A Bayesian Criterion for Cluster Stability

Hoyt Koepke[1] and Bertrand Clarke[2]*

[1]*Department of Statistics, University of Washington, Seattle, WA, USA*

[2]*Department of Medicine, University of Miami, Miami, FL, USA*

Received 23 April 2012; revised 21 November 2012; accepted 5 December 2012
DOI:10.1002/sam.11176
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: We present a technique for evaluating and comparing how clusterings reveal structure inherent in the data set. Our technique is based on a criterion evaluating how much point-to-cluster distances may be perturbed without affecting the membership of the points. Although similar to some existing perturbation methods, our approach distinguishes itself in five ways. First, the strength of the perturbations is indexed by a prior distribution controlling how close to boundary regions a point may be before it is considered unstable. Second, our approach is exact in that we integrate over all the perturbations; in practice, this can be done efficiently for well-chosen prior distributions. Third, we provide a rigorous theoretical treatment of the approach, showing that it is consistent for estimating the correct number of clusters. Fourth, it yields a detailed picture of the behavior and structure of the clustering. Finally, it is computationally tractable and easy to use, requiring only a point-to-cluster distance matrix as input. In a simulation study, we show that it outperforms several existing methods in terms of recovering the correct number of clusters. We also illustrate the technique in three real data sets. © 2013 Wiley Periodicals, Inc. Statistical Analysis and Data Mining, 2013

Keywords: clustering; stability; Bayesian; consistency; heatmap

1. STABILITY IN CLUSTERING

It is well known that clusterings can be unstable in the sense that multiple clusterings may claim to summarize a given data set equally well and we cannot tell which ones better reflect the intrinsic structure of the data. There are two main sources of instability. One kind of instability occurs when a data point can be reasonably assigned to more than one cluster. That is, the data point is in a boundary region and might plausibly be associated with two different clusters. Another kind of instability occurs when a cluster incorrectly indicates a centroid in the underlying distribution of data. These two sources of instability may hold for a single clustering in different regions: It is easy to imagine part of the boundary of a cluster passing through a region where there are data points whose cluster membership is indeterminate and another cluster centroid being located where there are few or no data points. The difficulty in assigning boundaries to clusterings means an overall evaluation of the stability of a clustering is desirable.

Correspondence to: Bertrand Clarke
(bclarke2@med.miami.edu)

© 2013 Wiley Periodicals, Inc.

There are many ways to do this; cluster stability is hardly a new topic and many authors have developed useful techniques. Sometimes these go by the name of cluster validation since they are an effort to argue that the clustering obtained represents the real distribution. Many established techniques are based on data perturbation. The idea is to perturb the data set, usually by subsampling the data or adding noise, then reclustering the perturbed data. For examples, see refs. 1–5 among others. The primary idea behind this sense of cluster stability is that a clustering solution should be resistant to changes in the data that one would expect to occur in a real-data generator. More specifically, the stability of the clustering is determined by analyzing the similarity of the clustering across data perturbation runs or between the original data and the perturbed data, usually using a type of index designed for this purpose such as the Hubert–Arabie adjusted rand index [6], or the variation of information index [7]. Typically, this is simply the average of the stability indices produced by the sequence of data perturbation runs, but other summaries are possible.

A second approach is to use the silhouette score, see ref. 8, to choose the number K of clusters in a clustering directly. The idea is that for each K, a clustering of size

K has already been found by some technique, such as K-means or hierarchical clustering. Then one defines the point-to-cluster distance from any fixed point to any fixed cluster as the average of the distances from that point to all the other points in a given cluster. The silhouette score for a point is then the difference between its point-to-cluster distances for the cluster it was assigned to and the next best cluster, scaled by the maximum of these two point-to-cluster distances. Thus each point is scored by its proximity to a boundary, with points on a boundary having a score near zero and points near the center of a concentrated cluster having a score near ± 1. The silhouette score for a clustering is the average of the silhouette scores for each point. Finally, K is chosen from the clustering with the smallest average silhouette distance. This is essentially what is done in the R-package pamk. This technique often works well but is limited because each point is only considered from the standpoint of its actual cluster and its next best cluster, not all the possible clusters.

A third way to assess the stability of clusterings is to look at how likely it is that some dispersion measure achieved by a clustering could be the result of spurious clusterings on unstructured data. For instance, the gap statistic of Tibshirani [9] uses the difference in total cluster spread, defined in terms of the sum of all pairwise distances in a cluster, between the actual data set and the clusterings of several reference distributions with no true clusters. For example, in Euclidean space with the squared error distance metric, the spread is the total empirical variance of all the clusters. The reference distributions are used to adjust for the dependence on K in the measure and to guard against spurious clusters. This follows easily if the reference distributions are uniform and have no real clusters. However, generating a suitable data set from a reference distribution is not easy, as the final value can dependent strongly on the distribution of points. While much of this is not well understood, Tibshirani [9] proposes using the uniform distribution within the bounding box of the original data set using principal components.

Most recently, there is an R-package clusterCons that does consensus clustering partially in an effort to ensure cluster stability and validation. The technique is based on resampling so that two data points that are in the same cluster over more bootstrap samples are more likely to be put in the same cluster. The technique is described in ref. 10, see also ref. 11. However, note that this is essentially a data perturbation method for stability being used as a way to choose a clustering in the first place, not really an evaluation of the clustering obtained.

From the Bayesian standpoint, consensus clusterings have been motivated by the same desire for stability but the approach has been quite different. Roughly, the Bayesian approach [12] is to start with several clusterings of a data set formed using different techniques; usually this is called an ensemble. Then, one can develop a distribution over all possible consensus clusterings which can be used to generate a consensus membership structure, see ref. 12 for details. Again, this is not so much a stability assessment technique as a way to use stability concepts to choose a good clustering in the first place. In general, our method could be used to evaluate the clustering output from clusterCons or Bayesian ensemble approaches and (hopefully) verify that the consensus clustering was better—or at least no worse—than the original clusterings on the full data set, at least in a stability sense.

By contrast, here, we present several evaluations of the stability of a clustering showing the stability of each point, of each cluster, and an overall assessment of the clustering stability. Its simplest form is for centroid-based clustering procedures. For this case, our stability criterion is based on assessing how much the distances from data points to cluster centroids can be perturbed while ensuring the data point is still closest to its assigned cluster's center. These perturbations are expressed in terms of factors on $d(x_i, \hat{\mu}_k)$'s where x_i is a data point $i = 1, \ldots, n$, d is a metric, and $\hat{\mu}_k$ is an estimate of the kth cluster centroid, $k = 1, \ldots, K$. Specifically, if we have a clustering $\hat{\mathcal{C}} = \hat{C}_K = (\hat{C}_1, \ldots, \hat{C}_K)$ of n data points into K nontrivial regions, we evaluate the stability of a fixed cluster $\hat{C}_k$ that has x_i as a member using sets of the form

$$\hat{S}_{ik} = \Big\{ (\lambda_1, \ldots, \lambda_K) |\ \forall \ell \neq k : \times \lambda_k d(x_i, \hat{\mu}_k) \leq \min_{\ell \neq k} \lambda_\ell d(x_i, \hat{\mu}_\ell) \Big\}, \qquad (1)$$

where the λ_k's are non-negative parameters and $\hat{\mu}_k$ is the centroid of $\hat{C}_k$.

More generally, we can form analogous sets when d is not a metric and the clustering is not centroid based. For instance, the distance dist used in place of d in Eq. (6) of Section 2.2 is derived from a metric but is not a metric itself. Nevertheless, the $\hat{S}_{ik}$'s in Eq. (1) are still well defined and can be used to assess stability. Indeed, all our method requires that

$$\hat{S}^*_{ik} = \Big\{ (\lambda_1, \ldots, \lambda_K) |\ \forall \ell \neq k : \times \lambda_k \delta(x_i, \hat{C}_k) \leq \min_{\ell \neq k}\ \lambda_\ell \delta(x_i, \hat{C}_\ell) \Big\}, \qquad (2)$$

be well defined, where δ merely assigns a 'dissimilarity' between points x_i and clusters C_k. That is, our methodology only requires the $n \times K$ inputs $\delta(x_i, \hat{C}_\ell)$; these can come from familiar dissimilarity measures such as average linkage.

While our experimental results and simulations (see Section 7.1) demonstrate that our method works with general dissimilarity measures, our formal results are only for the case that d is a metric. An examination of the proofs will reveal that the results will hold for some nonmetric d's, but it is hard to characterize them precisely. Loosely, as long as d (or δ) is strong enough to ensure probabilities accumulate in the clusters asymptotically correctly, versions of Theorem 1 in Section 3.1, the properties given in Section 4.2, and Theorem 3 in Section 4.4 should hold. However, Theorem 2 in Section 3.2 and the proposition in Section 4.3 are not likely to hold in much greater generality than we have shown here as they makes use of the full strength of a metric.

It is seen that for both forms, Eqs. 1 and 2, the larger the set is, the more stable $\hat{C}_k$ is and the more stable the $\hat{C}_k$'s are collectively the more desirable the overall clustering is. We measure the size of each $\hat{S}_{ik}$ (or $\hat{S}^*_{ik}$) by taking the λ_k's to be independently and identically distributed with a marginal distribution function $F(\cdot)$ and finding $F^K(\hat{S}_{ik})$, resp. $F^K(\hat{S}^*_{ik})$. (We omit the superscript K indicating the K-fold product of F's when no confusion will result, thus for instance we write $F^K(\hat{S}_{ik}) = F(\hat{S}_{ik})$ with mild abuse of notation.)

Thus, F is seen to play the role similar to that of a prior: F is chosen by the practitioner and used to make inferences about stability—our only assumptions are that F is continuous and has only non-negative support. It is a source of information entirely disjoint from the data and represents pre-experimental information about how much one should be able to perturb distances between points and centroids without too much change to the clustering. We emphasize that F is not a prior in the sense that one can use it to form a product with a likelihood. Nevertheless, we refer to the clustering stability measure we study here as 'Bayesian' and for convenience call F a prior.

For fixed i, the values $F(\hat{S}_{ik})$ can be regarded as a soft membership function. Obviously, for each k, $F(\hat{S}_{ik}) \geq 0$ and for each i, $\sum_k F(\hat{S}_{ik}) = 1$. So, an x_i that is stably inside a cluster $\hat{C}_k$ will have a high $F(\hat{S}_{ik})$ while an x_i that is far from $\hat{\mu}_k$ will give a small $F(\hat{S}_{ik})$. So, the size of $F(\hat{S}_{ik}) \in [0, 1]$ is an indicator of how plausible it is to regard x_i as a member of $\hat{C}_k$. Analogous reasoning applies to $\hat{S}^*_{ik}$.

To see this more precisely, suppose $\mathcal{C}(K, \mathbb{D})$ is a clustering function that partitions a set of n data points $\mathbb{D} = \{x_1, x_2, \ldots, x_n\}$ into a set $\{\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_K\}$ of K clusters. Our approach to stability is to create a new clustering function $\mathcal{C}(K, \mathbb{D}, \lambda^K)$, where $\lambda^K = (\lambda_1, \ldots, \lambda_K)$, by modifying $\mathcal{C}(K, \mathbb{D})$ to take a parameter λ^K that represents the perturbation of some aspect of the clustering. So, without loss of generality, suppose the clustering function $\mathcal{C}(K, \mathbb{D})$ returns an $n \times K$ assignment matrix of points to clusters, say

$$A = [a_{ik}] = \mathcal{C}(K, \mathbb{D}).$$

In the hard clustering case, $a_{ik} = 1$ if $x_i \in \hat{C}_k$ and 0 otherwise. For soft clustering, each row of A is a distribution over the clusters giving the partial membership of each point in each cluster.

Now suppose that for a given value λ^K, the perturbed clustering function $\mathcal{C}(K, \mathbb{D}, \lambda^K)$ gives the assignment matrix

$$\boldsymbol{A}^*(\lambda^K) = \left[\boldsymbol{a}^*_{ik}(\lambda^K)\right] = \mathcal{C}(K, \mathbb{D}, \lambda^K).$$

Conditional on λ^K, $\mathcal{C}(K, \mathbb{D}, \lambda^K)$ is a deterministic function. We can integrate out λ^K from the $[\boldsymbol{a}^*_{ik}(\lambda^K)]$'s with respect to F, and later take $F = F_\theta$ where θ will be a hyperparameter. The result is an $n \times K$ matrix $\boldsymbol{\Phi} = [\phi_{ik}]$ defined by

$$\boldsymbol{\Phi}(K, \mathcal{C}) = [\phi_{ik}] = \int \boldsymbol{A}^*(\lambda^K)\mathrm{d}F_\theta(\lambda^K). \tag{3}$$

We call $\boldsymbol{\Phi}$ the averaged assignment matrix. The integration spreads the binary membership matrix A across the clusters based on the behavior of those points under perturbation of the clustering function by λ. If one interprets rows of $\boldsymbol{A}^*(\lambda^K)$ as probability distributions over the clusters, that is, write $\boldsymbol{a}^*_{ik}(\lambda^K) = p(a_{ik}|\lambda^K)$ for fixed i, then $\boldsymbol{\Phi}$ is analogous to a Bayesian's marginal for the data.

The link between Eq. (1) and Eq. (3) that makes our approach feasible is taking $\boldsymbol{A}^*$ to be defined by indicator functions for the $\hat{S}_{ik}$'s. Specifically, we set

$$\boldsymbol{\Phi}(K, \mathcal{C}) = [\phi_{ik}] = \int \mathbb{I}_{\hat{S}_{ik}}(\lambda^K)\mathrm{d}F(\lambda^K), \tag{4}$$

so that when F concentrates at $\lambda^K = (1, \ldots, 1)$, $[\phi_{ik}] = [a_{ik}]$. We do not study the perturbed clusterings $\mathcal{C}(K, \mathbb{D}, \lambda^K)$ for individual values of λ^K, we study the average properties of the collection of perturbed clusterings $\{\mathcal{C}(K, \mathbb{D}, \lambda^K) |\ \lambda^K \in \mathbb{R}^{+,K}\}$ after integrating out λ^K as in Eq. (4). Note that, phrased this way, Eq. (4) is similar to the silhouette distance approach in ref. 8. The difference is that Eq. (4) involves the use of a prior F and, rather than using a ratio involving averaged distances for each point, Eq. (4) looks at a soft-clustering via the averaged assignment matrix. A more general form for Eq. (4) can be defined using S^*_{ik} in place of S_{ik}. Here, however, we focus on Eq. (4) because it is theoretically tractable; below we only use the S^*_{ik} version in two computed examples (see Sections 2.2 and 7.1).

As Eqs. 3 and 4 are matrices they provide a comprehensive assessment of how well each data point fits each

cluster. However, in practice we want to compare clusterings. So, for each point we use the entry in $\mathbf{\Phi}$ corresponding to its assigned cluster, then average these values over data points, that is, over $i = 1, \ldots, n$. We call the resulting scalar value the average pointwise stability APW. High values of the APW correspond to intuitively reasonable notions of stability and indicate desirable clusterings. That is, high APW can be used as a criterion to select among clusterings. In fact, our theoretical analysis below focuses on a special empirical case of the APW and its population form, see Eqs. 12 and 13. In Section 5, we explain the close relationship between these analytically tractable forms, the APW and the function of APW (calibrated against a baseline null distribution) that we use in practice.

We will see that an approach based on Eqs. 3 and 4 to evaluate the stability of a clustering has several advantages. First, if the stability of a clustering $\hat{\mathcal{C}}$ found using $\mathbb{D}$ is evaluated, it converges, asymptotically in n, to the population form of the stability of the clustering. Second, if one uses it to select a number of clusters, that is, uses our criterion to search for a high-stability clustering among a collection of candidate clusterings, one can prove that our stability criterion is consistent for the correct number of clusters, provided the corresponding modes in the data distribution are sufficiently distinguishable. Third, our criterion responds to how much mass a data set tends to put on boundary regions between clusters, giving smaller stability values as boundary regions become more populated. Fourth, in several generic cases when the stability of a clustering seems high, $\mathbf{\Phi}(K, \mathcal{C})$ approaches its maximal value of 1. This means we can use large values of our criterion as a way to choose clusterings and be secure that we have, in fact, found clusterings that match our intuitive idea of stability. Fifth, we establish a theorem showing that reasonable optimal clusterings under our criterion are not less stable, and often more stable, than other clusterings. This last result only holds in the limit of the population distribution concentrating on the cluster centroids, a condition which is very hard to relax in the general case.

From a performance standpoint, we compare our method to three conceptually different approaches to cluster stability, namely subsampling, silhouette distance, and the gap statistic. In a series of examples in Section 7, we find that with only a couple of explainable exceptions our method is better able to identify the correct number of clusters. Moreover, graphs of the clusterings chosen by our method on several examples appear reasonable and informative.

In addition, we use our method to do a stability analysis of two data sets. Our analysis leads us to suggest that in one data set there are more clusters than the apparent classes while in the other data set we find that fewer clusters than the apparent classes is reasonable. We comment that these two data sets were examined from a similar standpoint in ref. 13, who proposed two ways to evaluate how consistent a clustering was with subject-matter knowledge. At root, these evaluations are physically driven and provide scientific validation rather than stability. By contrast our method does not require any subject-matter information; it is a way to use stability concepts to derive information that may be germane to a subject-matter problem and amenable to downstream validation.

