Computational Statistics and Data Analysis 202 (2025) 108053

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A Bayesian cluster validity index

Onthada Preedasawakul [1], Nathakhun Wiroonsri *,[2]

Mathematics and Statistics with Applications Research Group (MaSA), Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

ARTICLE INFO

Keywords:
Cluster analysis
CVI
Dirichlet
Fuzzy c-means
K-means
MRI

ABSTRACT

Selecting the appropriate number of clusters is a critical step in applying clustering algorithms. To assist in this process, various cluster validity indices (CVIs) have been developed. These indices are designed to identify the optimal number of clusters within a dataset. However, users may not always seek the absolute optimal number of clusters but rather a secondary option that better aligns with their specific applications. This realization has led us to introduce a Bayesian cluster validity index (BCVI), which builds upon existing indices. The BCVI utilizes either Dirichlet or generalized Dirichlet priors, resulting in the same posterior distribution. The proposed BCVI is evaluated using the Calinski-Harabasz, CVNN, Davies–Bouldin, silhouette, Starczewski, and Wiroonsri indices for hard clustering and the KWON2, Wiroonsri–Preedasawakul, and Xie–Beni indices for soft clustering as underlying indices. The performance of the proposed BCVI with that of the original underlying indices has been compared. The BCVI offers clear advantages in situations where user expertise is valuable, allowing users to specify their desired range for the final number of clusters. To illustrate this, experiments classified into three different scenarios are conducted. Additionally, the practical applicability of the proposed approach through real-world datasets, such as MRI brain tumor images are presented. These tools are published as a recent R package 'BayesCVI'.

1. Introduction

Cluster analysis is a well-known unsupervised learning tool in statistical and machine learning. It is used to split observations into groups with similar behaviors (refer to the book by James et al. (2023) for a review). Researchers apply cluster analysis to solve problems in various fields, ranging from social science to outer space. There are different types of clustering algorithms, including centroid-based clustering (such as K-means and fuzzy c-means (FCM)), hierarchical clustering (which includes single linkage, complete linkage, group average agglomerative, and Ward's criterion), density-based clustering (such as DBSCAN, DENCLUE, and OPTICS), probabilistic clustering (such as EM), grid-based clustering (such as CLIQUE, MAFIA, ENCLUS, and OptiGrid), and spectral clustering (see Aggarwal and Reddy (2014) for more detailed information on these techniques). Recently, there has been significant attention given to deep learning clustering (Min et al., 2018) and 3D point cloud clustering (Xie et al., 2020; Guo et al., 2021; Chen et al., 2023). Some examples of 3D point cloud techniques include PointNet, PointNet++, DGCNN, and RandLA-Net.

* Corresponding author at: 126 Pracha Uthit Rd., Bangmod, Thung Khru, Bangkok 10140, Thailand.
E-mail addresses: o.preedasawakul@gmail.com (O. Preedasawakul), nathakhun.wir@kmutt.ac.th (N. Wiroonsri).

[1] This author has been supported by Petchra Pra Jom Klao Master's Degree Research Scholarship from King Mongkut's University of Technology Thonburi, Grant number: 32/2567.

[2] This author has been financially supported by National Research Council of Thailand (NRCT), Grant number: N42A660991 (2023).

https://doi.org/10.1016/j.csda.2024.108053
Received 9 April 2024; Received in revised form 26 July 2024; Accepted 23 August 2024

Available online 30 August 2024

Before the actual clustering process, most clustering algorithms require a necessary step known as clustering tendency assessment. This pre-clustering process aims to determine the existence of clusters in a dataset and, if present, to identify an appropriate unknown number of clusters (refer to Vysala and Gomes (2020); Kumar and Bezdek (2020) for more details). A cluster validity index (CVI) is commonly employed to perform this task (Akhanli and Hennig, 2020)). CVIs discussed in this work are all internal approaches, that is, each index is defined based on only a clustering result. External approaches such as Rand index (Rand, 1971), and Adjusted Rand index (Hubert and Arabie, 1985) are usually used for evaluating clustering algorithms on labeled datasets which is not our propose of this work. Although there are various existing CVIs, no one dominates others and their performances vary on different types of datasets.

In this work, we introduce a new concept of CVI that can be applied to any clustering algorithm, along with compatible CVIs. However, we specifically focus on two classic hard and soft clustering algorithms: K-means and FCM, to illustrate our idea. While it is impractical to list all existing CVIs here, Caliński and Harabasz (1974) (CH), Liu et al. (2013) (CVNN), Davies and Bouldin (1979) (DB), Rousseeuw (1987); Kaufman and Rousseeuw (2009) (SH), and Starczewski (2017) (STR) are notable examples for hard clustering, and Kwon et al. (2021) (KWON2), Xie and Beni (1991) (XB) for soft clustering (refer to Arbelaitz et al. (2013) and references therein for more examples). Recently, Wiroonsri (2024) and Wiroonsri and Preedasawakul (2023a) introduced correlation-based CVIs called the Wiroonsri index (WI) and Wiroonsri–Preedasawakul index (WP), which can accurately detect the optimal number of clusters and provide information about suboptimal numbers, enabling users to rank several options. These two indices are compatible with hard and soft clustering methods, respectively. The key concept in both works is allowing users to select the final number of clusters at a local peak rather than the global one, based on their experience and application needs. This concept motivated us to integrate Bayesian principles to incorporate users' experience and knowledge with these indices and other existing ones.

Why experience matters: In domains like business and healthcare, researchers possess valuable insights to estimate the number of clusters they expect. For example, one may anticipate the number of customer segments or types of lung cancer to range from 3 to 6, or understand that 2 or 3 colors are sufficient for visualizing Magnetic Resonance Imaging (MRI) images. Traditional CVIs might detect an optimal number of clusters that fall outside this expected range. This complication leads us to propose a connection between the Bayesian framework and CVIs.

When experience matters: We introduce a new Bayesian cluster validity index (BCVI) based on existing CVIs. We offer the option to use either a Dirichlet prior or a generalized Dirichlet (GD) prior for determining the optimal number of cluster candidates. This allows users to set parameters according to their experience in specific contexts. The posterior distribution remains unchanged, but the parameters are adjusted based on the data, allowing the final optimal number of clusters to be determined by considering both the data and knowledge/experience.

Bayesian analysis, a statistical approach widely used in various fields including medical diagnosis, is particularly relevant. Bayesian deep learning, introduced by Abdullah et al. (2022), has found success in healthcare applications such as disease diagnostics, medical imaging, clinical signal processing, and electronic health records. Additionally, Schmid et al. (2006) presented a novel approach for estimating kinetic parameters in DCE-MRI using adaptive Gaussian Markov random fields. Bayesian analysis also holds significance in business and marketing contexts (see Chandukala et al. (2011); Lohrke et al. (2018); Bianchi and Heo (2021) for examples). The Bayesian concept has previously been used in cluster analysis, where Bayesian clustering is used to group data points based on similarities by defining a prior distribution on the unknown partition (see Grazian (2023) for a comprehensive review). One of the main strengths of Bayesian clustering is its ability to handle complex data structures and dependencies among clusters by integrating prior knowledge and assumptions into the modeling process. Furthermore, Bayesian methods can automatically determine the number of clusters by estimating the posterior distribution on the partition. Probabilistic CVIs have been studied for over two decades. Bezdek et al. (1997) discussed probabilistic CVIs for normal mixtures, identifying five broad types including likelihood-based criteria, information-based criteria, Bozdogan's entropic complexity criteria, minimum information ratios, and other miscellaneous indices (see Bezdek et al. (1997) and references therein for further details). Additionally, Lee et al. (2008) introduced a probabilistic CVI compatible with iterative Bayesian fuzzy clustering. To the best of our knowledge, there is currently a lack of literature directly discussing Bayesian cluster validity indices, which further motivates the development of our BCVI.

Our BCVI is defined based on the assumption that a ratio, which is derived from an underlying index as defined in Section 3, follows a multinomial distribution given $\mathbf{p}$. Here, $\mathbf{p}$ follows a prior Dirichlet or GD distribution ($\mathbf{p} = (p_1, p_2, \dots, p_K)$, representing the probability that the actual number of groups is k for $k = 1, 2, \dots, K$). This allows users to set the parameters of the prior distribution based on their knowledge and intentions. As a result, the final number of clusters typically falls within the expected range, which is more relevant for users. We assess the performance of the BCVI with the underlying CH, CVNN, DB, SH, STR, WI, KWON2, WP, and XB through K-means and FCM on simulated datasets categorized into three cases, as well as real-world datasets including MRI images. Our evaluation presents the performance of BCVI alongside the nine original underlying indices outlined in the subsequent section. Although this is not a direct comparison test due to the novelty of the concept, it serves to confirm the claimed benefits of BCVI when choosing appropriate underlying CVIs.

The remaining sections of this work are organized as follows. Section 2 offers necessary background information about clustering algorithms, CVIs, and probability distributions used in this study. Our proposed index, along with its mathematical properties, is presented in Section 3. Section 4 delves into experimental results and potential applications to MRI images. Finally, Section 5 provides concluding remarks and discusses potential future directions. The proofs of the mathematical properties stated in Section 3 are presented in the Appendix A for readers who are interested.

2. Background

In this section, we provide the necessary definitions and background information, including clustering algorithms, cluster validity indices, and related probability distributions.

2.1. Cluster analysis and cluster validity index

In this subsection, we offer brief definitions of K-means and FCM, as well as six hard and three soft existing CVIs which will serve as underlying indices for our proposed BCVI in the subsequent section.

Let $n, k, p \in \mathbb{N}$ and denote $[n] = \{1, 2, \dots, n\}$. We establish the following notations used in this work. For $i \in [n]$ and $j \in [k]$, denote

1. $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$: Data points.
2. K: Actual number of clusters.
3. C_j: Set of data points in the j^{th} cluster.
4. v_j: j^{th} cluster centroid.
5. v_0: Centroid of the entire dataset.
6. $\bar{v}$: Centroid of all v_j.
7. $\mu = (\mu_{ij})$: Membership degree matrix where μ_{ij} denotes the degree to which a sample point x_i belongs to C_j.
8. $\|x - y\|$: Euclidean distance between x and y.
9. $\mathrm{Corr}(\cdot, \cdot)$ Correlation coefficient. In this work, we consider only the Pearson correlation.

2.1.1. K-means

K-means (MacQueen, 1967) is a simple yet efficient clustering algorithm. It operates by partitioning a dataset into k clusters, where k is a user-defined parameter. The algorithm commences by randomly initializing cluster centroids. Subsequently, it assigns each data point to the nearest centroid and updates the centroids based on the newly assigned points. This iterative process continues until the cluster centroids converge. The objective of K-means is to minimize the within-cluster sum of squares, expressed as:

$$\sum_{j=1}^{k} \sum_{x \in C_j} \|x - v_j\|^2.$$

2.1.2. Fuzzy C-means

FCM, introduced by Dunn (1973) and later refined by Bezdek et al. (1984), is a clustering technique utilized to group similar data points into user-specified c clusters. Each data point is assigned a membership degree, indicating the degree of belongingness to each cluster. The objective of FCM is to minimize the target function:

$$\sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^{m} \|x_i - v_j\|^2, \tag{1}$$

where $m > 1$ denotes the fuzziness parameter.

The optimization of (1) begins with random initialization of centroids v_j. Iteratively, the membership degrees are updated according to:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}}},$$

and centroids are updated as follows:

$$v_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} x_i}{\sum_{i=1}^{n} \mu_{ij}^{m}},$$

for $i \in [n]$ and $j \in [c]$. This iterative process continues until convergence is achieved.

2.1.3. Hard cluster validity indices

2.1.3.1. Calinski-Harabasz index The CH index (Caliński and Harabasz, 1974) is defined as:

$$\mathrm{CH}(k) = \frac{n-k}{k-1} \frac{\sum_{j=1}^{k} |C_j| \left\| v_j - v_0 \right\|}{\sum_{j=1}^{k} \sum_{x \in C_j} \left\| x - v_j \right\|}.$$

The largest value of $\mathrm{CH}(k)$ indicates a valid optimal partition.

2.1.3.2. Clustering validation index based on nearest neighbors The CVNN index (Liu et al., 2013) is an internal validation measure defined as

$$\mathrm{CVNN}(k, NN) = \frac{\mathrm{Sep}(k, NN)}{\max_{Kmin\leqslant k\leqslant Kmax} \mathrm{Sep}(k, NN)} + \frac{\mathrm{Com}(k)}{\max_{Kmin\leqslant k\leqslant Kmax} \mathrm{Com}(k)},$$

where Sep and Com are given as follows:

1. Intercluster Separation:

$$\mathrm{Sep}(k, NN) = \max_{j\in[k]} \frac{1}{|C_j|} \sum_{x\in C_j} \frac{q_x}{NN},$$

where NN is the input number of nearest neighbors and q_x denotes the number of nearest neighbors of x that are outside its own cluster.
2. Intracluster compactness:

$$\mathrm{Com}(k) = \sum_{j=1}^{k} \left[2\frac{\sum_{x,y\in C_j} \|x-y\|}{|C_j|(|C_j|-1)} \right],$$

where x and y are two different objects in C_j.

The smallest value of $\mathtt{CVNN}(k)$ indicates a valid optimal partition.

