This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2359431, IEEE Transactions on Pattern Analysis and Machine Intelligence

A Bayesian predictive model for clustering data of mixed discrete and continuous type

Paul Blomstedt, Jing Tang, Jie Xiong, Christian Granlund, and Jukka Corander

Abstract—Advantages of model-based clustering methods over heuristic alternatives have been widely demonstrated in the literature. Most model-based clustering algorithms assume that the data are either discrete or continuous, possibly allowing both types to be present in separate features. In this paper, we introduce a model-based approach for clustering feature vectors of mixed type, allowing each feature to simultaneously take on both categorical and real values. Such data may be encountered, for instance, in chemical and biological analyses, in the analysis of survey data, as well as in image analysis. Our model is formulated within a Bayesian predictive framework, where clustering solutions correspond to random partitions of the data. Using conjugate analysis, the posterior probability for each possible partition can be determined analytically, enabling the utilization of efficient computational search strategies for finding the posterior optimal partition. The derived model is illustrated using several synthetic and real data sets.

Index Terms—Bayes methods, predictive models, unsupervised learning, mixed distributions.

1 Introduction

Clustering is the unsupervised allocation of data items into classes based on the similarity of the items [1]. When no well-defined classes exist *a priori*, inferring the correct number of classes is also beneficial for the purpose of knowledge discovery. The majority of available clustering methods are heuristic in nature, being based on an evaluation of pairwise dissimilarities between items. Such methods include hierarchical clustering and K-means clustering, see e.g. [2]. A drawback of heuristic clustering methods is that they do not directly address the issue of determining the number of classes in the data [3]. An alternative clustering paradigm is formed by model-based clustering methods, which associate a probability distribution with each cluster [4], [5]. Clustering solutions are then essentially based on model fit, allowing the problem of finding an optimal allocation of items into classes as well as that of determining the number of classes to be solved simultaneously. Additionally, they enable the uncertainty relating to a classification of the data to be systematically evaluated.

Most model-based clustering algorithms assume that the data are either discrete or continuous, possibly allowing both types to be simultaneously present, but in separate features (e.g. [6], [7]). In this paper, we consider model-based clustering in the case where data of mixed discrete and continuous type may simultaneously be present within the *same* feature. Note, that there may be some danger of confusion with the term *mixed data*, as it has also been used to refer to data of the former type [7], which can be seen as a special case of the latter.

There are several instances in which data of mixed type may naturally arise. An obvious example is that of continuous data with either

- *P. Blomstedt is with the Department of Mathematics, Åbo Akademi University, Fänriksgatan 3, FI-20500 Turku, Finland and the Helsinki Institute for Information Technology (HIIT), P.O. Box 15400, FI-00076 Aalto University, Finland. E-mail: paul.blomstedt@helsinki.fi*
- *J. Tang is with the Institute for Molecular Medicine Finland (FIMM), P.O. Box 20, FI-00014 University of Helsinki, Finland. E-mail: jing.tang@helsinki.fi*
- *J. Xiong is with the Department of Mathematics and Statistics, P.O. Box 68, FI-00014 University of Helsinki, Finland. E-mail: jukka.corander@helsinki.fi*
- *C. Grandlund is with the Department of Mathematics, Åbo Akademi University. E-mail: cgranlun@gmail.com*
- *J. Corander is with the Department of Mathematics and Statistics, University of Helsinki and the Department of Mathematics, Åbo Akademi University. E-mail: jukka.corander@helsinki.fi*

missing or censored values corresponding to a discrete category. While handling such data is a standard feature in many clustering algorithms, a more challenging situation arises if several discrete categories such as "removed", "unknown", "incorrect", or "censored", all of which may carry information relevant for finding an optimal clustering solution, are simultaneously present in the data. Another typical example of mixed data is spectral data, arising for instance in many chemical and biological analyses. These type of data consist of non-negative real values with an excessive proportion of zeros, signifying the absence of a measured feature. Such data are sometimes referred to as *zero-inflated* data [8]. As a final example of mixed data, we mention image data consisting of vectors of pixel intensities which take values in a closed interval. In a gray-scale image the endpoints of the interval represent black and white, while values inside the interval represent various shades of gray. If pixel intensities are modeled on a continuous scale, an image with sharp contrasts may then display an excessive proportion of values at either or both endpoints.

In formulating our probabilistic clustering model, we utilize a Bayesian predictive framework for unsupervised classification introduced in [9], see also [10]. This framework has previously been applied in solving various classification problems in bioinformatics, see e.g. [11] and [12]. It is based on probabilistic modeling of random partitions of the data, rendering putative classes completely unlabeled and defined only through the individuals they contain. The number of classes is thus also determined by a partition. A closely related framework is that of product partition models, introduced by [13] and [14]. Although seemingly distinct from Bayesian clustering approaches based on Dirichlet process mixture (DPM) models [15], [16], a connection between product partition models and DPM models was established in [17].

The Bayesian predictive random partition framework, originally devised for categorical data, is here extended to handle mixed data. We show that the marginal likelihood can be factorized into a product of a discrete and a continuous component, allowing us to treat these components separately. Although the general model formulation does not require the continuous data to be of a specific distributional form, an explicit expression for the continuous component can be derived if the likelihood has an associated conjugate family of priors, so that the entire joint marginal likelihood can be determined analytically. This enables the utilization of computational strategies which are more efficient than standard Markov chain Monte Carlo methods for finding the posterior optimal partition [18], [19].

The rest of this paper is structured as follows. Section 2 gives a brief overview of the Bayesian predictive random partition framework for unsupervised classification. In Section 3, we develop a predictive model for mixed data and give an explicit expression for the marginal likelihood of a given clustering solution in the important special case that the continuous part of the data is modeled using a normal distribution. In Section 4, the derived model is illustrated using several synthetic and real data sets. The paper concludes with a discussion in Section 5.

2 Bayesian predictive framework for unsupervised classification

Consider a set N of n individuals $i \in N$, characterized by d-dimensional feature vectors $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)}) \in \mathbb{R}^d$. The entire data set of n observed feature vectors will be denoted $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})^{\mathrm{T}}$. The primary task of our analysis is that of partitioning N into k non-empty and non-overlapping clusters, represented by a partition $S = \{s_1, s_2, \ldots, s_k\}$, such that $s_c \cup_{c=1}^{k} = N$, $s_c \cap s_{c'} = \emptyset$, for all $c, c' = 1, \ldots, k$, $c \neq c'$. In this paper, unsupervised classification, or clustering, refers to the joint inference of both the number of clusters k and partition S although often, especially in the case of heuristic clustering methods, k is assumed to be given, conditional on which an optimal partition is sought.