In the rest of this paper, we make the case that our method for evaluating cluster stability is an improvement over many existing methods. In Section 2, before explaining the details of our approach, we show how our method can be used to visualize the stability of clusters, a property that other methods do not in general have. That is, we generate what we call pointwise stability graphs indicating regions of high or low stability; these depend on the choice of F. We also visualize $\mathbf{\Phi}$ via 'stability heatmaps' to indicate clusters with low or high stability. In Section 3, we formally state our stability criterion based on Eqs. 3 and 4 and give our first collection of theoretical results demonstrating that this criterion has an asymptotic limit (as $n \to \infty$). We also verify that our criterion gives consistent estimation for an optimal number of clusters. In Section 4, we continue this work but focus on theoretical arguments ensuring the optima from our criterion are intuitively reasonable. In Section 5, we discuss the general implementation of our method for choosing the number of clusters. In Section 6, we take up prior selection and in particular show how to estimate the hyperparameter for the shifted exponential prior family. In Section 7, we examine the performance of our method computationally. We use three sorts of synthetic data to verify that our method outperforms three existing methods and then present two stability analyses of real data sets. Finally, in Section 8, we briefly discuss the advantages, disadvantages, and applicability of the methods that we have studied.

2. GEOMETRIC MOTIVATION

In this section, we present a geometric motivation and interpretation for the averaged assignment matrix $\mathbf{\Phi}$ defined in Eqs. 3 and 4. Specifically, we show two ways $\mathbf{\Phi}$ can be used to visualize the stability of a clustering. The first of these is an averaged pointwise stability measure that we call a pointwise stability graph. It is only useful in two dimensions but helps reveal what the ϕ_{ik}'s mean. The second is a heatmap of certain elements of the averaged assignment matrix showing the interactions between the clusters in the sense of where points go under perturbation. This plot shows the interaction under perturbation between

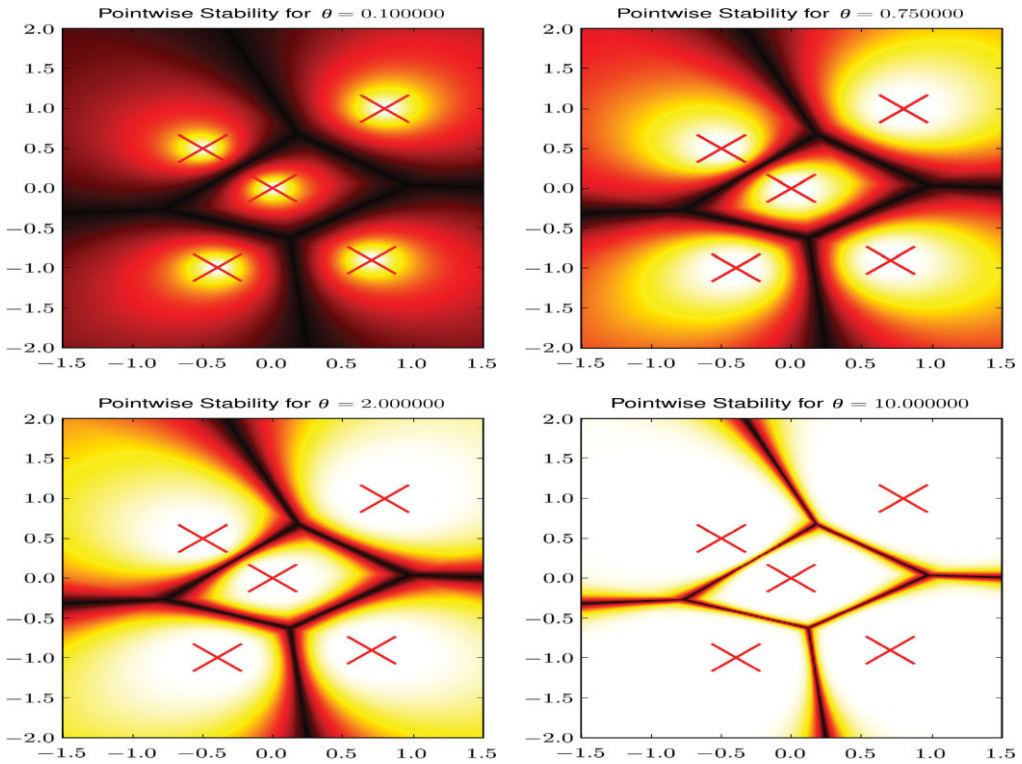


Fig. 1 The pointwise stability $\mathrm{PW}(x)$ for x ranging over $[-1.5, 1.5] \times [-2, 2]$, relative to fixed centroids indicated by red $\times$'s. Here, the prior controlling the scale of the perturbation values, the λ_k's, is a shifted exponential with location parameter 1 and scale parameter $\theta = 0.1, 0.75, 2, 10$ from upper left to lower right. White represents a pointwise stability of essentially 1 and black a pointwise stability of essentially 0 with intermediate colors representing intermediate pointwise stabilities. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

a fixed cluster and the other clusters; its usefulness is not dimensionally limited.

2.1. Properties of Pointwise Stability

Suppose, for data point x_i, h_i gives the index of the cluster to which it belongs, that is, $a_{i,h_i} = 1$. Then, the stability of x_i can be assessed by the (i, h_i) entry in the averaged assignment matrix:

$$\mathrm{PW}(x_i) = \mathrm{PW}_i = \phi_{i,h_i}.$$

Because we average over perturbations, that is, integrate out λ^K, PW_i is minimal when x_i is near the boundary of a cluster where perturbations in the distances are likely to send x_i closer to another cluster center. Likewise, $\mathrm{PW}(x_i)$ is maximal near cluster centers, where the perturbations, the λ_k's, have to be much larger to make the x_i change cluster membership. For general x, $\mathrm{PW}(x)$ can be evaluated from the set of K distances between x and each of the cluster centers; i.e., the $d(x, \hat{\mu}_k)$'s for $k = 1, \ldots, K$. That is,

$$\mathrm{PW}(x) = F\left(\left\{\lambda^K : \lambda_k d(x, \hat{\mu}_k) \leq \lambda_\ell d(x, \hat{\mu}_\ell)\, \forall \ell \neq k\right\}\right). \quad (5)$$

To illustrate how the pointwise stability behaves, consider the case of a centroid-based clustering in which a point is assigned to its nearest centroid. In two dimensions, a fixed collection of centroids taken as cluster centers partitions the space into a Voroni diagram by associating each point with its closest centroid. Now, $\mathrm{PW}(x)$ induces a type of 'soft' partitioning by associating each point x with a stability value that is low when x is near the boundary regions of the clusters and high near the cluster centers. How close a point must be to the boundary region before it is 'unstable' is controlled by the size of the perturbation factors λ^K; this in turn is controlled by the prior distribution $F(\lambda^K)$.

An example of this is shown in Fig. 1 where we have fixed five values $\hat{\mu}_k$, evaluated the pointwise stability $\mathrm{PW}(x)$ at each location in the rectangle, and assigned colors to values of $\mathrm{PW}(x)$. We call the result a pointwise stability graph for the clustering. The prior $F(\lambda^K)$ for each of the panels is a shifted exponential, that is, $f(\lambda|\theta) = \theta e^{-\theta(\lambda-1)}\mathbf{1}_{\lambda \geq 1}$, with $\theta = 0.1, 0.75, 2, 10$, respectively. Each panel in Fig. 1 shows five light colored regions surrounding the five cluster centers where $\mathrm{PW}(\cdot)$ is high. As x moves away from the cluster centers, the stability index in Eq. 5 decreases giving ever darker regions indicating instability.

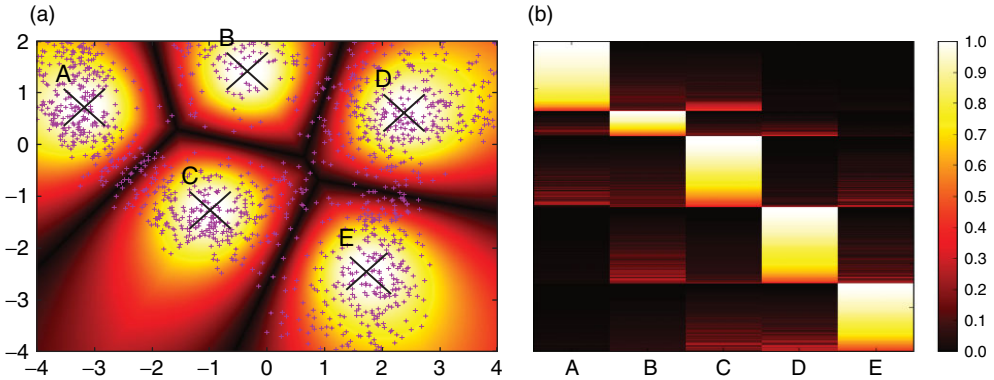


Fig. 2 Panel (a): A pointwise stability graph for the same data as used in Figure 1 but now the data points are also plotted along with the five cluster centers. Points on the boundary regions decrease the overall stability, while points near the cluster centers increase the overall stability. Panel (b): The heatmap for the data in Panel (a) showing how the points behave under perturbation. Note the decrease in stability from top to bottom in blocks on the main diagonal and the increase in stability from top to bottom in blocks off the main diagonal. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

These separate the clusters. On the other hand, as θ increases, the prior concentrates around 1 so more mass is on values of λ that are close to 1. Thus the dark regions between the clusters shrink, indicating that perturbations of points into those regions are more and more unlikely. For small values of θ, the perturbations are much larger so perturbations of points by the factors λ_k to regions outside their cluster is easier. This is indicated by the darker regions that become more extensive relative to the regions of cluster stability as θ gets smaller.

2.2. A Heatmap Plot of Clustering Behavior

For a visualization of stability that goes beyond two dimensions we propose a sorted heatmap plot of the averaged assignment matrix. Simply take the ϕ_{ik} from Eq. (4) and recall that i indexes the data points while k indexes the clusters. Therefore, in the matrix $\mathbf{\Phi}(K, \mathcal{C})$, we can sort over cluster first so that the first collection of rows corresponds to the data points in cluster 1, the second collection of rows corresponds to the data points in cluster 2, and so on up to K. Then, within each collection of rows representing a cluster, say k_0, sort the rows by putting the stabilities ϕ_{i,k_0} in decreasing order from top to bottom as i ranges over the data points in cluster k_0 (and k_0 ranges from 1 to K). Then, assign colors to ranges of values of the entries of the resulting matrix, higher values corresponding to lighter colors. We call the result a stability heatmap for a clustering.

An example of a stability heatmap can be derived from panel (a) of Fig. 2 and is shown in panel (b). Panel (a) indicates a level of pointwise stability between the top two panels of Fig. 1 because it used a data-driven value of θ (found using the technique of Section 6). Converting panel (a) to a heatmap is straightforward: The horizontal axis represents an ordering of the clusters from 1–5. Within the upper left block the $\phi_{i,1}$'s for x_i in cluster 1 are indicated by the gradual darkening from top to bottom. The block immediately below it represents the $\phi_{i,1}$'s for x_i in cluster 2 and the block immediately to its right represents the $\phi_{i,2}$'s for x_i in cluster 1. It is seen that the blocks on the main diagonal are quite light indicating that the clusters are quite stable even though within each block the color darkens a bit from the top to the bottom.

One can also look along the rows of the heatmap to see how much 'instability' there is between the clusters that the blocks on a row represent. Lighter colors in the blocks on the main diagonal (or equivalently darker colors in the blocks off the main diagonal) mean less instability in the sense that it takes larger perturbation factors to move a point from one cluster to another. For instance, in the fourth row of blocks it is seen that the second block has some light lines at the bottom suggesting an interaction, if mild, with the fourth and fifth blocks in that row.

In these heatmaps, comparisons must be done keeping in mind that only the rows sum to one; comparisons within columns do not in general. However, it is tempting to apply the same interpretation vertically to generate suggestions as to which clusters are unstable—provided a separate stability analysis were done to ensure the appearance of interchangability of points in the two clusters was not just an artifact of the method of construction of the heatmap.

EXAMPLE 1: MNIST Data Analysis. As an illustration of what the stability heatmap looks like in practice, consider the popular MNIST data set containing 1934 handwritten digits in 0 through 9, each given as a 32×32 pixel binary image, see ref. 14. This is a classification data set in which each binary image corresponds to a single digit label. So, we know unambiguously what the 10 clusters must represent

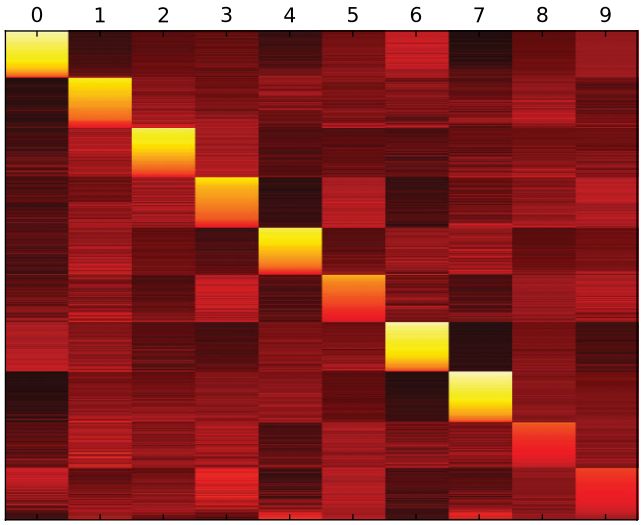


Fig. 3 Heatmap formed from the points in the MNIST data set treated as a clustering problem. Lighter blocks on the diagonal are more stable than darker blocks, while lighter blocks off the diagonal indicate mutual instability. The blocks on the main diagonal are generally brighter than any of the other blocks indicating that the images for each digit tend to be more similar to other digits in their class than to other those in other classes. However, many off-diagonal blocks are also bright, indicating many overlapping classes. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

and which images belong to each 'cluster'. If we treat this as a clustering in which cluster membership is given by class labels, we can evaluate its stability—in essence, we are asking which the classes are stable and which classes are easily confused.

To evaluate the ϕ_{ik}'s, we first specify a distance and prior. As the images are given as binary matrices $x = (x_{uv})$ and $y = (y_{uv})$ where $u, v = 1, \ldots, 32$, there is no obvious metric to use to form ϕ_{ik}. However, consider

$$\text{dist}(x, y) = \min_{\alpha,v\in\{0,1,\ldots,31\}} \sum_{u,v\in\{1,\ldots,32\}} |x_{u+\alpha,v+v} - y_{uv}|, \quad (6)$$

where we assume that $x_{u'v'} = 0$ if $u' > 32$ or $v' > 32$. Effectively dist takes the minimum L^1 norm (or, since the pixels are binary, Hamming distance) over all horizontal and vertical shifts. Clearly, dist is not a metric—a distance of zero does not imply equality—but it does ensure that x is properly aligned with y. We take the distance between a point x_i and a cluster—$\delta(x_i, C_k)$ in 2—to be the average of all distances, in the sense of Eq. (6), from x_i to the points in the cluster under consideration. This means the set $\hat{S}^*_{ik}$ can be defined. If we again assign a shifted exponential prior to the ten λ_k's, and choose θ as in Section 6, we can find the ϕ_{ik}'s.

Figure 3 shows the heatmap for the MNIST data set. Each column, numbered zero through nine, corresponds to a digit and the blocks on the main diagonal indicate the probability that a data point remains in its cluster under perturbation. It is seen that 0, 6, and and 7 are relatively stable digits while 8 and 9 are the least stable. Brighter colored blocks off the main diagonal show which instances of a digit are located on a boundary with at least one of the other digits. For example, many 0's are near the boundary of 6 but few 0's are near the boundaries of 1, 4, and 7. Similarly, 3 is often near the boundary of 2, 5, and 9. Another metric may yield better results, this example shows how the heatmap reveals the stability of a clustering.

3. THEORY FOR THE AVERAGED ASSIGNMENT MATRIX

To be more formal, suppose we have n independent and identical (IID) random d-dimensional variables $X_1, \ldots, X_n$ with outcomes $x_1, \ldots, x_n$ denoted $\mathbb{D}$. Write P to mean the probability of any of the X_i's. A clustering is a partition of collection of $\mathbb{D}$ into K sets $\hat{\mathcal{C}} = \hat{\mathcal{C}}_K = \{\hat{C}_1, \ldots, \hat{C}_K\}$. Regarding a clustering as a partition of the data can be equivalent to partitioning $\mathbb{R}^d$. For instance, if we write $\hat{\mu}_k$ to be the centroid of cluster $\hat{C}_k$ for $k = 1, \ldots, K$ and use a centroid-based clustering technique we can define, for any $x \in \mathbb{R}^d$,

$$x \in \hat{C}_k \Leftrightarrow \forall \ell \neq k \quad d(x, \hat{\mu}_k) \leq d(x, \hat{\mu}_\ell), \quad (7)$$

where d is a metric on $\mathbb{R}^d$. One standard centroid is the empirical conditional mean used in K-means clustering and defined by

$$\hat{\mu}_k = \frac{1}{|C_k|} \sum_{i\in\hat{C}_k} x_i. \quad (8)$$

Now $\hat{\mathcal{C}}_K$ can be regarded as a partition of the data or as a partition of $\mathbb{R}^d$ defined by the data and the context will make clear which is meant.

Note that the 'hat' on the $\mathcal{C}$ and C_k's indicates that the clustering is chosen using the data. If we regard the clustering as a set of regions in $\mathbb{R}^d$ then we can write the 'limits' of the $\hat{C}_k$'s as n increases as C_k's. So, $\mathcal{C}_K = \{C_1, \ldots, C_K\}$ is a population quantity, a disjoint and exhaustive partition of $\mathbb{R}^d$ into convex sets, reflecting the limiting behavior of $\hat{\mathcal{C}}_K$ with large n. Now, we can set $\mu_k = E(X\mathbb{I}_{C_k}(X))/P(C_k) = E(X|C_k)$ where $\mathbb{I}_A(U)$ is the indicator function for a random variable U to be in a set A. Thus, we also require that, parallel to Eq. (7),

$$x \in C_k \Leftrightarrow \forall \ell \neq k \quad d(x, \mu_k) \leq d(x, \mu_\ell). \quad (9)$$

In this case, we want to ensure that $\hat{C}_k \to C_k$ for each k. One natural criterion is

$$P(\hat{C}_k \triangle C_k) \to 0, \quad \text{as} \quad n \to \infty, \tag{10}$$

where $\hat{C}_k \triangle C_k$ is regarded as a set in $\mathbb{R}^d$, P indicates the probability on $\mathbb{R}^d$, and the mode of convergence is in the distribution P^n of $X_1, \ldots, X_n$ used to form the estimate $\hat{C}_k$ of C_k. If Eq. (10) holds, then $\hat{\mu}_k \to \mu_k$ in P. However, the converse fails in general. That is, if $\hat{\mu}_k \to \mu_k$ in P for all k for a centroid based clustering satisfying Eq. 7, then Eq. (10) does not necessarily hold for all k. We will get around this by assuming Eq. (9), that is, that the limiting clustering is centroid based as well.

