Journal of Classification
https://doi.org/10.1007/s00357-024-09481-3

ORIGINAL RESEARCH

Cluster Validation Based on Fisher's Linear Discriminant Analysis

Fabian Kächele[1] · **Nora Schneider**[1]

Accepted: 10 June 2024

Abstract
Cluster analysis aims to find meaningful groups, called clusters, in data. The objects within a cluster should be similar to each other and dissimilar to objects from other clusters. The fundamental question arising is whether found clusters are "valid clusters" or not. Existing cluster validity indices are computation-intensive, make assumptions about the underlying cluster structure, or cannot detect the absence of clusters. Thus, we present a new cluster validation framework to assess the validity of a clustering and determine the underlying number of clusters k^*. Within the framework, we introduce a new merge criterion analyzing the data in a one-dimensional projection, which maximizes the ratio of between-cluster-variance to within-cluster-variance in the clusters. Nonetheless, other local methods can be applied as a merge criterion within the framework. Experiments on synthetic and real-world data sets show promising results for both the overall framework and the introduced merge criterion.

Keywords Clustering · Cluster validation · Discriminant analysis · Number of clusters

1 Introduction

Cluster analysis has become an important tool for exploratory data analysis with theoretical and practical applications in a broad domain, including pattern recognition, image analysis, and information retrieval. It aims to partition objects of a given data set into meaningful groups, called clusters, to detect any hidden structure or to summarize large data sets. The objects within a cluster should be similar to each other and dissimilar to objects from other clusters. The fundamental issue concerning the validity and applicability of clustering results remains in finding the appropriate number of clusters k^*. While a division with an excessive number of clusters is non-intuitive and difficult to analyze, a division with too few clusters results in information loss. Thus, a good balance between accuracy and compressibility has to be found. Additionally, many popular clustering algorithms require the number of clusters as an input parameter. Therefore, methods for estimating the optimal number of clusters and

✉ Fabian Kächele
fabian.kaechele@kit.edu

[1] Institute for Operations Research, Analytics & Statistics, Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany

Published online: 04 July 2024

assessing the validity of a clustering are essential. We recommend, e.g., Rendón et al. (2011) or Wierzchoń (2018) for discussions and Hennig (2015) for a slightly philosophic perspective on the problem.

Despite the extensive research on the problem of determining the optimal number of clusters k^* (e.g., Dubes, 1987; Peck et al., 1989; Tibshirani et al., 2001; Sugar and James, 2003; von Luxburg, 2010 among others), the outcome is still unsatisfactory (Salvador and Chan, 2004; Omran et al., 2011) and the problem remains an active field of research, e.g., (Fu and Perry, 2020; Dangl and Leisch, 2019; Rossbroich et al., 2022). A common approach to obtain an estimate is by optimizing a validity index over different values for the number of clusters k. However, such approaches are computation-intensive, make assumptions about the underlying cluster structure, or cannot detect the absence of clusters, i.e., $k^* = 1$. Thus, the suitability of an approach depends on and varies with the data it is applied to. Another popular approach to determine the optimal number of clusters is introducing stopping rules for hierarchical clustering (Xu and Wunsch, 2005). Stopping rules define when the procedure should ideally stop, and the number of clusters of the resulting clustering is used as an estimate of the optimal number of clusters (Baker and Hubert, 1975; Milligan and Cooper, 1985; Cerdeira et al., 2012). For example, most lately Geng et al. (2019) presented an algorithm to assess the number of clusters in a network of communities. For a more general overview, we refer to, e.g., Hennig et al. (2015); Wiwie et al. (2015); Handl et al. (2005), and Hennig (2022) for recent applications of cluster evaluation.

Besides the problem of determining the optimal number of clusters k^*, the closely linked problem of assessing the overall validity of a clustering is of paramount importance. Several methods are proposed to test whether a clustering is valid or a product of randomness. See, e.g., Rand (1971); Bailey and Dubes (1982); Gates and Hansell (1983); Gordon (1998); Ingrassia and Punzo (2020); Ullmann et al. (2022) or Liu et al. (2008); Halkidi et al. (2001) for an overview. However, as in the former problem, most available methods are computation-intensive or need parametric assumptions. Closely related to the proposed method, existing local methods aim to decide whether two clusters are separated and thus can be used to fulfill both tasks given above. Note that they also can be applied as stopping rules for hierarchical clustering, i.e., as split or merge criteria. For example, Sneath (1977) proposed to test the distinctness of two clusters C_i and C_j by measuring the overlap of their projections onto the intercentroid line connecting the two cluster means. Further frequently used methods were proposed by Caliński and Harabasz (1974); Davies and Bouldin (1979) and Rousseeuw and Kaufman (1990) and are based on some measure of the compactness of a cluster. However, current research shows that the optimal choice of a criterion to compare a respective clustering is ambiguous (Arbelaitz et al., 2013; Wierzchoń, 2018).

Tackling the mentioned challenges of existing cluster validation methods, we present a new, intuitive, easy-to-implement, and powerful framework, which leverages local methods to estimate an appropriate number of clusters k^* based on a given clustering $\mathbf{C}$ and simultaneously assess the validity of the given clustering. This is especially useful when no knowledge of the underlying number of clusters is available. Further and in contrast to others, it does not require expensive reruns of the clustering algorithm nor any distributional assumptions. Within the framework, we propose a new, local merge criterion (i.e., local method) based on the variance of a projection. Here, we compare the variance of all cluster pairs $\{C_s, C_t\} \in \mathbf{C}$ in a one-dimensional subspace of the data to the variance of an artificial cluster consisting of points from C_s and C_t. The subspace is constructed by using the direction of Fisher's Linear Discriminant Analysis (Fisher, 1936), aiming to maximize the ratio of between to within variance in the clusters. Therefore, we call the proposed approach CVFLDA — an acronym of cluster validation based on Fisher's linear discriminant analysis. Further, we introduce a

nuance parameter to control the method's sensitivity. In contrast to existing approaches, the proposed method does not require computational extensive reruns of the clustering algorithm and is able to detect cases without an underlying cluster structure, i.e., $k^* = 1$. We show its applicability in a benchmarking study and demonstrate its superiority above other cluster evaluation methods in a simulation study as well as with real data set examples. Additionally, we provide a ready-to-use Python implementation of our method (see Appendix B).

The remaining of the paper is structured as follows: In Section 2, we introduce the new framework and the proposed merge criterion. Section 3 presents a benchmarking study based on simulated data. Then, in Section 4, we display the results of multiple applications on real-world data sets. Last, Section 5 summarises our contribution and points out possible further developments.

