Uncertainty Quantification in Bayesian clustering

Vacation Scholarship Project Report

Guanyu Chen
Supervised by Dr. Cecilia Balocchi
School of Mathematics
University of Edinburgh

August 2025

Abstract

With the ability to provide uncertainty in partition structure, Baysian Clustering method is highly appreciated. Nevertheless, the summarization of the posterior distribution over the clustering structure structureis always remains challenging. Previous studies only propose a single clustering estimate as representation of the posterior which leads to the ignorance of uncertainty and can even be unrepresentative in some multimodal posterior cases. However, the recent work of Balocchi and Wade (Balocchi & Wade, 2025) proposes a WASserstein Approximation for Bayesian clusterIng (WASABI) that approximates the posterior with multiple samples in a Wasserstein distance sense, under the Variation of Information (VI) metric on the partition space. In this project, we review the WASABI method and implement it with different metrics including Binder's loss (Binder, 1978; Dahl, 2006), the one minus Adjusted Rand index (omARI) (Hubert & Arabie, 1985), as well as the generalizations of VI loss and Binder's loss (Dahl et al., 2022). We also explore compare the performance of WASABI method under different metrics on different data sets, with a focus on Generalized 'n-invariant' Binder's loss.

Contents

1 Introduction

Aimed at uncovering the underlying structure of data by grouping similar data points, clustering is regarded as one of the most canonical tasks in unsupervised learning. In traditional clustering methods like k-means (Hartigan & Wong, 1979; Jain, 2010), however, this kind of methods only provide a single clustering estimate

without any uncertainty quantification. As an natural solution to the issues, Bayesian clustering methods combine the prior over the space of clustering with the loss to assess similarity between clusters, and provide a posterior distribution over the clustering structure. This posterior distribution can be used to to quantify beliefs and uncertainty for all possible clustering structures and patterns within.

However, summarizing this posterior distribution remains challenging due to the discrete nature of the partition space and the huge size of the space even for moderate sample sizes. For example, in Bayesian non-parametric model, the total number of way of partition is a Bell number, which means 20 data points can produce $B_{20} = 51,724,158,235,372$ possible partitions. As a result, various inference algorithms are used in approximating the posterior, among which Markov chain Monte Carlo (MCMC) is the most common tools to ease the computation, which provides substantial mostly unique clustering solutions that is drawing from the asymptotic exact approximation of the posterior.

Encountered with the overwhelming number of clustering solutions produced in inference procedure, a very natural task is to target a single clustering estimate. From the prospective of decision theory, the optimal Bayes estimator is found by the minimizing the posterior expected loss, which demands an appropriate loss function to meet the need.The most common used loss functions include Binder's loss (Binder, 1978; Dahl, 2006), the Variation of Information (VI) loss (Vinh et al., n.d.; Wade & Ghahramani, 2018), the adjusted Rand index (ARI) (Dahl et al., 2022; Hubert & Arabie, 1985) and the generalizations of VI loss and Binder's loss (Dahl et al., 2022). Nonetheless, as highlighted in the discussion of (Dahl et al., 2022), all these methods only concentrate on locating a single optimal clustering estimate.

Despite the single clustering solution provided, the Bayesian approach also provides important tools like the posterior similarity matrix, the elements of which represent the posterior probability of two data points clustering together, to provide uncertainty quantification via visualization and description. Except for the PSM, another significant tool is the usage of credible balls (Wade & Ghahramani, 2018), however, this require estimation of the bounds of the credible ball can sometimes be challenging thanks to the massive dimension of partition space.

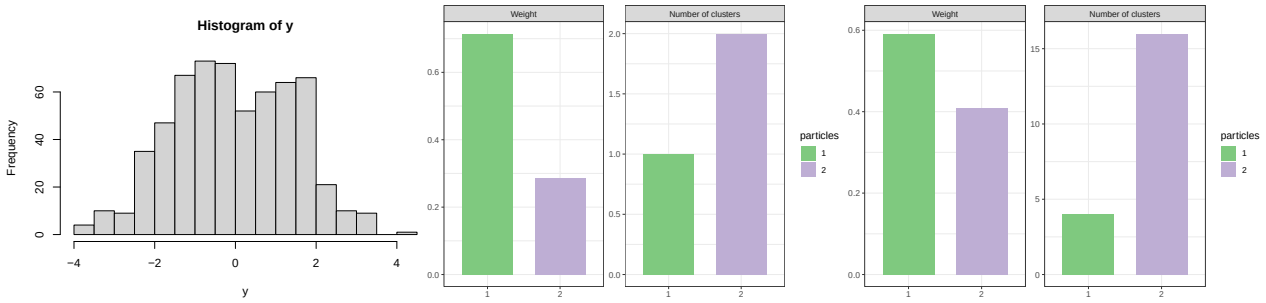
Though these tools are very powerful but still not sufficient enough to capture the uncertainty in the posterior distribution, in this case, Balocchi and Wade (Balocchi & Wade, 2025) propose a novel tool to enhance the understanding of this kind of uncertainty, which summarize the posterior with not one, but several optimal partitions, found by minimizing the Wasserstein distance equipped with an appropriate loss, named WASserstein Approximation for Bayesian clusterIng (WASABI). Except multiple optimal solutions, weights attached to each solution (known as particles) are also provided and different modes of clustering are also reflected. Thanks to the benefits of the VI loss discussed in (Wade & Ghahramani, 2018), the authors focus on the VI loss in their work.

However, other loss functions can be considered, in this project, we implement the WASABI method with other loss functions including Binder's loss, the one minus Adjusted Rand Index (omARI), as well as generalization of VI and Binder's loss and compare the performance of the corresponding WASABI estimators. The focus will fall on how the usage of Generalized Binder's loss affects the WASABI estimator compared to the original VI loss.

As a toy example, we can see how the change of loss function differs the corresponding WASABI estimator in **Example 1.** in (Balocchi & Wade, 2025).

From the plot we can see that with the use of generalized binder's loss function, the number of clusters of different particles differs drastically, which means that some other kinds of uncertainty may be captured with the use of different loss functions.

The code that generalize WASABI algorithm with other loss functions is available at https://github.com/guanyu-chen-gy/WASABI.ext.git, while the original version provide in (Balocchi & Wade, 2025) can be found on https://github.com/cecilia-balocchi/WASABI.



