Bayesian Analysis (2018) **13**, Number 2, pp. 559–626

Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)

Sara Wade* and Zoubin Ghahramani†

Abstract. Clustering is widely studied in statistics and machine learning, with applications in a variety of fields. As opposed to popular algorithms such as agglomerative hierarchical clustering or k-means which return a single clustering solution, Bayesian nonparametric models provide a posterior over the entire space of partitions, allowing one to assess statistical properties, such as uncertainty on the number of clusters. However, an important problem is how to summarize the posterior; the huge dimension of partition space and difficulties in visualizing it add to this problem. In a Bayesian analysis, the posterior of a real-valued parameter of interest is often summarized by reporting a point estimate such as the posterior mean along with 95% credible intervals to characterize uncertainty. In this paper, we extend these ideas to develop appropriate point estimates and credible sets to summarize the posterior of the clustering structure based on decision and information theoretic techniques.

Keywords: mixture model, random partition, variation of information, Binder's loss.

1 Introduction

Clustering is widely studied in statistics and machine learning, with applications in a variety of fields. Numerous models and algorithms for clustering exist, and new studies which apply these methods to cluster new datasets or develop novel models or algorithms are constantly being produced. Classical algorithms such as agglomerative hierarchical clustering or the k-means algorithm (Hartigan and Wong (1979)) are popular but only explore a nested subset of partitions or require specifying the number of clusters apriori. Moreover, they are largely heuristic and not based on formal models, prohibiting the use of statistical tools, for example, in determining the number of clusters.

Model-based clustering methods utilize finite mixture models, where each mixture component corresponds to a cluster (Fraley and Raftery (2002)). Problems of determining the number of clusters and the component probability distribution can be dealt with through statistical model selection, for example, through various information criteria. The expectation-maximization (EM) algorithm is typically used for maximum likelihood estimation (MLE) of the mixture model parameters. Given the MLEs of the parameters, the posterior probability that a data point belongs to a class can be computed through Bayes rule. The cluster assignment of the data point corresponds to the class with maximal posterior probability, with the corresponding posterior probability reported as a

*University of Warwick, s.wade@warwick.ac.uk
†University of Cambridge, zoubin@eng.cam.ac.uk

 https://doi.org/10.1214/17-BA1073

measure of uncertainty. Importantly, however, this measure of uncertainty ignores uncertainty in the parameter estimates. As opposed to MLE, Bayesian mixture models incorporate prior information on the parameters and allow one to assess uncertainty in the clustering structure unconditional on the parameter estimates.

Bayesian nonparametric mixture models assume that the number of components is infinite. As opposed to finite mixture models, this not only avoids specification of the number of components but also allows the number of clusters present in the data to grow unboundedly as more data is collected. Bayesian nonparametric mixture models induce a random partition model (Quintana (2006)) of the data points into clusters, and the posterior of the random partition reflects our belief and uncertainty of the clustering structure given the data.

However, an important problem in Bayesian nonparametric cluster analysis is how to summarize this posterior; indeed, often the first question one asks is what is an appropriate point estimate of the clustering structure based on the posterior. Such a point estimate is useful for concisely representing the posterior and often needed in applications. Moreover, a characterization of the uncertainty around this point estimate would be desirable in many applications. Even in studies of Bayesian nonparametric models where the latent partition is used simply as a tool to construct flexible models, such as in mixture models for density estimation (Lo (1984)), it is important to understand the behavior of the latent partition to improve understanding of the model. To do so, the researcher needs to be equipped with appropriate summary tools for the posterior of the partition.

Inference in Bayesian nonparametric partition models usually relies on Markov chain Monte Carlo (MCMC) techniques, which produce a large number of partitions that represent approximate samples from the posterior. Due to the huge dimension of the partition space and the fact that many of these partitions are quite similar differing only in a few data points, the posterior is typically spread out across a large number of partitions. Clearly, describing all the unique partitions sampled would be unfeasible, further emphasizing the need for appropriate summary tools to communicate our findings.

In a typical Bayesian analysis, the posterior of a univariate parameter of interest is often summarized by reporting a point estimate such as the posterior mean, median, or mode, along with a 95% credible interval to characterize uncertainty. In this paper, we aim to extend these ideas to develop summary tools for the posterior on partitions. In particular, we seek to answer the two questions: 1) What is an appropriate point estimate of the partition based on the posterior? 2) Can we construct a 95% credible region around this point estimate to characterize our uncertainty?

We first focus on the problem of finding an appropriate point estimate. A simple solution is to use the posterior mode. If the marginal likelihood of the data given the partition, that is with all mixture component parameters integrated out, and the prior of the partition are available in closed form, the posterior mode can be estimated based on the MCMC output by the sampled partition which maximizes the non-normalized posterior. In practice, a closed form for the marginal likelihood or prior is often unavailable, specifically, if conjugate priors for the component specific parameters do not exist

or are not utilized or hyperpriors are assigned to any hyperparameters. More generally, the posterior mode can be found by reporting the partition visited most frequently in the sampler. Yet this approach can be problematic, as producing reliable frequency counts is intractable due to the huge dimension of the partition space. In fact, in many examples, the MCMC chain does not visit a partition more than once. To overcome this, alternative search techniques have been developed to locate the posterior mode (Heller and Ghahramani (2005), Heard et al. (2006), Dahl (2009), Raykov et al. (2014)). However, it is well-known that the mode can be unrepresentative of the center of a distribution.

Alternative methods have been proposed based on the posterior similarity matrix. For a sample size of N, the elements of this N by N matrix represent the probability that two data points are in the same cluster, which can be estimated by the proportion of MCMC samples that cluster the two data points together. Then, classical hierarchical or partitioning algorithms are applied based on the similarity matrix (Medvedovic and Sivaganesan (2002), Medvedovic et al. (2004), Rasmussen et al. (2009), Molitor et al. (2010)). These methods have the disadvantage of being ad-hoc.

A more elegant solution is based on decision theory. In this case, one defines a loss function over clusterings. The optimal point estimate is that which minimizes the posterior expectation of the loss function. For example, for a real-valued parameter θ, the optimal point estimate is the posterior mean under the squared error loss $L_2(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$ and the posterior median under the absolute error loss $L_1(\theta, \widehat{\theta}) = |\theta - \widehat{\theta}|$.

The question to answer then becomes what is an appropriate loss function on the space of clusterings. The 0-1 loss function, a simple choice which leads to the posterior mode as the point estimate, is not ideal as it does not take into account the similarity between two clusterings. More general loss functions were developed by Binder (1978), and the so-called Binder's loss, which measures the disagreements in all possible pairs of observations between the true and estimated clusterings, was studied in a Bayesian nonparametric setting by Lau and Green (2007). Alternative loss functions considered in Bayesian nonparametrics can be found in Quintana and Iglesias (2003) and Fritsch and Ickstadt (2009).

In this paper, we propose to use the variation of information developed by Meilă (2007) as a loss function in a Bayesian nonparametric setting. Both the variation of information and Binder's loss possess the desirable properties of being metrics on the space of partitions and being *aligned* with the lattice of partitions. We provide a detailed comparison of these two metrics and discuss the advantages of the variation of information over Binder's loss as a loss function in Bayesian cluster analysis. Additionally, we propose a novel algorithm to locate the optimal partition, taking advantage of the fact that both metrics are aligned on the space of partitions.

Next, to address the problem of characterizing uncertainty around the point estimate, we propose to construct a credible ball around the point estimate. As both Binder's loss and the variation of information are metrics on the partition space, we can easily construct such a ball. Interestingly, the two metrics can produce very different credible balls, and we discuss this in detail. In existing literature, quantifications of uncertainty include reporting a heat map of the estimated posterior similarity matrix.

However, there is no precise quantification of how much uncertainty is represented by the posterior similarity matrix, and in a comparison with the 95% credible balls, we find that the uncertainty is under-represented by the posterior similarity matrix. Finally, we provide an algorithm to construct the credible ball and discuss ways to depict or report it.

The paper is organized as follows. Section 2 provides a review of Bayesian nonparametric clustering and existing point estimates of the clustering structure from a decision theoretic approach. In Section 3, we give a detailed comparison of two loss functions, Binder's loss and the variation of information, pointing out advantages of the latter. The optimal point estimate under the variation of information is derived in Section 4 and a novel algorithm to locate the optimal partition is proposed. In Section 5, we construct a credible ball around the point estimate to characterize posterior uncertainty and discuss how to compute and depict it. Finally, simulated and real examples are provided in Section 6.

2 Review

This section provides a review of Bayesian nonparametric clustering models and existing point estimates of the clustering in literature.

2.1 Bayesian nonparametric clustering

Mixture models are one of the most popular modeling tools in Bayesian nonparametrics. The data is assumed conditionally i.i.d. with density

$$f(y|P) = \int K(y|\theta)dP(\theta),$$

where $K(y|\theta)$ is a specified parametric density on the sample space with mixing parameter $\theta \in \Theta$ and P is a probability measure on Θ. In a Bayesian setting, the model is completed with a prior on the unknown parameter, which in this case, is the unknown mixing measure. In the most general setting, this parameter P can be any probability measure on Θ, requiring a nonparametric prior. Typically the nonparametric prior has discrete realizations almost surely (a.s.) with

$$P = \sum_{j=1}^{\infty} w_j \delta_{\theta_j} \text{ a.s.},$$

where it is often assumed that the weights (w_j) and atoms (θ_j) are independent and the θ_j are i.i.d. from some base measure P_0. Thus, the density is modeled with a countably infinite mixture model

$$f(y|P) = \sum_{j=1}^{\infty} w_j K(y|\theta_j).$$

Since P is discrete a.s., this model induces a latent partitioning $\mathbf{c}$ of the data where two data points belong to the same cluster if they are generated from the same mixture

component. The partition can be represented by $\mathbf{c} = (C_1, \ldots, C_{k_N})$, where C_j contains the indices of data points in the j^{th} cluster and k_N is the number of clusters in the sample of size N. Alternatively, the partition can be represented by $\mathbf{c} = (c_1, \ldots, c_N)$, where $c_n = j$ if the n^{th} data point is in the j^{th} cluster.

A key difference with finite mixture models is that the number of mixture components is infinite; this allows the data to determine the number of clusters k_N present in the data, which can grow unboundedly with the data. Letting $\mathbf{y}_j = \{y_n\}_{n \in C_j}$, the marginal likelihood for the data $y_{1:N}$ given the partition is

$$f(y_{1:N}|\mathbf{c}) = \prod_{j=1}^{k_N} m(\mathbf{y}_j) = \prod_{j=1}^{k_N} \int \prod_{n \in C_j} K(y_n|\theta) dP_0(\theta).$$

The posterior of the partition, which reflects our beliefs and uncertainty in the clustering given the data, is simply proportional to the prior times the marginal likelihood

$$p(\mathbf{c}|y_{1:N}) \propto p(\mathbf{c}) \prod_{j=1}^{k_N} m(\mathbf{y}_j), \tag{1}$$

where the prior of the partition is obtained from the selected prior on the mixing measure. For example, a Dirichlet process prior (Ferguson (1973)) for P with mass parameter α corresponds to

$$p(\mathbf{c}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \alpha^{k_N} \prod_{j=1}^{k_N} \Gamma(n_j),$$

where $n_j = |C_j|$ is the number of data points in cluster j. Various other priors developed in Bayesian nonparametric literature can be considered for the mixing measure P, such as the Pitman–Yor process (Pitman and Yor (1997)), also known as the two-parameter Poisson–Dirichlet process, or the normalized generalized Gamma process or more generally, a prior within the class of normalized completely random measures, Poisson–Kingman models (Pitman (2003)), or stick-breaking priors (Ishwaran and James (2001)). See Lijoi and Prünster (2011) for an overview.

In general, the marginal likelihood of the data given the partition or the prior of the partition used to compute the posterior in (1) may not be available in closed form. Moreover, there are

$$S_{N,k} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (k-j)^N,$$

a Stirling number of the second kind, ways to partition the N data points into k groups and

$$B_N = \sum_{k=1}^{N} S_{N,k},$$

a Bell number, possible partitions of the N data points. Even for small N, this number is very large, which makes computation of the posterior intractable for the simplest choice

of prior and likelihood. Thus, MCMC techniques are typically employed, such as the marginal samplers described by Neal (2000) with extensions in Favaro and Teh (2013) for normalized completely random measures and in Lomellí et al. (2016) for σ-stable Poisson–Kingman models; the conditional samplers described in Ishwaran and James (2001), Papaspiliopoulos and Roberts (2008), or Kalli et al. (2011), with extensions in Favaro and Teh (2013) for normalized completely random measures and in Favaro and Walker (2012) for σ-stable Poisson–Kingman models; or the recently introduced class of hybrid samplers for σ-stable Poisson–Kingman models in Lomellí et al. (2015). These algorithms produce approximate samples $(\mathbf{c}^m)_{m=1}^M$ from the posterior (1). Clearly, describing all the posterior samples is infeasible, and our aim is to develop appropriate summary tools to characterize the posterior.

Extensions of Bayesian nonparametric mixture models are numerous and allow one to model increasingly complex data. These include extensions for partially exchangeable data (Teh et al. (2006)), inclusion of covariates (MacEachern (2000)), time dependent data (Griffin and Steel (2006)), and spatially dependent data (Duan et al. (2007)) to name a few. See Müller and Quintana (2004) and Dunson (2010) for an overview. These extensions also induce latent clusterings of the observations, and the summary tools developed here are applicable for these settings as well.

2.2 Point estimation for clustering

Firstly, we seek a point estimate of the clustering that is representative of the posterior, which may be of direct interest to the researcher or, more generally, important for understanding the behavior of the posterior. From decision theory, a point estimate is obtained by specifying a loss function $L(\mathbf{c}, \widehat{\mathbf{c}})$, which measures the loss of estimating the true clustering $\mathbf{c}$ with $\widehat{\mathbf{c}}$. Since the true clustering is unknown, the loss is averaged across all possible true clusterings, where the loss associated to each potential true clustering is weighted by its posterior probability. The point estimate $\mathbf{c}^*$ corresponds to the estimate which minimizes the posterior expected loss,

$$\mathbf{c}^* = \underset{\widehat{\mathbf{c}}}{\operatorname{argmin}}\, \mathbb{E}[L(\mathbf{c}, \widehat{\mathbf{c}})|y_{1:N}] = \underset{\widehat{\mathbf{c}}}{\operatorname{argmin}} \sum_{\mathbf{c}} L(\mathbf{c}, \widehat{\mathbf{c}}) p(\mathbf{c}|y_{1:N}).$$

A simple choice for the loss function is the 0-1 loss, $L_{0-1}(\mathbf{c}, \widehat{\mathbf{c}}) = \mathbf{1}(\mathbf{c} \neq \widehat{\mathbf{c}})$, which assumes a loss of 0 if the estimate is equal to the truth and a loss of 1 otherwise. Under the 0-1 loss, the optimal point estimate is the posterior mode. However, this loss function is unsatisfactory because it doesn't take into account similarity between two clusterings; a partition which differs from the truth in the allocation of only one observation is penalized the same as a partition which differs from the truth in the allocation of many observations. Moreover, it is well-known that the mode can be unrepresentative of the center of a distribution. Thus, more general loss functions are needed.

However, constructing a more general loss is not straightforward because, as pointed out by Binder (1978), the loss function should satisfy basic principles such as invariance to permutations of the data point indices and invariance to permutations of the cluster labels for both the true and estimated clusterings. Binder notes that this first condition

implies that the loss is a function of the counts $n_{ij} = |C_i \cap \widehat{C}_j|$, which is the cardinality of the intersection between C_i, the set of data point indices in cluster i under $\mathbf{c}$, and $\widehat{C}_j$, the set of data point indices in cluster j under $\widehat{\mathbf{c}}$, for $i = 1, \ldots, k_N$ and $j = 1, \ldots, \widehat{k}_N$; the notation k_N and $\widehat{k}_N$ represents the number of clusters in $\mathbf{c}$ and $\widehat{\mathbf{c}}$, respectively. He explores loss functions satisfying these principles, starting with simple functions of the counts n_{ij}. The so-called Binder's loss is a quadratic function of the counts, which for all possible pairs of observations, penalizes the two errors of allocating two observations to different clusters when they should be in the same cluster or allocating them to the same cluster when they should be in different clusters:

$$\mathrm{B}(\mathbf{c}, \widehat{\mathbf{c}}) = \sum_{n<n'} l_1 \mathbf{1}(c_n = c_{n'}) \mathbf{1}(\widehat{c}_n \neq \widehat{c}_{n'}) + l_2 \mathbf{1}(c_n \neq c_{n'}) \mathbf{1}(\widehat{c}_n = \widehat{c}_{n'}).$$

If the two types of errors are penalized equally, $l_1 = l_2 = 1$, then

$$\mathrm{B}(\mathbf{c}, \widehat{\mathbf{c}}) = \frac{1}{2} \left(\sum_{i=1}^{k_N} n_{i+}^2 + \sum_{j=1}^{\widehat{k}_N} n_{+j}^2 - 2 \sum_{i=1}^{k_N} \sum_{j=1}^{\widehat{k}_N} n_{ij}^2 \right),$$

where $n_{i+} = \sum_j n_{ij}$ and $n_{+j} = \sum_i n_{ij}$. Under Binder's loss with $l_1 = l_2$, the optimal partition $\mathbf{c}^*$ is the partition $\mathbf{c}$ which minimizes

$$\sum_{n<n'} \left| \mathbf{1}(c_n = c_{n'}) - p_{n\,n'} \right|,$$

or equivalently, the partition $\mathbf{c}$ which minimizes

$$\sum_{n<n'} \left(\mathbf{1}(c_n = c_{n'}) - p_{n\,n'} \right)^2, \tag{2}$$

where $p_{n\,n'} = P(c_n = c_{n'} | y_{1:N})$ is the posterior probability that two observations n and n' are clustered together. This loss function was first studied in Bayesian nonparametrics by Lau and Green (2007). We note that in earlier work Dahl (2006) considered minimization of (2) but without the connection to Binder's loss and the decision theoretic approach.

Binder's loss counts the total number of disagreements, D, in the $\binom{N}{2}$ possible pairs of observations. The Rand index (Rand (1971)), a cluster comparison criterion, is defined as the number of agreements, A, in all possible pairs divided by the total number of possible pairs. Since $D + A = \binom{N}{2}$, Binder's loss and the Rand index, denoted $\mathrm{R}(\mathbf{c}, \widehat{\mathbf{c}})$, are related:

$$\mathrm{B}(\mathbf{c}, \widehat{\mathbf{c}}) = (1 - \mathrm{R}(\mathbf{c}, \widehat{\mathbf{c}})) \binom{N}{2},$$

and the point estimate obtained from minimizing the posterior expected Binder's loss is equivalent to the point estimate obtained from maximizing the posterior expected Rand's index. Motivated by this connection, Fritsch and Ickstadt (2009) consider maximizing the adjusted Rand index, introduced by Hubert and Arabie (1985) to correct the Rand index for chance. An alternative loss function is explored by Quintana and Iglesias (2003) specifically for the problem of outlier detection.

3 A comparison of the variation of information and Binder's loss

Meilă (2007) introduces the *variation of information* (VI) for cluster comparison, which is constructed from information theory and compares the information in two clusterings with the information shared between the two clusterings. More formally, the VI is defined as

$$\begin{aligned}\mathrm{VI}(\mathbf{c},\widehat{\mathbf{c}}) &= \mathrm{H}(\mathbf{c}) + \mathrm{H}(\widehat{\mathbf{c}}) - 2\mathrm{I}(\mathbf{c},\widehat{\mathbf{c}}) \\ &= -\sum_{i=1}^{k_N} \frac{n_{i+}}{N} \log\left(\frac{n_{i+}}{N}\right) - \sum_{j=1}^{\widehat{k}_N} \frac{n_{+j}}{N} \log\left(\frac{n_{+j}}{N}\right) - 2\sum_{i=1}^{k_N}\sum_{j=1}^{\widehat{k}_N} \frac{n_{ij}}{N} \log\left(\frac{n_{ij}N}{n_{i+}n_{+j}}\right),\end{aligned}$$

where log denotes log base 2. The first two terms represent the entropy of the two clusterings, which measures the uncertainty in bits of the cluster allocation of an unknown randomly chosen data point given a particular clustering of the data points. The last term is the mutual information between the two clusterings and measures the reduction in the uncertainty of the cluster allocation of a data point in $\mathbf{c}$ when we are told its cluster allocation in $\widehat{\mathbf{c}}$. The VI ranges from 0 to $\log(N)$. A review of extensions of the VI to normalize or correct for chance are discussed in Vinh et al. (2010). However, some desirable properties of the VI are lost under these extensions.

In this paper, we propose to use the VI as a loss function. Note that since $\mathrm{I}(\mathbf{c},\widehat{\mathbf{c}}) = \mathrm{H}(\mathbf{c}) + \mathrm{H}(\widehat{\mathbf{c}}) - \mathrm{H}(\mathbf{c},\widehat{\mathbf{c}})$, we can write

$$\begin{aligned}\mathrm{VI}(\mathbf{c},\widehat{\mathbf{c}}) &= \mathrm{H}(\mathbf{c}) + \mathrm{H}(\widehat{\mathbf{c}}) - 2\mathrm{H}(\mathbf{c}) - 2\mathrm{H}(\widehat{\mathbf{c}}) + 2\mathrm{H}(\mathbf{c},\widehat{\mathbf{c}}), \\ &= -\mathrm{H}(\mathbf{c}) - \mathrm{H}(\widehat{\mathbf{c}}) + 2\mathrm{H}(\mathbf{c},\widehat{\mathbf{c}}), \\ &= \sum_{i=1}^{k_N} \frac{n_{i+}}{N} \log\left(\frac{n_{i+}}{N}\right) + \sum_{j=1}^{\widehat{k}_N} \frac{n_{+j}}{N} \log\left(\frac{n_{+j}}{N}\right) - 2\sum_{i=1}^{k_N}\sum_{j=1}^{\widehat{k}_N} \frac{n_{ij}}{N} \log\left(\frac{n_{ij}}{N}\right).\end{aligned}$$

We provide a detailed comparison with an N-invariant version of Binder's loss, defined as

$$\tilde{\mathrm{B}}(\mathbf{c},\widehat{\mathbf{c}}) = \frac{2}{N^2}\mathrm{B}(\mathbf{c},\widehat{\mathbf{c}}) = \sum_{i=1}^{k_N} \left(\frac{n_{i+}}{N}\right)^2 + \sum_{j=1}^{\widehat{k}_N} \left(\frac{n_{+j}}{N}\right)^2 - 2\sum_{i=1}^{k_N}\sum_{j=1}^{\widehat{k}_N} \left(\frac{n_{ij}}{N}\right)^2.$$

Both loss functions are considered N-invariant as they only depend on N through the proportions n_{ij}/N. We focus on these two loss functions as they satisfy several desirable properties.

The first important property is that both VI and $\tilde{\mathrm{B}}$ are metrics on the space of partitions.

Property 1. *Both VI and $\tilde{B}$ are metrics on the space of partitions.*

A proof for VI can be found in Meilă (2007). For $\tilde{\mathrm{B}}$, the proof results from the fact that $\tilde{\mathrm{B}}$ can be derived as the Hamming distance between the binary representation of the clusterings.