A Bayesian predictive approach to the problem of unsupervised classification can be formulated in terms of a generic model learning task. For a detailed treatment of Bayesian

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2359431, IEEE Transactions on Pattern Analysis and Machine Intelligence

model selection, see [20]. Considering a finite collection $\mathcal{S}$ of partition solutions as the model space, an overall predictive belief model for the data is specified as

$$p(\mathbf{x}) = \sum_{S \in \mathcal{S}} \mathbb{P}(S) p(\mathbf{x}|S), \tag{1}$$

where $\mathbb{P}(S)$ is the prior probability of partition S and $p(\mathbf{x}|S)$ is the predictive belief model under partition S. The latter will be referred to as the *marginal likelihood* (or evidence) of partition S, as it is obtained by integrating $L_{\mathbf{x}}(\boldsymbol{\theta}_S)$, the likelihood of the model parameters $\boldsymbol{\theta}_S \in \Theta_S$ given the data $\mathbf{x}$, with respect to the prior density $\pi(\boldsymbol{\theta}_S)$,

$$p(\mathbf{x}|S) = \int_{\Theta_S} L_{\mathbf{x}}(\boldsymbol{\theta}_S) \pi(\boldsymbol{\theta}_S) d\boldsymbol{\theta}_S. \tag{2}$$

It can be shown [20] that under zero-one loss, the optimal partition $\hat{S}$ is obtained by maximization of the posterior probability for partition S,

$$\mathbb{P}(S|\mathbf{x}) = \frac{\mathbb{P}(S) p(\mathbf{x}|S)}{p(\mathbf{x})}, \tag{3}$$

such that

$$\hat{S} = \arg\max_{S \in \mathcal{S}} \{\mathbb{P}(S) p(\mathbf{x}|S)\}.$$

2.1 Prior distributions on partitions

We will now briefly consider some possible choices for prior distributions on the partitions $S \in \mathcal{S}$. Perhaps the most straightforward choice is to assume *a priori* that each S is equally probable. This leads to a prior of the form

$$\mathbb{P}(S) = \frac{1}{B_n},$$

where B_n, the n:th Bell number [21], counts the total number of ways in which a set of n items can be partitioned. Although uniform on the partitions themselves, a prior of the above type is highly non-uniform on the number of clusters k, implicitly placing large prior weight on solutions with many small clusters. If instead every k is assumed equally probable, the resulting prior is of the form

$$\mathbb{P}(S) = \frac{1}{n \cdot \mathrm{S}_2(n,k)},$$

where $\mathrm{S}_2(n,k)$ denotes the Stirling number of the second kind [22] and counts the number of ways to partition a set of n items into k non-empty subsets. Such a prior will in turn imply strong preferences among individual partitions with different k, giving higher prior probability to solutions with k close to 1 or n.

A strategy which differs from the above construction of uniform priors, is based on the use of a Chinese restaurant process (CRP) [23], which is a discrete-time stochastic process whose value at time point n is a partition of the set $\{1, \ldots, n\}$. This prior has, for all $k \in \{1, \ldots, n\}$, the form

$$\mathbb{P}(S) = \frac{\alpha^k \prod_{c=1}^{k} (n_c - 1)!}{\prod_{i=1}^{n} \alpha + i - 1}, \tag{4}$$

where $\alpha > 0$ is a concentration parameter controlling the tendency to place data items into separate clusters and n_c denotes the number of items in cluster c.

Any choice of prior distribution should appropriately reflect one's prior beliefs about the distribution over partitions. In particular with sparsely informative data sets, the prior may have a large impact on the outcome but on the other hand, for larger data sets where the data are sufficiently informative, the practical difference between different priors may be quite small. While the above choices serve as examples of fairly generic priors, examples of more problem-specific choices can be found in [24] and [25]. An experimental comparison of clusterings under different types of priors can be found in [26].

2.2 Connection to Dirichlet process mixture models

Under the CRP prior, see Equation (4), the predictive partition model introduced in this section satisfies the definition of a product partition model (PPM) [13], [14], which brings about an interesting connection to Dirichlet process mixture (DPM) models [15]. Namely, it is shown in [17] that integrating out the Dirichlet process in a DPM model leads to a clustering structure which is equivalent to that of a PPM with a CRP prior imposed on the

partition space. More specifically, the marginalized DPM will then be equivalent to the overall belief model (1). As noted by [27], when the focus is on clustering or generating a flexible partition model for prediction, it may be appealing to marginalize out the Dirichlet process in a DPM model in order to increase efficiency in computation and to simplify interpretation.

3 Predictive Model for Mixed Data

We begin our model formulation by assuming that each observation $x \in \mathbb{R}$ in a given cluster c and feature j is a realization of a random variable X with distribution function F. Define now $\mathcal{D} := \{x : F(x) - F(x-) > 0\} \subset \mathbb{R}$ to be the set of discontinuity points with respect to F, where $F(x-)$ denotes the left-hand limit of F at x. By the properties of the distribution function and by virtue of $\mathcal{D}$ being countable, X is discrete if

$$\mathbb{P}(X \in \mathcal{D}) = \sum_{x \in \mathbb{R}} [F(x) - F(x-)] = 1$$

and continuous if

$$\mathbb{P}(X \in \mathcal{D}) = \sum_{x \in \mathbb{R}} [F(x) - F(x-)] = 0.$$

Finally, if

$$\mathbb{P}(X \in \mathcal{D}) = \sum_{x \in \mathbb{R}} [F(x) - F(x-)] \in (0, 1)$$

then X is of *mixed* type. The case of singular distributions is not considered here, but see [28] for a further discussion about this topic.

A mixed distribution F can be decomposed as a mixture of a discrete and a continuous component:

$$F(x) = \mathbb{P}(X \in \mathcal{D}) F_D(x) + \mathbb{P}(X \in \mathcal{D}^c) F_C(x),$$

where

$$F_D(x) := F(x | X \in \mathcal{D})$$

and

$$F_C(x) := F(x | X \in \mathcal{D}^c)$$

are the distribution functions of the discrete and continuous components, respectively. For our current purposes, we may assume that $\mathcal{D}$ is finite with cardinality r, i.e. $\mathcal{D} = \{d_1, \ldots, d_r\}$. Without any further knowledge about the distributional form of the discrete component we may in a clustering context assume for it an r-valued categorical distribution as follows. Let $(Y_1 \ldots, Y_r)$ be a random vector such that

$$Y_l = \begin{cases} 1, & \text{if } X = d_l \\ 0, & \text{if } X \neq d_l \end{cases}$$

for $l = 1, \ldots, r$. The discrete component then has probability mass function

$$f_D(y_1 \ldots, y_r | \psi_1, \ldots, \psi_r) = \prod_{l=1}^{r} \psi_l^{y_l}, \qquad (5)$$

where $\psi_l = \mathbb{P}(Y_l = 1)$ and $\sum_{l=1}^{r} \psi_l = 1$. The continuous component is for the time being assumed to be an arbitrary continuous distribution with density $f_C(x|\eta)$ depending on a parameter η.