(a) Scatterplot of the data

(b) WASABI summary with traditional VI loss

(c) WASABI summary with GnBinder loss (a = 1.25)

Figure 1: Slightly bimodal example

2 Review

In the context of clustering, the observed data consists of measurements $\mathbf{y} = (y_1, ..., y_n)$ drawn from a heterogeneous population consisting of an unknown number homogeneous sub-populations. The observed $y_i \in \mathcal{Y}$ may be continuous, discrete, mixed, or more complex in nature (e.g. functional data). Each data point is associated with a discrete latent variable c_i (also called the allocation variable) indicating the group membership of the data point, i.e. $c_i = j$ if y_i belongs to the jth group, and $c_i = c_{i'}$ if y_i and $y_{i'}$ belong to the same sub-population. We are interested in obtaining estimates and understanding the uncertainty of clustering structure characterized by the latent $\rho_n = (c_1, ..., c_n)$. To achieve this, the Bayesian approach constructs a posterior distribution over the space of partitions,

$$\pi(\rho_n|\mathbf{y}) \propto p(\mathbf{y}|\rho_n)\pi(\rho_n) \tag{1}$$

where $\pi(\rho_n)$ is the prior over the partition space and $p(\mathbf{y}|\rho_n)$ is the likelihood in a model-based approach or defined based on a loss function in a loss-based approach (RIGON et al., 2023). For reviews on Bayesian clustering, please refer to (Wade & Ghahramani, 2018).

It is worth emphasizing that clustering is often referred to as an ill-posed problem, as it aims to discover unknown patterns or structures in the data. The notion of a cluster depends on the application at hand and can often be challenging to characterize formally. A unique clustering solution often does not exist (Hennig, 2015). Thus, one must carefully consider the model or loss employed and importantly, also characterize uncertainty in the clustering solution. To achieve the latter, Bayesian cluster analysis provides a formal framework through both the posterior distribution over the entire space of clusterings and by creating an ensemble of clustering solutions sampled from the posterior. Moreover, this also helps to mitigate sensitivity to local optimum which adversely impact all clustering algorithms due to the sheer size of the space.

2.1 Random Partition Models

A key ingredient in the Bayesian approach is the prior $\pi(\rho_n)$ over the clustering structure. This is referred to as *random partition model*,as it involves randomly assigning observations to clusters, or equivalently, randomly partitioning the indices of the data points {1,...,n} into Kn non-empty and mutually exhaustive sets C_j for $j = 1, ..., K_n$, where K_n represents the number of clusters. Thus, the partition ρ_n can also be presented as $\rho_n = C_1, ..., C_{K_n}$, where the sets satisfy $C_j \cap C_{j'} = \emptyset$ for $j \neq j'$ and $C_1 \cup ... \cup C_{K_n} = \{1, ..., n\}$; essentially, each C_j contains the indices of data points in the jth cluster. For ease of notation, the sample size is dropped from ρ_n and K_n, when the context is clear.

In practice, the number of clusters K is rarely known, and choosing this number is an important and difficult concern in cluster analysis. While information criteria or model selection tools, such as the Bayesian information criterion (BIC), are commonly used, this ignores uncertainty in the number of clusters and disregards information from the numerous models fits using different choices of K.Instead, our focus is

approaches that account for this uncertainty, namely, through 1) mixtures of finite mixtures, which naturally incorporate a prior on the unknown K (Miller & Harrison, 2018; Nobile & Fearnside, 2007; Richardson & Green, n.d.);2) sparse overfitted mixtures that specify an upper bound on the number of clusters and encourage extra components to be emptied through sparsity promoting priors (Malsiner-Walli et al., 2016; Rousseau & Mengersen, 2011); and 3) Bayesian nonparametric (BNP) mixtures (Müller, 2019) that allow the number of clusters to grow unboundedly with the data, with the Dirichlet process (DP) mixture (*On a Class of Bayesian Nonparametric Estimates*, n.d.) being the most-widely used example. For example, letting L denote an upper bound on the number of clusters and $\gamma > 0$, the random partition model for the overfitted mixture (corresponding to a Dirichlet prior on the weights of the mixture with L components) is

$$\pi(\rho) \;=\; \frac{\Gamma(\gamma L)\, L!}{\Gamma(\gamma L + n)(L-K)!} \prod_{j=1}^{K} \frac{\Gamma(n_j + \gamma)}{\Gamma(\gamma)},$$

with $n_j = |C_j|$ denoting the cluster size. In general, random partition models specify priors on the discrete, partially-ordered space of partitions, denoted by $\mathcal{P}$, which has massive dimension, even more so when accounting for the unknown number of clusters.

2.2 Loss Functions

The massive dimension makes computing the posterior in eq.(1) infeasible. Instead, MCMC is commonly used to obtain asymptotically exact draws, denoted by $\rho^{(t)}$ for $t = 1, ..., T$. The number of draws is typically in the tens to hundreds of thousands, and the $\rho^{(t)}$ are mostly unique (due to the size of the space), vary in the number of clusters, and suffer from label-switching.

To provide a single, representative clustering solution, the optimal Bayes estimator minimizes the posterior expected loss. For a partition-specific loss, (Wade & Ghahramani, 2018) recommend the Variation of Information (VI) (Vinh et al., n.d.), which defines the distance between any two partitions ρ_1 and ρ_2 as:

$$\begin{aligned} L_{\mathrm{VI}}(\rho_1, \rho_2) &= \mathrm{H}(\rho_1) + \mathrm{H}(\rho_2) - 2\mathrm{MI}(\rho_1, \rho_2) \\ &= \sum_{j_1=1}^{K_1} \frac{n_{j_1,+}}{n} \log_2\!\left(\frac{n_{j_1,+}}{n}\right) + \sum_{j_2=1}^{K_2} \frac{n_{+,j_2}}{n} \log_2\!\left(\frac{n_{+,j_2}}{n}\right) - 2\sum_{j_1=1}^{K_1}\sum_{j_2=1}^{K_2} \frac{n_{j_1,j_2}}{n} \log_2\!\left(\frac{n_{j_1,j_2}\, n}{n_{j_1,+}\, n_{+,j_2}}\right). \end{aligned}$$

where the first two terms represent the entropy of the two partitions and the last term is the mutual information between the partitions. The counts $n_{j1,j2}$ represent the cross-tabulation between the partitions, i.e. $n_{j1,j2}$ is the number of data points in cluster j_1 in ρ_1 and cluster j_2 in ρ_2, and K_1 and K_2 are the number of clusters in ρ_1 and ρ_2, respectively.