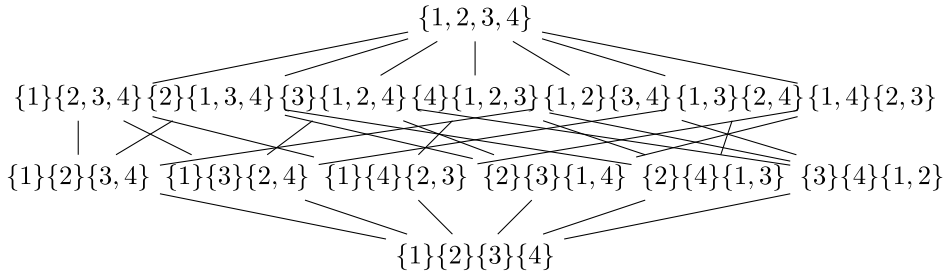


Figure 1: Hasse diagram for the lattice of partitions with a sample of size $N = 4$. A line is drawn from $\mathbf{c}$ up to $\widehat{\mathbf{c}}$ when $\mathbf{c}$ is covered by $\widehat{\mathbf{c}}$.

The next properties involve first viewing the space of partitions as a partially ordered set. In particular, consider the space of partitions $\mathbf{C}$ and the binary relation $\leq$ on $\mathbf{C}$ defined by set containment, i.e. for $\mathbf{c}, \widehat{\mathbf{c}} \in \mathbf{C}$, $\mathbf{c} \leq \widehat{\mathbf{c}}$ if for all $i = 1, \ldots, k_N$, $C_i \subseteq \widehat{C}_j$ for some $j \in \{1, \ldots, \widehat{k}_N\}$. The partition space $\mathbf{C}$ equipped with $\leq$ is a partially ordered set.

For any $\mathbf{c}, \widehat{\mathbf{c}} \in \mathbf{C}$, $\mathbf{c}$ is *covered* by $\widehat{\mathbf{c}}$, denoted $\mathbf{c} \prec \widehat{\mathbf{c}}$, if $\mathbf{c} < \widehat{\mathbf{c}}$ and there is no $\widehat{\widehat{\mathbf{c}}} \in \mathbf{C}$ such that $\mathbf{c} < \widehat{\widehat{\mathbf{c}}} < \widehat{\mathbf{c}}$. This covering relation is used to define the *Hasse diagram*, where the elements of $\mathbf{C}$ are represented as nodes of a graph and a line is drawn from $\mathbf{c}$ up to $\widehat{\mathbf{c}}$ when $\mathbf{c} \prec \widehat{\mathbf{c}}$. An example of the Hasse diagram for $N = 4$ is depicted in Figure 1.

The space of partitions possesses an even richer structure; it forms a lattice. This follows from the fact that every pair of partitions has a *greatest lower bound* and *least upper bound*; for a subset $\mathbf{S} \subseteq \mathbf{C}$, an element $\mathbf{c} \in \mathbf{C}$ is an upper bound for $\mathbf{S}$ if $\mathbf{s} \leq \mathbf{c}$ for all $\mathbf{s} \in \mathbf{S}$, and $\mathbf{c} \in \mathbf{C}$ is the least upper bound for $\mathbf{S}$, denoted $\mathbf{c} = \text{l.u.b.}(\mathbf{S})$, if $\mathbf{c}$ is an upper bound for $\mathbf{S}$ and $\mathbf{c} \leq \mathbf{c}'$ for all upper bounds $\mathbf{c}'$ of $\mathbf{S}$. A lower bound and the greatest lower bound for a subset $\mathbf{S} \subseteq \mathbf{C}$ are similarly defined, the latter denoted by g.l.b.($\mathbf{S}$). We define the operators $\wedge$, called the meet, and $\vee$, called the join, as $\mathbf{c} \wedge \widehat{\mathbf{c}} = \text{g.l.b.}(\mathbf{c}, \widehat{\mathbf{c}})$ and $\mathbf{c} \vee \widehat{\mathbf{c}} = \text{l.u.b.}(\mathbf{c}, \widehat{\mathbf{c}})$. Following the conventions of lattice theory, we will use $\mathbf{1}$ to denote the greatest element of the lattice of partitions, i.e. the partition with every observation in one cluster $\mathbf{c} = (\{1, \ldots, N\})$, and $\mathbf{0}$ to denote the least element of the lattice of partitions, i.e. the partition with every observation in its own cluster $\mathbf{c} = (\{1\}, \ldots, \{N\})$. See Nation (1991) for more details on lattice theory and the Supplementary Material (Wade and Ghahramani, 2017) for specific details on the lattice of partitions.

A desirable property is that both VI and $\tilde{\text{B}}$ are *aligned* with the lattice of partitions. Specifically, both metrics are *vertically aligned* in the Hasse diagram; if $\widehat{\widehat{\mathbf{c}}}$ is connected up to $\widehat{\mathbf{c}}$ and $\widehat{\mathbf{c}}$ is connected up to $\mathbf{c}$, then the distance between $\widehat{\widehat{\mathbf{c}}}$ and $\mathbf{c}$ is the vertical sum of the distances between $\widehat{\widehat{\mathbf{c}}}$ and $\widehat{\mathbf{c}}$ and between $\widehat{\mathbf{c}}$ and $\mathbf{c}$ (see Property 2). And, both metrics are *horizontally aligned*; the distance between any two partitions is the horizontal sum of the distances between each partition and the meet of the two partitions (see Property 3).

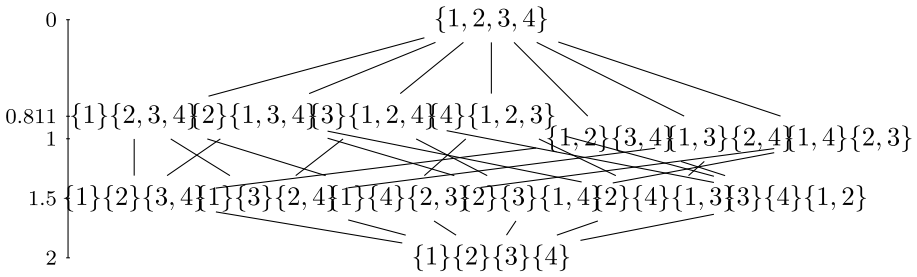


Figure 2: Hasse diagram stretched by VI with a sample of size $N = 4$. Note $2-\frac{3}{4}\log(3) \approx 0.811$. From the VI stretched Hasse diagram, we can determine the distance between any two partitions. Example: if $\mathbf{c} = (\{1,2\},\{3,4\})$ and $\widehat{\mathbf{c}} = (\{1\},\{3\},\{2,4\})$, then $\mathbf{c} \wedge \widehat{\mathbf{c}} = (\{1\},\{2\},\{3\},\{4\})$ and $d(\mathbf{c},\widehat{\mathbf{c}}) = d(\mathbf{c}\wedge\widehat{\mathbf{c}},\mathbf{1}) - d(\mathbf{c},\mathbf{1}) + d(\mathbf{c}\wedge\widehat{\mathbf{c}},\mathbf{1}) - d(\widehat{\mathbf{c}},\mathbf{1}) = 2 - 1 + 2 - 1.5 = 1.5$.

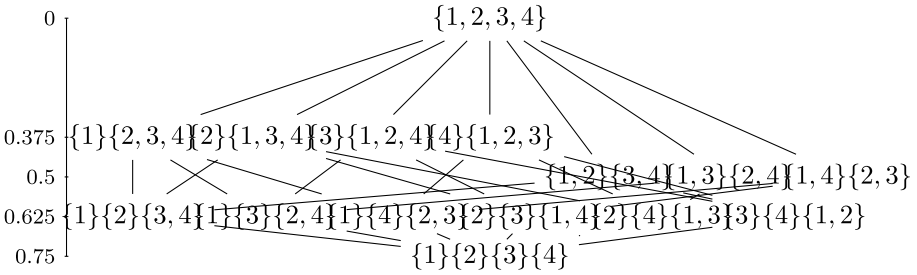


Figure 3: Hasse diagram stretched by $\tilde{\mathrm{B}}$ with a sample of size $N = 4$. From the $\tilde{\mathrm{B}}$ stretched Hasse diagram, we can determine the distance between any two partitions. Example: if $\mathbf{c} = (\{1,2\},\{3,4\})$ and $\widehat{\mathbf{c}} = (\{1\},\{3\},\{2,4\})$, then $\mathbf{c}\wedge\widehat{\mathbf{c}} = (\{1\},\{2\},\{3\},\{4\})$ and $d(\mathbf{c},\widehat{\mathbf{c}}) = d(\mathbf{c}\wedge\widehat{\mathbf{c}},\mathbf{1}) - d(\mathbf{c},\mathbf{1}) + d(\mathbf{c}\wedge\widehat{\mathbf{c}},\mathbf{1}) - d(\widehat{\mathbf{c}},\mathbf{1}) = 0.75 - 0.5 + 0.75 - 0.625 = 0.375$.

Property 2. *For both VI and $\tilde{B}$, if $\mathbf{c} \geq \widehat{\mathbf{c}} \geq \widehat{\widehat{\mathbf{c}}}$, then*

$$d(\mathbf{c},\widehat{\widehat{\mathbf{c}}}) = d(\mathbf{c},\widehat{\mathbf{c}}) + d(\widehat{\mathbf{c}},\widehat{\widehat{\mathbf{c}}}).$$

Property 3. *For both VI and $\tilde{B}$,*

$$d(\mathbf{c},\widehat{\mathbf{c}}) = d(\mathbf{c},\widehat{\mathbf{c}}\wedge\mathbf{c}) + d(\widehat{\mathbf{c}},\widehat{\mathbf{c}}\wedge\mathbf{c}).$$

Proofs can be found in the Supplementary Material. These two properties imply that if the Hasse diagram is stretched to reflect the distance between any partition and $\mathbf{1}$, the distance between any two partitions can be easily determined from the *stretched Hasse diagram*. Figures 2 and 3 depict the Hasse diagram for $N = 4$ in Figure 1 stretched according to VI and $\tilde{\mathrm{B}}$ respectively.

From the stretched Hasse diagram, we gain several insights into the similarities and differences between the two metrics. An evident difference is the scale of the two diagrams.

Property 4. *A distance on partitions satisfying Properties 2 and 3 has the property that for any two partitions* $\mathbf{c}$ *and* $\widehat{\mathbf{c}}$*,*

$$d(\mathbf{c}, \widehat{\mathbf{c}}) \leq d(\mathbf{1}, \mathbf{0}).$$

Thus,

$$VI(\mathbf{c}, \widehat{\mathbf{c}}) \leq \log(N) \quad \text{and} \quad \tilde{B}(\mathbf{c}, \widehat{\mathbf{c}}) \leq 1 - \frac{1}{N}.$$

A proof can be found in the Supplementary Material. In both cases, the bound on the distance between two clusterings depends on the sample size N. However, the behavior of this bound is very different; for VI, it approaches infinity as $N \to \infty$, and for $\tilde{\text{B}}$, it approaches one as $N \to \infty$. As N grows, the number of total partitions B_N increases drastically. Thus, it is sensible that the bound on the metric grows as the size of the space grows. In particular, $\mathbf{1}$ and $\mathbf{0}$ become more distant as $N \to \infty$, as there is an increasing number, $B_N - 2$, of partitions between these two extremes; for $\tilde{\text{B}}$, the loss of estimating one of these extremes with the other approaches the fixed number one, while for VI, the loss approaches infinity.

From the stretched Hasse diagram in Figures 2 and 3, we can determine the closest partitions to any $\mathbf{c}$. For example, the closest partitions to $\mathbf{1}$ are the partitions which split $\mathbf{1}$ into two clusters, one singleton and one containing all other observations; and the closest partitions to $(\{1\}, \{2\}, \{3, 4\})$ are the partition which merges the two smallest clusters $(\{1, 2\}, \{3, 4\})$ and the partition which splits the cluster of size two $(\{1\}, \{2\}, \{3\}, \{4\})$.

Property 5. *For both metrics VI and* $\tilde{B}$*, the closest partitions to a partition* $\mathbf{c}$ *are:*

- *if* $\mathbf{c}$ *contains at least two clusters of size one and at least one cluster of size two, the partitions which merge any two clusters of size one and the partitions which split any cluster of size two.*
- *if* $\mathbf{c}$ *contains at least two clusters of size one and no clusters of size two, the partitions which merge any two clusters of size one.*
- *if* $\mathbf{c}$ *contains at most one cluster of size one, the partitions which split the smallest cluster of size greater than one into a singleton and a cluster with the remaining observations of the original cluster.*

A proof can be found in the Supplementary Material. This property characterizes the set of estimated partitions which are given the smallest loss. Under both loss functions, the smallest loss of zero occurs when the estimated partition is equal to the truth. Otherwise, the smallest loss occurs when the estimated clustering differs from the truth by merging two singleton clusters or splitting a cluster of size two, or, if neither is possible, splitting the smallest cluster of size n into a singleton and a cluster of size $n - 1$. We further note that the loss of estimating the true clustering with a clustering which merges two singletons or splits a cluster of size two, is $\frac{2}{N}$ and $\frac{2}{N^2}$ for VI and $\tilde{\text{B}}$ respectively, which converges to 0 as $N \to \infty$ for both metrics, but at a faster rate for $\tilde{\text{B}}$.

Next, we note that the Hasse diagram stretched by $\tilde{\text{B}}$ in Figure 3 appears asymmetric, in the sense that $\mathbf{1}$ is more separated from the others when compared to the Hasse diagram stretched by VI in Figure 2.

Property 6. *Suppose N is divisible by k, and let $\mathbf{c}_k$ denote a partition with k clusters of equal size N/k.*

$$\tilde{B}(\mathbf{1}, \mathbf{c}_k) = 1 - \frac{1}{k} > \frac{1}{k} - \frac{1}{N} = \tilde{B}(\mathbf{0}, \mathbf{c}_k).$$

$$VI(\mathbf{1}, \mathbf{c}_k) = \log(k) \leq \log(N) - \log(k) = VI(\mathbf{0}, \mathbf{c}_k), \quad \textit{for } k \leq \sqrt{N},$$

and

$$VI(\mathbf{1}, \mathbf{c}_k) = \log(k) \geq \log(N) - \log(k) = VI(\mathbf{0}, \mathbf{c}_k), \quad \textit{for } k \geq \sqrt{N}.$$

Property 6 reflects the asymmetry apparent in Figure 3. In particular, for $\tilde{\text{B}}$, a partition with two clusters of equal size $\mathbf{c}_2$ will always be closer to the extreme $\mathbf{0}$ of each data point in its own cluster than the extreme $\mathbf{1}$ of everyone in one cluster. However, as the sample size increases, $\mathbf{c}_2$ becomes equally distant between the two extremes. For all other values of k, the extreme $\mathbf{0}$ will always be closer. This behavior is counter-intuitive for a loss function on clusterings. VI is much more sensible in this regard. If $k = \sqrt{N}$, $\mathbf{0}$ and $\mathbf{1}$ are equally good estimates of $\mathbf{c}_k$. For $k < \sqrt{N}$, $\mathbf{c}_k$ is better estimated by $\mathbf{1}$ and for $k > \sqrt{N}$, $\mathbf{c}_k$ is better estimated by $\mathbf{0}$; as the sample size increases, these preferences become stronger. In particular, note that loss of estimating $\mathbf{c}_2$ with $\mathbf{1}$ will always be smaller than estimating it with $\mathbf{0}$ for $N > 4$.

Additionally, we observe from Figure 3 that the partitions with two clusters of sizes one and three are equally distant between the two extremes under $\tilde{\text{B}}$. The following property generalizes this observation.

Property 7. *Suppose N is an even and square integer. Then, the partitions with two clusters of sizes $n = \frac{1}{2}(N - \sqrt{N})$ and $N - n$ are equally distant from $\mathbf{1}$ and $\mathbf{0}$ under $\tilde{B}$.*

This property is unappealing for a loss function, as it states that the loss of estimating a partition consisting of two clusters of sizes $\frac{1}{2}(N - \sqrt{N})$ and $\frac{1}{2}(N + \sqrt{N})$ with the partition of only one cluster or with the partition of all singletons is the same. Intuitively, however, $\mathbf{1}$ is a better estimate. The behavior of VI is much more reasonable, as partitions with two clusters will always be better estimated by $\mathbf{1}$ than $\mathbf{0}$ for $N > 4$ and partitions with $\sqrt{N}$ clusters of equal size are equally distant from $\mathbf{0}$ and $\mathbf{1}$.

Finally, we note that as both VI and $\tilde{\text{B}}$ are metrics on the space of clusterings, we can construct a ball around $\mathbf{c}$ of size ϵ, defined as:

$$B_\epsilon(\mathbf{c}) = \{\widehat{\mathbf{c}} \in \mathbf{C} : d(\mathbf{c}, \widehat{\mathbf{c}}) \leq \epsilon\}.$$

From Property 5, the smallest non-trivial ball will be the same for the two metrics. When considering the next smallest ball, differences emerge; a detailed example is provide in the Supplementary Material. In the authors' opinions, the VI ball more closely reflects our intuition of the closest set of partitions to $\mathbf{c}$.

4 Point estimation via the variation of information

As detailed in the previous section, both VI and $\tilde{\text{B}}$ share several desirable properties including being aligned with the lattice of partitions and coinciding in the smallest non-trivial ball around any clustering. However, in our comparison, differences also emerged. Particularly, we find that $\tilde{\text{B}}$ exhibits some peculiar asymmetries, preferring to split clusters over merging, and we find that the VI ball more closely reflects our intuition of the neighborhood of a partition. In light of this, we propose to use VI as a loss function in Bayesian cluster analysis. Under the VI, the optimal partition $\mathbf{c}^*$ is

$$\begin{aligned}\mathbf{c}^* &= \underset{\widehat{\mathbf{c}}}{\text{argmin}}\, \mathbb{E}[\text{VI}(\mathbf{c}, \widehat{\mathbf{c}})|\mathcal{D}] \\ &= \underset{\widehat{\mathbf{c}}}{\text{argmin}} \sum_{n=1}^{N} \log\Big(\sum_{n'=1}^{N} \mathbf{1}(\widehat{c}_{n'} = \widehat{c}_n)\Big) - 2\sum_{n=1}^{N} \mathbb{E}\Big[\log\Big(\sum_{n'=1}^{N} \mathbf{1}(c_{n'} = c_n, \widehat{c}_{n'} = \widehat{c}_n)\Big)|\mathcal{D}\Big], \quad (3)\end{aligned}$$

with $\mathcal{D}$ denoting the data. For a given $\widehat{\mathbf{c}}$, the second term in (3) can be approximated based on the MCMC output, and evaluating this term is of order $O(MN^2)$ (recall M is the number of MCMC samples). This may be computationally demanding if the number of MCMC samples is large and if (3) must be evaluated for a large number of $\widehat{\mathbf{c}}$. Alternatively, one can use Jensen's inequality, swapping the log and expectation, to obtain a lower bound on the expected loss which is computationally more efficient to evaluate:

$$\underset{\widehat{\mathbf{c}}}{\text{argmin}} \sum_{n=1}^{N} \log\Big(\sum_{n'=1}^{N} \mathbf{1}(\widehat{c}_{n'} = \widehat{c}_n)\Big) - 2\sum_{n=1}^{N} \log\Big(\sum_{n'=1}^{N} P(c_{n'} = c_n|\mathcal{D})\mathbf{1}(\widehat{c}_{n'} = \widehat{c}_n)\Big). \quad (4)$$

Similar to minimization of the posterior expected Binder's loss, minimization of (4) only depends on the posterior through the posterior similarity matrix, which can be pre-computed based on the MCMC output. In this case, computational complexity for a given $\widehat{\mathbf{c}}$ is reduced to $O(N^2)$.

Due to the huge dimensions of the partition space, computing the lower bound in (4) for every possible $\widehat{\mathbf{c}}$ is practically impossible. A simple technique to find the optimal partition $\mathbf{c}^*$ restricts the search space to some smaller space of partitions. The **R** package 'mcclust' (Fritsch (2012)), which contains tools for point estimation in Bayesian cluster analysis and cluster comparison, includes a function `minbinder()` that finds the partition minimizing the poster expected Binder's loss among the subset of partitions 1) visited in the MCMC chain or 2) explored in a hierarchical clustering algorithm with a distance of $1 - P(c_n = c_{n'}|\mathcal{D})$ and average or complete linkage. An alternative search algorithm developed in Lau and Green (2007), which is based on binary integer programming, is also implemented.

We propose a greedy search algorithm to locate the optimal partition $\mathbf{c}^*$ based on the Hasse diagram, which can be used for both VI and $\tilde{\text{B}}$. In particular, given some partition $\widehat{\mathbf{c}}$, we consider the l closest partitions that cover $\widehat{\mathbf{c}}$ and the l closest partitions that $\widehat{\mathbf{c}}$ covers. Here, the distance used to determine the closest partitions corresponds to the selected loss of VI or $\tilde{\text{B}}$. Next, the posterior expected loss $\mathbb{E}[L(\mathbf{c}, \widehat{\widehat{\mathbf{c}}})|\mathcal{D}]$ is computed for all proposed partitions $\widehat{\widehat{\mathbf{c}}}$, and we move in the direction of minimum posterior expected loss, that is the partition $\mathbf{c}'$ with minimal $\mathbb{E}[L(\mathbf{c}, \mathbf{c}')|\mathcal{D}]$ is selected. The algorithm stops when

no reduction in the posterior expected loss is obtained or when a maximum number of iterations has been reached. At each iteration, the computational complexity is $O(lN^2)$.

We have developed an **R** package 'mcclust.ext' (Wade (2015)), expanding upon the 'mcclust' package, that is currently available on the author's website[1] and includes functions `minbinder.ext()` and `minVI()` to find the partition minimizing the posterior expected Binder's loss and VI, respectively. In addition to implementing the search algorithms of `minbinder()` in 'mcclust' described previously, the greedy search algorithm is also included. As is common in greedy search algorithms, results are sensitive to both the starting value of $\widehat{\mathbf{c}}$ and the step size l. In practice, we recommend multiple restarts, for example, at different MCMC samples or the best partition found by the other search algorithms. A larger value of l will allow more exploration and reduce the need for multiple restarts, and we have chosen a default value of $l = 2N$ as this showed good exploration in the examples considered with little sensitivity to the initial value of $\widehat{\mathbf{c}}$. However, for larger datasets, this may be too expensive and multiple restarts with smaller l may be preferred.

An advantage of the greedy search algorithm over simply restricting to partitions visited in the chain is that partitions not explored in the MCMC algorithm can be considered; in fact, in almost all simulated and real examples, the clustering estimate is not among the sampled partitions and results in a lower expected loss than any sampled partition.

5 Credible balls of partitions

To characterize the uncertainty in the point estimate $\mathbf{c}^*$, we propose to construct a credible ball of a given credible level $1-\alpha$, $\alpha \in [0,1]$, defined as

$$B_{\epsilon^*}(\mathbf{c}^*) = \{\mathbf{c} : d(\mathbf{c}^*, \mathbf{c}) \leq \epsilon^*\},$$

where ϵ^* is the smallest $\epsilon \geq 0$ such that $P(B_\epsilon(\mathbf{c}^*)|\mathcal{D}) \geq 1-\alpha$. The credible ball is the smallest ball around $\mathbf{c}^*$ with posterior probability at least $1-\alpha$. It reflects the posterior uncertainty in the point estimate $\mathbf{c}^*$; with probability $1-\alpha$, we believe that the clustering is within a distance of ϵ^* from the point estimate $\mathbf{c}^*$ given the data. It can be defined based on any metric on the space of partitions, such as VI and $\tilde{\text{B}}$. If the smallest non-trivial ball under VI or $\tilde{\text{B}}$ has posterior probability of at least $1-\alpha$, the credible balls under the two metrics will coincide (see Property 5). Typically, however, they will be different.