Suppose now that we have observed some data for cluster c and feature j which we denote $\mathbf{x}_{cj} = (x^{(1)} \ldots, x^{(n_c)})^{\mathrm{T}}$. Further, denoting the mixture weight parameter $w = \mathbb{P}(X^{(i)} \in \mathcal{D}^c)$, for all $i = 1, \ldots, n_c$, and assuming the observations to be independent conditional on the parameters, we may now write the likelihood function of the parameters as

$$L_{\mathbf{x}_{cj}}(w, \psi_1, \ldots, \psi_r, \eta) = \prod_{i=1}^{n_c} \Big[(1-w) \prod_{l=1}^{r} \psi_l^{y_l^{(i)}} + w f_C(x^{(i)}|\eta) \Big]. \qquad (6)$$

As such, an expansion of (6) leads to 2^{n_c} terms, which is inhibitive for analytical calculations. However, by introducing the random variable

$$Y_{r+1} = \begin{cases} 1, & \text{if } X \in \mathcal{D}^c \\ 0, & \text{if } X \in \mathcal{D} \end{cases}$$

and denoting

$$p_l = \begin{cases} (1-w)\psi_l, & \text{if } l = 1, \ldots, r, \\ w, & \text{if } l = r+1 \end{cases}$$

such that $\sum_{l=1}^{r+1} p_l = 1$, the likelihood (6) may

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2359431, IEEE Transactions on Pattern Analysis and Machine Intelligence

be rewritten as

$$L_{\mathbf{x}_{cj}}(p_1,\ldots,p_{r+1},\eta) = \prod_{i=1}^{n_c}\left[p_1^{y_1^{(i)}}\cdots p_r^{y_r^{(i)}} + p_{r+1}f_C(x^{(i)}|\eta)y_{r+1}^{(i)}\right] = \prod_{i=1}^{n_c}\left[p_1^{y_1^{(i)}}\cdots p_r^{y_r^{(i)}}\left(p_{r+1}f_C(x^{(i)}|\eta)\right)^{y_{r+1}^{(i)}}\right] = \prod_{i=1}^{n_c}\left[p_1^{y_1^{(i)}}\cdots p_r^{y_r^{(i)}}p_{r+1}^{y_{r+1}^{(i)}}f_C(x^{(i)}|\eta)^{y_{r+1}^{(i)}}\right].$$

Finally, by defining $Z := X|X \in \mathcal{D}^c$, we write

$$L_{\mathbf{x}_{cj}}(p_1,\ldots,p_{r+1},\eta) = \prod_{l=1}^{r+1}p_l^{n_{cl}}\prod_{m=1}^{n_{c,r+1}}f_C(z^{(m)}|\eta) = \quad (7)$$

$$L_{\mathbf{y}_{cj}}(p_1,\ldots,p_{r+1})L_{\mathbf{z}_{cj}}(\eta), \quad (8)$$

where in the first equation, $n_{cl} = \sum_{i=1}^{n_c} y_l^{(i)}$ such that $\sum_{l=1}^{r+1} n_{cl} = n_c$, and $z^{(m)}$ denotes the m:th observed value in $\mathcal{D}^c$. The factorization in (8) is essential to our model formulation since it separates the likelihoods given the discrete part $\mathbf{y}_{cj}$ and the continuous part $\mathbf{z}_{cj}$ of the data. Note, that $L_{\mathbf{y}_{cj}}(p_1,\ldots,p_{r+1})$ also implicitly includes the mixture parameter $w = p_{r+1}$.

With the likelihood for a single cluster and feature now available to us, we may proceed with the derivation of the marginal likelihood (2) for a given partition S. The model parameters for cluster c and feature j are now explicitly indexed $p_{cj1},\ldots,p_{cj,r+1},\eta_{cj}$, and jointly denoted $\boldsymbol{\theta}_{cj}$. Analogously, the notation $\boldsymbol{\theta}_S$ is used for the entire set of parameters of partition S. Assuming the independence of clusters and features given the model parameters, the joint likelihood for $\boldsymbol{\theta}_S$ can be written as

$$L_{\mathbf{x}}(\boldsymbol{\theta}_S) = \prod_{c=1}^{k}\prod_{j=1}^{d}L_{\mathbf{y}_{cj}}(p_{cj1},\ldots,p_{cj,r+1})L_{\mathbf{z}_{cj}}(\eta_{cj}). \quad (9)$$

We further assume that $(p_{cj1},\ldots,p_{cj,r+1})$ and η_{cj} are mutually independent, such that the prior $\pi(\boldsymbol{\theta}_{cj})$ factorizes as

$$\pi(\boldsymbol{\theta}_{cj}) = \pi(p_{cj1},\ldots,p_{cjr})\pi(\eta_{cj}).$$

Thus,

$$\pi(\boldsymbol{\theta}_S) = \prod_{c=1}^{k}\prod_{j=1}^{d}\pi(p_{cj1},\ldots,p_{cj,r+1})\pi(\eta_{cj}). \quad (10)$$

Decomposing the parameter space as $\Theta_S = \mathrm{P}_S \times \mathrm{H}_S$ corresponding to the respective parameters of the discrete and continuous components, the marginal likelihood can be factorized as

$$\begin{aligned} p(\mathbf{x}|S) &= \int_{\Theta_S} L_{\mathbf{x}}(\boldsymbol{\theta}_S)\pi(\boldsymbol{\theta}_S)d\boldsymbol{\theta}_S \\ &= \int_{\mathrm{P}_S}\left[\prod_{c=1}^{k}\prod_{j=1}^{d}L_{\mathbf{y}_{cj}}(p_{cj1},\ldots,p_{cj,r+1}) \times \pi(p_{cj1},\ldots,p_{cj,r+1})\right]d\boldsymbol{p}_S \\ &\quad\times\int_{\mathrm{H}_S}\prod_{c=1}^{k}\prod_{j=1}^{d}L_{\mathbf{z}_{cj}}(\eta_{cj})\pi(\eta_{cj})d\boldsymbol{\eta}_S \\ &= p(\mathbf{y}|S)p(\mathbf{z}|S). \end{aligned} \quad (11)$$

This factorization enables us to treat the discrete and continuous components of the marginal likelihood separately.