Meanwhile, (Wade & Ghahramani, 2018) also introduce an "n-invariant" version of Binder's loss, which can be presented as

$$\begin{aligned} L_{\mathrm{nBinder}}(\rho_1, \rho_2) &= \frac{2}{n^2} L_{\mathrm{Binder}}(\rho_1, \rho_2) \\ &= \sum_{j_1=1}^{K_1} (\frac{n_{j_1,+}}{n})^2 + \sum_{j_2=1}^{K_2} (\frac{n_{+,j_2}}{n})^2 - 2\sum_{j_1=1}^{K_1}\sum_{j_2=1}^{K_2} (\frac{n_{j_1,j_2}}{n})^2. \end{aligned}$$

More generally, (Dahl et al., 2022) provides a generalized version of VI's loss and Binder's loss, where the generalized variation of information (GVI) can be defined as

$$L_{\mathrm{GVI}}(\rho_1, \rho_2) = a \sum_{j_1=1}^{K_1} \frac{n_{j_1,+}}{n} \log_2\!\left(\frac{n_{j_1,+}}{n}\right) + b \sum_{j_2=1}^{K_2} \frac{n_{+,j_2}}{n} \log_2\!\left(\frac{n_{+,j_2}}{n}\right) - (a+b) \sum_{j_1=1}^{K_1}\sum_{j_2=1}^{K_2} \frac{n_{j_1,j_2}}{n} \log_2\!\left(\frac{n_{j_1,j_2}\, n}{n_{j_1,+}\, n_{+,j_2}}\right).$$

and generalization of "n-invariant" Binder's loss can be defined as

$$L_{\text{GnBinder}}(\rho_1, \rho_2) = a \sum_{j_1=1}^{K_1} (\frac{n_{j_1,+}}{n})^2 + b \sum_{j_2=1}^{K_2} (\frac{n_{+,j_2}}{n})^2 - (a+b) \sum_{j_1=1}^{K_1} \sum_{j_2=1}^{K_2} (\frac{n_{j_1,j_2}}{n})^2.$$

It is worth pointing out that, in GnBinder loss a represents the cost of failing to cluster together two items which should in fact be clustered together, whereas b represents the cost of clustering two items which should in fact be separate (Dahl et al., 2022). For the sake of simplicity, we can set $a \in [0, 2]$ and $b = 2 - a$, and we will follow this custom throughout the report, that is

$$L_{\text{GVI}}(\rho_1, \rho_2) = a \sum_{j_1=1}^{K_1} \frac{n_{j_1,+}}{n} \log_2\left(\frac{n_{j_1,+}}{n}\right) + (2-a) \sum_{j_2=1}^{K_2} \frac{n_{+,j_2}}{n} \log_2\left(\frac{n_{+,j_2}}{n}\right) - 2 \sum_{j_1=1}^{K_1} \sum_{j_2=1}^{K_2} \frac{n_{j_1,j_2}}{n} \log_2\left(\frac{n_{j_1,j_2}\, n}{n_{j_1,+}\, n_{+,j_2}}\right).$$

and

$$L_{\text{GnBinder}}(\rho_1, \rho_2) = a \sum_{j_1=1}^{K_1} (\frac{n_{j_1,+}}{n})^2 + (2-a) \sum_{j_2=1}^{K_2} (\frac{n_{+,j_2}}{n})^2 - 2 \sum_{j_1=1}^{K_1} \sum_{j_2=1}^{K_2} (\frac{n_{j_1,j_2}}{n})^2.$$

Thus, the Bayes estimator under the VI loss (called the minVI partition) is:

$$\rho^* = \min_{\rho \in \mathcal{P}} \mathbb{E}[L_{\text{VI}}(\rho, \rho') \mid \mathbf{y}] \approx \min_{\rho \in \mathcal{P}} \frac{1}{T} \sum_{t=1}^{T} L_{\text{VI}}(\rho, \rho^{(t)}).$$

Similarly, the Bayes estimator under the Generalized "n-invariant" Binder loss (called the minB partition) is:

$$\rho^* = \min_{\rho \in \mathcal{P}} \mathbb{E}[L_{\text{GnBinder}}(\rho, \rho') \mid \mathbf{y}] \approx \min_{\rho \in \mathcal{P}} \frac{1}{T} \sum_{t=1}^{T} L_{\text{GnBinder}}(\rho, \rho^{(t)}).$$

While direct minimization cannot be performed, effective algorithms can be found in (Dahl et al., 2022; Wade & Ghahramani, 2018).

3 Classic WASABI Methods and Generalization

In (Balocchi & Wade, 2025), the WASABI approach use ideas from (Balocchi et al., 2023) to approximate the posterior $\pi(\rho|\mathbf{y})$ with a discrete distribution $q = \sum_{l=1}^{L} \omega_l \delta_{\rho_l}$ supported on a small number L of partitions. Unlike (Balocchi et al., 2023) that focuses on minimizing the Kullback-Leibler (KL) divergence, (Balocchi & Wade, 2025) use Wasserstein distance instead, which allows the use of metrics on the space of partitions. Here VI metric is used in Wasserstein distance, as a result, we will use the word "VI" to stress the original results from (Balocchi & Wade, 2025).

3.1 Theory

3.1.1 Review

The Wasserstein-VI distance is defined as follows, let $(\mathcal{P}, d_{\text{VI}})$ be the metric space on the set of partitions $\mathcal{P}$ embedded with the VI metric d_{VI}. Let p and q represent two distributions on the space of partitions $(\mathcal{P}, d_{\text{VI}})$. If $\mathcal{J}(p, q)$ is the set of couplings of p and q, i.e. the collection of distributions $J(\rho, \rho')$ on $\mathcal{P} \times \mathcal{P}$ with marginals p and q on the first and second factor respectively, then we define the Wasserstein-VI distance as

$$W_{\mathrm{VI}}(p,q) = \inf_{J\in\mathcal{J}} \sum_{\rho,\rho'\in\mathcal{P}} d_{\mathrm{VI}}(\rho,\rho')J(\rho,\rho').$$

Denote with $\mathcal{Q}_L = \{\sum_{l=1}^{L}\omega_\ell\delta_{\rho_\ell} : \sum_{\ell=1}^{L}\omega_\ell = 1, \omega_\ell \geq 0, \rho_\ell \in \mathcal{P}, \text{for } \ell = 1,..,L\}$ the collection of discrete distributions supported on L points.