From the MCMC output, we can obtain an estimate of ϵ^*, and thus the credible ball of level $1-\alpha$. First, the distance between all MCMC samples $\{\mathbf{c}^m\}$ and $\mathbf{c}^*$ is computed. For any $\epsilon \geq 0$,

$$P(B_\epsilon(\mathbf{c}^*)|\mathcal{D}) = \mathbb{E}[\mathbf{1}(d(\mathbf{c}^*, \mathbf{c}) \leq \epsilon)|\mathcal{D}] \approx \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}(d(\mathbf{c}^*, \mathbf{c}^m) \leq \epsilon),$$

and ϵ^* is the smallest $\epsilon \geq 0$ such that $\frac{1}{M}\sum_{m=1}^{M} \mathbf{1}(d(\mathbf{c}^*, \mathbf{c}^m) \leq \epsilon) \geq 1-\alpha$.

[1] https://www2.warwick.ac.uk/fac/sci/statistics/staff/academic-research/wade/.

To characterize the credible ball, we define the *vertical* and *horizontal bounds* of the credible ball. The vertical upper bounds consist of the partitions in the credible ball with the smallest number of clusters that are most distant from $\mathbf{c}^*$. The vertical lower bounds consist of the partitions in the credible ball with the largest number of clusters that are most distant from $\mathbf{c}^*$. The horizontal bounds consist of the partitions in the credible ball that are most distant from $\mathbf{c}^*$. The bounds are defined more formally below, where the notation $k(\mathbf{c})$ is used for the number of clusters in $\mathbf{c}$.

Definition 1 (Vertical upper bounds)**.** *The vertical upper bounds of the credible ball* $B_{\epsilon^*}(\mathbf{c}^*)$*, denoted* $v^u_{\epsilon^*}(\mathbf{c}^*)$*, are defined as*

$$v^u_{\epsilon^*}(\mathbf{c}^*) = \{\mathbf{c} \in B_{\epsilon^*}(\mathbf{c}^*) : k(\mathbf{c}) \leq k(\mathbf{c}') \,\forall\, \mathbf{c}' \in B_{\epsilon^*}(\mathbf{c}^*) \text{ and } d(\mathbf{c}, \mathbf{c}^*) \geq d(\mathbf{c}'', \mathbf{c}^*) \,\forall\, \mathbf{c}'' \in B_{\epsilon^*}(\mathbf{c}^*) \text{ with } k(\mathbf{c}) = k(\mathbf{c}'')\}.$$

Definition 2 (Vertical lower bounds)**.** *The vertical lower bounds of the credible ball* $B_{\epsilon^*}(\mathbf{c}^*)$*, denoted* $v^l_{\epsilon^*}(\mathbf{c}^*)$*, are defined as*

$$v^l_{\epsilon^*}(\mathbf{c}^*) = \{\mathbf{c} \in B_{\epsilon^*}(\mathbf{c}^*) : k(\mathbf{c}) \geq k(\mathbf{c}') \,\forall\, \mathbf{c}' \in B_{\epsilon^*}(\mathbf{c}^*) \text{ and } d(\mathbf{c}, \mathbf{c}^*) \geq d(\mathbf{c}'', \mathbf{c}^*) \,\forall\, \mathbf{c}'' \in B_{\epsilon^*}(\mathbf{c}^*) \text{ with } k(\mathbf{c}) = k(\mathbf{c}'')\}.$$

Definition 3 (Horizontal bounds)**.** *The horizontal bounds of the credible ball* $B_{\epsilon^*}(\mathbf{c}^*)$*, denoted* $h_{\epsilon^*}(\mathbf{c}^*)$*, are defined as*

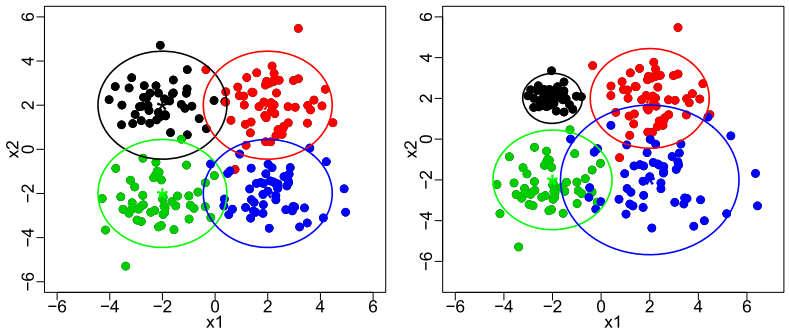
$$h_{\epsilon^*}(\mathbf{c}^*) = \{\mathbf{c} \in B_{\epsilon^*}(\mathbf{c}^*) : d(\mathbf{c}, \mathbf{c}^*) \geq d(\mathbf{c}', \mathbf{c}^*) \,\forall\, \mathbf{c}' \in B_{\epsilon^*}(\mathbf{c}^*)\}.$$

These bounds describe the extremes of the credible ball and with $1 - \alpha$ posterior probability, how different we believe the partition may be from $\mathbf{c}^*$. An example is provided in the Supplementary Material. In practice, we define the vertical and horizontal bounds based on the partitions in the credible ball with positive estimated posterior probability.

In existing literature, quantification of uncertainty in the clustering structure is typically described through a heat map of the estimated posterior similarity matrix. However, as opposed to the credible ball of Bayesian confidence level $1 - \alpha$, there is no precise quantification of how much uncertainty is represented by the posterior similarity matrix. Moreover, in the examples of Section 6, we find that in a comparison with the 95% credible balls, the uncertainty is under-represented by the posterior similarity matrix. Additionally, the credible balls have the added desirable interpretation of characterizing the uncertainty around the point estimate $\mathbf{c}^*$.

6 Examples

We provide both simulated and real examples to compare the point estimates from VI and Binder's loss and describe the credible ball representing uncertainty in the clustering estimate.



(a) Example 1: 4 clusters (b) Example 2: 4 clusters

Figure 4: The data is simulated from a mixture of four normals with locations $(\pm 2, \pm 2)'$ and colored by cluster membership. In (b) components having varying standard deviations.

6.1 Simulated examples

Two datasets of size $n = 200$ are simulated from:

$$X_i \overset{iid}{\sim} \sum_{j=1}^{4} \frac{1}{4} \mathrm{N}\left(\begin{bmatrix} (-1)^{\lfloor \frac{(j-1)}{2} \rfloor} 2 \\ (-1)^{j-1} 2 \end{bmatrix}, \begin{bmatrix} \sigma_j^2 & 0 \\ 0 & \sigma_j^2 \end{bmatrix} \right).$$

In the first example, $\sigma_j = 1$ for all components, while in the second example, components have varying standard deviations; $\sigma_j = 1$ for the two components located in the first and third quadrants, $\sigma_j = 0.5$ in the second quadrant, and $\sigma_j = 1.5$ in the fourth quadrant. The datasets for both examples are depicted in Figure 4 and colored by cluster membership.

We consider a Dirichlet process (DP) mixture model:

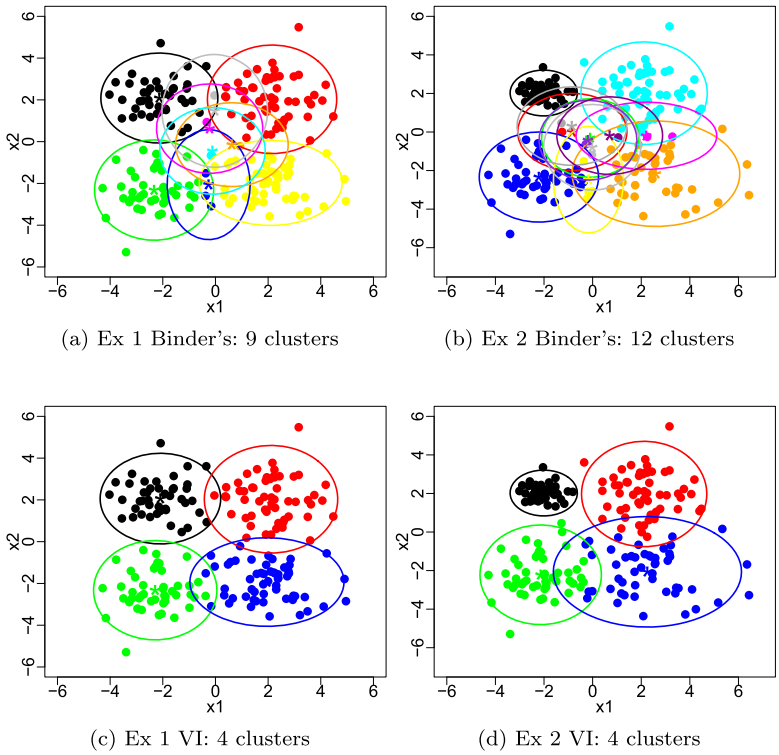
$$X_i | P \overset{iid}{\sim} \int \mathrm{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right) dP(\mu, \Sigma) \quad \text{and} \quad P \sim \mathrm{DP}(\alpha P_0), \tag{5}$$

where $\mu = (\mu_1, \mu_2)'$ and Σ is a diagonal matrix with diagonal elements (σ_1^2, σ_2^2). The base measure of the DP is the conjugate product of normal inverse gamma priors with parameters $(\mu_{0,i}, c_i, a_i, b_i)$ for $i = 1, 2$, i.e. P_0 has density

$$p_0(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \propto \prod_{i=1}^{2} \sqrt{\frac{c_i}{\sigma_i^2}} \exp\left(-\frac{c_i}{2\sigma_i^2} (\mu_i - \mu_{0,i})^2 \right) (\sigma_i^2)^{-a_i - 1} \exp\left(-\frac{b_i}{\sigma_i^2} \right).$$

The parameters were fixed to $\mu_{0,i} = 0$, $c_i = 1/2$, $a_i = 2$, $b_i = 1$ for $i = 1, 2$. The mass parameter α is given a $\mathrm{Gam}(1, 1)$ hyperprior.

A marginal Gibbs sampler is used for inference (Neal (2000)) with 10,000 iterations after a burn in period of 1,000 iterations. Trace plots and autocorrelation plots (not



(a) Ex 1 Binder's: 9 clusters (b) Ex 2 Binder's: 12 clusters

(c) Ex 1 VI: 4 clusters (d) Ex 2 VI: 4 clusters

Figure 5: Clustering estimate with color representing cluster membership for Binder's loss (first row) and VI (second row) with columns corresponding to examples.

shown) suggest convergence. Among partitions sampled in the MCMC, only one is visited twice and all others are visited once in the first example, while no partitions are visited more than once in the second example.

Figure 5 depicts the partition estimate found by the greedy search algorithm for Binder's loss and VI and for both examples (with multiple restarts and the default value of $l = 2N$); colors represent cluster membership with the posterior expected cluster-specific mean and variance represented through stars and ellipses, respectively. Tables in the Supplementary Material provide a comparison of the true partition with the estimates through a cross tabulation of cluster labels. In all examples, the four true clusters are visible; however, Binder's loss creates new small clusters for observations on the border between clusters where cluster membership is uncertain, overestimating the number of clusters. This effect is most extreme for the second example, where the fourth cluster (blue in Figure 4b) has increased overlap with the second and third clusters (red and green in Figure 4b), while the first cluster (black in Figure 4b) with decreased variance is well separated from the other clusters and identified in both estimates.

A further comparison of the true partition with the estimates under Binder's loss and VI, for both examples, is provided in Table 1. As expected, the $\tilde{\text{B}}$ estimate and

	Loss	k_N^*	N_I	$\mathbb{E}[\tilde{\text{B}}\|\mathcal{D}]$	$\tilde{\text{B}}(\mathbf{c}_t, \mathbf{c}^*)$	$\mathbb{E}[\text{VI}_{\text{LB}}\|\mathcal{D}]$	$\mathbb{E}[\text{VI}\|\mathcal{D}]$	$\text{VI}(\mathbf{c}_t, \mathbf{c}^*)$
Ex 1:	$\tilde{\text{B}}$	9	13	**0.062**	0.045	0.545	0.816	0.643
	VI	**4**	**9**	0.064	**0.044**	**0.426**	**0.77**	**0.569**
Ex 2:	$\tilde{\text{B}}$	12	18	**0.088**	0.056	0.846	1.068	0.764
	VI	**4**	**10**	0.093	**0.049**	**0.668**	**0.99**	**0.561**

Table 1: A comparison of the clustering estimate with $\tilde{\text{B}}$ or VI in terms of 1) number of clusters k_N^*; 2) number of data points incorrectly classified, denoted N_I; 3) expected $\tilde{\text{B}}$; 4) $\tilde{\text{B}}$ between the optimal and true clusterings; 5) expected lower bound of VI; 6) expected VI; and 7) VI between the optimal and true clusterings for both examples.

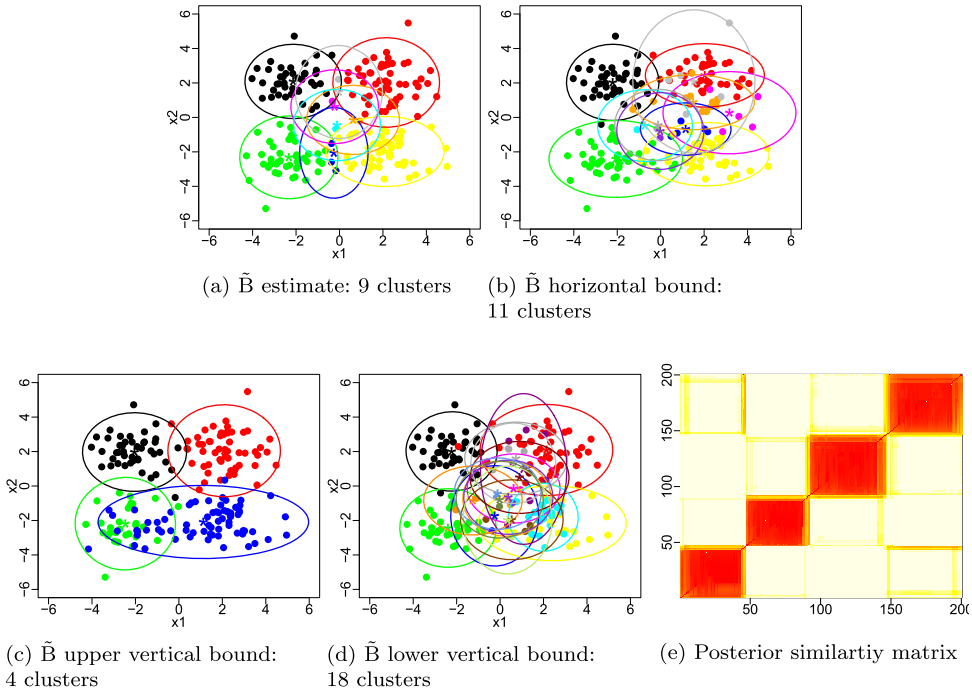
Ex 1	Loss	k_N^*	N_I	$\mathbb{E}[\tilde{\text{B}}\|\mathcal{D}]$	$\tilde{\text{B}}(\mathbf{c}_t, \mathbf{c}^*)$	$\mathbb{E}[\text{VI}_{\text{LB}}\|\mathcal{D}]$	$\mathbb{E}[\text{VI}\|\mathcal{D}]$	$\text{VI}(\mathbf{c}_t, \mathbf{c}^*)$
$N = 200$:	$\tilde{\text{B}}$	9	13	**0.062**	0.045	0.545	0.816	0.643
	VI	**4**	**9**	0.064	**0.044**	**0.426**	**0.77**	**0.569**
$N = 400$:	$\tilde{\text{B}}$	17	31	**0.068**	0.052	0.674	1.0	0.769
	VI	**4**	**18**	0.073	**0.044**	**0.505**	**0.933**	**0.54**
$N = 800$:	$\tilde{\text{B}}$	24	62	**0.068**	0.061	0.615	1.016	0.903
	VI	**4**	**47**	0.069	**0.056**	**0.477**	**0.943**	**0.742**
$N = 1600$:	$\tilde{\text{B}}$	41	93	**0.058**	**0.044**	0.551	0.898	0.719
	VI	**4**	**49**	0.0596	0.045	**0.403**	**0.814**	**0.629**

Table 2: Example 1 with increasing sample size: a comparison of the clustering estimate with $\tilde{\text{B}}$ or VI in terms of 1) number of clusters k_N^*; 2) number of data points incorrectly classified, denoted N_I; 3) expected $\tilde{\text{B}}$; 4) $\tilde{\text{B}}$ between the optimal and true clusterings; 5) expected lower bound of VI; 6) expected VI; and 7) VI between the optimal and true clusterings.

VI estimate achieve the lowest posterior expected loss for $\tilde{\text{B}}$ and VI, respectively, but interestingly, the VI estimate has the smallest distance from the truth for both $\tilde{\text{B}}$ and VI in both examples, with the greatest improvement in the second example. Furthermore, the number of incorrectly classified data points is greater for the $\tilde{\text{B}}$ estimate than the VI estimate.

Additional simulated experiments were performed to analyze the effect of increasing the sample size in the first example. The results are succinctly summarized in Table 2. As the sample size increases, more points are located on the border where cluster membership is uncertain. This results in an increasing number of clusters in the $\tilde{\text{B}}$ estimate (up to 41 clusters for $N = 1600$), while the VI estimate contains only four clusters for all sample sizes. In both estimates, the number of incorrectly classified data points increases with the sample size, however this number is smaller for the VI estimate in all sample sizes, with the difference between this number for Binder's and VI growing with the sample size. Furthermore, the VI estimate has improved VI distance with truth and improved or comparable $\tilde{\text{B}}$ distance with truth when compared with the $\tilde{\text{B}}$ estimate.

Further experiments were carried out to consider highly unbalanced clusters. In this case, the conclusions continue to hold; Binder's loss overestimates the number of clusters present, placing uncertain observations in new small clusters, and this effect becomes more pronounced with increased overlap between clusters (results not shown).



(a) $\tilde{B}$ estimate: 9 clusters

(b) $\tilde{B}$ horizontal bound: 11 clusters

(c) $\tilde{B}$ upper vertical bound: 4 clusters

(d) $\tilde{B}$ lower vertical bound: 18 clusters

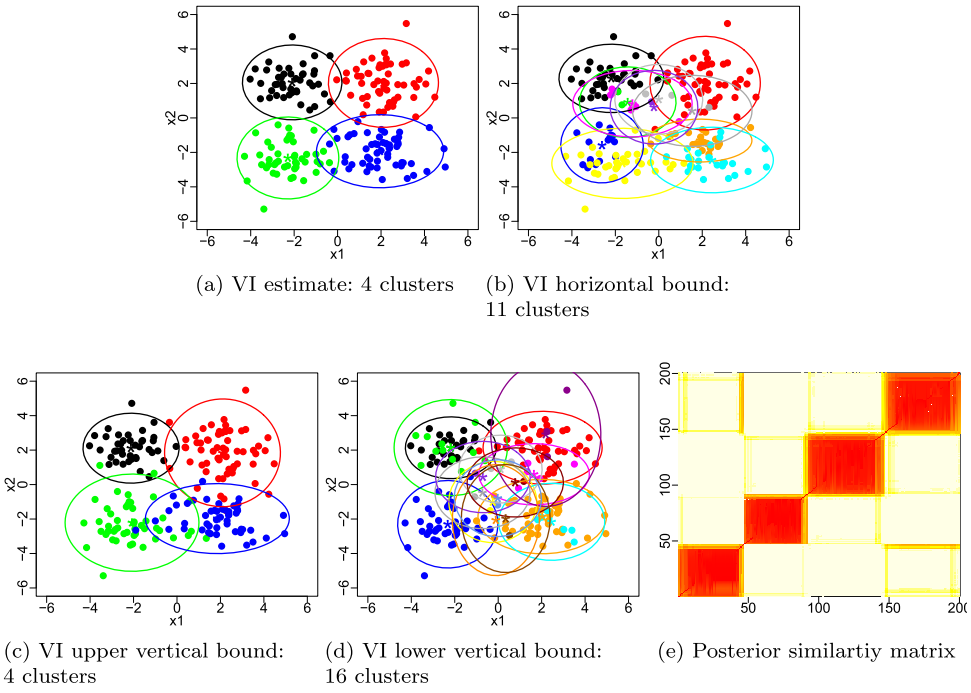
(e) Posterior similartiy matrix

Figure 6: Example 1: 95% credible ball with Binder's loss around $\mathbf{c}^*$ (a) represented by the (b) horizontal bound, (c) upper vertical bound, and (d) lower vertical bound, where color denotes cluster membership, and a heat map of the posterior similarity matrix (e).

For the first example, Figures 6 and 7 represent the 95% credible ball around the optimal partition for $\tilde{B}$ and VI, respectively, through the upper vertical bound, lower vertical bound, and horizontal bound, with data points colored according to cluster membership. Analogous plots for the second example are found in Figures 8 and 9. The Supplementary Material provides tables comparing the bounds with the true clustering through a cross tabulation of the true cluster labels with the cluster labels for each bound.

In the first example, we observe that elements of the 95% credible ball with positive estimated posterior probability have at least four clusters for both metrics and at most 18 clusters for $\tilde{B}$ or 16 clusters for VI, while the most distant elements contain 11 clusters for $\tilde{B}$ and VI (Table 3). For both metrics, these bounds reallocate uncertain data points on the border with these points either added to one of the four main clusters or to new small to medium-sized clusters. For example, in the $\tilde{B}$ upper bound, 19 elements of the third cluster (green in Figure 4a) are added to the fourth cluster (blue in Figure 4a) and in the $\tilde{B}$ lower bound, the fourth cluster (blue in Figure 4a) is split in two medium-sized clusters and several small clusters.

In the second example, the first cluster (black in Figure 4b) is stable in all bounds, while the 95% credible ball reflects posterior uncertainty on whether to divide the re-



(a) VI estimate: 4 clusters

(b) VI horizontal bound: 11 clusters

(c) VI upper vertical bound: 4 clusters

(d) VI lower vertical bound: 16 clusters

(e) Posterior similartiy matrix

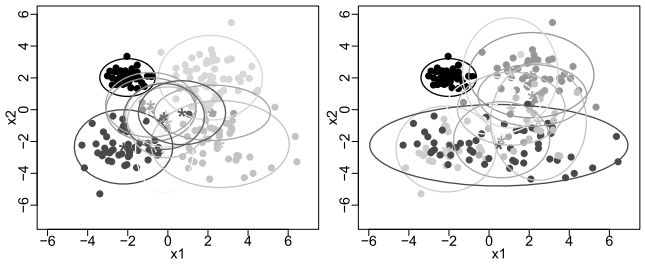
Figure 7: Example 1: 95% credible ball with VI around $\mathbf{c}^*$ (a) represented by the (b) horizontal bound, (c) upper vertical bound, and (d) lower vertical bound, where color denotes cluster membership, and a heat map of the posterior similarity matrix (e).