Now we can define a Bayesian notion of cluster stability. It has two forms, one for empirical clusterings and one for population clusterings. We start with the empirical form. Given K, let $\Lambda_1, \ldots, \Lambda_K$ be K IID non-negative real valued random variables with continuous marginal distribution F and outcomes denoted $\lambda_1, \ldots, \lambda_K$. Now, for fixed i, k, $\hat{\mathcal{C}}$, and $\mathbb{D}$, recall the set

$$\hat{S}_{ik} = \left\{\lambda^K : \forall \ell \neq k \ \lambda_j d(x_i, \hat{\mu}_k) \leq \lambda_\ell d(x_i, \hat{\mu}_\ell)\right\}, \tag{11}$$

where $\lambda^K = (\lambda_1, \ldots, \lambda_K)$. When x_i is close to $\hat{\mu}_k$, the set of λ_k and λ_ℓ's for $\ell \neq k$ for which $\mathbb{I}_{\hat{S}_{ik}}(\lambda^K)$ is one is large. Likewise, if the $\hat{\mu}_\ell$'s for $\ell \neq k$ are far from $\hat{\mu}_k$, the set of λ_k and the λ_ℓ's for $\ell \neq k$ for which $\mathbb{I}_{\hat{S}_{ik}}(\lambda^K)$ is one is large. That is, in the distribution F, when (i) x_i is near the centroid $\hat{\mu}_k$ of $\hat{C}_k$ or (ii) the $\hat{\mu}_\ell$'s are far from $\hat{\mu}_k$ we have $P(\mathbb{I}_{\hat{S}_{ik}}(\Lambda^K) = 1)$ is large and it is seen that (i) and (ii) mean cluster k is stable. (Note that we use $P = P_\Lambda$ for the probability associated with the Λ_k's as well as the X_i's, dropping the Λ when the context makes it clear which is meant.)

However, when $x_i \in \hat{C}_k$ is not close to $\hat{\mu}_k$, for example, is close to the boundary of $\hat{C}_k$, or the centroids of the $\hat{C}_\ell$'s (for $\ell \neq k$) are not far from $\hat{\mu}_k$ then $P(\mathbb{I}_{\hat{S}_{ik}}(\lambda^K) = 1)$ is small. In fact, if F is strictly increasing on $[0, \infty)$, $P(\mathbb{I}_{\hat{S}_{ik}}(\lambda^K) = 1)$ is (strictly) increasing as a function of $d(x_i, \hat{\mu}_k)$ and decreasing in the $d(x_i, \hat{\mu}_\ell)$'s (though not strictly because there are usually many values of ℓ). That is, when $\hat{C}_k$ is not stable in the sense of x_i only weakly representing it or there are other cluster centers close enough to compete effectively with $\hat{C}_k$ to represent x_i, $P(\mathbb{I}_{\hat{S}_{ik}}(\lambda^K) = 1)$ tends to be small. Thus, $P(\mathbb{I}_{\hat{S}_{ik}}(\lambda^K) = 1)$ is an assessment of how concentrated and separated the clusters are, that is, how stable $\hat{C}_k$ is as a summary of x_i. The exact trade-off between concentration and separation for $\hat{\mathcal{C}}_K$ depends on the specific choice of metric d and prior F.

We can combine these assessments over i and k to get an overall stability for the clustering. The empirical form of our Bayesian clustering stability criterion is, for fixed K, $\mathbb{D}$, and $\hat{\mathcal{C}}_K$, the average over instances of $\hat{S}_{ik}$ given by

$$Q_n(K) = \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{x_i \in \hat{C}_k\}} \times \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(x_i, \hat{\mu}_k) \leq \lambda_\ell d(x_i, \hat{\mu}_\ell)\}} \mathrm{d}F(\lambda^K). \tag{12}$$

For fixed K and $\mathcal{C}$ the population form of Eq. (12) replaces $\hat{S}_{ik}$ with its population form S_{ik} using the μ_k's in place of the $\hat{\mu}_k$'s and takes an expectation. The result is

$$Q_\infty(K) = \sum_{k=1}^{K} E\ \mathbb{I}_{\{X \in C_k\}} \times \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(X, \mu_k) \leq \lambda_\ell d(X, \mu_\ell)\}} \mathrm{d}F(\lambda^K). \tag{13}$$

For ease of notation, let

$$\hat{\phi}_k(x) = \mathbb{I}_{\{x \in \hat{C}_k\}} \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(x, \hat{\mu}_k) \leq \lambda_\ell d(x, \hat{\mu}_\ell)\}} \mathrm{d}F(\lambda^K)$$

and let

$$\phi_k(X) = E\ \mathbb{I}_{\{X \in C_k\}} \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(X, \mu_k) \leq \lambda_\ell d(X, \mu_\ell)\}} \mathrm{d}F(\lambda^K).$$

These are versions of ϕ_{ik} where the argument corresponding to i is replaced by x or X. Now Eqs. 12 and 13 become

$$Q_n(K) = \sum_{k=1}^{K} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_k(x_i) \right) \text{ and } Q_\infty(K) = \sum_{k=1}^{K} \phi_k(X). \tag{14}$$

3.1. Convergence of Q_n to Q_∞

In view of Eq. (14), we can show $Q_n(K)$ converges to $Q_\infty(K)$ largely because it is an instance of the law of large numbers. Moreover, although we have assumed Eq. 10, in fact Eq. (10) can be derived from the convergence properties of the cluster means $\hat{\mu}_k$ under Eqs. 7 and 9. These converence properties are built into the proof of the theorem below to make the result simpler to state.

THEOREM 1: Fix K and assume Eqs. 7 and 9. Then, if, $\forall \ell = 1, \ldots, K$, $\hat{\mu}_\ell \to \mu_\ell$, we have that

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_k(X_i) \to \phi_k(X), \tag{15}$$

and hence $Q_n(K) \to Q_\infty(K)$, as $n \to \infty$ in probability P^n, the distribution of $X_1, \ldots, X_n$.

Remark: It is easy to determine if Eq. (7) is satisfied because it is a property of the clustering procedure. However, Eq. (9) amounts to an assumption of the true clustering which is hard to verify and may not hold if the level sets of P around its modes are nonconvex. The hypotheses that the $\hat{\mu}_\ell$'s converge to their respective μ_ℓ's is also hard to verify, but can usually be safely assumed since it rests primarily on a law of large numbers applied to Eq. (8).

Proof: The proof follows from deriving bounds on $|Q_n(K) - Q_\infty(K)|$ so that the law of large numbers can be invoked. The details are given in Appendix A. ■

3.2. Choosing K Consistently

Having established that the empirical form of the stability assessment converges to the population form, we proceed to demonstrate that stability can be used to choose the number of clusters consistently. This is based on the idea that a more stable clustering is to be preferred over a less stable clustering and hence high values of the stability criterion are desired.

By changing our perspective, we can regard $Q_n(K)$ as a data dependent objective function for choosing K. Let K_1 be a relatively small positive integer and let $K_2 > K_1$ be a relatively large but finite integer. Let $[K_1, K_2]$ be the (compact) set of integers strictly between $K_1 - 1$ and $K_2 + 1$ and write

$$\hat{K} = \operatorname*{argmax}_{K \in [K_1, K_2]} Q_n(K). \tag{16}$$

Theorem 1 established that, pointwise in K, $Q_n(K) \to Q_\infty(K)$. Thus, for any bounded interval $[K_1, K_2]$ and $\epsilon > 0$, we have that

$$P\Big(\sup_{K \in [K_1, K_2]} |Q_n(K) - Q_\infty(K)| > \epsilon \Big) \to 0.$$

That is, $Q_n(K) \to Q_\infty(K)$ uniformly in probability on $[K_1, K_2]$. Let

$$K_{\text{opt}} = \operatorname*{argmax}_{K \in [K_1, K_2]} Q_\infty(K).$$

We have the following.

THEOREM 2: Suppose that $K_{\text{opt}} \in [K_1, K_2]$ is the unique maximum of $Q_\infty(K)$ over K. Then,

$$\operatorname*{argmax}_{K \in [K_1, K_2]} Q_n(K) \to \operatorname*{argmax}_{K \in [K_1, K_2]} Q_\infty(K), \tag{17}$$

that is $\hat{K} \to K_{\text{opt}}$, in P as $n \to \infty$.

Proof. Convergence Eq. (17) follows from a simple modification of Theorem 2.1 in ref. 15, see p. 2121. Specifically, Theorem 2.1 requires that the limiting objective function, here $Q_\infty(K)$, be continuous as a function of the parameter, here K. In fact, the continuity is only used at one step in the proof (p. 2122) to ensure a separation between values $Q_\infty(K_{\text{opt}})$ and $Q_\infty(K)$ for $K \neq K_{\text{opt}}$. This separation holds trivially for discrete parameters such as K. So, when $Q_\infty(\cdot)$ has a unique maximum in $[K_1, K_2]$ the step can be accomplished. So, the modified form of Theorem 2.1 can be used here to give Eq. (17).

4. THE STABILITY CRITERION

In this section we verify that it is reasonable to interpret Q_∞, and hence Q_n, as an assessment of stability in the sense that it is strongly influenced by how much mass the true distribution assigns on the boundary regions between clusters in a clustering. Indeed, we formalize the intuition that as a data point x_i moves closer to its assigned cluster center, stability increases; as x_i moves further from its cluster center, stability decreases. As a separate point, we argue that K's giving large values of $Q_\infty(\cdot)$ or its empirical version $Q_n(\cdot)$, that is, K_{opt} and $\hat{K}$, often correspond to an intuitively correct number of clusters.

4.1. Intuition behind Q_∞

As a simple motivating example, consider the case of two clusters in two dimensions. For simplicity, assume that these clusters have centroids at $(-1, 0)$ and $(1, 0)$, as shown in Fig. 4a. Thus the clusters here may be defined by the regions $C_1 = \{(x, y) : x \leq 0\}$ and $C_2 = \{(x, y) : x > 0\}$.

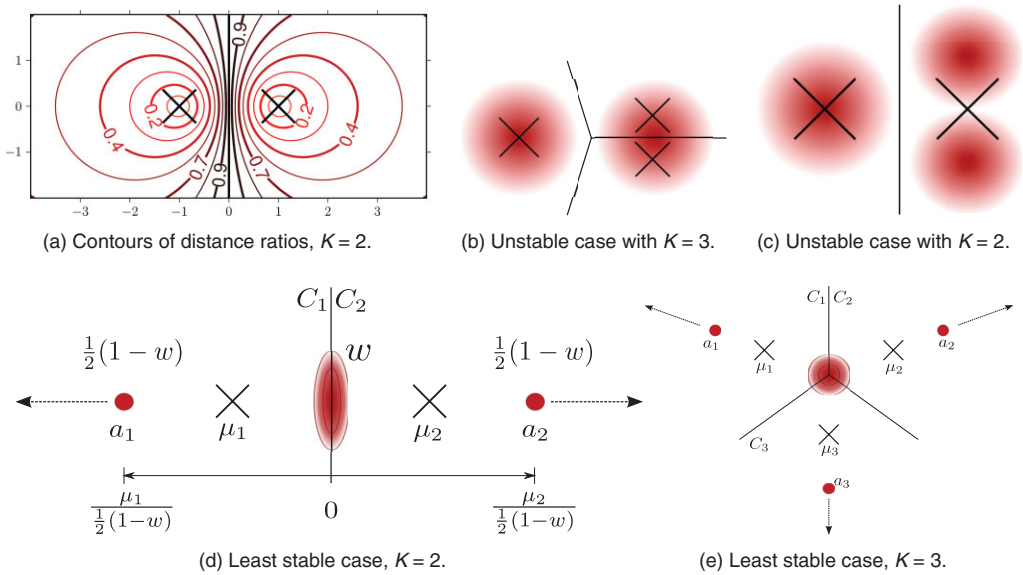
Now, in the $K = 2$ case, the stability criteria of Eq. (13) can be written as

$$\begin{aligned} Q_\infty(2) &= E_X P(D_1/D_2 \leq \Lambda_2/\Lambda_1) \mathbb{I}_{C_1}(X) \\ &\quad + E_X P(D_2/D_1 \leq \Lambda_1/\Lambda_2) \mathbb{I}_{C_2}(X) \qquad (18) \\ &= P_X(C_1) E_X(G_{12}(D_1/D_2)|C_1) \\ &\quad + P_X(C_2) E_X(G_{21}(D_2/D_1)|C_2), \qquad (19) \end{aligned}$$

where $\Lambda_1, \Lambda_2 \sim F$ are drawn from the perturbation prior F and

$$\begin{aligned} G_{12}(t) &= P(\Lambda_1/\Lambda_2 \geq t) \\ G_{21}(t) &= P(\Lambda_2/\Lambda_1 \geq t) \end{aligned} \tag{20}$$

are the survivor functions of the random variables Λ_1/Λ_2 and Λ_2/Λ_1, respectively. Here, $D_1 = d(X, \mu_1)$ and $D_2 =$



(a) Contours of distance ratios, $K = 2$. (b) Unstable case with $K = 3$. (c) Unstable case with $K = 2$.

(d) Least stable case, $K = 2$. (e) Least stable case, $K = 3$.

Fig. 4 (a) An illustration of our stability in 2d. Two centroids are placed at $(-1, 0)$ and $(1, 0)$, and contour lines show the level sets of D_1/D_2 for $x \leq 0$ and D_2/D_1 for $x > 0$. The stability decreases along the contours. (b) $K = 3$ is proposed, but $K = 2$ is correct. (c) $K = 2$ is proposed, but $K = 3$ is correct. (d) A case where $Q_\infty(2)$ achieve its minimum; as $w \to 0$, μ_1 is constant, but all mass becomes concentrated on boundary regions. (e) A case where $Q_\infty(3)$ achieves its minimum; mass becomes concentrated at a point equidistant to each of the three centroids. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$d(X, \mu_2)$ are regarded as random variables often truncated to C_1 and C_2, respectively.

Note that these survivor functions G_{12} and G_{21} are monotonically non-increasing, with $G_\cdot(0) = 1$, as $\Lambda_1, \Lambda_2 \geq 0$, and $G_\cdot(1) = 1/2$, as $P(\Lambda_1 \geq \Lambda_2) = 1/2$. Relating this to Eq. (19), G_{12} and G_{21} are monotonically non-increasing with increasing values of the distance ratio. Thus, as mass in the distribution of X is moved away from the centroids, the value of $Q_\infty(2)$ decreases to 1/2; as the distribution of X becomes more concentrated around the two centroids, $Q_\infty(2)$ increases to 1. Note that while the perturbation prior F determines the rate of increase or decrease, monotonicity is guaranteed for any F that is positive a.e. on $[0, \infty)$.

A two-dimensional illustration of this is shown in Fig. 4a. We here plot the contours of the ratio of distances to the two centroids, located at $(-1, 0)$ and $(1, 0)$ assuming the covariances of the two components in P are identical and have their eigenvalues along the x and y axes. It is seen that, for any F, the contribution of a data point to the stability assessment of the clustering is a function of its location in the contour plot. Contours near the boundary at $x = 0$ are less stable than those close to the centroids.

As a generality, the intuition of our stability measure closely follows that of Fischer's Linear Discriminant, defined as the ratio of within-class variance to between-class variance. In particular, when the mass of X is concentrated around the two centroids, Fischer's linear discriminant has its best error rates and and the $Q_\infty(K)$'s achieve their most stable values. Likewise, when the masss of X is spread widely around the two centroids, Fisher's linear discriminant gives poor classification and the $Q_\infty(K)$'s achieve their least stable values, that is, D_1/D_2 and D_2/D_1 are likely to be on contours far from their respective centroids.

Now, to motivate the intuition of how our method can be used to choose the correct number of clusters, we wish to consider simple examples: $K = 2$ is correct, but the candidate clustering contains three clusters, and $K = 3$ is correct but the candidate clustering contains two clusters. The first case is shown in Fig. 4b. Here, we would expect $Q_\infty(3) < Q_\infty(2)$, as one of the modes is split in the $K = 3$ case (shown), and mass near the boundary in the split mode decreases the overall stability. The second case is shown in Fig. 4c. Here, we would expect $Q_\infty(3) > Q_\infty(2)$, as little of the mass in the second cluster in the $K = 2$ case (shown) is near the most stable region of the cluster. In both of these cases, were the correct clusterings shown, most of the mass would be concentrated in regions of high stability. It is easy to see that the stability of a well chosen clustering will be higher than the stability of a poorly chosen clustering, an intuition which we explore depth in the subsequent sections.

Finally, consider a case in which $Q_\infty(2)$ achieves $1/2$, its lower bound. In Fig. 4d, we show a single mode concentrated along the border of C_1 and C_2 and two modes with mass w at $2\mu_1/(1-w)$ and $2\mu_2/(1-w)$. As $w \to 0$, all the mass concentrates in regions where $D_1 \simeq D_2$ and thus $E_X G_{12}(D_1/D_2) \simeq E_X G_{21}(D_2/D_1) \simeq 1/2$. Here, one could argue that one or three clusters is correct (depending

on the mass at a_1 and a_2), and it is clear that $Q_\infty(K)$ would be higher in both cases. Similarly, for $K = 3$, the lower bound of $Q_\infty(3)$, $1/3$, is achieved by analogously concentrating most of the mass at a point equidistant to each of the three centroids Fig. 4e.