The WASABI-VI posteior is defined as the discrete distribution $q^*_{\mathrm{VI}} = \sum_{\ell=1}^{L}\omega^*_\ell\delta_{\rho^*_\ell} \in \mathcal{Q}_L$ that best approximates the posterior in a Wasserstein-VI sense.

$$q^*_{\mathrm{VI}} = \underset{q\in\mathcal{Q}_L}{\arg\min} W_{VI}\left(\pi(\cdot \mid \mathbf{y}), q(\cdot)\right). \tag{2}$$

In this case, the quality of the approximation is quantified by the Wasserstein-VI distance, $W_{\mathrm{VI}}(\pi(\cdot \mid \mathbf{y}), q^*(\cdot))$, which measures minimal amount of uncertainty lost when summarizing L partitions.

In general, the WASABI-VI posterior q^*, the solution to (2) is found by identifying a set of "centers" or "particles" $^* = \{\rho^*_1, ..., \rho^*_{\mathrm{L}}\}$ that minimize

$$\sum_{\ell=1}^{L}\sum_{\rho\in\mathcal{N}_\ell} d_{\mathrm{VI}}(\rho,\rho^*_\ell)\pi(\rho \mid \mathbf{y}) \tag{3}$$

where $\mathcal{N}_\ell = \{\rho : d_{\mathrm{VI}}(\rho,\rho^*_\ell) < d_{\mathrm{VI}}(\rho,\rho^*_{\ell'}) \text{ for all } \ell' \neq \ell\}$ corresponds to the set of partitions that are closer to ρ^*_ℓ than to any other center, with the name "region of attraction" of the center ρ^*_ℓ. The optimal weights ω^*_ℓ associated to each center are found by $\omega^*_\ell = \sum_{\rho\in\mathcal{N}_\ell}\pi(\rho \mid \mathbf{y})$.

This means that the WASABI-VI posterior is supported on the centers ρ^*_ℓ that minimize (3), and the probability ω^*_ℓ associated with each of them is equal to the posterior mass of its region of attraction.

However, due to the massive dimension of $\mathcal{P}$, in practice, the objective function in (3) is intractable, in this case, we should consider a similar approximation of the posterior when MCMC estimate of posterior distribution is provided. Let $\hat{\pi}(\rho \mid \mathbf{y})$ be the approximation to the posterior provided by the MCMC draws, $\hat{\pi}(\rho \mid \mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T}\delta_{\rho^{(t)}}$. Then the WASABI-VI posterior

$$q^* = \underset{q\in\mathcal{Q}_L}{\arg\min} W_{\mathrm{VI}}\left(\hat{\pi}(\cdot \mid \mathbf{y}), q(\cdot)\right). \tag{4}$$

With the properties of the Wasserstein distance between two discrete distributions, the solution to (4), i.e. the WASABI-VI posterior, is supported on the "centers" or "particles" $\rho^* = \{\rho^*_1, ..., \rho^*_L\}$ which minimize

$$\frac{1}{T}\sum_{\ell=1}^{L}\sum_{\rho^{(t)}\in\hat{\mathcal{N}}_\ell} d_{\mathrm{VI}}(\rho^{(t)},\rho^*_\ell) \tag{5}$$

where $\hat{\mathcal{N}}_\ell = \{\rho^{(t)} : d_{\mathrm{VI}}(\rho^{(t)},\rho^*_\ell) < d_{\mathrm{VI}}(\rho^{(t)},\rho^*_{\ell'}) \text{ for all } \ell' \neq \ell\}$ is referred as the "empirical region of attraction" of ρ^*_l. The optimal weights $\omega^*_\ell = |\hat{\mathcal{N}}_\ell|/T$.

Described by (Canas & Rosasco, 2012) and first suggested by (Pollard, 1982), the connection between k-means algorithm and the approximation in a Wasserstein sense (equipped with the Euclidean distance) of an empirical distribution with a discrete distribution supported on k points, shows that the k-means algorithm provides a locally optimal solution to this problem. So in (Balocchi & Wade, 2025), a locally optimal set of partitions that minimize (5) can be obtained via a k-medoids-like algorithm, which generalizes k-means beyond the Euclidean distance, to the set of MCMC samples $\{\rho^{(t)} : t = 1, ..., T\}$, with $k = L$ and distance equal to the VI distance. In other words, the MCMC-sampled partitions are clustered into L groups, and each of these groups is represented by one of the centers ρ^*_ℓ, which are the same centers that support the WASABI-VI posterior q^* with weights determined by the size of each group.

3.1.2 Generalization

Similarly, we can define the Wasserstein distance on top of other metrics. And we should always aware that the only reason why the above procedures, (3) and (5) can be directly applied to the process finding the other WASABI posterior is that the properties of Wasserstein distance used to prove the propositions in (Balocchi & Wade, 2025) are general, in other words, do not rely on the metric used in partition space!

Without loss of generality, we can use Generalized "n-invariant" Binder's loss for example, let $(\mathcal{P}, d_{\text{GnBinder}})$ be the metric space on the set of partitions $\mathcal{P}$ embedded with the metric Generalized "n-invariant" Binder's loss d_{GnBinder}. Let p and q represent two distributions on the space of partitions $(\mathcal{P}, d_{\text{GnBinder}})$. If $\mathcal{J}(p, q)$ is the set of couplings of p and q, i.e. the collection of distributions $J(\rho, \rho')$ on $\mathcal{P} \times \mathcal{P}$ with marginals p and q on the first and second factor respectively, then the Wasserstein-GnBinder distance is

$$W_{\text{GnBinder}}(p, q) = \inf_{J \in \mathcal{J}} \sum_{\rho, \rho' \in \mathcal{P}} d_{\text{GnBinder}}(\rho, \rho') J(\rho, \rho').$$

Then the WASABI-GnBinder posteior is defined as the discrete distribution $q^*_{\text{GnBinder}} = \sum_{\ell=1}^{L} \omega^*_\ell \delta_{\rho^*_\ell} \in \mathcal{Q}_L$ that best approximates the posterior in a Wasserstein-GnBinder sense.

$$q^*_{\text{GnBinder}} = \underset{q \in \mathcal{Q}_L}{\arg\min} W_{\text{GnBinder}} \left(\pi(\cdot \mid \mathbf{y}), q(\cdot) \right). \tag{6}$$

Now, as opposed to (2), the quality of the approximation is quantified by this new kind of Wasserstein distance, the Wasserstein-GnBinder distance, $W_{\text{GnBinder}}(\pi(\cdot \mid \mathbf{y}), q^*(\cdot))$, that measures minimal amount of uncertainty lost when summarizing L partitions.

