admixture pro-

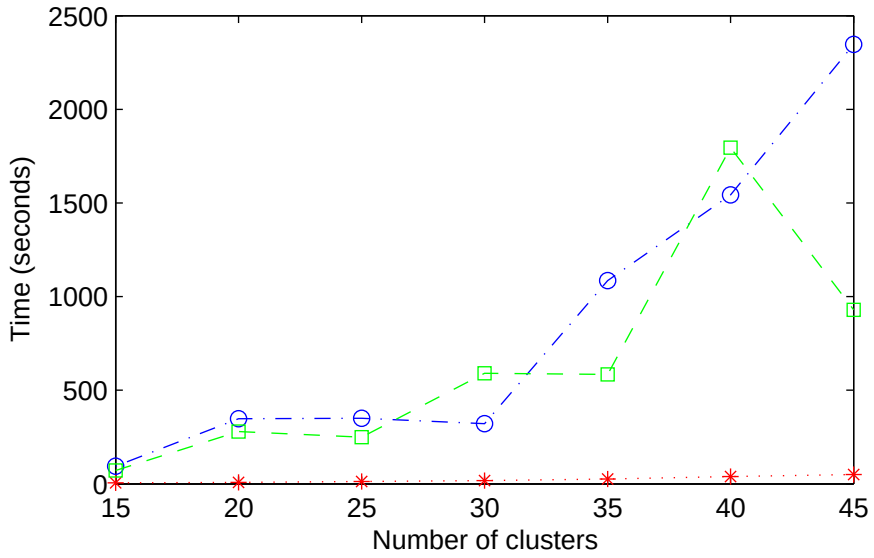

Figure 1: Time taken by different algorithms on artificial data sets. The x-axis shows the complexity level of the data set in terms of the number of clusters and the y-axis shows the average time for five data sets at each level in seconds. Because the times taken by the two versions of the greedy algorithm were roughly equal, they are represented by just one curve (red asterisks). The MCMC initialized by a single cluster is represented by a green square. The MCMC initialized with the correct number plus additional 10 clusters is represented by a blue circle. The times recorded for the MCMC correspond to the times when they reached a model with equal or higher marginal likelihood than the highest value found by the greedy algorithm. If the MCMC never found such a model, the recorded time is the time taken by the MCMC to reach the model with the highest marginal likelihood in that run.

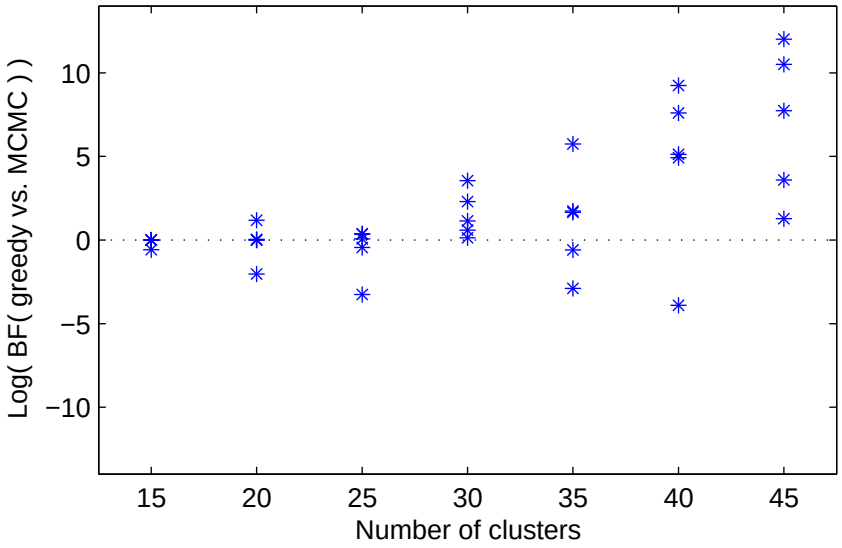

Figure 2: Logarithm of the Bayes factor of M_1 vs. M_2, where M_1 is the model with the highest marginal likelihood value found in the two runs of the greedy algorithm and M_2 is the corresponding model obtained from the MCMC runs. The Bayes factor is shown for each analyzed synthetic data set. Thus, five values are shown for each data set size.

portions, we employ a simulation strategy where artificial reference individuals generated according to the no-admixture null hypothesis are analyzed with the presented admixture model, and the results are compared with the results for the individuals in the data set. To illustrate the advantages of our approach, we present a re-analysis of an extensive real data set (Rosenberg et al., 2005) consisting of human marker data for 1056 individuals collected from 52 native populations from around the world.