4.2. Properties of $Q_\infty(\cdot)$ for General K

The foregoing intuition extends to higher dimensions and to general K. To see this, consider the clustering $\mathcal{C} = \{C_1, \ldots, C_K\}$. Write $\mu_k = E(X|C_k)$ and $D_k = d(X, \mu_k)$ for $k = 1, \ldots, K$ and let $R_{ij} = \Lambda_j/\Lambda_i$ for $i, j = 1, \ldots, K$. Now, analogously to Eq. (20), consider the leave-out-one survivor functions defined for $k = 1$ and $k = 2$ as

$$\begin{aligned} G_{\widehat{1,1}}(t_2, \ldots, t_K) &= G_{12,\ldots,1K}(t_2, \ldots, t_K) \\ &= P(R_{12} \geq t_2, \ldots, R_{1K} \geq t_K), \\ G_{\widehat{2,2}}(t_1, t_3, \ldots, t_K) &= G_{21,23,\ldots,2K}(t_1, t_3, \ldots, t_K) \\ &= P(R_{21} \geq t_1, R_{23} \geq t_3, \ldots, R_{2K} \geq t_K), \end{aligned}$$

where $\widehat{1,1}$, for instance, means that the first cluster is left out; $\widehat{2,2}$ is similar and we use the hat-notation for omission without further comment. In general, the leave-out-k survivor function is

$$\begin{aligned} G_{\widehat{k,k}}(t_1, \ldots, \hat{t}_k, \ldots, t_K) &= G_{k1,\ldots,\widehat{kk},\ldots,KK}(t_1, \ldots, \hat{t}_k, \ldots t_K) \\ &= P(R_{k1} \geq t_1, \ldots, \widehat{R_{kk} \geq t_k}, \ldots, R_{KK} \geq t_K). \end{aligned}$$

Now, parallel to Eq. (18) we have

$$\begin{aligned} Q_\infty(K) &= E_X P_\Lambda(D_1/D_2 \leq \Lambda_2/\Lambda_1, \ldots, D_1/D_K \\ &\quad \leq \Lambda_K/\Lambda_1)\mathbb{I}_{C_1} + E_X P_\Lambda(D_2/D_1 \\ &\quad \leq \Lambda_1/\Lambda_2, D_2/D_3 \leq \Lambda_3/\Lambda_2, \ldots, D_2/D_K \\ &\quad \leq \Lambda_K/\Lambda_2)\mathbb{I}_{C_2} + \cdots + E_X P_\Lambda(D_K/D_1 \\ &\quad \leq \Lambda_1/\Lambda_K, \ldots, D_K/D_{K-1} \leq \Lambda_{K-1}/\Lambda_K)\mathbb{I}_{C_K} \\ &= P(C_1)E(G_{\widehat{1,1}}(D_1/D_2, \ldots, D_1/D_K)|C_1) \\ &\quad + P(C_2)E(G_{\widehat{2,2}}(D_2/D_1, D_2/D_3, \ldots, \\ &\quad \times D_2/D_K)|C_2) \\ &\quad + \cdots + P(C_K)E(G_{\widehat{K,K}}(D_K/D_1, \ldots, \\ &\quad \times D_K/D_{K-1})|C_K). \end{aligned} \tag{21}$$

Also, it is easy to see that for each k, $G_{\widehat{k,k}}(\mathbf{0}_{K-1}) = 1$ where $\mathbf{0}_{K-1}$ is a vector of 0's of length $K - 1$.

We can also see that $G_{\widehat{k,k}}(\mathbf{1}_{K-1}) = 1/K$. By the definition of the R_{1k}'s,

$$\begin{aligned} G_{\widehat{1,1}}(\mathbf{1}_{K-1}) &= P(\Lambda_2/\Lambda_1 \geq 1, \ldots, \Lambda_K/\Lambda_1 \geq 1) \\ &= P(\Lambda_2 \geq \Lambda_1, \ldots, \Lambda_K \geq \Lambda_1) \\ &= P(\Lambda_1 \leq \min(\Lambda_2, \ldots, \Lambda_K)). \end{aligned}$$

However, assuming continuous $\Lambda_\cdot$'s,

$$\sum_{k=1}^{K} P(\Lambda_k \leq \min(\Lambda_1, \ldots, \hat{\Lambda}_k, \ldots, \Lambda_K) = 1.$$

So, by symmetry, all terms in the sum are equal. That is, $P(\Lambda_k \leq \min(\Lambda_1, \ldots, \hat{\Lambda}_k, \ldots, \Lambda_K) = 1/K$. The $G_{\widehat{k,k}}$'s for $k \neq 1$ are similar.

Now, it is easy to see that on C_k, $0 \leq D_k/D_\ell \leq 1$ for $\ell \neq k$. In addition, as $(s_1, \ldots, s_{K-1}) \in \mathbb{R}^{K-1}$ increases from $\mathbf{0}_{K-1}$ to $\mathbf{1}_{K-1}$ along a curve in the $K-1$ unit cube we see that each $G_{\widehat{k,k}}(s_1, \ldots, s_{K-1})$ decreases from 1 to $1/K$. (Here, increasing means that $(s_1, \ldots, s_{K-1}) \leq (s'_1, \ldots, s'_{K-1}) \iff \forall k\ s_k \leq s'_k$.) Thus, we can generalize the bounds on $Q_\infty(\cdot)$ to

$$1/K \leq Q_\infty(K) \leq 1.$$

As in the $K = 2$ case, these bounds are achievable. Indeed, if we consider a sequence of distributions P that concentrate at K centroids we get $D_k/D_\ell \to 0$ on C_k for $\ell \neq k$. Therefore, Eq. (21) gives

$$\begin{aligned} Q_\infty(K) &\to P(C_1)G_{\widehat{1,1}}(\mathbf{0}_{K-1}) + \cdots + P(C_K)G_{\widehat{K,K}}(\mathbf{0}_{K-1}) \\ &= \sum_{k=1}^{K} P(C_k) = 1, \end{aligned}$$

corresponding to the highly stable case of the mass of P concentrating at the centroids μ_k.

In addition, Fig. 4e shows the generic form of a sequence of P's for which the lower bound is achievable. There are three sectors but the argument is analogous for four or more sectors. The density of P concentrates on a shrinking disc at the center and each sector has a point, labeled a_1, a_2, and a_3 with mass near it so that $E_P(X|C_k) = \mu_k$ is constant as P varies. As P concentrates at the center, and the points a_i move to infinity in the direction of the arrows, the mass near the a_k's decreases so that $D_k/D_\ell \to 1$ on C_k, for $k \neq \ell$. Now, Eq. (21) gives

$$\begin{aligned} Q_\infty(K) &\to P(C_1)G_{\widehat{1,1}}(\mathbf{1}_{K-1}) + \cdots + P(C_K)G_{\widehat{K,K}}(\mathbf{1}_{K-1}) \\ &= (1/K)\sum_{k=1}^{K} P(C_k) = 1/K, \end{aligned}$$

the lowest stability value. Parallel to the $K = 2$ case, depending on the mass at a_1, a_2, a_3, one can argue that a single central cluster or four clusters are reasonable depending on how much mass there is at the three points.

Thus, even though the range for general K is $[1/K, 1]$, the high values still correspond to stability with little mass near the boundary regions while the low values correspond to instability in the sense that most of the mass is on the boundary regions. Indeed, whenever $D_k/D_\ell \to 0$ on C_k for $k \neq \ell$, $Q_\infty(K) \to 1$; see Proposition 1 below for the two natural cases in which this happens.

Finally for this subsection, we comment on the role of the distribution $F(\cdot)$ of the Λ_k's. The more F concentrates as a point, the easier it is for an inequality like $\Lambda_1 D_1 \leq \min_{\ell \neq 1} \Lambda_\ell D_\ell$ to be satisfied when $D_1 < \min_{\ell \neq 1} D_\ell$. Likewise, the more the mass of the Λ_k's is spread out, the harder it is for $\Lambda_1 D_1 \leq \min_{\ell \neq 1} \Lambda_\ell D_\ell$ to be satisfied when $D_1 < \min_{\ell \neq 1} D_\ell$ (see Fig. 1).

4.3. Verification That Two Stable Settings Have High Stability

Our overall strategy will be to choose K, and more generally clusterings, so that $Q_n(K)$ and thus $Q_\infty(K)$ will be large. We have informally observed that large values of $Q_n(K)$ and $Q_\infty(K)$ correspond to clusters that are concentrated around their centroids or cluster centers that are separated (provided that the degree of concentration does not decrease too quickly as the centroids separate). To strengthen this intuition, we present the following.

PROPOSITION 1: Suppose P has K distinct centroids μ_k for $k = 1, \ldots, K$, with $\forall k : P(C_k) > 0$. Then, (i) if P concentrates at the μ_k's, that is, $d(X, \mu_k)\mathbb{I}_{C_k} \to 0$ in P-probability we have that

$$Q_\infty(K) \to 1;$$

and (ii) if a sequence of P's satisfies $\min_{k \neq \ell} d(\mu_k, \mu_\ell) \to \infty$ and the $d(X, \mu_k)\mathbb{I}_{C_k}$'s are bounded above in probability, that is, $d(X, \mu_k)\mathbb{I}_{X_k} = \mathcal{O}_P(1)$, then again we get $Q_\infty(K) \to 1$.

Remark: In the special case that d is Euclidean distance, the hypothesis of (i) means that the conditional variance of X, given C_k, goes to zero, if it exists, by a uniform integrability argument. Also, by Markov's inequality, (ii) implies that $Q_\infty(K)$ goes to one if all the conditional variances given C_k are bounded above and the μ_k's separate. Essentially, this ensures there will be a region separating the centroids from each other on which there is vanishingly small probability.

Proof: To prove (i), suppose that the μ_k's are fixed and distinct, and let $\epsilon > 0$. Then, $P(d(X, \mu_k) \geq \epsilon | C_k) \to 0$ and for $\ell \neq k$, $P(d(X, \mu_\ell) \geq \epsilon | C_k) \to 1$. It is seen that the integrand of $\phi_k(X)$ is uniformly integrable for any sequence of P's because it is bounded by one. So, as P concentrates, the definition of $\phi_k(X)$ gives that

$$\phi_k(X) \to E\mathbb{I}_{X \in C_k} \int \mathbb{I}_{\{\forall \ell \neq k : 0 \leq \lambda_\ell d(X, \mu_\ell)\}} \mathrm{d}F(\lambda^K) \to E\mathbb{I}_{X \in C_k} \times \int \mathbb{I}_{\{\forall \ell \neq k : 0 \leq \epsilon\}} \mathrm{d}F(\lambda^K) = P(C_k), \tag{22}$$

and therefore $Q_\infty(K) \to \sum_{k=1}^K P(C_k) = 1$.
To prove (ii), note that for $\ell \neq k$, if $X \in C_k$ that

$$d(X, \mu_\ell) \to \infty, \tag{23}$$

and that there is a B so that

$$P(d(X, \mu_k) \leq B) \geq 1 - \eta$$

for any preassigned $\eta > 0$, as P varies. Now write,

$$\begin{aligned} \phi_k(X) = {} & E\mathbb{I}_{\{x \in C_k\}} \mathbb{I}_{d(X, \mu_k) \leq B} \\ & \times \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(x_i, \mu_k) \leq \lambda_\ell d(x, \mu_\ell)\}} \mathrm{d}F(\lambda^K) \\ & + E\mathbb{I}_{\{x \in C_k\}} \mathbb{I}_{d(X, \mu_k) > B} \\ & \times \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(x_i, \mu_k) \leq \lambda_\ell d(x, \mu_\ell)\}} \mathrm{d}F(\lambda^K). \end{aligned} \tag{24}$$

By the bounded convergence theorem, the second term goes to zero because the integral and first indicator function are both bounded by one and $\mathbb{I}_{d(X, \mu_k) > B} \to 0$ in probability as $\eta \to \infty$. The first term is bounded below by

$$E\mathbb{I}_{\{x \in C_k\}} \mathbb{I}_{d(X, \mu_k) \leq B} \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k B \leq \lambda_\ell d(x, \mu_\ell)\}} \mathrm{d}F \times (\lambda_1, \ldots, \hat{\lambda}_k, \ldots, \lambda_K) \mathrm{d}F(\lambda_k), \tag{25}$$

where $\hat{\lambda}_k$ means λ_k is omitted from the integration. For fixed λ_k, Eq. (23) implies that

$$\mathbb{I}_{d(X, \mu_k) \leq B} \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k B \leq \lambda_\ell d(x, \mu_\ell)\}} \mathrm{d}F \times (\lambda_1, \ldots, \hat{\lambda}_k, \ldots, \lambda_K) \to 1 \tag{26}$$

in probability. (To see this, consider the set $\{\min(\lambda_1, \ldots, \hat{\lambda}_k, \ldots, \lambda_K) > \eta\}$, note the right hand side of Eq. (26) is bigger

than $\epsilon > 0$ for n large enough, and then let $\eta \to 0$.) Using Eq. (26) in Eq. (25) gives

$$E\mathbb{I}_{\{x\in C_k\}}\mathbb{I}_{d(X,\mu_k)\le B}\int \mathbb{I}_{\{\lambda_k<\xi \quad \text{and} \quad \forall \ell\neq k \lambda_\ell>\eta\}}\mathrm{d}F \times (\lambda_1,\ldots,\hat{\lambda}_k,\ldots,\lambda_K)\mathrm{d}F(\lambda_k) + E\mathbb{I}_{\{x\in C_k\}}\mathbb{I}_{d(X,\mu_k)\le B}\int \mathbb{I}_{\{\lambda_k>\xi \quad \text{and} \quad \forall \ell\neq k \lambda_\ell>\eta\}}\mathrm{d}F \times (\lambda_1,\ldots,\hat{\lambda}_k,\ldots,\lambda_K)\mathrm{d}F(\lambda_k) \quad (27)$$

as an asymptotic lower bound on the first term of Eq. (24). Clearly, as $\xi \to 0$, the first term in Eq. (27) goes to zero and the second term goes to $P(C_k)$. ■

4.4. Q_∞ for a Mixture Distribution

In this section, we verify an asymptotic property of our stability criterion under the extra assumption that the probability P is a mixture of distributions. Suppose the density p of P can be written as

$$p(x) = \sum_{k=1}^{K_T} \alpha_k \psi_k(x|\mu_k,\sigma_k), \quad (28)$$

where

$$\mu_k = \int x\psi_k(x|\mu_k,\sigma_k)\mathrm{d}\mu(x),$$

and $\alpha_k \in (0,1)$ for $k = 1,\ldots,K_T$ with $\sum_k \alpha_k = 1$. Here, σ_k is a measure of concentration for ψ_k with the property that

$$\psi_k(\cdot|\mu_k,\sigma_k) \to \delta_{\mu_k} \quad \text{as} \quad \sigma_k \to 0,$$

where δ_{μ_k} is unit mass at μ_k. In the special case that the ψ_k's are normal it is reasonable to take $\sigma_k = \sqrt{\mathrm{Var}_k(X)}$ where the variance is taken in the distribution with density ψ_k. However, we do not require the ψ_k's to be normal.

As before we take $\mathcal{C} = (C_1,\ldots,C_{K_T})$ to be the correct clustering. For instance, if $d(\cdot,\cdot)$ is squared error and the ψ_k's are normal with K_T modes then $\mathcal{C}$ could be the usual partition of the X-space into K_T classes based on, say, Fisher's Linear discriminant analysis. That is, $(C_1,\ldots,C_{K_T})$ would roughly correspond to regions around the K_T modes in Eq. (28) where each of the ψ_k was highest. This is similar to the optimality of K-means clustering as established in ref. [16].

Note that partitioning the range of X is equivalent to rewriting Eq. (28) in disjoint form. That is, it is equivalent to write

$$p(x) = \sum_{k=1}^{K_T} \tilde{\alpha}_k \tilde{\psi}_k(x|\tilde{\mu}_k,\tilde{\sigma}_k) \quad (29)$$

where

$$\tilde{\psi}_k(x|\tilde{\mu}_k,\tilde{\sigma}_k) = \frac{p(x|\mu_k,\sigma_k)\mathbb{I}_{C_k}}{P(C_k)},$$

are the disjoint, normalized forms of the ψ_k's and retain

$$\tilde{\mu}_k = \int x\tilde{\psi}_k(x|\tilde{\sigma}_k,\tilde{\sigma}_k)\mathrm{d}\mu(x)$$

and

$$\tilde{\psi}_k(\cdot|\tilde{\mu}_k,\tilde{\sigma}_k) \to \delta_{\tilde{\mu}_k} \quad \text{as} \quad \tilde{\sigma}_k \to 0.$$

The distinction between Eq. (28) and the disjoint form Eq. (29) is important because we evaluate the performance of a clustering using Eq. (28) but in fact when we have a candidate clustering, we only have a partition of the data that gives a partition of the X-space and the data on the partition elements only correspond to a coarse estimate of the terms in Eq. (29).

For $\mathcal{C}$, write

$$Q_\infty(K_T,\sigma) = \sum_{k=1}^{K_T} E_\sigma \mathbb{I}_{C_k}\int \mathbb{I}_{\{\forall \ell\neq k:\lambda_k d(X,\mu_k)\le\lambda_\ell d(X,\mu_\ell)\}}\mathrm{d}F(\lambda^{K_T}) \quad (30)$$

where $\sigma = (\sigma_1,\ldots,\sigma_{K_T})$ and the component means $\mu_1,\ldots,\mu_{K_T}$ are understood in the notation E_σ even though not explicitly indicated. Analogously, for another clustering $\mathcal{C}' = (C'_1,\ldots,C'_{K'})$, write

$$Q'_\infty(K',\sigma') = \sum_{k=1}^{K'} E_{\sigma'} \mathbb{I}_{C'_k}\int \mathbb{I}_{\{\forall \ell\neq k:\lambda_k d(X,\mu'_k)\le\lambda_\ell d(X,\mu'_\ell)\}}\mathrm{d}F(\lambda^{K'}) \quad (31)$$

where the centroids are $\mu'_k = E_{\sigma'} X\mathbb{I}_{C'_k}$ for $k = 1,\ldots,K'$ and $\sigma' = (\sigma'_1,\ldots,\sigma'_{K_T})$ (the same parametrization, but possibly different values).

