Bayesian cluster analysis: Point estimation and credible balls

Sara Wade and Zoubin Ghahramani

May 14, 2015

Abstract

Clustering is widely studied in statistics and machine learning, with applications in a variety of fields. As opposed to classical algorithms 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 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, it is difficult to assess statistical properties, such as uncertainty on the number of clusters.

Bayesian nonparametric clustering or random partition models (Quintana [2006]) are becoming increasingly popular, as they overcome many of the drawbacks of classical algorithms. A Bayesian nonparametric treatment of the clustering problem involves assigning a prior over the space of all possible partitions of the data and computing the posterior of the partition given the data. Thus, instead of returning a single clustering solution, Bayesian nonparametric models provide a posterior over the entire space of clusterings, expressing our belief and uncertainty in the clustering structure given the data.

However, an important problem in Bayesian 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 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 infeasible, 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 the 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 likelihood of the data given the partition 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 likelihood or prior is often unavailable, for example due to the presence of hyperparameters that cannot be marginalized. In general, 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]). 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 de-

fines 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 $L_2(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$; the posterior median under the absolute error loss $L_1(\theta, \widehat{\theta}) = |\theta - \widehat{\theta}|$; and the posterior mode under the 0-1 loss $L_{0-1}(\theta, \widehat{\theta}) = \mathbf{1}(\theta \neq \widehat{\theta})$.

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 credible balls 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 modelling tools in Bayesian nonparametrics. The data are assumed conditionally i.i.d. with density

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

where $K(y|\theta)$ is a parametric density on the sample space with parameter θ. A nonparametric prior is placed on the mixing measure P, and typically this prior has discrete realizations almost surely (a.s.). In this case,

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

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 modelled 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. An advantage of the Bayesian nonparametric approach is that the number of clusters k_N is determined by and can grow with the data. Marginalizing over the random probability measure, the data $y_{1:N}$ is modelled as

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

where $\mathbf{y}_j = \{y_n\}_{n \in C_j}$.

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

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

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 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 two-parameter Poisson-Dirichlet process (Pitman and Yor [1997]) or the normalized generalized Gamma process or more generally, the class of Poisson-Kingman models (Pitman [2003]), the class of normalized completely random measures, or the class of stick-breaking priors (Ishwaran and James [2001]). See Lijoi and Prünster [2011] for an overview.

In general, the likelihood or the prior may not be available in closed form. Moreover, there are

$$S_{N,k} = \frac{1}{k!}\sum_{j=0}^{k}(-1)^j \left(\begin{array}{c} k \\ j \end{array} \right) (k-j)^N,$$

a Stirling number of the second kind, ways to partition the N data points in to 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] or the conditional samplers described in Ishwaran and James [2001], Kalli et al. [2011], or Papaspiliopoulos and Roberts [2008]. 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 clustering(s) 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:

$$\mathbf{c}^* = \underset{\mathbf{c}}{\operatorname{argmax}}\, p(\mathbf{c}|y_{1:N}).$$

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 point 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 $n_{i\,j}$, which count the number of data points in cluster i under $\mathbf{c}$ and 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 $n_{i\,j}$. 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

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

where $n_{i+} = \sum_j n_{i\,j}$ and $n_{+\,j} = \sum_i n_{i\,j}$. 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 a 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

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

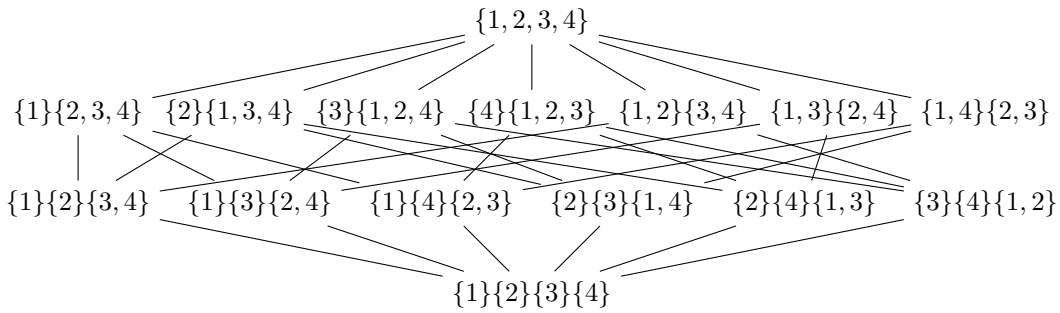


Figure 1: Hasse diagram for the lattice of partitions with a sample of size $N = 4$. A line is drawn from $\mathbf{c}$ up to $\widehat{\mathbf{c}}$ when $\mathbf{c}$ is covered by $\widehat{\mathbf{c}}$.

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 3.1 *Both VI and $\tilde{B}$ are metrics on the space of partitions, that is they satisfy:*

1. *Non-negativity: $d(\mathbf{c}, \widehat{\mathbf{c}}) \geq 0$ and $d(\mathbf{c}, \widehat{\mathbf{c}}) = 0$ if and only if $\mathbf{c} = \widehat{\mathbf{c}}$ under a permutation of cluster labels,*
2. *Symmetry: $d(\mathbf{c}, \widehat{\mathbf{c}}) = d(\widehat{\mathbf{c}}, \mathbf{c})$,*
3. *Triangle inequality: for any $\mathbf{c}, \widehat{\mathbf{c}}, \widehat{\widehat{\mathbf{c}}}$,*

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

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.

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, meaning that the following properties are satisfied:

1. Reflexivity: $\mathbf{c} \leq \mathbf{c}$,
2. Antisymmetry: if $\mathbf{c} \leq \widehat{\mathbf{c}}$ and $\widehat{\mathbf{c}} \leq \mathbf{c}$ then $\mathbf{c} = \widehat{\mathbf{c}}$ under a permutation of cluster labels,

3. Transitivity: if $\mathbf{c} \leq \widehat{\mathbf{c}}$ and $\widehat{\mathbf{c}} \leq \widehat{\widehat{\mathbf{c}}}$, then $\mathbf{c} \leq \widehat{\widehat{\mathbf{c}}}$.

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* (g.l.b.) and *least upper bound* (l.u.b.), where an element $\mathbf{c} \in \mathbf{C}$ is an upper bound for a subset $\mathbf{S} \subseteq \mathbf{C}$ if $\mathbf{s} \leq \mathbf{c}$ for all $\mathbf{s} \in \mathbf{S}$, and $\mathbf{c} \in \mathbf{C}$ is the least upper bound, if it exists, for a subset $\mathbf{S} \subseteq \mathbf{C}$ 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 are similarly defined). In general, a partially ordered set satisfying these properties forms a lattice; that is, for any $\mathbf{c}, \widehat{\mathbf{c}}, \widehat{\widehat{\mathbf{c}}} \in \mathbf{C}$

1. $\mathbf{c} \wedge \mathbf{c} = \mathbf{c}$ and $\mathbf{c} \vee \mathbf{c} = \mathbf{c}$,
2. $\mathbf{c} \wedge \widehat{\mathbf{c}} = \widehat{\mathbf{c}} \wedge \mathbf{c}$ and $\mathbf{c} \vee \widehat{\mathbf{c}} = \widehat{\mathbf{c}} \vee \mathbf{c}$,
3. $\mathbf{c} \wedge (\widehat{\mathbf{c}} \wedge \widehat{\widehat{\mathbf{c}}}) = (\mathbf{c} \wedge \widehat{\mathbf{c}}) \wedge \widehat{\widehat{\mathbf{c}}}$ and $\mathbf{c} \vee (\widehat{\mathbf{c}} \vee \widehat{\widehat{\mathbf{c}}}) = (\mathbf{c} \vee \widehat{\mathbf{c}}) \vee \widehat{\widehat{\mathbf{c}}}$,
4. $\mathbf{c} \wedge (\mathbf{c} \vee \widehat{\mathbf{c}}) = \mathbf{c}$ and $\mathbf{c} \vee (\mathbf{c} \wedge \widehat{\mathbf{c}}) = \mathbf{c}$,

where the operators $\wedge$ and $\vee$ are defined as $\mathbf{c} \wedge \widehat{\mathbf{c}} = \text{g.l.b.}(\mathbf{c}, \widehat{\mathbf{c}})$ and $\mathbf{c} \vee \widehat{\mathbf{c}} = \text{l.u.b.}(\mathbf{c}, \widehat{\mathbf{c}})$ and equality holds under a permutation of cluster labels. In this case, the operator $\wedge$ is called the meet and the operator $\vee$ is called the join. 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 their own cluster $\mathbf{c} = (\{1\}, \ldots, \{N\})$. See Nation [1991] for more details on lattice theory.

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 3.2). 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 3.3).