2 Cluster Validation

Consider a multivariate data matrix $\mathbf{X} = \left(\mathbf{x}_1^\top, ..., \mathbf{x}_n^\top\right) \in \mathbb{R}^{n \times d}$ consisting of d attributes observed on n objects. The n-dimensional vector $\mathbf{y}$ contains the cluster labels for the respective objects and k is the number of distinct clusters in the corresponding clustering $\mathbf{C}$. Formally $\mathbf{C} = \{C_1, ..., C_k\}$ is a clustering such that

$$C_s \cap C_t = \emptyset, \ s! = t \text{ for all cluster pairs} \in \mathbf{C} \text{ and } \bigcup_{\ell=1}^{k} C_\ell = \{\mathbf{x}_1, ..., \mathbf{x}_n\}.$$

We assume that the initial number of clusters in the given clustering $\mathbf{C}$ is greater or equal to the optimal number of clusters k^*, i.e., $k \geq k^*$. The optimal number of clusters k^* is the true underlying number of clusters in the data, which is usually unknown. In the following, we propose a new method to audit the clustering $\mathbf{C}$ and return modified cluster labels $\mathbf{y}'$ that correspond to a new clustering $\mathbf{C}' = \left\{C'_1, ..., C'_{\hat{k}^*}\right\}$. The new clustering includes the estimated number of clusters $\hat{k}^*$ in case of a non-valid, initial clustering $\mathbf{C}$ with $k \geq k^*$.

2.1 Cluster Validation Framework

In this section, we introduce the proposed cluster validation framework. Remember that the framework only provides a step-by-step procedure in which any desired merge criterion can be applied.

The framework iteratively evaluates for every cluster pair $(C_s, C_t) \in \mathbf{C}$ whether they are two true clusters, i.e., well separated, or (part of) one cluster that was wrongly split by a clustering algorithm. To do so, we propose a merge criterion that compares the variance of a new, merged cluster and the variance of both initial clusters in a one-dimensional projection. Next, we sort the cluster pairs in descending order based on their indication of stemming from the same single cluster, which is measured by our criterion. Finally, we merge the first cluster pair in the list, meaning the two clusters with the lowest combined scatter (thus being close together) with respect to the sum of their individual scatters, and repeat the procedure. This evaluation can be carried out by any merge criteria (local method). The described process is repeated for the modified clustering until there are only non-mergeable clusters left, i.e., the merge criteria rejects the hypothesis of two clusters stemming from the same original

cluster for each cluster pair. In the end, the method returns the modified clustering and the corresponding labels, which consist of the estimate for the optimal number of clusters. For the k-means algorithm used in our simulation study, Melnykov and Michael (2020) discuss the effect of merging clusters whereas the merging of clusters generated from other algorithms is discussed, e.g., by Melnykov (2016) and Li (2005). The pseudocode of the framework is given in Algorithm 1.

Algorithm 1 Cluster Validation Framework.

Require: Clustering $\mathbf{C}$, false_clusters = True
  $\mathbf{C} \rightarrow \mathbf{C}'$
  while false_clusters **do**
    false_clusters = False, $list$ = empty
    for every cluster pair $k = (C'_s, C'_t) \in \mathbf{C}'$ **do**
      Calculate merge criteria for k
      if (C'_s, C'_t) are false clusters **then**:
        false_clusters = True
        Add k to $list$
      end if
    end for
    Sort $list$ by the possibility of being one cluster
    Update $\mathbf{C}'$ by merging first cluster pair from $list$
  end while
  return Number of clusters $\hat{k}^*$ and new clustering $\mathbf{C}'$

2.2 Merge Criterion

Besides the simple validation framework, we propose a new, suitable merge criterion to use therein. For our variance-based approach, we apply a one-dimensional projection in which the variance between both clusters C_s and C_t is maximized and the variance within each cluster is minimized. Thus, intuitively, the projected data maximizes the contrast between clusters. Working in such one-dimensional subspaces has already proven successful for clustering problems by Peña and Prieto (2001) and Delaigle et al. (2019). As we only consider the variance of the data (or groups of data), we center it for simpler math and calculate the scatter, i.e., the scaled covariance matrix by

$$\Sigma = (n-1) \times \text{cov}(\mathbf{X}) = \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top = \sum_{i=1}^{n} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top, \tag{1}$$

where $\boldsymbol{\mu}$ represents the d-dimensional mean vector of the data and $\tilde{\mathbf{x}}_i$ the centered data points for $i \in \{1, \ldots, n\}$. Now, remember that we can project a point $\tilde{\mathbf{x}}_i$ on a one-dimensional subspace with the help of vector $\mathbf{z} \in \mathbb{R}^{d \times 1}$ with $\mathbf{z}^\top \mathbf{z} = \mathbf{1}$, i.e.,

$$\dot{x}_i = \mathbf{z}^\top \tilde{\mathbf{x}}_i,$$

and re-project it to $\hat{\mathbf{x}}_i$ by

$$\hat{\mathbf{x}}_i = \mathbf{z}\mathbf{z}^\top \tilde{\mathbf{x}}_i$$

for $i \in \{1, \ldots, n\}$. The squared Euclidean norm of this re-projected vector is given by

$$\| \hat{\mathbf{x}}_i \|_2^2 = \| \mathbf{z}\mathbf{z}^\top \tilde{\mathbf{x}}_i \|_2^2 = \mathbf{z}^\top \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \mathbf{z}$$

and the summation of the squared Euclidean norm of all n samples is given by

$$\sum_{i=1}^{n} \parallel \hat{\mathbf{x}}_i \parallel_2^2 = \sum_{i=1}^{n} \mathbf{z}^\top \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \mathbf{z} = \mathbf{z}^\top \left(\sum_{i=1}^{n} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \right) \mathbf{z}, \tag{2}$$

for $i \in \{1, \ldots, n\}$. It follows (see Eqs. 1 and 2) that the former expression is the variance of the resulting projection, or more precisely, the variance of the corresponding entry in the scatter matrix Σ. Thus, the latter is the sum of the squared Euclidean distance of the reconstructions.

We now define the within-Scatter Σ_W of a cluster pair C_s and C_t by

$$\Sigma_W = \sum_{\ell \in s,t} \Sigma^\ell = \sum_{\ell \in s,t} \sum_{i=1}^{n_\ell} (\mathbf{x}_i^{(\ell)} - \boldsymbol{\mu}_\ell)(\mathbf{x}_i^{(\ell)} - \boldsymbol{\mu}_\ell)^\top,$$

where n_ℓ is the number of samples and $\boldsymbol{\mu}_\ell = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \mathbf{x}_i^{(\ell)}$ the mean of a cluster C_ℓ, $\ell \in \{s, t\}$. Equally, the between-Scatter Σ_B of the two clusters is defined by

$$\Sigma_B = \sum_{\ell=1}^{2} n_\ell (\boldsymbol{\mu}_\ell - \boldsymbol{\mu})(\boldsymbol{\mu}_\ell - \boldsymbol{\mu})^\top$$

where n_ℓ again is the size of cluster C_ℓ, $\ell \in \{s, t\}$ for the two clusters at hand.