	Loss	Upper		Lower		Horizontal	
		k_N^u	$d(\mathbf{c}^*, \mathbf{c}_u)$	k_N^l	$d(\mathbf{c}^*, \mathbf{c}_l)$	k_N^h	$d(\mathbf{c}^*, \mathbf{c}_h)$
Ex 1:	$\tilde{\text{B}}$	4	0.097	18	0.097	11	0.097
	VI	4	1.02	16	1.152	11	1.213
Ex 2:	$\tilde{\text{B}}$	4	0.137	19	0.131	10	0.137
	VI	3	1.043	16	1.342	6	1.403

Table 3: A summary of the credible bounds with $\tilde{\text{B}}$ or VI in terms of the number of clusters and distance to the clustering estimate for the upper vertical, lower vertical, and horizontal bounds and for both examples.

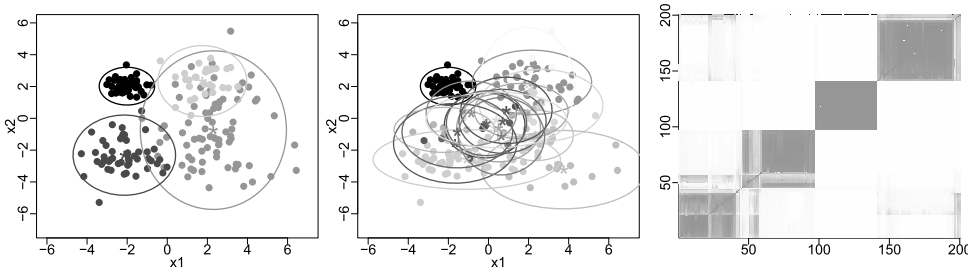
maining data points into 3 to 18 clusters for $\tilde{\text{B}}$ and 2 to 15 clusters for VI (Table 3). Notice the high uncertainty in the fourth cluster with increased variance (blue in Figure 4b). Additionally, note the greater uncertainty around the optimal estimate in Example 2, as the horizontal distance in Table 3 is greater for Example 2 for both metrics.

Figures 6–9 also present heat maps of the posterior similarity matrix for both examples. In general, the posterior similarity matrix appears to under-represent the uncertainty; indeed, one would conclude from the similarity matrix that there is only



(a) $\tilde{\mathrm{B}}$ estimate: 12 clusters

(b) $\tilde{\mathrm{B}}$ horizontal bound: 10 clusters



(c) $\tilde{\mathrm{B}}$ upper vertical bound: 4 clusters

(d) $\tilde{\mathrm{B}}$ lower vertical bound: 19 clusters

(e) Posterior similartiy matrix

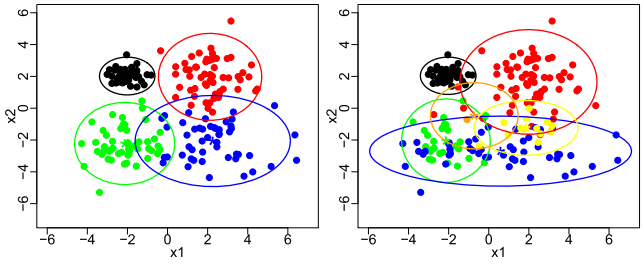
Figure 8: Example 2: 95% credible ball with Binder's loss around $\mathbf{c}^*$ (a) represented by the (b) horizontal bound, (c) upper vertical bounds (only one of two shown for conciseness), and (d) lower vertical bound, where color denotes cluster membership, and a heat map of the posterior similarity matrix (e).

uncertainty in allocation of a few data points in Example 1. Moreover, the 95% credible ball gives a precise quantification of the uncertainty.

6.2 Galaxy example

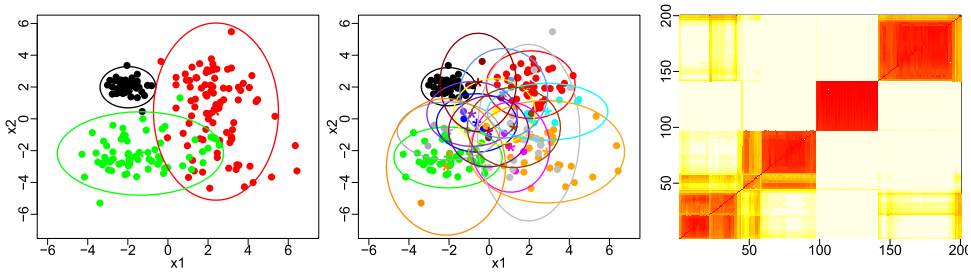
We consider an analysis of the galaxy data (Roeder (1990)), available in the MASS package of **R**, which contains measurements of velocities in km/sec of 82 galaxies from a survey of the Corona Borealis region. The presence of clusters provides evidence for voids and superclusters in the far universe. The data is modeled with a DP mixture (5). The parameters were selected empirically with $\mu_0 = \bar{x}, c = 1/2, a = 2, b = s^2$, where $\bar{x}$ represents the sample mean and s^2 represents the sample variance. The mass parameter α is given a Gam$(1, 1)$ hyperprior.

With 10,000 samples after 1,000 burn in, the posterior mass is spread out over 9,636 partitions, emphasizing the need for appropriate summary tools. Figure 10 plots the point estimate of the partition found by the greedy search algorithm for Binder's loss and VI (with multiple restarts and the default value of $l = 2N$). The data values are



(a) VI estimate: 4 clusters

(b) VI horizontal bound: 6 clusters



(c) VI upper vertical bound: 3 clusters

(d) VI lower vertical bound: 16 clusters

(e) Posterior similarity matrix

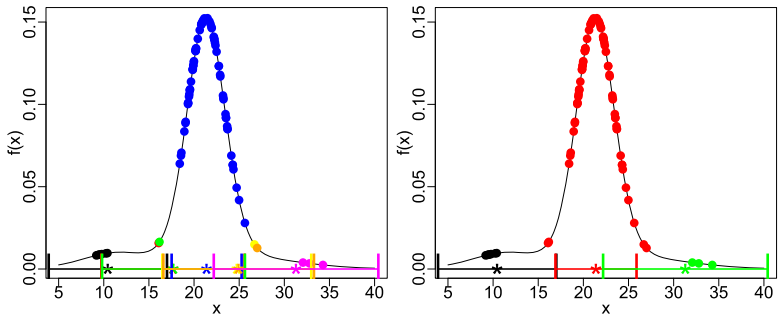
Figure 9: Example 2: 95% credible ball with VI around $\mathbf{c}^*$ (a) represented by the (b) horizontal bound, (c) upper vertical bound, and (d) lower vertical bound, where color denotes cluster membership and a heat map of the posterior similarity matrix (e).

plotted against the estimated density values from the DP mixture model and colored according to cluster membership, with correspondingly colored stars and bars along the x-axis representing the posterior mean and variance within cluster. Again, we observe that Binder's loss places observations with uncertain allocation into singleton clusters, with a total of 7 clusters, 4 of which are singletons, while the VI solution contains 3 clusters. Table 4 compares the point estimates in terms of the posterior expected $\tilde{\mathrm{B}}$, lower bound of VI, and VI; as anticipated, the $\tilde{\mathrm{B}}$ solution has the smallest posterior expected $\tilde{\mathrm{B}}$ and the VI solution has the smallest posterior expected VI.

Loss	k_N^*	$\mathbb{E}[\tilde{\mathrm{B}}\|\mathcal{D}]$	$\mathbb{E}[\mathrm{VI}_{\mathrm{LB}}\|\mathcal{D}]$	$\mathbb{E}[\mathrm{VI}\|\mathcal{D}]$
$\tilde{\mathrm{B}}$	7	**0.218**	0.746	1.014
VI	3	0.237	**0.573**	**0.939**

Table 4: Galaxy example: a comparison of the optimal partition with Binder's loss and VI in terms of posterior expected $\tilde{\mathrm{B}}$, lower bound to VI, and VI.

The 95% VI credible ball contains all partitions with a VI distance less than 1.832. Figure 11 depicts the 95% credible ball through the upper vertical, lower vertical, and horizontal bounds, which are further described and summarized in Table 5 and



(a) Binder's loss: 7 clusters (b) VI: 3 clusters

Figure 10: Galaxy example: optimal clustering estimate with color representing cluster membership for Binder's loss and VI, with correspondingly colored stars and bars along the x-axis representing the posterior mean and variance within cluster.

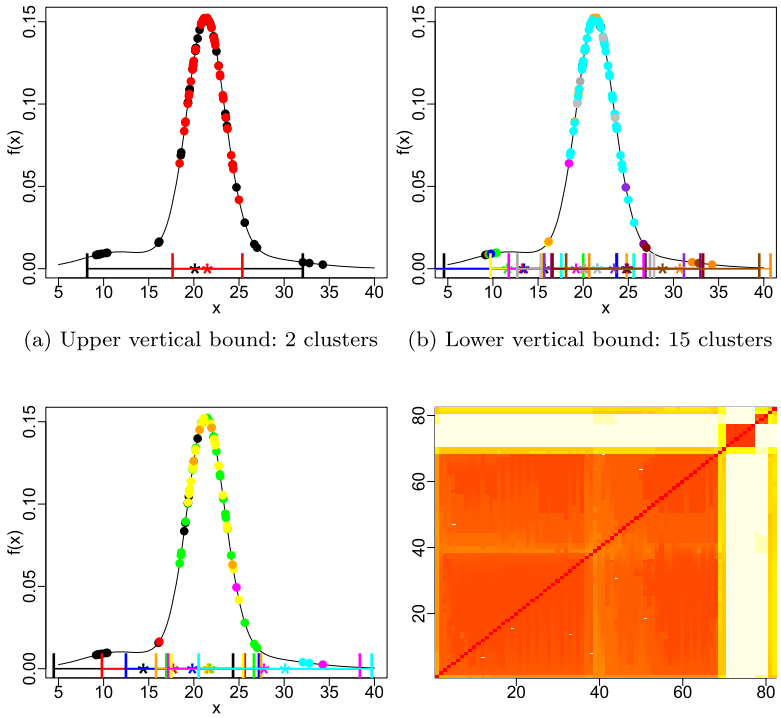
	Upper		Lower		Horizontal	
	k_N^u	$d(\mathbf{c}^*, \mathbf{c}_u)$	k_N^l	$d(\mathbf{c}^*, \mathbf{c}_l)$	k_N^h	$d(\mathbf{c}^*, \mathbf{c}_h)$
Galaxy	2	1.364	15	1.669	8	1.832

Table 5: Galaxy example: a summary of the credible bounds with VI in terms of the number of clusters and distance to the clustering estimate for the upper vertical, lower vertical, and horizontal bounds.

in cross tabulation tables in the Supplementary Material. We observe a large amount of variability around the optimal partition. With 95% posterior probability, we believe that, on one extreme, the data could be modeled using only 2 components, one with a large variance to account for outliers (black cluster in Figure (11a)). On the other extreme, the data could be further split into one medium sized cluster and many, 14 to be precise, smaller clusters. The horizontal bound, the most extreme partition in the credible ball, splits the largest cluster (red in Figure 10b) into two medium sized clusters and four small clusters and reallocates some of its data points to the first cluster (black in Figure 10b). Figure 11d emphasizes that the posterior similarity matrix under-represents the uncertainty around the point estimate in comparison to the credible ball.

7 Discusssion

Bayesian cluster analysis provides an advantage over classical cluster analysis, in that the Bayesian procedure returns a posterior distribution over the entire partition space, reflecting uncertainty in the clustering structure given the data, as opposed to returning a single solution or conditioning on the parameter estimates and number of clusters. This allows one to assess statistical properties of the clustering given the data. However, due to the huge dimension of the partition space, an important problem in Bayesian cluster analysis is how to appropriately summarize the posterior. To address this problem, we



(a) Upper vertical bound: 2 clusters (b) Lower vertical bound: 15 clusters

(c) Horizontal bounds: 8 clusters (d) Posterior similarity matrix

Figure 11: Galaxy example: 95% credible ball with VI represented by the (a) upper vertical bound, (b) lower vertical bound, and (c) horizontal bound, where color denotes cluster membership, with correspondingly colored stars and bars along the x-axis representing the posterior mean and variance within cluster, and (d) a heat map of the posterior similarity matrix.

have developed tools to obtain a point estimate of clustering based on the posterior and describe uncertainty around this estimate via the 95% credible ball.

Obtaining a point estimate through a formal decision theory framework requires the specification of a loss function. Previous literature focused on Binder's loss. In this work, we propose to use an information theoretic measure, the variation of information, and provide a detailed comparison of the two metrics. We find that Binder's loss exhibits peculiar asymmetries, preferring to split over merge clusters, and the variation of information is more symmetric in this regard. This behavior of Binder's loss causes the optimal partition to overestimate the number of clusters, allocating uncertain data points to small additional clusters. In addition, we have developed a novel greedy search algorithm to locate the optimal partition, allowing one to explore beyond the space of partitions visited in the MCMC chain.

To represent uncertainty around the point estimate, we construct 95% credible balls around the point estimate and depict the credible ball through the upper vertical, lower

vertical, and horizontal bounds. In addition to a heat map of the posterior similarity matrix, which is often reported in literature, the 95% credible ball enriches our understanding of the uncertainty present. Indeed, it provides a precise quantification of the uncertainty present around the point estimate, and in examples, we find that an analysis based on the posterior similarity matrix leads one to be over certain in the clustering structure. The developed posterior summary tools for Bayesian cluster analysis are available[2] through an **R** package 'mcclust.ext' (Wade (2015)), expanding upon the existing **R** package 'mcclust' (Fritsch (2012)).

In future work, we aim to extend these ideas to Bayesian feature allocation analysis, an extension of clustering which allows observations to belong to multiple clusters (Griffiths and Ghahramani (2011)). A further direction of research will be to explore posterior consistency for the number of clusters based on the VI estimate for Bayesian nonparametric mixture models; this is of particular interest in light of the negative results of Miller and Harrison (2013) and Miller and Harrison (2014) and the positive results in our simulation studies (Table 2). Finally, scalability issues of Bayesian nonparametric mixture models are an important concern for very large datasets. To scale with large sample sizes, a number of papers have avoided exploration of the posterior on partitions through MCMC and focused on finding a point estimate of the partition, often through MAP inference (Heller and Ghahramani (2005), Dahl (2009), Raykov et al. (2014)) or the DP-means algorithm and its extensions (Kulis and Jordan (2012), Jiang et al. (2012), Broderick et al. (2013)). One direction of future research is to develop an algorithm to find the point estimate which minimizes the posterior expected VI that avoids MCMC. Of course, while gaining in scalability, we lose the uncertainty in the clustering structure.

Supplementary Material

Supplementary material for Bayesian cluster analysis: Point estimation and credible balls (DOI: 10.1214/17-BA1073SUPP; .pdf).

References

Binder, D. (1978). "Bayesian Cluster Analysis." *Biometrika*, 65: 31–38. MR0501592. doi: https://doi.org/10.1093/biomet/65.1.31. 561, 564

Broderick, T., Kulis, B., and Jordan, M. (2013). "MAD-Bayes: MAP-based asymptotic derivations from Bayes." In *Proceedings of the 30th International Conference on Machine Learning*, 226–234. 583

Dahl, D. (2006). "Model-based clustering for expression data via a Dirichlet process mixture model." In Do, K., Müller, P., and Vannucci, M. (eds.), *Bayesian Inference for Gene Expression and Proteomic*, 201–218. Cambridge University Press. MR2706330. 565

[2]Through the author's website https://www2.warwick.ac.uk/fac/sci/statistics/staff/academic-research/wade/.

Dahl, D. (2009). "Modal clustering in a class of product partition models." *Bayesian Analysis*, 4: 243–264. MR2507363. doi: https://doi.org/10.1214/09-BA409. 561, 583

Duan, J., Guindani, M., and Gelfand, A. (2007). "Generalized spatial Dirichlet processes." *Biometrika*, 94: 809–825. MR2416794. doi: https://doi.org/10.1093/biomet/asm071. 564

Dunson, D. (2010). "Nonparametric Bayes applications to biostatistics." In Hjort, N., Holmes, C., Müller, P., and Walker, S. (eds.), *Bayesian nonparametrics*. Cambridge University Press. MR2730665. 564

Favaro, S. and Teh, Y. (2013). "MCMC for normalized random measure mixture models." *Statistical Science*, 28: 335–359. MR3135536. doi: https://doi.org/10.1214/13-STS422. 564

Favaro, S. and Walker, S. (2012). "Slice sampling σ-stable Poisson–Kingman mixture models." *Journal of Computational and Graphical Statistics*, 22: 830–847. MR3173745. doi: https://doi.org/10.1080/10618600.2012.681211. 564

Ferguson, T. (1973). "A Bayesian analysis of some nonparametric problems." *Annals of Statistics*, 1: 209–230. MR0350949. 563

Fraley, C. and Raftery, A. (2002). "Model-based clustering, discriminant analysis, and density estimation." *Journal of the American Statistical Association*, 97: 611–631. MR1951635. doi: https://doi.org/10.1198/016214502760047131. 559

Fritsch, A. (2012). *mcclust: Process an MCMC Sample of Clusterings*. URL http://cran.r-project.org/web/packages/mcclust/mcclust.pdf 571, 583

Fritsch, A. and Ickstadt, K. (2009). "Improved criteria for clustering based on the posterior similarity matrix." *Bayesian Analysis*, 4: 367–392. MR2507368. doi: https://doi.org/10.1214/09-BA414. 561, 565

Griffin, J. and Steel, M. (2006). "Order-based dependent Dirichlet processes." *Journal of the American Statistical Association*, 10: 179–194. MR2268037. doi: https://doi.org/10.1198/016214505000000727. 564

Griffiths, T. and Ghahramani, Z. (2011). "The Indian buffet process: An introduction and review." *Journal of Machine Learning Research*, 12: 1185–1224. MR2804598. 583

Hartigan, J. and Wong, M. (1979). "Algorithm AS 136: A k-means clustering algorithm." *Journal of the Royal Statistical Society, Series C*, 28: 100–108. MR0405726. 559

Heard, N., Holmes, C., and Stephens, D. (2006). "A quantitative study of gene regulation involved in the immune response of anopheline mosquitos: An application of Bayesian hierarchical clustering of curves." *Journal of the American Statistical Association*, 101: 18–29. MR2252430. doi: https://doi.org/10.1198/016214505000000187. 561

Heller, K. and Ghahramani, Z. (2005). "Bayesian hierarchical clustering." In *Proceedings of the 22nd International Conference on Machine Learning*, 297–304. 561, 583

Hubert, L. and Arabie, P. (1985). "Comparing partitions." *Journal of Classification*, 2: 193–218. 565

Ishwaran, H. and James, L. (2001). "Gibbs campling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96: 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 563, 564

Jiang, K., Kulis, B., and Jordan, M. (2012). "Small-variance asymptotics for exponential family Dirichlet process mixture models." In *Advances in Neural Information Processing Systems*, 3158–3166. 583

Kalli, M., Griffin, J., and Walker, S. (2011). "Slice sampling mixture models." *Statistics and Computing*, 21: 93–105. MR2746606. doi: https://doi.org/10.1007/s11222-009-9150-y. 564

Kulis, B. and Jordan, M. (2012). "Revisiting K-means: New algorithms via Bayesian nonparametrics." In *Proceedings of the 29th International Conference on Machine Learning*, 513–520. 583

Lau, J. and Green, P. (2007). "Bayesian model-based clustering procedures." *Journal of Computational and Graphical Statistics*, 16: 526–558. MR2351079. doi: https://doi.org/10.1198/106186007X238855. 561, 565, 571

Lijoi, A. and Prünster, I. (2011). "Models beyond the Dirichlet process." In Hjort, N., Holmes, C., Müller, P., and Walker, S. (eds.), *Bayesian Nonparametrics*, 80–136. Cambridge, UK: Cambridge University Press. MR2730661. 563

Lo, A. (1984). "On a class of Bayesian nonparametric estimates: I. Density estimates." *Annals of Statistics*, 12: 351–357. MR0733519. doi: https://doi.org/10.1214/aos/1176346412. 560

Lomellí, M., Favaro, S., and Teh, Y. (2015). "A hybrid sampler for Poisson–Kingman mixture models." In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*. 564

Lomellí, M., Favaro, S., and Teh, Y. (2016). "A marginal sampler for σ-stable Poisson–Kingman mixture models." *Journal of Computational and Graphical Statistics*. To appear. 564

MacEachern, S. (2000). "Dependent Dirichlet processes." *Technical Report, Department of Statistics, Ohio State University*. 564

Medvedovic, M. and Sivaganesan, S. (2002). "Bayesian infinite mixture model based clustering of gene expression profiles." *Bioinformatics*, 18: 1194–1206. 561

Medvedovic, M., Yeung, K., and Bumgarner, R. (2004). "Bayesian mixture model based clustering of replicated microarray data." *Bioinformatics*, 20: 1222–1232. 561

Meilă, M. (2007). "Comparing clusterings – an information based distance." *Journal of Multivariate Analysis*, 98: 873–895. MR2325412. doi: https://doi.org/10.1016/j.jmva.2006.11.013. 561, 566

Miller, J. and Harrison, M. (2013). "A simple example of Dirichlet process mixture inconsistency for the number of components." In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc. 583

Miller, J. and Harrison, M. (2014). "Inconsistency of Pitman–Yor process mixtures for the number of components." *Journal of Machine Learning Research*, 15: 3333–3370. MR3277163. 583

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010). "Bayesian profile regression with an application to the national survey of children's health." *Biostatistics*, 11: 484–498. 561

Müller, P. and Quintana, F. (2004). "Nonparametric Bayesian data analysis." *Statistical Science*, 19: 95–110. MR2082149. doi: https://doi.org/10.1214/088342304000000017. 564

Nation, J. (1991). *Notes on Lattice Theory*. http://www.math.hawaii.edu/~jb/books.html. 567

Neal, R. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistcs*, 9: 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 564, 574

Papaspiliopoulos, O. and Roberts, G. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*, 95(1): 169–186. MR2409721. doi: https://doi.org/10.1093/biomet/asm086. 564

Pitman, J. (2003). "Poisson Kingman partitions." In *Statistics and Science: a Festschrift for Terry Speed*, 1–34. Beachwood: IMS Lecture Notes. MR2004330. doi: https://doi.org/10.1214/lnms/1215091133. 563

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator." *Annals of Probability*, 25: 855–900. MR1434129. doi: https://doi.org/10.1214/aop/1024404422. 563

Quintana, F. (2006). "A predictive view of Bayesian clustering." *Journal of Statistical Planning and Inference*, 136: 2407–2429. MR2279815. doi: https://doi.org/10.1016/j.jspi.2004.09.015. 560

Quintana, F. and Iglesias, P. (2003). "Bayesian clustering and product partition models." *Journal of the Royal Statistical Society: Series B*, 65: 557–574. MR1983764. doi: https://doi.org/10.1111/1467-9868.00402. 561, 565

Rand, W. (1971). "Objective criteria for the evaluation of clustering methods." *Journal of the American Statistical Association*, 66: 846–850. 565

Rasmussen, C., De la Cruz, B., Ghahramani, Z., and Wild, D. (2009). "Modeling and visualizing uncertainty in gene expression clusters using Dirichlet process mixtures." *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 6: 615–628. 561

Raykov, Y., Boukouvalas, A., and Little, M. (2014). "Simple approximate MAP Inference for Dirichlet processes." Available at https://arxiv.org/abs/1411.0939. MR3572859. doi: https://doi.org/10.1214/16-EJS1196. 561, 583

Roeder, K. (1990). "Density estimation with confidence sets exemplified by superclusters and voids in galaxies." *Journal of the American Statistical Association*, 85: 617–624. 579

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). "Hierarchical Dirichlet process." *Journal of the American Statistical Association*, 101: 1566–1581. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 564

Vinh, N., Epps, J., and Bailey, J. (2010). "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance." *Journal of Machine Learning Research*, 11: 2837–2854. MR2738784. 566

Wade, S. (2015). *mcclust.ext: Point estimation and credible balls for Bayesian cluster analysis*. URL https://www.researchgate.net/publication/279848500_mcclustext-manual. 572, 583

Wade, S. and Ghahramani, Z. (2017). "Supplementary material for Bayesian cluster analysis: Point estimation and credible balls." *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1073SUPP. 567

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/I036575/1].