Analogue to (3), we see that the WASABI-GnBinder posterior q^*, the solution to (6) is found by identifying a set of "centers" or "particles" $^* = \{\rho^*_1, ..., \rho^*_{\text{L}}\}$ that minimize

$$\sum_{\ell=1}^{L} \sum_{\rho \in \mathcal{N}_\ell} d_{\text{GnBinder}}(\rho, \rho^*_\ell) \pi(\rho \mid \mathbf{y}) \tag{7}$$

where $\mathcal{N}_\ell = \{\rho : d_{\text{GnBinder}}(\rho, \rho^*_\ell) < d_{\text{GnBinder}}(\rho, \rho^*_{\ell'}) \text{ for all } \ell' \neq \ell\}$ corresponds to the set of partitions that are closer to ρ^*_ℓ than to any other center, with the name "region of attraction" of the center ρ^*_ℓ. The optimal weights ω^*_ℓ associated to each center are found by $\omega^*_\ell = \sum_{\rho \in \mathcal{N}_\ell} \pi(\rho \mid \mathbf{y})$.

Also, because of the same reason that $\mathcal{P}$ holds a massive dimension, the intractable expression (7) can be found by MCMC draws.

Moreover, since the general Wasserstein distance's properties used in proving (5) do not rely on the specific metric used, following the same procedure, we use MCMC draws to find out the WASABI-GnBinder posterior, which is supported supported on the "centers" or "particles" $\rho^* = \{\rho^*_1, ..., \rho^*_L\}$ that minimize

$$\frac{1}{T} \sum_{\ell=1}^{L} \sum_{\rho^{(t)} \in \widehat{\mathcal{N}}_\ell} d_{\text{GnBinder}}(\rho^{(t)}, \rho^*_\ell) \tag{8}$$

where $\widehat{\mathcal{N}}_\ell = \{\rho^{(t)} : d_{\text{GnBinder}}(\rho^{(t)}, \rho^*_\ell) < d_{\text{GnBinder}}(\rho^{(t)}, \rho^*_{\ell'}) \text{ for all } \ell' \neq \ell\}$ is referred as the "empirical region of attraction" of ρ^*_l. The optimal weights $\omega^*_l = |\widehat{\mathcal{N}}_\ell|/T$.

Moreover, when we applied the k-medoids-like algorithm in (Balocchi & Wade, 2025), we can find a locally optimal solution that minimize (8).

The above example that used General "n-invariant" Binder's loss as a specific case but one can always follow the same process and define Wassertein distance as well as the corresponding WASABI posterior to do uncertainty quantification under whatever metric in partition space you may like.

3.2 Generalized Algorithm

In (Balocchi & Wade, 2025), the search algorithm to find the L centers that approximate WASABI-VI posterior use a similar structure to the k-medoids algorithm, and it can be promoted to find out WASABI posterior no matter what loss function is used, with a minor modification in some of the step, which is nothing but use specified metric to substitute VI distance in the original algorithm. A more general version can be described as follows.

In short, given an initialization of the centers, the algorithm alternates between the $N-update$ step, where the regions of attraction are updated, and the *search* step, where the new center for each region of attraction corresponds to the partition that minimize specific loss function, which is minVI in the original algorithm, of each group. And in operation of $N-update$ and *search*, we are no longer merely using only VI distance in the computation, but any specified metric. This can be described in the following diagram.

Here is the algorithm:

Algorithm 1: Find the approximate WASABI posterior

Input: MCMC samples $\{\rho^{(t)}\}_{t=1}^{T}$, number of particles L, initialization method *init*, tolerance ϵ, loss function Q

Output: $\{\rho_1^*, \dots, \rho_L^*\}, \mathbf{w} = (w_1, \dots, w_L)$

Initialize $\rho_{1:L}^*$ using method *init*;

repeat

  N-update step:;

  Compute $Q(\rho^{(t)}, \rho_\ell^*)$ for all t and ℓ;

  Assign each $\rho^{(t)}$ to closest center ρ_ℓ^* and update region of attraction $\mathcal{N}_\ell$;

  if *any region of attraction* $\mathcal{N}_\ell$ *is empty* **then**

    Replace ρ_ℓ^* with a distant partition $\rho^{(t)}$ and set $\mathcal{N}_\ell = \{\rho^{(t)}\}$;

  search step:;

  for $\ell = 1, \dots, L$ **do**

    Update centers: $\rho_\ell^* \leftarrow \text{minQ}(\mathcal{N}_\ell)$;

    Update centers' expected-Q: $l_\ell \leftarrow \text{EQ}(\rho_\ell^*, \mathcal{N}_\ell)$;

    Update weights: $w_\ell \leftarrow |\mathcal{N}_\ell|/T$;

  Update loss: $W \leftarrow \sum_{\ell=1}^{L} w_\ell \cdot l_\ell$;

until *change of* W *is less than* ϵ;

3.3 Generalization of VIC and EVIC

In (Balocchi & Wade, 2025), as a tool for the comparison of two specific partitions, the VI contribution is defined as follows,

$$VIC_i(\rho_1, \rho_2) = \frac{1}{n}\left[\log_2\left(\frac{\sum_{i'=1}^{n} 1(c_{1,i} = c_{1,i'})}{n}\right) + \log_2\left(\frac{\sum_{i'=1}^{n} 1(c_{2,i} = c_{2,i'})}{n}\right) - 2\log_2\left(\frac{\sum_{i'=1}^{n} 1(c_{1,i} = c_{1,i'},\, c_{2,i} = c_{2,i'})}{n}\right)\right].$$