As stated earlier, the optimal projection $\mathbf{z}$ maximizes the scatter between the clusters

$$\max_{\mathbf{z}} \mathbf{z}^\top \Sigma_B \mathbf{z} \tag{3}$$

and simultaneously minimizes the within-scatter Σ_W

$$\min_{\mathbf{z}} \mathbf{z}^\top \Sigma_W \mathbf{z}. \tag{4}$$

Therefore, Eqs. 3 and 4 can be merged into one optimization problem by

$$\max_{\mathbf{z}} \frac{\mathbf{z}^\top \Sigma_B \mathbf{z}}{\mathbf{z}^\top \Sigma_W \mathbf{z}}, \tag{5}$$

which corresponds to the Fisher criterion and reduces to a generalized eigenvalue problem, where $\mathbf{z}$ is the eigenvector with the largest eigenvalue of the matrix $A = \Sigma_W^{-1} \Sigma_B$ (see Hastie et al., 2009). Note that this approach is similar to Fisher's linear discriminant analysis (Fisher, 1936), assuming the clusters are the classes in the data.

For the one-dimensional projection $\dot{x}_i^{(\ell)}$ of an object $\mathbf{x}_i^{(\ell)} \in C_\ell$, $\ell \in \{s, t\}$ and $i \in \{1, \ldots, n_\ell\}$ follows

$$\dot{x}_i^{(\ell)} = \mathbf{z}^\top \mathbf{x}_i^{(\ell)} = \mathbf{z}^\top (\boldsymbol{\mu}_\ell + \mathbf{x}_i^{(\ell)} - \boldsymbol{\mu}_\ell) = \mathbf{z}^\top \boldsymbol{\mu}_\ell + \mathbf{z}^\top \tilde{\mathbf{x}}_i^{(\ell)} \tag{6}$$

where $\boldsymbol{\mu}_\ell$ is the centroid of C_ℓ and $\mathbf{z}^\top \tilde{\mathbf{x}}_i^{(\ell)}$ the length of the projection.

With the projection of each cluster at hand, we can now form a new merged cluster C_m used for comparison in our merge criterion. Several different ways to do so may come to mind, including modeling decisions such as the size of the merged cluster, the balance with regard to both initial clusters, and the choice of elements from each initial cluster. The merged cluster consists of the cluster-halves of both clusters with the smallest (Euclidean) distance to the centroid of the respective other cluster in the projection. From simulations, we conclude that an equally weighted merged cluster seems to be beneficial, i.e., C_m is constructed so that 50% stem from C_s and the other 50% stem from C_t. Assuming that cluster C_s has more samples than C_t, we can select all $\lfloor n_{C_t}/2 \rfloor$ points from the cluster-half of C_t closest to C_s.

For selecting points from C_s, we need further selecting criteria to ensure that both C_s and C_t are represented by the same number of points in C_m. We investigated two strategies to select $\lfloor n_{C_s}/2 \rfloor$ points from C_s: Randomly sampling points (without replacement) of C_s from the closest cluster-half and selecting the samples with the closest distance to the centroid of C_t. Empirical results show that the first option leads to more robustness for unbalanced cluster sizes (see Appendix A.1 for details). Thus, by applying this strategy C_m has the same size as the smaller cluster of C_s and C_t.

In the next step, the sample variances of both initial clusters C_s, C_t, and the artificial, merged cluster C_m are calculated in direction of $\mathbf{z}$, i.e., in the one-dimensional projection. Recall that the variance of a cluster, e.g., cluster C_ℓ with n points $\mathbf{X}^{(\ell)} = (\mathbf{x}_1^{(\ell)\top}, \ldots, \mathbf{x}_n^{(\ell)\top})$ in direction of $\mathbf{z}$ is equivalent to the variance of the respective projection lengths (see Eq. 6):

$$\begin{aligned} \text{var}(\dot{x}^{(\ell)}) &= \text{var}(\mathbf{z}^\top \boldsymbol{\mu}_\ell + \mathbf{z}^\top \tilde{\mathbf{x}}^{(\ell)}) \\ &= \text{var}(\mathbf{z}^\top \tilde{\mathbf{x}}^{(\ell)}), \end{aligned}$$

where $\boldsymbol{\mu}_\ell$ represents the mean of the cluster and $\tilde{\mathbf{x}}^{(\ell)}$ the centered data of cluster $C_\ell,\ \ell \in \{s, t\}$.

The derived decision-making policy is straightforward. If the merged cluster C_m has a larger variance than both original clusters C_s and C_t, the clusters are separated enough to be considered true clusters. Otherwise, the variance of the merged cluster indicates that C_s and C_t are (part of) one cluster, i.e., they are a pair of false clusters.

We compare the variances of the merged cluster and the original clusters in the one-dimensional projection by introducing an additional parameter, the margin of safety $\lambda \geq 0$. λ is defined as a multiplier for the expected standard deviation of the variance and acts as a sensitivity parameter. The variance $S^2_{(\ell)}$ of a cluster $C_\ell,\ \ell \in \{s, t\}$ and the corresponding standard deviation of the variance $SD_{(\ell)}$ are given by

$$\begin{aligned} S^2_{(\ell)} &= \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} (\dot{x}_i^{(\ell)} - \bar{\dot{x}}^{(\ell)})^2 \\ SD_{(\ell)} &= \sqrt{\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \left((\dot{x}_i^{(\ell)} - \bar{\dot{x}}^{(\ell)})^2 - S^2_{(\ell)} \right)^2}, \end{aligned}$$

where n_ℓ is the number of points in the cluster and $\bar{\dot{x}}^{(\ell)}$ the cluster mean in the projection. Consequently, two clusters C_s and C_t are assumed to be separated if

$$S^2_{(s)} + \lambda \times SD_{(s)} < S^2_{(m)} \tag{7}$$

and

$$S^2_{(t)} + \lambda \times SD_{(t)} < S^2_{(m)}. \tag{8}$$

If one cluster pair is not well separated, our framework suggests merging the cluster pair for which $S^2_{(m)}/(S^2_{(t)} + S^2_{(s)})$ is minimized per iteration, as proposed in Algorithm 1.

The parameter λ controls the trade-off between detecting two clusters and merging two clusters. While large values result in two clusters being merged more easily, small values result in predicting more distinct clusters. The right choice of λ depends on the underlying data structure. We provide empirical results of the effect of λ on the performance of our proposed method on different synthetic datasets in Sections 3 and Appendix A.2. Generally, we recommend choosing λ based on the underlying data structure and the trade-off between

the costs associated with having too many clusters versus having too few clusters. If overestimating the number of clusters in the data is expensive, λ should be set to a higher number. Otherwise, if an underestimation is more hurtful, λ should be chosen smaller.