One may argue that it is more reasonable to compare $Q'_\infty(K',\sigma')$ to a version of $Q_\infty(K_T,\sigma)$ taking expectations using 29 rather than Eq. (28). However, the differences are small and do not affect the main results of this section comparing the optimal clustering $\mathcal{C}$ and an alternative clustering $\mathcal{C}'$. The reason is that a result only seems to be feasible when the component variances go to zero. When

this holds, the support of the mixture components outside C_k's go to zero as well because the components concentrate on their centroids.

THEOREM 3: Fix $\mathcal{C}'$ with $P(C'_k) > 0$ for all $k = 1, \ldots, K'$. We have the following three properties of Q_∞: (i) If $K' \geq K_T + 1$ and each C_k has at most one of the μ_k's in it, then

$$\lim_{\sigma \to 0} Q_\infty(K_T, \sigma) \geq \lim_{\sigma' \to 0} Q'_\infty(K', \sigma').$$

(ii) If $K' \leq K_T - 1$, then

$$\lim_{\sigma \to 0} Q_\infty(K_T, \sigma) > \lim_{\sigma' \to 0} Q'_\infty(K', \sigma').$$

(iii) In general,

$$\lim_{\sigma \to 0} Q_\infty(K_T, \sigma) \geq \lim_{\sigma' \to 0} Q'_\infty(K', \sigma').$$

Remark: The theorem is weak because it uses limits as $\sigma, \sigma' \to 0$. However, it is not clear which stronger statements are true. For instance, we have been unable to show

$$\exists \sigma_0 \ \forall \sigma_k, \sigma'_k < \sigma_0 \quad Q_\infty(K_T, \sigma) \geq Q'_\infty(K', \sigma'),$$

that is, the main point of the theorem is true pre-asymptotically (in σ, σ'), although we conjecture it is true. The limitation in proving this seems to be dealing with the possible differences between the μ_k's for $\mathcal{C}$ and the μ'_k's for $\mathcal{C}'$. More precisely, we have not been able to identify reasonable hypotheses that preserve the inequality of the result; we comment on this in a remark after the proof. On the other hand, if no limit is taken over σ then the result may be false or meaningless. Indeed, Ray and Lindsay [17], see Section 1.2, observes that a mixture of two distinct normals can be unimodal. *A fortiori*, a mixture of K normals may have strictly fewer than K modes. In these cases it is not clear whether two components of the mixture represent two meaningful sub-populations or the unimodal mixture itself represents a single population with dispersion greater than a single normal component permits. Thus, whether $Q_\infty(K_T, \sigma)$ where K_T is the number of components in a mixture Eq. (28) ideally should be bigger or smaller than $Q'_\infty(K', \sigma')$ is a question of physical modeling not statistical evaluation. So, reducing a clustering to its modal structure by taking limits over σ may be effectively necessary for evaluation of stability in general.

Proof: We begin with (i). First consider a cluster $C'_{k'}$ in $\mathcal{C}'$ that does not contain any of the μ_k's from the C_k's. The component in Eq. (31) corresponding to $C'_{k'}$ is

$$E_{\sigma'} \mathbb{I}_{C'_{k'}} \int \mathbb{I}_{\{\forall \ell \neq k : \lambda'_{k'} d(X, \mu'_{k'}) \leq \lambda'_\ell d(X, \mu'_\ell)\}} \mathrm{d}F(\lambda^{K'}) \leq P_{\sigma'}(C'_{k'}) \to 0, \tag{32}$$

as $\sigma' \to 0$. This means that the terms in Eq. (31) corresponding to clusters with none of the μ_k's in them contribute zero asymptotically as $\sigma' \to 0$.

Now, suppose $K' = K_T + K$ where $K \geq 1$ and that, by relabeling if necessary, $C'_1, \ldots, C'_{K_T}$ have exactly one μ_k in each of them and $C'_{K_T+1}, \ldots, C'_{K_T+K}$ do not have any of the μ_k's in them and for $k = 1, \ldots, K_T$ $\mu_k \in C'_k$. We have that

$$Q'_\infty(K', \sigma') = \sum_{k=1}^{K_T} E_{\sigma'} \mathbb{I}_{C'_k} \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(X, \mu'_k) \leq \lambda_\ell d(X, \mu'_\ell)\}} \mathrm{d}F(\lambda^{K'}) + \sum_{k=K_T+1}^{K_T+K} E_{\sigma'} \mathbb{I}_{C'_k} \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(X, \mu'_k) \leq \lambda_\ell d(X, \mu'_\ell)\}} \mathrm{d}F(\lambda^{K'}). \tag{33}$$

By Eq. (32), the terms in the second sum are $o(1)$ as $\sigma' \to 0$. For the kth term in the first sum in Eq. (33) we have that

$$\int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(X, \mu'_k) \leq \lambda_\ell d(X, \mu'_\ell)\}} \mathrm{d}F(\lambda^{K'}) \to \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(\mu_k, \mu'_k) \leq \lambda_\ell d(\mu_k, \mu'_\ell)\}} \mathrm{d}F(\lambda^{K'}), \leq 1 \tag{34}$$

and that $P(C'_k) \to \alpha_k$ as $\sigma' \to 0$. Using Eq. (34) in Eq. (33) gives

$$\lim_{\sigma' \to 0} Q'_\infty(K', \sigma') \leq \sum_{k=1}^{K_T} \alpha_k = \lim_{\sigma \to 0} Q_\infty(K_T, \sigma), \tag{35}$$

as claimed in (i).

Next, we show (ii). To begin, consider the special case that $K' = K_T - 1$ and that $\mathcal{C}'$ has exactly one element, say C'_1 (relabel if necessary) with exactly two μ_k's in it and the other $C'_2, \ldots, C'_{K'}$ have exactly one μ_k i each of them so that $\mu_{k+1} \in C'_k$ for $k = 2, \ldots, K'$. Then, as $\sigma' \to 0$,

$$\begin{aligned} Q'_\infty(K', \sigma') &= E_{\sigma'} \mathbb{I}_{C'_1 \cap C_1} \\ &\times \int \mathbb{I}_{\{\forall \ell \neq 1 : \lambda_1 d(X, \mu'_1) \leq \lambda_\ell d(X, \mu'_\ell)\}} \mathrm{d}F(\lambda^{K'}) + E_{\sigma'} \mathbb{I}_{C'_1 \cap C_2} \\ &\times \int \mathbb{I}_{\{\forall \ell \neq 1 : \lambda_1 d(X, \mu'_1) \leq \lambda_\ell d(X, \mu'_\ell)\}} \mathrm{d}F(\lambda^{K'}) + E_{\sigma'} \mathbb{I}_{C'_1 \setminus (C_1 \cup C_2)} \\ &\times \int \mathbb{I}_{\{\forall \ell \neq 1 : \lambda_1 d(X, \mu'_1) \leq \lambda_\ell d(X, \mu'_\ell)\}} \mathrm{d}F(\lambda^{K'}) \end{aligned}$$

$$
\begin{aligned}
&+ \sum_{k'=2}^{K'} E_{\sigma'} \mathbb{I}_{C'_{k'}} \int \mathbb{I}_{\{\forall \ell \neq k' : \lambda_{k'} d(X,\mu'_{k'}) \leq \lambda_\ell d(X,\mu'_\ell)\}} \mathrm{d}F(\lambda^{K'}) \\
&\longrightarrow \alpha_1 \int \mathbb{I}_{\{\forall \ell \neq 1 : \lambda_1 d(\mu_1,\mu'_1) \leq \lambda_\ell d(\mu_1,\mu'_\ell)\}} \mathrm{d}F(\lambda^{K'}) \\
&+ \alpha_2 \int \mathbb{I}_{\{\forall \ell \neq 1 : \lambda_1 d(\mu_2,\mu'_1) \leq \lambda_\ell d(\mu_2,\mu'_\ell)\}} \mathrm{d}F(\lambda^{K'}) \\
&+ o(1) \\
&+ \sum_{k'=2}^{K'} \alpha_{k'+1} \int \mathbb{I}_{\{\forall \ell \neq k' : \lambda_{k'} d(\mu_{k'+1},\mu'_{k'}) \leq \lambda_\ell d(\mu_{k'+1},\mu'_\ell)\}} \\
&\times \mathrm{d}F(\lambda^{K'}) \qquad (36) \\
&< \sum_{k=1}^{K_T} \alpha_k = \lim_{\sigma \to 0} Q_\infty(K_T, \sigma). \qquad (37)
\end{aligned}
$$

More generally, the same reasoning holds if $K' \leq K_T - 1$ and some C'_k's have no μ_k's in them and other C'_k's have 1, 2, or more μ_k's in them. Specifically, C'_k's with no μ_k's in them are $o(1)$ as $\sigma' \to 0$; C'_k's with exactly one μ_k in them contribute at most α_k (if $\mu'_k = \mu_k$) and C'_k's with two or more μ_k's in them contribute terms of the form of the first two terms in the limit above, that is, are strictly bounded above by a sum of the corresponding α_k's.

For part (iii), begin by writing $K' = K_1 + K_2 + K_3$ and partition the clusters in $\mathcal{C}'$ into sets of size K_1, K_2 and K_3 where the first K_1 members C'_k's with means μ'_k do not contain any of the μ_k's, the second K_2 members C''_k's with means μ''_k contain exactly one of the μ_k's and the third collection of K_3 members C'''_k's with means μ'''_k contain two or more of their μ_k's. This means we have that

$$
\begin{aligned}
Q'_\infty(K', \sigma') &= \sum_{k=1}^{K_1} E_{\sigma'} \mathbb{I}_{C'_k} \int \mathbb{I}_{\{\forall \ell \neq k' : \lambda_k d(X,\mu'_k) \leq \lambda_\ell d(X,\mu^*_\ell)\}} \\
&\times \mathrm{d}F(\lambda^{K'}) \\
&+ \sum_{k=1}^{K_2} E_{\sigma'} \mathbb{I}_{C''_k} \int \mathbb{I}_{\{\forall \ell \neq k' : \lambda_k d(X,\mu''_k) \leq \lambda_\ell d(X,\mu^*_\ell)\}} \mathrm{d}F(\lambda^{K'}) \\
&+ \sum_{k=1}^{K_3} E_{\sigma'} \mathbb{I}_{C'''_k} \int \mathbb{I}_{\{\forall \ell \neq k' : \lambda_k d(X,\mu'''_k) \leq \lambda_\ell d(X,\mu^*_\ell)\}} \mathrm{d}F(\lambda^{K'}), \qquad (38)
\end{aligned}
$$

in which μ^*_ℓ indicates that all the μ'_k's, μ''_k's, and μ'''_k's are included.

As in Eq. (32), if $K_1 \geq 1$, the first sum in Eq. (38) is $o(1)$ as $\sigma' \to 0$.

As in bounding Eq. (36) by Eq. (37), if $K_2 \geq 1$, the second sum in Eq. (38) converges to

$$
\sum_{k=1}^{K_2} \alpha''_k \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(\mu_k,\mu''_k) \leq \lambda_\ell d(\mu_k,\mu^*_\ell)\}} \mathrm{d}F(\lambda^{K'}) \leq \sum_{k=1}^{K_2} \alpha''_k \qquad (39)
$$

with equality if the cluster means of the clusters $\mathcal{C}'$ are the same as for $\mathcal{C}$, as $\sigma' \to 0$. Note that in Eq. (39) the notation α''_k is used to indicate the weight of the component of the mixture distribution appropriate to the C''_k and that we have used μ_k generically to mean the limit of X on C_k as $\sigma' \to 0$.

The third sum in Eq. (38) converges like the first two terms in Eq. (36), that is, to a number strictly less than

$$
\sum_{k=1}^{K_3} \alpha'''_k,
$$

as $\sigma' \to 0$, (provided $K_3 \geq 1$), where the α'''_k's are the components of the mixture distribution appropriate to the C'''_k's. Since there are $K_2 + K_3$ weights α_k and all of them appear in the limits of the second and third sums in Eq. (38) we have that

$$
\lim_{\sigma \to 0} Q'_\infty(K', \sigma') \leq \sum_{k=1}^{K_T} \alpha_k = \lim_{\sigma \to \infty} Q_\infty(K_T, \sigma),
$$

so that (iii) follows. ■

Remark: It is seen that Eq. (34) is strict unless all $\mu_k = \mu'_k$ for $k = 1, \ldots, K_T$. Thus, equality in Eq. (34) holds if $C_k = C_{k'}$ and for some carefully chosen $\mathcal{C}'$'s but in general is not typical for $\mathcal{C}'$. This means that Eq. (35) is typically strict in practice. Using this and part (ii) suggests that we will often get strict equality in part (iii), but we do not have a proof of this.

Indeed, observe that if $K_T = K'$ and the clusters in $\mathcal{C}$ correspond to the clusters in $\mathcal{C}'$, we can compare the limits of $Q'_\infty(K_T, \sigma')$ and $Q_\infty(K_T, \sigma)$ directly as $\sigma, \sigma' \to 0$. In fact,

$$
\begin{aligned}
Q_\infty(K_T, \sigma) &= \sum_{k=1}^{K_T} E_\sigma \mathbb{I}_{C_k} \\
&\times \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(X,\mu_k) \leq \lambda_\ell d(X,\mu_\ell)\}} \mathrm{d}F(\lambda^{K_T}) \to \sum_{k=1}^{K_T} \alpha_k = 1,
\end{aligned}
$$

as in Proposition 1. By contrast, if $\mu_k, \mu'_k \in C_k, C'_k$, then

$$
\begin{aligned}
Q'_\infty(K_T, \sigma') &= \sum_{k=1}^{K_T} E_\sigma \mathbb{I}_{C'_k} \\
&\times \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(X,\mu'_k) \leq \lambda_\ell d(X,\mu'_\ell)\}} \mathrm{d}F(\lambda^{K_T})
\end{aligned}
$$

$$\to \sum_{k=1}^{K_T} \alpha_k \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(\mu_k, \mu'_k) \leq \lambda_\ell d(\mu_k, \mu'_\ell)\}} \mathrm{d}F(\lambda^{K_T})$$

$$\leq \sum_{k=1}^{K_T} \alpha_k = 1 = \lim_{\sigma \to 0} Q(K_T, \sigma),$$

with equality if and only if all $\mu_k = \mu'_k$. If we were to reverse the roles of $\mathcal{C}$ and $\mathcal{C}'$, we would get the reverse inequality in the limit when $K' = K_T$, but there is an asymmetry in that $\mathcal{C}$ is assumed to be optimal while $\mathcal{C}'$ is not. Further, if $\mu_k = \mu'_k$ but $\mathcal{C} \neq \mathcal{C}'$, then we cannot reasonably distinguish between clusterings with the same number of clusters asymptotically as $\sigma, \sigma' \to 0$.

5. CHOOSING THE NUMBER OF CLUSTERS

Recall the definition of the pointwise stability PW_i in Eq. (2.1) and the definition of $Q_n(K)$ in Eq. (12). Taking an average over the data points gives the average pointwise stability, $\mathrm{APW} = (1/n) \sum_{i=1}^{n} \mathrm{PW}_i$. Now,

$$\begin{aligned}
\mathrm{APW} &\approx E(\mathrm{APW} | \forall i \ x_i \text{ is in its correct } C_k) \\
&= \frac{1}{n} \sum_{i=1}^{n} E(\mathrm{PW}_i | x_i \text{ is in its correct } C_k) \\
&\approx E(Q_n(K) | \forall i \ x_i \text{ is in its correct } C_k) \\
&\approx Q_n(K) \approx Q_\infty(K),
\end{aligned}$$

at least for large n. So, our empirical stability criterion Q_n is approximated by the APW. Given this, we shift attention from Q_n or Q_∞ to the APW and show how the APW can be used—after rescaling—to find an estimate for the correct number of clusters. Note that the asymptotic equivalence of APW and Q_n means that, by Theorem 2, our estimate of K will be consistent. Moreover, in Section 7 we will observe that our method is also usually more efficient that several other methods.

5.1. Scaling the Averaged Pointwise Stability

Again let h_i be the index of the cluster to which x_i belongs. Write the APW of a clustering $\hat{\mathcal{C}} = (\hat{C}_1, \ldots, \hat{C}_K)$ as

$$\mathrm{APW} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{PW}(h_i, d(x_i, \hat{C}_1), ..., d(x_i, \hat{C}_K)), \qquad (40)$$

in which

$$\mathrm{PW}(h_i, d(x_i, \hat{C}_1), ..., d(x_i, \hat{C}_K)) = F_\theta(\{\lambda^K \ : \ \lambda_{h_i} d(x_i, \hat{C}_{h_i}) \leq \lambda_\ell d(x_i, \hat{C}_\ell) \, \forall \ell \neq h_i\}), \qquad (41)$$

where θ is a hyperparameter in the prior F.

Although the APW is consistent for the correct number of clusters, in practice we propose a statistic formed by scaling the APW by a baseline APW value that we denote $\mathrm{APW}_{\mathrm{base}}$. Assessing the APW relative to a baseline value indicates how stable a clustering is relative to an arbitrary assignment of points to clusters. In effect, we compare the APW to what its null hypothesis value would be if it were used as a statistic in a non-parametric permutation test. This allows us to associate uncertainty information to the APW and calibrate for spurious structure in the data. For instance, if there is no clustering at all, we expect $\mathrm{APW} \simeq \mathrm{APW}_{\mathrm{base}}$. For this reason, use $\log(\mathrm{APW}/\mathrm{APW}_{\mathrm{base}})$ as our statistic; if it is not positive, then the clustering is unstable.