3.1 Marginal likelihood of the discrete component

We will now derive an explicit expression for the marginal likelihood $p(\mathbf{y}|S)$ under the assumption that the discrete component has a likelihood function of the form

$$L_{\mathbf{y}_{cj}}(p_{cj1},\ldots,p_{cj,r+1}) = \prod_{l=1}^{r+1}p_{cjl}^{n_{cjl}}.$$

A conjugate prior for $(p_{cj1},\ldots,p_{cj,r+1})$ is the Dirichlet distribution $\mathrm{Dir}(\lambda_{cj1},\ldots,\lambda_{cj,r+1})$ with density

$$\pi(p_{cj1},\ldots,p_{cj,r+1}) = \frac{\Gamma(\sum_{l=1}^{r+1}\lambda_{cjl})}{\prod_{l=1}^{r+1}\Gamma(\lambda_{cjl})}\prod_{l=1}^{r+1}p_{cjl}^{\lambda_{cjl}-1}.$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2359431, IEEE Transactions on Pattern Analysis and Machine Intelligence

The marginal likelihood for the discrete component can then be written as

$$\begin{aligned} p(\mathbf{y}|S) &= \int_{\mathrm{P}_S} \left[\prod_{c=1}^{k} \prod_{j=1}^{d} L_{\mathbf{y}_{cj}}(p_{cj1}, \ldots, p_{cj,r+1}) \right. \\ &\qquad \left. \times \pi(p_{cj1}, \ldots, p_{cj,r+1}) \right] d\boldsymbol{p}_S \\ &= \int_{\mathrm{P}_S} \left[\prod_{c=1}^{k} \prod_{j=1}^{d} \frac{\Gamma(\sum_{l=1}^{r+1} \lambda_{cjl})}{\prod_{l=1}^{r+1} \Gamma(\lambda_{cjl})} \right. \\ &\qquad \left. \times \prod_{l=1}^{r+1} p_{cjl}^{\lambda_{cjl}+n_{cjl}-1} \right] d\boldsymbol{p}_S \\ &= \prod_{c=1}^{k} \prod_{j=1}^{d} \frac{\Gamma(\sum_{l=1}^{r+1} \lambda_{cjl})}{\Gamma(\sum_{l=1}^{r+1} n_{cjl} + \lambda_{cjl})} \\ &\qquad \times \prod_{l=1}^{r+1} \frac{\Gamma(n_{cjl} + \lambda_{cjl})}{\Gamma(\lambda_{cjl})}, \end{aligned} \tag{12}$$

which is a product of standard Dirichlet-multinomial densities. The hyper-parameters $\lambda_{cj1}, \ldots, \lambda_{cj,r+1}$ can be chosen according to auxiliary information about the features in the data. When such information is not available, a common choice is to use $\lambda_{cjl} = 1/(r+1)$, for all $l = 1, \ldots, r+1$.

3.2 Marginal likelihood of the continuous component

So far, no assumptions have been made about the distributional form of the density of the continuous component $f_C(\cdot|\eta_{cj})$, see (7). While in principle, any form suitable for a specific problem at hand could be used, we will here only deal with the case of a normal density. This choice is firstly due to its wide applicability and secondly, it enables efficient analytical calculations through the use of conjugate priors on the parameters.

Defining $\eta_{cj} = (\mu_{cj}, \tau_{cj})$, where μ_{cj} and τ_{cj} denote the expected value and precision, respectively, the likelihood of the parameters of cluster c and feature j is

$$\begin{aligned} L_{\mathbf{z}_{cj}}(\mu_{cj}, \tau_{cj}) &= \prod_{m_c=1}^{n_{cj,r+1}} f_C(z_{m_cj}|\eta_{cj}) \\ &= \prod_{m_c=1}^{n_{cj,r+1}} (2\pi)^{-1/2} \tau_{cj}^{1/2} \\ &\qquad \times \exp\Big\{ -\frac{\tau_{cj}}{2}(z_{m_cj} - \mu_{cj})^2 \Big\}. \end{aligned}$$

A conjugate prior for parameter τ_{cj} is the gamma distribution with shape α_{cj} and inverse scale β_{cj}. Conditionally on τ_{cj}, a conjugate prior for μ_{cj} is the normal distribution with mean μ_{cj0} and precision $\rho_{cj}\tau_{cj}$. The joint prior therefore has a normal-gamma density

$$\begin{aligned} \pi(\mu_{cj}, \tau_{cj}) &= \pi(\mu_{cj}|\tau_{cj})\pi(\tau_{cj}) \\ &= (2\pi)^{-1/2} (\rho_{cj}\tau_{cj})^{1/2} \\ &\qquad \times \exp\Big\{ -\frac{\rho_{cj}\tau_{cj}}{2}(\mu_{cj} - \mu_{cj0})^2 \Big\} \\ &\qquad \times \frac{\beta_{cj}^{\alpha_{cj}}}{\Gamma(\alpha_{cj})} \tau_{cj}^{\alpha_{cj}-1} \exp(-\beta_{cj}\tau_{cj}). \end{aligned}$$

Using the conjugate property of the normal-gamma distribution (details omitted), and decomposing the parameter space as $\mathrm{H}_S = \mathrm{M}_S \times \mathrm{T}_S$, we obtain the marginal likelihood for the continuous component in an explicit form

$$\begin{aligned} p(\mathbf{z}|S) &= \int_{\mathrm{M}_S} \int_{\mathrm{T}_S} \left[\prod_{c=1}^{k} \prod_{j=1}^{d} L_{\mathbf{z}_{cj}}(\mu_{cj}, \tau_{cj}) \right. \\ &\qquad \left. \times \pi(\mu_{cj}|\tau_{cj})\pi(\tau_{cj}) \right] d\boldsymbol{\mu}_S d\boldsymbol{\tau}_S \\ &= \prod_{c=1}^{k} \prod_{j=1}^{d} (2\pi)^{-n_{cj,r+1}/2} \Big(\frac{\rho_{cj}}{\rho_{cj*}} \Big)^{1/2} \\ &\qquad \times \frac{\Gamma(\alpha_{cj*})}{\Gamma(\alpha_{cj})} \frac{\beta_{cj}^{\alpha_{cj}}}{\beta_{cj*}^{\alpha_{cj*}}}, \end{aligned} \tag{13}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2359431, IEEE Transactions on Pattern Analysis and Machine Intelligence

where

$$\begin{aligned}\rho_{cj*} &= \rho_{cj} + n_{cj,r+1}\\ \alpha_{cj*} &= \alpha_{cj} + \frac{n_{cj,r+1}}{2}\\ \beta_{cj*} &= \beta_{cj} + \frac{1}{2}\sum_{m_c=1}^{n_{cj,r+1}}(z_{m_cj} - \bar{z}_{cj})^2\\ &\quad + \frac{n_{cj,r+1}\rho_{cj}(\bar{z}_{cj} - \mu_{cj})^2}{2\rho_{cj*}}\\ \bar{z}_{cj} &= \frac{1}{n_{cj,r+1}}\sum_{m_c=1}^{n_{cj,r+1}}(z_{m_cj}).\end{aligned}$$

As in the case of the hyper-parameters of the discrete component, the hyper-parameters for $\pi(\mu_c, \tau_{cj})$ can be chosen according to any available information about the features. In the absence of such information, we have chosen to set

$$\mathbb{E}(\mu_{cj}) = \mu_{cj0} = \bar{z}_j = \frac{1}{n}\sum_{i=1}^{n} z_j^{(i)}$$

and

$$\begin{aligned}\mathbb{E}(\tau_{cj}) &= \frac{\alpha_{cj}}{\beta_{cj}} = (s_j^2)^{-1}\\ &= \left(\frac{1}{n-1}\sum_{i=1}^{n}(z_j^{(i)} - \bar{z}_j)^2\right)^{-1},\end{aligned}$$

where $\bar{z}_j$ and s_j^2 denote the sample mean and variance, respectively, of all observed values of feature j (i.e. before any division of the data into clusters). In practice, we have normalized the continuous data for each feature such that $\bar{z}_j = 0$ and $s_j^2 = 1$, and set $\mu_{cj0} = 0$ and $\alpha_{cj} = \beta_{cj} = \rho_{cj} = 1$.