Following the same idea, we can derive Generalized VI contribution (GVIC), Generalized "n-invariant" Binder contribution (GnBC) as follows,

$$GVIC_i(\rho_1,\rho_2) = \frac{1}{n}\left[a\log_2\left(\frac{\sum_{i'=1}^{n}1(c_{1,i}=c_{1,i'})}{n}\right)+(2-a)\log_2\left(\frac{\sum_{i'=1}^{n}1(c_{2,i}=c_{2,i'})}{n}\right) - 2\log_2\left(\frac{\sum_{i'=1}^{n}1(c_{1,i}=c_{1,i'},\, c_{2,i}=c_{2,i'})}{n}\right)\right].$$

$$GnBC_i(\rho_1,\rho_2) = a\left(\frac{\sum_{i'=1}^{n}1(c_{1,i}=c_{1,i'})}{n}\right)^2+(2-a)\left(\frac{\sum_{i'=1}^{n}1(c_{2,i}=c_{2,i'})}{n}\right)^2 - 2\left(\frac{\sum_{i'=1}^{n}1(c_{1,i}=c_{1,i'},\, c_{2,i}=c_{2,i'})}{n}\right)^2$$

Similarly, for VI contribution by group (VICG),

$$VICG_k(\rho_1,\rho_2) = \frac{n_{k_1,k_2}}{n}\left[\log_2\left(\frac{n_{k_1,+}}{n}\right)+\log_2\left(\frac{n_{+,k_2}}{n}\right)-2\log_2\left(\frac{n_{k_1,k_2}}{n}\right)\right].$$

we can derive Generalized VI contribution by group (GVICG) and Generalized "n-invarinat" Binder contribution by group (GnBCG),

$$GVICG_k(\rho_1,\rho_2) = \frac{n_{k_1,k_2}}{n}\left[a\log_2\left(\frac{n_{k_1,+}}{n}\right)+(2-a)\log_2\left(\frac{n_{+,k_2}}{n}\right)-2\log_2\left(\frac{n_{k_1,k_2}}{n}\right)\right],$$

$$GnBCG_k(\rho_1,\rho_2) = \frac{n_{k_1,k_2}}{n}\left[a\left(\frac{n_{k_1,+}}{n}\right)^2+(2-a)\left(\frac{n_{+,k_2}}{n}\right)^2-2\left(\frac{n_{k_1,k_2}}{n}\right)^2\right].$$

For EVI contribution (EVIC),

$$EVIC_i(\rho^*) = \frac{1}{n}\left\{\log_2\left(\frac{\sum_{i'=1}^{n}1(c_i^*=c_{i'}^*)}{n}\right)+\mathbb{E}\left[\log_2\left(\frac{\sum_{i'=1}^{n}1(c_i=c_{i'})}{n}\right)\Big|\mathbf{y}\right] - 2\,\mathbb{E}\left[\log_2\left(\frac{\sum_{i'=1}^{n}1(c_i^*=c_{i'}^*,\, c_i=c_{i'})}{n}\right)\Big|\mathbf{y}\right]\right\},$$

which provides a measure of uncertainty in each point's cluster allocation, we can also derive versions of EnVI and EnBC as analogues for Generalized VI and Generalized "n-invariant" Binder loss,

$$EGVIC_i(\rho^*) = \frac{1}{n}\left\{a\log_2\left(\frac{\sum_{i'=1}^{n}1(c_i^*=c_{i'}^*)}{n}\right)+(2-a)\mathbb{E}\left[\log_2\left(\frac{\sum_{i'=1}^{n}1(c_i=c_{i'})}{n}\right)\Big|\mathbf{y}\right] - 2\,\mathbb{E}\left[\log_2\left(\frac{\sum_{i'=1}^{n}1(c_i^*=c_{i'}^*,\, c_i=c_{i'})}{n}\right)\Big|\mathbf{y}\right]\right\},$$

$$EnBC_i(\rho^*) = \left\{a\left(\frac{\sum_{i'=1}^{n}1(c_i^*=c_{i'}^*)}{n}\right)^2+\mathbb{E}\left[(2-a)\left(\frac{\sum_{i'=1}^{n}1(c_i=c_{i'})}{n}\right)^2\Big|\mathbf{y}\right] - 2\,\mathbb{E}\left[\left(\frac{\sum_{i'=1}^{n}1(c_i^*=c_{i'}^*,\, c_i=c_{i'})}{n}\right)^2\Big|\mathbf{y}\right]\right\}.$$

Similarly the MCMC based estimator can be found as

$$\widehat{EGVIC}_i(\rho^*) = \frac{1}{n}\left[a\log_2\left(\frac{\sum_{i'=1}^n 1(c_i^* = c_{i'}^*)}{n}\right) + (2-a)\frac{1}{T}\sum_{t=1}^T \log_2\left(\frac{\sum_{i'=1}^n 1(c_i^{(t)} = c_{i'}^{(t)})}{n}\right) - 2\frac{1}{T}\sum_{t=1}^T \log_2\left(\frac{\sum_{i'=1}^n 1(c_i^* = c_{i'}^*,\, c_i^{(t)} = c_{i'}^{(t)})}{n}\right)\right],$$

and

$$\widehat{EnBC}_i(\rho^*) = \left\{a\left(\frac{\sum_{i'=1}^n 1(c_i^* = c_{i'}^*)}{n}\right)^2 + \frac{1}{T}\sum_{t=1}^T\left[(2-a)\left(\frac{\sum_{i'=1}^n 1(c_i^{(t)} = c_{i'}^{(t)})}{n}\right)^2\right] - 2\left[\left(\frac{\sum_{i'=1}^n 1(c_i^* = c_{i'}^*,\, c_i^{(t)} = c_{i'}^{(t)})}{n}\right)^2\right]\right\}.$$

and the WASABI approximation can be written as

$$\widehat{\widehat{EGVIC}}_i(\rho^*) = \frac{1}{n}\left[a\log_2\left(\frac{\sum_{i'=1}^n 1(c_i^* = c_{i'}^*)}{n}\right) + \sum_{\ell=1}^L w_\ell \log_2\left(\frac{\sum_{i'=1}^n 1(c_{\ell,i}^* = c_{\ell,i'}^*)}{n}\right) - 2\sum_{\ell=1}^L w_\ell \log_2\left(\frac{\sum_{i'=1}^n 1(c_i^* = c_{i'}^*,\, c_{\ell,i}^* = c_{\ell,i'}^*)}{n}\right)\right],$$