To find a suitable baseline value for APW, we first used a bootstrap procedure on the $n \times K$ individual distance measures instead. For $b = 1, 2, ..., N_K$, we created a new $n \times K$ point-to-cluster distance matrix with entries drawn at random from the original distances. We then computed the average pointwise stability on this sampled baseline distance matrix, giving APW_b.

Then, for each value of b we find

$$S_{K,b} = \log \frac{\mathrm{APW}}{\mathrm{APW}_b} \qquad (42)$$

giving N_K stability scores for the clustering $\hat{C}_1, \ldots, \hat{C}_K$.

Effectively, $S_{K,b}$ actually represents an outcome of a random variable S_K that is a function of the data points and it is as if we have N_K independent outcomes of it. In principle, we have this for each K in a range, so we can describe a technique for choosing K in terms of the distributions of the S_K's implied by the values $S_{K,b}$ for $b = 1, \ldots, N_K$. Although we have permitted N_K to depend on K, in practice we have set $N_K = 100$ in all our calculations below and verified that this was large enough to identify the K's of stable clusterings unambiguously. Had $N_K = 100$ not been large enough, we would just choose N_K larger until we either got unambiguous results or accepted that unambiguous results are unattainable.

5.2. Maximizing $E(S_K)$ to Find the Number of Clusters

For each K, we have implicitly defined a random variable S_K and by bootstrapping we have a sample of size N_K from it. Treating this sample as IID, we find the smallest K achieving $\max_K E(S_K)$. The idea is that we want as few clusters as possible (to avoid dividing large clusters) but we want the largest value of $E(S_K)$ because it represents the K for which the APW is largest relative to its baseline, that is, the cluster of size K that is most stable.

One procedure that approximately does this in the cases we have examined is the following.

1. Choose K^* that maximizes $E(S_K)$ or, more exactly, its empirical estimate $(1/N_K)\sum_{b=1}^{N_K} S_{K,b}$.

2. Choose K^{**} as the smallest element in the set $\{2, 3, ..., K^*\}$ such that a one sided t-test does not show a statistical significant difference between $E(S_{K^{**}})$ and $E(S_{K^*})$ using the N_{K^*} data points from S_{K^*} (formed by the N_{K^*} bootstrap samples) and the $N_{K^{**}}$ data points from $S_{K^{**}}$. In other words, for each K satisfying $2 \le K \le K^* - 1$, do the one sided t-test H_0: $ES_{K^*} \le ES_{k^{**}}$ vs. H_1: $ES_{K^*} > ES_{K^{**}}$ and choose the lowest K^{**} for which the null is not rejected at, say, the 0.05 level. If the null is rejected for $K = 2, ..., K^* - 1$, set $K^{**} = K^*$.

3. Set $\hat{K} = K^{**}$ if the 2.5% quantile of $S_{K^{**}}$ is positive; otherwise, set $\hat{K} = 1$.

This procedure produces an estimate $\hat{K}$ of the true value of K, assuming (i) prior selection has already been done, for instance, as in Section 6 and (ii) that for each K we already have identified an optimal clustering. This is possible for clustering techniques such as K-means and most hierarchical methods.

An illustration of this procedure is provided in Fig. 5. We call Panel (a) a stability curve because it shows the relative stability of a series of clusterings of size K via a series of boxplots, one for each candidate value of K, using the values $S_{K,b}$ for $b = 1, \ldots, N_K = 100$. The midline of each box is the median for the corresponding value of K. Clearly, choosing all N_K's to be 100 is enough to see that $K = 4$ has the highest value of the estimate of $E(S_K)$ and that the estimate of $E(S_3)$ is below the estimate of $E(S_4)$ so $\hat{K} = 4$, the correct value.

Panel (b) shows the data clustered into four clusters and (c) the stability heatmap for the clustering. Panels (d) and (f) are similar, but for neighboring values $K = 3, 5$. The three clusters in panel (d) do not look like a good summary for the data and the heatmap (e) indicates that the cluster represented by the lower left block is not very stable. This corresponds to cluster three in the plot of the data. By contrast, it is seen in Panel (g) that five clusters is relatively stable at the cost of splitting the points on the right into three clusters (5, 1, and 4) when (b) shows two is enough. However, panel (g) shows that fewer of the clusters in (f) are stable compared with (c).

6. PRIOR SELECTION

The general theory presented here holds for any prior density $w(\lambda|\theta)$ equipped with a hyperparameter θ that can be tuned. The interpretation of the key stability quantities assumes $w(\lambda|\theta)$ is continuous in θ and that as a function of θ the density is smoothly deformable. So, the range of choices for $w(\lambda|\theta)$ that can give good results in principle is relatively unconstrained. Nevertheless, for computing, we must fix a family of priors. In our examples to follow, as in the computations shown in Section 2, we have used the exponential family shifted by one. So, in this section we explain our method for hyperparameter selection in general and then we explain how it applies to the shifted exponential.

6.1. The General Method

Assuming that the prior F to be used on the λ_k's $(k = 1, \ldots, K)$ is IID and depends smoothly on a finite dimensional real parameter θ, it is enough to identify a function of θ we can maximize. The natural choice is the cumulative stability $\sum_{K\in[K_1,K_2]} E(S_K)$. That is, we find θ by maximizing the difference between the log stability of the clustering and the log baseline stability for the K-cluster clustering, averaging over K and the sample space. As this is a population quantity we use the natural empirical analog. That is, we write

$$\begin{aligned}
\frac{1}{K_2 - K_1}\sum_{K=K_1}^{K_2} E(S_K) &\approx \frac{1}{K_2 - K_1} \\
&\quad \times \sum_{K=K_1}^{K_2}\left(\frac{1}{N_K}\sum_{b=1}^{N_K}\log\frac{\mathrm{APW}}{\mathrm{APW}_b}\right) \\
&= \frac{1}{K_2 - K_1}\sum_{K=K_1}^{K_2}\left(\frac{1}{N_K}\sum_{b=1}^{N_K}\log\frac{(1/n)\sum_{i=1}^{n}\mathrm{PW}(x_i)}{(1/n)\sum_{i=1}^{n}\mathrm{PW}_i(b)}\right) \\
&= \frac{1}{K_2 - K_1}\sum_{K=K_1}^{K_2}\log\frac{1}{n}\sum_{i=1}^{n}\mathrm{PW}(x_i) \\
&\quad -\frac{1}{K_2 - K_1}\sum_{K=K_1}^{K_2}\left(\frac{1}{N_K}\sum_{b=1}^{N_K}\log\frac{1}{n}\sum_{i=1}^{n}\mathrm{PW}_i(b)\right), \quad (43)
\end{aligned}$$

where $\mathrm{PW}(x_i)$ is defined in Eq. (5) and $\mathrm{PW}_i(b)$ is a summand in Eq. (41). Both $\mathrm{PW}(x_i)$ and $\mathrm{PW}_i(b)$ depend on F and hence on θ so the right hand side of Eq. (43) can be maximized over θ to approximate the maximum of the left hand side over θ. When θ is one-dimensional, as here, this can be done easily by a simple Brent bisection algorithm; as noted in Section 5.1, we took $N_K = 100$. All examples in this paper use this method.

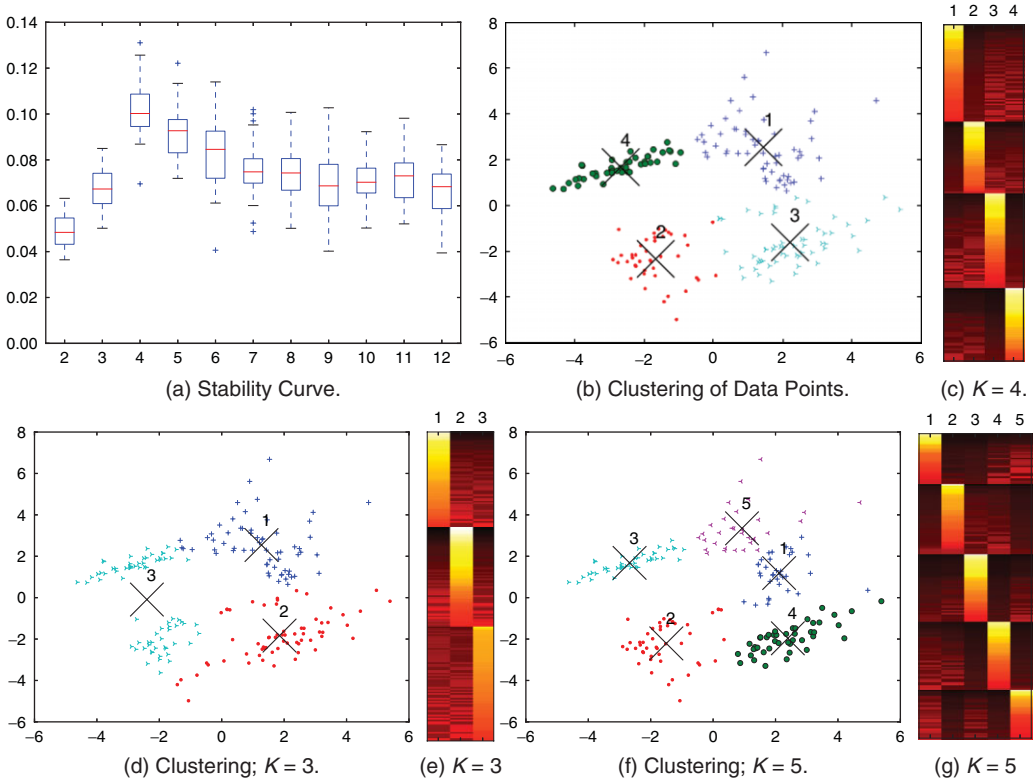


Fig. 5 Two dimension toy example with 200 data points generated from a mixture model with four components. Panel (a) shows, for each K, the boxplot for the bootstrap sample for S_K and it is easy to see $\hat{K} = 4$. Panel (b) shows the correct clustering and (c) its stability heatmap. Panels (d) and (f) are the same as (b) but the clusterings are optimal for incorrect K's, with heatmaps as in (e) and (g). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

6.2. Shifted Exponential Prior

The most useful prior we have found thus far for scale perturbations on the relationships among the $d(x_i, \hat{C}_k)$'s is the exponential distribution with the location shifted by one. That is, we set

$$f(\lambda|\theta) = \theta e^{-\theta(\lambda-1)} \mathbb{I}_{\lambda \geq 1}, \tag{44}$$

as the density of F_θ. We defend this choice through a series of arguments; for ease of exposition, when possible, we henceforth drop the index i in the ϕ_{ik}'s since the data are IID.

First, Eq. (44) is among the simplest priors we can choose that has a satisfactory interpretation. There are two parameters, θ and the shift (chosen to be one) that we think make sense to include. First, the interpretation of θ is straightforward: If $\theta \to \infty$ then the priors assign unit mass at the shift value, reducing the ϕ_k's to a function of the point-to-cluster distances namely $\phi_{ik} = a_{ik} = \mathbb{I}_{\forall \ell \neq k:\ d_{ik} \leq d_{i\ell}}$ (see 4), a sensible limit. If $\theta \to 0$ then ratios of the form λ_j/λ_k have a distribution more and more like a ratio of independent uniforms on $[1, \infty)$, representing significant perturbation that tends to overwhelm the original distance measures. In between zero and ∞ we have values of θ that are sensitive to the point-to-cluster distances and adapt to the relative location of the data points and cluster centers, as well as to the dimension of the data. This means that θ has a reasonable interpretation.

If we retain the interpretation of θ but use a simpler prior such as the exponential distribution without the shift, that is, $\theta e^{-\theta\lambda}$ for $\lambda \in (0, \infty)$ we can derive

$$\phi_{ik} = \frac{d^{-1}(x_i, C_k)}{\sum_{\ell=1}^{K} d^{-1}(x_i, C_\ell)}, \tag{45}$$

see ref. 18. This is superficially reasonable: If d_1 is small then the corresponding entry in ϕ_k will be close to one, its maximal value. On the other hand, if all d_k's are similar in size then the corresponding entry in ϕ_k will be near $1/K$, its minimal value. However, a shift of zero is unreasonable because the ϕ_k's do not depend on θ so the stability measure cannot adapt to the clustering. In such cases, two clusterings with the same point-to-cluster distances will be seen as equally stable even when (for instance) the points are arranged very differently or the dimension increases.

If we retain the interpretation of θ but use a more complicated prior, for example, a shifted Gamma prior, the results from stability assessments were no better than what we found with the shifted exponential—but were much harder to compute. Normal priors performed quite badly, see ref. 18. Indeed, the main point of Proposition 2 below is to generalize from Eq. 45 to the shifted exponential (for computational purposes). This is not easy; generalizing further would be even more difficult. On the other hand, a sufficiently general search over priors might result in a better choice than Eq. (44).

We set the shift to one for convenience. As noted in our interpretation of θ, the role of the shift is in the ratios λ_j/λ_k. So, any nonzero shift would have the same qualitative effect. The main change would be in the numerical value of θ found to be optimal in Eq. (43). Overall, the shifted exponential prior is the simplest, interpretable, computationally efficient prior that we found that gave good performance.

To proceed with the analysis using a shifted exponential prior, denote the corresponding averaged assignment matrix with a superscript LE:

$$\begin{aligned}\mathbf{\Phi}^{\mathrm{LE}} &= [\phi_{ik}^{\mathrm{LE}}]_{i=1,\ldots,n;k=1,\ldots,K} \\ &= \left(\int \mathbb{I}_{\forall \ell \neq k:\ \lambda_k d_{ik} \leq \lambda_\ell d_{i\ell}} \prod_{\ell=1}^{K} f(\lambda_\ell|\theta) \mathrm{d}\lambda^K \right)_{i=1,\ldots,n;k=1,\ldots,K}.\end{aligned}$$

To implement a procedure for finding $\mathbf{\Phi}^{\mathrm{LE}}$ computationally, we first find a relatively convenient expression for the ϕ_k^{LE}'s in terms of the $d_k = d(\cdot, C_k)$'s.

PROPOSITION 2: Let $\mathbf{d} = (d_1, \ldots, d_K)$ be a list of K distances and let $\psi : \{1, \ldots, K\} \to \{1, \ldots, K\}$ be the bijection that puts $\mathbf{d}$ in sorted order, that is, $d_{\psi(1)} \leq \ldots \leq d_{\psi(K)}$. Define

$$\begin{aligned}B_k &= \theta \sum_{j=1}^{k} \frac{1}{d_{\psi(j)}} \\ C_k &= \begin{cases} 1 & k = 1 \\ \exp\left[-\theta \sum_{j=1}^{k-1} \left(\frac{d_{\psi(k)}}{d_{\psi(j)}} - 1 \right) \right] & k \in \{2, \ldots, K\} \end{cases} \\ D_k &= \sum_{j=k+1}^{K} C_j \left[B_{j-1} \left(\frac{B_{j-1} d_{\psi(j)}}{\theta} + 1 \right) \right]^{-1}.\end{aligned}$$

Then,

$$\phi_k^{\mathrm{LE}} = \frac{\theta}{d_k} \left(\frac{C_{\psi^{-1}(k)}}{B_{\psi^{-1}(k)}} - D_{\psi^{-1}(k)} \right).$$

Remark: Proposition 2 can be generalized to non-IID λ^k's, see ref. 18, in which case λ^K has density $f(\lambda^K|\theta_1, \ldots \theta_K) = \prod_k f(\lambda_k|\theta_k)$ so K values of θ must be estimated, one for each cluster.

Proof: The technique of proof is to use Fubini's Theorem on ϕ_k, integrate over some of the λ's, and then break up the domain of integration to recognize the B's, C's and D's. The details are given in Appendix B. ■

Proposition 2 justifies Algorithm 1 below which essentially pre-computes the common terms. The running time of this algorithm is linear in K apart from the call to order the entries of d, which runs in $\mathcal{O}(K \log K)$ time. Thus, the overall running time is $\mathcal{O}(K \log K)$ for each data point.

Algorithm 1: Procedure for the calculation of the ϕ^{LE}'s:

Input: A vector $\mathbf{d}$ of K distances and a real parameter θ.
Output: A probability vector ϕ^{LE} of partial memberships.
$K \leftarrow \mathrm{length}(\mathbf{d})$, $\psi \leftarrow \mathrm{argsort}(\mathbf{d})$, Initialize d^s, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, $\mathbf{r}$ as vectors of length K;
for $k = 1$ **to** K **do** $d_j^s \leftarrow d_{\psi(k)}$, $r_k \leftarrow \theta^2/d_k^s$, $B_1 \leftarrow r_1$, $C_1 \leftarrow 1$;
for $j = 2$ **to** K **do** $B_k \leftarrow B_{k-1} + r_k$;
$C_k \leftarrow \exp((j-1)\theta - B_{k-1} d_k^s)$;
$D_K \leftarrow 0$;
for $k = K-1$ **to** 1 **step** -1 **do** $D_k \leftarrow D_{k+1} + C_{k+1}/(B_k(B_k r_{k+1} + 1))$;
for $k = 1$ **to** K **do** $\phi_{\psi(k)}^{LE} \leftarrow r_k(C_k/B - k - D_k)$;
return ϕ^{LE}

We use this procedure in Section 7, without further comment, to choose appropriate values for θ in a variety of examples involving simulated and real data.