4 Illustrations

4.1 Search algorithm

In the following illustrations, we use a greedy stochastic search algorithm to search the model space $\mathcal{S}$ for an optimal partition $\hat{S}$. The algorithm is similar to the one used in [19], where a detailed description of the algorithm can be found. In short, after an initial partition has been obtained using complete linkage clustering, the search then moves in the model space in a greedy fashion by application of the following operators:

1) Move an item to another cluster.
2) Merge two clusters.
3) Split a cluster.

Each state $S \in \mathcal{S}$ is evaluated in terms of its posterior probability (3) and convergence takes place when none of the operators lead to an improvement in this probability. Throughout the illustrations, a CRP prior (4) with $\alpha = 1$ will be used as a prior over the partition space.

4.2 Analysis of synthetic data

First, we use the predictive random partition model (RPM) with mixed likelihood introduced in Sections 2 and 3 to cluster simulated data, which enables us to validate its performance as the data generating mechanism and the true partition of the data are known. To evaluate performance, we use the adjusted Rand index (ARI) [29], which measures the similarity between two partitions. While in theory, it is defined in the interval $[-1, 1]$, it will mostly yield values between 0 and 1, where the former denotes a degree of similarity corresponding to a random allocation of items into clusters and the latter indicates that two partitions are equal.

As few model-based methods are explicitly designed to handle mixed data (note the possible ambiguity with the term *mixed*, see Section 1), we will compare the clustering solutions for the simulated data with those obtained using two standard heuristic clustering methods, the K-means algorithm (e.g. [30]) and Ward's method for hierarchical clustering [31]. In [19], however, a quasi-likelihood approach was introduced for clustering fuzzy $[0, 1]$-valued feature vectors, also allowing the data to be discrete at 0 and 1. This approach was shown in number of simulations to perform favorably to both a discrete Bayesian clustering model using binarized data and the popular Bayesian model-based clustering algorithm AutoClass [6]. In order to include the quasi-likelihood approach in our comparison, we initially restrict ourselves to data with values in the interval $[0, 1]$. Also included in the comparison are non-mixed Gaussian and binarized RPMs.

All simulated data sets have 20 clusters while the number of features is varied from 10 to 90.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2359431, IEEE Transactions on Pattern Analysis and Machine Intelligence

For each cluster c and feature j, a uniformly distributed value

$$p_{cj} \sim \text{Unif}(0, 1)$$

was first generated, whereupon n_c values were simulated from a binomial distribution $\text{Bin}(2, p_{cj})$. Finally, all values equal to 2 were replaced by values generated from a normal distribution $\text{N}(\mu_{cj}, \sigma_{cj}^2)$ with mean and standard distributed as

$$\mu_{cj} \sim \text{Unif}(0, 1), \quad \sigma_{cj} \sim \text{Unif}(0.01, 0.1). \quad (14)$$

The number of items n_c, $c = 1, \ldots, 20$, in each cluster was drawn uniformly at random from the set $\{10, 11, \ldots, 40\}$. Table 1 shows the results for a series of 10 simulations per number of features. Note, that the results for Ward's method and K-means are given conditional on the correct number of clusters while the results for the other clustering methods are given conditional on the inferred number of clusters. Comparing the results for the mixed and non-mixed forms of RPM reveals that explicitly accounting for the mixed structure of the data is particularly beneficial when the number of features is relatively small. As the number increases, the patterns in the data become more distinguishable even when the "incorrect" model is used to form clusters.

In a second simulated example, we add one more discrete category and no longer restrict ourselves to the interval $[0, 1]$. Here, the data for each cluster and feature were generated from a mixture of $\text{Bin}(2, \psi_{cj})$ and $\text{N}(\mu_{cj}, \sigma_{cj}^2)$ with mixture weight w_{cj}, such that

$$w_{cj}, \psi_{cj} \sim \text{Unif}(0, 1)$$

independently, and

$$\mu_{cj} \sim \text{Unif}(5, 10), \quad \sigma_{cj} = 1.$$

The results are shown in Table 2. It is important to note that unlike the mixed RPM, the Gaussian RPM and heuristic algorithms treat the discrete categories as numerical data. While having the discrete categories on a nominal scale instead of a numerical scale would have no effect on the performance of the mixed RPM, this could seriously affect the performance of the other clustering methods as the nominal values would either have to be omitted or carefully re-coded on a numerical scale.

Table 1
Average ARI (sd's in parentheses) for simulated data sets with 20 clusters and a varying number of features. Each result is based on a series of 10 simulations.

No. of features	**Clustering method**		
	Mixed RPM	Gaussian RPM	Binarized RPM
10	0.693 (0.082)	0.131 (0.033)	0.102 (0.018)
30	0.969 (0.034)	0.587 (0.086)	0.548 (0.079)
50	0.997 (0.008)	0.949 (0.029)	0.927 (0.029)
70	1.000 (0.000)	0.994 (0.008)	0.992 (0.004)
90	1.000 (0.000)	0.998 (0.002)	0.995 (0.003)
	Quasi-likelihood	Ward	K-means
10	0.106 (0.025)	0.110 (0.025)	0.114 (0.025)
30	0.492 (0.094)	0.344 (0.070)	0.458 (0.117)
50	0.773 (0.080)	0.626 (0.047)	0.781 (0.059)
70	0.971 (0.017)	0.835 (0.041)	0.900 (0.055)
90	0.996 (0.004)	0.908 (0.024)	0.891 (0.054)

Table 2
Average ARI (sd's in parentheses) for second simulation experiment.

No. of features	**Clustering method**	
	Mixed RPM	Gaussian RPM
10	0.237 (0.084)	0.159 (0.045)
30	0.858 (0.116)	0.663 (0.080)
50	0.985 (0.037)	0.929 (0.036)
70	1.000 (0.000)	0.984 (0.007)
90	1.000 (0.000)	0.997 (0.004)
	Ward	K-means
10	0.127 (0.032)	0.148 (0.047)
30	0.476 (0.055)	0.595 (0.049)
50	0.759 (0.045)	0.863 (0.044)
70	0.888 (0.018)	0.891 (0.047)
90	0.953 (0.022)	0.883 (0.049)

4.3 Amphetamine profiling

In this illustration the mixed data model is applied to classifying amphetamine samples into production batches based on their associated impurity profiles. Amphetamine is known to be produced in batches, each of which contains a unique chromatographic profile of impurities comprising different amounts of decomposi-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2359431, IEEE Transactions on Pattern Analysis and Machine Intelligence

tion products, reaction by-products and unchanged starting materials [32]. These profiles can be used in forensic investigations to link together samples from the same batch.