3 Benchmarking Study

In this section, we present the results of simulations on synthetic data. We evaluated and compared the performance of the CVFLDA to well-known validation methods on synthetic, simulated data. Similar to other studies (Milligan and Cooper, 1985; Tibshirani et al., 2001; Arbelaitz et al., 2013), we apply a clustering algorithm to data with a set of different values for k (number of clusters) to obtain various clusterings. In our experiment, we use the k-means-algorithm with a range of $k \in [2, 15]$ to obtain different initial clusterings and evaluate cluster validity indices on them. The estimate for the underlying number of clusters in the data $\hat{k}^*$ of each index is given by its minimal value. However, one of the strengths of our framework is that it does not require several clusterings (one for each k) of the "raw"-data to estimate the underlying number of clusters. Therefore, it comes with a much lower computational cost than its competitors. As our framework requires only an initial clustering with $k \geq k^*$, we investigate the effect of $k \gg k^*$ within our framework in Appendix A.3. The computer code for the CVFLDA is given in Appendix B. Note again that the result naturally depends on the initial clustering given to the algorithm.

3.1 Baseline Methods

Comparing our proposed method to state-of-art approaches is twofold. First, we compare the performances of our framework and merge criterion with the "traditional" approach of optimizing validity indices and the gap statistic. Thus, we calculate the Calinski-Harabasz (CH), Davies-Bouldin (DB), and Silhouette indices (SIL) which are top performers in comparative studies by Milligan and Cooper (1985) and Arbelaitz et al. (2013) for each k. For the estimate obtained by the gap statistic (GAP) (Tibshirani et al., 2001), we use ten reference distributions. Second, we evaluate the performance of our merge criterion versus the criterion by Sneath (1977) ($w = \sqrt{3}$ as recommended) by applying both within our framework for a fixed $k = 15$. See the Supplementary material linked in Appendix B for a detailed description of the used baseline methods.

3.2 Simulation Setup

In the simulation presented in this section, we generate synthetic data sets that contain either no cluster structure or multivariate clusters. For the cases without a cluster ($k^* = 1$), we draw $n = 300$ objects from a uniform distribution over the unit (hyper-) cube embedded in 2, 4, and 8 dimensions. For the cluster-structured data sets, we follow the data generation process used by Milligan and Cooper (1985); Tibshirani et al. (2001) and Arbelaitz et al. (2013). The generated data covers variations of the following aspects: dimensions (2, 4, 8), number of clusters (2, 4, 6, 8), overlap, and respective cluster sizes for the first cluster ($n \in [100, 200, 400]$). The cluster centers are randomly located in the hypercube window defined by the interval $I_C = [0, 50] \times [0, 20] \times \cdots \times [0, 20]$. Coincidentally, the variances for the respective attributes are randomly chosen from the interval $I_{\text{var}} = [0.25, 16]$. The corresponding coordinates for all cluster pairs had to be at least $\left(f \times (\sqrt{\text{var}_{C_i}} + \sqrt{\text{var}_{C_j}})\right)$

apart. For the value of the separation factor f, we sample from a uniform distribution over the interval [1.37, 1.88]. The factor allows the creation of clusters that are close to each other but still clearly separable. Potential overlaps in other dimensions were permitted in any form. We used the described data generation process to simulate different types of clusters; multivariate Gaussian clusters with and without correlated attributes, skewed and heavy-tailed clusters (see Supplementary Materials for details). Different from Milligan and Cooper (1985), we do not use truncated distributions. For each cluster type, we generated three different cluster structures (mean and variances) for every combination of dimensions and cluster numbers and repeated each experiment 30 times per cluster structure to get reliable results. In total, 3240 data sets with cluster structure and 300 data sets with a random structure, i.e., no clusters, were evaluated per cluster structure. Further, we apply all methods on high-dimensional data and clusterings from hierarchical clustering. From a practitioner's perspective, the difference in computational burden may be interesting. We report this in the Supplementary Material.

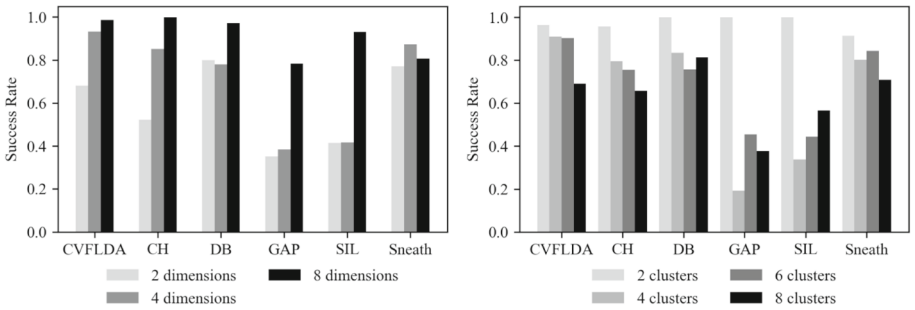
3.3 Simulation Results

Results for multivariate normal clusters with independent attributes are summarized in Table 1. The estimates of the $\text{CVFLDA}_{\lambda=0}$ obtain the highest success rate, while $\text{CVFLDA}_{\lambda=1}$ closely follows with the lowest mean and variance of the absolute deviation from the underlying number of clusters. The sub-index indicates the chosen margin of safety λ. DB and SIL fail to correctly predict the cluster numbers in almost half of all cases, while gap statistics and CH index yield respectable results. This impression is confirmed by the high variance of the absolute difference between the DB and SIL. Interestingly, CH also shows a relatively high variance despite its good success rate. This indicates few but severe estimation errors. Further, the low mean difference of the gap statistics implies symmetric estimations of the number of clusters, whereas DB and SIL permanently underestimate the number of clusters.

Table 1 Mean difference, mean absolute difference, variance of absolute difference, and success rate of the estimated optimal number of clusters from different cluster validation techniques

	Mean Difference	Mean absolute Difference	Variance absolute Difference	Success Rate
$\text{CVFLDA}_{\lambda=0}$	**0.0037**	0.2401	0.7670	**0.8775**
$\text{CVFLDA}_{\lambda=1}$	-0.1275	**0.1898**	**0.3593**	0.8750
$\text{CVFLDA}_{\lambda=2}$	-0.1951	0.2160	0.4132	0.8664
$\text{CVFLDA}_{\lambda=3}$	-0.2324	0.2472	0.5003	0.8596
$\text{CVFLDA}_{\lambda=5}$	-0.3185	0.3235	0.7256	0.8404
$\text{CVFLDA}_{\lambda=10}$	-0.5701	0.5719	1.5362	0.7432
CH	-0.0207	0.4040	1.6105	0.8506
DB	-0.9420	0.9420	1.7065	0.5059
GAP	0.0648	0.2525	0.4110	0.8167
SIL	-0.9284	0.9284	2.1473	0.5870
Sneath	-0.3355	0.3355	0.5686	0.7914

Results are based on 3240 synthetic, random data sets with cluster structure ($k^* > 1$). The best result is printed **bold**



(a) Effect of dimensionality. (b) Effect of the number of clusters.

