Received: 1 September 2021 | Revised: 24 November 2021 | Accepted: 8 December 2021

DOI: 10.1002/wics.1575

ADVANCED REVIEW

WIREs COMPUTATIONAL STATISTICS WILEY

Stability estimation for unsupervised clustering: A review

Tianmou Liu[1] | **Han Yu**[2] | **Rachael Hageman Blair**[3]

[1]Institute for Artificial Intelligence and Data Science, State University of New York at Buffalo, Buffalo, New York, USA

[2]Roswell Park Comprehensive Cancer Center, Buffalo, New York, USA

[3]Department of Biostatistics, Institute for Artificial Intelligence and Data Science, State University of New York at Buffalo, Buffalo, New York, USA

Correspondence
Rachael Hageman Blair, Department of Biostatistics, Institute for Artificial Intelligence and Data Science, State University of New York at Buffalo, Buffalo, NY 14260-1660, USA.
Email: hageman@buffalo.edu

Funding information
Rachael Hageman Blair was supported by the NSF DMS 1557589. Han Yu was supported by the National Cancer Institute Cancer Center Support (Grant P30CA016056) and National Cancer Institute IOTN Moonshot (Grant U24CA232979).

Edited by: Nicole Lazar, Commissioning Editor and David Scott, Co-Editor-in-Chief

[Correction added on 21 January 2022, after first online publication: The copyright line was changed.]

Abstract

Cluster analysis remains one of the most challenging yet fundamental tasks in unsupervised learning. This is due in part to the fact that there are no labels or gold standards by which performance can be measured. Moreover, the wide range of clustering methods available is governed by different objective functions, different parameters, and dissimilarity measures. The purpose of clustering is versatile, often playing critical roles in the early stages of exploratory data analysis and as an endpoint for knowledge and discovery. Thus, understanding the quality of a clustering is of critical importance. The concept of *stability* has emerged as a strategy for assessing the performance and reproducibility of data clustering. The key idea is to produce perturbed data sets that are very close to the original, and cluster them. If the clustering is stable, then the clusters from the original data will be preserved in the perturbed data clustering. The nature of the perturbation, and the methods for quantifying similarity between clusterings, are nontrivial, and ultimately what distinguishes many of the stability estimation methods apart. In this review, we provide an overview of the very active research area of cluster stability estimation and discuss some of the open questions and challenges that remain in the field.

This article is categorized under:

Statistical Learning and Exploratory Methods of the Data Sciences > Clustering and Classification

KEYWORDS

clustering, model selection, resampling, stability, unsupervised learning, validation

1 | INTRODUCTION

The overall objective of clustering is to form groups of items that are highly similar to each other, and dissimilar to other groups (Kaufman & Rousseeuw, 2009; Sarle, 1990). The broad utility of clustering has charged the research community and led to a continuous surge of clustering techniques that have been developed in statistics, machine learning, pattern recognition, and many other fields. Depending on the problem at hand, there are different branches of clustering, including unsupervised, supervised, semi-supervised (Bair, 2013), fuzzy (Yang, 1993), and soft (Peters et al., 2013).

Of these, the clustering of unlabeled data is the most common, yet challenging and sometimes illusive, areas of unsupervised learning (Grira et al., 2004; Steinbach et al., 2004). At the forefront of these challenges is the fact that there are no labels in unsupervised clustering, which makes performance measures and the assessment of cluster quality and reproducibility problematic.

There are several other factors that complicate the clustering process. Although stability does not attempt to address all of these issues collectively, we briefly outline other points of variability and uncertainty that occur in the clustering process. For a more in-depth treatment, see Jain et al. (1999), Shirkhorshidi et al. (2014), Steinbach et al. (2004), and Wang et al. (2012). Unsupervised clustering requires subjective decisions to be made by the investigator in the selection of measures that would define how *similar* items are. Often this decision is guided by the type of data that is being clustered, for example, continuous, binary, categorical, or a mixture thereof, and convenience of default built-in dissimilarity measures in clustering routine implementations. Choi et al. (2010) put this issue into perspective when they examined 76 measures for similarity of binary data alone. Arguably, there are at least as many for continuous data. Another point of uncertainty is with the investigator's selection of the clustering algorithm. Different clustering algorithms have different optimization functions and therefore differ in how they establish *similarity* within a grouping (Rand, 1971; Rokach & Maimon, 2005). Estivill-Castro (2002) attribute a large number of clustering algorithms as a reflection of the fact that a *cluster* cannot be precisely defined. Once an algorithm is selected, often parameters that determine the number of clusters for a given algorithm must be selected. For example, in *k*-means, the number of clusters (k) has to be prespecified, and in hierarchical clustering, the user determines the height at which the dendrogram is cut to result in a clustering. Surveys of clustering algorithms and dissimilarity measures have been performed for specific application domains, for example, magnetoencephalography (Guggenmos et al., 2018), X-rays (Iwasaki et al., 2017), structural chemistry (Adamson & Bush, 1975), image retrieval, and segmentation (Puzicha et al., 1999), among others. Shirkhorshidi et al. (2015) conducted a comparison study on 12 frequently used similarity measures for continuous data. In this study, the adaptivity of different measures to different datasets and different clustering methods was examined (Shirkhorshidi et al., 2015). Their findings indicate that clustering algorithms link items together to form groups based on similarity, but the process of linking (sometimes known as linkage) is driven by different objective functions and model assumptions inherent to the selected clustering method. Due to the combinatorial explosion of linking different dissimilarity measures with clustering algorithms, a comprehensive survey would be intractable. Moreover, the results would likely vary according to the dataset or simulation.

Despite these challenges and limitations, clustering remains a main staple in analytics across fields. In an analytic pipeline, clustering can be used as a first step when used in connection with exploratory data analysis (Tukey, 1977), for example, subgroup identification (Dubes & Jain, 1980), identifying nearest neighbors, imputation, outlier removal, and anomaly detection (Ahmed et al., 2016; Chandola et al., 2009). Clustering can also be used as an endpoint, for the purpose of knowledge and discovery with an aim to provide insights into a domain-specific application. The versatile nature of clustering has inspired a range of methodological developments around the topics of reproducibility and performance.

1.1 | Why clustering stability?

With so many options available to the investigator and no labels or gold standard; there are very few heuristics available to assess cluster quality (e.g., Gnanadesikan et al., 1977; Rousseeuw, 1987). These simple measures do not account for the uncertainty and variation in the data itself that could be due to sampling or measurement error. Data is finite and represents a sample from an underlying population; with distributions that are unknown. In small datasets, structure, and patterns can be easily induced by sampling. On the other hand, real structure and patterns can be missed because of insufficient data for detection. This phenomenon broadly applies to datasets with different structures and sampling processes. Characterizing the reproducibility of a clustering is an important concept and the overarching aim of stability analysis.

A natural solution is to validate a clustering by taking an independent sample, or many independent samples, from the underlying population. However, this is rarely possible due to limitations such as time and expense. As an alternative to collecting new data for validation, cluster stability methods rely on perturbations to the original dataset. Stability measures capture how well partitions and clusters are preserved under perturbations to the original dataset. The underlying premise is that a good clustering of the data will be reproduced over an ensemble of perturbed datasets that are nearly identical to the original data. Stability measures the quality of preservation of clustering solutions across perturbed datasets.

Stability has enjoyed popularity in a range of fields and application areas. For example, clinical studies have included stability to identify stable disease clusters based on phenotype (Newby et al., 2014), to characterize subgroups of disease in longitudinal studies (Loza, Adcock, et al., 2016; Loza, Djukanovic, et al., 2016) and to identify stable clusters of symptoms to promote improved patient care (Kim et al., 2005). Stability has also been applied to high-dimensional omics studies that are characteristically noisy and challenging to cluster, including gene expression data from microarrays (Bertoni & Valentini, 2005; Giancarlo & Utro, 2012; Giurcăneanu & Tăbuş, 2004; Kerr & Churchill, 2001; Smolkin & Ghosh, 2003a, 2003b), multilayer omics data (Hidalgo & Ma, 2018) and single-cell RNA data (Peyvandipour et al., 2020; Tang et al., 2021; Zhang et al., 2020). Stability has also been used in the field of marketing analysis and segmentation (Dolnicar & Lazarevski, 2009' Müller & Hamm, 2014; Hajibaba et al., 2019), e-tourism (Dolnicar, 2002, 2020), for vehicular networks (Abboud & Zhuang, 2015; Mammu et al., 2013), structural chemistry (Erdmann & Schwarz, 2007), among others. In fact, any field which utilizes clustering can effectively utilize stability to characterize and improve solutions, thus the impact on applications is likely to continue to increase.