Invited comment on Article by Wade and Ghahramani

Stefano Monni*

Professor Wade and Professor Ghahramani have written an interesting paper that deals with the very important question of how to summarize the posterior distribution of partitions in nonparametric models. The summary of the posterior they propose is in the form of a point estimate and an associated credible ball, which quantifies the uncertainty of the estimate, using a decision-theoretic approach.

I enjoyed reading the paper and would like to make two comments. The first one is a clarification. The second is concerned with the graphical representation of the credible ball.

The authors consider the variation of information (VI) and Binder's loss; show that these criteria for comparison of clusters are metrics/distances; give proofs of some of their properties, such as vertical and horizontal collinearity; determine their bounds and their scales; obtain the closest cluster to a given cluster according to those distances. All these properties (and a few more others) are already described and proven in the paper of Meilă (2007), where the VI distance is introduced. Indeed, the authors explicitly refer to that paper for some proofs. However, they do so only for the VI. Because of this, some readers may be left with the impression that the properties of the Binder's loss and some of the advantages of the VI over the Binder's loss when comparing clusters were never described in detail before this paper. Thus, I think it is important to stress that Meilă does give details of the properties of the Binder's loss. To be precise, Meilă considers a number of criteria useful to compare clusters, among which the Mirkin metric (Mirkin and Chernyi, 1970). The latter metric is equal to twice the Binder's loss $B(\mathbf{c}, \hat{\mathbf{c}})$, and its N-invariant version is equal to the N-invariant Binder's loss $\tilde{B}(\mathbf{c}, \hat{\mathbf{c}})$. In fact Meilă provides a quite explicit comparison of the VI with the Mirkin metric. With this clarified, the entire Section 3 of the paper under discussion should only be seen as a review of the comparison of Binder's loss with the VI as criteria for comparison of clusters. To be fair, in other sections of the paper, the contrasts between these two distances are analyzed further (for example in the description of the credible balls) although not at the same level of formality.

The fact that the Binder's loss and the VI are distances is brought to bear in the definition of the credible ball, which is the most interesting part of the paper. The credible ball is a very useful concept and I agree with the authors that it allows a characterization of uncertainty of the point estimate. Naturally, since the ball is a subset of the partition space, one is faced, yet again, with the problem of summarizing a subset of partitions. It is perhaps for this reason that the authors introduce the concept of vertical and horizontal bounds. If one looks at the graphical representations of the credible balls presented in the paper (Figures 7 and 8 for instance), one will undoubtedly

*Department of Mathematics, American University of Beirut, sm150@aub.edu.lb

find them pleasing and informative. However, some questions should be asked about such plots. Namely, I'm concerned about graphically representing the credible ball when the vertical and horizontal bounds consist of more than one partition. Indeed, the definitions of the bounds do not prevent such occurrences and, in fact, examples of such bounds are explicitly given in section 2 of the supplementary material. If the number of objects to cluster is large, it is quite plausible that the horizontal and vertical bounds too are sets of large size. The problem of representing a credible ball has turned into the problem of representing the bounds, which appears to be just as intricate, if not identical. In the paper it is stated that what is used in practice to define the bounds is the subset of partitions in the credible ball that have positive estimated posterior probabilities, but, even so, the bounds will hardly contain one partition. I would be very interested in knowing what the authors suggest should be done when the sets of the bounds are large. One can perhaps just depict one representative partition for each of the bounds that is selected on the basis of additional considerations. As a selection criterion one could employ the value of the posterior probability or of the expected posterior loss, but it is difficult to see whether this could really work well. To put it another way, I'm suggesting the authors should think of a refinement of the definition of bounds.

Wade and Ghahramani propose the credible ball as an elegant alternative to the posterior similarity matrix in assessing the uncertainty of a cluster estimate. They state that the posterior similarity matrix under-represents this uncertainty, when compared with the credible ball. While I'm sure that they will agree with me that much more evidence is necessary to conclude whether this is true in general, I suspect that the difficulty I see in representing the credible ball may limit its success. The heatmap of the posterior similarity matrix continues to be in my view a very valid (if not irreplaceable) tool for assessing the uncertainty of cluster estimates. However, I hope to hear from Wade and Ghahramani that my concerns about the graphical representation of the credible ball are misplaced.

References

Meilă, M. (2007). "Comparing clusterings-an information based distance." *Journal of Multivariate Analysis*, 98(5): 873–895. MR2325412. doi: https://doi.org/10.1016/j.jmva.2006.11.013. 588

Mirkin, B. G. and Chernyi, L. B. (1970). "Measurement of the distance between distinct partitions of a finite set of objects." *Automation and Remote Control*, 31(5): 786–792. MR0300907. 588

Invited comment on Article by Wade and Ghahramani

Giorgio Paulon*, Lorenzo Trippa†, and Peter Müller ‡

We thank the authors for an interesting discussion of estimates and uncertainty summaries for random partitions. A coherent description of uncertainties is one of the strengths of the Bayesian approach, but it is difficult to summarize and report it in the case of a random partition. The clever and elegant approach of Wade and Ghahramani addresses this critical gap in the literature. However, the approach relies on loss functions that ignore the underlying inference problem that gave rise to the random partition. In other words, the loss functions are generic inference losses that ignore the context of the scientific question that the investigators are trying to address. In this discussion we would like to elaborate on the authors' related comment that alternative loss functions could be tailored to specific problems.

We assume that the inference problem and sampling model include cluster-specific parameters, $\boldsymbol{\theta}^\star_j$, $j = 1, \ldots, k_N$. For example, if $\boldsymbol{\theta}^\star_j$ were the mean times to progression for patients in a clinical trial, the clusters would describe patient subpopulations with different mean time to progression. A summary of the random partition should then focus on partitions with meaningfully different $\boldsymbol{\theta}^\star_j$'s. Similarly, in some contexts, one might prefer avoiding inclusion and reporting of small clusters. Inspired by Xu *et al.* (2016) who use a determinantal point process to favor configurations with diverse cluster-specific parameters, we propose the following loss function. The loss function formalizes a tradeoff between reporting clusters that are representative of the posterior and, with the second term, favoring partitions with clusters C_j that are diverse:

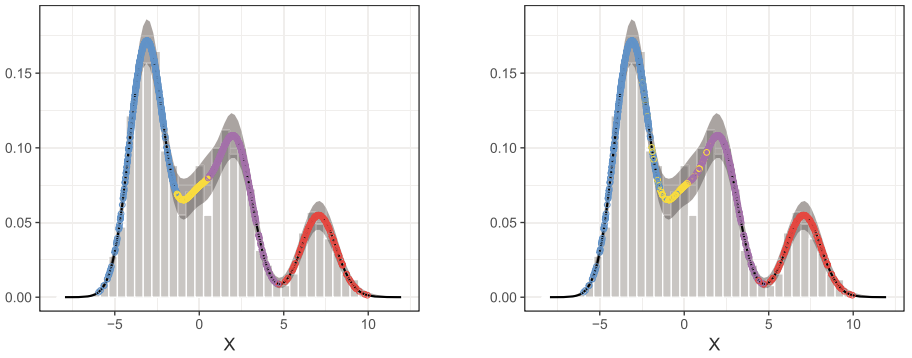
$$L_{rep}(\boldsymbol{c}, \hat{\boldsymbol{c}}, \boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}^\star) = \frac{1}{N} \sum_{n=1}^{N} \left(\theta^\star_{c_n} - \hat{\theta}^\star_{\hat{c}_n} \right)^2 - \lambda \det(\widehat{\Phi}),$$

where $[\widehat{\Phi}_{ij}]_{i,j} = \phi_\tau(\hat{\theta}^\star_i, \hat{\theta}^\star_j)$ for some kernel $\phi_\tau(x, y)$, e.g. the squared exponential $\phi_\tau(x, y) = \exp\{-0.5[(x - y)/\tau]^2\}$. That is, $\det(\widehat{\Phi})$ is the volume of a parallelotope spanned by the columns of $\widehat{\Phi}$, which is zero when $\theta^\star_i = \theta^\star_j$ for any $i \neq j$, and maximized when they are very distinct. Of course the squared distance in the loss can be replaced by a different distance, e.g. one that allows for asymmetric costs of misfit. The second component of the loss function could also be modified to mirror specific goals, for example penalizing configurations that include small clusters. The point here is that, in general, the particular application should drive the choice of the loss function.

*Department of Statistics & Data Science, University of Texas, Austin, TX, giorgio.paulon@utexas.edu

†Dana-Faber Cancer Institute, Boston, MA, ltrippa@jimmy.harvard.edu

‡Department of Statistics & Data Science, University of Texas, Austin, TX, pmueller@math.utexas.edu



(a) under VI (b) optimal $\hat{c}$ under L_{rep}

Figure 1: Optimal $\hat{c}$ for the normal mixture example. The histogram shows the data; the black curve shows the estimated posterior mean of the random probability measure, along with pointwise 95% credible intervals. Color shows the estimated cluster membership for x_i.

We compared L_{rep} with the VI loss and also with the squared loss (Dahl, 2006) in the following example. Let $\mathrm{N}(x;\ m, s)$ denote a normal p.d.f. with location m and scale s evaluated at x, and let $\boldsymbol{\mu} = (-3, -3.5, -2.6, 0, 1.8, 2.4, 7.1)$. We simulated $N = 1000$ observations from a mixture of 7 normals, $p(x_i \mid \boldsymbol{\mu}) \propto \sum_{j=1}^{7} \mathrm{N}(x_i;\ \mu_j, 1)$. We fit the data using a Dirichlet process mixture of normals model. In this case, only four components of the mixture are likely to be practically meaningful. The three values around -3 and the two around 2 are not meaningfully different (relative to the variances in the normal kernels). Inference summaries under L_{rep} and VI loss are shown in Figure 1. In this example the posterior mode for k_N is $\hat{k}_N = 7$. But both loss functions penalize excessive complexity and shrink the reported partition to the 4 groups shown in the figure. Although the VI loss does not explicitly favor easy interpretation, it does surprisingly well in this example. We used an implementation that restricted the search for the Bayes estimate of the partition under L_{rep} to the simulated partitions only, which might explain the counter-intuitive lack of monotonicity in the cluster membership in Figure 1a. One could alternatively use better search algorithms such as, for example, the sequentially-allocated latent structure optimization (SALSO) in the `sdols` R package (Dahl and Müller, 2017). We do not show the results obtained under squared loss or Binder's loss, since both clearly overfit the data reporting $k_N = 53$ components.

Next we investigate a scenario with a small number of observations. We compare the same two loss functions with a dataset from a clinical trial for sarcoma patients with binary endpoints (tumor response) (León-Novelo *et al.*, 2013). The goal of the study is to cluster $N = 10$ different sarcomas subtypes. That is, the experimental units for the random partition are the disease subtypes. The sampling model is binomial sampling, $x_i|\pi_i \sim \mathrm{Bin}(M_i, \pi_i)$ for the number of tumor responses x_i for a given number of patients M_i under each sarcoma subtype, $i = 1, \ldots, N$. The number of patients, M_i for each subtype are moderately small, between 2 and 29. We implement inference

using a Dirichlet process mixture of probit models. Inference under VI and squared loss reports 10 singleton clusters, a partition which is difficult to interpret, also because of the negligible differences between estimated cluster-specific response rates. See Figure 2 for a summary of the posterior estimated response rates π_i. In contrast, the desired preference for interpretable structure is explicitly included in L_{rep}, leading us to report $\hat{\boldsymbol{c}} = (1, 1, 1, 1, 1, 1, 1, 1, 2, 1)$, which appears more plausible in the light of the estimated response probabilities (the singleton cluster is Ewings' sarcoma).

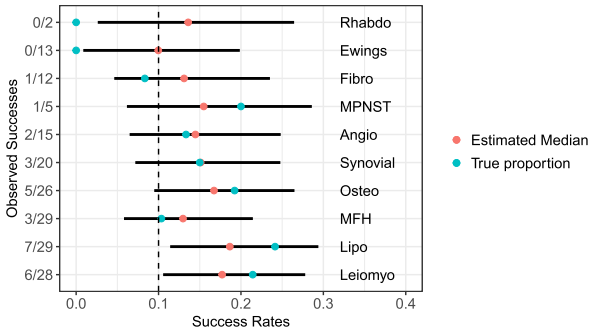


Figure 2: 90% posterior credible intervals of the Binomial success probabilities π_i for each sarcoma. For reference the dashed vertical line marks 0.1.

There are two more aspects of inference for random partitions that we would like to briefly discuss. Both are related to the underlying data analysis problem. In many applications the main inference target is not the entire partition, but only a special subset. Assume, for example, that in an analysis of clinical trial data cluster-specific parameters $\boldsymbol{\theta}_j^\star$ are interpreted as treatment effects. An important problem is to find the subset of patients who most benefit from the treatment under consideration, that is, the subset with the largest $\boldsymbol{\theta}_j^\star$. This is known as subgroup analysis. Let $B = C_{j^\star}$ denote the subset C_j with the largest $\boldsymbol{\theta}_j^\star$. Characterizing uncertainty on a random partition now reduces to reporting uncertainty on B. Schnell *et al.* (2016) develop a clever approach to determine a pair of subsets (D, S) such that $p(D \subseteq B \subseteq S \mid data) > 1-\alpha$. Subgroup analysis is in general not necessarily linked with random partitions and involves several other issues. The point here is to emphasize that relevant uncertainty on a random partition need not treat all subsets symmetrically. Investigators might only be concerned about a particular subset.

Finally, we would like to bring up one more aspect about summaries of clustering uncertainty, related to reproducibility. Above, we used a decision theoretic framework to summarize a random partition with a good estimate that is constructed to be representative of the posterior distribution. Additionally, we report uncertainty measures that mirror the distance between the selected configuration and a fictitious latent partition. Although primarily meant to summarize the posterior distribution, these uncertainty measures are also vaguely related to the (frequentist) variability of the estimate $\hat{c}$. Indeed, consider repeating the entire experiment de novo, including both, data generation and analysis. It remains unclear how different the estimated configuration $\hat{c}$ might turn

out. In most Bayesian estimation problems of key parameters, including means or medians, estimating this variability is unnecessary to express uncertainty, and the focus is exclusively on the posterior distribution of the parameter of interest. But clustering is an attempt to organize data points into conveniently created categories. An underlying true unknown partition might be useless or not exist at all. These considerations lead us to suggest the report of replicability measures that could contrast $\hat{c}$ and estimates under independent replicates, possibly including variations in the sample size. An extended set of uncertainty metrics could scrutinize the main drivers of variability, including limitations in the measurement of the statistical units (low sample size for each sarcoma subtype, in the previous application), data preprocessing, clustering methods, and experimental designs.

References

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In M. Vannucci, K.-A. Do, and P. Müller, editors, *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press. MR2706330. 591

Dahl, D. B. and Müller, P. (2017). *sdols: Summarizing Distributions of Latent Structures*. R package version 1.4. 591

León-Novelo, L. G., Müller, P., Arap, W., Kolonin, M., Sun, J., Pasqualini, R., and Do, K.-A. (2013). Semiparametric Bayesian inference for phage display data. *Biometrics*, **69**(1), 174–183. MR3058064. doi: https://doi.org/10.1111/j.1541-0420.2012.01817.x. 591

Schnell, P. M., Tang, Q., Offen, W. W., and Carlin, B. P. (2016). A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics*, **72**, 1026–1036. MR3591587. doi: https://doi.org/10.1111/biom.12522. 592

Xu, Y., Müller, P., and Telesca, D. (2016). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics*, **72**, 955–964. MR3545688. doi: https://doi.org/10.1111/biom.12482. 590

Invited comment on Article by Wade and Ghahramani

Nial Friel* and Riccardo Rastelli†

Abstract. We present a discussion of the paper "Bayesian cluster analysis: point estimation and credible balls" by Sara Wade and Zoubin Ghahramani. We believe that this paper contributes substantially to the literature on Bayesian clustering by filling in an important methodological gap, by providing a means to assess the uncertainty around a point estimate of the optimal clustering solution based on a given loss function. In our discussion we reflect on the characterisation of uncertainty around the Bayesian optimal partition, revealing other possible alternatives that may be viable. In addition, we suggest other important extensions of the approach proposed which may lead to wider applicability.

Keywords: Bayesian clustering, greedy optimisation, latent variable models, Markov chain Monte Carlo.

We congratulate the authors, Wade and Ghahramani (W&G hereafter), on a wonderful article which is an excellent contribution to the area of Bayesian cluster analysis. Here the authors address the problem of appropriately summarising a partition based on a posterior. This is a crucial issue arising in a variety of clustering contexts. While Markov chain Monte Carlo techniques, for example, can be used to efficiently sample the cluster membership variables from the posterior distribution of a variety of mixture models, it is not immediately clear then how one can reasonably summarise such information. Similarly to other previous papers, notably Lau and Green (2007), the authors define the optimal partition as the one minimising the posterior expectation of a suitable loss function, and propose a greedy algorithm to estimate such an optimal solution. Somewhat surprisingly, there has been very little in the literature around how one might assess the uncertainty in this point estimate. W&G address this crucial gap by introducing a strategy to characterise the uncertainty around the optimal partition using an adaptation of the credible intervals approach. We consider this to be a major contribution and expect it stimulate future developments in this field.

We have recently worked on the same problem and published our findings in Rastelli and Friel (2017) (hereafter referred to as R&F). Similarly to W&G, we rely on a decision theoretic framework to summarise a collection of partitions, however, differently from their approach, our contribution is primarily focused on the computational aspects of the problem. Our method is implemented in the `R` package `GreedyEPL` available on `CRAN`. In this discussion we compare our findings to those of W&G mainly focusing on

*School of Mathematics and Statistics and Insight Centre for Data Analytics, University College Dublin, Ireland, nial.friel@ucd.ie

†Institute for Statistics and Mathematics, WU Vienna University of Economics and Business, Austria, riccardo.rastelli@wu.ac.at

the following aspects: the choice of loss function used and the ensuing computational complexity; alternatives to the credible balls approach; the wider applicability of the methods proposed.

1 Choice of loss function and computational efficiency

As W&G clearly point out in their paper, commonly used loss functions such as the $0-1$ loss or the squared error loss are not ideally suited to compare partitions, due to the discrete nature of the variables and because of the lack of total order in the space. This leads to the important issue of finding an appropriate and reasonable loss function to compare partitions. A popular choice in this context is Binder's loss, primarily for two main reasons: its close connection to the Rand index; but also since the corresponding optimal partition can be estimated via the posterior similarity matrix, which itself can be routinely estimated by Markov chain Monte Carlo, for example. The posterior similarity matrix is an $N \times N$ matrix with element n, n' (denoted $p_{n,n'}$ in W&G) equal to the posterior probability that observations n and n' are allocated to the same cluster and where N denotes the size of the dataset. The Variation of Information (VI) loss does not possess such a representation in terms of the posterior similarity matrix and as such it turns out that this brings with it an increased computational overhead. However, W&G neatly sidestep this problem by exploiting Jensen's inequality to obtain a lower bound for the VI loss which relies only on the posterior similarity matrix. This input is interesting, though we note that the effect that this approximation has on the estimated optimal partition is not clear.

In R&F, the approach we advocate does not rely on the posterior similarity matrix representation and does not involve any approximation. In fact, our method may be used with any loss function, $\mathcal{L}(\mathbf{a}, \mathbf{z})$ that depends on the two partitions, $\mathbf{a}$ and $\mathbf{z}$ through the counts n_{ij}, denoting the number of data points allocated to group i in partition $\mathbf{a}$ and to group j in partition $\mathbf{z}$, which can conceptually be considered as depending on the contingency table defined by both partitions. Binders' loss and VI loss are included in this family, along with other known losses such as the normalised VI and the normalised information distance. Moreover, since our approach does not require the posterior similarity matrix, its computational complexity in N is decreased to a linear order (See Figure 1 of Rastelli and Friel (2017)). However, the computational cost of our approach also becomes increasingly costly as sample size of partitions drawn from the posterior increases.

Additionally, R&F empirically assess the effect of the various loss functions on simulated data and in particular we refer the reader to Figure 3 of Rastelli and Friel (2017). The main take home message is that the VI loss typically achieved the best results in terms of the number of estimated groups, while the other loss functions, including Binders loss, the normalised VI loss and the normalised information distance often exhibit unreasonable behaviour and overestimation of the number of groups. However, our findings also reveal that the VI loss tends to be biased towards an overestimation of the number of groups. This seems not to be case with the results presented in W&G. We wonder if the approximation the authors introduce may have an impact on the estima-

tion of the number of groups? All things considered, we deem the research question of finding an optimal loss function and associated computational strategy still very open.

2 Quantifying the uncertainty around the estimated Bayes partition

In our experience, the marginal posteriors for the cluster membership variables generally exhibit some degree of multimodality, even after labeling issues have been taken into account. This is one important reason why often Markov chain Monte Carlo sampling methods generally struggle to explore the discrete search space efficiently.

We believe that the same multimodality may also have non-negligible effects on the characterisation of the uncertainty around the estimated optimal Bayes partition. In a nutshell: if, by definition, the credible ball has to include 95% of the posterior mass, it will contain most of the relevant modes, but it may also include many of irrelevant partitions "between" them, in the sense of the loss considered. This would result in a quite heterogeneous set which may be hard to characterise, and where the horizontal and vertical bounds may not be so relevant to the clustering problem. We present a small experiment here to illustrate this point. Here we simulated a data set by sampling from a uniform distribution in the square $[-1, 1] \times [-1, 1]$. We assumed the data followed a Gaussian mixture model and then obtained a posterior sample of partitions using the `R` package `bayesm`. We then applied the methodology proposed by W&G to assess the uncertainty in the estimated Bayes partition and present the output of this experiment in Figure 1. In this case, while the optimal Bayes partition seems very reasonable, having found three contiguous group, the bounds of the credible ball appear quite diverse and "distant" from the actual optimal solution (particularly the horizontal one). We feel that these bounds do not necessarily convey much information regarding which partitions are inside the ball and which are not. Of course, this is a situation where the model is mis-specified, as is the usual case in practice, and this may partially explain the results in Figure 1.