Article II, A Bayesian method for identification of stock mixtures from molecular marker data

The usual goal of fisheries management is a mixture of conservation and exploitation goals, and the purpose of the analysis of stock mixtures is to provide a quantitative basis for decision making in fisheries management (Hilborn and Walters, 1992; McAllister and Kirkwood, 1998). In the analysis of stock mixture composition, Bayesian mixture models based on observed genotypes of fish have recently (e.g. Pella and Masuda, 2001) replaced the earlier applied maximum likelihood approach based on latent class mixture models. The Bayesian framework is especially suitable for the analysis of stock mixtures, by offering tools to combine biological background knowledge with observations made about a population (e.g. DNA samples), and to produce probabilistic characterizations of the related uncertainties (e.g. the number of putative stocks).

In Article II, we introduce a novel Bayesian model formulation and its implementation for the analysis of stock mixtures. The method is unique by enabling the exploitation of partial baseline information (i.e. samples with known stock assignment from some, but not necessarily all of the possible stocks), and *a priori* assignment of sampled individuals into sampling units, which the clustering algorithm will consider as inseparable. Another advantage of the method over some alternatives (Dawson and Belkhir, 2001) is the ability to deal with missing observations. The utilization of the baseline samples is achieved by updating the prior distributions for the allele frequency parameters in those stocks from which baseline samples are available. The *a priori* assignment of individuals into sampling units is achieved by modifying the prior distribution for the possible partitions by setting a zero probability to those partitions in which the sampling units are not assigned as a unity in some stock. We illustrate the behavior of the introduced method in several different biological situations by carrying out an extensive simulation study which is based partly on empirical molecular data for the Baltic Sea stock mixtures of Atlantic salmon (Koljonen et al., 2002).

Article III, Bayesian search of functionally divergent protein subgroups and their function specific residues

The rapid growth of protein sequence databases necessitates the development of novel methods for automatic extractions of interesting features among the sequences. One task of particular interest is the classification of proteins based

on their functions. These functions are related to the underlying amino acid sequences and reflect the evolutionary relationships among the sequences, as the residues (i.e. amino acids in the sequence) critical for the function of a protein are more conserved in evolution and consequently, show less variation among a group of related proteins than other parts in a sequence (Casari et al., 1995; Lichtarge et al., 2003). Article III presents a model formulation for the identification of evolutionarily related protein subgroups with their group specific conserved residues. The components of the model include: 1) clustering of the sequences, 2) separation of data dimensions into signal and noise, and 3) specification of the conserved residues for each cluster. Informative priors are derived for the model parameters using available biological knowledge about the characteristics of data. The method is illustrated with two real protein families, urease (Heger and Holm, 2003) and crotonase (A. Sivakumar, personal communication). These analyses show that, among other things, by considering the Bayes factor of signal vs. noise for each sequence position separately, our method is able to highlight the parts of the sequence which are in the 3-dimensional structure close to the active site of the protein.

Article IV, Bayesian clustering of fuzzy feature vectors using a quasi-likelihood approach

In Section 2, the partition model for unsupervised classification was derived for vectors of discrete binary features. In Article IV, we present a modification of the partition model which allows it to be used with continuous features with values in the interval $[0, 1]$. The modification is achieved by replacing the binomial likelihood for binary features with the exponential of the binomial quasi-likelihood (Wedderburn, 1974) for continuous features. The resulting model can be interpreted as a fuzzy formulation of the original model where the sums of observed values replace the observed counts of ones and the sums of observed complementary values (one minus observed values) replace the observed counts of zeros. Such 'fuzzy' ones and zeros are further given weights according to the amount of information they provide about the corresponding mean value, as specified by the dispersion parameter in the quasi-likelihood function. The derived model is illustrated by performing an unsupervised classification for 952 fatty acid profiles of *Bacillus* bacteria (Slabbinck et al., 2008). Such profiles can be obtained from gas chromatographic analyses, and they reflect the chemical composition of the sampled bacteria.

Article V, Bayesian modeling of recombination events in bacterial populations

Admixture models (I; Falush et al., 2003) attempt to identify proportions of the genome which originate from different ancestral sources. However, these models are not able to pinpoint the areas in the genome which have originated from the different sources. On the other hand, Bayesian models based on segmenting the investigated genome sequences, and inferring a phylogeny (i.e. an evolutionary

tree) for each segment, have gained popularity in the recent years (Suchard et al., 2003; Minin et al., 2005). However, the inference with such models is computationally demanding, and for this reason they are mainly suitable for the analysis of data sets with a limited number of strains. In Article V, we consider the identification of the recombinant segments along with their origins for multilocus DNA sequences for a group of bacterial strains. Our goal is to develop a method capable for the analysis of large data sets commonly encountered in microbiology.