2.1.3.3. Davies and Bouldin index The DB's measure (Davies and Bouldin, 1979) is defined as:

$$\mathtt{DB}(k) = \frac{1}{k} \sum_{i=1}^{k} R_{i,qt},$$

where

$$R_{i,qt} = \max_{j\in[k]\setminus\{i\}} \left\{ \frac{S_{i,q} + S_{j,q}}{M_{ij,t}} \right\},$$

$$S_{i,q} = \left(\frac{1}{|C_i|} \sum_{x\in C_i} \|x - v_i\|^q \right)^{1/q}$$

for $q, t \geqslant 1$ and $i \neq j \in [k]$, and

$$M_{ij,t} = \left(\sum_{s=1}^{p} |v_{is} - v_{js}|^t \right)^{1/t}.$$

The smallest value of $\mathtt{DB}(k)$ indicates a valid optimal partition.

2.1.3.4. Silhouette index For $i \in [n]$, $l \in [k]$, and $x_i \in C_l$, let

$$a(i) = \frac{1}{|C_l| - 1} \sum_{y\in C_l} \|x_i - y\| \text{ and } b(i) = \min_{r\neq l} \frac{1}{|C_r|} \sum_{y\in C_r} \|x_i - y\|.$$

The silhouette value of one data point x_j is defined as:

$$s(j) = \begin{cases} \dfrac{b(j) - a(j)}{\max\{a(j), b(i)\}} & \text{if } |C_j| > 1 \\ 0 & \text{if } |C_j| = 1 \end{cases}.$$

The silhouette index (Rousseeuw, 1987; Kaufman and Rousseeuw, 2009) is defined as

$$\mathrm{SH}(k) = \frac{1}{n} \sum_{i=1}^{n} s(i).$$

The largest value of $\mathtt{SH}(k)$ indicates a valid optimal partition.

2.1.3.5. Starczewski index The STR index (Starczewski, 2017) is defined as

$$\mathtt{STR}(k) = [E(k) - E(k-1)][D(k+1) - D(k)],$$

where $D(k) = \dfrac{\max_{i,j\in[k]} \|v_i - v_j\|}{\min_{i,j\in[k]} \|v_i - v_j\|}$, and $E(k) = \dfrac{\sum_{i=1}^{n} \|x_i - v_0\|}{\sum_{j=1}^{k} \sum_{x\in C_j} \|x - v_j\|}$.
The largest value of $\mathtt{STR}(k)$ indicates a valid optimal partition.

2.1.3.6. Wiroonsri index The WI index (Wiroonsri, 2024) is defined as follows: For $m \in \{2,3,\dots,n-1\}$ and $k \in \{2,3,\dots,m\}$,

Case 1: $\max_{2\leqslant l\leqslant m} \mathtt{NCI1}(k) < +\infty$

$$\mathtt{NCI}_m(k) = \begin{cases} \min_{2\leqslant l\leqslant m} \{\mathtt{NCI1}(l) | \mathtt{NCI1}(l) > -\infty\} & \text{if } \mathtt{NCI1}(k) = -\infty \\ \mathtt{NCI1}(k) & \text{otherwise,} \end{cases}$$

Case 2: $\max_{2\leqslant l\leqslant m} \mathtt{NCI1}(k) = +\infty$

$$\mathtt{NCI}_m(k) = \begin{cases} \min_{2\leqslant l\leqslant m} \{\mathtt{NCI1}(l) | \mathtt{NCI1}(l) > -\infty\} + \mathtt{NCI2}(k) & \text{if } \mathtt{NCI1}(k) = -\infty \\ \max_{2\leqslant l\leqslant m} \{\mathtt{NCI1}(l) | \mathtt{NCI1}(l) < +\infty\} + \mathtt{NCI2}(k) & \text{if } \mathtt{NCI1}(k) = +\infty \\ \mathtt{NCI1}(k) + \mathtt{NCI2}(k) & \text{otherwise,} \end{cases}$$

where

$$\mathtt{NCI1}(k) = \frac{(\mathtt{NC}(k) - \mathtt{NC}(k-1))(1 - \mathtt{NC}(k))}{\max\{0, \mathtt{NC}(k+1) - \mathtt{NC}(k)\}(1 - \mathtt{NC}(k-1))} \tag{2}$$

and

$$\mathtt{NCI2}(k) = \frac{\mathtt{NC}(k) - \mathtt{NC}(k-1)}{1 - \mathtt{NC}(k-1)} - \frac{\mathtt{NC}(k+1) - \mathtt{NC}(k)}{1 - \mathtt{NC}(k)}, \tag{3}$$

with $\mathtt{NC} = \mathrm{Corr}(\vec{d}, \vec{c}(k))$, $\mathtt{NC}(1) = \frac{\mathrm{SD}(\vec{d}_v)}{\max \vec{d}_v - \min \vec{d}_v}$. Note that we let

$$\vec{d}_v = (\|x_i - v_0\|)_{i\in[n]}, \tag{4}$$

$$\vec{d} = (\|x_i - x_j\|)_{i,j\in[n]} \tag{5}$$

be a vector of length $\binom{n}{2}$ containing distances of all pairs of data points, and

$$\vec{c}(k) = (\|v_i(k) - v_j(k)\|)_{i,j\in[n]}$$

be a vector of the same length containing the distances of all pairs of corresponding centroids of clusters in which two points are located. The largest value of $\mathtt{WI}(k)$ indicates a valid optimal partition.

2.1.4. Soft cluster validity indices

2.1.4.1. KWON2 index The KWON2 index (Kwon et al., 2021) is defined as follows:

$$\mathtt{KWON2}(k) = \frac{w_1 \left[w_2 \sum_{j=1}^{k} \sum_{i=1}^{n} \mu_{ij}^{2^{\sqrt{\frac{m}{2}}}} \|x_i - v_j\|^2 + \dfrac{\sum_{j=1}^{k} \|v_j - v_0\|^2}{\max_j \|v_j - v_0\|^2} + w_3 \right]}{\min_{i\neq j} \|v_i - v_j\|^2 + \frac{1}{k} + \frac{1}{k^{m-1}}}$$

where $w_1 = \dfrac{n-k+1}{n}$, $w_2 = \left(\dfrac{k}{k-1}\right)^{\sqrt{2}}$ and $w_3 = \dfrac{nk}{(n-k+1)^2}$.
The smallest value of $\mathtt{KWON2}(k)$ indicates a valid optimal partition.

2.1.4.2. Wiroonsri and Preedasawakul index The WP index (Wiroonsri and Preedasawakul, 2023a) is defined in three cases. Let $m \in \{2,3,\dots,n-1\}$ and $k \in \{2,3,\dots,m\}$,

Case 1: $\max_{2\leqslant l\leqslant p} \mathtt{WPCI1}(k) < +\infty$ and $\exists l \in [p]\backslash\{1\}$ such that $|\mathtt{WPCI1}(l)| < \infty$.

$$\mathtt{WP}_p(k) = \begin{cases} \min_{2\leqslant l\leqslant p} \{\mathtt{WPCI1}(l) | \mathtt{WPCI1}(l) > -\infty\} & \text{if } \mathtt{WPCI1}(k) = -\infty \\ \mathtt{WPCI1}(k) & \text{otherwise.} \end{cases}$$

Case 2: $\max_{2\leqslant l\leqslant p} \mathtt{WPCI1}(k) = +\infty$ and $\exists l \in \{2,3,\dots,p\}$ such that $|\mathtt{WPCI1}(l)| < \infty$.

$$\mathtt{WP}_p(k) = \begin{cases} \min_{2\leqslant l\leqslant p} \{\mathtt{WPCI1}(l) | \mathtt{WPCI1}(l) > -\infty\} + \mathtt{WPCI2}(k) & \text{if } \mathtt{WPCI1}(k) = -\infty \\ \max_{2\leqslant l\leqslant p} \{\mathtt{WPCI1}(l) | \mathtt{WPCI1}(l) < +\infty\} + \mathtt{WPCI2}(k) & \text{if } \mathtt{WPCI1}(k) = +\infty \\ \mathtt{WPCI1}(k) + \mathtt{WPCI2}(k) & \text{otherwise.} \end{cases}$$

Case 3: $\forall l \in \{2,3,\dots,p\}$, $|\texttt{WPCI1}(l)| = +\infty$.

$$\texttt{WP}_p(k) = \texttt{WPCI2}(k),$$

where $\texttt{WPCI1}(k)$ and $\texttt{WPCI2}(k)$ are defined similarly to (2) and (3), respectively, with NC replaced by WPC with $\texttt{WPC}(k) = \text{Corr}(\vec{d}, \vec{v}(k))$, $\texttt{WPC}(1) = \frac{\texttt{SD}(\vec{d}_v)}{\max \vec{d}_v - \min \vec{d}_v}$, $\vec{d}_v$ and $\vec{d}$ are as in (5) and (4), respectively,

$$o_i(k,\gamma) = \frac{\sum_{j=1}^{k} \mu_{ij}^{\gamma} v_j}{\sum_{j=1}^{k} \mu_{ij}^{\gamma}} \quad \text{and} \quad \vec{v}(k) = (\|o_i(k,\gamma) - o_j(k,\gamma)\|)_{i,j\in[n]}.$$

The largest value of $\texttt{WP(k)}$ indicates a valid optimal partition.

2.1.4.3. Xie and Beni index The XB index (Xie and Beni, 1991) is defined as follows:

$$\texttt{XB}(k) = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n} \mu_{ij}^2 \|x_i - v_j\|^2}{n \cdot \min_{j\neq l} \{\|v_j - v_l\|^2\}}.$$

The smallest value of $\texttt{XB}(k)$ indicates a valid optimal partition.

2.2. Dirichlet and generalized Dirichlet distributions

In this subsection, we state the definitions of the Dirichlet and GD distributions and their necessary properties.

2.2.1. Dirichlet distribution

The Dirichlet distribution (Olkin and Rubin, 1964) with parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_K)$ where $\alpha_k > 0$ has the probability density function given by

$$f(x_1,\dots,x_K|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1},$$

for $0 \leqslant x_k \leqslant 1$ for all k and $\sum_{k=1}^{K} x_k = 1$ where the multivariate beta function is defined as

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}.$$

The next lemma provides the mean and variance of the Dirichlet distribution.

Lemma 2.1. *Let $\mathbf{X} = (X_1,\dots,X_K)$ have a Dirichlet distribution with parameters $\alpha_1,\dots,\alpha_K$. Then*

$$\mathbb{E}[X_k] = \frac{\alpha_k}{\alpha_0}, \quad \text{and} \quad Var(X_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)},$$

where

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k. \tag{6}$$

2.2.2. Generalized Dirichlet distribution

The GD distribution (Wong, 1998) with parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_{K-1})$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_{K-1})$ where $\alpha_k, \beta_k > 0$ has the probability density function given by

$$f(x_1,\dots,x_{K-1}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{k=1}^{K-1} \frac{x_k^{\alpha_k - 1}(1 - x_1 - \cdots - x_k)^{\gamma_k}}{B(\alpha_k,\beta_k)}$$

for $0 \leqslant x_k \leqslant 1$, and $x_1 + x_2 + \cdots + x_{K-1} \leqslant 1$ where $\gamma_k = \beta_k - \alpha_{k+1} - \beta_{k+1}$ for $k \in [K-2]$ and $\gamma_{K-1} = \beta_{K-1} - 1$, and $B(\cdot,\cdot)$ is the beta function.

The next lemma is proved in Properties 1 and 2 in Wong (1998).

Lemma 2.2. *Let $s_1,\dots,s_{K-1} \in \mathbb{N}_0$, $\mathbf{X} = (X_1,\dots,X_{K-1})$ have a GD distribution with parameters $\alpha_1,\dots,\alpha_{K-1}$, $\beta_1,\dots,\beta_{K-1}$ and $\delta_k = \sum_{i=k+1}^{K-1} s_i$ for $k \in [K-1]$, the general moment function of $\mathbf{X}$ is given by*

$$\mathbb{E}[X_1^{s_1} X_2^{s_2} \dots X_{K-1}^{s_{K-1}}] = \prod_{k=1}^{K-1} \frac{\Gamma(\alpha_k+\beta_k)\Gamma(\alpha_k+s_k)\Gamma(\beta_k+\delta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)\Gamma(\alpha_k+\beta_k+s_k+\delta_k)}. \tag{7}$$

Additionally, the sth moment of $X_K := 1 - X_1 - \cdots - X_{K-1}$ is given by

$$\mathbb{E}[X_K^s] = \prod_{k=1}^{K-1} \frac{\Gamma(\alpha_k+\beta_k)\Gamma(\beta_k+s)}{\Gamma(\beta_k)\Gamma(\alpha_k+\beta_k+s)}. \tag{8}$$

3. Our proposed index

In this section, we state our proposed BCVI and discuss some theoretical properties in the three following subsections. Subsection 3.1 is devoted for prior and posterior distributions related to BCVI. The definition of BCVI is defined in Subsection 3.2. Then some mathematical properties of BCVI are discussed in Subsection 3.3. We begin this section by stating our assumption of the distribution of a dataset based on an underlying CVI.