Different methodologies for cluster stability have emerged over the past 30+ years and used to offset some of the clustering challenges and limitations described above. These stability methodologies differ fundamentally in how small perturbations to the original dataset are generated, and how similarity between clustering is measured. Some foundational issues that have been addressed with stability include an estimate of confidence to an item's membership to a cluster, an estimate of confidence to cluster, and an overall estimate of confidence for a clustering of a dataset. Synonymous with cluster stability is its utility in the selection of the optimal number of clusters, which herein we refer to as *model selection*. Although not all stability methodologies lend themselves to the problem of model selection, we discuss stability methods that have and have not been developed for this purpose, and later discuss some controversy around this topic. To the author's knowledge, there has not been a review on cluster stability in over a decade (Von Luxburg, 2009). The review by Von Luxburg (2009) had a restricted focus on the stability of *k*-means. This review covers the fundamentals of cluster stability approaches, advancements, and open challenges that exist in this area.

2 | A CASE STUDY OF CLUSTERING STABILITY

In order to motivate the broad range of stability methods, we will demonstrate some of the core concepts and output from the various approaches. For simplicity, we consider a classic dataset describing three species of iris flowers: setosa, versicolor, and virginica (Fisher, 1936). The data was taken from the machine learning repository (Asuncion & Newman, 2007) and contains 150 observations (species) and four continuous variables describing measurements of the

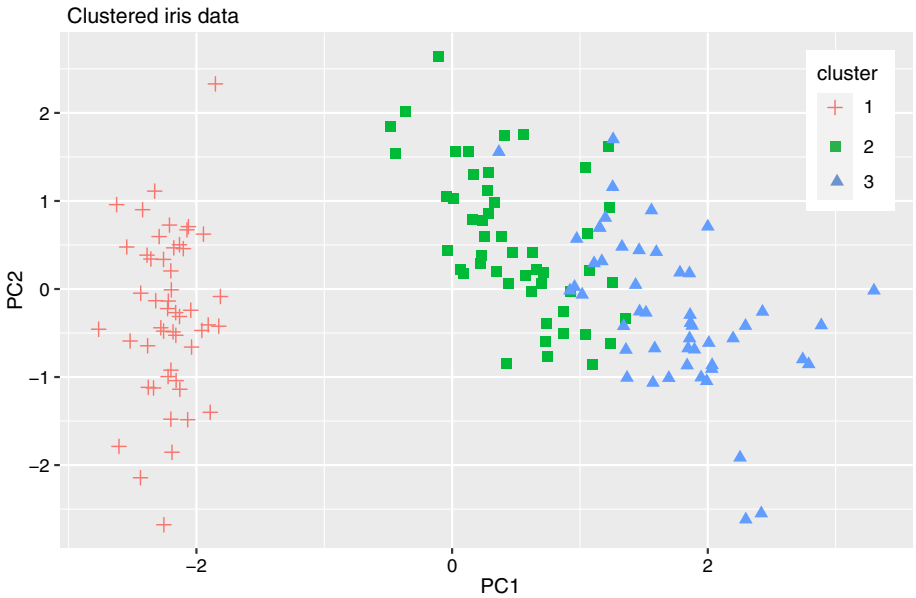

FIGURE 1 The first two principal components of the iris data, which describe 84% of the variation. The clusters found using *k*-means ($k = 3$) are indicated by shape and color

flower attributes. The iris data is one of the most widely used datasets for classification and pattern recognition. One reason for this is the separation characteristics of the clusters. There is one well-separated cluster and two clusters the are more overlapping (Figure 1). These clusters were inferred using k-means ($k = 3$) and align reasonably well with the species labels (adjusted rand index = 0.62; Steinley, 2004), with setosa well-separated, and versicolor and virginica overlapping.

In the sections that follow, we will use the iris data to demonstrate some of the clustering concepts and methods. Our selection of methods is based on the availability of packages for estimating stability in the R programming language (Table 1). Source code is available in the Supporting Information. The objective of this case study is to use stability to characterize the iris clustering. Importantly, the case study is not intended to be a comparison of cluster stability methods, which would require a comprehensive panel of benchmarking datasets and pairing with clustering algorithms. We also emphasize that although there are three species of flowers, it is reasonable to expect to see k selected as two or three, due to the cluster separation. In practice, we would not have label information to guide us on the number of groups in a truly unsupervised setting.

3 | APPROACHES TO CLUSTERING STABILITY

Several methods have been developed for cluster stability. Organizationally, we have broken these methods down into the following three categories: resampling for stability estimation (Section 3.1), cluster validation via data splitting and

TABLE 1 Stability methods and implementations for unsupervised clustering

Reference	Stability item	Model selection	Clustering method	Implementation
Bootstrapping				
Bootstrap technique by Jain and Moreau (1987)	Overall	Yes	k-means, HC	Not found
ANOVA method by Kerr and Churchill (2001)	Overall	No	model-based	MAANOVA (R)
BagClust1& 2 by Dudoit and Fridlyand (2003)	Observation, cluster, overall	Yes	PAM	Clue (R)
Cluster-wise assessment by Hennig (2007)	Cluster	Yes	general	fpc (R)
Clustering instability by Fang and Wang (2012)	Overall	Yes	k-means	fpc (R)
Bootstrap Jaccard by Yu et al. (2019)	Observation, cluster, overall	Yes	k-means	Bootcluster (R)
Cluster validation via data splitting and subsampling				
Clest by Dudoit and Fridlyand (2002)	Overall	Yes	general	RSKC (R)
Figure of Merit by Levine and Domany (2001)	Overall	Yes	general	Not found
Model explorer algorithm (Ben-Hur et al., 2002)	Overall	Yes	general	Not found
Stability-based validation by Lange et al. (2004)	Overall	Yes	general	Not found
Prediction strength by Tibshirani and Walther (2005)	Observation, overall	Yes	k-means	fpc (R)
Alternative methods				
Loevinger method (Bertrand & Mufti, 2006)	Overall, cluster-level	Yes	k-means	Not found
Matrix manipulation (Steinley, 2008)	Cluster, overall	Yes	k-means	Not found
Optimal transport alignment (Li et al., 2019)	Observation, cluster, overall	Yes	general	OTclust (R)

subsampling (Section 3.2), and alternative methods that do not adhere to these classic approaches (Section 3.3). These approaches are also summarized in Table 1 together with known software implementations.

As a starting point, we introduce some common themes across these methods, as well as some important terminology. Notably, the majority of methods can be classified as resampling-based (Section 3.1) and validation-based (Section 3.2). A schematic depicting the general estimation process is shown in Figure 3. The distinguishing feature between these branches of approaches is how the perturbed datasets (second layers of Figure 3), by which stability is estimated, are generated from the original data, $X^0 \in \mathbf{R}^{N\times p}$.

Resampling (Figure 2a,b) aims to construct data replications through random sampling of the data. Popular strategies include bootstrapping (Efron & Tibshirani, 1994) and subsampling (Politis et al., 1999). Both are used widely in statistics and have desirable asymptotic properties. In the context of cluster stability, the bootstrap procedures resample the data with replacement, thereby producing datasets of identical size to the original. On the contrary, subsampling for cluster stability is the random sampling of the data without replacement to create a dataset of smaller size. These datasets, $\{X^1, X^2, ..., X^B\}$, are different representations of perturbed data. Each dataset is clustered to generate a partitions, $\{P^1, P^2, ..., P^B\}$, that defines a set of clusters. From this stage, we will discuss stability defined through consensus style aggregation (Figure 3a) and through comparison of membership changes between bootstrapped partitions and the original data partition (Figure 3b). In the consensus style clustering, we will focus our discussion on methods with output measures of stability that can provide guidance on cluster quality.

The validation-based approaches proceed in a similar way, but cast the problem into a supervised framework. Specifically, the data is usually split into a learning and validation set. Figure 3c shows a simple example where the data is split and clustered. The validation set is then projected onto the partitions generated by the other learning set. High stability measurements suggest that the clustering arising from the projection onto the learning set partitions is nearly identical to the clustering of the validation set.

Both resampling and validation approaches capture the degree of *similarly* (or lack thereof) between clusterings. How similarity is defined is another distinguishing factor between stability estimation methods. The problem of comparing clusters from different partitions is in itself a major challenge. For example, if the data is clustered, and a resampled dataset is clustered, quantifying how similar these clusterings are is non-trivial. Some stability estimation methods require a mapping between clusterings, followed by the comparisons of clusters. This mapping will most likely not be perfect for all clusters, unless the clusterings being compared are identical. Once the mapping is performed, the cluster compositions can be compared with a dissimilarity measures, for example, Jaccard coefficient. An alternative to remapping is to assess the changes in pair-wise membership of the items being clustered. A clustering can be represented mathematically using a binary *co-membership matrix* with entries of 1 if items i and j belong to the same cluster,