Property 3.2 *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 3.3 *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 Appendix. 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

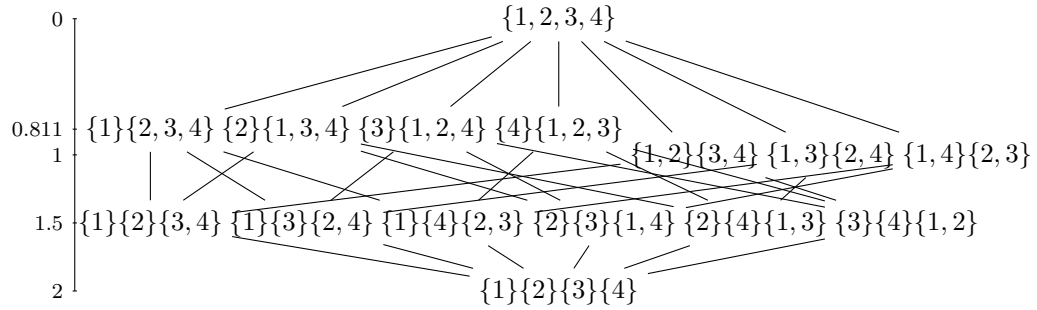


Figure 2: Hasse diagram stretched by VI with a sample of size $N = 4$. Note $2 - \frac{3}{4}\log(3) \approx 0.811$. From the VI streched Hasse diagram, we can determine the distance between any two partitions. Example: if $\mathbf{c} = (\{1,2\},\{3,4\})$ and $\widehat{\mathbf{c}} = (\{1\},\{3\},\{2,4\})$,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

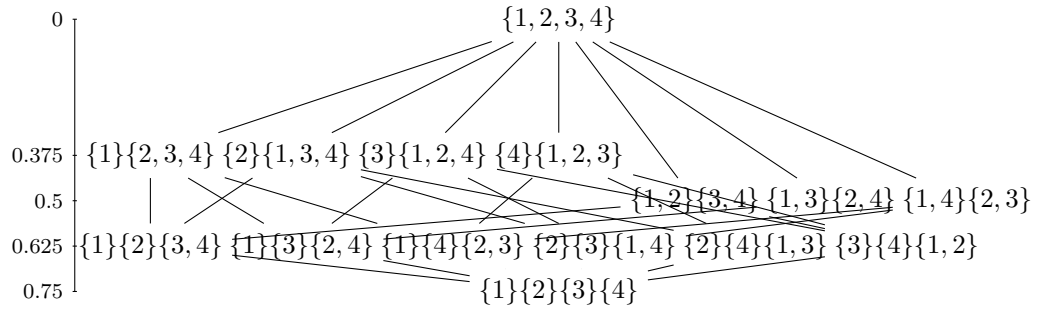


Figure 3: Hasse diagram stretched by $\tilde{\mathrm{B}}$ with a sample of size $N = 4$. From the $\tilde{\mathrm{B}}$ streched Hasse diagram, we can determine the distance between any two partitions. Example: if $\mathbf{c} = (\{1,2\},\{3,4\})$ and $\widehat{\mathbf{c}} = (\{1\},\{3\},\{2,4\})$, then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

$N = 4$ in Figure 1 stretched according VI and $\tilde{\mathrm{B}}$ respectively. As an example, consider $\mathbf{c} = (\{1,2\},\{3,4\})$ and $\widehat{\mathbf{c}} = (\{1\},\{3\},\{2,4\})$; their meet is $\mathbf{c} \wedge \widehat{\mathbf{c}} = (\{1\},\{2\},\{3\},\{4\})$, and the VI distance is $\mathrm{VI}(\mathbf{c},\widehat{\mathbf{c}}) = \mathrm{VI}(\mathbf{c}\wedge\widehat{\mathbf{c}},\mathbf{1}) - \mathrm{VI}(\mathbf{c},\mathbf{1}) + \mathrm{VI}(\mathbf{c}\wedge\widehat{\mathbf{c}},1) - \mathrm{VI}(\widehat{\mathbf{c}},1) = 2 - 1 + 2 - 1.5 = 1.5$.

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 3.4 *A distance on partitions satisfying Properties 3.2 and 3.3 has the property that for any two partitions* $\mathbf{c}$ *and* $\widehat{\mathbf{c}}$*,*

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

Thus,

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

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 are $N \to \infty$, as there are 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 2 $(\{1\}, \{2\}, \{3\}, \{4\})$.

Property 3.5 *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.*

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

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

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

and

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

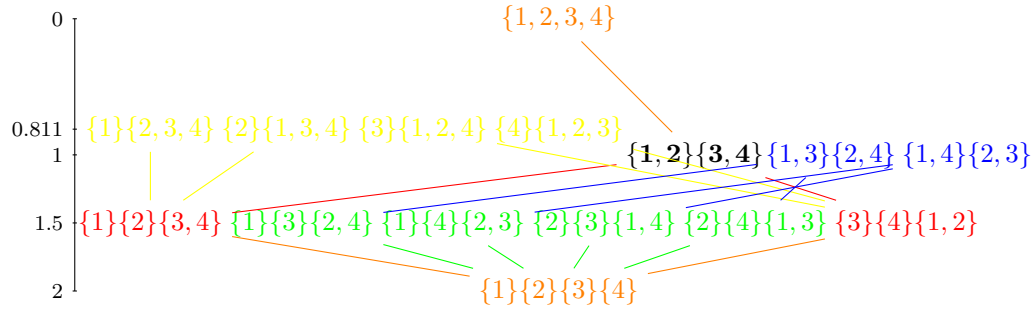


Figure 4: Example of the VI ball around $\mathbf{c} = (\{1, 2\}, \{3, 4\})$, with the rainbow color indicating increasing distance from $\mathbf{c}$. The smallest non-trivial credible ball contains all the red clusterings, the next smallest contains the red and orange clusterings, and so on.

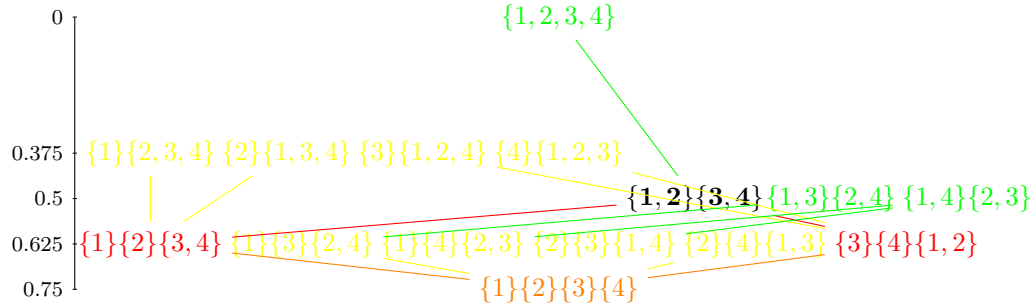


Figure 5: Example of the $\tilde{\mathrm{B}}$ ball around $\mathbf{c} = (\{1, 2\}, \{3, 4\})$, with the rainbow color indicating increasing distance from $\mathbf{c}$. The smallest non-trivial credible ball contains all the red clusterings, the next smallest contains the red and orange clusterings, and so on.

Property 3.6 reflects the asymmetry apparent in Figure 3. In particular, for $\tilde{\mathrm{B}}$, a partition with two clusters of equal size $\mathbf{c}_2$ will always be closer to the extreme $\mathbf{0}$ of everyone in their 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 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 1 and 3 are equally distant between the two extremes under $\tilde{\mathrm{B}}$. The following property generalizes this observation.

Property 3.7 *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 distance 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 $\tilde{\mathrm{B}}$ are metrics on the space of clusterings, we can construct a ball around $\mathbf{c}$ of size ϵ, defined as:

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

Interestingly, the balls differ between the two metrics even for the simple example with $N = 4$ in Figures 4 and 5, which consider a ball around $\mathbf{c} = (\{1, 2\}, \{3, 4\})$. In these figures, partitions are rainbow colored by increasing distance to $\mathbf{c}$; thus, the smallest non-trivial ball, i.e. the smallest ball around $\mathbf{c}$ with at least two partitions, contains all red clusterings, the next smallest ball contains all red and orange clusterings, and so on. In general, from Property 3.5, the smallest non-trivial ball will be the same for the two metrics, which is confirmed in the figures, as the set of red clusterings coincide. When considering the next smallest ball, differences emerge; the VI ball includes the red clusterings and the orange clusterings $\mathbf{0}$ and $\mathbf{1}$, and the $\tilde{\mathrm{B}}$ ball includes the red clusterings and the orange clustering $\mathbf{0}$. Note that $\mathbf{1}$ is only included in the $\tilde{\mathrm{B}}$ ball around $\mathbf{c}$ when it is expanded to include all clusterings and is considered as distant to $\mathbf{c}$ as the partitions $\mathbf{c} = (\{1, 3\}, \{2, 4\})$ and $\mathbf{c} = (\{1, 4\}, \{2, 3\})$. 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 $\tilde{\mathrm{B}}$ 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 $\tilde{\mathrm{B}}$ 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 loss function in Bayesian cluster analysis. Under the VI, the optimal partition $\mathbf{c}^*$ is

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

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. This, however, 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 efficient to evaluate:

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

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 pre-computed based on the MCMC output.

Due to the huge dimensions of the partition space, computing (3) or (4) for every possible $\widehat{\mathbf{c}}$ is practically impossible. An simple technique to find the optimal partition $\mathbf{c}^*$ restricts the search space to some smaller space of partitions, for example, the partitions visited in the MCMC chain. Alternative algorithms have been developed in Quintana and Iglesias [2003] and Lau and Green [2007].

We propose a greedy search algorithm to locate the optimal partition $\mathbf{c}^*$ based on the Hasse diagram, which can be used to 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 is computed for all proposed partitions and we move in the direction of minimum posterior expected loss. The algorithm stops when no reduction in the posterior expected loss is obtained or when a maximum number of iterations has been reached. In practice, we may initialize the algorithm with say a particular sample of the MCMC, e.g. the last sample, or with the MCMC sample which minimizes the posterior expected loss. The algorithm is summarized below.