and

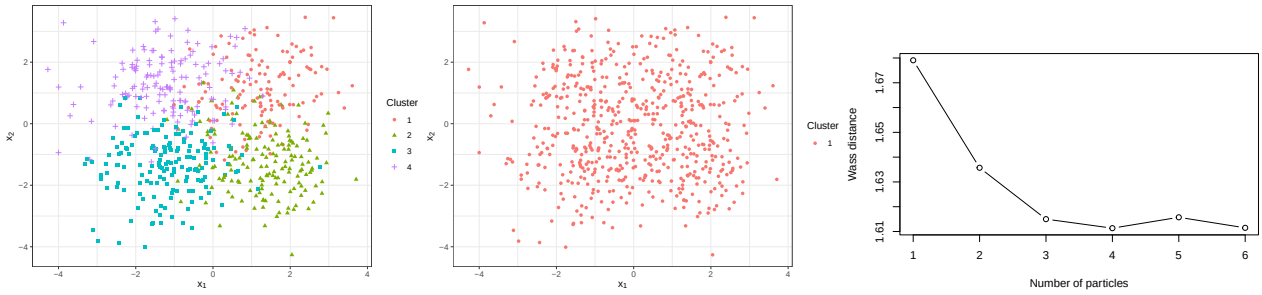
$$\widehat{\widehat{EnBC}}_i(\rho^*) = \left\{a\left(\frac{\sum_{i'=1}^n 1(c_i^* = c_{i'}^*)}{n}\right)^2 + \sum_{\ell=1}^L \omega_\ell\left[(2-a)\left(\frac{\sum_{i'=1}^n 1(c_{\ell,i}^* = c_{\ell,i'}^*)}{n}\right)^2\right] - 2\sum_{\ell=1}^L \omega_\ell\left[\left(\frac{\sum_{i'=1}^n 1(c_i^* = c_{i'}^*,\, c_{\ell,i}^* = c_{\ell,i'}^*)}{n}\right)^2\right]\right\}.$$

4 Experiment

In this part, we will focus on Two-dimensional extension of the bimodal example, which is **Example 2.** in (Balocchi & Wade, 2025).

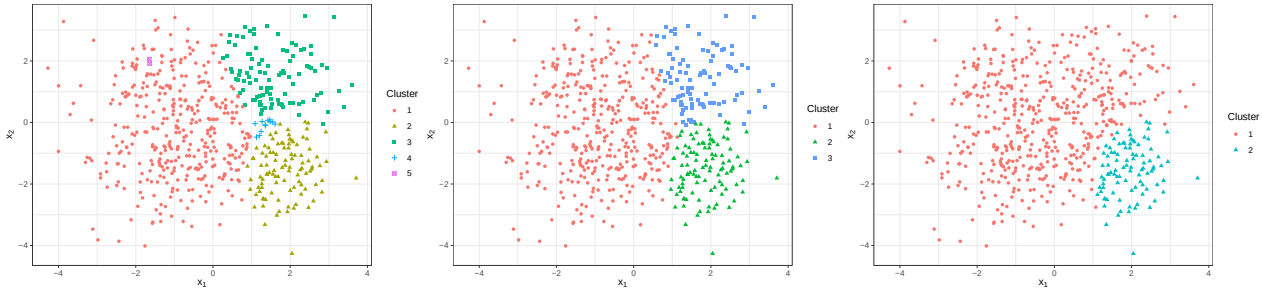
Consider a two-dimensional Gaussian mixture with four components, each with mean $(\pm m, \pm m)$ located in one of the four quadrants and diagonal covariance matrix with unit variance. Figure 2(a) displays the data colored by the partition used in the data-generating process for $m = 1.25$. MCMC samples are obtained by fitting a diagonal location-scale Dirichlet process mixture (DPM) using the BNPmix R package (Corradin et al., 2021). The minVI partition (Figure 2(b)) merges only one component, which does not reflect the axis-aligned elliptical clusters of the DPM model. This highlights that a single-point estimate can be misleading when the posterior is multimodal. WASABI elbow plot (Figure 2(c)) suggests that multiple estimators are useful for describing the posterior.

Moreover, we can check how the WASABI estimator changes with respect to other loss functions. In this experiment, we will focus on Generalized "n-invariant" Binder's loss. First of all, we should figure out what value of parameter a should be used for discussion. We do this by verifying salso estimate (Dahl et al., 2022).



(a) Scatterplot of the data (b) minVI partition (c) WASABI elbow plot for VI

Figure 2: Two-dimensional extension of the bimodal example

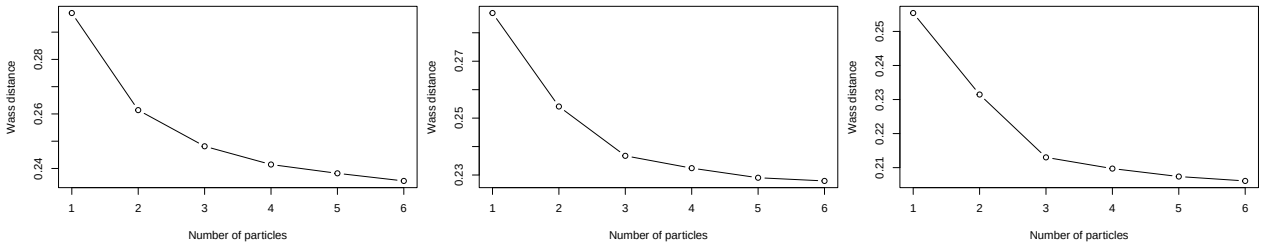


(a) salso estimate for a = 1.17 (b) salso estimate for a = 1.3 (c) salso estimate for a = 1.5

Figure 3: Salso estimate changes with respect to a

From 3 we can see that the number of clusters decreases as a grows larger, this is not surprising because a represents the cost of failing to cluster together two items which should in fact be clustered together, in this case.

To see the changes, for the above chosen values of a, we draw the elbow plots (Figure 4) and visualize the output WASABI approximation (Figure 5).



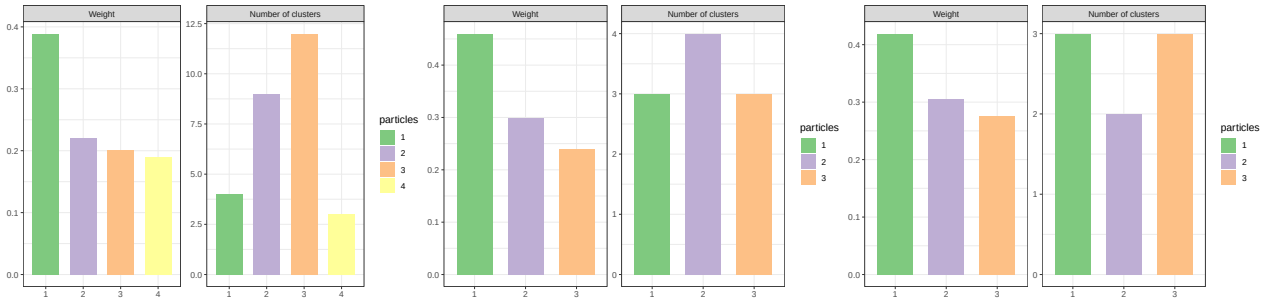
(a) a = 1.15 (b) a = 1.3 (c) a = 1.5

Figure 4: Elbow plots for WASABI-GnBinder estimate

From Figure 4, we can tell that $L = 4$ particles achieves a balance between parsimony and minimizing the objective in (a), while $L = 3$ holds the same benefit in (b) and (a).