Alternatively, one may instead consider an approach based on the idea of high posterior density regions, and simply list all of the partitions that have posterior probability above a certain threshold. This method would include all of the relevant partitions regardless of their distance from the Bayes partition (in the sense of the distance induced by the Hasse diagram), providing a good representation of what the possible optimal alternatives look like. From a computational perspective, both methods are straightforward to implement once the posterior values and the distances to the Bayes solution are available for all of the partitions sampled.

3 Wider application of mixture models

W&G propose applications of their methodology to Gaussian mixture models. We would like to conclude our discussion by remarking that the method they proposed may be applied in more general mixture modelling contexts, thereby widening their applicability.

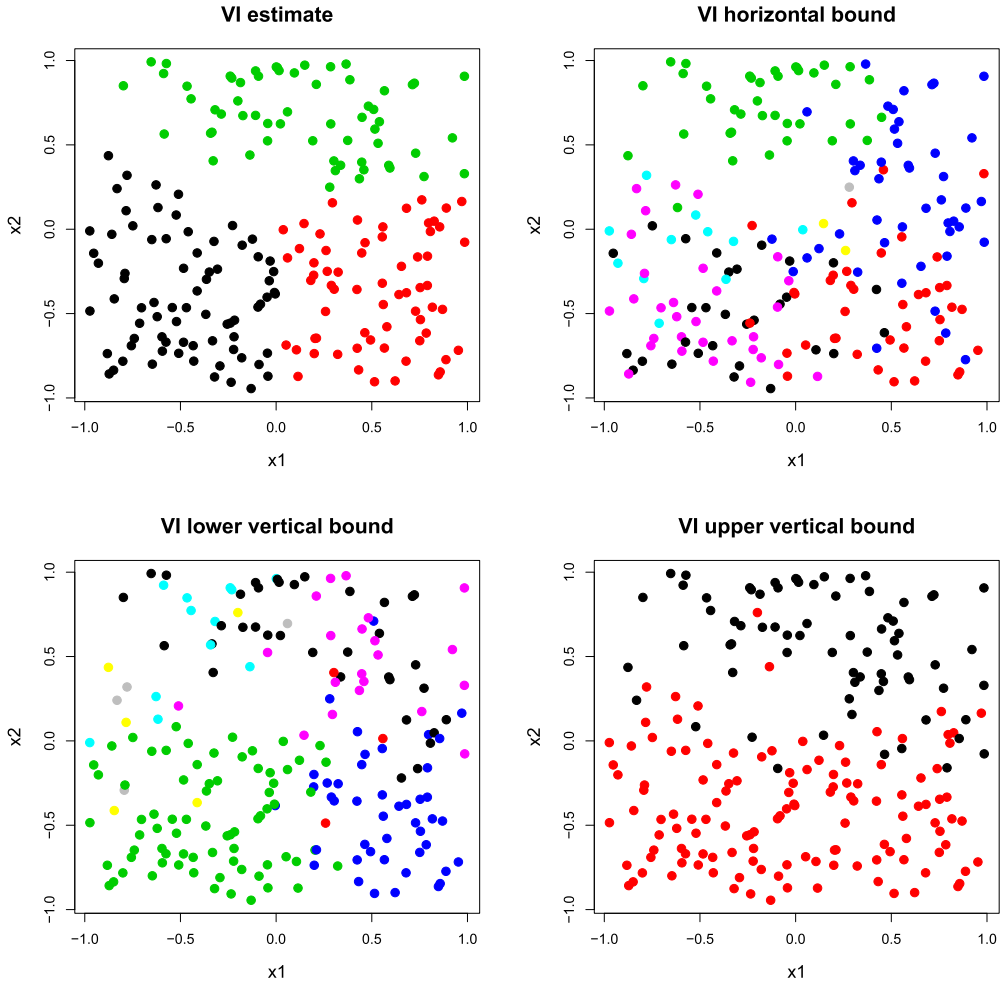


Figure 1: VI loss optimal clustering and credible ball bounds for the simulated uniform data proposed.

For instance, recent research has focused much on mixture models for network data (Daudin et al., 2008). Computationally efficient Markov chain Monte Carlo sampling strategies for network clustering models have been proposed by McDaid et al. (2013) and Wyse and Friel (2012). In R&F, we propose several applications of the decision theoretic framework to Gaussian mixture models, but also to stochastic block models for networks, and to latent block models for bipartite networks. Furthermore, mixed-membership models (Airoldi et al., 2008) extend the basic clustering structures to partial memberships, where nodes of the network may distribute their affiliation among the groups. Extending the decision theoretic framework proposed by W&G to these contexts would be a great next step forward.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). "Mixed membership stochastic blockmodels." *Journal of Machine Learning Research*, 9(Sep): 1981–2014. 597

Daudin, J. J., Picard, F., and Robin, S. (2008). "A mixture model for random graphs." *Statistics and Computing*, 18(2): 173–183. MR2390817. doi: https://doi.org/10.1007/s11222-007-9046-7. 597

Lau, J. W. and Green, P. J. (2007). "Bayesian model-based clustering procedures." *Journal of Computational and Graphical Statistics*, 16(3): 526–558. MR2351079. doi: https://doi.org/10.1198/106186007X238855. 594

McDaid, A. F., Murphy, T. B., Friel, N., and Hurley, N. J. (2013). "Improved Bayesian inference for the stochastic block model with application to large networks." *Computational Statistics & Data Analysis*, 60: 12–31. MR3007016. doi: https://doi.org/10.1016/j.csda.2012.10.021. 597

Rastelli, R. and Friel, N. (2017). "Optimal Bayesian estimators for latent variable cluster models." *Statistics and Computing*. doi: https://doi.org/10.1007/s11222-017-9786-y. 594, 595

Wyse, J. and Friel, N. (2012). "Block clustering with collapsed latent block models." *Statistics and Computing*, 22(2): 415–428. MR2865026. doi: https://doi.org/10.1007/s11222-011-9233-4. 597

Contributed comment on Article by Wade and Ghahramani

William Weimin Yoo*

Abstract. I begin my discussion by giving an overview of the main results. Then I proceed to touch upon issues about whether the credible ball constructed can be interpreted as a confidence ball, suggestions on reducing computational costs, and posterior consistency or contraction rates.

Keywords: Bayesian clustering, variation of information, Binder's loss, credible ball, overfitted mixtures, Bayes Lepski.

The authors should be congratulated for producing such an interesting and important work. In the present paper, Wade and Ghahramani (2017) investigated the issues of point estimation and uncertainty quantification for Bayesian clustering analysis. Here, the data density is modelled as a countably infinite mixture and latent variables attaching to each observation are introduced to represent cluster membership. A common prior for the mixing distribution is the Dirichlet process, and they used this as the default prior in the simulations and real data analysis. They derived point estimators through decision theory by considering two different clustering losses/metrics, i.e., Binder's loss (N-invariant version) and variation of information (VI). They endowed the space of partitions with a lattice by including partial order and the covering relation, and this enables them to compare properties of these two metrics and define a consistent notion of closeness between partitions. This latter notion was further used to develop a method to construct credible ball over partitions using the aforementioned metrics. The optimization problem needed to find the point estimate (for VI) is computational demanding and the search space is very high-dimensional. To scale up computations, the authors proposed a greedy search algorithm.

I start my discussion by asking the question whether the credible balls constructed can be interpreted as confidence balls in the frequentist sense? Specifically, do the 95% credible balls based on Binder's loss or VI with their vertical and horizontal bounds, have also approximate 95% frequentist coverage probability (contains the true clustering 95% of the time)? For finite dimensional parameters, we have the Bernstein-von Mises theorem to ensure this equivalence; however in the nonparametric setting as in this paper, this equivalence breaks down and it is in general not true that Bayesian credible ball is also a frequentist confidence ball. It would be very interesting if we can give some theoretical guarantees on coverage for the VI credible ball, or maybe compare the extent of its uncertainty in a simulation with a confidence ball over partitions constructed based on non-Bayesian methods (if there are any). In complex models, it is straightforward to use Markov Chain Monte Carlo (MCMC) samples to construct credible balls, as compared to frequentist methods which rely on complicated asymptotic normality analysis or bootstrap, and hence such comparisons and coverage guarantees

*Mathematical Institute, Leiden University, The Netherlands, yooweimin0203@gmail.com

will provide good incentives for statisticians (particularly non-Bayesians) to use the methods proposed in this paper to do clustering in their own work.

A recurring theme that came up when designing algorithms in the paper is the ability to scale to massive datasets and to speed up computations. Instead of using infinite mixtures which entails searching over the entire partition space, one can use overfitted mixtures as investigated in Rousseau and Mengersen (2011), where one intentionally overfit the model by choosing a larger but finite number of components than necessary and use some sparsity-inducing priors to zero out the unnecessary components. Alternatively, by observing in Table 2 that the number of clusters for the VI credible ball stays constant for the different sample sizes considered, its robust property suggests that we could first try to estimate the correct number of clusters, through MAP (Maximum a posteriori) or the recently proposed Bayes Lepski's method (Yoo and van der Vaart (2018)), and only explore the part of the partition space corresponding to this estimated number of clusters.

I totally agree with the authors that we need results on posterior consistency and contraction rates, in order to fully resolve the ambiguity caused by the positive results of the present paper and the negative results of Miller and Harrison (2014). Question of interests include characterizing the rate at which the number of clusters estimated under the VI posterior approaches the true number, and whether this rate is optimal. In addition, it would also be interesting to study miss-classification errors and how they grow with sample size or depend on the chosen loss function. A deeper understanding of these issues will help statisticians choose the right priors and design algorithms to control these errors.

The present paper proposes a very promising method to obtain point estimate and uncertainty quantification for Bayesian cluster analysis, which is a great improvement in terms of interpretability over posterior similarity matrices commonly considered in the literature. I envision that the lattice-based framework introduced here can be extended to other settings as well, e.g., multiple membership clusters, and I am certain this work will further spur research in these areas.

References

Miller, J. and Harrison, M. (2014). "Inconsistency of Pitman-Yor process mixtures for the number of components." *Journal of Machine Learning Research*, 15: 3333–3370. MR3277163. 600

Rousseau, J. and Mengersen, K. (2011). "Asymptotic behaviour of the posterior distribution in overfitted mixture models." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73(5): 689–710. MR2867454. doi: https://doi.org/10.1111/j.1467-9868.2011.00781.x. 600

Wade, S. and Ghahramani, Z. (2017). "Bayesian Cluster Analysis: Point Estimation and Credible Balls." *Bayesian Analysis*, 1–29. Advance publication. 599

Yoo, W. W. and van der Vaart, A. W. (2018). "The Bayes Lepski's Method and Credible Bands through Volume of Tubular Neighborhoods." arXiv:1711.06926 [math.ST]. 600

Contributed comment on Article by Wade and Ghahramani*

Sylvia Frühwirth-Schnatter†, Bettina Grün‡, and Gertraud Malsiner-Walli§

We would like to congratulate the authors on addressing the difficult problem of summarizing the posterior distribution of partitions. The high dimensionality of the partition space and the low support for any single partition make this problem very challenging. To our knowledge, their approach is the first one, which tries to systematically estimate bounds for confidence regions of the partition posterior. In this comment, we would like to emphasize that their proposed procedure is not only useful for Bayesian nonparametric mixture models, but can also prove very useful for finite mixture models.

1 Sparse finite mixture models

As opposed to common belief which is also expressed in the introduction of the paper, the number of clusters in the data is not necessarily fixed a priori for finite mixtures and can be estimated from the data, in particular when using sparse finite mixture models (Malsiner-Walli et al., 2016, 2017; Frühwirth-Schnatter and Malsiner-Walli, 2018). The authors' procedure for summarizing uncertainty in the posterior of the partitions is particularly appealing for such sparse finite mixture models where the number of data clusters is random. Data clusters in this context refer to clusters of data points induced by the partitions. Sparse finite mixture models are based on an overfitting finite mixture distribution with the number K of components exceeding the number of data clusters, in combination with a very small value for the hyperparameter e_0 of the Dirichlet prior on the mixture weights. Such a setting encourages partitions with less clusters than there are components, implying that during Markov chain Monte Carlo (MCMC) sampling data points are only assigned to a subset of the components and some components are left empty. Sampling from the posterior of the partitions for sparse finite mixture models is straightforward as standard MCMC sampling schemes developed for finite mixtures can be used.

2 Illustration using example 1 of the paper

To illustrate how the proposed inference tools can be used for post-processing the partitions sampled from a sparse finite mixture model, we fit a sparse finite mixture model

*This research was funded by the Austrian Science Fund (FWF): P28740.

†Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien, Austria, sylvia.fruehwirth-schnatter@wu.ac.at

‡Department of Applied Statistics, Johannes Kepler University Linz, Austria, Bettina.Gruen@jku.at

§Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien, Austria, gertraud.malsiner-walli@wu.ac.at

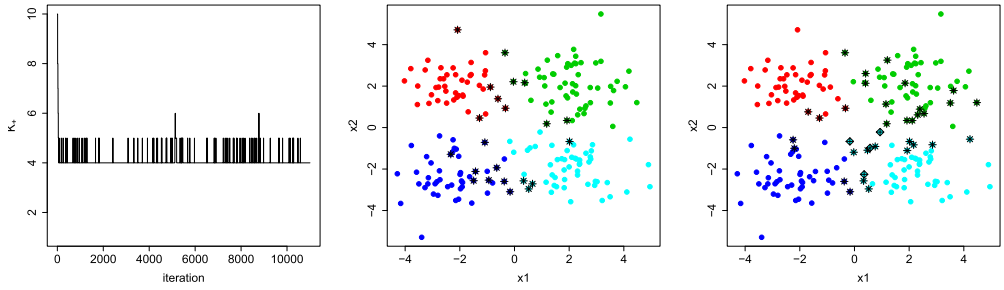


Figure 1: Left: trace plot of the number of data clusters K_+ including burn-in. Middle and right: scatter plots of the data indicating the final partition $c^\star$ using four different colors except for the data points marked with black stars which belong to different clusters in the upper and lower bounds using $\alpha = 0.50$ (middle) and $\alpha = 0.05$ (right).

Partition	k_N^*	N_I	Cluster sizes	ARI
Upper bound $\alpha = 0.05$	4	17	65, 50, 45, 40	0.78
Upper bound $\alpha = 0.50$	4	16	60, 59, 42, 39	0.80
$c^\star$	**4**	**6**	**56, 54, 47, 43**	**0.92**
Lower bound $\alpha = 0.50$	5	13	55, 51, 48, 45, 1	0.84
Lower bound $\alpha = 0.05$	5	30	50, 48, 47, 41, 14	0.71

Table 1: The final partition $c^\star$ and boundary partitions of the credible balls. For each partition, the number of clusters k_N^*, the number of misclassified data points N_I, the cluster sizes (in decreasing order) and the adjusted Rand indices (ARI) are reported.

with $K = 10$ in combination with $e_0 = 0.01$ to the data set of their example 1. The priors on the component means and variances follow Frühwirth-Schnatter (2006). Gibbs sampling with data augmentation is initialized by assigning data points to all available components and 10,000 posterior samples are drawn after a burn-in of 1,000 iterations.

For each partition drawn during MCMC sampling, the number of data clusters K_+ induced by the non-empty components is determined and the corresponding trace plot is shown in Figure 1. During burn-in, most components become empty and the sampler iterates between partitions with 4 and 5 data clusters. These partitions (excluding the burn-in) are summarized based on the VI loss using the R package mcclust.exe.

Table 1 shows characteristics of the estimated final partition $c^\star$ and reports the upper and lower bounds for $\alpha = 0.50$ and $\alpha = 0.05$. If α decreases, the cluster size of the largest cluster increases for the upper bounds and the cluster size of the smallest data cluster increases for the lower bounds. This behavior might be expected from the order relation discussed in Property 5 of the paper. The adjusted Rand index (ARI) measures the correspondence between the true clustering and each of the partitions. The partition which minimizes the expected VI loss has the highest ARI. The ARI decreases with decreasing α for both, the lower as well as the upper bounds. Figure 1 illustrates the final partition in a scatter plot of the data using different colors for the data clusters identified. In addition, data points which are not consistently allocated to

the same clusters in the final and boundary partitions using either $\alpha = 0.50$ or $\alpha = 0.05$ are marked with black stars. These data points could be regarded as "uncertain" in their cluster membership.

3 Final remarks

The close relationship between Bayesian cluster analysis based on finite and infinite mixtures is again demonstrated by indicating how inference tools developed for the infinite case also prove useful in the finite case. For finite mixtures, the proposed inference tools have a number of advantages for post-processing samples from the partition posterior: (1) no model selection needs to be performed, (2) no identified model where label-switching is resolved is required and (3) uncertainty estimates for the partition posterior are readily available based on the credible balls. We hope that future work on Bayesian cluster analysis follows our example and develops and demonstrates inference tools not only for the infinite mixture case, but also considers the finite case.

References

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. New York: Springer. MR2265601. 602

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2018). "From Here to Infinity – Sparse Finite versus Dirichlet Process Mixtures in Model-Based Clustering." URL http://arxiv.org/pdf/1706.07194.pdf. 601

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). "Model-Based Clustering based on Sparse Finite Gaussian Mixtures." *Statistics and Computing*, 26(1): 303–324. MR3439375. doi: https://doi.org/10.1007/s11222-014-9500-2. 601

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017). "Identifying Mixtures of Mixtures Using Bayesian Estimation." *Journal of Computational and Graphical Statistics*, 26(2): 285–295. MR3640186. doi: https://doi.org/10.1080/10618600.2016.1200472. 601

Contributed comment on Article by Wade and Ghahramani

Roberto Casarin* and Stefano Tonellato†

Abstract. This article discusses the Wade and Ghahramani's (2017) paper on a new estimator for clustering structures based on the variation of information (VI) metric. The present discussion focuses on the estimation of concentration parameter of the Dirichlet process. In estimating the clustering structure, the concentration parameter is integrated out and the marginal posterior distribution of the random partition is used to evaluate the posterior loss. Here we propose to use the optimal VI for model selection.

MSC 2010 subject classifications: Primary 62G05, 62F15, 60G57, 60G09.

Keywords: Bayesian nonparametrics, Dirichlet process prior, model selection, variation of information criterion.

1 Introduction

The authors are to be congratulated on their excellent intuition, which has culminated in the development of a new Bayesian point estimator for clustering structure which can find applications in many Bayesian nonparametrics studies. Their Bayesian approach to clustering estimation is inspired by the paper of Meilă (2007). The proposed model provides an alternative to the Dahl (2006) method widely used in the Bayesian nonparametric literature.

In the application to the galaxy data we assume the same DP mixture model as in equation (5) of the paper and the same prior setting $\mu_0 = \bar{x}$, $c = 1/2$, $a = 2$ and $b = s^2$. Instead of estimating α we assume the concentration parameter α takes values in the finite set $A = \{\alpha_1, \ldots, \alpha_n\}$ and for each element α_j of this regular grid we evaluate the partition posterior distribution $p(\mathbf{c}|y_{1:N}, \alpha_j)$ given by

$$p(\mathbf{c}|y_{1:N}, \alpha_j) \propto \frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j + N)} \alpha_j^{k_N} \prod_{j=1}^{k_N} \Gamma(n_{n_j}) m(\mathbf{y}_j),$$

where $m(\mathbf{y}_j)$ is the marginal likelihood of the observations in the j-th partition. For each value of $\alpha \in A$ we run the Gibbs sampler as in the algorithm 8 of Neal (2000) and find the optimal value of the VI criterion at α (VIC_α) as

$$\text{VIC}_\alpha = \min_{\hat{\mathbf{c}}} \int L(\mathbf{c}, \hat{\mathbf{c}}) p(\mathbf{c}|y_{1:N}, \alpha) d\mathbf{c}.$$

*Department of Economics, University Ca' Foscari of Venice, Cannaregio 873, 30121, Venezia, Italy, r.casarin@unive.it

†Department of Economics, University Ca' Foscari of Venice, Cannaregio 873, 30121, Venezia, Italy, stone@unive.it

c	VIC_α			VI
	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 1.5$	
1/2	0.61	0.72	0.83	0.74
1/10	0.23	0.32	0.41	0.29

Table 1: Optimal VIC_α for $\alpha \in \{0.5, 1, 1.5\}$ and different values of c.

The VIC_α obtained are given in Table 1 for $\alpha \in \{0.5, 1, 1.5\}$. The optimal VI value with α integrated out, using a $\mathcal{G}a(1, 1)$ prior, is reported in the last column. The minimumVIC_α is attained for $\alpha = 0.5$ and it is always smaller than the integrated VI. The result suggests the VIC_α is favoring smaller values of the concentration parameter.

2 Conclusion

In their paper, the authors sketch a number of possible extensions. We would suggest as further research line also the combination of posterior clustering probabilities obtained from the different models. It is clear that this is an exciting and stimulating work. We are therefore very pleased to be able to propose the vote of thanks to the authors for their work.

References

Dahl, D. B. (2006). "Model-based clustering for expression data via a Dirichlet process mixture model." In Do, K.-A., P. Muller, P., and Vannucci, M. (eds.), *Bayesian Inference for Gene Expression and Proteomics*, 201–218. Cambridge University Press. MR2706330. 604

Meilă, M. (2007). "Comparing clusterings—an information based distance." *Journal of Multivariate Analysis*, 98(5): 873–895. MR2325412. doi: https://doi.org/10.1016/j.jmva.2006.11.013. 604

Neal, R.M. (2000). "Markov Chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9:249–265. 604

Wade, S. and Ghahramani, Z. (2017). "Bayesian Cluster Analysis: Point Estimation and Credible Balls." *Bayesian Analysis*. 604

Acknowledgments

This research used the multiprocessor cluster system at University Ca' Foscari of Venice (SCSCF).

Contributed comment on Article by Wade and Ghahramani

Eduard Belitser* and Nurzhan Nurushev†‡

We would like to congratulate the authors on an impressive paper that solves an open problem on uncertainty quantification in cluster analysis from a practical point of view. Let us first summarize some key ides of the present paper. Due to the huge dimension of the partition space in Bayesian nonparametric cluster analysis, one of the main problems in Bayesian cluster analysis is how to appropriately summarize the posterior. This problem in the present paper is addressed by providing tools to obtain a point estimate of clustering based on the posterior and describe uncertainty around this estimate via the 95% credible ball. The computation of the point estimate $\mathbf{c}^*$ is based on the greedy search algorithm and Hasse diagram, which can be used for both the variation of information and Binder's loss. In simulation study the authors construct a credible ball of a given credible level $1-\alpha, \alpha \in [0,1]$, defined as $B_{\epsilon^*}(\mathbf{c}^*) = \{\mathbf{c} : d(\mathbf{c}^*, \mathbf{c}) \le \epsilon^*\}$, where ϵ^* is the smallest ϵ such that $P(B_\epsilon(\mathbf{c}^*)|\mathcal{D}) \ge 1-\alpha$.

However, the practical results of the present paper leads to the question of whether the credible ball $B_{\epsilon^*}(\mathbf{c}^*)$ is "optimal". Namely, does the credible ball $B_{\epsilon^*}(\mathbf{c}^*)$ lead to *confidence*? The point estimate $\mathbf{c}^*$ can be very close to, or far away from the true clustering $\mathbf{c}$, without us knowing the actual distance. One would like to have some sort of quantification for the reliability of the estimator $\mathbf{c}^*$, which can be seen as the problem of constructing *confidence* balls for $\mathbf{c}^*$. Confidence balls are a type of set estimates intended to quantify the accuracy of the estimator. The size of the ball quantifies the level of uncertainty of the estimator $\mathbf{c}^*$.