7. COMPUTED EXAMPLES

In this section we demonstrate the effectiveness of our approach in comparison with several other popular clustering stability methods, namely the gap statistic, subsampling, and the silhouette distance. While numerous other methods exist, we argue these are a good representation of the variety of approaches to cluster stability since their motivations are disjoint from each other and from our proposed method. The gap statistic is the standard version from ref. 9, with the bounding box for the uniform null distribution chosen using the principle components of the sample. Likewise, the silhouette score is the standard one from ref. 8. The subsampling method we chose as representative of the numerous variants in the literature, was to divide the data randomly into three equal groups, then for given K, find a clustering into K clusters for each of the three distinct pairs formed by excluding one of the groups. Then, using the variation of information metric, we scored the clustering on each pair with the original clustering an the full data set, discarding points from the full data set that were not in the pair. The overall stability was then the average of the three pairwise

scores. This procedure was repeated 100 times on randomly chosen partitions of the data into the three groups, with the final score being the average score over 100 runs. The estimate of K is the $\hat{K}$ with the lowest final score. Code for the instantiation of all these methods, and the new methods in this paper, is available on request from the first author, pending the release of an R package.

7.1. Synthetic Data

To compare our method for choosing K with these three standard methods, we first generate synthetic data for clustering purposes using the method described in detail in ref. 18. This method works by first generating points from a mixture of normal components where the components are separated so its clear what the clusters are. Then, a series of transformation is applied to the data points to situate them in a random non-linear, non-orthogonal coordinate system. Next, the points in the K clusters are reallocated so that each component in the final clustering is separated by a drop in the mixture model density $p(x)$ down to or below $\beta \min(p(m_{k_1}), p(m_{k_2}))$, where the minimum is taken over all components k_1 and k_2 with m_{k_1} and m_{k_2} denoting the modes of these components. The parameter $\beta \leq 1$ is set by the user to control how much the density between each pair of clusters must drop, effectively controlling the separation between the clusters. The other user set parameter, the 'severity', controls how far from linearity a coordinate transformation will be on average. The resulting clusters are unimodal but shaped far differently, for example, rarely convex, often looking like a drop of paint that had been smeared haphazardly. This makes the components of the new data set difficult to model by standard techniques, challenging both the clustering and validation stages—although, because we generated the data, we know what the correct clusters are.

For our experiments, we used three classes of normal mixtures in in 2, 5, 10, and 20 dimensions, and with 100, 150, 200, and 300 data points each, respectively. These classes were as follows.

$T1$: This class has four mixture components and uses significant nonlinear transformations to give the components more difficult shapes.

$T2$: This class has five mixture components, but fewer sheering and scaling operations so the clusters are better modeled by a centroid-based method such as k-means.

$T3$: This class is formed by taking samples from a single unimodal normal density. This tests the ability of the methods to detect when there is no actual clustering, that is, $K = 1$. To make this a bit more difficult, the standard deviation of each dimension increases linearly from 1 to 2 over the dimensions. Thus, in the 2d example, $X \sim N(0, \Sigma)$, where $\Sigma = ((1, 0), (0, 2))$. In the 20d case, the standard deviation for dimension 10 is 1.5, and the standard deviation for dimension 20 is 2.

For classes $T1$ and $T2$, we fixed the severity of the nonlinear transformations and then adjusted the β parameter controlling the component separation until the results were able to distinguish between the methods. For the first class, the resulting β's were 0.55, 0.5, 0.45, and 0.35; for the second class, β was 0.7, 0.6, 0.55, and 0.45. It is seen that these values are roughly in the middle of [0, 1] indicating that the separation is not so low as to make clustering easy nor so high as to make clustering impossible.

To test the clustering evaluation procedures on a given sample, we used K-means to find a candidate clusterings with K centroids. To reduce problems from a bad clustering, K-means was seeded from perturbed values of the known modes, then run $10p$ times, where p is the dimension. Of these runs, the one with the lowest cost function was chosen as the candidate clustering. The clustering procedure was the same for baseline and subsampled distributions.

Tables 1, 2, and 3 show the results of our simulations for the three classes of data. We labeled the results of our method 'perturbations' because the underlying motivation is the perturbation of distances by factors λ_k. Note that in addition to the four stability methods already discussed, we included a variation on our method called 'perturbations with average linkage'. Average linkage is not a metric and so does not fall under the hypotheses of our theoretical results, however, average linkage is arguably the most popular choice of linkages for hierarchical clustering. In this case, we set

$$d(x_i, \hat{C}_k) = \sqrt{\frac{1}{|\hat{C}_k|} \sum_{j \in \hat{C}_k} \|x_i - x_j\|_2^2}, \tag{46}$$

where $\|\cdot\|$ is the Euclidean distance.

Table 1 is for data from class $T1$. The column headed 'Data set' indicates the key features of the data (class, dimension, correct number of clusters, and sample size). For each of the five methods we did 100 runs, that is, formed 100 estimates $\hat{K}$ using each method to estimate $K = K_{\text{true}} = 4$. The last 12 columns show the sampling distribution of $\hat{K}$ for the methods. It is seen that in all but one setting the perturbation method using the Euclidean distance is best in the sense that its sampling distribution is most peaked around the correct number of clusters. The

Table 1. Sampling distributions for $\hat{K}$ for $T1$ data examples. The two perturbation methods are seen to be most concentrated around the true value $K = 4$ in all cases.

Data set	Method $\hat{K} \rightarrow$	1	2	3	4	5	6	7	8	9	10	11	12
$T1$, $2d$, $K = 4$, $n = 100$	Gap	16	10	3	57	12	2	0	0	0	0	0	0
$T1$, $2d$, $K = 4$, $n = 100$	Subsampling	0	15	9	69	7	0	0	0	0	0	0	0
$T1$, $2d$, $K = 4$, $n = 100$	Silhouette	0	0	0	68	18	8	2	2	1	1	0	0
$T1$, $2d$, $K = 4$, $n = 100$	Pertubations	0	0	2	85	9	3	1	0	0	0	0	0
$T1$, $2d$, $K = 4$, $n = 100$	Pert. Avg. Linkage	0	0	3	87	8	2	0	0	0	0	0	0
$T1$, $5d$, $K = 4$, $n = 150$	Gap	6	4	0	37	31	11	10	1	0	0	0	0
$T1$, $5d$, $K = 4$, $n = 150$	Subsampling	0	18	6	62	13	1	0	0	0	0	0	0
$T1$, $5d$, $K = 4$, $n = 150$	Silhouette	0	2	2	67	19	4	5	1	0	0	0	0
$T1$, $5d$, $K = 4$, $n = 150$	Perturbations	0	2	6	78	11	2	1	0	0	0	0	0
$T1$, $5d$, $K = 4$, $n = 150$	Pert. Avg. Linkage	0	6	12	70	12	0	0	0	0	0	0	0
$T1$, $10d$, $K = 4$, $n = 200$	Gap	0	2	0	36	27	21	6	5	2	1	0	0
$T1$, $10d$, $K = 4$, $n = 200$	Subsampling	0	12	3	73	11	1	0	0	0	0	0	0
$T1$, $10d$, $K = 4$, $n = 200$	Silhouette	0	3	0	60	22	10	1	3	1	0	0	0
$T1$, $10d$, $K = 4$, $n = 200$	Perturbations	0	7	1	87	2	3	0	0	0	0	0	0
$T1$, $10d$, $K = 4$, $n = 200$	Pert. Avg. Linkage	1	12	12	66	7	2	0	0	0	0	0	0
$T1$, $20d$, $K = 4$, $n = 400$	Gap	0	6	0	32	35	15	8	2	2	0	0	0
$T1$, $20d$, $K = 4$, $n = 400$	Subsampling	0	2	0	87	9	2	0	0	0	0	0	0
$T1$, $20d$, $K = 4$, $n = 400$	Silhouette	0	1	0	82	10	7	0	0	0	0	0	0
$T1$, $20d$, $K = 4$, $n = 400$	Perturbations	0	7	1	89	2	1	0	0	0	0	0	0
$T1$, $20d$, $K = 4$, $n = 400$	Pert. Avg. Linkage	0	13	9	71	4	2	0	0	0	0	0	1

one exception is in two dimensions where the perturbation method comes second to the perturbation method with average linkage. We attribute this to the fact that clusters from the $T1$ class of data are not as well-summarized by centroids as they are by average linkage, especially in low dimensions. (As dimension increases, there is more and more 'space' between points so centroid methods can become more representative.) Nevertheless, in all cases, one of the two perturbation methods is always best. The gap statistic tends to be appropriately located but has a much higher spread tending to be skewed upward for higher dimensions. The subsampling method is also appropriately located but tends to be skewed to one side or the other, at least for the sample sizes we used. The silhouette method was similar, but appeared to be skewed upward too often.

Table 2 is for data from class $T2$. These data sets were generally a bit easier to cluster and are more amenable to centroid methods than $T1$ data. It is seen that the perturbation method is again best in all cases but 20 dimensions, where subsampling does best, though in some other settings the silhouette method is a close second. This leads us to suggest that in higher dimensions the behavior of stability methods for clustering may be qualitatively different their behavior in lower dimensions because in the limit of high dimensions all points become equidistant from each other, see ref. 19 which shows

$$\lim_{d\to\infty} \frac{\text{dist}_{\max} - \text{dist}_{\min}}{\text{dist}_{\min}} \to 0$$

where $\text{dist}_{\max}$, $\text{dist}_{\min}$ are the maximal and minimal distances between data points in d dimensions. This means that distance measures become uninformative about the clustering, a problem not suffered by subsampling methods since they are not distance based. Nevertheless, all the methods are again appropriately located but the sampling distribution of the perturbation method is more concentrated around the true value of K than for the other methods, except for $d = 20$ where it is a close second.

Data from class $T3$ is used as a sanity check to verify that when there is only one cluster, the stability method will not suggest it is unstable implying two or more clusters would be more reasonable. Table 3 shows the results of this search for spurious clusters. It is seen that the perturbation methods are essentially the only ones that put appreciable, sometimes very high, weight on choosing a single cluster, except in two dimensions where the gap statistic does best. This may due to the fact that a random uniform distribution, used as the baseline distribution for the gap statistic, is, here, a better null clustering to compare against for low dimensions than for higher dimensions, that is, the uniform is a poor baseline for normal data in higher dimensions. However, this seems unimportant because all five methods do poorly in this 2d case.

Overall, this simulation study shows that our method gives results than are comparable or better than the other given methods, sometimes by wide margin. This improvement is not uniform—nor should we expect it to be. It is just overall better than other methods for a large class of problems with easy to moderately difficult data and moderate dimension size. In the very few cases where

Table 2. Sampling distributions for $\hat{K}$ for $T2$ data examples. The perturbation method is generally more highly concentrated around the true $K = 5$, except for 20 dimensions where it is a close second to subsampling.

Data set	Method $\hat{K} \rightarrow$	1	2	3	4	5	6	7	8	9	10	11	12
$T2$, $2d$, $K = 5$, $n = 100$	Gap	16	33	26	7	17	1	0	0	0	0	0	0
$T2$, $2d$, $K = 5$, $n = 100$	Subsampling	0	32	11	7	43	7	0	0	0	0	0	0
$T2$, $2d$, $K = 5$, $n = 100$	Silhouette	0	1	1	6	52	19	11	5	3	1	0	1
$T2$, $2d$, $K = 5$, $n = 100$	Perturbations	0	0	4	19	64	10	2	1	0	0	0	0
$T2$, $2d$, $K = 5$, $n = 100$	Pert. Avg. Linkage	0	4	10	25	53	6	2	0	0	0	0	0
$T2$, $5d$, $K = 5$, $n = 150$	Gap	1	4	2	1	54	31	6	1	0	0	0	0
$T2$, $5d$, $K = 5$, $n = 150$	Subsampling	0	15	2	5	60	17	1	0	0	0	0	0
$T2$, $5d$, $K = 5$, $n = 150$	Silhouette	0	0	0	2	63	26	9	0	0	0	0	0
$T2$, $5d$, $K = 5$, $n = 150$	Perturbations	0	1	1	8	84	6	0	0	0	0	0	0
$T2$, $5d$, $K = 5$, $n = 150$	Pert. Avg. Linkage	0	4	5	19	65	7	0	0	0	0	0	0
$T2$, $10d$, $K = 5$, $n = 200$	Gap	0	2	0	0	60	22	13	2	1	0	0	0
$T2$, $10d$, $K = 5$, $n = 200$	Subsampling	0	7	0	1	82	8	2	0	0	0	0	0
$T2$, $10d$, $K = 5$, $n = 200$	Silhouette	0	1	0	0	83	7	6	3	0	0	0	0
$T2$, $10d$, $K = 5$, $n = 200$	Perturbations	0	3	4	7	84	1	1	0	0	0	0	0
$T2$, $10d$, $K = 5$, $n = 200$	Pert. Avg. Linkage	0	4	7	10	72	3	4	0	0	0	0	0
$T2$, $20d$, $K = 5$, $n = 400$	Gap	0	0	0	0	43	35	14	7	1	0	0	0
$T2$, $20d$, $K = 5$, $n = 400$	Subsampling	0	2	0	0	95	2	1	0	0	0	0	0
$T2$, $20d$, $K = 5$, $n = 400$	Silhouette	0	0	0	0	90	9	1	0	0	0	0	0
$T2$, $20d$, $K = 5$, $n = 400$	Perturbations	0	2	1	3	92	2	0	0	0	0	0	0
$T2$, $20d$, $K = 5$, $n = 400$	Pert. Avg. Linkage	0	4	4	16	71	5	0	0	0	0	0	0

Table 3. Sampling distributions for $\hat{K}$ for $T3$ data examples. Outside of low dimensions where no method works well, the perturbation methods give the best results.

Data set	Method $\hat{K} \rightarrow$	1	2	3	4	5	6	7	8	9	10	11	12
$T3$, $2d$, $K = 1$, $n = 100$	Gap	14	82	4	0	0	0	0	0	0	0	0	0
$T3$, $2d$, $K = 1$, $n = 100$	Subsampling	0	95	5	0	0	0	0	0	0	0	0	0
$T3$, $2d$, $K = 1$, $n = 100$	Silhouette	0	80	1	1	0	1	1	3	3	1	0	9
$T3$, $2d$, $K = 1$, $n = 100$	Perturbations	0	10	28	20	18	10	7	1	5	1	0	0
$T3$, $2d$, $K = 1$, $n = 100$	Pert. Avg. Linkage	0	63	29	3	3	2	0	0	0	0	0	0
$T3$, $5d$, $K = 1$, $n = 150$	Gap	1	96	3	0	0	0	0	0	0	0	0	0
$T3$, $5d$, $K = 1$, $n = 150$	Subsampling	0	100	0	0	0	0	0	0	0	0	0	0
$T3$, $5d$, $K = 1$, $n = 150$	Silhouette	0	39	5	2	1	0	0	1	6	5	12	29
$T3$, $5d$, $K = 1$, $n = 150$	Perturbations	38	3	14	13	8	8	4	1	5	1	3	2
$T3$, $5d$, $K = 1$, $n = 150$	Pert. Avg. Linkage	80	6	3	3	1	3	0	1	0	1	2	0
$T3$, $10d$, $K = 1$, $n = 200$	Gap	0	100	0	0	0	0	0	0	0	0	0	0
$T3$, $10d$, $K = 1$, $n = 200$	Subsampling	0	100	0	0	0	0	0	0	0	0	0	0
$T3$, $10d$, $K = 1$, $n = 200$	Silhouette	0	2	0	0	0	0	0	0	1	10	29	58
$T3$, $10d$, $K = 1$, $n = 200$	Perturbations	97	3	0	0	0	0	0	0	0	0	0	0
$T3$, $10d$, $K = 1$, $n = 200$	Pert. Avg. Linkage	99	1	0	0	0	0	0	0	0	0	0	0
$T3$, $20d$, $K = 1$, $n = 400$	Gap	0	100	0	0	0	0	0	0	0	0	0	0
$T3$, $20d$, $K = 1$, $n = 400$	Subsampling	0	100	0	0	0	0	0	0	0	0	0	0
$T3$, $20d$, $K = 1$, $n = 400$	Silhouette	0	0	0	0	0	0	0	0	0	6	21	73
$T3$, $20d$, $K = 1$, $n = 400$	Perturbations	100	0	0	0	0	0	0	0	0	0	0	0
$T3$, $20d$, $K = 1$, $n = 400$	Pert. Avg. Linkage	100	0	0	0	0	0	0	0	0	0	0	0

perturbation methods are not best, they are a very close second.

7.2. Stability Analysis of the Wisconsin Breast Cancer Data Set

In this section we give a second—and easy—example of how our method can be applied to real data. We use the Wisconsin breast cancer data set first described in ref. 20 and available from ref. 21. Like the MNIST data set from Section 2.2, this is a classification data set in which each patient is classified as having malignant or benign breast cancer. The sample size is 569 and there are a total of ten explanatory variables representing measurements made on cell nuclei. The two classes are fairly well separated. Indeed, it is possible to find the two classes by clustering.