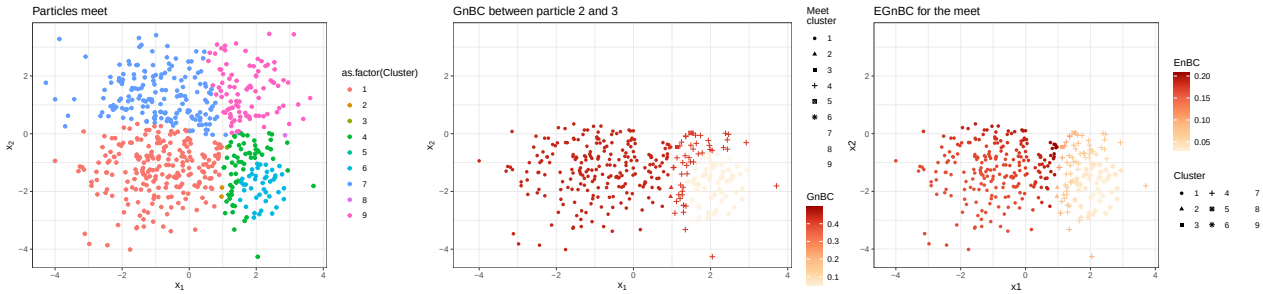
From figure 3 and 4, we see that some of the WASABI particles with the same value of "a" display a larger number of clusters and the number of particles changes for different number of clusters. In this case, when we do WASABI estimate, it is worth considering different kinds of loss functions for the WASABI estimator for uncertainty quantification.

Finally, using GnBinder($a = 1.5$) as a particular example (6), we can look at the visualization that use GnBC (b) and EGnBC (c).



(a) a = 1.15 (b) a = 1.3 (c) a = 1.5

Figure 5: Elbow plots for WASABI-GnBinder estimate



(a) Particle meet (b) GnBC for the meet (c) EGnBC for the meet

Figure 6: Visualization for WASABI-GnBinder(a = 1.5) estimator

5 Acknowledgments

I would like to express my sincere gratitude to Dr. Cecilia Balocchi for her patient guidance, insightful suggestions, and generous support throughout this project. I am also grateful to the School of Mathematics for providing me with the opportunity to gain valuable research experience under respectable Dr. Balocchi's supervision. This has been a truly rewarding and memorable experience, for which I am deeply thankful.

References

Balocchi, C., Deshpande, S. K., George, E. I., & Jensen, S. T. (2023). Crime in Philadelphia: Bayesian Clustering with Particle Optimization. *Journal of the American Statistical Association*, *118*(542), 818–829. https://doi.org/10.1080/01621459.2022.2156348

Balocchi, C., & Wade, S. (2025). *Understanding uncertainty in Bayesian cluster analysis* (arXiv:2506.16295). arXiv. https://doi.org/10.48550/arXiv.2506.16295

Binder, D. A. (1978). Bayesian Cluster Analysis. *Biometrika*, *65*(1), 31–38. https://doi.org/10.2307/2335273

Canas, G., & Rosasco, L. (2012). Learning Probability Measures with respect to Optimal Transport Metrics. *Advances in Neural Information Processing Systems*, *25*.

Corradin, R., Canale, A., & Nipoti, B. (2021). BNPmix: An R Package for Bayesian Nonparametric Modeling via Pitman-Yor Mixtures. *Journal of Statistical Software*, *100*, 1–33. https://doi.org/10.18637/jss.v100.i15

Dahl, D. B. (2006). *10 Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model*.

Dahl, D. B., Johnson, D. J., & Müller, P. (2022). Search Algorithms and Loss Functions for Bayesian Clustering. *Journal of Computational and Graphical Statistics*, *31*(4), 1189–1201. https://doi.org/10.1080/10618600.2022.2069779

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the*

Royal Statistical Society. Series C (Applied Statistics), *28*(1), 100–108. https://doi.org/10.2307/2346830

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, *64*, 53–62. https://doi.org/10.1016/j.patrec.2015.04.009

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Malsiner-Walli, G., Frühwirth-Schnatter, S., & Grün, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Stat Comput*, *26*(1-2), 303–324. https://doi.org/10.1007/s11222-014-9500-2

Miller, J. W., & Harrison, M. T. (2018). Mixture Models With a Prior on the Number of Components. *Journal of the American Statistical Association*.

Müller, P. (2019). Bayesian Nonparametric Mixture Models. In *Handbook of Mixture Analysis*. Chapman and Hall/CRC.

Nobile, A., & Fearnside, A. T. (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, *17*(2), 147–162. https://doi.org/10.1007/s11222-006-9014-7

On a Class of Bayesian Nonparametric Estimates: I. Density Estimates on JSTOR. (n.d.). https://www.jstor.org/stable/2241 Retrieved August 29, 2025, from https://www.jstor.org/stable/2241054

Pollard, D. (1982). Quantization and the method ofk-means. *IEEE Trans. Inform. Theory*, *28*(2), 199–205. https://doi.org/10.1109/TIT.1982.1056481

Richardson, S., & Green, P. J. (n.d.). *On Bayesian Analysis of Mixtures with an Unknown Number of Components*.

RIGON, T., HERRING, A. H., & DUNSON, D. B. (2023). A generalized Bayes framework for probabilistic clustering. *Biometrika*, *110*(3), 559–578. https://doi.org/10.1093/biomet/asad004

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

Vinh, N. X., Vinh, N. X., Epps, J., Epps, J., & Bailey, J. (n.d.). *Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance*.

Wade, S., & Ghahramani, Z. (2018). Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion). *Bayesian Analysis*, *13*(2), 559–626. https://doi.org/10.1214/17-BA1073