Greedy search algorithm based on the Hasse diagram:

- Initialize $\widehat{\mathbf{c}}$.
- For $i = 1, \dots I$
  - Find the l closest partitions that cover $\widehat{\mathbf{c}}$ and the l closest partitions that $\widehat{\mathbf{c}}$ covers.
  - Compute $\mathbb{E}[L(\mathbf{c}, \widehat{\widehat{\mathbf{c}}})|\mathcal{D}]$ for all $2l$ partitions and select the partition $\mathbf{c}'$ with minimal $\mathbb{E}[L(\mathbf{c}, \mathbf{c}')|\mathcal{D}]$.
  - If $\mathbb{E}[L(\mathbf{c}, \mathbf{c}')|\mathcal{D}] < \mathbb{E}[L(\mathbf{c}, \widehat{\mathbf{c}})|\mathcal{D}]$, set $\widehat{\mathbf{c}} = \mathbf{c}'$. Otherwise, STOP.
- end

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 3.5). Typically, however, they will be different, see the discussion at the end of Section 3.

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

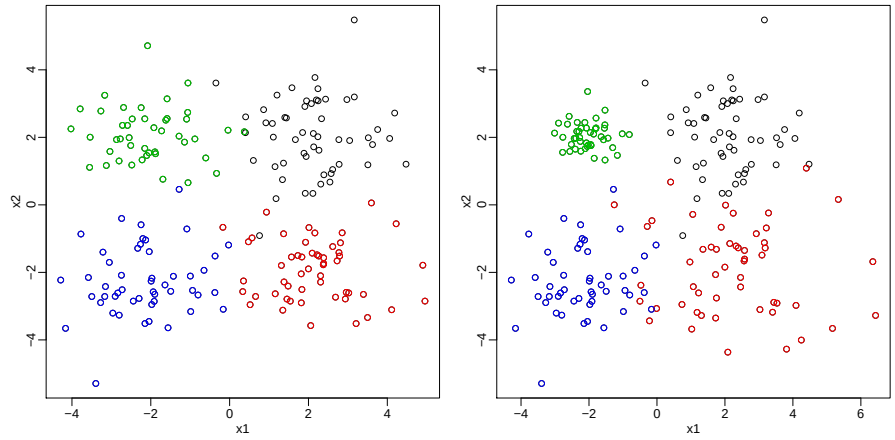
$$\frac{1}{M}\sum_{m=1}^{M} \mathbf{1}(d(\mathbf{c}^*, \mathbf{c}^m) \leq \epsilon) \geq 1-\alpha.$$

To characterize the credible ball, we define the *vertical* and *horizontal bounds* based on the credible ball. The vertical upper bounds consist of the partitions in the credible ball with the smallest number of clusters which are most distant from $\mathbf{c}^*$. The vertical lower bounds consist of the partitions in the credible ball with the largest number of clusters which are most distant from $\mathbf{c}^*$. The horizontal bounds consist of the partitions in the credible ball which are most distant from $\mathbf{c}^*$.

For example, suppose $N = 4$ and the optimal point estimate under both VI and $\tilde{\text{B}}$ is $(\{1,2\},\{3,4\})$. Further suppose that the 95% credible ball with the VI and $\tilde{\text{B}}$ metric consists of the red, orange, and yellow partitions depicted in Figures 4 and 5, respectively. For VI, the upper vertical bound is $\mathbf{1}$, the lower vertical bound is $\mathbf{0}$, and the horizontal bounds are the yellow partitions, i.e. those with one singleton and one cluster of size $N-1=3$. For $\tilde{\text{B}}$, the upper vertical bounds are the partitions with one singleton and one cluster of size $N-1=3$, the lower vertical bound is $\mathbf{0}$, and the horizontal bounds are the yellow partitions, i.e. those with one singleton and one cluster of size $N-1=3$ and those with two singletons and one cluster of size $N-2=2$ which are not covered by $\mathbf{c}^*$.

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 the 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 $\mathbf{c}^*$.



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

Figure 6: The data are simulated from a mixture of four normals with locations $(\pm 2, \pm 2)'$ and colored by cluster membership. In (a) the standard deviations of all components are equal to 1, and in (b) the standard deviations are 1 in the first and third quadrants, 0.5 in second quadrant, and 1.5 in the fourth quandrant.

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.

6.1 Simulated examples

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

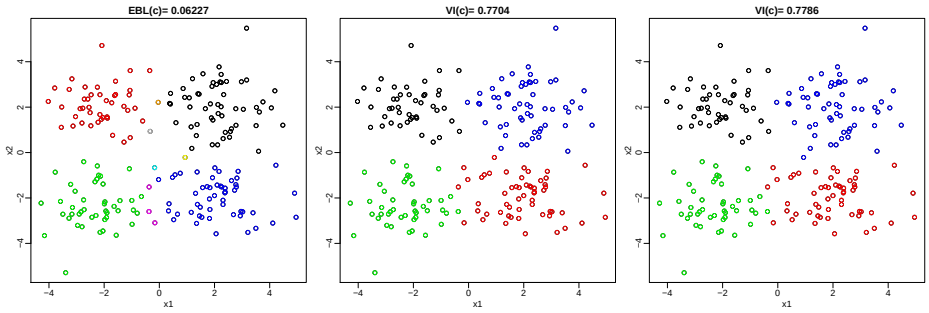
$$X_i \overset{iid}{\sim} \sum_{j=1}^{4} \frac{1}{4} \mathrm{N}\left(\begin{bmatrix} (-1)^{\lfloor \frac{(j-1)}{2} \rfloor} 2 \\ (-1)^{j-1} 2 \end{bmatrix}, \begin{bmatrix} \sigma_j^2 & 0 \\ 0 & \sigma_j^2 \end{bmatrix} \right).$$

For the first example, $\sigma_j = 1$ for all components, and for the second, $\sigma_j = 1$ 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 data for both examples are depicted in Figure 6 and colored by cluster membership.

We consider a Dirichlet process (DP) mixture model:

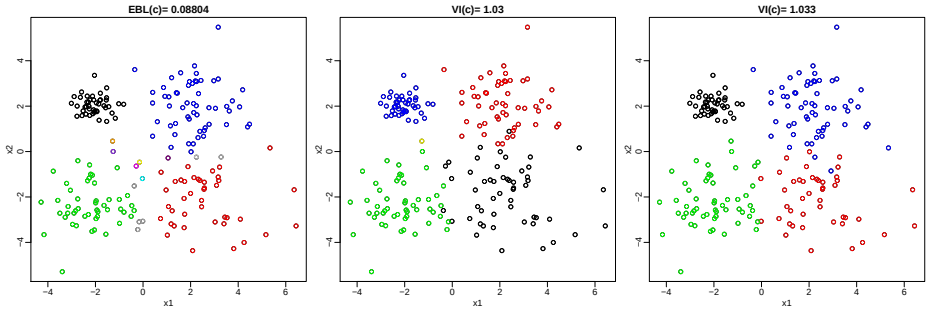
$$\begin{aligned} X_i | P &\overset{iid}{\sim} \int \mathrm{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right) dP(\mu, \Sigma), \\ P &\sim \mathrm{DP}(\alpha P_0), \end{aligned} \tag{5}$$

where $\mu = (\mu_1, \mu_2)'$ and Σ is a diagonal matrix with diagonal elements (σ_1^2, σ_2^2).



(a) Binder's loss: 9 clusters (b) VI: 4 clusters (c) Modified VI: 4 clusters

Figure 7: Example 1: optimal clustering estimate with color representing cluster membership for Binder's loss, VI, and the modified VI.



(a) Binder's loss: 12 clusters (b) VI: 5 clusters (c) Modified VI: 4 clusters

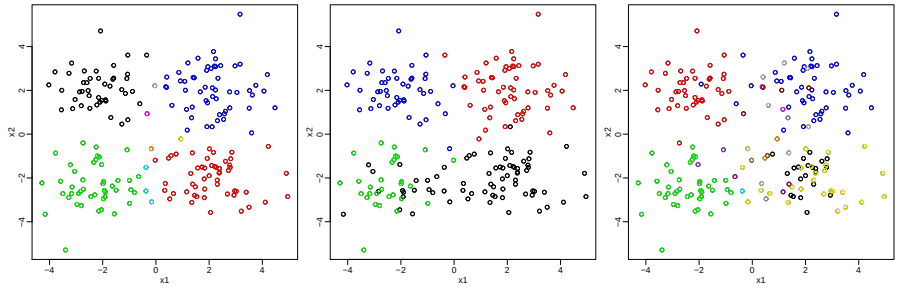
Figure 8: Example 2: optimal clustering estimate with color representing cluster membership for Binder's loss, VI, and the modified VI.