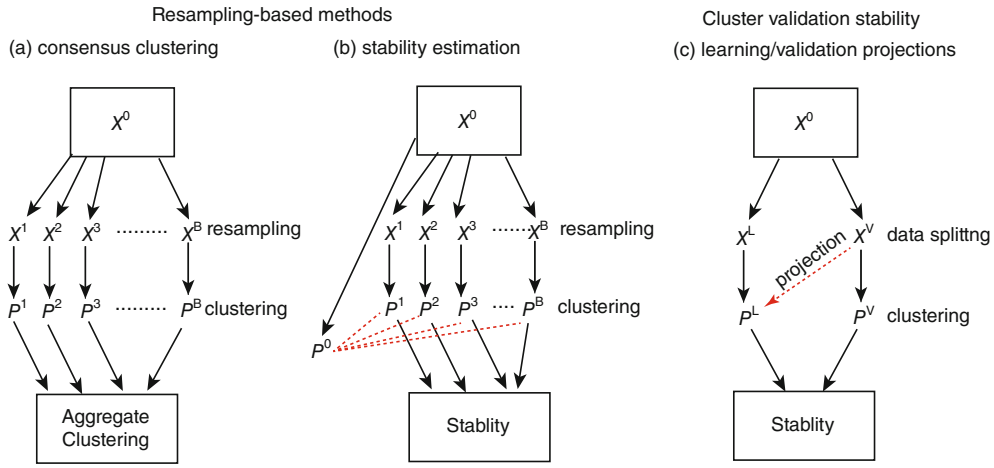

FIGURE 2 Some common approaches to cluster stability are depicted. (a) Data is resampled and clustered. Select consensus methods aggregate clustering partitions from bootstrap replications, P^b, to form a clustering that is often accompanied by a measure of stability. (b) Another resampling approach creates comparisons between the original clustering and the bootstrap replication clusterings (red dotted lines) to derive a measure of stability for the original data, X^0. (c) A cluster validation approach divides data into a learning set, X^L, and a validation/test set, X^V. Cluster partitions P^L and P^V are obtained, and often the validation data is projected onto the learning set partitions (red dotted arrow) to produce a clustering, which is then compared to the clustering obtained with P^V

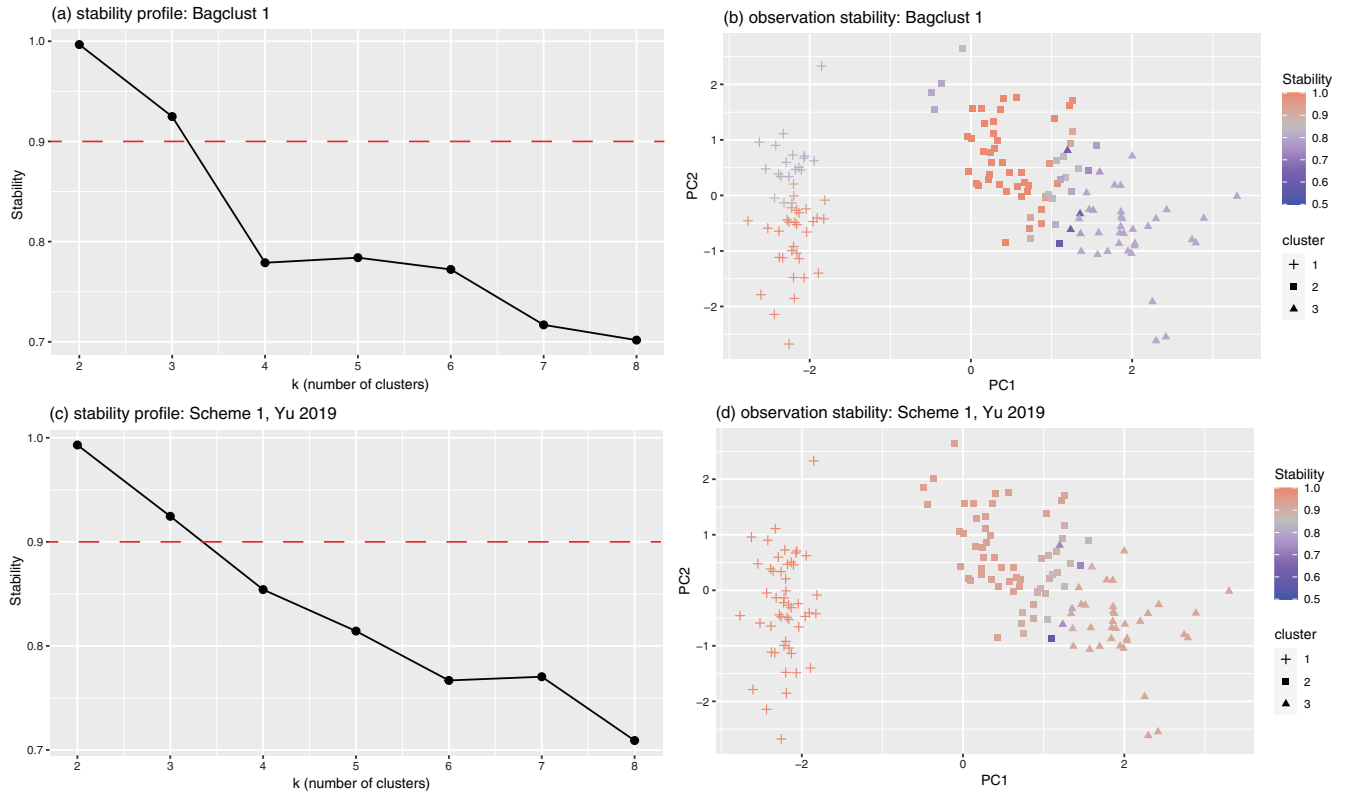

FIGURE 3 The stability of the iris data as characterized using resampling methods. (a) Stability estimates were generated using Bagclust 1 for different values of k (Dudoit & Fridlyand, 2003). (b) The observation level stability generated by Bagclust for $k = 3$. (c) Stability estimates generated using scheme 1 of the approach by Yu et al. (2019) for different values of k. (d) The observation level stability generated by scheme 1 for $k = 3$

and 0 otherwise. Changes in this matrix capture the stability of a clusterings at the level of the individual items being clustered, and can be aggregated to a cluster measure, or an overall measure of stability.

3.1 | Resampling for stability estimation

Bootstrapping is a simple procedure that enables the generation of replicate datasets of the same size and is relatively easy to implement. For a dataset $X^0 \in \mathbf{R}^{N \times p}$ with N observations, the data are resampled with replacement to generate bootstrap replications, $\{X^1, X^2, \ldots X^B\}$, that are the same size as the data (Efron & Tibshirani, 1994). Inherently within a given bootstrap replication, an observation x_i can occur once, multiple times, or not at all. The bootstrap has been applied to the area of clustering stability in rather unique ways.

Felsenstein (1985) developed one of the first approaches to bootstrap based clustering when taking a consensus style approach to the inference of phylogenetic trees, which inherently have hierarchical dendrogram structure. Although not stable per se, the early application sought to quantify uncertainty in dendrogram structures. The bootstrap was used in this context to identify high occurring branches (e.g., 95% of the trees). Jain and Moreau (1987) combined the bootstrap method with the Davies and Bouldin Criterion, which is a function of the cluster dispersion and between cluster separation. Briefly, the within-cluster dispersion is defined as a function of the sum of distances from every point in the cluster to the cluster center divided by the number of points in the cluster, and the between-cluster dispersion is defined as a function of the distance between two clusters centers. Jain and Moreau (1987) bootstrap this stability measure and average overall k clusters. The emphasis of their work was to model selection for the optimal value of k (number of clusters) that provided the most stable partitions of the data. The optimal k was identified as the smallest varying measure that minimizes the criterion, thereby reflecting the most stable clusterings. Notably, this application is amendable to more general clustering algorithms although examples were limited to k-means and hierarchical clustering with

various linkage functions. The statistics were also shown to be effective for the comparisons of clustering algorithms (Jain & Moreau, 1987).

Leisch (1999) developed a unique combination of partitioning and hierarchical clustering methods. This approach applies k-means to B bootstrap replication of the dataset to *stabalize* the cluster centers. These $B \times k$ centers are the primary output of interest from the k-means algorithm. The collection of centers serves as input to a hierarchical clustering routine, and the original data points are mapped to the closest center. The dendrogram is cut to create k clusters, and the mapped data points that align with hierarchical assignment are assigned accordingly. A rationalization of this approach is that by using hierarchical clustering on the bootstrapped centers, the concentration shifts to the centers, and potentially reduces background noise. The unique integration of hierarchical clustering and k-means also captures some of the advantages of these methods, and in some ways, lessens their limitations. For example, hierarchical clustering is inherently flexible with respect to definitions of dissimilarity and linkage, but can be computationally intensive. On the other hand, k-means uses a Euclidean distance to measure the distance of observation to center and is sensitive to random initializations, but is fast computationally. Similar to Felsenstein (1985), this approach does not provide an output measurement of stability, but rather aims to produce a more stable clustering.

A model-based approach assigning confidence to clusters was developed by Kerr and Churchill (2001) that aimed specifically at resampling residuals from ANOVA models for microarray studies. The focus was k-means although the method and application is applicable to more general settings; provided that there is a suitable experimental design. The re-sampled residuals are used to estimate new gene abundance measures, and these are then clustered and compared to the original data clustering. Microarray studies have been known to produce exceptionally noisy data, and this approach reveals whether or not the clusters generated from the data are robust to noise. Resampling from the ANOVA error distribution in this manner also enables model adjustments for important covariates, which other model-free methods described in this review cannot do explicitly. There have been other stability methods proposed specifically in the area of gene expression studies such as an iterative method for coupled two-way clustering (Getz et al., 2000), bootstrapping approaches for algorithm selection (Yeung et al., 2001), consensus clustering (Monti et al., 2003), and stability estimates that account for chip design (Smolkin & Ghosh, 2003a, 2003b).