Fig. 1 Results of different cluster validation techniques over 3240 synthetic, random data sets broken down by dimensionality and number of underlying clusters, aggregated over all other parameters. Note that we use the CVFLDA with $\lambda = 2$

Last, applying Sneath's criterion within the proposed framework also leads to respectable results, demonstrating the good, general applicability of the proposed, simple framework.

For a more detailed analysis, we report the success rate exemplary of $\text{CVFLDA}_{\lambda=2}$ and the benchmark methods with respected dimensions and number of clusters in Fig. 1a and b. Note that for higher dimensions and a smaller number of clusters, the generated clusters are generally more separated. However, as the plots suggest, our method is more stable for a change in these factors. The same plot for $\lambda \in \{0, 1, 2, 3, 5, 10\}$ is given in Appendix A.2 and a table providing individual results for each setting is given in the Supplementary Material.

For the analysis of cases where no clusters are present in the data, i.e., where $k^* = 1$, we compared the proposed framework with the gap statistics in 2, 4 and 8 dimensions. Remember that classical indices cannot be used in this setting since they are mathematically not defined $k^* = 1$. Results are shown in Table 2.

While Sneath's criterion embedded in our framework correctly identified the absence of clusters in 100% of the 300 sample data sets, the gap statistic was successful in 93%. When applying our criterion, the number of correctly identified clusters grows while increasing the margin of safety λ, i.e., the correct cluster number was estimated in only 58% of the

Table 2 Mean difference, mean absolute difference, variance of absolute difference, and success rate of the estimated optimal number of clusters from different cluster validation techniques

Method	Mean Difference	Mean absolute Difference	Variance absolute Difference	Success Rate
$\text{CVFLDA}_{\lambda=0}$	5.016	5.016	38.976	0.580
$\text{CVFLDA}_{\lambda=1}$	1.166	1.166	13.038	0.886
$\text{CVFLDA}_{\lambda=2}$	0.023	0.023	0.062	0.986
$\text{CVFLDA}_{\lambda=3}$	0.003	0.003	0.003	0.996
$\text{CVFLDA}_{\lambda=5}$	**0.000**	**0.000**	**0.000**	**1.000**
$\text{CVFLDA}_{\lambda=10}$	**0.000**	**0.000**	**0.000**	**1.000**
GAP	0.073	0.073	0.075	0.930
Sneath	**0.000**	**0.000**	**0.000**	**1.000**

Results are based on 300 random data sets without cluster structure ($k^* = 1$). Best results are printed **bold**

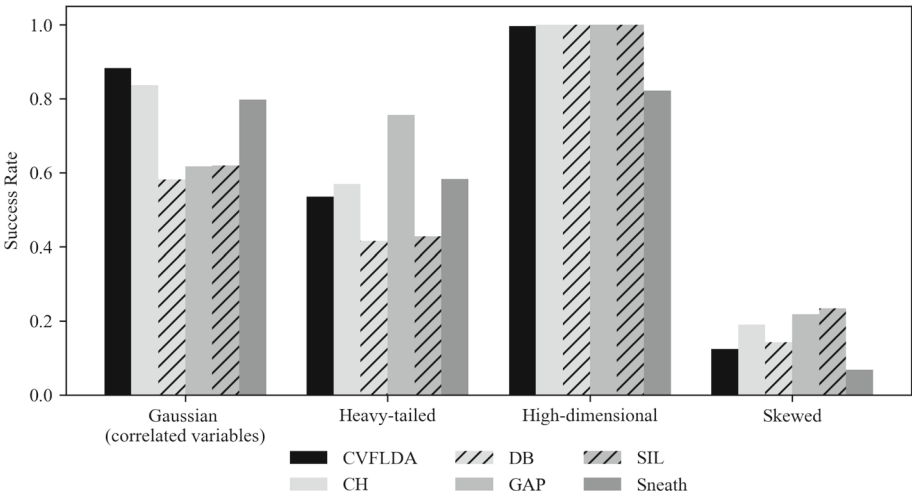


Fig. 2 Results of further cluster types, aggregated over all other parameters. Note that we that we use the CVFLDA with $\lambda = 2$. Further detailed results are reported in the Supplementary Material

datasets for $\text{CVFLDA}_{\lambda=0}$, but in 100% of the datasets for $\text{CVFLDA}_{\lambda=5}$ and $\text{CVFLDA}_{\lambda=10}$. We observe that the margin of safety λ indeed acts as a nuance parameter to control the method's sensitivity, meaning an increasing value of λ decreases the chance of overestimating the number of clusters while reducing its sensitivity. From our simulations on Gaussian clusters with uncorrelated variables, we observe that for this underlying data structure, $\lambda = 2$ is a reasonable choice. The success rate in settings with cluster structure is only slightly smaller than with $\lambda = 1$, but the absence of clusters is detected significantly more reliably. For cluster structures with larger overlaps (e.g., Gaussian clusters with correlated variables, heavy-tailed and skewed clusters), smaller values for λ, i.e., $\lambda = 1$, seem to perform better. These observations reflect a general trade-off in clustering. The risk of detecting clusters in random noise increases when the risk of not detecting (overlapping) clusters decreases. Hence, methods that identify random noise well usually perform worse in detecting clusters, particularly when there are overlaps in the data and vice versa.

Table 3 Characteristics of real-world data sets used in the study

Dataset	# Instances	# Attributes	# Clusters
Ecoli	336	7	8
Glass	214	9	7
Haberman	306	3	2
Iris	150	4	3
Palmer Penguin	344	6	3
Seeds	210	7	3
Transfusion	748	4	2
Vertebral Column	310	6	3
WineQuality Red	1599	8	10
Yeast	1484	8	10

Our method remains competitive on various cluster types, as summarized in Fig. 2 and reported in detail in the Supplementary Material. The CVFLDA performs well with correlated data and delivers consistently good results, closely followed by the Calinski-Harabasz index, which only struggles with a higher number of clusters. For heavy-tailed data, the GAP statistic seems to be superior, followed by the CH index and CVFLDA. Furthermore, our experiments show that one should be careful with the CVFLDA on skewed data, as the success rate is very low in this case. However, the success rates of all other methods also decrease massively, indicating a need for further research on cluster validation with skewed data. Finally, CVFLDA tends to be slightly weaker than its competitors on high-dimensional data but still achieves success rates of $> 90\%$ in our experiments.