The data consist of the impurity profiles of 233 amphetamine samples, obtained using gas-chromatography mass-spectroscopy (GC–MS), see e.g. [33]. This method produces a signal over a retention time interval, where impurities appear as peaks. The raw data has 3826 measurement values for each profile, see Fig. 1. An initial preprocessing of the data was done as follows. The baseline for each profile (i.e. the level at which measurements are noise) was first estimated using asymmetric least squares smoothing [34]. This typically yields a non-linear smooth function over the retention time interval. A common phenomenon in chromatographic analyses is that the repeatability standard deviation of a measured value at any given retention time point is linearly proportional to the magnitude of this value [35]. Therefore, to control the level of noise across the entire profile, we divided the raw data by the estimated baseline. Since the resolution of the raw data was unnecessarily high to yield meaningful features, the 3820 first values of the baseline corrected profiles were divided into intervals of 20 measurements, after which the highest value was selected from each of the resulting 191 intervals to represent an impurity. Finally, to separate signal from noise, all values less than 3 were set to 0.

With no knowledge of the true partition of the data, we first tested the approach on simulated data. For a simulated data set of k clusters, we randomly selected k profiles from the baseline corrected data. Denoting these profiles $\boldsymbol{\mu}_c$, $c = 1, \ldots, k$, we simulated for each cluster n_c vectors from a multivariate normal distribution with mean vector $\boldsymbol{\mu}_c$ and covariance matrix $\Sigma_c = [.25\,\mathrm{diag}(\boldsymbol{\mu}_c)]^2$. As a final step we reduced the number of features to 191 and applied the same cutoff value of 3 as above. A visual comparison between a simulated data set ($k = 10$ and $n_c = 23$ for all c) and the real data set suggests that the simulation captures the characteristics of the data reasonably well, see Fig. 2. Clustering 100 simulated data sets, each with $k = 5$ and $n_c = 20$ for all c, resulted

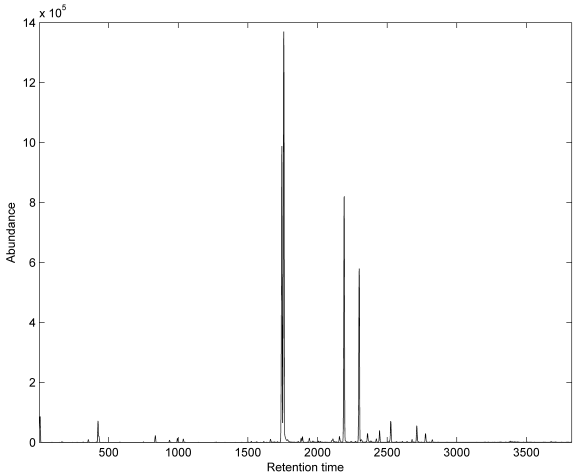


Figure 1. Impurity profile of amphetamine sample.

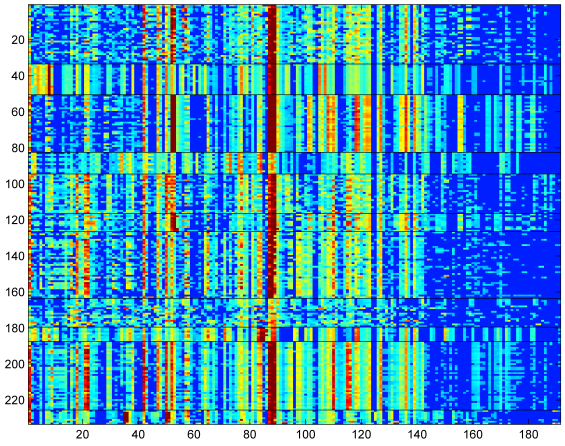


Figure 3. Clustering solution for amphetamine data.

in an average ARI of 0.984 with a standard deviation of 0.058. The average ARI slightly less than 1 is explained by the fact that whenever two similar profiles, presumably from the same batch, were selected in the simulation as mean vectors for two separate clusters, the resulting clusters would eventually be joined together by the clustering algorithm and lead to a reduced ARI. Finally, the clustering solution for the real data is shown in Fig. 3.

4.4 Clustering of hand-written digits

The MNIST database [36], available from http://yann.lecun.com/exdb/mnist/, is a database

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2359431, IEEE Transactions on Pattern Analysis and Machine Intelligence

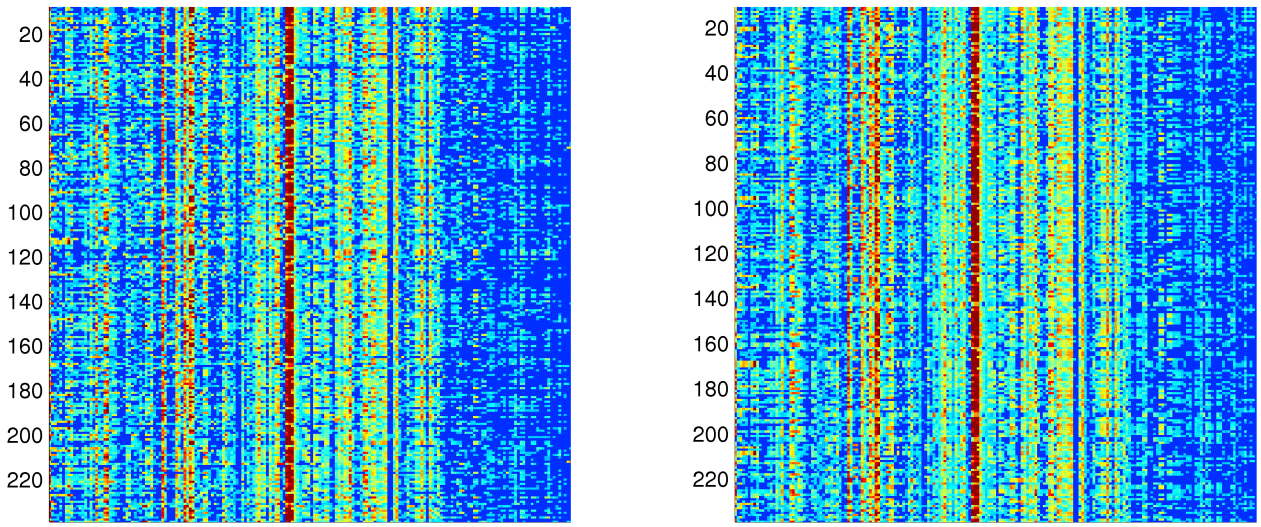


Figure 2. Comparison between real (left panel) and simulated (right panel) amphetamine data. Rows correspond to individual samples and columns to features. Zero values are shown in blue and the highest values in red.

consisting of 60,000 training examples and 10,000 test examples of images of hand-written digits. Each item in the database has 784 features corresponding to 28×28 pixels with values ranging from 0 to 255. The data exhibit a high proportion of 0's ($\sim 80\%$), the remaining proportion being roughly uniformly distributed over $\{1, \ldots, 255\}$, with a slight concentration of data points on values close to 255. Although being inherently discrete, modeling the data on a continuous scale enables a more parsimonious parametrization to be used. The data can thus be interpreted as having a structure of discrete 0's mixed with uniformly distributed continuous values, which lends itself naturally to the clustering framework introduced in this paper. Another model structure which could reasonably be motivated by the characteristics of the data is to assume discrete values at both end points of the support, with a Gaussian continuous component to account for subtle label-specific variations in intermediate values.