Dudoit and Fridlyand (2003) developed two resampling methods, known as *Bagclust1* and *Bagclust2*, that aggregate over bootstrap replications of the data. In *Bagclust1*, each bootstrap sample is clustered by a user-specified clustering algorithm. The bootstrap cluster labels have to first be aligned with the clusters in the original data that maximize the overlap of their observations. Note that the mapping of clusters between the re-sampled data and the original data will be imperfect, and vary across the bootstrap replications of the data. In *Bagclust2*, a new dissimilarity matrix $\boldsymbol{M}$ is formed. Each entry $\boldsymbol{M}$ is defined as $1 - a_{ij}/m_{ij}$ where m_{ij} is the number of times both points i and j appeared in the same bootstrap dataset. And a_{ij} is defined as the number of times both points i and j are clustered into the same cluster. Note that $0 \leq a_{ij} \leq m_{ij} \leq B$ where B is the total number of bootstrap datasets generated. However, for the dissimilarity matrix $\boldsymbol{M}_{ij}$ to be a well-defined distance measure, it needs to satisfy $\boldsymbol{M}_{ij} \neq 0$ if $i \neq j$. A limitation of *Bagclust2* is that it is not guaranteed to satisfy this property. It is possible for two different point i and j to be clustered into the same cluster every time they both appears in the bootstrap dataset making $\boldsymbol{M}_{ij} = 1 - a_{ij}/m_{ij} = 1 - 1 = 0$. The algorithm performed well with the use of PAM clustering (Kaufman & Rousseeuw, 2009), and was found to be less sensitive to noise when applied to high-dimensional microarray data.

The resampling methods *Bagclust1* and *Bagclust2* can also evaluate stability at the level of the observation (Dudoit & Fridlyand, 2003). However, the method relies on the projection of resampled clusters to those in the original dataset. This projection can be an issue when a cluster is broken down into multiple smaller clusters in a resampled clustering, or when smaller clusters merge together. For example, when a cluster does not show in a resampled solution, the method would ignore the cluster, which may lead to overestimation of clustering stability. A similar strategy was adopted by Hennig (2007) in 2007 to assess a cluster-wise measure of stability. In that setting, the mapping is performed by identifying the most similar cluster, via Jaccard coefficient, when comparing the data clusters to the clusters in bootstrap replications. These Jaccard coefficients are then aggregated to produce a measure of stability for each cluster. An immediate advantage of this approach is that there could be scenarios where there is good structure as indicated by highly stably clusters, and some clusters that are unstable and could be disregarded. An overall measure of stability would average this out and important grouping could be left undiscovered.

These limitations that arise from projection can be addressed by using the change in pairwise co-membership. Fang and Wang (2012) developed a stability approach that leveraged co-memberships for the purpose of estimating overall stability in order to perform model selection. The algorithm generates B pairs of bootstrap datasets resampled from the original data (the actual number of bootstrap dataset generated is therefore $2B$). The same clustering method and

clustering parameters are applied on each pair of bootstrap dataset X_b and $\widetilde{X}_b$, $1 \le b \le B$. Cluster instability is calculated as a function of the number of pairs of points (x_i, x_j) in X_0 belong to the same clustering in X_b but not in $\widetilde{X}_b$, or do not belong to the same cluster in X_b but do in $\widetilde{X}$. The clustering instability is averaged over all B bootstrap pairs and the optimal k is chosen where the clustering instability is minimized. Fang and Wang (2012) developed two additional measurements, one is the standard error of clustering instability and the other is instability converging path. The estimated standard error algorithm first generates C bootstrap datasets X_c, $1 \le c \le C$. For each X_c, $1 \le c \le C$, the clustering instability is calculated according to the algorithm above (as a result, $C \times 2B$ bootstrap datasets are generated). The sample standard error of the C instabilities are estimated from the clustering instability. The instability path is the instabilities for a certain fixed k estimated by increasing the number of bootstrap dataset B gradually. The convergence of instability paths for different k's can be used for model selection.

Recently, Yu et al. (2019) proposed a stability method based on Jaccard coefficient and bootstrap. Two fundamentally different algorithmic schemes were established based on the Jaccard dissimilarity of pairwise membership changes between bootstrap replications of the data. The differences between these two schemes lies with the assignment of a *reference set*, by which the bootstrapped clusterings are compared. Scheme 1 generates clusterings from each of the B bootstrap samples. Each of the bootstrap clusterings is compared with the original dataset to generate a Jaccard-based measure of similarity using co-membership changes between clusterings. Similar to Dudoit and Fridlyand (2003), stability can be derived at the level of the observation. Additional levels of stability can be derived at the cluster level and as a measure of overall performance. In contrast, Scheme 2 does not use the original clustering as a reference, and alternatively performs an exhaustive search for an optimal clustering by systematically assigning the reference as the bootstrap clusterings. Scheme 2 is designed for scenarios when the researcher lacks confidence in the structural integrity of the original dataset. Yu et al. (2019) used these approaches for the derivation of a stability profile that captures stability across a range of k values, which can be used for the determination of the number of clusters. Applications are primarily to k-means with potential extensions to other clustering methods.

Despite its simplicity, bootstrapping clustering stability has common limitations that are the direct result of resampling with replacement. Resampling with replacement can lead to bias in the estimates that are being bootstrapped. In fact, the average number of distinct observations in each bootstrap replication is approximately 0.632 (Efron & Tibshirani, 1997). In center-based methods, such as k-means, the duplications of observations create bias in the estimation of the centers. In methods that rely on a dissimilarity matrix, the repetition of observations creates an analogous set off issues with construction. Most methods do not explicitly describe the solution to overcome this problem. A popular idea is to *jitter* the observations, which is the addition of a negligible amount of random noise so as to make the observations in the bootstrap replications unique. The *jitter* approach effectively amounts to two forms of perturbation to the original dataset, the re-sampling and the addition of random noise. Hennig (2007) combines bootstrap with jittering for cluster stability estimates. Although this may ease the computational issues resulting from observations being represented multiple times, for example, calculation of a dissimilarity matrix, it does not correct the issues with bias. Additional bias can be incurred when cluster memberships are re-projected onto the bootstrap centers (Yu et al., 2019); due to the fact that observations that may have participated one or more times in cluster parameter estimation (e.g., centers k-means). This creates a systematic upward bias in stability estimates that is analogous to issues that arise when using bootstrap to estimate generalization error in supervised learning (Efron & Tibshirani, 1997).

For the iris data, we applied Bagclust 1 (Dudoit & Fridlyand, 2003) with 50 bootstrap replications (Figure 3a,b) using the clue package. This bootstrapping process was performed over a range of k-values ($k = 2, \ldots 8$) to create a stability profile (Figure 3a). The stability profile suggests that two or three clusters are both fairly stable with stability levels above 0.90. In general, there is not a perfect way to select k from a stability profile of this type. While the level itself is important; it is also the shape of the stability profile. Figure 3a shows a clear drop from $k = 3$ to $k = 4$, sometimes referred to as an *elbow* in the plot. In this case, $k = 3$ would be a good choice, as the structure in the data breaks down as indicated by the drop in stability. Although $k = 2$ is more stable, due to the separability of setosa, the high stability value for $k = 3$ indicates some weaker substructure in the data. For $k = 3$, the stability levels for the individual observations are shown in Figure 3b. The pattern of stability suggests that across some of the clusterings of the bootstrap replications some of the original clusters (Figure 1) are actually split. The stability profile for Scheme 1 by Yu et al. (2019) is shown in Figure 3c for 50 bootstrap replications across the same range of k values using the bootcluster package. Yu et al. (2019) suggest using stability values above 0.9 for the selection of k resulting in $k = 3$ as optimal. This threshold was also suggested by Tibshirani and Walther (2005). Notably, the stability profile does not exhibit the same elbow as seen in Figure 3a. The observation level stability for Scheme 1 is shown in (Figure 3d) for $k = 3$. Unlike Bagclust 1, the

observation stability is more consistent with the original clustering of the iris data (Figure 1). Specifically, the well separated setosa is highly stable, and the unstable points sit at the cluster boundary of the overlapping clusters. The methods developed by Fang and Wang (2012) and Hennig (2007) were implemented using the fpc package. The outputs do not lend themselves to the same level of visualizations, but rather yield summary values. Fang and Wang (2012) selected $k = 2$ as optimal. Hennig (2007) selected a more complex model, with five clusters, as optimal. Although the implementation of this method, in the R package fpc, is flexible to work with a range of clustering algorithms, a challenge was the need to set a number of tuning parameters, which may be data specific (Hennig, 2007).