To summarize, we recommend performing a careful descriptive analysis of the available data before selecting a cluster validation method. Such an analysis helps to understand the results of the cluster validation indices better and avoids drawing false conclusions from the data. For example, CVFLDA results are trustworthy in Gaussian settings but should be supplemented or replaced in settings with heavy-tailed and skewed data. In general, we recommend using several validation indices simultaneously for more comprehensive results and informed decisions.

4 Application on Real World Data

Supplementary to the simulations, we evaluate our validation approach on different data sets from the University of California - Irvine (UCI) machine learning repository (Dua and Graff, 2017) and the Palmer Penguin dataset.[1] In general, interpreting the cluster validation results from these data sets should be done with caution since they are usually intended for supervised learning and consequently not well adapted for the clustering problem (Arbelaitz et al., 2013). Detailed information about the selected data sets can be found in Table 3.

On each data set, we tested the different approaches 15 times to account for randomness included in the clustering, the construction of C_m and the evaluation of the gap statistic. The established methods are evaluated on the result of a k-means clustering with $k \in [2, k^* + 10]$, where k^* is the real number of clusters. As an input for our validation framework, we use the result of a single clustering with $k = k^* + 10$. Experimental results are displayed in Table 4.

In general, we observe that the performance of the validation methods largely depends on the data set. In detail, we see that the gap statistic performs best with an average mean absolute deviation of 1.980 and an average variance absolute deviation of 0.500 over all datasets. In this setting, our method follows closely behind with an average mean absolute deviation of 2.026 and an average variance absolute deviation of 0.768. Note, however, that the results largely depend on the dataset at hand. Further, observe that the CH has an average variance close to zero, but an average mean absolute deviation of 4.186 over all datasets. Hence, it is very stable but not precise. Also, Sneath's criteria embedded in our framework results in a very low average variance of 0.054 but simultaneously displays a higher average mean absolute deviation of 3.393.

[1] https://allisonhorst.github.io/palmerpenguins

Table 4 Mean absolute difference and variance of absolute of the estimated optimal number of clusters from different cluster validation techniques for real-world data sets

Data set	Method	Mean Absolute Difference	Variance Absolute Difference
Ecoli	CVFLDA	1.533	0.782
	CH	5.000	0.000
	DB	6.000	0.000
	GAP	3.467	1.182
	SIL	7.000	0.000
	Sneath	7.000	0.000
Glass	CVFLDA	3.800	2.960
	CH	3.933	0.062
	DB	5.133	8.382
	GAP	2.000	1.067
	SIL	2.133	0.116
	Sneath	4.400	0.240
Haberman	CVFLDA	0.733	0.196
	CH	2.000	0.000
	DB	1.000	0.000
	GAP	0.000	0.000
	SIL	1.000	0.000
	Sneath	1.000	0.000
Iris	CVFLDA	1.267	0.729
	CH	0.000	0.000
	DB	1.000	0.000
	GAP	3.333	1.156
	SIL	1.000	0.000
	Sneath	1.000	0.000
Palmer Penguin	CVFLDA	1.733	0.196
	CH	8.933	0.062
	DB	8.733	0.196
	GAP	1.000	0.000
	SIL	1.000	0.000
	Sneath	2.000	0.000
Seeds	CVFLDA	0.800	0.427
	CH	0.000	0.000
	DB	0.000	0.000
	GAP	0.000	0.000
	SIL	1.000	0.000
	Sneath	1.600	0.240

Table 4 continued

Data set	Method	Mean Absolute Difference	Variance Absolute Difference
Transfusion	CVFLDA	1.333	0.489
	CH	9.000	0.000
	DB	4.733	14.062
	GAP	0.000	0.000
	SIL	5.333	19.022
	Sneath	1.000	0.000
Vertebral Column	CVFLDA	1.933	0.062
	CH	1.000	0.000
	DB	0.000	0.000
	GAP	4.000	0.400
	SIL	0.000	0.000
	Sneath	1.933	0.062
Winequality Red	CVFLDA	1.800	1.093
	CH	4.000	0.000
	DB	4.000	0.000
	GAP	4.000	0.000
	SIL	4.000	0.000
	Sneath	5.000	0.000
Yeast	CVFLDA	5.333	0.755
	CH	8.000	0.000
	DB	4.333	0.622
	GAP	2.000	1.200
	SIL	5.933	0.996
	Sneath	9.000	0.000

5 Conclusion

In this paper, we presented a new cluster validation technique based on a simple pairwise comparison of clusters and a merge criterion defined on a one-dimensional projection of the data. The used projection is similar to Fisher's Linear Discriminant Analysis, aiming to maximize the ratio of between-variance to within variance in the clusters. However, we emphasize that the proposed framework can be applied to other merge criteria as well. In general, we conclude that the proposed validation technique is especially useful when no knowledge of the underlying number of clusters is available. In cases with no cluster structure in the data, it is able to detect this absence. Otherwise, it returns an improved clustering with an estimate of the optimal number of clusters $\hat{k}^*$ based on the initial clustering. In the paper, we demonstrated the performance of the new cluster validation method on simulated and real-world data and compared it with other well-known validity indices. Last, ready-to-use computer code is provided.

For future research, further improvements of the method, e.g., the development of a kernelized version, seem to be promising. Also, an application within spectral clustering approaches seems possible.

A Ablation Study

In this section, we provide further experimental results of design components of our framework. Specifically, we first discuss the strategy of how the merged cluster C_m is constructed. Then, we demonstrate the effect of the parameter λ in the merge criterion. Finally, we investigate the robustness of the performance of our method to the initial clustering and the performance of our method under overlapping cluster structures.

A.1 Construction of C_m

Figure 3 below visualizes the success rate for different construction principles of the merged cluster C_m in a two-dimensional setting and several degrees of imbalance of the cluster sizes. Specifically, we increase the number of points in C_1 with respect to the total number of samples. We generate two Gaussian clusters with mean $\boldsymbol{\mu}_1 = (0, 0)$ and $\boldsymbol{\mu}_2 = (4, 0)$ and the identity matrix as covariance matrix. The sample size n reflects the total number of objects combined in both clusters, so $n = n_{C_1} + n_{C_2}$. We observe that balancing (random half), meaning that C_m is constructed by selecting 50% of the points randomly (without replacement) from the closest cluster-half (distance to the centroid of the other cluster in the projection) of both clusters, outperforms the other two methods. Especially in situations with unbalanced cluster sizes, the method outperforms its competitors, and hence, we recommend using this construction principle for C_m. Balancing (closest), which selects the closest points in ascending order (by distance), has the worst performance overall, whereas taking the full closest half of each cluster lies in between. Note that the resulting cluster C_m in the latter case does not necessarily consist of 50% of the points stemming from each of the initial clusters as in both former cases.