Clustering of the strains in the data set provides a starting point for the analysis. Each strain is then investigated in turn more closely by finding an optimal segmentation for its genome sequence. Within each segment, we utilize the partition model by conditioning the remaining strains to be assigned to the clusters indicated by the initial cluster analysis. This strategy enables a computationally efficient implementation, which allows the analysis of data sets comprising hundreds of strains, while taking into account the uncertainty related to the nucleotide values characteristic for the different clusters. A marginal posterior probability distribution is calculated for the origin of each base in the sequence by considering marginal likelihoods of those models which have the most significant impact on this distribution. Highlights of the analysis of a real data set consisting of multilocus genome sequences for strains from genus *Burkholderia* (Baldwin et al., 2005) illustrate that the method is able to produce novel insights about recombination patterns between bacterial species.

5 Acknowledgements

I would like to thank Professor Jukka Corander for his guidance throughout this work. Together with Professor Elja Arjas he created the conditions which made it possible to concentrate on working efficiently while at the same time enjoying the process.

I have had a pleasure to collaborate with people from various backgrounds. I would like to thank jointly all my collaborators. I am especially grateful to professors Bernard De Baets and Peter Dawyndt for hosting me in Belgium in the fall 2006. Other people from outside Finland who have contributed to this work include Adam Baldwin, William P. Hanage, Chris Dowson, and Esh Mahenthiralingam. Also the input from Petri Törönen, Liisa Holm and Samu Mäntyniemi has been invaluable for some parts of the presented work. Thank you all!

I have had many enjoyable conversations about Bayesian statistics and various other subjects with my fellow Ph.D. students Jing Tang, Matti Pirinen, Jukka Kohonen, Jukka Sirén, among others.

Finally, I would like to thank my wife Maija for her support and my daughter Tuuli for brightening my days. Also my parents and my sister have encouraged and supported me during the past few years.

This work was mainly funded by Graduate School in Computational Methods and Information Technology (ComMIT). Financial help was also received from

the Centre of Population Genetic Analyses supported by the Academy of Finland (grant no. 53297), Research funds of the University of Helsinki, Tekes (grant no. 40672/01) and Academy of Finland (grant no. 121301).

References

[1] M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Dover, New York, 1965.

[2] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2:1152–1174, 1974.

[3] A. Baldwin, E. Mahenthiralingam, K. M. Thickett, D. Honeybourne, M. C. J. Maiden, J. R. Govan, D. P. Speert, J. L. LiPuma, P. Vandamme, and C. G. Dowson. Sequence typing for the Burkholderia cepacia complex: a novel scheme that provides both species and strain differentiation. *Journal of Clinical Microbiology*, 43:4665–4673, 2005.

[4] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

[5] M. A. Beaumont and B. Rannala. The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5:251–261, 2004.

[6] J. S. Bernardo and A. F. M. Smith. *Bayesian Theory.* Wiley, Chichester, 1994.

[7] P. J. Brown, T. Fearn, and M. Vannucci. The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. *Biometrika*, 86:635–648, 1999.

[8] P. J. Brown, M. Vannucci, and T. Fearn. Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64:519–536, 2002.

[9] G. Casari, C. Sander, and A. Valencia. A method to predict functional residues in proteins. *Nature Structural Biology*, 2:171–178, 1995.

[10] The Uniprot Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 36:190–195, 2008.

[11] J. Corander, M. Gyllenberg, and T. Koski. Bayesian model learning based on a parallel MCMC strategy. *Statistics and Computing*, 16:355–362, 2006.

[12] J. Corander, M. Gyllenberg, and T. Koski. Random partition models and exchangeability for Bayesian identification of population structure. *Bulletin of Mathematical Biology*, 69:797–815, 2007.

[13] J. Corander, J. Sirén, and E. Arjas. Bayesian spatial modeling of genetic population structure. *Computational Statistics*, 23:111–129, 2008a.

[14] J. Corander, J. Tang, P. Marttinen, and J. Sirén. Enhanced Bayesian modelling in BAPS for learning genetic structures of populations, 2008b. Submitted.

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

[16] J. Corander, P. Waldmann, and M. J. Sillanpää. Bayesian analysis of genetic differentiation between populations. *Genetics*, 163:367–374, 2003.