Let us specify the optimality framework for confidence balls. Assume that any partition $\mathbf{c}$ belongs to some functional class $\mathbf{C}_\beta$ indexed by unknown structural parameter $\beta \in \mathcal{B}$ (e.g., number of clusters). Denote the probability measure of the data $\mathcal{D}$ by $\mathrm{P}_{\mathbf{c}} = \mathrm{P}_{\mathbf{c}}^{(N)}$, the minimax concentration rate over $\mathbf{C}_\beta$ by $r_{N,\beta}$. The goal is to construct such a confidence ball $B_{C\epsilon^*}(\mathbf{c}^*) = \{\mathbf{c} : d(\mathbf{c}^*, \mathbf{c}) \le C\epsilon^*\}$ that for any $\alpha_1, \alpha_2 \in (0,1]$ there exist $C, c > 0$ such that

$$\sup_{\mathbf{c}\in\mathbf{C}_0} \mathrm{P}_{\mathbf{c}}\big(\mathbf{c} \notin B_{C\epsilon^*}(\mathbf{c}^*)\big) \le \alpha_1, \quad \sup_{\mathbf{c}\in\mathbf{C}_1} \mathrm{P}_{\mathbf{c}}\big(\epsilon^* \ge c r_{N,\beta}\big) \le \alpha_2, \tag{1}$$

for some $\mathbf{C}_0, \mathbf{C}_1 \subseteq \mathbf{C}_\beta$ and all $\beta \in \mathcal{B}$. The minimax concentration rate $r_{N,\beta}$ is a benchmark for the effective radius of the confidence ball $B_{C\epsilon^*}(\mathbf{c}^*)$. The first expression in (1) is called *coverage relation* and the second *size relation*. It is desirable to have the coverage and size relations to be hold for the biggest $\mathbf{C}_0, \mathbf{C}_1$. For example, if we insist on overall uniformity $\mathbf{C}_0 = \mathbf{C}_1 = \mathbf{C}_\beta$, then the results in Li (1989) and Cai and Low

*Department of Mathematics, VU Amsterdam, e.n.belitser@vu.nl
†Korteweg-de Vries Institute for Mathematics, University of Amsterdam, n.nurushev@uva.nl
‡Research funded by the Netherlands Organisation for Scientific Research NWO.

(2004) (more refined versions are in Baraud (2004) say basically that the radius of confidence ball cannot be of a bigger order than $N^{1/4}$. Many good confidence balls cannot be optimal in this sense (called "honest" in some papers), e.g., in sparse normal means model. Instead, it makes sense to sacrifice in the set $\mathbf{C}_0 = \mathbf{C}_\beta \backslash \mathbf{C}'$, by removing a preferably small portion of "deceptive parameters" $\mathbf{C}'$ from $\mathbf{C}$ so that the optimal minimax rate becomes attainable in the size relation with interesting (preferably "massive") sets $\mathbf{C}_1$ (see Belitser and Nurushev (2017) for details). To the best of our knowledge, it is not known whether it is possible to construct a confidence ball simultaneously with a good coverage and optimal size adaptively to some scale $\mathbf{C}_\beta$ in the studied model. This is a challenging problem and of great importance to our understanding of uncovering partitions $\mathbf{c}$.

Admittedly, the above optimality framework is formulated from the frequentists perspective whereas the authors pursue a purely Bayesian approach. An advantage is that such a framework allows to compare different procedures. We wonder whether the authors could come up with a general idea of how to compare different Bayesian procedures from the purely Bayesian perspective. We understand that the authors were mainly focused on the practical results related to the construction of credible balls, but we hope this comment will inspire the authors and other people to work on this interesting problem in the future.

We would like to finish with the question to the authors whether the point estimates studied in the present paper can be used for the Hamming loss function. If it was possible then it might be interesting to create a new simulation study in the stochastic block model and compare the radius ϵ^* of credible ball $B_{C\epsilon^*}(\mathbf{c}^*)$ (based on the Hamming loss function) with the minimax rate for community detection problem studied in Zhang and Zhou (2016). Then one could answer the question whether the radius ϵ^* of credible ball $B_{C\epsilon^*}(\mathbf{c}^*)$ is optimal in this sense or not.

References

Baraud, Y. (2004). "Confidence balls in Gaussian regression." *Annals of Statistics*, 32(2): 528–551. MR2060168. doi: https://doi.org/10.1214/009053604000000085. 607

Belitser, E. and Nurushev, N. (2017). "Needles and straw in a haystack: robust confidence for possibly sparse sequences." https://arxiv.org/abs/1511.01803. 607

Cai, T. T. and Low, M. G. (2004). "An adaptation theory for nonparametric confidence intervals." *Annals of Statistics*, 32(5): 1805–1840. MR2102494. doi: https://doi.org/10.1214/009053604000000049. 606

Li, K.-C. (1989). "Honest Confidence Regions for Nonparametric Regression." *Annals of Statistics*, 17(3): 1001–1008. MR1015135. doi: https://doi.org/10.1214/aos/1176347253. 606

Zhang, A. Y. and Zhou, H. H. (2016). "Minimax rates of community detection in stochastic block models." *Annals of Statistics*, 44(5): 2252–2280. MR3546450. doi: https://doi.org/10.1214/15-AOS1428. 607

Contributed comment on Article by Wade and Ghahramani

Reza Mohammadi*

I would first like to congratulate Dr Wade and Professor Ghahramani for their excellent exposition of the Bayesian nonparametric cluster analysis by developing point estimates and credible sets to summarize the posterior of the clustering structure. Their method is based on a greedy search algorithm to locate the optimal partition based on Hasse diagram, which can be used for both the *variation of information* and the *Binder's loss*. Here, I would like to contribute to the discussion by suggesting a comparison with the Bayesian parametric methods of finite mixture distributions based on the trans-dimensional Markov chain Monte Carlo (MCMC) algorithms.

Comparison with finite mixture distributions

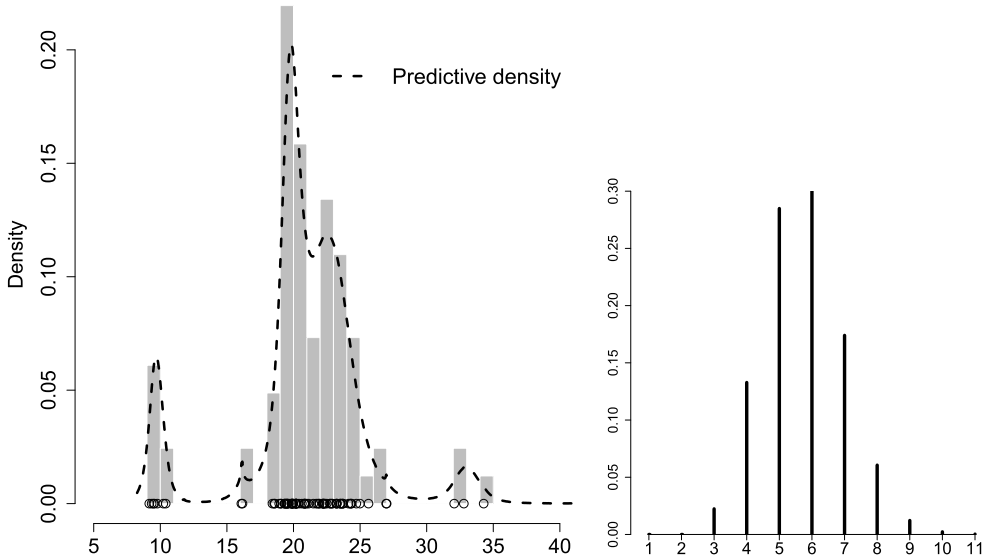
This paper illustrates the high potential of the Bayesian nonparametric cluster analysis. Here, I focus on the Bayesian parametric approaches for finite mixture distributions based on trans-dimensional MCMC sampling algorithms.

In the Bayesian analysis of finite mixture distributions with an unknown number of components, the main problem is sampling from the posterior distributions. Since the number of components is unknown, it requires advanced search algorithms which can potentially move in the model space. Transdimensional search algorithms explore the model space when the model does not have a fixed dimension; common ones are the reversible jump MCMC by Green (1995) and birth-death MCMC Stephens (2000). In the context of finite mixture distributions, these methods have been used by Green (1995); Stephens (2000); Mohammadi et al. (2013) and in the case of graphical models Mohammadi and Wit (2015).

To compare the performance of the Bayesian nonparametric method, proposed in the paper, with the Bayesian parametric method based on finite mixture distributions, we apply the finite mixture of normal distribution for galaxy data with the same scenario as in subsection 6.2 of the paper. We run the birth-death MCMC algorithm proposed by Stephens (2000) using the R-package bmixture (Mohammadi, 2018), function *bmixnorm()*, with 10K samples with 10K burnin.

Figure 1(b) shows that the data came from 5 or 6 clusters. The sum of the estimated posterior probability of the number of clusters for the case of 4 up to 8 clusters is 0.95, which can be considered as the 95 percent confident interval for the number of clusters; The results are comparable with the results in Table 5 of this paper.

*Dept. of Operation Management, University of Amsterdam, Netherlands, a.mohammadi@uva.nl; url: http://www.uva.nl/profile/a.mohammadi



(a) Data histogram and predictive density (b) Posterior probibility of clusters

Figure 1: (a) Histogram of galaxy data and estimated density based on the finite mixture of Normal distribution and (b) the estimation of the posterior distribution of the number of clusters.

It would be quite useful if the authors could comment on the comparison of their approach to finite mixture distributions as Bayesian parametric approaches and the possibility of replacing greedy search algorithm in their Bayesian nonparametric framework with the trans-dimensional MCMC algorithms.

References

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82(4): 711–732. MR1380810. doi: https://doi.org/10.1093/biomet/82.4.711. 608

Mohammadi, A., Salehi-Rad, M., and Wit, E. C. (2013). "Using mixture of Gamma distributions for Bayesian analysis in an M/G/1 queue with optional second service." *Computational Statistics*, 28: 683–700. MR3064474. doi: https://doi.org/10.1007/s00180-012-0323-3. 608

Mohammadi, A. and Wit, E. C. (2015). "Bayesian Structure Learning in Sparse Gaussian Graphical Models." *Bayesian Analysis*, 10(1): 109–138. MR3420899. doi: https://doi.org/10.1214/14-BA889. 608

Mohammadi, R. (2018). *bmixture: Bayesian Estimation for Finite Mixture of Distributions*. R package version 0.6. 608

Stephens, M. (2000). "Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods." *Annals of Statistics*, 40–74. MR1762903. doi: https://doi.org/10.1214/aos/1016120364. 608

Contributed comment on Article by Wade and Ghahramani

Julyan Arbel*, Riccardo Corradin†, and Michał Lewandowski*

Abstract. We propose a simulation study to emphasise the difference between Variation of Information and Binder's loss functions in terms of number of clusters estimated by means of (1) the use of the Markov chain Monte Carlo (MCMC) output only and (2) a "greedy" method.

Wade and Ghahramani's paper is a very neat contribution to Bayesian cluster analysis in at least two respects: (i) by formalizing cluster credible coverage via Hasse diagrams, and (ii) by recasting the problem in a decision theory framework, with tangible improvements brought by the Variation of Information (VI) loss function (Meilă, 2007) over Binder's (Binder, 1978; Dahl, 2006).

We propose a simulation study implementing two algorithms provided by Wade and Ghahramani's package mcclust.ext for finding the argument minimizing the posterior expected loss: (1) the *draw algorithm*, which restricts the minimization problem to the MCMC output, and (2) the *greedy algorithm*, which is more reliable as it also scans the neighbouring clusters of the MCMC output, but with a larger computational cost. While increasing the sample size, we point out the radically different behavior of the number of clusters estimated under VI and Binder, especially with the greedy algorithm.

Our simulation study is based on the same data generation as in the first example of Section 6.1 in Wade and Ghahramani (2017): a mixture of four Gaussian distributions equally weighted with means $(\pm 2, \pm 2)$ and identity covariance matrix. We estimated the model using a marginal approach provided by BNPmix[1] R package. We synthesised the output with mcclust.ext package.[2] The Dirichlet process mixture model was estimated with mass parameter fixed to 1, and by specifying an independent base measure on locations and scales, with a 0-vector prior mean for the location component and an identity matrix prior mean for the scale component (25 000 iterations with 5 000 burn-in period). We considered four different sample sizes $n = \{20, 40, 100, 300\}$.

The results are shown in Figure 1. With the draw algorithm, the cluster estimates under both losses are quite close in terms of number of clusters. In contrast, the greedy algorithm leads to cluster estimates obtained via Binder's loss function with excessive size, while that obtained via VI remains coherent with the number of components of the model (four).

Similarly to the authors' finding, ours' indicates that Binder's loss function exhibits an undesirable property of overestimating the number of clusters (Miller and Harrison,

*Univ. Grenoble Alpes, Inria, CNRS, LJK, 38000 Grenoble, France, julyan.arbel@inria.fr; michal.lewandowski@inria.fr

†DISMEQ, University of Milano Bicocca, 20126 Milano MI, Italy, riccardo.corradin@unimib.it

[1]Package available at https://github.com/rcorradin/BNPmix, can be installed via devtools.

[2]Code of the simulation study available at https://github.com/rcorradin/WGdiscussion.

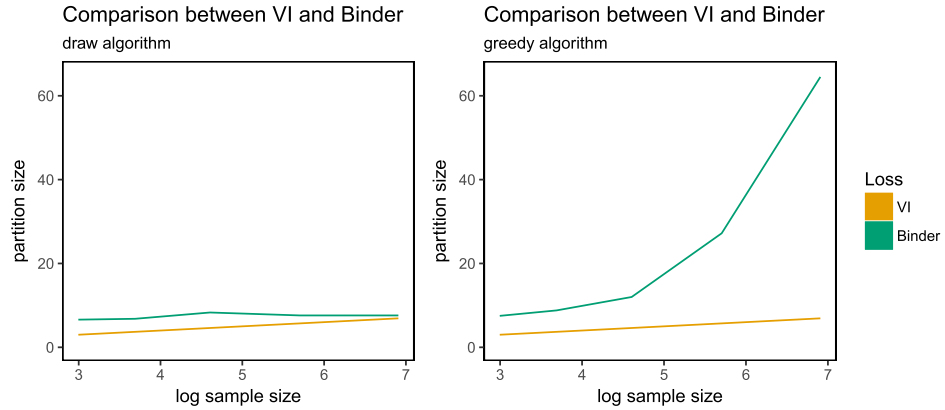


Figure 1: Size of the cluster estimate under VI (yellow line) and Binder (green light). Left: *draw algorithm*. Right: *greedy algorithm*.

2013, 2014). Variation of Information tends to lessen this problem. As alluded to by the authors, a theoretical study of the asymptotic behavior of the VI estimator would be very timely. Especially in light of the recent contribution by Rajkowski (2016) about the asymptotic behavior of the cluster estimator under the $0-1$ loss (MAP estimator).

References

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika*, 65(1):31–38. MR0501592. doi: https://doi.org/10.1093/biomet/65.1.31. 611

Dahl, D. B. (2006). Model-based clustering for expression data via a dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, pages 201–218. MR2706330. 611

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895. MR2325412. doi: https://doi.org/10.1016/j.jmva.2006.11.013. 611

Miller, J. W. and Harrison, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in neural information processing systems*, pages 199–206. 611

Miller, J. W. and Harrison, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *The Journal of Machine Learning Research*, 15(1):3333–3370. MR3277163. 611

Rajkowski, Ł. (2016). Analysis of MAP in CRP Normal-Normal model. *arXiv preprint arXiv:1606.03275*. 612

Wade, S. and Ghahramani, Z. (2017). Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Analysis*. 611

Contributed comment on Article by Wade and Ghahramani

Bernardo Nipoti* and Weining Shen†

We vividly congratulate the authors for providing very interesting insight on how to summarize posterior belief on the space of partitions, and characterize its uncertainty. In this comment we focus on the latter point and we build upon the authors' definition and characterization of credible balls of partitions. Specifically, we would like to suggest that an alternative formulation of such quantities might be considered and might provide additional insight, useful to the difficult task of summarizing a posterior belief on the space of partitions. We articulate our comment in two points.

1. Where to center? Commonly adopted measures of posterior uncertainty, such as highest posterior density or quantile-based posterior intervals, are defined independently of any posterior estimator, which might vary based on the choice of the loss function. Similarly, while the authors define a credible ball of partitions $B_{\epsilon^*}(\mathbf{c}^*)$ as centered around a point estimate $\mathbf{c}^*$, we would like to observe that such definition could be tweaked so to make it independent of any point estimator. This could be achieved by centering the credible ball around the true partition $\mathbf{c}_0$ and by averaging across all possible true partitions. That is, following the same idea used in Section 2.2 of Wade and Ghahramani (2017), we can define a credible ball B_{ϵ^*} of credible level $1-\alpha$ as

$$B_{\epsilon^*} = \{\mathbf{c} : \mathbb{E}_{\mathbf{c}_0}[d(\mathbf{c},\mathbf{c}_0) \mid \mathcal{D}] \leq \epsilon^*\} \approx \left\{\mathbf{c} : \frac{1}{M}\sum_{m=1}^{M} d(\mathbf{c},\mathbf{c}^m) \leq \epsilon^*\right\},$$

where $\{\mathbf{c}^m\}_{m=1}^M$ is the set of partitions visited by the Markov chain Monte Carlo (MCMC) algorithm. An estimate of ϵ^* in this case is obtained from the MCMC output by choosing ϵ^* as the smallest $\epsilon \geq 0$ such that the average of pairwise distances satisfies $\frac{2}{M(M-1)}\sum_{n=1}^{M}\sum_{m>n}^{M}\mathbf{1}(d(\mathbf{c}^n,\mathbf{c}^m) \leq \epsilon) \geq 1-\alpha$. Compared with the authors' credible ball definition, the alternative construction is computationally more expensive, but it does not rely on the correct estimation of a posterior point estimator. It would be interesting to investigate the theoretical properties (e.g. frequentist coverage) of these two different definitions.

2. To center or not to center? Alternatively, credible balls could be defined without resorting to the idea of centering them at any given partition, being it a point estimate $\mathbf{c}^*$ or the true $\mathbf{c}_0$. A simple approach consists in sorting the MCMC samples via a certain measure, such as entropy. More specifically, if we call $H(\mathbf{c})$ the entropy of a partition $\mathbf{c} = (C_1,\ldots,C_{k_N})$, defined as $H(\mathbf{c}) = \log(N) - 1/N\sum_{j=1}^{k_N} n_j\log(n_j)$, where $n_j = |C_j|$ is the number of data points in cluster j, then a entropy-based credible ball $B_{\epsilon^*}^{(H)}$ can

*School of Computer Science and Statistics, Trinity College Dublin, Ireland, nipotib@tcd.ie
†Department of Statistics, University of Califorinia, Irvine, United States, weinings@uci.edu

be defined as

$$B_{\epsilon^*}^{(H)} = \{\mathbf{c} : H_L \leq H(\mathbf{c}) \leq H_U\},$$

where $H_L := H(\mathbf{c}_H^{(l)})$ and $H_U := H(\mathbf{c}_H^{(u)})$, and $\mathbf{c}_H^{(l)}$ and $\mathbf{c}_H^{(u)}$ are empirical quantiles of the sorted set of observed partitions, such that $M^{-1}\sum_{m=1}^{M}\mathbf{1}\{H_L \leq H(\mathbf{c}^{(m)}) \leq H_U\} \geq 1-\alpha$. As a by-product a point estimate $\mathbf{c}_H^*$ can be obtained by considering the posterior median partition. Following Example 2 in Section 6.1 of the main paper, we considered the simulated data from a mixture of four bivariate normal distributions with unequal covariance matrices, used the posterior samples output from the R package "mcclust.ext", and obtained a quantile-based 95% credible ball $B_{\epsilon^*}^{(H)}$. The plots for the posterior median $\mathbf{c}_H^*$, upper bound $\mathbf{c}_H^{(u)}$ and lower bound $\mathbf{c}_H^{(l)}$ are presented in Figure 1. Compared with Figure 9 in the main paper, the clustering results look quite similar, with a lower bound $\mathbf{c}_H^{(l)}$ showing a moderately smaller number of clusters than the VI lower vertical bound. For both methods, the effect of outliers and over-estimation of the number of clusters in the lower bound is apparent. This is expected since Dirichlet mixture priors are known to provide consistent density estimation (Shen et al., 2013), while overestimating the number of clusters (Miller and Harrison, 2013). It would be interesting to investigate and compare entropy-based and VI vertical bounds, using other prior distributions known to have better inferential properties in terms of clustering, such as normalized random measures (Barrios et al., 2013).

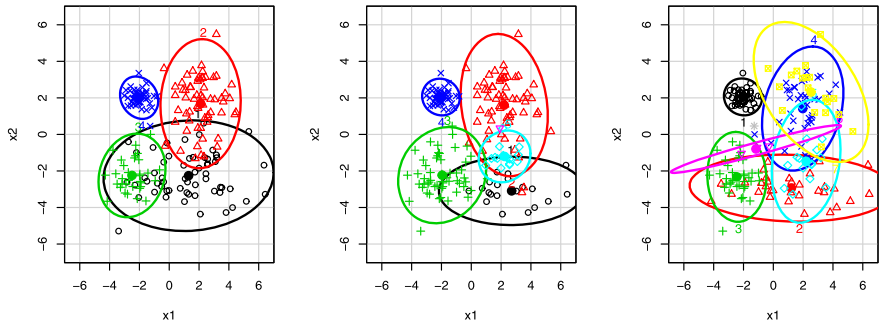


Figure 1: Posterior median $\mathbf{c}_H^*$ (4 clusters, left), upper bound $\mathbf{c}_H^{(u)}$ (5 clusters, middle) and lower bound $\mathbf{c}_H^{(l)}$ (12 clusters, right) of a quantile-based 95% credible ball $B_{\epsilon^*}^{(H)}$.

References

Barrios, E., Lijoi, A., Nieto-Barajas, L. E., and Prünster, I. (2013). "Modeling with normalized random measure mixture models." *Statistical Science*, 313–334. MR3135535. doi: https://doi.org/10.1214/13-STS416. 614

Miller, J. W. and Harrison, M. T. (2013). "A simple example of Dirichlet process mixture inconsistency for the number of components." In *Advances in neural information processing systems*, 199–206. 614

Shen, W., Tokdar, S. T., and Ghosal, S. (2013). "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures." *Biometrika*, 100(3): 623–640. MR3094441. doi: https://doi.org/10.1093/biomet/ast015. 614

Wade, S. and Ghahramani, Z. (2017). "Bayesian cluster analysis: Point estimation and credible balls." *Bayesian Analysis*. 613

Contributed comment on Article by Wade and Ghahramani

Federico Castelletti* and Stefano Peluso†

We congratulate the authors for the insightful paper. We find particularly interesting the adoption of the two ball sizes as concise and easily interpretable measure of uncertainty associated to point estimates on non-standard and large supports. In the present discussion we want to highlight the applicability of the method proposed by the authors beyond the space of random partitions, to Directed Acyclic Graphs (DAGs) and to their Markov equivalence classes, namely to Essential Graphs (EGs).