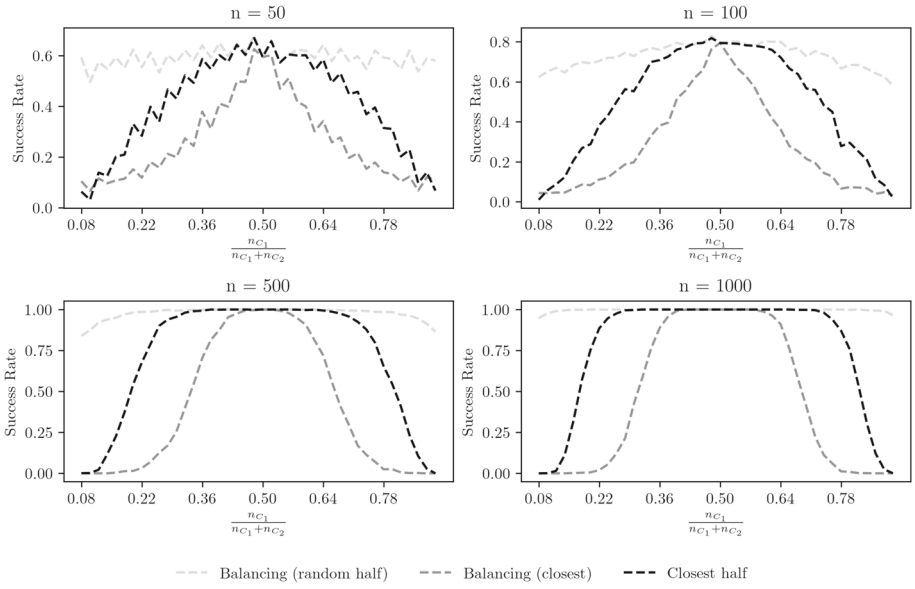


Fig. 3 Success rate of CVFLDA with $\lambda = 2$ for different construction methodologies of C_m over the total number of points and cluster size ratios. All results are averaged out of 1000 simulation runs

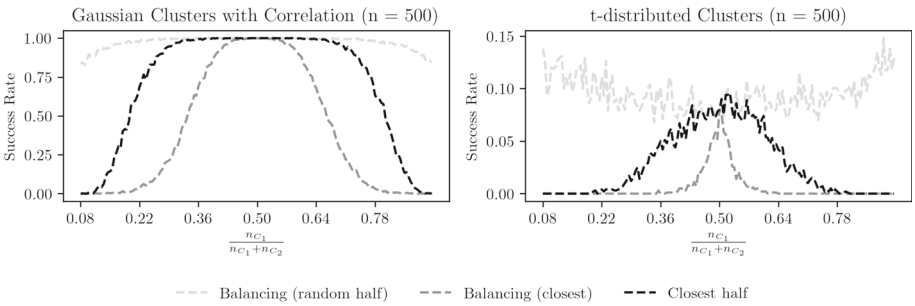
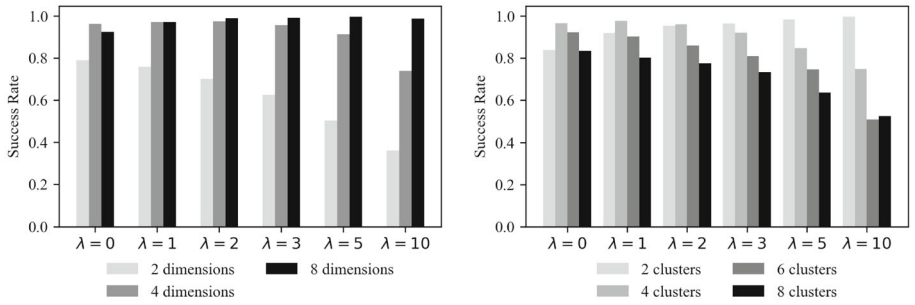


Fig. 4 Success rate of CVFLDA with $\lambda = 2$ for different construction methodologies of C_m over the total number of points and cluster size ratios. All results are averaged out of 1000 simulation runs

To show that this effect is robust for different distributions, we repeated this experiment with Gaussian clusters, with a randomly chosen covariance matrix, and with t-distributed clusters (df = 5). Results are shown in Fig. 4 and are consistent with the findings from standard Gaussian clusters. Note, that performance on t-distributed data decreases, as with heavy-tailed data the two distributions are more overlapping.

A.2 Effect of λ

In Fig. 5, we show the effect of different values of the safety margin parameter λ on synthetic data with Gaussian clusters from Section 3. We observe that increasing values of λ result in a worse performance for lower dimensional data or if many clusters are present. If the number of true clusters increases there are more overlaps in the cluster structure due to our data-generating process. Similarly, in higher dimensions, the clusters are more distinct. This shows once again that larger values of λ perform worse, especially when the cluster structure has overlaps.



(a) Effect of dimensionality. (b) Effect of cluster number.

Fig. 5 Results of CVFLDA with different values of λ over 3240 synthetic, random data sets broken down by dimensionality and number of underlying clusters ($k^* > 1$), aggregated over all other parameters

A.3 Effect of Initial Overestimation

Our proposed framework assumes that the initial clustering has at least as many clusters as the underlying partitioning, i.e., $k \geq k^*$. However, this does not impose a strong restriction

on applying our framework. We tested the effect of different levels of initial overestimation ($k \gg k^*$) on the accuracy for randomly generated multivariate normal clusters in varying dimensions (2, 4, 8), number of clusters (2, 4, 6, 8), and sizes similar to Section 3. As Fig. 6 suggests, the accuracy does not vary significantly with the level of initial overestimation, and hence one can apply our framework as long as the initial clustering overestimates the underlying partitioning to some extent.