	$\mathbb{E}[\tilde{\mathrm{B}}\|\mathcal{D}]$	$\tilde{\mathrm{B}}(\mathbf{c}_t, \mathbf{c}^*)$	$\mathbb{E}[\mathrm{VI}\|\mathcal{D}]$	$\mathrm{VI}(\mathbf{c}_t, \mathbf{c}^*)$
$\mathbf{c}^*$ from $\tilde{\mathrm{B}}$	**0.062**	0.045	0.816	0.643
$\mathbf{c}^*$ from VI	0.064	**0.044**	**0.77**	**0.569**
$\mathbf{c}^*$ from mVI	0.064	0.049	0.779	0.620

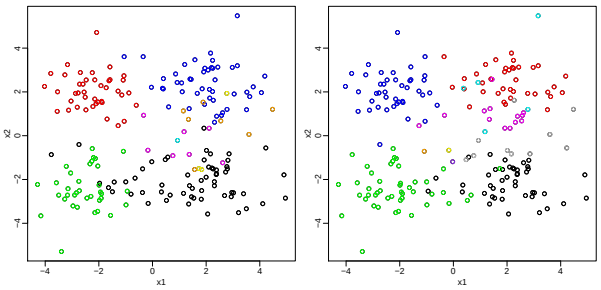
Table 1: Example 1: a comparison of the optimal partition with $\tilde{\mathrm{B}}$, VI, or the modified VI in terms of expected $\tilde{\mathrm{B}}$, $\tilde{\mathrm{B}}$ between the optimal and true clusterings, expected VI, and VI between the optimal and true clusterings.

	$\mathbb{E}[\tilde{\mathrm{B}}\|\mathcal{D}]$	$\tilde{\mathrm{B}}(\mathbf{c}_t, \mathbf{c}^*)$	$\mathbb{E}[\mathrm{VI}\|\mathcal{D}]$	$\mathrm{VI}(\mathbf{c}_t, \mathbf{c}^*)$
$\mathbf{c}^*$ from $\tilde{\mathrm{B}}$	**0.088**	**0.056**	1.068	0.764
$\mathbf{c}^*$ from VI	0.099	**0.056**	**1.03**	**0.646**
$\mathbf{c}^*$ from mVI	0.099	0.062	1.033	0.648

Table 2: Example 2: a comparison of the optimal partition with $\tilde{\mathrm{B}}$, VI, or the modified VI in terms of expected $\tilde{\mathrm{B}}$, $\tilde{\mathrm{B}}$ between the optimal and true clusterings, expected VI, and VI between the optimal and true clusterings.



(a) $\tilde{\mathrm{B}}$ estimate: 9 clusters (b) $\tilde{\mathrm{B}}$ upper vertical bound: 4 clusters (c) $\tilde{\mathrm{B}}$ lower vertical bound: 18 clusters



(d) $\tilde{\mathrm{B}}$ horizontal bounds: 8 and 11 clusters

Figure 9: Example 1: 95% credible ball with Binder's loss around c^* (a) represented by (b) the upper vertical bound, (c) the lower vertical bound, and (d) the horizontal bounds, where color denotes cluster membership.

The base measure of the DP is 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

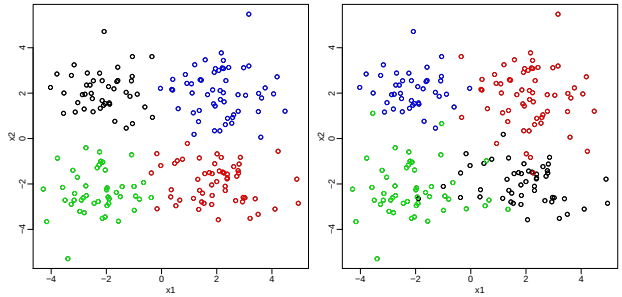
$$p_0(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \propto \prod_{i=1}^{2} \sqrt{\frac{c_i}{\sigma_i^2}} \exp\left(-\frac{c_i}{2\sigma_i^2}(\mu_i - \mu_{0,i})^2\right) (\sigma_i^2)^{-a_i-1} \exp\left(-\frac{b_i}{\sigma_i^2}\right).$$

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 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 shown) suggest convergence.

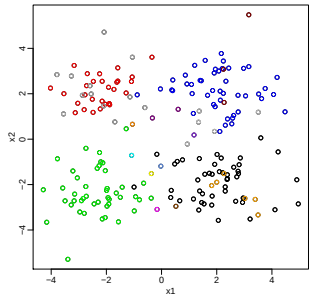
Among partitions sampled in the MCMC, only one is visited twice and all others are visited once in the first example, while no partitions are visited more than once in the second. Thus, an estimate of posterior mode based on frequency counts is not reliable.

For the first example, Figure 7 depicts the optimal partition found by the greedy search algorithm for Binder's loss (9 clusters), VI (4 clusters), and the modified VI (4 clusters) where the lower bound to the expected VI in (4) is minimized. The four true clusters are visible in all solutions; however, Binder's loss creates new small clusters for observations located on the border between clusters whose cluster membership is uncertain, overestimating the number of

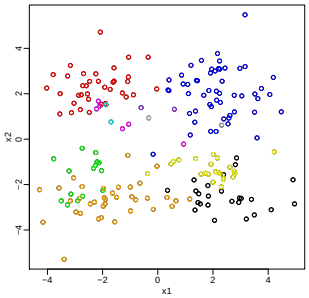


(a) VI estimate: 4 clusters

(b) VI upper vertical bound: 4 clusters



(c) VI lower vertical bound: 16 clusters



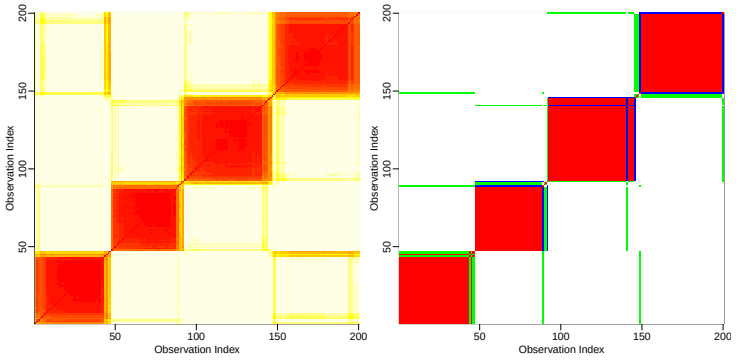
(d) VI horizontal bound: 11 clusters

Figure 10: Example 1: 95% credible ball with VI around c^* (a) represented by (b) the upper vertical bound, (c) the lower vertical bound, and (d) the horizontal bound, where color denotes cluster membership.

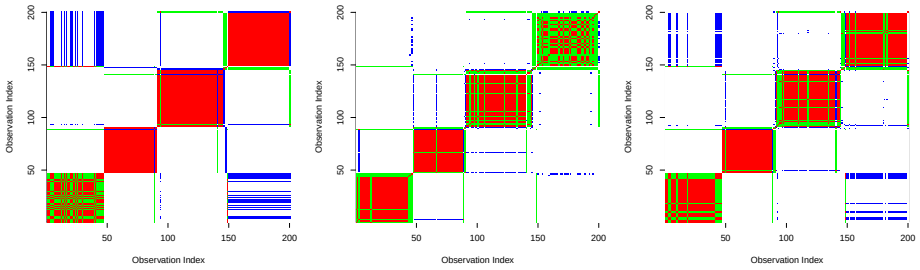
clusters. As expected, the $\tilde{\text{B}}$ estimate and VI estimate achieve the lowest posterior expected loss for $\tilde{\text{B}}$ and VI, respectively, but interestingly, the VI estimate has the smallest distance from the truth for both $\tilde{\text{B}}$ and VI (see Table 1).

For the second example, the optimal partition found by the greedy search algorithm for Binder's loss (12 clusters), VI (5 clusters), and the modified VI (4 clusters) are depicted in Figure 8 and compared in Table 2. Again, we observe that the four main clusters are present in all three point estimates, but Binder's loss allocates uncertain observations on the borders to their own clusters, overestimating the number of clusters present.

For the first example, Figures 9 and 10 represent the 95% credible ball around the optimal partition for $\tilde{\text{B}}$ and VI, respectively, through the upper vertical bound, the lower vertical bound, and the horizontal bounds, with data points colored according to cluster membership. Figures 11 and 12 provide an alternative visualization of the optimal partition and the bounds of the 95% credible ball. In these figures, the optimal partition and bounds are compared with the true partition through a color map of the $N \times N$ matrix with red indicating two data points in the same cluster for both the true and optimal partition; white indicating two data points in different clusters for both the true and optimal partition; green indicating two data points in the same cluster for



(a) Posterior similarity matrix (b) $\tilde{B}$ estimate

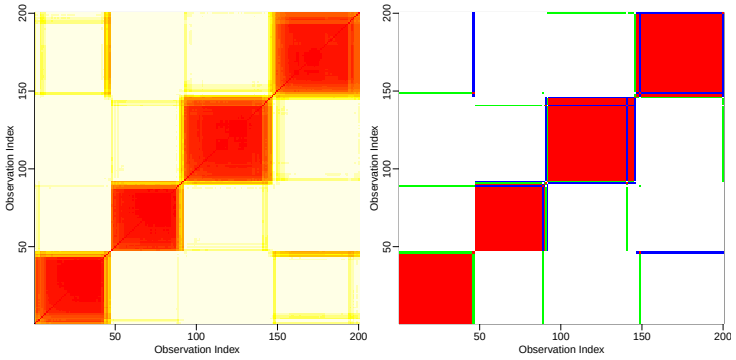


(c) $\tilde{B}$ upper vertical bound (d) $\tilde{B}$ lower vertical bound (e) $\tilde{B}$ horizontal bound