3.2 | Cluster validation via data splitting and subsampling

In unsupervised learning, a common approach is to re-cast the problem at hand as a supervised problem. In this new setting, a gold standard is created by which performance can be measured. A family of clustering stability approaches center on this idea, and generate perturbed datasets through splitting (e.g., into a learning and test set) and subsampling approaches (Figure 3c). When the sample size is adequate this is a feasible approach as it eliminates bias that can arise from re-sampling based approaches. However, important structures may be missed when the sample size is too small. In the methods that follow, the common theme is to re-cast the clustering problem as a classification problem where the objective is to predict the cluster labels. The performance of this prediction characterizes the stability of the clustering. The nature of the setup prohibits the detection of the trivial case that there is no structure in the dataset ($k = 1$). Some of the below approach work around this with an alternative measure, and some require the investigator to rule this out a priori before proceeding. These methods also share a theme of contrasting stability to a reference null distribution for the purpose of model selection.

A prediction-based method known as *Clest* was developed by Dudoit and Fridlyand (2002) for the purpose of selecting the optimal number of clusters. The algorithm requires that the user specifies the maximum number of clusters, M. For each k, $2 \le k \le M$, *Clest* randomly splits the original dataset, $X \in \mathbf{R}^{N \times p}$, into two non-overlapping sets a total of B times, into a learning set L^b and a test set T^b, $1 \le b \le B$. A clustering procedure is then applied to the learning set L^b to get a partition $P(L^b)$. The problem is then cast into a supervised learning problem by passing the cluster labels and the learning set into a classifier $C(T^b)$ to build a predictive model. The classifier is then applied to the test set to get a test set partition. The same clustering algorithm used with learning set L^b is then applied on test set T^b to get another test set clustering membership $P(T^b)$. Note that both the classifier and the clustering algorithm is chosen according to the researcher's preferences. Dudoit and Fridlyand (2002) demonstrated *Clest* using partition around medoids (PAM; Van der Laan et al., 2003) with the a linear discriminant analysis classifier (Lachenbruch & Goldstein, 1979).

In order to generate a measure of stability, an index $s_{k,b}$ is calculated by comparing the test set partitions according to an external index measure of agreement between partitions, for example, Rand Index (Rand, 1971), Jaccard coefficient (Jain & Dubes, 1988), and FM index (Fowlkes & Mallows, 1983). For each k, the median of the indices over B bootstrap experiments are recorded as t_k. The same procedure is then applied to B null reference datasets generated to have no signal ($k = 1$) and the index measures are recorded as t_k^0. From these measures, a p-value, p_k, is calculated as the proportion of the index measures, $t_{k,b}$, $1 \le b \le B_0$, that are at least as large as the t_k. Finally, let $d_k = t_k - t_k^0$ denote the difference between the two statistics. The optimal k is chosen such that p_k is less than a preset threshold, and d_k larger than a preset threshold. If there is no such k's, then no structure is found ($k = 1$). If there are more than one k's, choose where d_k is maximized. Figure 4a shows the stability profile estimated by Clest for the original iris data and a generated null reference set. The separation is maximized at $k = 2$ and was significant ($p < 10^{-10}$), whereas $k = 3$ was not significant ($p = 0.10$).

Lange et al. (2004) developed a framework that improves upon *Clest* (Dudoit & Fridlyand, 2002). Similar to *Clest*, the data is divided into a learning set and test set. Right away, there are two fundamental distinctions in the setup. First, the learning and test set are equal in size, as the imbalance can be problematic and lose important signal, especially in small datasets. Second, the splitting is performed with a generative splitting scheme. The authors also emphasize the importance of the selection of the classifier itself, as we are trying to capture the stability of the data distributions, and want to thereby avoid influence by the classifier. It is therefore suggested that the potential classifiers should be selected to minimize a loss function. The approach itself is then largely similar to *Clest*. First, the two datasets generated from the original are clustered. Then, the learning set is used to predict the cluster labels. The fit model is then used to predict the labels for the test set. A normalized agreement function is calculated to estimate the distance of the two sets of solutions (cluster labels) for the test set. For the model selection component, the process is applied to null reference

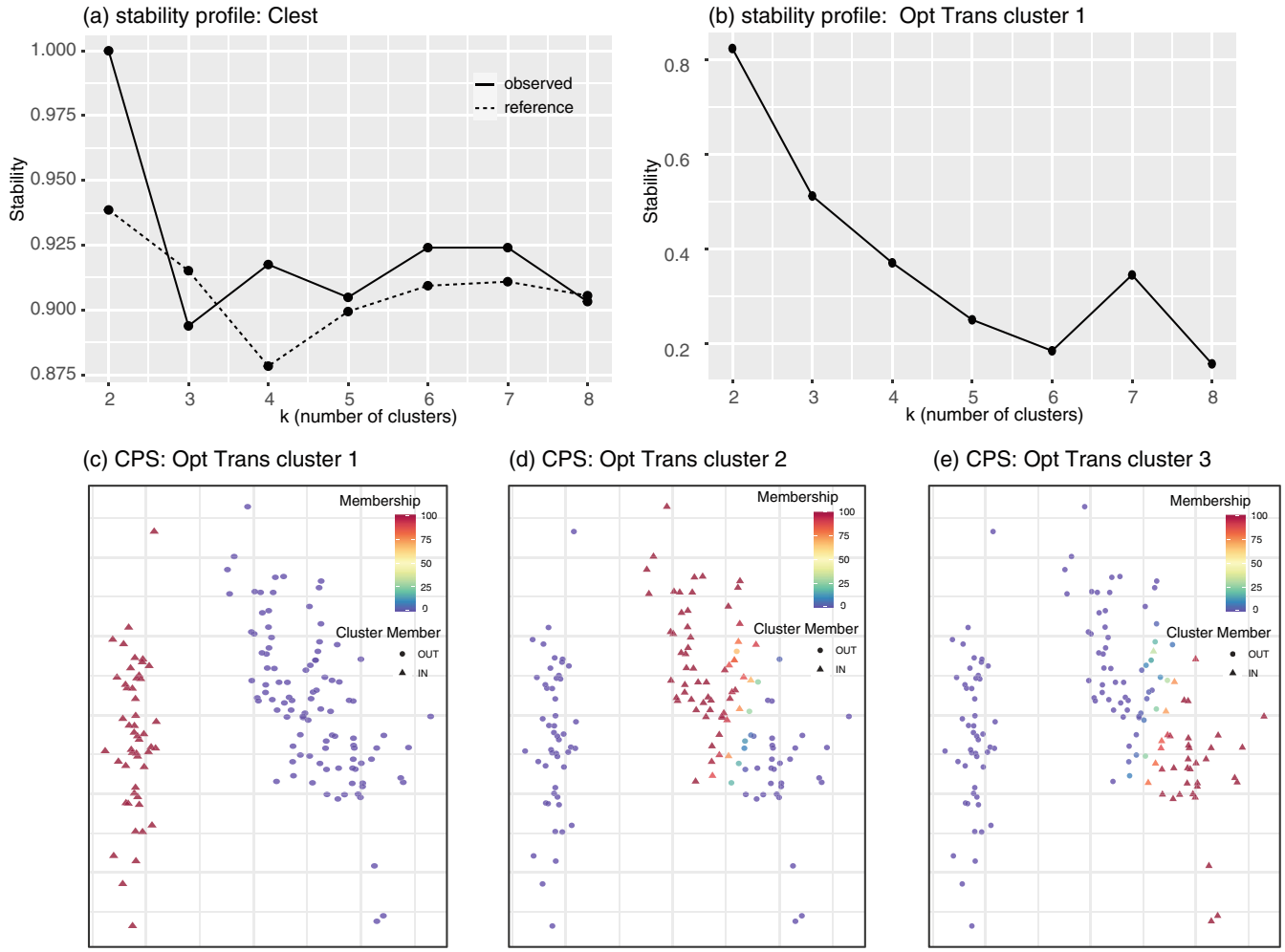

FIGURE 4 (a) The stability profile of the iris data was estimated using the Clest algorithm (Dudoit & Fridlyand, 2002). (b) Stability profile using optimal transport alignment (OTA; Li et al., 2019). (c) Visualization of the covering point set (CPS) for the OTA algorithm cluster 1, which is well separated. (d) CPS for the OTA algorithm cluster 2, and (e) CPS for the OTA algorithm for cluster 3

datasets with no signal, $\widehat{S}(R_k)$, a random labeling algorithm is used to assign each observation to a class with probability $1/k$. The same process of clustering and classification is then applied to the reference null sets, and the optimal k is selected to be the minimum of the ratio: $\widehat{S}(A_k)/\widehat{S}(R_k)$.

Levine and Domany (2001) introduced a method that derives a *Figure of Merit*, $M(V) \in [0,1]$, for a clustering procedure with parameterization V. The Figure of Merit is a function of how well the co-membership matrices generated from the clustering subsamples of the data agree with the clustering of the original data. For model selection, different parameterizations are explored to estimate an optimal Figure of Merit that has highly stable co-memberships (entries close to one). The Figure of Merit can also be used to select between competing clustering methods. A limitation of this approach is that it can produce local maxima; resulting in trivial or misleading solutions. Ben-Hur et al. (2002) developed a similar approach that is based on the subsampling. In this setting, the subsets are randomly generated to have a fixed proportion of points that overlap between them (0.8 in application). For a given k, their approach clusters pairs of these subsamples and computes the similarity between the labels of the points common to both sets. This subsampling–clustering–computing procedure is carried out multiple times to generate a distribution of similarities for the given k. Similarity distribution curves are plotted onto the same graph and the optimal k is selected as the point where curves make the biggest gap/transition, and produces an *elbow* in the profile. The success of this approach depends critically on the selection of the hyperparamater that controls the proportion of overlap between the subsamples, and the preset number of iterations (Ben-Hur et al., 2002). The authors provide some guidelines for their tuning, see Ben-Hur et al. (2002) for details.