Let $\mathbf{x} = (x_1, x_2, \dots, x_n)$ denote a dataset of size $n \in \mathbb{N}$. Let $K \in \mathbb{N}$ be the maximum number of clusters to be considered, and let $\mathbf{p} = (p_2, p_3, \dots, p_K)$, where p_k, $k = 2, 3, \dots, K$ represents the probability that the dataset consists of k groups. Let $r_k(\mathbf{x})$ be a ratio adjusted from a CVI defined as

$$r_k(\mathbf{x}) = \begin{cases} \dfrac{GI(k) - \min_j GI(j)}{\sum_{i=2}^{K}(GI(i) - \min_j GI(j))} & \text{for Condition A,} \\ \dfrac{\max_j GI(j) - GI(k)}{\sum_{i=2}^{K}(\max_j GI(j) - GI(i))} & \text{for Condition B,} \end{cases} \tag{9}$$

where GI represents an arbitrary CVI.

Condition A: The largest value of the GI indicates the optimal number of clusters.

Condition B: the smallest value of the GI indicates the optimal number of clusters. It is clear that $0 \leqslant r_k(\mathbf{x}) \leqslant \mathbf{1}$. Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)} \tag{10}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant for the probability density function.

3.1. Prior and posterior of $\mathbf{p}$

In this section, we explore two different options for the prior distribution of $\mathbf{p}$, namely Dirichlet and GD priors.

3.1.1. Dirichlet prior and posterior

Here, we assume that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \dots, \alpha_K)$ with the probability density function

$$\pi(\mathbf{p}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=2}^{K} p_k^{\alpha_k - 1}. \tag{11}$$

Theorem 3.1. *Let $K \in \mathbb{N}$ and $\mathbf{r}(\mathbf{x}) = (r_2(\mathbf{x}), \dots, r_K(\mathbf{x}))$, where $r_k(\mathbf{x})$ is defined as in (9). Assuming that $\mathbf{x}$ follows the distribution described in (10), the posterior distribution of $\mathbf{p}$ has the probability density function:*

$$\pi(\mathbf{p}|\mathbf{x}) = \frac{1}{B(\boldsymbol{\alpha} + n\mathbf{r}(\mathbf{x}))} \prod_{k=2}^{K} p_k^{\alpha_k + nr_k(\mathbf{x}) - 1}.$$

In particular, it follows a Dirichlet distribution with parameters $\boldsymbol{\alpha} + n\mathbf{r}(\mathbf{x})$.

Proof. See Appendix A.1. □

The next corollary follows directly from the above theorem and Lemma 2.1.

Corollary 3.1. *For $k = 2, 3, \dots, K$, the posterior means and variances of p_k are given, respectively, by*

$$\mathbb{E}[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n},$$

and

$$\mathrm{Var}(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)},$$

where α_0 *is as in* (6) *with* k *from* 2 *to* K.

3.1.2. Generalized Dirichlet prior and posterior

In this subsection, we consider a GD prior distribution for $\mathbf{p}$, characterized by parameters $\boldsymbol{\alpha} = (\alpha_2, \dots, \alpha_{K-1})$ and $\boldsymbol{\beta} = (\beta_2, \dots, \beta_{K-1})$, denoted as $\mathrm{GD}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. The probability density function of this prior distribution is given by

$$\pi(\mathbf{p}) = \prod_{k=2}^{K-1} \frac{p_k^{\alpha_k - 1}(1 - p_2 - \cdots - p_k)^{\gamma_k}}{B(\alpha_k, \beta_k)}, \tag{12}$$

where $\gamma_k = \beta_k - \alpha_{k+1} - \beta_{k+1}$ for $k = 2, 3 \dots, K-2$ and $\gamma_{K-1} = \beta_{K-1} - 1$.

Theorem 3.2. *Let* $K \in \mathbb{N}$ *and* $\mathbf{r}(\mathbf{x}) = (r_2(\mathbf{x}), \dots, r_K(\mathbf{x}))$ *where* $r_k(\mathbf{x})$ *is as in* (9). *Assume that* $\mathbf{x}$ *has a distribution following* (10) *and* $\mathbf{p}$ *has a prior distribution following* (12). *Then the posterior distribution of* $\mathbf{p}$ *has the probability density function given by*

$$\pi(\mathbf{p}|\mathbf{x}) = \prod_{k=2}^{K-1} \frac{p_k^{\alpha'_k - 1}(1 - p_2 - \cdots - p_k)^{\gamma'_k}}{B(\alpha'_k, \beta'_k)},$$

where

$$\alpha'_k = \alpha_k + nr_k(x) \tag{13}$$

and

$$\beta'_k = \beta_k + \sum_{i=k+1}^{K} nr_i(x) \tag{14}$$

for $k = 2, \dots, K-1$, $\gamma'_k = \beta'_k - \alpha'_{k+1} - \beta'_{k+1}$ *for* $k = 2, 3 \dots, K-2$ *and* $\gamma'_{K-1} = \beta'_{K-1} - 1$.

In particular, it has a GD distribution with parameters $\boldsymbol{\alpha}' = (\alpha'_2, \dots, \alpha'_{K-1})$ *and* $\boldsymbol{\beta}' = (\beta'_2, \dots, \beta'_{K-1})$.

Proof. See Appendix A.2. □

Next, we compute the posterior means and variances of $\mathbf{p}$.

Corollary 3.2. *For* $k = 2, 3, \dots, K-1$, *the posterior means of* p_i *are given by*

$$\mathbb{E}[p_k|\mathbf{x}] = \frac{\alpha'_k}{\alpha'_k + \beta'_k} \prod_{i:i<k} \frac{\beta'_i}{\alpha'_i + \beta'_i} \quad \text{and} \quad \mathbb{E}[p_K|\mathbf{x}] = \prod_{k=2}^{K-1} \frac{\beta'_k}{\alpha'_k + \beta'_k},$$

and the variances are given by

$$\mathrm{Var}(p_k|\mathbf{x}) = \frac{(\alpha'_k + 1)\alpha'_k}{(\alpha'_k + \beta'_k + 1)(\alpha'_k + \beta'_k)} \prod_{i:i<k} \frac{(\beta'_i + 1)\beta'_i}{(\alpha'_i + \beta'_i + 1)(\alpha'_i + \beta'_i)} - \frac{\alpha'^2_k}{(\alpha'_k + \beta'_k)^2} \prod_{i:i<k} \frac{\beta'^2_i}{(\alpha'_i + \beta'_i)^2},$$

and

$$\mathrm{Var}(p_K|\mathbf{x}) = \prod_{k=2}^{K-1} \frac{(\beta'_k + 1)\beta'_k}{(\alpha'_k + \beta'_k + 1)(\alpha'_k + \beta'_k)} - \prod_{k=2}^{K-1} \frac{\beta'^2_k}{(\alpha'_k + \beta'_k)^2},$$

where α'_k *and* β'_k, *are given in* (13) *and* (14), *respectively.*

Proof. See Appendix A.3. □

3.2. Definition

In this subsection, we present the definition of our proposed BCVI. For a data point $\mathbf{x} = (x_1, x_2, \dots, x_n)$, let $f(\mathbf{x}|\mathbf{p})$ denote the conditional probability density function given $\mathbf{p}$, as specified in (10), where p_k represents the probability that the actual number of clusters is k. We further assume that the prior distribution of $\mathbf{p}$ follows either a Dirichlet or GD distribution, as described in the previous section. The BCVI is then defined as follows.

Definition 3.3. For $k = 2, 3, \dots, K$,

$$\texttt{BCVI}(k) = \mathbb{E}[p_k|\mathbf{x}] \tag{15}$$

where $\mathbb{E}[p_k|\mathbf{x}]$ is computed according to either Corollary 3.1 or Corollary 3.2.

Though any index may be taken as an underlying CVI in (9), in this work, we focus on testing our proposed BCVI based solely on the six CVIs mentioned in the previous section.

Remark 3.4.

1. Since p_k represents the probability that the actual number of clusters is k, we can use these probabilities to construct a confidence set. For instance, if $p_{i_1} + p_{i_2} + p_{i_3}$, we are confident that the actual number of clusters lies within the set $\{i_1, i_2, i_3\}$.
2. BCVI is only meaningful when the underlying CVI can provide a ranking of the optimal numbers of clusters. If the underlying CVI only indicates the best option, p_k is not a valid probability.

3.3. Mathematical properties

In this subsection, we consider some properties of our BCVI. It is reasonable to assume that α_k is of the same order with respect to n for all k.

3.3.1. Dirichlet prior

By (15) and Corollary 3.1, for $k = 2, 3, \dots, K$, the BCVI is given by

$$\texttt{BCVI}(k) = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}. \tag{16}$$

The following proposition analyzes the behavior of BCVI when n is large according to the order of α_k in each situation.

Proposition 3.5. *For $k = 2, 3, \dots, K$, the following holds.*

1. *If $\max_k \alpha_k = o(n)$ as $n \to \infty$, then*

$$\lim_{n\to\infty} \mathit{BCVI}(k) = r_k(\mathbf{x}).$$

2. *If $\alpha_k = O(n)$, i.e. $\alpha_k/n \to c_k$ as $n \to \infty$ for some $c_k > 0$ for all k, then*

$$\lim_{n\to\infty} \mathit{BCVI}(k) = \frac{c_k + r_k(\mathbf{x})}{\sum_{j=2}^{K} c_j + 1}.$$

3. *If $\alpha_k = O(g(n))$, $g(n)/n \to \infty$, and $\alpha_k/g(n) \to c_k$ as $n \to \infty$, then*

$$\lim_{n\to\infty} \mathit{BCVI}(k) = \frac{c_k}{\sum_{j=2}^{K} c_j}.$$

Proof. See Appendix A.4. □

Remark 3.6. Proposition 3.5 implies that if α_k is of a larger order than n, then the underlying index has no impact on BCVI when n is large. Similarly, if α_k is of the same order as n, then the underlying index has the same impact on BCVI regardless of the sample size. Therefore, selecting $\alpha_k = o(n)$ such that $\alpha_k \to \infty$ as $n \to \infty$, such as $\alpha_k = O(\sqrt{n})$, makes the most sense. However, in cases where users have strong beliefs about the expected number of clusters, $\alpha_k = O(n)$ may also be chosen.

The next two propositions provide the properties of BCVI for some specific $\boldsymbol{\alpha}$.

Proposition 3.7. *If α_k, $k = 2, 3, \dots, K$ are all equal, then BCVI and the underlying CVI, i.e., GI, are equivalent.*

Proof. See Appendix A.5. □

Proposition 3.8. *Let $s \in \mathbb{N}$ and $i_1, i_2, \dots, i_s \in \{2, 3, \dots, K\}$ be local peaks of GI(k). If for any $j \in [s]$, $\alpha_{i_j-1}, \alpha_{i_j}, \alpha_{i_j+1}$ are equal, then $i_1, i_2, \dots, i_s$ remain local peaks with respect to BCVI(k).*

Proof. See Appendix A.6. □

3.3.2. Generalized Dirichlet prior

By (15) and Corollary 3.2, the BCVI is

$$\texttt{BCVI}(k) = \frac{\alpha_k + nr_k(x)}{\alpha_k + \beta_k + \sum_{i=k}^{K} nr_i(x)} \prod_{i:i<k} \frac{\beta_i + \sum_{j=i+1}^{K} nr_j(x)}{\alpha_i + \beta_i + \sum_{j=i}^{K} nr_j(x)} \tag{17}$$

for $k = 2, 3, \dots, K-1$, and

$$\texttt{BCVI}(K) = \prod_{k=2}^{K-1} \frac{\beta_k + \sum_{i=k+1}^{K} nr_i(x)}{\alpha_k + \beta_k + \sum_{i=k}^{K} nr_i(x)}. \tag{18}$$

The following proposition examines the behavior of BCVI when n is large relative to the orders of α_k and β_k in each scenario.

Proposition 3.9.

1. *If* $\max_k \alpha_k = o(n)$ *and* $\max_k \beta_k = o(n)$ *as* $n \to \infty$, *then for* $k = 2, 3, \dots, K$,

$$\lim_{n\to\infty} \texttt{BCVI}(k) = r_k(x).$$

2. *If* $\alpha_k = O(n)$ *and* $\beta_k = O(n)$, *i.e.* $\alpha_k/n \to c_k$ *and* $\beta_k/n \to d_k$ *as* $n \to \infty$ *for some* $c_k, d_k > 0$ *for all* k, *then for* $k = 2, 3, \dots, K-1$,

$$\lim_{n\to\infty} \texttt{BCVI}(k) = \frac{c_k + r_k(x)}{c_k + d_k + \sum_{i=k}^{K} r_i(x)} \prod_{i:i<k} \frac{d_i + \sum_{j=i+1}^{K} r_j(x)}{c_i + d_i + \sum_{j=i}^{K} r_j(x)},$$

and

$$\lim_{n\to\infty} \texttt{BCVI}(K) = \prod_{k=2}^{K-1} \frac{d_k + \sum_{i=k+1}^{K} r_i(x)}{c_k + d_k + \sum_{i=k}^{K} r_i(x)}.$$

3. *If* $\alpha_k = O(g(n))$ *and* $\beta_k = O(g(n))$, $g(n)/n \to \infty$, *and* $\alpha_k/g(n) \to c_k$ *and* $\beta_k/g(n) \to d_k$ *as* $n \to \infty$, *then for* $k = 2, 3, \dots, K-1$,

$$\lim_{n\to\infty} \texttt{BCVI}(k) = \frac{c_k}{c_k + d_k} \prod_{i:i<k} \frac{d_i}{c_i + d_i},$$

and

$$\lim_{n\to\infty} \texttt{BCVI}(K) = \prod_{k=2}^{K-1} \frac{d_k}{c_k + d_k}.$$

Proof. See Appendix A.7. □

It is evident that we can interpret the above proposition similarly to Remark 3.6 regarding the Dirichlet case.