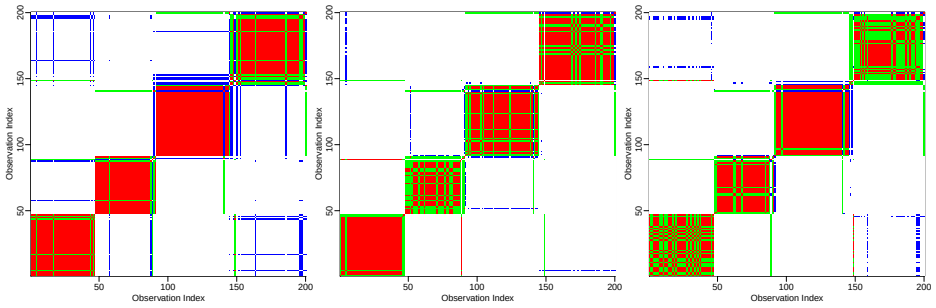
Figure 11: Example 1: (a) Heat map of the posterior similarity matrix. (b) Optimal $\tilde{B}$ partition compared with the true partition through a color map the $N \times N$ matrix with red indicating two data points in the same cluster for both the true and optimal partition; white indicating two data points in different clusters for both the truth and optimal; green indicating two data points in the same cluster for the truth and different clusters for the optimal; and blue indicating two data points in the same cluster for the optimal and different clusters for the truth. (c),(d),(e) Representation of the 95% credible ball with $\tilde{B}$ through a color map of the $N \times N$ matrix comparing the bound with the truth (only one of two horizontal bounds shown for conciseness).

the truth and different clusters for the optimal partition; and blue indicating two data points in the same cluster for the optimal partition and different clusters for the truth. Thus, red and white have a positive interpretation, while blue and green have a negative interpretation. Observations have been sorted by hierarchical clustering. Analogous plots for the second example are found in Figures 13, 14, 15 and 16.

We observe that for the first example 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 8 and 11 clusters for $\tilde{B}$ and 11 clusters for VI. For both metrics, these bounds reallocate uncertain data points on the border or in some cases merge or split one of the four main clusters. In the second example,



(a) Posterior similarity matrix (b) VI partition estimate

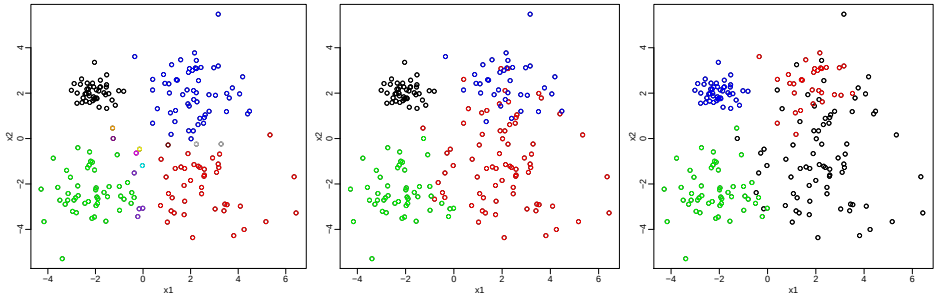


(c) VI upper vertical bound (d) VI lower vertical bound (e) VI horizontal bound

Figure 12: Example 1: (a) Heat map of the posterior similarity matrix. (b) Optimal VI partition compared with the true partition through a color map the $N \times N$ matrix with red indicating two data points in the same cluster for both the true and optimal partition; white indicating two data points in different clusters for both the truth and optimal; green indicating two data points in the same cluster for the truth and different clusters for the optimal; and blue indicating two data points in the same cluster for the optimal and different clusters for the truth. (c),(d),(e) Representation of the 95% credible ball with VI through a color map of the $N \times N$ matrix comparing the bound with the truth.

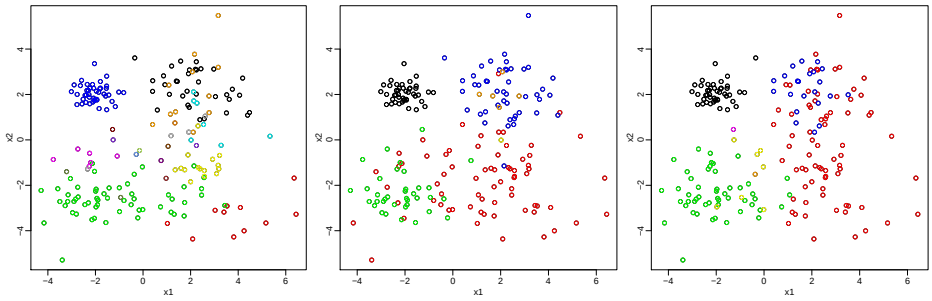
the green cluster in Figure 6b is stable in all bounds, while the 95% credible ball reflects posterior uncertainty on whether to divide the remaining observations into 3 to 18 clusters for $\tilde{\text{B}}$ and 2 to 15 clusters for VI. As a consequence of the asymmetric nature of $\tilde{\text{B}}$ discussed in Section 3, the number of clusters in the $\tilde{\text{B}}$ upper or lower vertical bound is greater than or equal to the number of clusters in the VI upper or lower, respectively, vertical bound in all examples.

Figures 11, 12, 15, and 16 provide a comparison of uncertainty represented by the posterior similarity matrix with the uncertainty represented by the 95% credible ball. In general, the posterior similarity matrix appears to under-represent the uncertainty; indeed, one would conclude from the similarity matrix that there is only uncertainty in allocation of a few data points in Example 1. Moreover, the 95% credible ball gives a precise quantification of the uncertainty.



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

(b) $\tilde{\text{B}}$ upper vertical bounds: 4 clusters



(c) $\tilde{\text{B}}$ lower vertical bound: 19 clusters

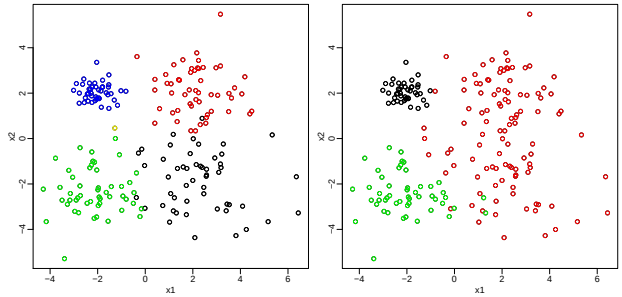
(d) $\tilde{\text{B}}$ horizontal bounds: 6 and 7 clusters

Figure 13: Example 2: 95% credible ball with Binder's loss represented by (a) the upper vertical bounds, (b) the lower vertical bound, and (c) the horizontal bounds (only two of three horizontal bounds shown for conciseness), where color denotes cluster membership.

Additionally, we note that for both metrics, there is greater uncertainty around the optimal estimate in Example 2; for Example 1, the 95% credible ball contains partitions with $\tilde{\text{B}}$-distance less than 0.097 or a VI-distance of less than 0.841, while for Example 2, the 95% credible ball contains partitions that with a $\tilde{\text{B}}$-distance of less than 0.188 and a VI-distance of less than 0.985.

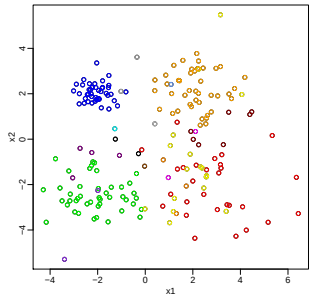
The greedy search algorithm was performed with different starting points and different values of l, which controls the amount of local exploration at each iteration. We experimented starting the search at the last MCMC sample or the MCMC sample which minimizes the criteria. The latter is clearly a better starting point, but requires additional computation for initialization. On the other hand, when initializing with the last sampled partition, more iterations are typically required to locate the optimal partition in the greedy search. In most cases, the algorithm converged to the same solution for both initializations, depending on the choice of l. The optimal partitions reported are found via the greedy search algorithm initialized at the MCMC sample which minimizes the criteria with $l = 200$.

An advantage of the greedy search algorithm is that partitions not explored in the MCMC algorithm can be considered; for example, in all simulated and real examples, the $\tilde{\text{B}}$ estimate is not among the sampled partitions and results

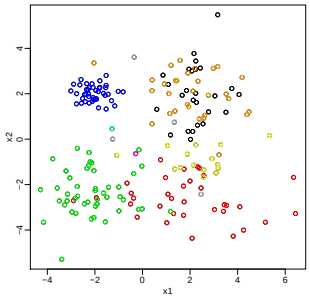


(a) VI estimate: 5 clusters

(b) VI upper vertical bound: 3 clusters



(c) VI lower vertical bound: 16 clusters



(d) VI horizontal bound: 10 clusters

Figure 14: Example 2: 95% credible ball with VI represented by (a) the upper vertical bound, (b) the lower vertical bound, and (c) the horizontal bound, where color denotes cluster membership.