A method known as *prediction strength* was developed by Tibshirani and Walther (2005) for the purpose of estimating the optimal number of clusters. The prediction strength algorithm seeks to capture and validate the structure of the through k-fold cross-validation. In a theme similar to Dudoit and Fridlyand (2002) and Lange et al. (2004), the clustering problem is cast into a supervised learning framework, but not as a classification problem. Following k-fold cross-validation, $k-1$ folds are combined into a learning set, and a kth fold is retained as the test set. The learning set data is clustered to obtain a parameter set for the clustering (e.g., means or centroids). The test set is then projected into the clustering space to obtain a set of cluster labels. The test set is then clustered on its own. The prediction strength, a surrogate for stability, is a function of the co-membership changes when comparing the test set projection onto the learning set clustering, and the clustering of the test set on its own. Note that this method requires the specification of a threshold in order to select the optimal number of clusters. Empirically, the choice of threshold was suggested to fall above 0.80 or 0.90, and the importance of examination prediction strength profile across a range of k values is emphasized (Tibshirani & Walther, 2005). The *gap statistic* arose out of earlier work by Tibshirani et al. (2001) and avoided the need to identify a threshold value for model selection. The gap statistic relies on contrasting the within-cluster sum of pairwise dissimilarity of a dataset to bootstrapped versions of a null reference with no signal. Although not a measure of stability per se, exploiting the within cluster dissimilarity differences between the structured data and random null reference distributions is a unique approach to model selection that avoids the need to set a predefined threshold.

3.3 | Alternative methods

There have been a handful of promising methods that do not fit the mold of resampling or cluster validation. Bertrand and Mufti (2006) developed a series of stability rules based on Loevinger's measures (Loevinger, 1947) using subsamples of the data. Loevinger's measure is in the range $[0,1]$, with a value of 0 if sets E and F are independent, and value 1 if $E \subseteq F$. One aspect of this approach that sets it apart is the *proportionate stratified sampling* (Hansen et al., 1953) process that is utilized, where each cluster in the original dataset is resampled without replacement by a fraction, f, which is recommended to be at least 0.7. Uniquely, this setup ensures that each cluster structure in the original dataset is preserved to some degree. The stability rules are designed to determine if a cluster is in isolation and/or cohesion with another cluster are aggregated over the resampled data. The significance of the rule is assessed using a p-value that contrasts the corresponding measures with a dataset that has no structure. Estimates are sensitive to the hyper-parameter for the sampling and it is emphasized that the p values should be interpreted cautiously due to a number of reasons that influence the distribution of stability measures under the null hypothesis (Bertrand & Mufti, 2006). This approach provides a strong level of interpretation to the user because the resulting stability measures capture the stability of individual clusters and the partition, thereby allowing the identification and prioritization of stable homogenous clusters. Although this method implements a form of resampling, the stratification of the clusters distinguishes it from the previous methods described.

Steinley (2008) proposed a stability estimation based on matrix manipulation that is primarily designed for k-means. The k-means algorithm is run on the same dataset with different random initializations. The co-membership matrices from these clusterings are then aggregated to form a consensus matrix. The consensus matrix is clustered and reordered using an optimization algorithm that maximizes within-block co-occurrences. The partitioning of the consensus matrix into block diagonal form represents the *most stable* partition that can be achieved with k-means clustering. These blocks can also be further examined to assess the degree of overlapping between clusters, and to quantify the membership of an item to a given cluster. These measures can be interpreted probabilistically, and similar to Yu et al. (2019), provides insights into stability that go beyond summary statistics and capture the item and cluster level.

An optimal transport framework (Villani, 2003) for cluster stability was recently developed by Li et al. (2019). The approach utilizes the optimal transport alignment (OTA) algorithm, which operates on the ensemble of clustering partitions. OTA-based stability introduced several novel aspects to the stability field. Although the algorithm relies on bootstrapping, it can be generalized to alternative approaches that generate perturbed data. Mean partitions are estimated using OTA, and cluster alignment is performed across the bootstrap replications to generate an ensemble of clusters, C_i, which are most similar and representative. Although this type of cluster mapping can be problematic, the uncertainty of this process is quantified using a cluster alignment matrix. The authors also introduce a measure that is analogous to confidence intervals that is known as covering point set (CPS), which further captures cluster separability based on the collection of aligned clusters. A distinguishing feature of OTA stability is the inherent flexibility with respect to clustering method, which is due to the fact that the algorithm is performed independently of the clustering method used to

generate the partitions and the individual data points. Additionally, most stability estimations described in this review claim to extend well to other clustering methods. However, this may require some additional algorithm modifications and considerations. On the other hand, a major strength of OTA is that it only operates on an ensemble of partitions that could arise from any clustering method, and how these partitions are generated is disregarded. In order to emphasize this flexibility, Li et al. (2019) demonstrate this method on a range of simulations and real-world dataset; they also examine a range of clustering methods, including *k*-means, hierarchical clustering, model-based clustering (mclust; Scrucca et al., 2016), DBscan (Ester et al., 1996). Recently, an `OTclust` pipeline was further developed for applications to biomedical data. Specifically, additional applications were developed for omics data, sub-group identification and the selection of data generation technologies were described and connected with the R package (Zhang et al., 2020).

The iris data was examined using the `OTclust package` (Li et al., 2019; Zhang et al., 2020). Figure 4b shows the stability for *k*-means as a measure of *overall tightness* across a range of *k* values. The stability profile summarizes the consistency of the clustering for different *k* values across bootstrap replications. In this case, the profile suggests that $k = 2$ is the most stable clustering (Figure 4b). However, to better illustrate the idea behind the covering point set, we examined the results for $k = 3$ in order to visualize more variation in the CPS matrix. Figure 4a shows the well-separated cluster (triangles) which are in close to 100% of the bootstrap replications, whereas the points (circles) in the other clusters are never members. On the other hand, the overlapping clusters (Figure 4d,e) show strong memberships away from the boundary (red triangles), but the points at the points proximal to the cluster boundary appear more often in the corresponding CPS. Visualizations of the CPS offer insights into the observation level stability, similar to the findings of Bagclust 1 (Dudoit & Fridlyand, 2003) and Scheme 1 (Yu et al., 2019; Figure 3c,d).

4 | DISCUSSION

Clustering is one of the most widely utilized tool in data mining. The most fundamental uses can be found in exploratory data analysis (EDA) and data preprocessing. More complex applications aim to group data for the purpose of extracting knowledge discovery in databases (KDD). Both EDA and KDD can be easily compromised due to the fact that there exists no gold standard by which to assess the quality and reproducibility of a clustering. Thus, the area of cluster stability remains an important one that continues to be widely studied. The sheer number of methods of stability estimation reflects the different interpretations of the notion of *stability* and the fact that the ground truth is unknown. Although stability will not guarantee an optimal clustering; it does suggest reproducibility and confidence in the results.

This review focused on the general problem of stability estimation for unsupervised clustering. An immediate challenge is that there are many clustering methods to choose from. The problem of selecting a clustering algorithm is not a new one (Rice, 1976); and is universal across all areas of data mining. The selection of clustering method, and in some cases choices of dissimilarity and linkage, can be rather subjective in nature and largely data dependent. Clustering methods differ in their optimization and some require choices for dissimilarity measures. The selection of method will naturally influence different aspects of the feature space in a clustering problem, leading to different groupings. These subjective choices naturally propagate into the stability measures. Stability estimates will also be sensitive to cluster size, cluster size imbalance, and the heterogeneity of the data in the clusters. In practice, it is entirely feasible that these factors may lead to the discovery of highly stable erroneous clusters, or highly unstable true structures.

An additional layer of complexity is that methodologies capture different aspects of stability, some at the level of the observation, cluster, and as an overall measure. Coupling different clusters with different stability methods may give conflicting results. This phenomenon was already witnessed with the simple iris example. Not surprisingly, the majority of stability methods focus on a specific clustering method, and discuss the generalization to alternative clustering methods. For example, the majority of stability methods have been developed around *k*-means, one of the most popular clustering algorithms. This center-based approach has been shown to work well with both bootstrapping and model validation approaches. Methods that rely on the aggregation of co-membership matrices may generalize better as the co-membership matrix can be derived from any clustering algorithm. The modifications needed to generalize and adapt the different stability approaches to alternative clustering methods and dissimilarities is not always straight forward, and may require additional research. Comprehensive examinations of the generalization properties of stabilities would enable an investigator to identify the most suitable set of techniques for a stability analysis, and may play a role in the decision to proceed with certain clustering algorithms that are more compatible with stability measures.