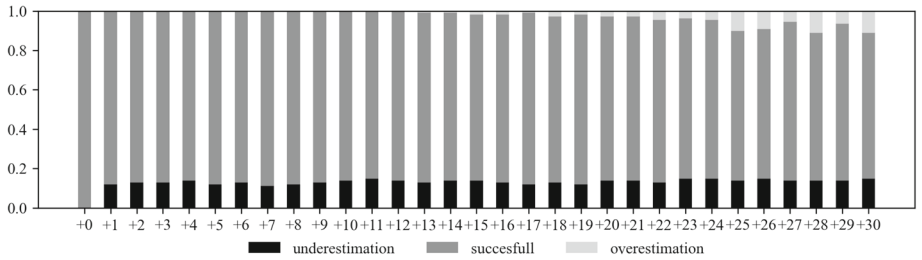
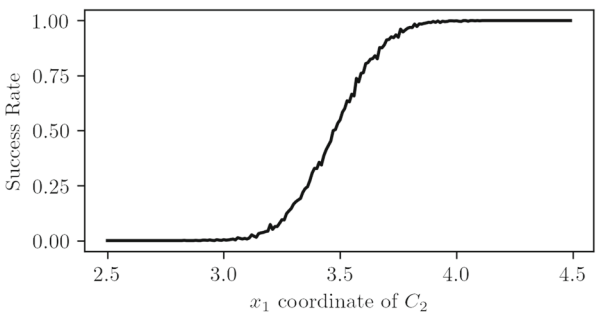


Fig. 6 Accuracy of the CVFLDA with $\lambda = 2$ for different levels of initial overestimation

A.4 Effect of Overlapping Cluster Structure

Naturally, when the "true" clusters are overlapping (meaning that the corresponding densities significantly overlap) finding the underlying true cluster structure and validating a clustering is challenging. We observe that the performance of our proposed method increases if the "true" clusters are more distinct, i.e., less overlapping. We generated two two-dimensional standard Gaussian clusters C_1 and C_2, where we keep the mean of C_1 fixed in the origin and the mean of C_2 moves along the x_1-axis. We provide the results in Fig. 7. Intuitively, if two clusters are distinct there is a gap between them consisting of low density of points. In this case, the variance of the merged cluster is expected to be larger, which is why they are more easily identified as true clusters.

Fig. 7 Success rate of CVFLDA for two Gaussian clusters where their overlap is controlled by the location of the center of C_2. All results are averaged out of 1000 simulation runs



B Code Availability

Ready-to-use computer code (Python) is available at the following web address: https://github.com/NoraSchneider/CVFLDA.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00357-024-09481-3.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availibility The data analyzed during the current study are made available by Dua and Graff (2017) and under https://allisonhorst.github.io/palmerpenguins/.

Declarations

Ethical Approval The authors comply with all ethical standards. No research involving Human Participants and/or Animals was conducted.

Conflict of Interest The authors declare no competing interests.

References

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition, 46*(1), 243–256.

Bailey, T. A., & Dubes, R. (1982). Cluster validity profiles. *Pattern Recognition, 15*(2), 61–83.

Baker, F. B., & Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association, 70*(349), 31–38.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods, 3*(1), 1–27.

Cerdeira, J., Martins, M., & Silva, P. (2012). A combinatorial approach to assess the separability of clusters. *Journal of Classification, 29*, 7–22.

Dangl, R., & Leisch, F. (2019). Effects of resampling in determining the number of clusters in a data set. *Journal of Classification 37*.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1*(2), 224–227.

Delaigle, A., Hall, P., & Pham, T. (2019). Clustering functional data into groups by using projections. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81*(2), 271–304.

Dua, D., & Graff, C. (2017). UCI machine learning repository.

Dubes, R. C. (1987). How many clusters are best? - An experiment. *Pattern Recognition, 20*(6), 645–663.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics, 7*(2), 179–188.

Fu, W., & Perry, P. O. (2020). Estimating the number of clusters using cross-validation. *Journal of Computational and Graphical Statistics, 29*(1), 162–173.

Gates, M. A., & Hansell, R. I. C. (1983). On the distinctness of clusters. *Journal of Theoretical Biology, 101*(2), 263–273.

Geng, J., Bhattacharya, A., & Pati, D. (2019). Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association, 114*(526), 893–905.

Gordon, A. D. (1998). Cluster validation. In C. Hayashi, K. Yajima, H.-H. Bock, N. Ohsumi, Y. Tanaka, & Y. Baba (Eds.), *Data science, classification, and related methods, Tokyo* (pp. 22–39). Springer Japan.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems 17*.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics 21*(15), 3201–3212.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction* (2 ed.). Springer.
Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters 64*, 53–62. Philosophical Aspects of Pattern Recognition.
Hennig, C. (2022). An empirical comparison and characterisation of nine popular clustering methods. *Advances in Data Analysis and Classification*.
Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (2015). *Handbook of cluster analysis (1th* (edition). New York: Chapman and Hall/CRC.
Ingrassia, S., & Punzo, A. (2020). Cluster validation for mixtures of regressions via the total sum of squares decomposition. *Journal of Classification 37*(2), 526–547.
Li, J. (2005). Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics, 14*(3), 547–568.
Liu, Y., Hayes, D. N., Nobel, A., & Marron, J. S. (2008). Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association, 103*(483), 1281–1293.
Melnykov, V. (2016). Merging mixture components for clustering through pairwise overlap. *Journal of Computational and Graphical Statistics, 25*(1), 66–90.
Melnykov, V., & Michael, S. (2020). Clustering large datasets by merging k-means solutions. *Journal of Classification, 37*, 97–123.
Milligan, G., & Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159–179.
Omran, M. G. H., Engelbrecht, A. P., & Salman, A. (2011). An overview of clustering methods. *Intelligent Data Analysis, 11*, 583–605.
Peck, R., Fisher, L., & Ness, J. V. (1989). Approximate confidence intervals for the number of clusters. *Journal of the American Statistical Association, 84*(405), 184–191.
Peña, D., & Prieto, F. J. (2001). Cluster identification using projections. *Journal of the American Statistical Association, 96*(456), 1433–1445.
Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846–850.
Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. M. (2011). Internal versus external cluster validation indexes. *International Journal of computers and communications, 5*(1), 27–34.
Rossbroich, J., Durieux, J., & Wilderjans, T. F. (2022). Model selection strategies for determining the optimal number of overlapping clusters in additive overlapping partitional clustering. *Journal of Classification*.
Rousseeuw, P. J., & Kaufman, L. (1990). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.
Salvador, S., & Chan, P. (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *16th IEEE international conference on tools with artificial intelligence*, pp. 576–584. IEEE.
Sneath, P. (1977). A method for testing the distinctness of clusters: A test of the disjunction of two clusters in Euclidean space as measured by their overlap. *Journal of the International Association for Mathematical Geology, 9*(2), 123–143.
Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association, 98*(463), 750–763.
Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411–423.
Ullmann, T., Hennig, C., & Boulesteix, A.-L. (2022). Validation of cluster analysis results on validation data: A systematic framework. *WIREs Data Mining and Knowledge Discovery, 12*(3), e1444.
von Luxburg, U. (2010). Clustering stability: An overview. *Foundations and Trends in Machine Learning, 2*(3), 235–274.
Wierzchoń, S. T. (2018). *Modern algorithms of cluster analysis*. Springer International Publishing.
Wiwie, C., Baumbach, J., & Röttger, R. (2015). Comparing the performance of biomedical clustering methods. *Nature Methods, 12*, 1033–1038.
Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645–678.