For this illustration, we limit ourselves to a subset of 3,147 items of the MNIST test set, corresponding to all items labeled 0, 1, or 2. Unlike a supervised classification task, where each item is classified into one of a number of *pre-specified* classes, our objective here is to find a clustering solution which in an unsupervised manner is able to discriminate between the different labels. Additionally, we want to find subclasses corresponding to sets of prototypical handwritings within classes. Fig. 4 shows the pixel-wise cluster means for the solution of 26 clusters obtained using the discrete-uniform model. The discrete-Gaussian-discrete model resulted in 27 clusters.

To investigate the discriminative properties of the clustering solutions, we calculated the theoretical classification error, i.e. the classification error that would result from associating each cluster with a label. The label of a cluster was determined according to the most frequently occurring true label among the items in that cluster. This induced a classification of the data from which the error was calculated as the proportion of misclassified data items. The classification error for the discrete-uniform model was 0.0089. Using the discrete-Gaussian-discrete model resulted in a slightly higher classification error of 0.0146, again underlining the significance of carefully choosing an appropriate model structure for a given problem. We finally note, that while many supervised clas-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2014.2359431, IEEE Transactions on Pattern Analysis and Machine Intelligence

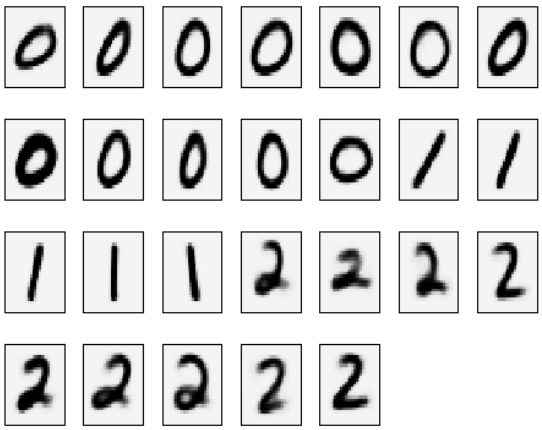


Figure 4. Obtained cluster means for labels 0, 1 and 2 of the MNIST test data set.

sification methods have been reported to yield more accurate results on the MNIST data (see e.g. [37]), the results indicate that the clusters found by the unsupervised models are sensible and could probably still be improved upon in supervised mode. The possibility of extending the present clustering framework to supervised classification is briefly discussed at the end of the concluding section.

5 Discussion

In this paper, we have developed a clustering model for data with mixed categorical and continuous features, which utilizes a Bayesian predictive random partition framework introduced in [9]. We have shown that explicitly accounting for a mixed structure in the data is beneficial, in particular when the number of features is relatively small. The developments build upon the notion of mixed distributions, which have traditionally found use for instance in actuarial science for modeling insurance claims [38] and in geophysical sciences for modeling precipitation [39], [40]. For a brief overview of the properties of mixed distributions, see [41]. In Section 3, we show that the marginal likelihood for a given clustering solution can be factorized into a product of a discrete and a continuous component, which enables these two components to be treated separately. We also give an analytical expression for the marginal likelihood under the assumption that the discrete part of the data follows a categorical distribution and the continuous part is normally distributed. The categorical distribution is a generic choice in the sense that each discrete category has its own parameter, while with additional background information about the distributional form of the discrete data, a more efficient parametrization could possibly be achieved. The normal distribution, on the other hand, is a convenient choice for the continuous data, both due to its wide applicability and its conjugate properties enabling analytical posterior inference, although the general model formulation is not restricted to this choice. For example, in Section 4.4 a uniform distribution was assigned to the continuous component.

Several extensions could be made to the predictive model developed in Section 3. First of all, we have for notational clarity assumed that the number of discrete categories remains constant over all features. This is also a reasonable assumption in many real-world applications, such as those of Section 4. However, an extension to the more general case of a varying number of categories would be relatively straightforward. Also, the continuous part could in principle have a different parametric family for every feature. Most importantly yet, the Bayesian predictive model formulation allows a direct extension to supervised and semi-supervised classification [42]. The latter refers to a supervised classification task where test items may be allocated to classes other than those specified by the training data. For instance in the case of the amphetamine data in Section 4.3, one may wish to compare a set of seized drug samples with a database of samples from identified clandestine laboratories, while allowing for the possibility that some of the samples originate from a yet unidentified laboratory.

Acknowledgments

The authors would like to thank Pekka Marttinen and Jukka Kohonen for their helpful input on matters of implementation and the National Bureau of Investigation Forensic Laboratory,

Finland (NBI) for providing the amphetamine data. Three anonymous referees are gratefully acknowledged for their insightful comments which helped to improve the manuscript. The research was funded by Academy of Finland grant no. 251170 and ERC grant no. 239784.

REFERENCES

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[3] C. Fraley and A. E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," *The Computer Journal*, vol. 41, no. 8, pp. 578–588, 1998.

[4] H. Bock, "Probabilistic models in cluster analysis." *Computational Statistics and Data Analysis*, vol. 23, pp. 5–28, 1996.

[5] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *JASA*, vol. 97, no. 458, pp. 611–631, 2002.

[6] P. Cheeseman and J. Stutz, "Bayesian classification (AutoClass): Theory and results," in *Advances in knowledge discovery and data mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. Am. Assoc. for Artificial Intelligence, 1996, pp. 153–180.

[7] I. Morlini, "A latent variables approach for clustering mixed binary and continuous variables within a gaussian mixture model," *Adv. Data Anal. Classif.*, vol. 6, pp. 5–28, 2012.

[8] J. Tang, J. Tao, H. Urakawa, and J. Corander, "T-BAPS: A Bayesian statistical tool for comparison of microbial communities using terminal-restriction fragment length polymorphism (T-RFLP) data," *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1, 2007, doi: 10.2202/1544-6115.1303.

[9] J. Corander, M. Gyllenberg, and T. Koski, "Bayesian unsupervised classification framework based on stochastic partitions of data and a parallel search strategy," *Advances in Data Analysis and Classification*, vol. 3, pp. 3–24, 2009.

[10] ——, "Random partition models and exchangeability for Bayesian identification of population structure," *Bull. of Math. Biology*, vol. 69, pp. 797–815, 2007.