In general, the problem of selecting the number of clusters (model selection) in a dataset is fraught with challenges. As demonstrated with the iris data, the use of a stability value or stability profile to select the optimal number of clusters may not give a clear solution. Although several stability methods are designed for model selection; their use for this purpose has been controversial. Ben-David et al. (2006) examine the theoretical properties of stability for center-based and spectral clustering. Their findings show that stability is not well suited for model selection unless the objective function has a global minimizer. Otherwise, stability can be induced by symmetries of data which is unrelated to clustering parameters. Shamir and Tishby (2008) showed that stability based on model validation is a meaningful measure for larger sample sizes. The theoretical and empirical justification is based on the fact that stability for model validation is a form of generalization error that does not degrade with increasing sample size. Ben-David and Von Luxburg (2008) explore cluster stability as a function of the cluster boundary and find that it is not possible to guarantee global convergence. They support the use of cluster stability as a *red flag* that something is wrong when low stability arises. However, they argue against the use of clustering to instill confidence in a grouping due to many potential issues, including small sample sizes, geometric instability, and local optima. Von Luxburg (2009) discussed at length about stability based on k-means. Specifically, it was emphasized that stability measures should only be convincing when the underlying distribution can be represented by center-based clusters. Von Luxburg (2009) also supported the use of stability to signal a potentially unstable clustering and dataset, and reviewed the theoretical properties that had been derived up until then. To the author's knowledge, there has not been a comprehensive benchmarking of stability approaches for model selection verse more classical clustering heuristics, for example, the silhouette plots (Rousseeuw, 1987), Calinski–Harabasz index (Maulik & Bandyopadhyay, 2002), gap statistic (Tibshirani et al., 2001), or profiles of the within-cluster dissimilarity.

A common limitation across stability estimations that are based on bootstrapping is the inherent bias that can arise from the resampling process. This problem is commonplace when using the bootstrap to estimate generalization error. Breiman (1996) brought this issue into sight for supervised learning, suggesting that the *left out* (out-of-bag) observations from the resampling serve as test cases for error estimates. The rationale is that given the training set is used to construct the predictor, the most accurate error estimate should be a test set independent of the training set. Breiman (1996) showed this empirically for bagged predictors and random forests Breiman (2001). Efron and Tibshirani (1997) developed general estimators for generalization error that are based on out-of-bag estimation. The use of out-of-bag estimation has not been developed for stability estimation, but is a promising direction that could potentially ease of overcome the bias from the resampling procedures.

Although stability is not directly comparable between clustering methods; it has been shown to be useful in selecting a clustering method for a given dataset. Ensemble clustering is a related area of research that aims to aggregate clustering solutions, from different methods, for a given dataset (Strehl & Ghosh, 2002; Topchy et al., 2005; Vega-Pons & Ruiz-Shulcloper, 2011). Ensemble clustering can be used as a tool for meta-analysis because it operates directly on the cluster partitions that arise from different clustering methods. The aggregation is thereby independent of the original data features and the individual clustering algorithms used to create the ensemble of partitions. Of the methods described in this review, the OTA framework described in Section 3.3 combines properties of both stability methods and ensemble methods, as it operates directly on the ensemble of partitions generated (as in ensemble clustering) from perturbed versions of the datasets (as in stability methods; Li et al., 2019). Note that the core fundamental problem that these research areas share is how to compare and combine clustering solutions across an ensemble. Ensemble clustering accounts for uncertainty in the clustering technique selected, while stability estimation focuses on the uncertainty in the data itself. Notably, this does not imply that ensemble clusters are necessarily stable. In fact, it is entirely possible that combining clustering solutions in an ensemble would yield unstable clusters that are not reproducible. The quality of a clustering solution arising from an ensemble will be largely a function of the quality of the individual partitions being combined. To the author's knowledge, direct bridging of these methodologies has not been explored, although may be of mutual benefit.

There are a number of unsupervised learning methods, outside of clustering, which leverage definitions of stability. Networks are graphical models that represent directed and undirected associations (edges) between random variables (nodes). Learning network structure from data is a challenging problem. Markov Chain Monte Carlo methods capture a form of stability by sampling an ensemble of graphs from the posterior distribution, which can be summarized using Bayesian Model Averaging, or alternative consensus methods. However, this is not directly relatable to the cluster stability, as it does not arise from perturbations or resampling of the dataset. Meinshausen and Bühlmann (2010) developed a method known as stability selection, which can be used in connection with inference of Gaussian Graphical Models (GGMs). Community (aka module) detection is the process of partitioning a graph into groups with high

similarity, which can be veiled as a veiled clustering problem. Recently, Tian et al. (2021) extended a bootstrap framework for clustering to the problem of module detection.

5 | CONCLUDING REMARKS

Quantifying stability can instill confidence in a given clustering when labels or a gold standard are unavailable. Cluster stability has served as a surrogate for performance and reproducibility for a range of applications. We reviewed some of the most widely used and highly relevant methods developed in the last 30+ years. The stability measure, although still being studied theoretically, has wide applications and extensions such as quantifying stability at the level of the observation, cluster, and overall, as well as model selection for the optimal number of clusters. Although an already active field, the area of clustering stability is rich with interesting open questions and the capability for more in-depth theoretical formulations and cross-field applications.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Tianmou Liu: Data curation (equal); formal analysis (equal); investigation (equal); writing – original draft (equal); writing – review and editing (equal). **Han Yu:** Data curation (equal); validation (equal); writing – original draft (supporting); writing – review and editing (supporting). **Rachael Hageman Blair:** Conceptualization (equal); funding acquisition (equal); project administration (equal); supervision (lead); writing – original draft (lead); writing – review and editing (lead).

DATA AVAILABILITY STATEMENT

Data available in article supplementary material

ORCID

Rachael Hageman Blair https://orcid.org/0000-0001-8538-2447

RELATED WIRES ARTICLE

Cluster ensembles

REFERENCES

Abboud, K., & Zhuang, W. (2015). Stochastic modeling of single-hop cluster stability in vehicular ad hoc networks. *IEEE Transactions on Vehicular Technology*, *65*(1), 226–240.

Adamson, G. W., & Bush, J. A. (1975). A comparison of the performance of some similarity and dissimilarity measures in the automatic classification of chemical structures. *Journal of Chemical Information and Computer Sciences*, *15*(1), 55–58.

Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, *60*, 19–31.

Asuncion, A., & Newman, D. (2007). *UCI machine learning repository*. Center for Machine Learning and Intelligent Systems.

Bair, E. (2013). Semi-supervised clustering methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(5), 349–361.

Ben-David, S., & Von Luxburg, U. (2008). Relating clustering stability to properties of cluster boundaries. In *21st Annual Conference on Learning Theory (COLT 2008)*. Omnipress. pp. 379–390.

Ben-David, S., Von Luxburg, U., & Pál, D. (2006). A sober look at clustering stability. In *COLT'06: Proceedings of the 19th annual conference on Learning Theory* (pp. 5–19). Springer.

Ben-Hur, A., Elisseeff, A., & Guyon, I. (2002). A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing*, *7*, 6–17.

Bertoni, A., & Valentini, G. (2005). Random projections for assessing gene expression cluster stability. In *Proceedings of the 2005 IEEE International Joint Conference on Neural Networks* (Vol. 1, pp. 149–154). IEEE.

Bertrand, P., & Mufti, G. B. (2006). Loevinger's measures of rule quality for assessing cluster stability. *Computational Statistics & Data Analysis*, *50*(4), 992–1015.

Breiman, L. (1996). *Out-of-bag estimation. Technical report*. Statistics Department, University of California Berkeley.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *41*(3), 1–58.

Choi, S.-S., Cha, S.-H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, *8*(1), 43–48.

Dolnicar, S. (2002). A review of data-driven market segmentation in tourism. *Journal of Travel & Tourism Marketing*, *12*(1), 1–22.

Dolnicar, S. (2020). Market segmentation for e-tourism. In Z. Xiang, M. Fuchs, U. Gretzel, & W. Höpken (Eds.), *Handbook of e-tourism* (pp. 1–15). Springer.

Dolnicar, S., & Lazarevski, K. (2009). Methodological reasons for the theory/practice divide in market segmentation. *Journal of Marketing Management*, *25*(3–4), 357–373.

Dubes, R., & Jain, A. K. (1980). Clustering methodologies in exploratory data analysis. *Advances in Computers*, *19*, 113–228.

Dudoit, S., & Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, *3*(7), 1–21.

Dudoit, S., & Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, *19*(9), 1090–1099.

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The 632+ bootstrap method. *Journal of the American Statistical Association*, *92*(438), 548–560.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap: Chapman & Hall/CRC monographs on statistics & applied probability*. CRC Press.

Erdmann, T., & Schwarz, U. S. (2007). Impact of receptor-ligand distance on adhesion cluster stability. *The European Physical Journal E*, *22*(2), 123–137.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). ACM Press.

Estivill-Castror, V. (2002). Why so many clustering algorithms: A position paper. *ACM SIGKDD Explorations Newsletter*, *4*(1), 65–75.