In Consonni and La Rocca (2012) an Objective Bayes model selection procedure is proposed for DAGs, later extended in Consonni et al. (2017) to covariate-adjusted model selection, and to EGs in Castelletti et al. (2018). Similarly to the space of random partitions, exhaustive enumeration of DAGs and EGs is not feasible (Madigan et al., 1996), since the number of DAGs and EGs grows super-exponentially in the number of nodes (Gillispie and Perlman, 2002). Therefore the posterior probability associated to DAGs and EGs for non-trivial dimensions is only available up to a normalizing constant, requiring appropriate Markov Chain Monte Carlo (MCMC) procedures to perform posterior inference.

The need of summarizing MCMC visits on the space of DAGs and EGs through a point estimate and an uncertainty measure raises, among others, difficulties similar to those in the paper under discussion: the 0-1 loss function adopted in Consonni et al. (2017) resulted in the choice of the modal DAG as point estimate, whilst the point estimator of the EG in Castelletti et al. (2018) is based on the inclusion of edges with marginal posterior inclusion probability higher than 50%. Also, in both cases no measure of uncertainty on the whole structure has been provided, but only MCMC-based uncertainty measures on graph features (as on the number of chain components and v-structures), or rough uncertainty measures based on variation of the edge inclusion probability threshold.

Potentially, both estimates can be improved and the uncertainty around the estimated graph can be measured following the lines suggested in Wade et al. (2017). First, the 0-1 loss function can be replaced by a graph-equivalent of the first metric proposed in the current paper: the error in the classification to the correct cluster becomes the error in the inclusion of an edge. A ball at level $(1-\alpha)$ would be represented by the set of visited EGs whose distance with respect to the point-estimated graph $\hat{\mathcal{G}}$ is less than the threshold ϵ^*. Then, *vertical* and *horizontal bounds*, as defined in the paper, can be also introduced. Coherently, one can for instance consider the graph with the smallest (largest) number of edges that are most distant from $\hat{\mathcal{G}}$ as the vertical lower (upper) bound. Similarly for the horizontal bound.

*Università Cattolica del Sacro Cuore, Milan, Italy, federico.castelletti@unicatt.it
†Università Cattolica del Sacro Cuore, Milan, Italy, stefano.peluso@unicatt.it

Second, the information-based metric from Meilă (2007) is strictly related to the Hamming distance between the binary representations of the clusterings. It could be interesting to find in the graphical context a metric related to the Structural Hamming Distance (the number of edge insertions, deletions or flips needed to transform one graph into another) among (classes of) graphs. Finally, metrics alternative to the 0-1 loss introduce an additional layer of complexity for DAGs and EGs: the estimated graph might lie outside of its support, and therefore some projection onto the proper space could be necessary (Castelletti et al., 2018).

Third, the adoption of the VI metric as a measure of distance between clusterings can be also interesting when model selection of DAG models can be performed in an *interventional* setting (Hauser and Bühlmann, 2015). In general, the size of a Markov equivalence class of DAGs is used as a measure of complexity of causal learning (He et al., 2013). Assuming faithfulness of the observed data to *some* true DAG model, interventions on variables from randomized experiments can be used to improve the identifiability of such a model and then reduce the size of the estimated equivalence class. The problem of optimal choice of *intervention targets* is carried out in He and Geng (2008) from a design of experiments perspective. Specifically, each Markov equivalence class can be partitioned into DAG sub-classes, each with common edges orientations on the intervened nodes. A different intervention target induces a different partition of DAGs, and the VI metric among DAG partitions suggests maximum-entropy-based criteria for optimized choices of targets.

References

Castelletti, F., Consonni, G., Della Vedova, M., and Peluso, S. (2018). "Learning Markov Equivalence Classes of Directed Acyclic Graphs: an Objective Bayes Approach." *Bayesian Analysis*. 616, 617

Consonni, G. and La Rocca, L. (2012). "Objective Bayes Factors for Gaussian Directed Acyclic Graphical Models." *Scandinavian Journal of Statistics*, 39: 743–756. MR3000846. doi: https://doi.org/10.1111/j.1467-9469.2011.00785.x. 616

Consonni, G., La Rocca, L., and Peluso, S. (2017). "Objective Bayes Covariate-Adjusted Sparse Graphical Model Selection." *Scandinavian Journal of Statistics*, 44(3): 741–764. MR3687971. doi: https://doi.org/10.1111/sjos.12273. 616

Gillispie, S. B. and Perlman, M. D. (2002). "The size distribution for Markov equivalence classes of acyclic digraph models." *Artificial Intelligence*, 141: 137–155. MR1935281. doi: https://doi.org/10.1016/S0004-3702(02)00264-3. 616

Hauser, A. and Bühlmann, P. (2015). "Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs." *Journal of the Royal Statistical Society. Series B (Methodology)*, 77: 291–318. MR3299409. doi: https://doi.org/10.1111/rssb.12071. 617

He, Y. and Geng, Z. (2008). "Active learning of causal networks with intervention experiments and optimal designs." *Journal of Machine Learning Research*, 9: 2523–2547. MR2460892. 617

He, Y., Jia, J., and Yu, B. (2013). "Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs." *The Annals of Statistics*, 41: 1742–1779. MR3127848. doi: https://doi.org/10.1214/13-AOS1125. 617

Madigan, D., Andersson, S., Perlman, M., and Volinsky, C. (1996). "Bayesian Model Averaging and Model Selection for Markov Equivalence Classes of Acyclic Digraphs." *Communications in Statistics: Theory and Methods*, 2493–2519. MR1439312. doi: https://doi.org/10.1214/aos/1031833662. 616

Meilă, M. (2007). "Comparing clusterings – an information based distance." *Journal of Multivariate Analysis*, 98: 873–895. MR2325412. doi: https://doi.org/10.1016/j.jmva.2006.11.013. 617

Wade, S., Ghahramani, Z., et al. (2017). "Bayesian Cluster Analysis: Point Estimation and Credible Balls." *Bayesian Analysis*. 616

Contributed comment on Article by Wade and Ghahramani

Bent Natvig*

In Wade et al. (2018), considering Bayesian cluster analysis, it is stated in the introduction that a more elegant solution, than the one presented in some papers, is based on decision theory. Furthermore, it is stated that the question to answer then becomes what is an appropriate loss function over clusterings?

This leads my mind to a paper, Natvig and Tvete (2007), on Bayesian hierarchical space-time modeling of earthquake data. Our aim was to get some insight into where and when large earthquakes occur, or otherwise stated, we were interested in the clustering of large earthquakes. We would like to judge our model more specifically, based upon its ability to avoid two errors by not predicting the large earthquakes and signal false alarms.

In that paper we took a coarse point of view considering grid cells of 50×50 km and time periods of 4 months, which seems suitable for predictions. We discussed different alternatives of a Bayesian hierarchical space-time model inspired by the paper Wikle et al. (2001). For each time period the observations were the magnitudes of the largest observed earthquake within each grid cell. As data we applied parts of an earthquake catalogue provided by The Northern California Earthquake Center where we limited ourselves to the area 32–37 degrees N and 115–120 degrees W, and for the time period January 1981 through December 1999 containing the Landers and Hector Mine earthquakes, respectively measuring 7.3 and 7.1 on the Richter scale. Based on space-time model alternatives one step earthquake predictions for the time periods containing these two events for all grid cells are arrived at.

We constructed a specially designed loss function, weighted over the X spatial cells, that penalizes both these errors. The weight for a given spatial cell, x, for a given prediction period,t, is dependent both upon the magnitude of the observed earthquake, M(x,t), and the distance between the observed earthquake and the predicted value $\hat{M}(x,t,j)$ for sample j. To take signalized earthquakes in neighbouring cells into account we let, suppressing the time notation t, $\hat{M}(x,j)$ be a spatially weighted average of the predictions, $\{M^*(x,j)\}_{x=1}^{X}$, where each of the two, three, or four predictions at a neighbouring cell has half the weight of the one at x and the rest given weight zero. The specially designed loss function, for a given predicted period, is given by:

$$L_2(M,\hat{M}) = \left[\frac{1}{X}\sum_{x=1}^{X}\frac{1}{N}\sum_{j=1}^{N} w(\hat{M}(x,j),M(x))(\hat{M}(x,j)-M(x))^2\right]^{1/2},$$

*Department of Mathematics, University of Oslo, bent@math.uio.no

$$w(\hat{M}(x,j),M(x)) = \begin{cases} 0.2 & \text{if } M(x)\in[0,1) \text{ and } |\hat{M}(x,j)-M(x)|\geq 3, \\ 0.24 & \text{if } M(x)\in[1,2) \text{ and } |\hat{M}(x,j)-M(x)|\geq 2.3, \\ 0.28 & \text{if } M(x)\in[2,3) \text{ and } |\hat{M}(x,j)-M(x)|\geq 1.6, \\ 0.36 & \text{if } M(x)\in[3,4) \text{ and } |\hat{M}(x,j)-M(x)|\geq 1.0, \\ 0.48 & \text{if } M(x)\in[4,5) \text{ and } |\hat{M}(x,j)-M(x)|\geq 0.8, \\ 0.64 & \text{if } M(x)\in[5,6) \text{ and } |\hat{M}(x,j)-M(x)|\geq 0.7, \\ 0.80 & \text{if } M(x)\in[6,7) \text{ and } |\hat{M}(x,j)-M(x)|\geq 0.6, \\ 1.00 & \text{if } M(x)\geq 7 \text{ and } |\hat{M}(x,j)-M(x)|\geq 0.5, \\ 0 & \text{otherwise.} \end{cases}$$

The weights $w(\hat{M}(x,j),M(x)))$ are subjectively designed to punish a lack of ability to predict large earthquakes. This is reflected in the equation above by giving a weight equal to 0.2 when $M(x)\in[0,1)$ increasing to 1, in nondecreasing steps, when $M(x)\geq 7$. The weights are also designed to punish a lack of ability to predict the large earthquakes more accurately than the small ones. Hence, to have a loss when $M(x)\in[0,1)$ the absolute difference between the predicted and the maximal earthquake must be greater or equal to 3. This threshold decreases, in nonincreasing steps, to 0.5 when $M(x)\geq 7$. The basic idea behind the equation above should be reflected in any modification of it.

Let $\hat{M}^p$ be the spatial average of the pth percentiles in the simulated prediction distribution $\{M^*(x,j)\}_{j=1}^N$. We then also compute another specially designed loss function:

$$L_3(M,\hat{M}^p) = \left[\frac{1}{X}\sum_{x=1}^{X} w(\hat{M}^p(x),M(x))(\hat{M}^p(x)-M(x))^2\right]^{1/2}.$$

L_3 is different from L_2 in that we consider the pth percentile rather than all the sampled predictions. This is a sensible loss function to consider in the search of predictions giving the smallest overall loss. It can be shown that the losses L_2 and L_3 are quite small for the less extreme period, somewhat larger for the Hector Mine period and again larger for the Landers period. This is as expected due to the earthquake activity in the various periods. Due to the fact that all weights are less than or equal to 1, the L_2 values are much smaller than the L_1 values given in Natvig and Tvete (2007). Obviously, one will obtain large losses when predicting periods where there is a high earthquake activity, and small losses in less extreme periods. W feel that the estimated loss values are not alarmingly high.

References

Natvig, B. and Tvete, I. F. (2007). "Bayesian Hierarchical Space-time Modeling of Earthquake Data." *Methodology and Computing in Applied Probability*, 9: 89–114. MR2364983. doi: https://doi.org/10.1007/s11009-006-9008-0. 619, 620

Wade, S. and Ghahramani, Z. (2018). "Bayesian Cluster Analysis: Point Estimation and Credible Balls." *Bayesian Analysis*, 13. 619

Wikle, C., Milliff, R., Nychka, D., and Berliner, L. (2001). "Spatio-temporal hierarchial Bayesian modeling: tropical occean surface winds." *Journal of the American Statistical Association*, 96: 382–397. MR1939342. doi: https://doi.org/10.1198/016214501753168109. 619

Rejoinder

Sara Wade* and Zoubin Ghahramani†

We sincerely thank the discussants for the interesting and insightful comments. Our paper investigates approaches to summarize the posterior distribution over the space of partitions. The massive dimension of the partition space and its categorical nature combined with the low posterior probability of any single partition make this a challenging problem. As highlighted by the discussants, there are a number of relevant extensions and open problems.

Properties. As noted by **Monni**, Meilă (2007) details various properties of the VI and other cluster comparison measures, including an equivalent version of Binder's loss. The aim of Section 3 was to highlight the asymmetry of Binder's loss compared to VI, accumulating in Properties 6 and 7. This was not discussed in Meilă (2007), and for completeness, we include a through review of some relevant properties described by Meilă (2007), such as vertical and horizontal collinearity.

Credible balls. To characterize uncertainty in the point estimate, we defined credible balls around the point estimate based on a chosen metric, e.g. VI or Binder's loss. We agree with **Monni** that the posterior similarity matrix is an important tool for assessing uncertainty but emphasize that the credible ball provides additional information that enriches our understanding of uncertainty around the point estimate. We proposed to summarize these balls based on the vertical and horizontal bounds, and as noted by **Monni** (and in the paper), these bounds may consist of more than one partition. In theory, this could be a large number, but in practice, we restrict the bounds to partitions with positive posterior probability, which in our experience provides at most a handful of partitions for each bound. This restriction is for computational purposes, but also serves as a *refinement* of the bound. We encourage further research into the construction of credible balls of partitions. **Friel and Rastelli** propose an interesting idea based on highest posterior density (HPD) regions. In practice, the MCMC may not visit any partition more than once; thus, the suggested HPD region, which considers partitions with posterior probability over a threshold, would contain all or no partitions in this setting. Moreover, relevant summaries of the HPD regions may be needed in practice. **Nipoti and Shen** outline some alternative ideas for defining credible balls. First, they consider centering the ball at the true partition, not the point estimate; this however comes at a computational cost, as it requires evaluating pairwise distances between all MCMC samples. Second, they propose a construction of entropy-based credible balls. These ideas are certainly interesting and more work is needed to investigate and compare the credible balls in theory and simulations. We suspect that the entropy-based credible balls could be quite different from VI credible balls. For example, consider the case with $N = 4$ depicted in Figure 1. We note that the Hasse diagram stretched by the

*University of Warwick, s.wade@warwick.ac.uk
†University of Cambridge, zoubin@eng.cam.ac.uk

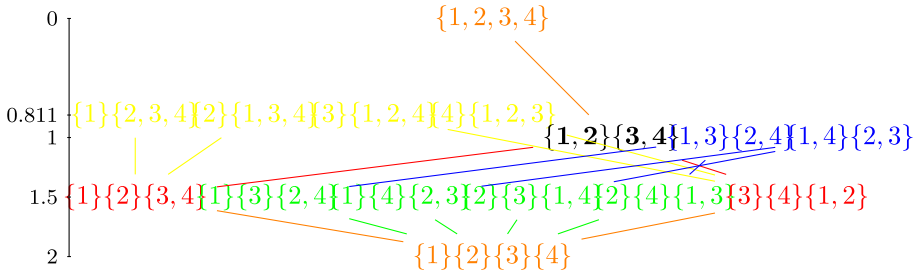


Figure 1: Example of the VI ball around $\mathbf{c} = (\{1,2\},\{3,4\})$, with rainbow color indicating increasing distance from $\mathbf{c}$. The smallest non-trivial credible ball contains all the red clusterings, the next smallest contains the red and orange clusterings, and so on.

entropy results in the same relative positions of the partitions, with the values on the y-axis simply rescaled to be between 0 and $\log(4)$. Suppose that the VI point estimate of $\mathbf{c} = (\{1,2\},\{3,4\})$ has posterior probability of 0.5, then the 50% VI credible ball would contain only that partition. On the other hand, the 50% entropy credible ball would contain the black and blue partitions, that is, all partitions with two clusters of equal size; the blue partitions, however, have the greatest VI distance from $\mathbf{c} = (\{1,2\},\{3,4\})$, and the VI credible ball would have to be extended to 100% credibility to contain these partitions. Another interesting direction is discussed by **Paulon, Trippa, and Müller**, where it may be relevant to understand uncertainty of a subset of the partition.

Computations. **Friel and Rastelli** in Rastelli and Friel (2017) expand upon our work and develop an alternative technique for optimizing the posterior expected loss, which is linear in N but increasingly expensive in the number of MCMC samples. Also, their approach does not require the approximation of the posterior expected loss through Jensen's inequality. As they highlight, more work is needed to understand the impact of this approximation, particularly, as their studies suggest that optimizing the lower bound to the posterior expected VI may actually perform better at recovering the true number of clusters. **Arbel, Corradin, and Lewandowski** point out the improvements in optimization of the greedy search algorithm over restricting to the MCMC samples (at a computational cost). Related to this, we note that the simple proposed approach to locate the credible bounds restricts to the MCMC samples. Similar improvements could be expected here by searching outside of the MCMC samples; this would however require a novel algorithm to locate the bounds as well as an appropriate refinement of the bounds (because, as highlighted by **Monni**, without restriction to the MCMC samples the bounds could contain a large number of partitions). One simple approach could be to restrict the search to partitions that are vertically aligned with the point estimate; this, however, also has its limitations.

Asymptotics. A study of the asymptotic properties of the proposed estimators and credible balls is an important and timely research direction, especially in light of the negative results of Miller and Harrison (2014) on posterior inconsistency for the number

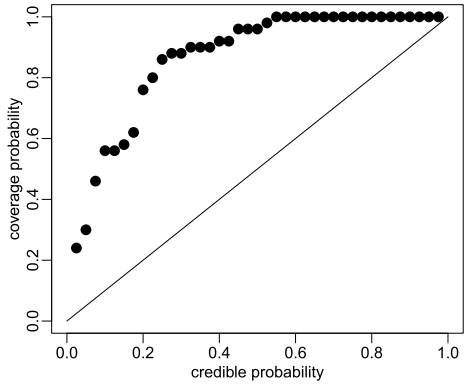


Figure 2: Credible probability against estimated coverage probability from 50 replicated experiments of Example 1 with $N = 200$; the solid represents the optimal setting when the $1 - \alpha$ credible ball achieves $1 - \alpha$ coverage.

of clusters and the positive results of the experiments in Table 2, as well as the recent results for the MAP estimator Rajkowski (2016), pointed out by **Arbel, Corradin, and Lewandowski**. **Belitser and Nurushev** and **Yoo** raise an interesting question on optimality of the proposed $(1 - \alpha) * 100\%$ credible balls; specifically, do they also have $(1-\alpha)*100\%$ frequentist coverage probability? We carried out a small experiment by simulating 50 datasets with the same data-generating mechanism as described in Example 1. We first note that $\widehat{k}_N = 4$ for the VI estimate in 90% of the experiments (although only one starting point was considered for the greedy search, and this may be improve with multiple restarts). Figure 2 depicts the estimated coverage probabilities as a function of the credible probability. This suggests that, in this setting, VI credible balls can be interpreted as confidence balls, although they are not optimal and have quite large coverage probabilities. The reasons for this could be a combination of the nonparametric model and the data-generating mechanism, as well as the credible ball definition, and it would be interesting to extend this simulation study for parametric models, large sample sizes, and other data-generating mechanisms. Overall, a deeper understanding of the frequentist coverage of the credible balls and other asymptotic properties is needed. **Belitser and Nurushev**: to the best of our knowledge, we do not know of results on minimax rates in the community detection problem for Binder's loss or VI, but note that Binder's loss can be viewed as the Hamming distance between the binary representation of clusterings. And, we would be intrigued to understand if the work of Zhang and Zhou (2016) could help to shed light on the coverage of the proposed credible balls for the community detection problem.

Applications. Our paper focused on Bayesian nonparametric mixture models and experiments considered Gaussian mixtures, but as highlighted by the discussants, the proposed tools have applications beyond this. **Friel and Rastelli** and **Belitser and Nurushev** discuss stochastic block models for networks. **Frühwirth–Schnatter, Grün, and Malsiner–Walli** and **Yoo** discuss sparse finite mixture models and **Frühwirth–**

Schnatter, Grün, and Malsiner–Walli extend Example 1 by using the proposed summary tools for sparse finite mixtures. **Mohammadi** considers finite mixture models and trans-dimensional MCMC, which allows exploration of the space of partitions. The proposed summary tools are relevant in this case as well, and instead of "replacing" the trans-dimensional MCMC, the greedy search algorithm would be used to find a point estimate of the partition based on those explored in the trans-dimensional MCMC. In addition to considering marginal properties, such as the posterior on the number of clusters, we have developed tools to further understand the posterior on the clustering structure. **Castelletti and Peluso** describe an interesting application and extension to DAGs and EGs.

Loss functions. We have proposed and motivated the use of VI as a general loss functions for partitions, and developed tools for summarizing MCMC samples of partitions. However, in some applications one may be interested in more problem-specific loss functions. An interesting example for clinical trials is provided by **Paulon, Trippa, and Müller**, where the loss function is a combination of the squared error loss for the true and estimated parameters of each patient and a penalization term that encourages cluster parameters to be distinct. Another interesting example is provided by **Natvig and Tvete**, where a problem-specific loss function is designed for earthquake data.

Model selection. **Casarin and Tonellato** propose an interesting use of the posterior expected VI as a model selection tool to identify hyperparameters. Although this requires fitting several models, it can result in a lower posterior expected VI, compared with the hierarchical model with hyperpriors on the hyperparameters. **Casarin and Tonellato** have outlined a promising research direction, and we would, for example, be interested in the use of the posterior expected VI as a model selection tool to compare nonparametric priors on partitions, beyond the the Dirichlet process (Lijoi and Prünster (2011), Barrios et al. (2013)) or the sparse finite mixture models investigated by **Frühwirth–Schnatter, Grün, and Malsiner–Walli**.

References

Barrios, E., Lijoi, A., Nieto-Barajas, L., and Prünster, I. (2013). "Modeling with normalized random measure mixture models." *Statistical Science*, 313–334. MR3135535. doi: https://doi.org/10.1214/13-STS416. 625

Lijoi, A. and Prünster, I. (2011). "Models beyond the Dirichlet process." In Hjort, N., Holmes, C., Müller, P., and Walker, S. (eds.), *Bayesian Nonparametrics*, 80–136. Cambridge, UK: Cambridge University Press. MR2730661. 625

Meilă, M. (2007). "Comparing clusterings – an information based distance." *Journal of Multivariate Analysis*, 98: 873–895. MR2325412. doi: https://doi.org/10.1016/j.jmva.2006.11.013. 622

Miller, J. and Harrison, M. (2014). "Inconsistency of Pitman-Yor process mixtures for the number of components." *Journal of Machine Learning Research*, 15: 3333–3370. MR3277163. 623

Rajkowski, L. (2016). "Analysis of MAP in CRP normal-normal model." ArXiv preprint arXiv:1606.03275. 624

Rastelli, R. and Friel, N. (2017). "Modeling with normalized random measure mixture models." *Statistics and Computing*. doi: https://doi.org/10.1007/s11222-017-9786-y. 623

Zhang, A. and Zhou, H. (2016). "Minimax rates of community detection in stochastic block models." *Annals of Statistics*, 44: 2252–2280. MR3546450. doi: https://doi.org/10.1214/15-AOS1428. 624