[11] W. P. Hanage, C. Fraser, J. Tang, T. Connor, and J. Corander, "Hyper-recombination, diversity and antibiotic resistance in the pneumococcus," *Science*, vol. 324, pp. 1454–1457, 2009.

[12] C. Chewapreecha, S. R. Harris, N. J. Croucher, C. Turner, P. Marttinen, L. Cheng, A. Pessia, D. M. Aanensen, A. E. Mather, A. J. Page, S. J. Salter, D. Harris, F. Nosten, D. Goldblatt, J. Corander, J. Parkhill, P. Turner, and S. D. Bentley, "Dense genomic sampling identifies highways of pneumococcal recombination," *Nature Genetics*, vol. 46, no. 3, pp. 305–309, 2014.

[13] J. A. Hartigan, "Partition models," *Comm. Statistics – Theory and Methods*, vol. 19, pp. 2745–2756, 1990.

[14] D. Barry and J. A. Hartigan, "Product partition models for change point problems," *Annals of Statistics*, vol. 20, pp. 260–279, 1992.

[15] C. E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974.

[16] R. Moroni, P. Blomstedt, L. Wilhelm, T. Reinikainen, E. Sippola, and J. Corander, "Markov Chain sampling methods for Dirichlet process mixture models," *J. Computational and Graphical Statistics*, vol. 9, pp. 249–265, 2000.

[17] F. A. Quintana and P. L. Iglesias, "Bayesian clustering and product partition models," *JRSS B*, vol. 65, no. 2, pp. 557–574, 2003.

[18] P. Marttinen, J. Corander, P. Törönen, and L. Holm, "Bayesian search of functionally divergent protein subgroups and their function specific residues," *Bioinformatics*, vol. 22, pp. 2466–2474, 2006.

[19] P. Marttinen, J. Tang, B. De Baets, P. Dawyndt, and J. Corander, "Bayesian clustering of fuzzy feature vectors using a quasi-likelihood approach," *IEEE TPAMI*, vol. 31, no. 1, pp. 74–85, 2009.

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

[21] G. C. Rota, "The number of partitions of a set," *Am. Math. Monthly*, vol. 71, no. 5, pp. 498–504, 1964.

[22] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover, 1964, 10th printing.

[23] J. Pitman, *Combinatorial Stochastic Processes*, ser. Lecture Notes in Mathematics. Berlin: Springer-Verlag, 2006, vol. 1875, lectures from the 32nd Summer School on Probability Theory, Saint-Flour, July 7–24, 2002.

[24] J. Corander and P. Marttinen, "Bayesian identification of admixture events using multi-locus molecular markers," *Molecular Ecology*, vol. 15, pp. 2833–2843, 2006.

[25] J. Corander, J. Sirén, and E. Arjas, "Bayesian spatial modelling of genetic population structure," *Computational Statistics*, vol. 23, pp. 111–129, 2008.

[26] J. Kohonen and J. Corander, "Computing exact clustering posteriors with subset convolution," *Comm. Statistics – Theory and Methods*, 2014, in press.

[27] J.-H. Park and D. B. Dunson, "Bayesian generalized product partition model," *Statistica Sinica*, vol. 20, pp. 1203–1226, 2010.

[28] L. H. Koopmans, "Some simple singular and mixed probability distributions," *Am. Math. Monthly*, vol. 76, no. 3, pp. 297–299, 1969.

[29] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, pp. 193–218, 1985.

[30] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, Inc., 2001.

[31] J. Ward, J. H., "Hierarchical grouping to optimize an objective function," *JASA*, vol. 58, no. 301, pp. 236–244, 1963.

[32] M. D. Cole, *The Analysis of Controlled Substances*. John Wiley & Sons, Ltd., 2003.

[33] D. Harris, *Quantitative Chemical Analysis*, 8th ed. W. H. Freeman & Company, 2010.

[34] P. H. C. Eilers and H. F. M. Boelens, "Baseline correction with asymmetric least squares smoothing," http://www.science.uva.nl/~hboelens/publications/draftpub/Eilers_2005.pdf, 2005, online; accessed 20-May-2013.

[35] P. Blomstedt, R. Gauriot, N. Viitala, T. Reinikainen, and J. Corander, "Bayesian predictive modeling and comparison of oil samples," *J. Chemometrics*, vol. 28, no. 1, pp. 52–59, 2014.

[36] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-

based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.

[37] D. C. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," http://arxiv.org/abs/1202.2745, 2012, arXiv preprint.

[38] R. Kaas, M. Goovaerts, J. Dhaene, and M. Denuit, *Modern Actuarial Risk Theory: Using R*, 2nd ed. Springer, 2009.

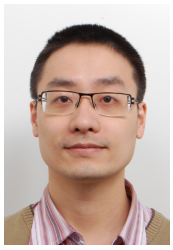
[39] A. Feuerverger, "On some methods of analysis for weather experiments," *Biometrika*, vol. 66, no. 3, pp. 655–658, 1979.

[40] B. Kedem, L. S. Chiu, and G. R. North, "Estimation of mean rain rate: Application to satellite observations," *J. Geophysics Research*, vol. 95, pp. 1965–1972, 1990.

[41] S. Kumar, "Mixed type distributions," http://www.stat.uiowa.edu/~nshyamal/22S175/DI.pdf, 2004, online; accessed 20-May-2013.

[42] J. Corander, Y. Cui, T. Koski, and J. Sirén, "Have I seen you before? Principles of Bayesian predictive classification revisited," *Statistics and Computing*, vol. 23, pp. 59–73, 2013.

Paul Blomstedt received his MSc degree in statistics from the University of Helsinki in 2007 and his PhD degree in statistics from Åbo Akademi University in 2013. Since then he has worked as postdoctoral researcher at the University of Helsinki, and more recently at the Helsinki Institute for Information Technology (HIIT), where his research focuses on the application of Bayesian models in machine learning.



Jing Tang is a senior researcher at the Institute for Molecular Medicine Finland (FIMM), aiming at developing computational modeling approaches to understand molecular interactions between drugs and diseases. He obtained the PhD degree in Statistics from the University of Helsinki in 2009.

Jie Xiong received his BSc degree in applied mathematics from Beijing Normal University in 2009. He received his MSc in Bioinformatics from the University of Helsinki in 2011 and continues to work there toward a PhD degree. His research interests are Bayesian inference, machine learning and computational statistics.

Christian Granlund received his MSc in applied mathematics from Åbo Akademi University in 2012 and has since then been involved in an industrial project focusing on statistical thermodynamics.

Jukka Corander is a professor of statistics at the University of Helsinki. He received his MSc and PhD degrees in statistics in 1995 and 2000, respectively, from Stockholm University. His main academic interests are graphical models, probabilistic classification theory, stochastic computation theory, statistical models for molecular evolution and population genetics, and data mining with molecular biological data.