4. Experimental results

In this section, we demonstrate the benefits of our proposed BCVI by comparing it to the underlying indices as listed in Section 2. Our experiment is divided into two distinctive parts, each described in a separate subsection: artificial datasets and real-world datasets, including MRI datasets. We test our BCVI using two clustering algorithms: K-means and FCM with a fuzziness parameter of 2 for hard and soft clustering, respectively. For the results presented in this entire section, we run either K-means or FCM for a total of 20 rounds and select the one with the smallest objective function. This is to make sure that the result is appropriate for testing the CVIs. To facilitate our experiment, we utilize the CRAN packages called "factoextra" (Kassambara and Mundt, 2020), "fpc" (Hennig, 2024), and our recent packages called "UniversalCVI" (Wiroonsri and Preedasawakul, 2023b), and "BayesCVI" (Preedasawakul and Wiroonsri, 2024) within the RStudio environment (RStudio Team, 2020).

The artificial datasets (K1-1 to K3-2, C1-1 to C3-2) are newly generated from the Gaussian distribution and the uniform distribution, except for K1-2 and C1-2, which are from Barton (2019). The datasets are intentionally named K and C to correspond with K-means and FCM, respectively. The real-world datasets include BWIS, DERM, and SEED from UCI (Lichman, 2013), and MRI brain tumor datasets (TUMOR1 from Nickparvar (2022), and the remainder from Chakrabarty (2019)). Since the WP underlying index is not directly applicable to big data, we resize the MRI images to 85×85 pixels before applying the BCVI while implementing the other existing CVIs to the full images. The selection of the main clustering algorithm on each dataset is based on accuracy when setting the number of clusters to be the actual number of groups. To ensure that the datasets are appropriate for fairly testing and comparing the performance of CVIs, we first verify that the main clustering algorithms are applicable to them. We check the accuracy of K-means and FCM on those labeled datasets, as shown in Table 1, using the 'AccClust' function in the package 'UniversalCVI'. Note that we

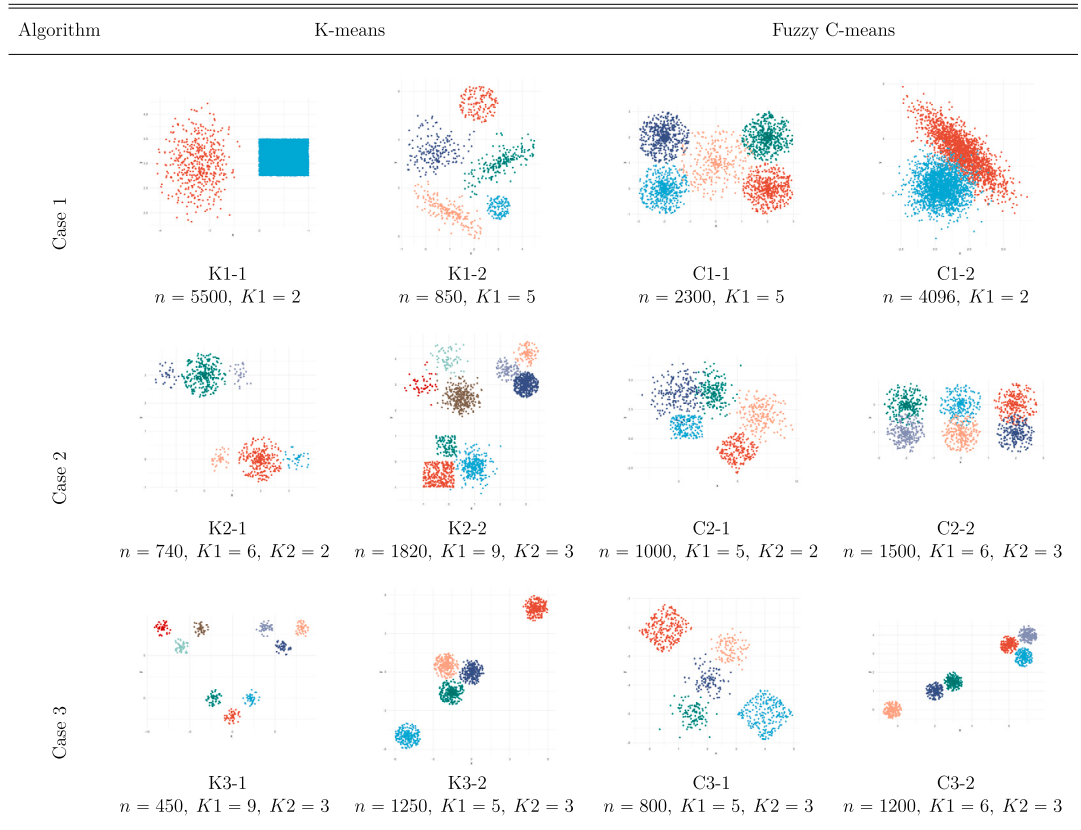
Table 1
Accuracy of clustering algorithms on labeled datasets.

Artificial Datasets				Real-world datasets			
Kmeans		Fuzzy C-means		Kmeans		Fuzzy C-means	
Data	Acc	Data	Acc	Data	Acc	Data	Acc
K1-1	0.9993	C1-1	0.9583	BWIS	0.9104		
K1-2	0.9471	C1-2	0.9531				
K2-1	0.8878	C2-1	0.9460	DERM	0.7514	SEED	0.9190
K2-2	0.9637	C2-2	0.9413				
K3-1	1.0000	C3-1	0.9988	SEED	0.9190		
K3-2	1.0000	C3-2	0.9917				

Table 2
Dirichlet prior parameters.

Data Type	$\boldsymbol{\alpha}$	Kmax	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Artificial	α_1	10	20	20	20	5	5	5	0.5	0.5	0.5
	α_2		0.5	0.5	0.5	5	5	5	20	20	20
	α_3		5	5	5	20	20	20	0.5	0.5	0.5
Real-world	α_1	10	25	25	2	2	0.5	0.5	0.5	0.5	0.5
	α_2		0.5	0.5	2	2	25	25	25	25	25
	α_3		2	2	25	25	0.5	0.5	0.5	0.5	0.5
MRI	α_1	8	25	25	2	2	0.5	0.5	0.5	-	-
	α_2		0.5	0.5	2	2	25	25	25	-	-
	α_3		2	2	25	25	0.5	0.5	0.5	-	-

Note: The parameters shown in this table will be multiplied by $\sqrt{n}$ where n is the number of data points.



Note: n, $K1$, and $K2$ are the number of data points, the first and secondary options for the number of clusters, respectively.

Fig. 1. Artificial datasets. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 3
Hard BCVI on artificial datasets.

Data	α	K1	K2	BCVI(CH)			Prob	CH			BCVI(CVNN)			Prob	CVNN			BCVI(DB)			Prob	DB		
				1	2	3		1	2	3	1	2	3		1	2	3	1	2	3		1	2	3
	α_1			4	3	2	50.07				3	2	4	63.20				4	3	2	55.90			
K1-1	α_2	2	-	10	9	8	62.65	10	9	8	9	10	8	43.58	3	7	5	9	8	10	60.76	9	8	4
	α_3			7	6	5	55.98				7	5	6	61.92				6	7	5	52.04			
	α_1			4	3	2	60.69				4	2	3	69.32				4	3	2	64.12			
K1-2	α_2	5	-	10	8	9	69.04	10	8	7	8	10	9	62.77	4	**5**	8	9	10	8	65.75	4	6	**5**
	α_3			7	5	6	68.23				**5**	6	7	65.86				6	5	7	68.10			
	α_1			2	3	4	66.04				2	3	4	77.14				2	3	4	74.22			
K2-1	α_2	6	2	10	9	8	68.79	2	10	9	9	10	8	60.14	2	3	4	9	10	8	58.69	2	3	**6**
	α_3			**6**	7	5	64.98				5	6	7	62.52				**6**	5	7	66.90			
	α_1			3	4	2	58.69				3	4	2	68.20				3	2	4	70.67			
K2-2	α_2	9	3	10	9	8	65.12	10	**9**	8	8	10	9	51.58	3	4	6	**9**	10	8	61.90	3	2	4
	α_3			7	5	6	63.04				6	7	5	67.08				5	7	6	54.28			
	α_1			3	4	2	64.22				4	3	2	67.02				3	4	2	67.97			
K3-1	α_2	9	3	**9**	10	8	76.02	**9**	10	8	**9**	8	10	68.46	**9**	8	7	**9**	8	10	71.47	**9**	3	8
	α_3			7	6	5	65.68				7	6	5	70.44				7	5	6	66.49			
	α_1			4	3	2	58.76				4	3	2	64.85				3	2	4	71.00			
K3-2	α_2	5	3	8	9	10	65.12	**5**	6	7	10	9	8	63.37	**5**	4	3	8	9	10	55.85	3	2	**5**
	α_3			**5**	6	7	68.65				**5**	6	7	64.31				**5**	6	7	65.68			

Data	α	K1	K2	BCVI(SH)			Prob	SH			BCVI(STR)			Prob	STR			BCVI(WI)			Prob	WI		
				1	2	3		1	2	3	1	2	3		1	2	3	1	2	3		1	2	3
	α_1			3	2	4	68.98				**2**	3	4	66.25				**2**	5	4	72.76			
K1-1	α_2	2	-	8	3	9	58.55	3	8	7	8	10	9	53.40	8	**2**	3	5	2	8	61.18	5	**2**	8
	α_3			3	7	6	56.69				6	7	5	49.04				5	2	7	72.11			
	α_1			4	3	2	62.38				4	3	2	66.65				2	4	3	66.23			
K1-2	α_2	5	-	8	10	9	65.78	**5**	4	6	10	9	8	62.63	7	**5**	4	8	10	9	66.85	8	2	**5**
	α_3			**5**	6	7	69.80				**7**	5	6	68.68				**5**	6	7	64.88			
	α_1			2	4	3	72.50				2	3	4	70.40				2	3	4	73.25			
K2-1	α_2	6	2	10	9	8	61.14	2	4	3	8	9	10	59.97	2	**6**	8	8	9	10	63.30	2	8	**6**
	α_3			7	5	6	66.16				**6**	5	7	69.43				**6**	5	7	63.25			
	α_1			3	2	4	72.53				3	2	4	62.37				3	2	4	63.88			
K2-2	α_2	9	3	8	9	10	53.04	3	2	4	10	9	8	67.60	10	3	**9**	**9**	8	10	60.50	3	5	**9**
	α_3			5	7	6	61.29				5	7	6	56.88				5	7	6	62.48			
	α_1			3	4	2	68.22				3	4	9	63.00				3	4	2	65.89			
K3-1	α_2	9	3	8	9	10	69.46	3	7	8	**9**	10	8	81.42	**9**	3	6	**9**	8	10	77.84	**9**	3	4
	α_3			7	5	6	68.24				6	5	9	61.54				7	6	5	62.19			
	α_1			2	4	3	66.94				**5**	3	2	69.44				3	4	2	57.54			
K3-2	α_2	5	3	9	10	8	56.40	**5**	2	7	**5**	10	8	63.13	**5**	3	2	10	8	9	68.38	**5**	10	3
	α_3			**5**	7	6	69.19				**5**	7	6	76.56				**5**	6	7	66.61			

Note: K1 and K2 (if exists) are the optimal and the second optimal numbers of clusters, respective. The table displays the first three ranks for the final number of clusters according to each CVI and the posterior probability of the first three ranks based on BCVI. For BCVI, the results are bold for K1 only if they appear at the first rank. For underlying CVIs, the results are bold for K1 and underlined for K2. The $\boldsymbol{\alpha}$ values in case 1,2, and 3 are given in Table 2 (these also apply to all the tables below).

intend to consider only the datasets with 75% accuracy or above. This criterion is set because if the main clustering algorithm is unable to find the correct groups, then it is not suitable for testing the performance of CVIs.

As previously mentioned, our experiment for the proposed BCVI is conducted using only the Dirichlet prior. Therefore, we must first set its parameters according to (11). The BCVI results presented in this section are computed based on the parameters outlined in Table 2. It is important to note that for each dataset, we propose three parameter options corresponding to three different scenarios: when the user prefers a small ($\boldsymbol{\alpha}_1$), large ($\boldsymbol{\alpha}_2$), and moderate ($\boldsymbol{\alpha}_3$) number of groups, respectively. This approach is advantageous for users with an approximate idea of how many clusters they expect. To determine the parameter values, we refer to Proposition 3.5 and Remark 3.6 that $\boldsymbol{\alpha}$ of order $O(n^s)$ where $0 < s < 1$ is the most appropriate. In addition, it is not suitable to set the multiplier of n^s for each k to be comparable to n.

4.1. Artificial datasets

In this subsection, we assess the performance of our BCVI on artificial datasets categorized into three distinct cases, as outlined below, in order to highlight the advantages of our proposed Bayesian approach. These datasets comprise both benchmark datasets and simulated datasets, as mentioned previously.