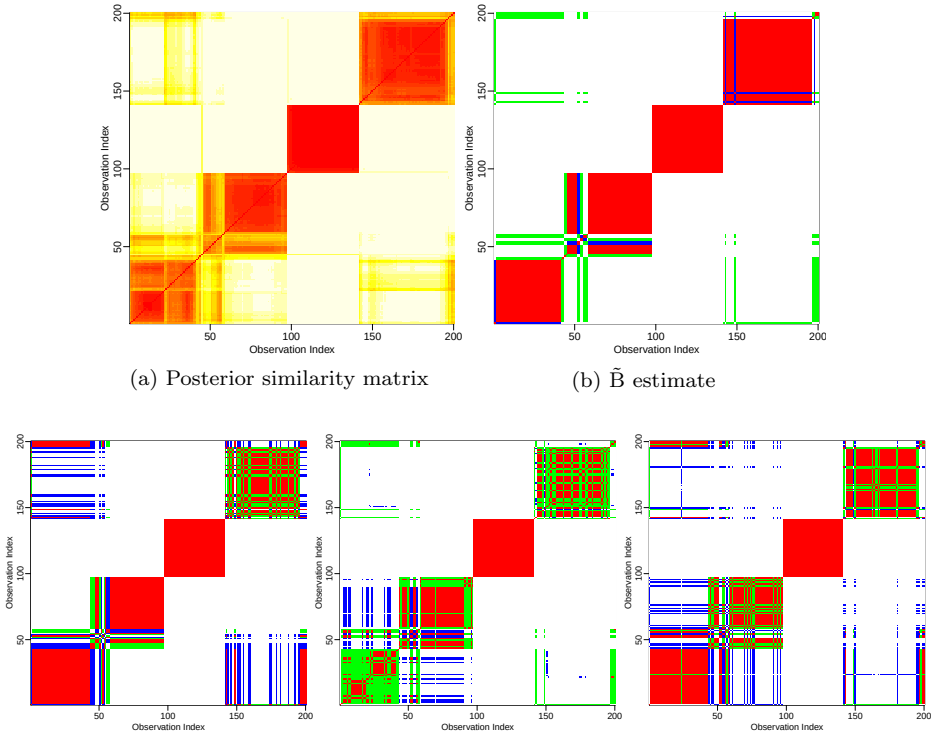
in a lower expected loss than any sampled partition. Finally, we note that the optimal partition for VI and the modified VI are quite similar, differing in the allocation of one data point for the first example, a handful of data points in the second example, and no data points in the real example. However, computation time under the modified VI is significantly reduced by at least a 600 fold decrease in all examples.

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 are modelled 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. In



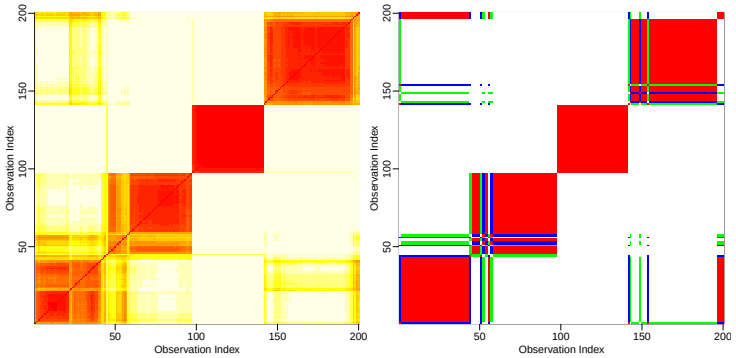
(a) Posterior similarity matrix (b) $\tilde{B}$ estimate

(c) $\tilde{B}$ upper vertical bound (d) $\tilde{B}$ lower vertical bound (e) $\tilde{B}$ horizontal bound

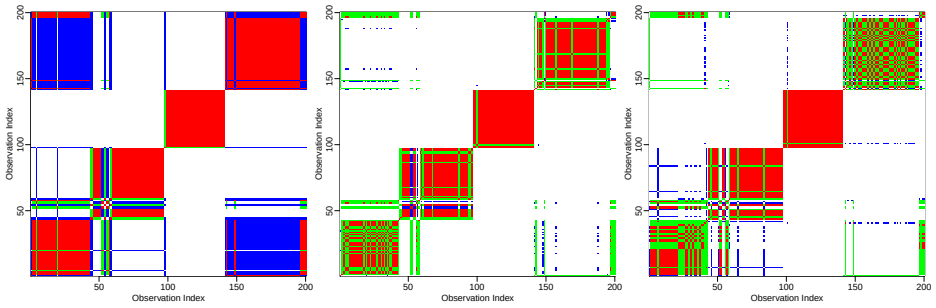
Figure 15: Example 2: (a) Heat map of the posterior similarity matrix. (b) Optimal $\tilde{B}$ partition compared with the true partition through a color map the $N \times N$ matrix with red indicating two data points in the same cluster for both the true and optimal partition; white indicating two data points in different clusters for both the truth and optimal; green indicating two data points in the same cluster for the truth and different clusters for the optimal; and blue indicating two data points in the same cluster for the optimal and different clusters for the truth. (c),(d),(e) Representation of the 95% credible ball with $\tilde{B}$ through a color map of the $N \times N$ matrix comparing the bound with the truth (only one of two upper bounds and one of three horizontal bounds shown for conciseness).

Figure 17, we plot the point estimate of the partition found by the greedy search algorithm for Binder's loss and VI. The data values are plotted against the estimated density values from the DP mixture model and colored according cluster membership. Again, we observe that Binder's loss prefers to place observations with uncertain allocation into singleton clusters, with the optimal partition containing 7 clusters, 4 of which are singletons, while the VI solution contains 3 clusters.

Table 3 compares the point estimates from the different criteria in terms of the posterior expected $\tilde{B}$ and VI; as anticipated, the $\tilde{B}$ solution has the smallest posterior expected $\tilde{B}$ and the VI solution has the smallest posterior expected VI. Interestingly, the VI estimate and the optimal partition found by minimizing



(a) Posterior similarity matrix (b) VI estimate

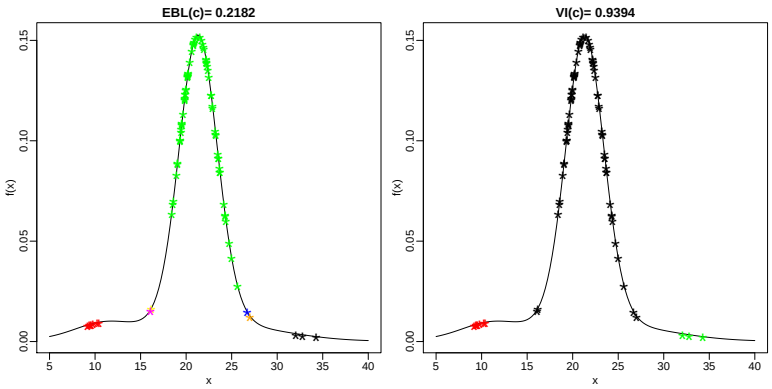


(c) VI upper vertical bound (d) VI lower vertical bound (e) VI horizontal bound

Figure 16: Example 2: (a) Heat map of the posterior similarity matrix. (b) Optimal VI partition compared with the true partition through a color map the $N \times N$ matrix with red indicating two data points in the same cluster for both the true and optimal partition; white indicating two data points in different clusters for both the truth and optimal; green indicating two data points in the same cluster for the truth and different clusters for the optimal; and blue indicating two data points in the same cluster for the optimal and different clusters for the truth. (c),(d),(e) Representation of the 95% credible ball with VI through a color map of the $N \times N$ matrix comparing the bound with the truth.

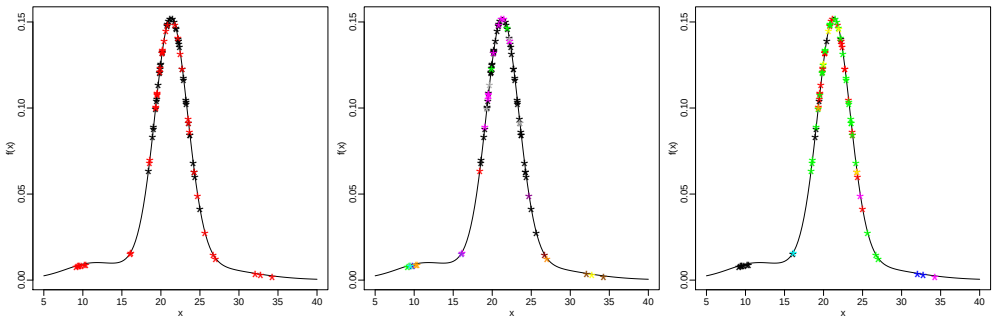
the lower bound to the posterior expected VI in (4) are equivalent, while the latter requires significantly less computation time (a 2600 fold decrease).

The 95% VI credible ball around the represenative VI partition contains all partitions with a VI distance less than 1.267. Figures 18 and 19 summarize the 95% credible ball through the upper vertical, lower vertical, and horizontal bounds. 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 modelled using only 2 components with a large variance for one component to account for outliers (red cluster in Figure (18a)). On the other extreme, the data could be further split in many, 15 to be precise, smaller clusters. Figure 19 emphasizes that the posterior similarity matrix under-represents the uncertainty around the point estimate. Following our comparison of $\tilde{\text{B}}$ and VI in Section



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

Figure 17: Galaxy example: optimal clustering estimate with color representing cluster membership for Binder's loss and VI. The optimal solution for VI and the modified VI criteria are equivalent.