[17] Hal Daumé III. Fast search for Dirichlet process mixture models. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AIStats)*, San Juan, Puerto Rico, 2007.

[18] K. J. Dawson and K. Belkhir. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research*, 78:59–77, 2001.

[19] B. de Finetti. Funzione caratteristica di un fenomeno aleatorio. *Memorie della R. Accademia dei Lincei*, 4:86–133, 1930.

[20] B. de Finetti. Sur la condition d'équivalence partielle. *Actualités Scientifiques et Industrielles*, 737:5–18, 1938.

[21] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley, New York, 2nd edition, 2001.

[22] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3:87–112, 1972.

[23] D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587, 2003.

[24] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman & Hall/CRC, Boca Raton, 2nd edition, 2004.

[25] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London, 1996.

[26] A. Heger and L. Holm. Sensitive pattern discovery with 'fuzzy' alignments of distantly related proteins. *Bioinformatics*, 19:i130–i137, 2003.

[27] R. Hilborn and C. Walters. *Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty.* Chapman & Hall, New York, 1992.

[28] M. Jakobsson, S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H. C. Funq, Z. A. Szpiech, J. H. Deqnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg, and A. B. Sinqleton. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 21:998–1003, 2008.

[29] H. Jeffreys. *Theory of Probability.* Claredon Press, Oxford, 1961.

[30] Beatrix Jones, Carlos Carvalho, Adrian Dobra, Chris Hans, Chris Carter, and Mike West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20:388–400, 2005.

[31] R. Kass and A. E. Raftery. Bayes factors. *Journal of American Statistical Association*, 90:773–795, 1995.

[32] M.-L. Koljonen, J. Tähtinen, M. Säisä, and J. Koskiniemi. Maintenance of genetic diversity of Atlantic salmon by captive breeding programmes and the geographic distribution of microsatellite variation. *Aquaculture*, 212:69–92, 2002.

[33] P. S. Laplace. *Essai Philosophique sur les Probabilitiés.* Courcier, Paris, 1814.

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

[35] O. Lichtarge, H. Yao, D. M. Kristensen, S. Madabushi, and I. Mihalek. Accurate and scalable identification of functional sites by evolutionary tracing. *Journal of Structural and Functional Genomics*, 4:159–166, 2003.

[36] M. McAllister and G. Kirkwood. Bayesian stock assessment: a review and example application using the logistic model. *ICES Journal of Marine Science*, 55:1031–1060, 1998.

[37] V. N. Minin, K. S. Dorman, F. Fang, and M. A. Suchard. Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, 21:3034–3042, 2005.

[38] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. Technical Report 9815, University of Toronto, 1998.

[39] J. Pella and M. Masuda. Bayesian methods for analysis of stock mixtures from genetic characters. *Fishery Bulletin*, 99:151–167, 2001.

[40] W. Perks. Some observations on inverse probability, including a new indifference rule. *Journal of the Institute of Actuaries*, 73:285–334, 1947.

[41] B. D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, 1996.

[42] N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. Genetic structure of human populations. *Science*, 298:2381–2385, 2002.

[43] S. A. Sisson. Transdimensional Markov chains: A decade of progress and future perspectives. *Journal of American Statistical Association*, 100:1077–1089, 2005.

[44] B. Slabbinck, B. De Baets, P. Dawyndt, and P. De Vos. Genus-wide Bacillus species identification through proper artificial neural network experiments on fatty acid profiles. *Antonie van Leeuwenhoek*, 2008. doi: 10.1007/s10482-008-9229-z.

[45] A. J. Stam. Generation of a random partition of a finite set by an urn model. *Journal of Combinatorial Theory, Series A*, 35:231–240, 1983.

[46] M. A. Suchard, R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer. Inferring spatial phylogenetic variation along nucleotide sequences: A multiple changepoint model. *Journal of American Statistical Association*, 98:427–437, 2003.

[47] J. Tang, W. P. Hanage, C. Fraser, and J. Corander. Identifying currents in the gene pool, 2008. Submitted.

[48] Z. Tu and S. Zhu. Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:657–673, 2002.

[49] R. Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61:439–447, 1974.

[50] D. J. Wilkinson. Bayesian methods in bioinformatics and computational systems biology. *Briefings in Bioinformatics*, 8:109–116, 2007.

[51] S. L. Zabell. W. E. Johnson's sufficientness postulate. *The Annals of Statistics*, 10:1090–1099, 1982.