Table 4
Soft BCVI on artificial datasets.

Data	α	K1	K2	BCVI(KWON2)			Prob	KWON2			BCVI(WP)			Prob	WP			BCVI(XB)			Prob	XB		
				1	2	3		1	2	3	1	2	3		1	2	3	1	2	3		1	2	3
	α_1			4	3	2	58.39				2	4	3	66.44				4	3	2	67.87			
C1-1	α_2	5	-	10	9	8	61.84	4	**5**	6	9	8	10	56.81	2	4	**5**	10	8	9	50.67	4	**5**	3
	α_3			**5**	6	7	62.93				**5**	7	6	59.91				**5**	6	7	64.63			
	α_1			3	4	2	52.00				**2**	3	4	58.15				3	2	4	54.06			
C1-2	α_2	2	-	10	8	9	62.03	10	8	9	9	10	8	55.86	9	**2**	5	10	8	9	62.39	3	10	7
	α_3			7	6	5	59.63				5	7	6	59.65				7	6	5	57.21			
	α_1			4	3	2	63.00				2	4	3	68.09				4	2	3	72.50			
C2-1	α_2	5	2	10	9	8	66.04	4	**5**	10	9	10	8	61.97	2	**5**	9	10	9	8	59.54	4	**5**	2
	α_3			**5**	6	7	66.68				**5**	7	6	65.66				**5**	6	7	63.69			
	α_1			4	3	2	59.93				2	4	3	65.26				4	3	2	60.49			
C2-2	α_2	6	3	9	8	10	65.24	9	8	7	9	10	8	59.89	2	4	**6**	9	8	10	65.03	9	4	8
	α_3			7	6	5	64.64				**6**	5	7	64.67				**6**	7	5	64.30			
	α_1			4	3	2	66.37				3	4	2	68.09				4	2	3	74.21			
C3-1	α_2	5	3	9	8	10	66.91	4	**5**	9	9	8	10	66.44	**5**	3	9	9	8	10	62.16	4	2	3
	α_3			**5**	7	6	65.50				**5**	6	7	64.25				**5**	7	6	62.41			
	α_1			3	4	2	58.96				3	2	4	62.79				2	3	4	72.22			
C3-2	α_2	6	3	10	9	8	65.98	**6**	7	10	8	9	10	60.63	**6**	3	8	10	8	9	54.75	2	3	**6**
	α_3			**6**	7	5	68.19				**6**	5	7	69.71				**6**	5	7	66.15			

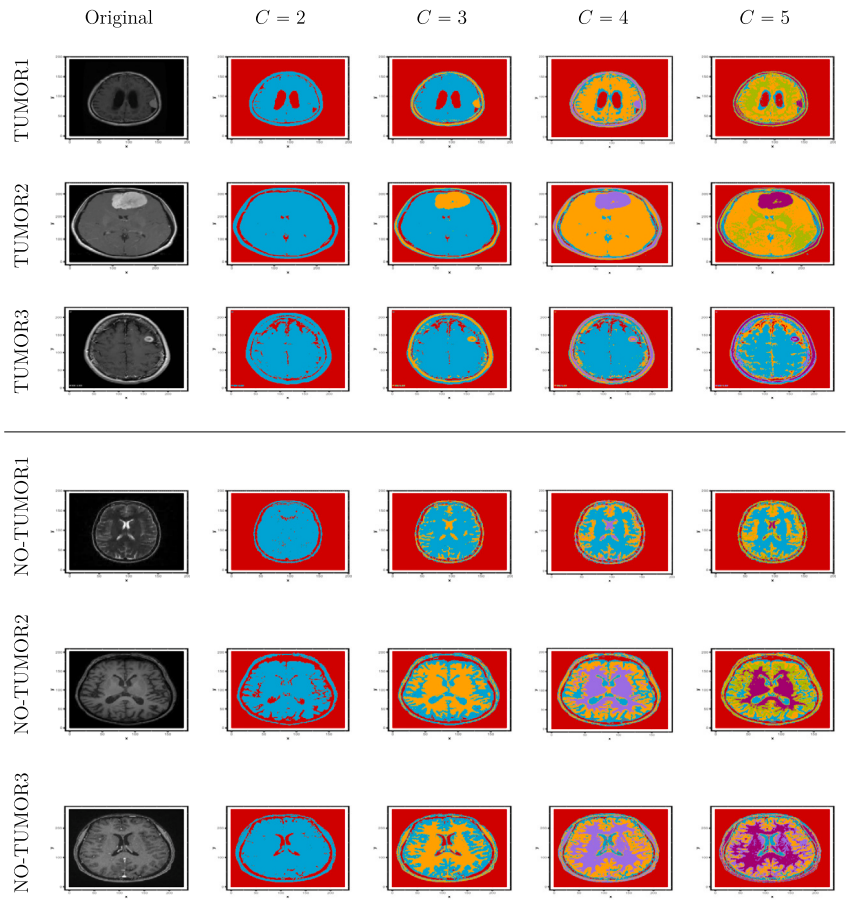


Fig. 2. MRI Datasets.

Table 5
Hard BCVI on real-world datasets.

Data	α	K	BCVI(CH)			Prob	CH			BCVI(CVNN)			Prob	CVNN			BCVI(DB)			Prob	DB		
			1	2	3		1	2	3	1	2	3		1	2	3	1	2	3		1	2	3
BWIS	α_1	2	**2**	3	4	87.03				**2**	3	5	81.89				**2**	3	4	80.28			
	α_2		6	7	8	50.92	**2**	3	4	7	6	10	52.60	**2**	3	5	10	9	8	54.83	**2**	10	4
	α_3		4	5	2	83.14				5	4	2	81.21				4	5	2	71.13			
DERM	α_1	6	3	2	4	84.11				3	2	5	82.73				2	3	4	83.07			
	α_2		**6**	7	8	53.12	3	4	2	7	6	8	53.38	5	4	3	**6**	9	8	52.9	5	2	3
	α_3		4	5	3	83.59				5	4	3	83.59				4	5	2	77.48			
SEED	α_1	3	2	3	4	88.39				**3**	2	5	86.23				2	3	4	87.32			
	α_2		6	7	8	53.35	2	**3**	4	6	9	10	54.80	**3**	2	6	6	9	7	54.10	2	**3**	4
	α_3		4	5	2	84.78				5	4	3	80.36				4	5	2	76.40			

Data	α	K	BCVI(SH)			Prob	SH			BCVI(STR)			Prob	STR			BCVI(WI)			Prob	WI		
			1	2	3		1	2	3	1	2	3		1	2	3	1	2	3		1	2	3
BWIS	α_1	2	**2**	3	4	89.22				**2**	3	8	82.32				**2**	3	10	76.98			
	α_2		6	10	7	50.15	**2**	3	4	8	10	9	54.85	8	**2**	4	10	6	9	62.61	10	6	4
	α_3		4	5	2	80.92				4	5	8	76.54				4	5	10	75.77			
DERM	α_1	6	3	2	4	81.08				2	3	5	87.35				2	3	4	78.17			
	α_2		**6**	7	9	54.36	4	3	5	9	6	7	51.73	5	2	3	**6**	10	8	57.84	4	**6**	10
	α_3		4	5	3	82.48				5	4	2	80.20				4	5	6	77.91			
SEED	α_1	3	2	3	4	88.38				2	3	4	87.42				**3**	2	8	84.40			
	α_2		8	6	7	53.18	2	**3**	4	9	6	7	54.31	2	**3**	4	8	10	7	58.55	8	10	**3**
	α_3		4	5	2	85.62				4	5	2	77.08				4	5	8	80.45			

Case 1: The underlying index incorrectly detects the true number of groups
Case 2: The underlying index originally leads to one of the secondary options
Case 3: The underlying index correctly detects the true number of groups, but users seek a secondary option because the optimal one is either too small or too large

The 12 datasets are depicted in Fig. 1: six for hard and soft CVIs each, categorized into the three previously mentioned cases. The parameters are configured for hypothetical scenarios based on users' preferences: two to four groups ($\boldsymbol{\alpha}_1$), eight to ten groups ($\boldsymbol{\alpha}_2$), and five to seven groups ($\boldsymbol{\alpha}_3$). Table 3, and 4 display the first three ranks for the final number of clusters according to each CVI, where the true optimal and second optimal options are bold and underlined, respectively. The parameters for BCVI are determined according to Table 2, with the rationale explained thereafter. From this point, we denote BCVI(GI) for BCVI with any underlying index GI. From Table 3, and 4, for the hard and soft clustering results, respectively, in cases 1 and 2, all the underlying indices incorrectly detect the number of clusters except for CVNN on K1-2. However, BCVI(WI) and BCVI(WP) provide the correct decision for all cases when the true number of clusters falls into the correct parameter scenario. On the other hand, BCVI(CH), BCVI(CVNN), BCVI(DB), BCVI(SH), BCVI(STR), BCVI(KWON2), and BCVI(XB) still provide incorrect decisions in some cases. This implies that the WI and WP exhibit a local peak at the true number of clusters as pointed out in Wiroonsri (2024) and Wiroonsri and Preedasawakul (2023a). In case 3, we illustrate when the underlying index can accurately detect the number of clusters, but that number is either too large or too small to implement. By adjusting the parameters to align with the requirements, only BCVI(DB), BCVI(WI) and BCVI(WP) guide the final decision toward a secondary choice in all cases. Additionally, we calculate the posterior probability of the first three ranks, as shown in Table 3 and 4. It indicates that we can be approximately 60% to 80% confident that the actual number of clusters falls within the first three ranks.

4.2. Real-world and MRI datasets

In this subsection, we apply our BCVI to real-world datasets: BWIS, DERM, and SEED (Table 5), along with MRI brain tumor images shown in Fig. 2. For hard clustering, we consider CH, CVNN, DB, SH, STR, and WI indices, while for soft clustering, we consider KWON2, WP, and XB indices. We intentionally select datasets where the underlying indices incorrectly detect the true numbers of classes. This is to underscore that our BCVI can rectify incorrect estimations when users possess prior knowledge about their expected number of classes. Table 5 and 6 display the first three ranks for the final number of clusters based on each CVI, where the true optimal option is bolded. The parameters for BCVI are determined according to Table 2, with the rationale explained thereafter.

First, we discuss the results for the real-world datasets, except MRI images. BCVI(WI) and BCVI(WP) have the capability to correct erroneous underlying CVI decisions when the parameters are appropriately configured. Again, the BCVI with other underlying CVIs correct parts but not all the cases. However, it is worth noting that CH, CVNN, SH, and DB correctly detect the true number of classes for some datasets prior to applying BCVI. Additionally, we can have approximately 80% confidence that the true number of clusters falls within the first three ranks of BCVI when selecting an appropriate $\boldsymbol{\alpha}$.

Table 6
Soft BCVI on real-world and MRI datasets.

Data	α	K	BCVI(KWON2)			Prob	KWON2			BCVI(WP)			Prob	WP			BCVI(XB)			Prob	XB		
			1	2	3		1	2	3	1	2	3		1	2	3	1	2	3		1	2	3
	α_1		**3**	2	4	86.18				**3**	2	7	81.30				2	3	4	88.32			
SEED	α_2	3	8	7	10	53.54	4	**3**	2	7	10	9	61.64	7	10	5	8	7	9	52.81	2	**3**	4
	α_3		4	5	3	78.43				5	4	7	78.50				4	5	2	79.71			
	α_1		**3**	2	5	55.75				**3**	2	4	66.85				**3**	2	4	81.45			
TUMOR1	α_2	3	7	8	6	72.70	7	8	6	8	6	7	71.56	8	4	6	6	7	8	59.13	**3**	2	4
	α_3		5	4	7	66.90				4	5	8	61.65				4	3	5	65.25			
	α_1		**3**	2	5	56.65				**3**	2	5	75.19				**3**	2	5	76.74			
TUMOR2	α_2	3	8	7	6	71.93	8	7	5	8	5	7	61.46	5	**3**	8	8	7	6	51.81	**3**	5	4
	α_3		5	4	8	67.04				5	4	3	76.21				5	4	3	75.83			
	α_1		**3**	2	5	56.32				**3**	2	5	80.88				**3**	2	5	82.73			
TUMOR3	α_2	3	8	7	6	72.73	8	7	6	7	3	8	63.49	**3**	5	7	7	8	3	56.4	**3**	5	2
	α_3		5	4	8	66.61				5	3	4	77.42				5	4	3	74.11			
	α_1		3	2	8	54.24				**2**	3	4	65.89				**2**	3	4	74.28			
NO-TUMOR1	α_2	2	8	7	6	74.34	8	7	6	8	6	7	63.32	4	**2**	8	6	7	4	60.68	4	6	**2**
	α_3		5	4	8	67.55				4	5	8	72.48				4	5	6	75.06			
	α_1		3	2	5	54.80				**2**	3	4	68.52				**2**	3	4	74.86			
NO-TUMOR2	α_2	2	8	7	6	73.56	8	7	6	8	7	6	62.78	8	4	**2**	8	6	7	64.74	**2**	4	8
	α_3		5	4	8	67.58				4	5	8	73.68				4	2	5	65.34			
	α_1		3	2	5	54.38				**2**	3	4	71.56				**2**	3	4	79.09			
NO-TUMOR3	α_2	2	6	7	8	73.83	6	7	8	6	7	8	62.00	**2**	6	4	6	8	2	54.06	**2**	4	3
	α_3		5	4	6	67.89				4	5	2	69.05				4	5	2	76.27			