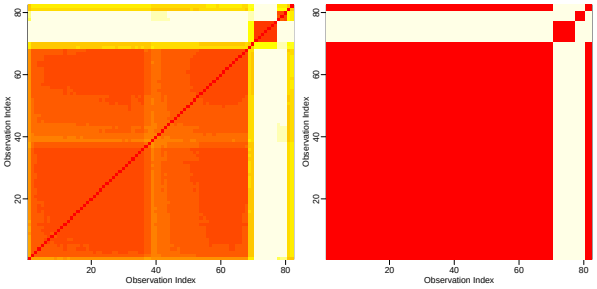
(a) VI upper vertical bound: 2 clusters (b) VI lower vertical bound: 15 clusters (c) VI horizontal bounds: 8 clusters

Figure 18: Galaxy example: 95% credible ball with VI represented by (a) the upper vertical bound, (b) the lower vertical bound, and (c) the horizontal bound, where color denotes cluster membership.

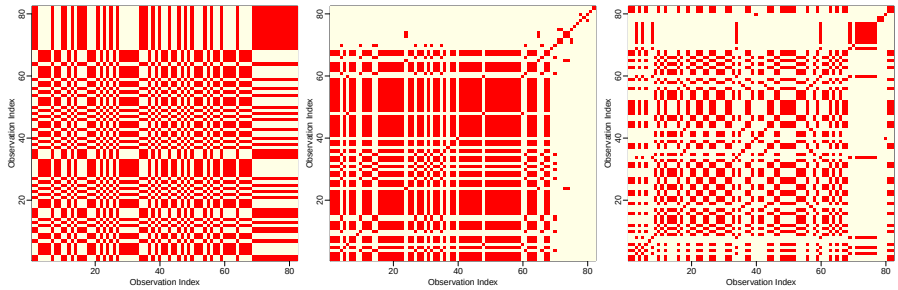
	$\mathbb{E}[\tilde{\mathrm{B}}\|\mathcal{D}]$	$\mathbb{E}[\mathrm{VI}\|\mathcal{D}]$
$\mathbf{c}^*$ from $\tilde{\mathrm{B}}$	**0.218**	1.014
$\mathbf{c}^*$ from VI	0.237	**0.939**

Table 3: Galaxy example: a comparison of the optimal partition with Binder's loss and VI in terms of posterior expected $\tilde{\mathrm{B}}$ and VI. The optimal solution for VI and the modified VI criteria are equivalent.

3 and the results of the simulated examples in Section 6.1, only the 95% VI credible ball is reported.



(a) Posterior similarity matrix (b) VI estimate



(c) VI upper vertical bound (d) VI lower vertical bound (e) VI horizontal bound

Figure 19: Galaxy example: (a) Heat map of the posterior similarity matrix. (b) Optimal VI partition (from the greedy search algorithm) depicted through a color map the binary $N \times N$ matrix with red indicating two data points in the same cluster for the optimal partition and white indicating two data points in different clusters. (c),(d),(e) Representation of the 95% credible ball with VI depicted through a color map of the binary $N \times N$ matrix.

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. 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 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 via 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, particularly focusing on their behavior on the lattice of partitions. We find that Binder's loss exhibits peculiar asymmetries, placing a smaller loss on the partition which splits two equally sized clusters into many singletons compared

with a larger loss on the partition which merges these two clusters. 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 observations to their own cluster. In addition, we have also proposed a novel greedy search algorithm based on the Hasse diagram 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. As opposed to the posterior similarity matrix, the 95% credible ball 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 will be shared through an **R** package 'mcclust.ext', expanding upon the existing **R** package 'mcclust' (Fritsch [2012]), which contains tools for point estimation in Bayesian cluster analysis and cluster comparison.

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 (see Griffiths and Ghahramani [2011] for an overview).

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

References

D.A. Binder. Bayesian Cluster Analysis. *Biometrika*, 65:31–38, 1978.

D.B. Dahl. Model-based clustering for expression data via a Dirichlet process mixture model. In K.A. Do, P. Müller, and M. Vannucci, editors, *Bayesian Inference for Gene Expression and Proteomic*, pages 201–218. Cambridge University Press, 2006.

D.B. Dahl. Modal clustering in a class of product partition models. *Bayesian Analysis*, 4:243–264, 2009.

J.A. Duan, M. Guindani, and A.E. Gelfand. Generalized spatial Dirichlet processes. *Biometrika*, 94:809–825, 2007.

D.B. Dunson. Nonparametric Bayes applications to biostatistics. In N.L. Hjort, C. Holmes, P. Müller, and S.G. Walker, editors, *Bayesian nonparametrics*. Cambridge University Press, 2010.

S. Favaro and Y.W. Teh. MCMC for normalized random measure mixture models. *Statistical Science*, 28:335–359, 2013.

T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973.

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

A. Fritsch and K. Ickstadt. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4:367–392, 2009.

J.E. Griffin and M. Steel. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 10:179–194, 2006.

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

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

N.A. Heard, C.C. Holmes, and D.A. Stephens. 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, 2006.

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

L. Hubert and P. Arabie. Comparing Partitions. *Journal of Classification*, 2: 193–218, 1985.

H. Ishwaran and L.F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173, 2001.

M. Kalli, J.E. Griffin, and S.G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21:93–105, 2011.

J.W. Lau and P.J. Green. Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16:526–558, 2007.

A. Lijoi and I. Prünster. Models beyond the Dirichlet process. In N.L. Hjort, C.C. Holmes, P. Müller, and S.G. Walker, editors, *Bayesian Nonparametrics*, pages 80–136, Cambridge, UK, 2011. Cambridge University Press.

A.Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics*, 12:351–357, 1984.

S.N. MacEachern. Dependent Dirichlet processes. *Technical Report, Department of Statistics, Ohio State University*, 2000.

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

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

M. Meilă. Comparing clusterings – an information based distance. *Journal of Multivariate Analysis*, 98:873–895, 2007.

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

P. Müller and F.A. Quintana. Nonparametric Bayesian data analysis. *Statistical Science*, 19:95–110, 2004.

J.B. Nation. *Notes on Lattice Theory.* 1991. http://www.math.hawaii.edu/ jb/books.html.

R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistcs*, 9:249–265, 2000.

O. Papaspiliopoulos and G.O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1): 169–186, 2008.

J. Pitman. Poisson Kingman partitions. In *Statistics and Science: a Festschrift for Terry Speed*, pages 1–34, Beachwood, 2003. IMS Lecture Notes.

J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900, 1997.

F.A. Quintana. A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference*, 136:2407–2429, 2006.

F.A. Quintana and P.L. Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B*, 65:557–574, 2003.

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

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

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

Y.W. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet process. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

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

Appendix I: Proofs

Proof of Property 3.2. First note that if $\mathbf{c} \geq \widehat{\mathbf{c}}$, or equivalently $\mathbf{c} \wedge \widehat{\mathbf{c}} = \widehat{\mathbf{c}}$, then for all nonzero $n_{i\,j}$, $n_{i\,j} = n_{+\,j}$. Thus, if $\mathbf{c} \geq \widehat{\mathbf{c}}$,

$$\begin{aligned} \mathrm{VI}(\mathbf{c}, \widehat{\mathbf{c}}) &= \sum_{i=1}^{k_N} \frac{n_{i+}}{N} \log\left(\frac{n_{i+}}{N}\right) + \sum_{j=1}^{\widehat{k}_N} \frac{n_{+\,j}}{N} \log\left(\frac{n_{+\,j}}{N}\right) - 2\sum_{j=1}^{\widehat{k}_N} \frac{n_{+\,j}}{N} \log\left(\frac{n_{+\,j}}{N}\right) \\ &= \sum_{i=1}^{k_N} \frac{n_{i+}}{N} \log\left(\frac{n_{i+}}{N}\right) - \sum_{j=1}^{\widehat{k}_N} \frac{n_{+\,j}}{N} \log\left(\frac{n_{+\,j}}{N}\right). \end{aligned} \tag{6}$$

Let n_i denote the number of observations in cluster i under $\mathbf{c}$; $\widehat{n}_i$ denote the number of observations in cluster i under $\widehat{\mathbf{c}}$; and $\widehat{\widehat{n}}_i$ denote the number of observations in cluster i under $\widehat{\widehat{\mathbf{c}}}$. Then, from (6), if $\mathbf{c} \geq \widehat{\mathbf{c}} \geq \widehat{\widehat{\mathbf{c}}}$,

$$\begin{aligned}
\mathrm{VI}(\mathbf{c}, \widehat{\widehat{\mathbf{c}}}) &= \sum_{i=1}^{k_N} \frac{n_i}{N} \log\left(\frac{n_i}{N}\right) - \sum_{i=1}^{\widehat{\widehat{k}}_N} \frac{\widehat{\widehat{n}}_i}{N} \log\left(\frac{\widehat{\widehat{n}}_i}{N}\right) \\
&= \sum_{i=1}^{k_N} \frac{n_i}{N} \log\left(\frac{n_i}{N}\right) - \sum_{i=1}^{\widehat{k}_N} \frac{\widehat{n}_i}{N} \log\left(\frac{\widehat{n}_i}{N}\right) + \sum_{i=1}^{\widehat{k}_N} \frac{\widehat{n}_i}{N} \log\left(\frac{\widehat{n}_i}{N}\right) - \sum_{i=1}^{\widehat{\widehat{k}}_N} \frac{\widehat{\widehat{n}}_i}{N} \log\left(\frac{\widehat{\widehat{n}}_i}{N}\right) \\
&= \mathrm{VI}(\mathbf{c}, \widehat{\mathbf{c}}) + \mathrm{VI}(\widehat{\mathbf{c}}, \widehat{\widehat{\mathbf{c}}})
\end{aligned}$$

For $\tilde{\mathrm{B}}$ the proof is similar, since if $\mathbf{c} \geq \widehat{\mathbf{c}}$,

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

∎

Proof of Property 3.3. The meet between two clusterings $\mathbf{c}$ and $\widehat{\mathbf{c}}$ will have at most $k_N \cdot \widehat{k}_N$ clusters with n_{ij} data points in cluster i, j for $i = 1, \ldots, k_N$, $j = 1, \ldots, \widehat{k}_N$ (some n_{ij} may equal zero, resulting in less than $k_N \cdot \widehat{k}_N$ clusters).