Fang, Y., & Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, *56*(3), 468–477.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, *39*, 783–791.

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

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*(383), 553–569.

Getz, G., Levine, E., & Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 12079–12084.

Giancarlo, R., & Utro, F. (2012). Algorithmic paradigms for stability-based cluster validity and model selection statistical methods, with applications to microarray data analysis. *Theoretical Computer Science*, *428*, 58–79.

Giurcăneanu, C. D., & Tăbuş, I. (2004). Cluster structure inference based on clustering stability with applications to microarray data analysis. *EURASIP Journal on Advances in Signal Processing*, *2004*(1), 1–17.

Gnanadesikan, R., Kettenring, J. R., & Landwehr, J. M. (1977). Interpreting and assessing the results of cluster analyses. *Bulletin of the International Statistical Institute*, *47*(2), 451–463.

Grira, N., Crucianu, M., & Boujemaa, N. (2004). Unsupervised and semi-supervised clustering: A brief survey. *A Review of Machine Learning Techniques for Processing Multimedia Content*, *1*, 9–16.

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for meg: A comparison of dissimilarity measures. *NeuroImage*, *173*, 434–447.

Hajibaba, H., Grün, B., & Dolnicar, S. (2019). Improving the stability of market segmentation analysis. *International Journal of Contemporary Hospitality Management*, *32*, 1393–1411.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory. Vol. I. Methods and applications*. Wiley.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis*, *52*(1), 258–271.

Hidalgo, S. J. T., & Ma, S. (2018). Clustering multilayer omics data using muncut. *BMC Genomics*, *19*(1), 1–13.

Iwasaki, Y., Kusne, A. G., & Takeuchi, I. (2017). Comparison of dissimilarity measures for cluster analysis of x-ray diffraction data from combinatorial libraries. *NPJ Computational Materials*, *3*(1), 1–9.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Jain, A. K., & Moreau, J. (1987). Bootstrap technique in cluster analysis. *Pattern Recognition*, *20*(5), 547–568.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis* (Vol. 344). Wiley.

Kerr, M. K., & Churchill, G. A. (2001). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(16), 8961–8965.

Kim, H.-J., McGuire, D. B., Tulman, L., & Barsevick, A. M. (2005). Symptom clusters: Concept analysis and clinical implications for cancer nursing. *Cancer Nursing*, *28*(4), 270–282.

Lachenbruch, P. A., & Goldstein, M. (1979). Discriminant analysis. *Biometrics*, *35*, 69–85.

Lange, T., Roth, V., Braun, M. L., & Buhmann, J. M. (2004). Stability-based validation of clustering solutions. *Neural Computation*, *16*(6), 1299–1323.

Leisch, F. (1999). *Bagged clustering*. Adaptive Information Systems and Modelling in Economics and Management Science, WU Vienna University of Economics and Business.

Levine, E., & Domany, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, *13*(11), 2573–2593.

Li, J., Seo, B., & Lin, L. (2019). Optimal transport, mean partition, and uncertainty assessment in cluster analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *12*(5), 359–377.

Loevinger, J. E. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, *61*(4), i–49.

Loza, M. J., Adcock, I., Auffray, C., Chung, K. F., Djukanovic, R., Sterk, P. J., Susulic, V. S., Barnathan, E. S., Baribaud, F., & Silkoff, P. E. (2016). Longitudinally stable, clinically defined clusters of patients with asthma independently identified in the adept and u-biopred asthma studies. *Annals of the American Thoracic Society*, *13*(Suppl 1), S102–S103.

Loza, M. J., Djukanovic, R., Chung, K. F., Horowitz, D., Ma, K., Branigan, P., Barnathan, E. S., Susulic, V. S., Silkoff, P. E., Sterk, P. J., Baribaud, F., & ADEPT (Airways Disease Endotyping for Personalized Therapeutics) and U-BIOPRED (Unbiased Biomarkers for the Prediction of Respiratory Disease Outcome Consortium) Investigators. (2016). Validated and longitudinally stable asthma phenotypes based on cluster analysis of the adept study. *Respiratory Research*, *17*(1), 1–21.

Mammu, A. S. K., Hernandez-Jayo, U., & Sainz, N. (2013). Cluster-based mac in vanets for safety applications. In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1424–1429). IEEE.

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(12), 1650–1654.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *72*(4), 417–473.

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, *52*(1), 91–118.

Müller, H., & Hamm, U. (2014). Stability of market segmentation with cluster analysis—A methodological approach. *Food Quality and Preference*, *34*, 70–78.

Newby, C., Heaney, L. G., Menzies-Gow, A., Niven, R. M., Mansur, A., Bucknall, C., Chaudhuri, R., Thompson, J., Burton, P., Brightling, C., on behalf of the British Thoracic Society Severe Refractory Asthma Network (2014). Statistical cluster analysis of the British thoracic society severe refractory asthma registry: Clinical outcomes and phenotype stability. *PLoS One*, *9*(7), e102987.

Peters, G., Crespo, F., Lingras, P., & Weber, R. (2013). Soft clustering–fuzzy and rough approaches and their extensions and derivatives. *International Journal of Approximate Reasoning*, *54*(2), 307–322.

Peyvandipour, A., Shafi, A., Saberian, N., & Draghici, S. (2020). Identification of cell types from single cell data using stable clustering. *Scientific Reports*, *10*(1), 1–12.

Politis, D. N., Romano, J. P., & Wolf, M. (1999). *Subsampling*. Springer.

Puzicha, J., Buhmann, J. M., Rubner, Y., & Tomasi, C. (1999). Empirical evaluation of dissimilarity measures for color and texture. In *Proceedings of the Seventh IEEE International Conference on Computer Vision* (Vol. 2, pp. 1165–1172). IEEE.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850.

Rice, J. R. (1976). The algorithm selection problem. *Advances in computers*, *15*, 65–118.

Rokach, L., & Maimon, O. (2005). Clustering methods. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 321–352). Springer.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Sarle, W. S. (1990). *Algorithms for clustering data*. Prentice Hall.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 289.

Shamir, O., & Tishby, N. (2008). Cluster stability for finite samples. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20). Curran Associates, Inc.

Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLoS One*, *10*(12), e0144059.

Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y., & Herawan, T. (2014). Big data clustering: A review. In *International Conference on Computational Science and Its Applications* (pp. 707–720). Springer.

Smolkin, M., & Ghosh, D. (2003a). Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, *36*(4), 892–895.

Smolkin, M., & Ghosh, D. (2003b). Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, *4*(1), 1–7.

Steinbach, M., Ertöz, L., & Kumar, V. (2004). The challenges of clustering high dimensional data. In L. T. Wille (Ed.), *New directions in statistical physics* (pp. 273–309). Springer.

Steinley, D. (2004). Properties of the Hubert-Arable adjusted rand index. *Psychological Methods*, *9*(3), 386–396.

Steinley, D. (2008). Stability analysis in k-means clustering. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 255–273.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*(Dec), 583–617.

Tang, M., Kaymaz, Y., Logeman, B. L., Eichhorn, S., Liang, Z. S., Dulac, C., & Sackton, T. B. (2021). Evaluating single-cell cluster stability using the Jaccard similarity index. *Bioinformatics*, *37*(15), 2212–2214.

Tian, M., Blair, R. H., Mu, L., Bonner, M., Browne, R., & Yu, H. (2021). A framework for stability-based module detection in correlation graphs. *Statistical Analysis and Data Mining*, *14*(2), 129–143.

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, *14*(3), 511–528.

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.

Topchy, A., Jain, A. K., & Punch, W. (2005). Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(12), 1866–1881.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Addison-Wesley.

Van der Laan, M., Pollard, K., & Bryan, J. (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, *73*(8), 575–584.

Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, *25*(03), 337–372.

Villani, C. (2003). *Topics in optimal transportation (Graduate Studies in Mathematics)* (Vol. 58). American Mathematical Society.

Von Luxburg, U. (2009). Clustering stability: An overview. *Foundations and Trends in Machine Learning*, *2*(3), 235–274.

Wang, J., Wang, S.-T., & Deng, Z.-H. (2012). Survey on challenges in clustering analysis research. *Control and Decision*, *27*(3), 321–328.

Yang, M.-S. (1993). A survey of fuzzy clustering. *Mathematical and Computer Modelling*, *18*(11), 1–16.

Yeung, K. Y., Haynor, D. R., & Ruzzo, W. L. (2001). Validating clustering for gene expression data. *Bioinformatics*, *17*(4), 309–318.

Yu, H., Chapman, B., Di Florio, A., Eischen, E., Gotz, D., Jacob, M., & Blair, R. H. (2019). Bootstrapping estimates of stability for clusters, observations and model selection. *Computational Statistics*, *34*(1), 349–372.

Zhang, L., Lin, L., & Li, J. (2020). Cps analysis: Self-contained validation of biomedical data clustering. *Bioinformatics*, *36*(11), 3516–3521.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Liu, T., Yu, H., & Blair, R. H. (2022). Stability estimation for unsupervised clustering: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(6), e1575. https://doi.org/10.1002/wics.1575