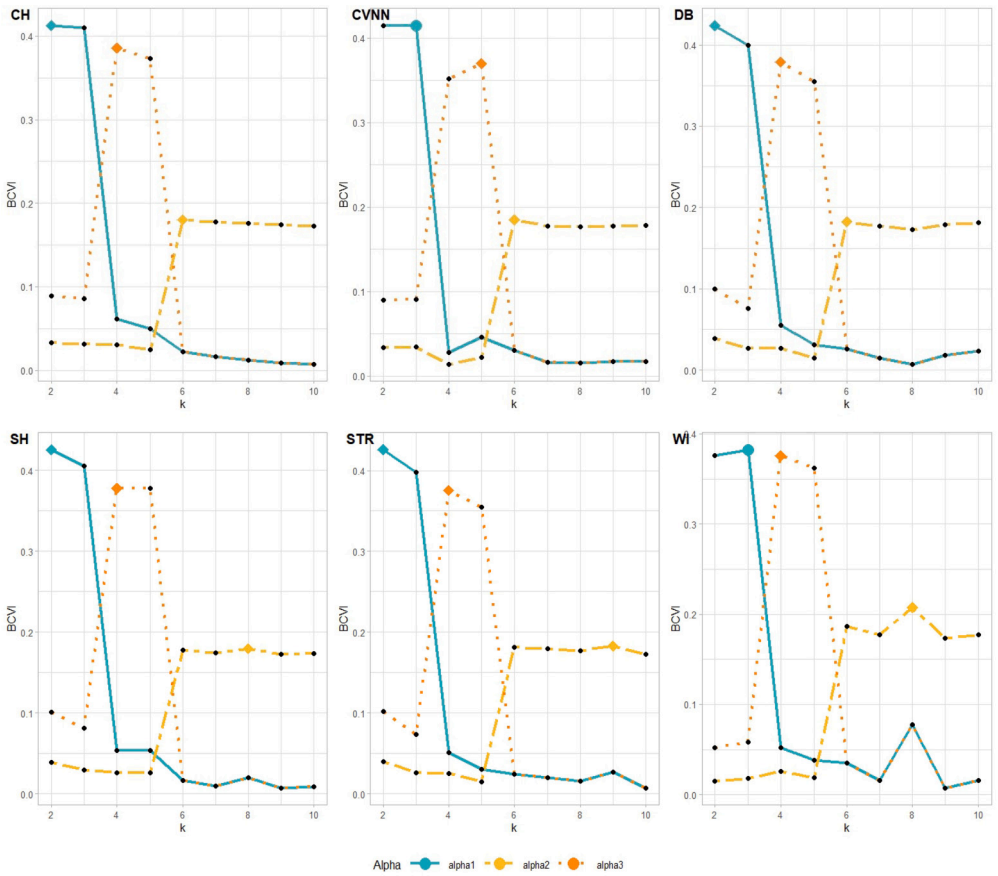


Fig. 3. Hard BCVIs based on different $\boldsymbol{\alpha}$ on the SEED dataset. A bullet indicates that BCVI detects the correct number of clusters. Conversely, a diamond indicates an incorrect detection. This also applies to the remaining figures.

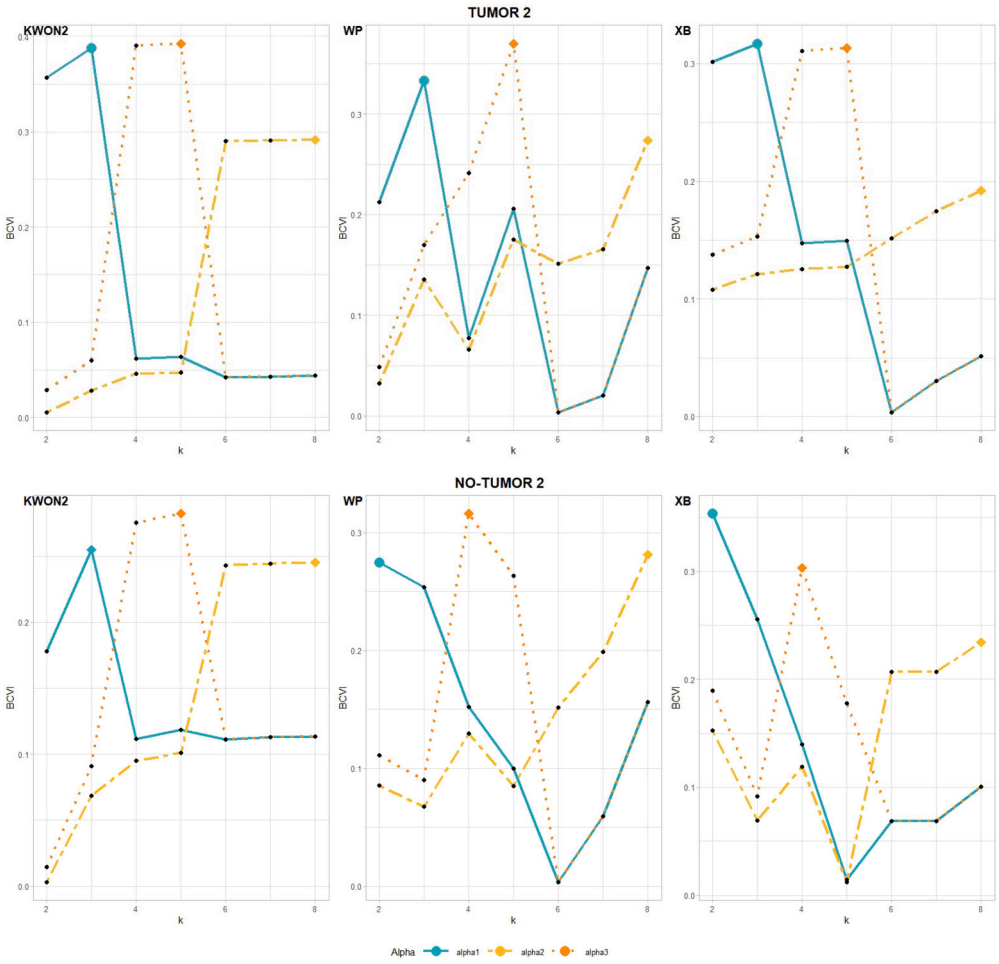


Fig. 4. Soft BCVIs based on different $\boldsymbol{\alpha}$ on the TUMOR2 and NO-TUMOR2 MRI image datasets.

Table 7
Computation times (in seconds) for running BCVI with $\boldsymbol{\alpha}_1$ and six hard underlying CVIs through K-means clustering for K from 2 to Kmax as in Table 2.

Type	Data	n	CH	CVNN	DB	SH	STR	WI
Artificial	K1-1	5500	0.63	4.06	0.99	0.97	0.70	4.52
	K1-2	850	0.08	0.12	0.06	0.08	0.06	0.16
	K2-1	740	0.05	0.09	0.11	0.06	0.05	0.11
	K2-2	1820	0.08	0.37	0.14	0.17	0.11	0.47
	K3-1	450	0.02	0.04	0.03	0.06	0.03	0.04
	K3-2	1250	0.07	0.23	0.07	0.11	0.07	0.23
Real-world	BWIS	569	0.21	0.25	0.23	0.11	0.26	0.41
	DERM	358	0.09	0.11	0.11	0.06	0.11	0.17
	SEED	210	0.04	0.06	0.04	0.06	0.03	0.04

To detect tumors in MRI images, it is evident from Fig. 2 that three colors are most appropriate for images with tumors, while two colors suffice for those without tumors. Additionally, having too many colors can needlessly complicate the images. Therefore, we set the parameters to be large at two and three groups and limit the maximum number of clusters considered to be eight. From Table 6, it is apparent that none of the existing indices can accurately detect the correct numbers of clusters for all six images. However, by applying BCVI(WP) and BCVI(XB) with $\boldsymbol{\alpha}_1$ from Table 2, we successfully identify the numbers of clusters for all six images.

According to Proposition 3.8, this suggests that the underlying WI and WP indices possess correct local peaks for all the datasets. On the other hand, CH, CVNN, DB, SH, STR, XB, and KWON2 are unable to handle this task perfectly.

As an example, we plot the BCVI values with different $\boldsymbol{\alpha}$ in Figs. 3 and 4 for SEED, TUMOR2, and NO-TUMOR2 datasets to illustrate the behaviors of BCVI based on the prior parameters. It is clear that meaningful secondary peaks for underlying indices are important for BCVI to function appropriately.

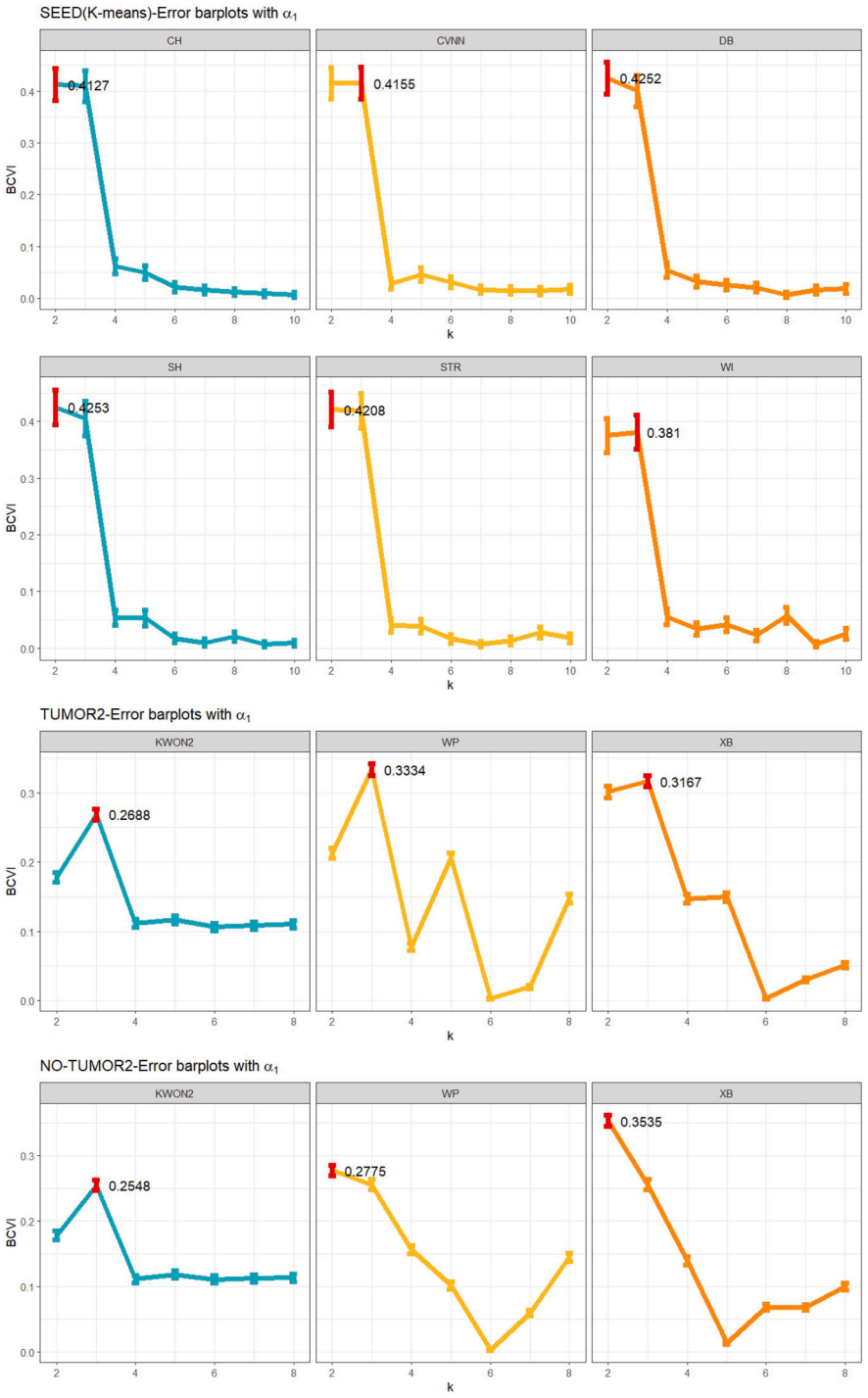


Fig. 5. BCVIs plots with two standard deviation error bars.

Additionally, we provide a few examples of plots of the BCVI values with two standard deviation error bars for the $\boldsymbol{\alpha}$ corresponding to our intended numbers of groups in Fig. 5. These standard deviations are computed directly from Corollary 3.1. This aims to demonstrate the variation of BCVI values concerning k. For the SEED dataset, though BCVI(WP) can correctly detect the number of groups, we can see that the error bars for two and three groups are intersected. Therefore, we can not say very confidently that three is the real peak.

Table 8
Computation times (in seconds) for running BCVI with $\boldsymbol{\alpha}_1$ and three soft underlying CVIs through FCM with $m = 2$ for C from 2 to Kmax as in Table 2.