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

The last line follows from the fact that the pairs $(\mathbf{c}, \mathbf{c} \wedge \widehat{\mathbf{c}})$ and $(\widehat{\mathbf{c}}, \mathbf{c} \wedge \widehat{\mathbf{c}})$ are vertically aligned, see the proof of Property 3.2. The proof for $\tilde{\mathrm{B}}$ is similar. ∎

Proof of Property 3.4. First note that from Property 3.2

$$d(\mathbf{1}, \mathbf{0}) = d(\mathbf{1}, \mathbf{c}) + d(\mathbf{c}, \mathbf{c} \wedge \widehat{\mathbf{c}}) + d(\mathbf{0}, \mathbf{c} \wedge \widehat{\mathbf{c}}), \tag{7}$$

and

$$d(\mathbf{1}, \mathbf{0}) = d(\mathbf{1}, \widehat{\mathbf{c}}) + d(\widehat{\mathbf{c}}, \mathbf{c} \wedge \widehat{\mathbf{c}}) + d(\mathbf{0}, \mathbf{c} \wedge \widehat{\mathbf{c}}). \tag{8}$$

Combining (7) and (8), we have

$$2d(\mathbf{1}, \mathbf{0}) = d(\mathbf{1}, \mathbf{c}) + d(\mathbf{1}, \widehat{\mathbf{c}}) + d(\mathbf{c}, \mathbf{c} \wedge \widehat{\mathbf{c}}) + d(\widehat{\mathbf{c}}, \mathbf{c} \wedge \widehat{\mathbf{c}}) + 2d(\mathbf{0}, \mathbf{c} \wedge \widehat{\mathbf{c}}),$$

which, using Property 3.3, implies that

$$d(\mathbf{1}, \mathbf{0}) = \frac{1}{2}\left(d(\mathbf{1}, \mathbf{c}) + d(\mathbf{1}, \widehat{\mathbf{c}}) + d(\mathbf{c}, \widehat{\mathbf{c}})\right) + d(\mathbf{0}, \mathbf{c} \wedge \widehat{\mathbf{c}}).$$

Since d is a metric, the triangle inequality states that

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

Adding $d(\mathbf{c}, \widehat{\mathbf{c}})$ to either side, we have

$$d(\mathbf{c}, \widehat{\mathbf{c}}) \leq \frac{1}{2} \left(d(\mathbf{1}, \mathbf{c}) + d(\mathbf{1}, \widehat{\mathbf{c}}) + d(\mathbf{c}, \widehat{\mathbf{c}}) \right).$$

Thus,

$$\begin{aligned} d(\mathbf{c}, \widehat{\mathbf{c}}) \leq & \frac{1}{2} \left(d(\mathbf{1}, \mathbf{c}) + d(\mathbf{1}, \widehat{\mathbf{c}}) + d(\mathbf{c}, \widehat{\mathbf{c}}) \right) + d(\mathbf{0}, \mathbf{c} \wedge \widehat{\mathbf{c}}) - d(\mathbf{0}, \mathbf{c} \wedge \widehat{\mathbf{c}}) \\ & = d(\mathbf{1}, \mathbf{0}) - d(\mathbf{0}, \mathbf{c} \wedge \widehat{\mathbf{c}}) \leq d(\mathbf{1}, \mathbf{0}). \end{aligned}$$

The last two statements results from $\mathrm{VI}(\mathbf{1}, \mathbf{0}) = \log(N)$ and $\tilde{\mathrm{B}}(\mathbf{1}, \mathbf{0}) = 1 - \frac{1}{N}$. ∎

Proof of 3.5. From Property 3.3, the distance between $\mathbf{c}$ and any $\widehat{\mathbf{c}} \neq \mathbf{c}$ will be bounded below by the distance between $\mathbf{c}$ and their meet $\mathbf{c} \wedge \widehat{\mathbf{c}}$,

$$\mathrm{VI}(\mathbf{c}, \widehat{\mathbf{c}}) \geq \mathrm{VI}(\mathbf{c}, \mathbf{c} \wedge \widehat{\mathbf{c}}).$$

If $\mathbf{c} \wedge \widehat{\mathbf{c}} = \mathbf{c}$, then $\widehat{\mathbf{c}} > \mathbf{c}$, and there exists a $\mathbf{c}^m$ such that $\widehat{\mathbf{c}} \geq \mathbf{c}^m \succ \mathbf{c}$. From Property 3.2, $\mathrm{VI}(\mathbf{c}, \widehat{\mathbf{c}}) \geq \mathrm{VI}(\mathbf{c}, \mathbf{c}^m)$. Otherwise, $\mathbf{c} \wedge \widehat{\mathbf{c}} < \mathbf{c}$, and there exists a $\mathbf{c}^s$ such that $\mathbf{c} \wedge \widehat{\mathbf{c}} \leq \mathbf{c}^s \prec \mathbf{c}$. From Property 3.2, $\mathrm{VI}(\mathbf{c}, \mathbf{c} \wedge \widehat{\mathbf{c}}) \geq \mathrm{VI}(\mathbf{c}, \mathbf{c}^s)$. Thus the closest partitions to $\mathbf{c}$ will be among those which cover $\mathbf{c}$ or those which $\mathbf{c}$ covers.

Let's first consider the partitions which cover $\mathbf{c}$. If $\mathbf{c}^m \succ \mathbf{c}$, then $\mathbf{c}^m$ is obtained from $\mathbf{c}$ by merging two clusters i and j, and

$$\mathrm{VI}(\mathbf{c}, \mathbf{c}^m) = \frac{1}{N} \left((n_i + n_j) \log(n_i + n_j) - n_i \log(n_i) - n_j \log(n_j) \right), \tag{9}$$

which is minimized when i and j correspond to the two smallest clusters in $\mathbf{c}$. Next consider the partitions which $\mathbf{c}$ covers. If $\mathbf{c}^s \prec \mathbf{c}$, then $\mathbf{c}^s$ is obtained from $\mathbf{c}$ by splitting a cluster i of size $n_i > 1$ into two clusters i_1, i_2 of sizes n_{i_1} and n_{i_2}, and

$$\mathrm{VI}(\mathbf{c}, \mathbf{c}^s) = \frac{1}{N} \left(n_i \log(n_i) - n_{i_1} \log(n_{i_1}) - n_{i_2} \log(n_{i_2}) \right), \tag{10}$$

which is minimized when i is the smallest cluster of size $n_i > 1$ and $n_{i_1} = 1$ and $n_{i_2} = n_i - 1$. Thus, the closest partitions to $\mathbf{c}$ will be among those which merge the two smallest clusters or split the smallest cluster i of size $n_i > 1$ into a singleton and a cluster of size $n_i - 1$.

To compare (9) and (10), we first consider the case when $\mathbf{c}$ contains at least two singletons. In this case, the closest partitions which cover $\mathbf{c}$ merge two singletons, and (9) reduces to

$$\frac{1}{N} \left(2 \log(2) - 1 \log(1) - 1 \log(1) \right) = \frac{2}{N}.$$

Letting i be the index of the smallest cluster of size $n_i > 1$, the closest partition which $\mathbf{c}$ covers splits cluster i into a singleton and a cluster of size $n_i - 1$, and (10) reduces to

$$\frac{1}{N} \left(n_i \log(n_i) - (n_i - 1) \log(n_i - 1) \right).$$

For $n_i = 2$, these distances are equal, and the closest partitions to $\mathbf{c}$ will be those which merge any two singletons or split any cluster of size $n_i = 2$. For $n_i > 2$, the partitions which merge any two singletons are the closest.

Next, suppose that $\mathbf{c}$ contains at most one singleton, and let n_i, n_j denote the sizes of the smallest two clusters, where n_i corresponds to the size of the smaller cluster unless a singleton is present. The closest partitions which cover $\mathbf{c}$ merge any two clusters of sizes n_i and n_j, and (9) is

$$\frac{1}{N}\left((n_i + n_j)\log(n_i + n_j) - n_i \log(n_i) - n_j \log(n_j)\right). \tag{11}$$

The closest partitions which $\mathbf{c}$ covers split any cluster of size n_i into a singleton and a cluster of size $n_i - 1$, (10) reduces to

$$\frac{1}{N}\left(n_i \log(n_i) - (n_i - 1)\log(n_i - 1)\right). \tag{12}$$

In this case, (12) will be less than (11), and the closest partitions are those which split the smallest cluster i into a singleton and cluster of size $n_i - 1$. The proof for $\tilde{\mathrm{B}}$ is similar. ∎