Type	Data	n	KWON2	WP	XB
Artificial	C1-1	2300	8.30	10.45	8.07
	C1-2	4096	18.93	25.11	18.50
	C2-1	1000	3.60	4.64	3.57
	C2-2	1500	8.00	6.59	5.20
	C3-1	800	4.44	3.52	2.87
	C3-2	1200	3.17	3.65	2.72
Real-world	SEED	210	0.99	1.25	1.00
MRI	TUMOR1	85×85	20.80	33.96	21.36
	TUMOR2	85×85	18.21	28.72	18.14
	TUMOR3	85×85	19.31	30.18	18.87
	NO-TUMOR1	85×85	23.46	34.96	23.05
	NO-TUMOR2	85×85	21.31	32.56	21.33
	NO-TUMOR3	85×85	21.30	32.41	21.77

4.3. Time complexity

In this subsection, we compare time complexities of BCVI with all the underlying CVIs discussed in this work. The computation times were recorded from $K = 2$ to $K = Kmax$ according to Table 2. We show the results only for $\boldsymbol{\alpha}_1$ since the time complexity is consistent regardless of the parameter $\boldsymbol{\alpha}$. The experiments were implemented in R version 4.2.2 with Rstudio on Windows 11, utilizing an AMD Ryzen 5 4600H (3.0 GHz) and 16 GB of RAM.

Table 7 shows the computation time in seconds for hard BCVI with each underlying CVI through the K-means clustering with 20 initial starts. CH, DB, SH, and STR are much faster than CVNN and WI. Among the four, CH, and STR have slightly better computation times. Table 8 shows the computation time in seconds for soft BCVI with each underlying CVI through the FCM with fuzziness $m = 2$ and 20 initial starts. The computation times of BCVI with all the underlying CVIs are a bit long due to large data sizes. However, KWON2 and XB are superior to WP in term of time complexity.

5. Conclusion

In this work, we propose a new concept that directly applies the Bayesian framework to CVIs. This connection is unique and has not been explored before in the literature. Specifically, we introduce BCVI, which can be used in conjunction with any existing CVIs. This is advantageous for users who have prior knowledge of the expected number of clusters in their datasets. Although we only test BCVI with K-means and FCM, it can be applied to any clustering algorithms compatible with any CVIs. BCVI calculates the posterior probability of the actual number of clusters, with the prior distribution being either Dirichlet or GD. The main features of our BCVI are as follows:

1. **Novel and unique concept:** BCVI allows users to blend their knowledge with a dataset's pattern to identify the final number of clusters.
2. **Flexibility:** BCVI allows users to determine and flexibly set parameters according to their needs and select any clustering algorithms and underlying CVIs of their choice.

We present the results of BCVI together with nine underlying CVIs, namely CH, CVNN, DB, KWON2, SH, STR, WI, WP, and XB. While this is not a comparative test due to the new concept, the results affirm that BCVI offers several primary advantages, especially when underlying indices provide meaningful secondary options:

1. **Correcting erroneous results:** BCVI can lead to the correct number of clusters in cases where the underlying CVI is incorrect. However, this requires users to select appropriate parameters based on their knowledge.
2. **Providing alternative options:** BCVI can suggest alternative suboptimal numbers of clusters if the optimal one is not suitable for users in their context.

These advantages are especially advantageous when the expected range is definite. For example, we are aware that a few colors are enough for MRI image pre-processing. The limitations of BCVI are that:

1. It relies on the quality of underlying indices.
2. It is only effective when underlying indices are present, providing meaningful options for ranking local peaks for the final number of clusters.
3. It requires users to determine prior parameter values.

Therefore, BCVI performs especially well based on the underlying WI and WP indices which always provide secondary options. However, the other CVIs are superior to the WI and WP in term of computation times. Future research directions include studying the statistical inference of BCVI, exploring alternative prior distributions with the same support as Dirichlet distribution, testing BCVI with additional clustering algorithms and underlying CVIs, and applying it to diverse real-world applications.

Acknowledgements

All the authors would like to thank anonymous reviewers, the editor-in-chief and the associate editor for extremely helpful comments and detailed suggestions that led to a big improvement of the paper. Nathakhun would like to also thank National Research Council of Thailand (NRCT), Grant number: N42A660991 (2023) for the project financial support. Onthada has been supported by Petchra Pra Jom Klao Master's Degree Research Scholarship from King Mongkut's University of Technology Thonburi, Grant number: 32/2567.

Appendix A. Proofs of mathematical theorems and properties

A.1. Proof of Theorem 3.1

Starting from (10) and (11), the joint distribution of $(\mathbf{x}, \mathbf{p})$ is given by

$$f(\mathbf{x}, \mathbf{p}) = \frac{C(\mathbf{p})}{B(\boldsymbol{\alpha})} \prod_{k=2}^{K} p_k^{\alpha_k + nr_k(\mathbf{x})-1}.$$

Integrating over $\mathbf{p}$ to obtain the joint probability density function, the marginal of $\mathbf{x}$ is

$$m(\mathbf{x}) = \frac{B(\boldsymbol{\alpha} + n\mathbf{r}(\mathbf{x}))}{B(\boldsymbol{\alpha})} C(\mathbf{p}).$$

Therefore, the posterior probability density function is

$$\pi(\mathbf{p}|\mathbf{x}) = \frac{f(\mathbf{x}, \mathbf{p})}{m(\mathbf{x})} = \frac{1}{B(\boldsymbol{\alpha} + n\mathbf{r}(\mathbf{x}))} \prod_{k=2}^{K} p_k^{\alpha_k + nr_k(\mathbf{x})-1}. \quad \square$$

A.2. Proof of Theorem 3.2

From (10) and (12), and utilizing the fact that $p_K = 1 - p_2 - \cdots - p_{K-1}$, the joint distribution of $(\mathbf{x}, \mathbf{p})$ is given by

$$f(\mathbf{x}, \mathbf{p}) = \frac{C(\mathbf{p})}{\prod_{k=2}^{K-1} B(\alpha_k, \beta_k)} \prod_{k=2}^{K-2} p_k^{\alpha_k + nr_k(\mathbf{x})-1} (1 - p_2 - \cdots - p_k)^{\gamma_k} \left[p_{K-1}^{\alpha_{K-1} + nr_{K-1}(x)-1} \left(1 - p_2 - \cdots - p_{K-1}\right)^{\beta_{K-1} + nr_K(x)-1} \right].$$

Since

$$\begin{aligned} \gamma_k &= \beta_k - \alpha_{k+1} - \beta_{k+1} \\ &= \left(\beta_k + \sum_{i=k+1}^{K} nr_i(x) \right) - (\alpha_{k+1} + nr_{k+1}(x)) - \left(\beta_{k+1} + \sum_{i=k+2}^{K} nr_i(x) \right), \end{aligned}$$

for $k = 2, \ldots, K-2$ and

$$\beta_{K-1} + nr_K(x) - 1 = \left(\beta_{K-1} + \sum_{i=K}^{K} nr_i(x) \right) - 1,$$

taking an integral over $\mathbf{p}$ to the joint probability density function, the marginal of $\mathbf{x}$ is

$$m(\mathbf{x}) = C(\mathbf{p}) \prod_{k=2}^{K-1} \frac{B\left(\alpha_k + nr_k(x), \beta_k + \sum_{i=k+1}^{K} nr_i(x)\right)}{B\left(\alpha_k, \beta_k\right)}.$$

Therefore, the posterior probability density function can be written as

$$\pi(\mathbf{p}|\mathbf{x}) = \frac{f(\mathbf{x}, \mathbf{p})}{m(\mathbf{x})} = \prod_{k=2}^{K-1} \frac{p_k^{\alpha'_k} (1 - p_2 - \cdots - p_k)^{\gamma'_k}}{B(\alpha'_k, \beta'_k)},$$

where $\gamma'_k = \gamma_k$ for $k = 2, \ldots, K-2$, $\gamma'_{K-1} = \beta'_{K-1} - 1$ and α'_k and β'_k, are given in (13) and (14), respectively. $\square$

A.3. Proof of Corollary 3.2

For $k = 2, \dots, K-1$, we apply (7) in Lemma 2.2 with $s_k = 1$ and $s_l = 0$ for $l \neq k$, yielding

$$\begin{aligned}\mathbb{E}[p_k|\mathbf{x}] &= \frac{\Gamma(\alpha'_k+\beta'_k)\Gamma(\alpha'_k+1)\Gamma(\beta'_k)}{\Gamma(\alpha'_k)\Gamma(\beta'_k)\Gamma(\alpha'_k+\beta'_k+1)}\prod_{i:i<k}\frac{\Gamma(\alpha'_i+\beta'_i)\Gamma(\alpha'_i)\Gamma(\beta'_i+1)}{\Gamma(\alpha'_i)\Gamma(\beta'_i)\Gamma(\alpha'_i+\beta'_i+1)}\prod_{i:i>k}\frac{\Gamma(\alpha'_i+\beta'_i)\Gamma(\alpha'_i)\Gamma(\beta'_i)}{\Gamma(\alpha'_i)\Gamma(\beta'_i)\Gamma(\alpha'_i+\beta'_i)}\\ &= \frac{\alpha'_k}{\alpha'_k+\beta'_k}\prod_{i:i<k}\frac{\beta'_i}{\alpha'_i+\beta'_i},\end{aligned}$$

where we utilize $\delta_i = 1$ for $i < k$ and $\delta_i = 0$ for $i > k$. Similarly, for $s_k = 2$ and $s_l = 0$ for $l \neq k$, we have

$$\begin{aligned}\mathbb{E}[p_k^2|\mathbf{x}] &= \frac{\Gamma(\alpha'_k+\beta'_k)\Gamma(\alpha'_k+2)\Gamma(\beta'_k)}{\Gamma(\alpha'_k)\Gamma(\beta'_k)\Gamma(\alpha'_k+\beta'_k+2)}\prod_{i:i<k}\frac{\Gamma(\alpha'_i+\beta'_i)\Gamma(\alpha'_i)\Gamma(\beta'_i+2)}{\Gamma(\alpha'_i)\Gamma(\beta'_i)\Gamma(\alpha'_i+\beta'_i+2)}\prod_{i:i>k}\frac{\Gamma(\alpha'_i+\beta'_i)\Gamma(\alpha'_i)\Gamma(\beta'_i)}{\Gamma(\alpha'_i)\Gamma(\beta'_i)\Gamma(\alpha'_i+\beta'_i)}\\ &= \frac{(\alpha'_k+1)(\alpha'_k)}{(\alpha'_k+\beta'_k+1)(\alpha'_k+\beta'_k)}\prod_{i:i<k}\frac{(\beta'_i+1)(\beta'_i)}{(\alpha'_i+\beta'_i+1)(\alpha'_i+\beta'_i)}.\end{aligned}$$

For the term p_K, applying (8) from Lemma 2.2 with $s = 1$ and $s = 2$, we respectively obtain

$$\mathbb{E}[p_K|\mathbf{x}] = \prod_{k=2}^{K-1}\frac{\Gamma(\alpha'_k+\beta'_k)\Gamma(\beta'_k+1)}{\Gamma(\beta'_k)\Gamma(\alpha'_k+\beta'_k+1)} = \prod_{k=2}^{K-1}\frac{\beta'_k}{\alpha'_k+\beta'_k}$$

and

$$\mathbb{E}[p_K^2|\mathbf{x}] = \prod_{k=2}^{K-1}\frac{\Gamma(\alpha'_k+\beta'_k)\Gamma(\beta'_k+2)}{\Gamma(\beta'_k)\Gamma(\alpha'_k+\beta'_k+2)} = \prod_{k=2}^{K-1}\frac{(\beta'_k+1)\beta'_k}{(\alpha'_k+\beta'_k+1)(\alpha'_k+\beta'_k)}. \quad \square$$

A.4. Proof of Proposition 3.5

From (16) and that $\alpha_k/n \to 0$, $\alpha_k \to c_k$, and $\alpha_k/g(n) \to c_k$ as $n \to \infty$ for all k, we obtain

1. $$\lim_{n\to\infty}\texttt{BCVI}(k) = \lim_{n\to\infty}\frac{\alpha_k+nr_k(\mathbf{x})}{\alpha_0+n} = \lim_{n\to\infty}\frac{\frac{\alpha_k}{n}+r_k(\mathbf{x})}{\frac{\sum_{j=2}^K\alpha_j}{n}+1} = r_k(\mathbf{x}),$$

2. $$\lim_{n\to\infty}\texttt{BCVI}(k) = \lim_{n\to\infty}\frac{\alpha_k+nr_k(\mathbf{x})}{\alpha_0+n} = \lim_{n\to\infty}\frac{\frac{\alpha_k}{n}+r_k(\mathbf{x})}{\frac{\sum_{j=2}^K\alpha_j}{n}+1} = \frac{c_k+r_k(\mathbf{x})}{\sum_{j=2}^K c_j+1},$$

and

3. $$\lim_{n\to\infty}\texttt{BCVI}(k) = \lim_{n\to\infty}\frac{\alpha_k+nr_k(\mathbf{x})}{\alpha_0+n} = \lim_{n\to\infty}\frac{\frac{\alpha_k}{g(n)}+\frac{nr_k(\mathbf{x})}{g(n)}}{\frac{\sum_{j=2}^K\alpha_j}{g(n)}+\frac{n}{g(n)}} = \frac{c_k}{\sum_{j=2}^K c_j}. \quad \square$$

A.5. Proof of Proposition 3.7

Assume that $\alpha_k = \alpha$ for all k. From (16), we have

$$\texttt{BCVI}(k) = \frac{\alpha+nr_k(\mathbf{x})}{(K-1)\alpha+n}$$

which is a function of k only through $r_k(x)$ defined in (9). Since $r_k(x)$ is adjusted monotonously from the underlying index GI(k), BCVI(k) and GI(k) yield exactly the same ranking of the preferred numbers of groups. $\square$

A.6. Proof of Proposition 3.8

For $j \in [s]$, assume that $\alpha_{i_j-1} = \alpha_{i_j} = \alpha_{i_j+1} = \alpha$. Then, by (16), we have $\texttt{BCVI}_{i_j-1} = \frac{\alpha+nr_{i_j-1}(\mathbf{x})}{(K-1)\alpha+n}$, $\texttt{BCVI}_{i_j} = \frac{\alpha+nr_{i_j}(\mathbf{x})}{(K-1)\alpha+n}$, and $\texttt{BCVI}_{i_j+1} = \frac{\alpha+nr_{i_j+1}(\mathbf{x})}{(K-1)\alpha+n}$. Since i_j is a local peak, r_{i_j} is greater than both r_{i_j-1} and r_{i_j+1}. This implies that BCVI still has a local peak at i_j. $\square$

A.7. Proof of Proposition 3.9

Using (17) and (18) and the assumptions 1, 2, and 3 in the statement, we obtain the following.

1. Dividing both the numerator and the denominator of (17) by n^{k-1} for $k = 2, \dots, K-1$ and of (18) by n^{K-2}, we have

$$\begin{aligned}\lim_{n\to\infty} \texttt{BCVI}(k) &= \lim_{n\to\infty} \frac{\frac{\alpha_k}{n} + r_k(x)}{\frac{\alpha_k}{n} + \frac{\beta_k}{n} + \sum_{i=k}^{K} r_i(x)} \prod_{i:i<k} \frac{\frac{\beta_i}{n} + \sum_{j=i+1}^{K} r_j(x)}{\frac{\alpha_i}{n} + \frac{\beta_i}{n} + \sum_{j=i}^{K} r_j(x)} \\ &= \frac{r_k(x)}{\sum_{i=k}^{K} r_i(x)} \prod_{i:i<k} \frac{\sum_{j=i+1}^{K} r_j(x)}{\sum_{i=k}^{K} r_i(x)} = r_k(x),\end{aligned}$$

and

$$\begin{aligned}\lim_{n\to\infty} \texttt{BCVI}(K) &= \lim_{n\to\infty} \prod_{k=2}^{K-1} \frac{\beta_k + \sum_{i=k+1}^{K} n r_i(x)}{\alpha_k + \beta_k + \sum_{i=k}^{K} n r_i(x)} \\ &= \lim_{n\to\infty} \prod_{k=2}^{K-1} \frac{\frac{\beta_k}{n} + \sum_{i=k+1}^{K} r_i(x)}{\frac{\alpha_k}{n} + \frac{\beta_k}{n} + \sum_{i=k}^{K} r_i(x)} \\ &= \prod_{k=2}^{K-1} \frac{\sum_{i=k+1}^{K} r_i(x)}{\sum_{i=k}^{K} r_i(x)} = r_K(x).\end{aligned}$$

2. Dividing both the numerator and denominator of (17) by n^{k-1} for $k = 2, \dots, K-1$ and of (18) by n^{K-2}, we obtain

$$\begin{aligned}\lim_{n\to\infty} \texttt{BCVI}(k) &= \lim_{n\to\infty} \frac{\frac{\alpha_k}{n} + r_k(x)}{\frac{\alpha_k}{n} + \frac{\beta_k}{n} + \sum_{i=k}^{K} r_i(x)} \prod_{i:i<k} \frac{\frac{\beta_i}{n} + \sum_{j=i+1}^{K} r_j(x)}{\frac{\alpha_i}{n} + \frac{\beta_i}{n} + \sum_{j=i}^{K} r_j(x)} \\ &= \frac{c_k + r_k(x)}{c_k + d_k + \sum_{i=k}^{K} r_i(x)} \prod_{i:i<k} \frac{d_i + \sum_{j=i+1}^{K} r_j(x)}{c_i + d_i + \sum_{j=i}^{K} r_j(x)},\end{aligned}$$

and

$$\begin{aligned}\lim_{n\to\infty} \texttt{BCVI}(K) &= \lim_{n\to\infty} \prod_{k=2}^{K-1} \frac{\frac{\beta_k}{n} + \sum_{i=k+1}^{K} r_i(x)}{\frac{\alpha_k}{n} + \frac{\beta_k}{n} + \sum_{i=k}^{K} r_i(x)} \\ &= \prod_{k=2}^{K-1} \frac{d_k + \sum_{i=k+1}^{K} r_i(x)}{c_k + d_k + \sum_{i=k}^{K} r_i(x)}.\end{aligned}$$

3. Dividing both the numerator and the denominator of (17) by $g^{k-1}(n)$ for $k = 2, \dots, K-1$, and dividing both the numerator and denominator of (18) by $g^{K-2}(n)$, we obtain

$$\begin{aligned}\lim_{n\to\infty} \texttt{BCVI}(k) &= \lim_{n\to\infty} \frac{\frac{\alpha_k}{g(n)} + \frac{n r_k(x)}{g(n)}}{\frac{\alpha_k}{g(n)} + \frac{\beta_k}{g(n)} + \frac{n}{g(n)} \sum_{i=k}^{K} r_i(x)} \prod_{i:i<k} \frac{\frac{\beta_i}{g(n)} + \frac{n}{g(n)} \sum_{j=i+1}^{K} r_j(x)}{\frac{\alpha_i}{g(n)} + \frac{\beta_i}{g(n)} + \frac{n}{g(n)} \sum_{j=i}^{K} r_j(x)} \\ &= \frac{c_k}{c_k + d_k} \prod_{i:i<k} \frac{d_i}{c_i + d_i},\end{aligned}$$

and

$$\lim_{n\to\infty} \texttt{BCVI}(K) = \lim_{n\to\infty} \prod_{k=2}^{K-1} \frac{\frac{\beta_k}{g(n)} + \frac{n}{g(n)} \sum_{i=k+1}^{K} r_i(x)}{\frac{\alpha_k}{g(n)} + \frac{\beta_k}{g(n)} + \frac{n}{g(n)} \sum_{i=k}^{K} r_i(x)} = \prod_{k=2}^{K-1} \frac{d_k}{c_k + d_k}. \quad \square$$

References

Abdullah, A.A., Hassan, M.M., Mustafa, Y.T., 2022. A review on Bayesian deep learning in healthcare: applications and challenges. IEEE Access 10, 36538–36562. https://doi.org/10.1109/ACCESS.2022.3163384.

Aggarwal, C.C., Reddy, C.K., 2014. Data clustering. Algorithms Appl. 93.

Akhanli, S.E., Hennig, C., 2020. Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. Stat. Comput. 30, 1523–1544.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I., 2013. An extensive comparative study of cluster validity indices. Pattern Recognit. 46, 243–256. https://doi.org/10.1016/j.patcog.2012.07.021.

Barton, T., 2019. Clustering benchmarks. https://github.com/deric/clustering-benchmark/.

Bezdek, J.C., Ehrlich, R., Full, W., 1984. Fcm: the fuzzy c-means clustering algorithm. Comput. Geosci. 10, 191–203. https://doi.org/10.1016/0098-3004(84)90020-7.

Bezdek, J.C., Li, W., Attikiouzel, Y., Windham, M.P., 1997. A geometric approach to cluster validity for normal mixtures. Soft Comput. 1, 166–179.

Bianchi, G., Heo, C.Y., 2021. A Bayesian statistics approach to hospitality research. Curr. Issues Tour. 24, 3141–3150. https://doi.org/10.1080/13683500.2021.1896486.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat. 3, 1–27. https://doi.org/10.1080/03610927408827101. https://www.tandfonline.com/doi/abs/10.1080/03610927408827101.

Chakrabarty, N., 2019. Brain mri images for brain tumor detection. https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection.
Chandukala, S.R., Dotson, J.P., Brazell, J.D., Allenby, G.M., 2011. Bayesian analysis of hierarchical effects. Mark. Sci. 30, 123–133.
Chen, H., Xie, T., Liang, M., Liu, W., Liu, P.X., 2023. A local tangent plane distance-based approach to 3d point cloud segmentation via clustering. Pattern Recognit. 137, 109307. https://doi.org/10.1016/j.patcog.2023.109307.
Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1, 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.
Dunn, J.C., 1973. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. J. Cybern. 3, 32–57. https://doi.org/10.1080/01969727308546046.
Grazian, C., 2023. A review on Bayesian model-based clustering. arXiv:2303.17182.
Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M., 2021. Deep learning for 3d point clouds: a survey. IEEE Trans. Pattern Anal. Mach. Intell. 43, 4338–4364. https://doi.org/10.1109/TPAMI.2020.3005434.
Hennig, C., 2024. fpc: flexible procedures for clustering. https://CRAN.R-project.org/package=fpc. r package version 2.2-12.
Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193–218.
James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J., 2023. An Introduction to Statistical Learning: With Applications in Python. Springer Nature.
Kassambara, A., Mundt, F., 2020. factoextra: extract and visualize the results of multivariate data analyses. https://CRAN.R-project.org/package=factoextra. r package version 1.0.7.
Kaufman, L., Rousseeuw, P.J., 2009. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons.
Kumar, D., Bezdek, J.C., 2020. Clustering tendency assessment for datasets having inter-cluster density variations. In: 2020 International Conference on Signal Processing and Communications (SPCOM), pp. 1–5.
Kwon, S.H., Kim, J., Son, S.H., 2021. Improved cluster validity index for fuzzy clustering. Electron. Lett. 57, 792–794.
Lee, S.W., Kim, Y.S., Bien, Z., 2008. A probabilistic cluster validity index for agglomerative Bayesian fuzzy clustering. In: 2008 International Conference on Computational Intelligence for Modelling Control & Automation, pp. 368–373.
Lichman, M., 2013. Uci machine learning repository. http://archive.ics.uci.edu/datasets.
Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., Wu, S., 2013. Understanding and enhancement of internal clustering validation measures. IEEE Trans. Cybern. 43, 982–994. https://doi.org/10.1109/TSMCB.2012.2220543.
Lohrke, F.T., Carson, C.M., Lockamy, A., 2018. Bayesian analysis in entrepreneurship decision-making research: review and future directions. Manag. Decis. 56, 972–986.
MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, J. (Eds.), Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp. 281–297.
Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., Long, J., 2018. A survey of clustering with deep learning: from the perspective of network architecture. IEEE Access 6, 39501–39514. https://doi.org/10.1109/ACCESS.2018.2855437.
Nickparvar, M., 2022. Brain tumor mri dataset. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset.
Olkin, I., Rubin, H., 1964. Multivariate beta distributions and independence properties of the Wishart distribution. Ann. Math. Stat. 35, 261–269. https://doi.org/10.1214/aoms/1177703748.
Preedasawakul, O., Wiroonsri, N., 2024. BayesCVI: Bayesian Cluster Validity Index. R package version 1.0.0.
Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. 66, 846–850. https://doi.org/10.1080/01621459.1971.10482356.
Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7. https://www.sciencedirect.com/science/article/pii/0377042787901257.
RStudio Team, 2020. RStudio: Integrated Development Environment for R. RStudio. PBC, Boston, MA.
Schmid, V.J., Whitcher, B., Padhani, A.R., Taylor, N.J., Yang, G.Z., 2006. Bayesian methods for pharmacokinetic models in dynamic contrast-enhanced magnetic resonance imaging. IEEE Trans. Med. Imaging 25, 1627–1636. https://doi.org/10.1109/TMI.2006.884210.
Starczewski, A., 2017. A new validity index for crisp clusters. Pattern Anal. Appl. 20, 687–700.
Vysala, A., Gomes, D.J., 2020. Evaluating and validating cluster results. arXiv:2007.08034.
Wiroonsri, N., 2024. Clustering performance analysis using a new correlation-based cluster validity index. Pattern Recognit. 109910. https://doi.org/10.1016/j.patcog.2023.109910.
Wiroonsri, N., Preedasawakul, O., 2023a. A correlation-based fuzzy cluster validity index with secondary options detector. https://doi.org/10.48550/arXiv.2308.14785. arXiv:2308.14785, 2023.
Wiroonsri, N., Preedasawakul, O., 2023b. UniversalCVI: Hard and Soft Cluster Validity Indices. R package version 1.1.2.
Wong, T.T., 1998. Generalized Dirichlet distribution in Bayesian analysis. Appl. Math. Comput. 97, 165–181. https://doi.org/10.1016/S0096-3003(97)10140-0.
Xie, X., Beni, G., 1991. A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. 13, 841–847. https://doi.org/10.1109/34.85677.
Xie, Y., Tian, J., Zhu, X.X., 2020. Linking points with labels in 3d: a review of point cloud semantic segmentation. IEEE Geosci. Remote Sens. Mag. 8, 38–59. https://doi.org/10.1109/MGRS.2019.2937630.