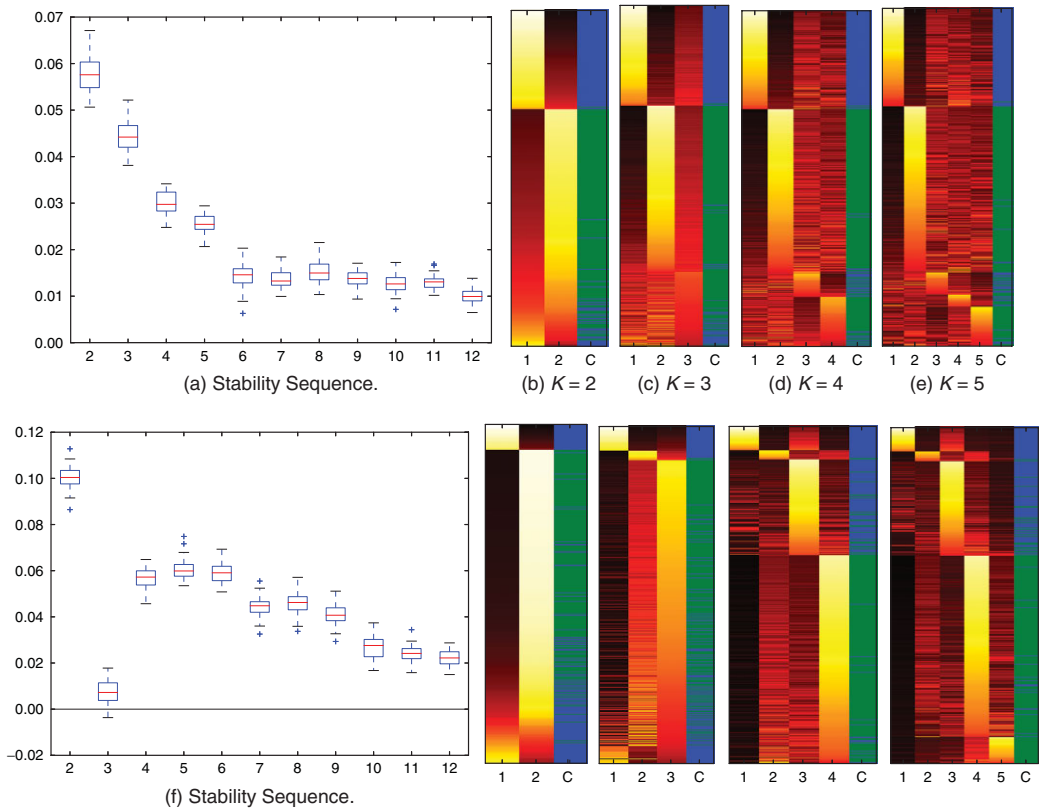


Fig. 6 Top row: Stability analysis of complete linkage clustering of the studentized breast cancer data set using the cosine distance. As can be seen in the $K = 2$ stability heatmap (b), this clustering has only a handful of points that are in the boundary region between the two clusters. The $(2, 1)$ block is quite light at the bottom indicating points that are falsely clustered. Using blue to indicate malignant instances and green to indicate the benign instances it is seen that there are blue lines in the lower green block corresponding to points that are falsely clustered. Furthermore, the heatmaps (c), (d), and (e) for clusterings with $K = 3$, $K = 4$, and $K = 5$ indicate that there are two significantly stable clusters, one in the malignant class and one in the benign class. The other clusters are all on the boundary regions. Bottom row: Stability analysis of complete linkage clustering of the raw breast cancer data set using the cosine distance. Again, the clustering for $K = 2$ does not separate the classes distinctly as seen by the light patch at the bottom of the $(2, 1)$ block, even though the clusters themselves, and the stability score, indicate overall stability. The stability heatmaps for $K = 3, 4, 5$ suggest there are up to two significantly stable clusters in the benign class and three significantly stable clusters in the malignant category. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

We first found two complete linkage hierarchical clusterings for the data, measuring the distance between points by the the cosine distance. The first used the data 'as is' while in the second we studentized all ten explanatory variables.

The stability analysis for the two complete linkage clusterings is shown in Fig. 6 and takes δ to be complete linkage. (We used complete linkage to form the clusters because it permits large, diffuse clusters, but switched to average linkage for the stability analysis because it favors more compact clusters, a desirable property for stability.) Panels (a) and (f) show that for both forms of the data $K = 2$ is the most stable. However, panel (a) shows that studentization gives a clear fall-off of stability with increasing K while the raw data shows a sharp drop from $K = 2$ to $K = 3$ and a sharp rise from $K = 3$ to $K = 4$ leading to a local maximum at $K = 5$, even though the global maximum is at $K = 2$.

The stability heatmaps for the clusterings with $K = 2$, 3, 4, and 5 are shown next to their respective stability sequences. On the right of each heatmap we have added an extra column coding malignant cases in blue and benign cases in green. A blue line in the lower, mostly green, region or a green line in the upper, mostly blue region indicates a data point that has been put in a cluster different from its class. This provides a visual assessment of how far wrong the clustering is and is another way to represent information in the off-diagonal blocks of the heatmap. It is seen that for $K = 2$ with studentized data, the clustering recovers the classes better than it does for the raw data. It is also seen that the misclustered data points account for the instability under perturbations (the blue lines in the green region), indicating they are close to the boundary between the two classes. The heatmaps for $K = 2$ also suggest that the clustering on the raw data is more stable than on the

studentized data for $K = 2$, even though it matches the class labels worse.

For $K \geq 3$, the heatmaps for the two clusterings are even more revealing. For $K = 3$, the clustering on the studentized data is poor: Only two of the clusters show any stability. However, with the raw data, the split in the lower, green, benign region looks much more stable. For $K = 4, 5$ the heatmaps for the studentized data continue to show great instability. However, the local maximum from the raw data suggests that a splitting the blue region into two clusters and the green region into two or three clusters continues to reveal good stability. Additionally, the upper left block and the $(2, 2)$ block in the heatmaps for $K = 4, 5$ shows there is a subclass of the malignant tumors that is well separated and does not share a boundary with the other clusters, as evidenced by the dark regions in blocks $(1, 2)$ and $(2, 1)$. Other similar interactions between clusters, that is, along a row, may also be observed. In particular, this stability analysis suggests that there are subclasses, particularly within the malignant classes. This may be more useful than the simple stability scores.

7.3. Analysis of the Yeast Data Set

As a third example of how our method can be applied, we use the yeast data set first described in ref. 22 and available from ref. 21. This too is a classification data set. The sample size is 1484 and there are ten classes and eight explanatory variables. Clustering on this data set is quite difficult, as pointed out by Nakai and Kanehisa [22] as the classes do not separate easily. In fact, in contrast to the breast cancer data set where we found stable subclasses within the two apparent classes, our analysis here finds seven or eight classes is more stable than the ten apparent classes suggesting that some of the classes overlap significantly.

Our analysis here begins by rounding the data, that is, transforming by the inverse covariance matrix, and then finding the K-means clusterings for a range of K. Then we generated the stability sequence in Panel (a) of Fig. 7. This shows $K = 8$ is most stable, but values five through nine are not bad.

To investigate further, we plotted stability heatmaps for $K = 6, 7, 8, 9$ in panel (b)–(e) of Fig. 7. As can be seen, when $K = 6$, the blocks on the main diagonal are not very light and on each row there is not a lot of difference between the cluster from the main diagonal and its competing clusters. For $K = 7$, the blocks off the main diagonal are a little darker than those on the main diagonal and the contrast is stronger for $K = 8$ and $K = 9$. However, there is little (if any) improvement in the contrast between the blocks on and off the main diagonal in moving from $K = 8$ to $K = 9$. Note that the bars along the right hand side of the heatmaps indicate how well the clustering tends to reproduce the class labels. While the classes do not separate well, the clusters tend to consist of one or two characteristic classes, indicating that we have indeed recovered some structure, an observation borne out by noting that many of the clusters are well separated. Furthermore, a practitioner can easily note from this plot which clusters are well separated and which clusters border each other—and may hence be regarded as overlapping to the point that merging them should be considered. Such structure is likely more interesting in practice than the simple stability scores.

In this example we have presented the stability analysis using sphered the data, rather than raw or studentized data, because K-means clustering with the rounding of data gave the most reasonable stability sequence and the heatmap of the most stable clustering showed clear distinctions between the classes. Although we did not present them here, the stability sequences for the raw or studentized data had a maximum at two and then declined. This seemed unrealistic especially because the corresponding heatmaps did not display distinct well-separated components as clearly as panels (b)–(e) of Fig. 7. As a generality is remains unclear when to sphere, studentize or use the data 'as is' so in this example we have only shown the most stable version we found.

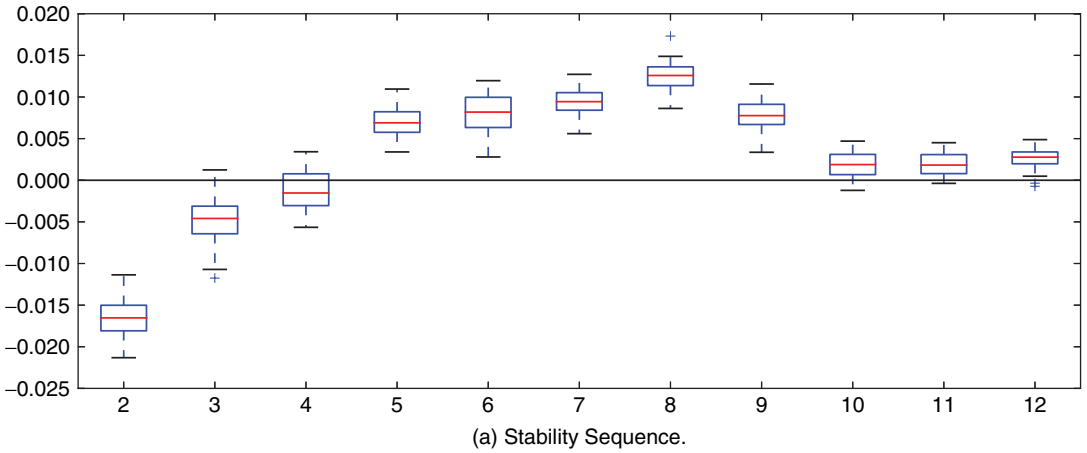
7.4. Experimental Conclusions

To conclude this section, we comment briefly on the overall performance of each of the methods we used here, for the case that p is not too large relative to n.

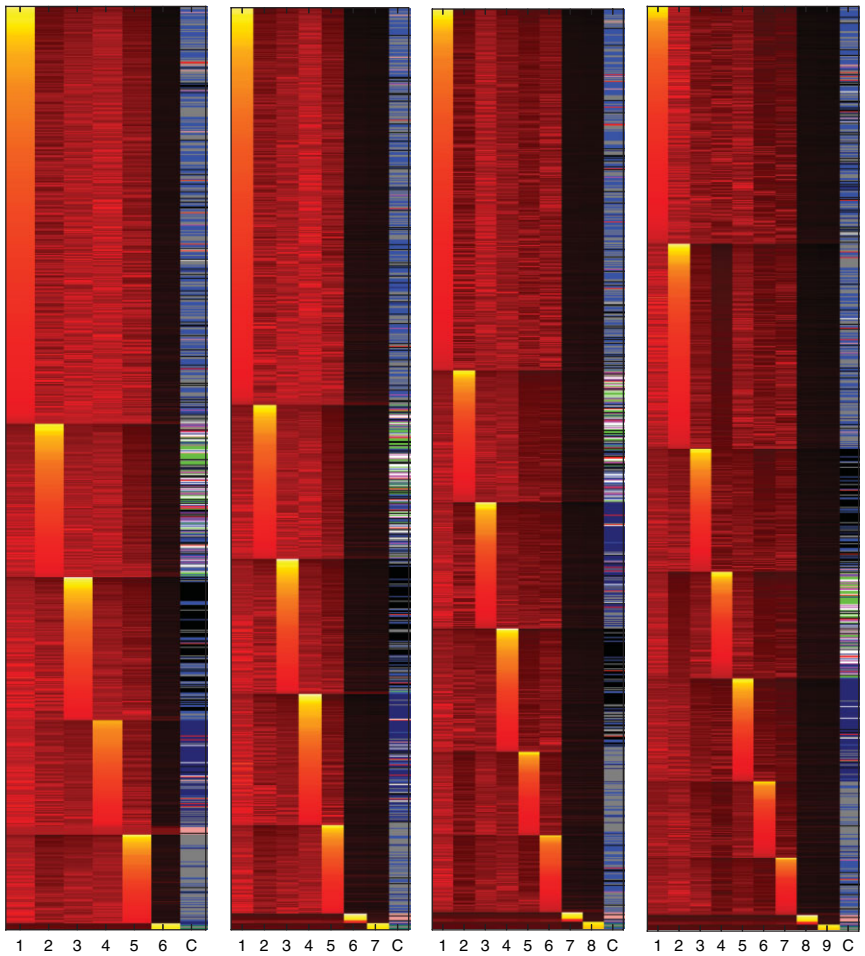
Gap Statistic: The original gap statistic performs quite poorly in comparison to other methods. There seem to be two reasons. First, the gap statistic does not seem to capture the *most* statistically significant clustering, but rather the *first* statistically significant clustering and so it does not distinguish well between two clusters and two well-separated groups of clusters (with one or both of the groups containing well-separated clusters). By focusing on boundary regions defined by scaling perturbations, our method better accounts for this situation. It should be noted, however, that better use of the plot of the gap statistic as a function of K may lead to better choices for K.

Second, when there is a single cluster that is long and thin, breaking it into two separate clusters gives a substantial drop in the dispersion measure on which the gap statistic is based. This can cause two clusters to appear more stable than one. For instance, as seen in the $T3$ data (where $K = 1$ is correct) in Section 7.1, the gap statistic routinely chooses $\hat{K} = 2$ even though the boundary region between the two clusters would clearly indicate $K = 2$ gives an unstable clustering.

Subsampling: The subsampling method employed here generally performs best among the methods we compared



(a) Stability Sequence.



(b) $K = 6$. (c) $K = 7$. (d) $K = 8$. (e) $K = 9$.

Fig. 7 Top: The stability sequence for K-means clusterings for $K = 2, \ldots, 12$ for the yeast data set. The most stable choice is $K = 8$. Bottom: The stability heatmaps for $K = 6, 7, 8, 9$. The original ten classes have the following color codes: CYT is blue, ERL is dark green, EXC is magenta, ME1 is bright green, ME2 is white, ME3 is dark blue, MIT is black, NUC is gray, POX is pink, and VAC is red. These classes do not separate well but there are distinct groups forming each cluster, particularly at $K = 8$. The bar on the right of each heatmap shows that most stable blocks tend to be from the same class but the tendency is not strong. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

against. However, it suffers several deficiencies. First, subsampling is not able to choose $K = 1$ when a single cluster is correct. Second, it suffers from the same two effects that make using the gap statistic ineffective. Specifically, a single long, thin cluster can be consistently partitioned into two clusters under subsampling. Indeed, the only case in which subsampling outperforms our method is with the 20 dimensional $T2$ data in Section 7.1 in which the mixture components are tightly packed but have shapes not too far from a standard normal. Furthermore, two well-separated groups of clusters can bias the most stable clustering toward low values of $\hat{K}$. In low dimensions, where we expect this effect to be stronger, subsampling consistently underestimates the number of clusters. On the other hand, our results show subsampling handles these cases better than the gap statistic.

Silhouette: The silhouette method is based on an intuition that is similar to our method. However, our perturbation method outperforms it in every example, particularly in higher dimensions. We believe that our perturbation method has three distinct advantages over silhouette. First, the silhouette method only considers the distance between a point and its closest two clusters; thus it does not account for points close to the boundary of three or more clusters, a less stable case. Second, the silhouette statistic does not inherently include uncertainty information. This makes it more difficult to choose between competing maxima; it is unclear if the difference between two similar values is because of inherent structure or noise. Finally, this prevents the silhouette method from detecting the case where $K = 1$, as there is no way to compare against the $K = 1$ case meaningfully.

Perturbations with other distances: We used a minimum of norms and an average linkage distance as the point-to-cluster distance in Section 2.2 and in Sections 7.2 and 7.3. In the first, the minimum norm was necessary for physical reasons and the heatmap did not indicate high stability; the MNIST data likely does not have high cluster stability. In the breast cancer and yeast data, the average Euclidean distance did not always perform as well as might be hoped. On the other hand, it is known that average linkage tends to underestimate the number of clusters. However, using average linkage still outperformed the gap statistic in every case. Furthermore, it gives us a way to apply our method when the notion of a centroid is not clear.

8. CONCLUSIONS

We have proposed a new technique for assessing the stability of a clustering. This technique is based on evaluating the probability of a set of multiplicative factors on the point-to-cluster distances that preserve their relative sizes. In practice, if we have an optimal clustering of size K, for each K in a range, then we compare the probability to a baseline formed by taking bootstrap samples. We then choose the value of K giving the largest value compared to the baseline. We have shown that this method has appealing theoretical properties such as consistency for K and that it has behavior that matches what we would intuitively want from a stability method. Indeed, as our examples show, our methods provide not just an examination of overall stability but also an examination of the stability of each cluster.

While our theory is limited to using point-to-cluster distances that are metrics, the method itself is not. We show this is two examples (Sections 2.2 and 7.2) and argue that when centroid-based clustering is not appropriate it will be necessary to use distances that respond to the shapes of the clusters. This is not atypical: Silhouette distances can become unrepresentative if the clusters are far from convex and our perturbation methods can overstate the instability of nonconvex clusterings.

We have not explicitly examined the performance of our stability technique with high dimensional data—the highest dimension we have considered is 20 –nor have we studied our method in the case $p > n$. However, some guidance can be given. First, $p > n$ is permitted by our theory; the question is how much larger than n we can let p be while still getting useful results. Second, for fixed n and L^p-metrics, clustering becomes unstable as p increases relative to n unless the spread of the qth coordinates (for $q \leq p$) in the data points stays large enough; see ref. 23. Moreover, even when the data points are sufficiently distinct the geometry of a high-dimensional data set may be better described by an ultrametric or other technique intended for use in higher dimensions [24].

The implication of these two points is that the appropriate metric for stability assessment is a function of p and n as well as the spread of the data. For p not too large (relative to n), the case studied here, commonly used metrics or linkages based on them will be appropriate. However, when p is large enough, ultrametrics may be appropriate; it is unclear whether the size of p alone or whether the size of p relative to n is the defining feature of this case, see ref. 25, Sec. 5 and ref. 26, Sections 2 and 3.

Overall, the usefulness of our method for a given data set rests on having made a reasonable choice for d to assess the point-to-cluster distances and for the prior to assess how perturbable the distances are. Reasonable choices depend on the dimension p, how close the data points are to each other and the cluster centroids. As long as d and the prior accurately reflect the structure of the data, our method should give good results; our theoretical results hold for very general priors and distance measures.

APPENDIX A. PROOF OF THEOREM 1

Let

$$\phi_{1,k}(X_i) = \mathbb{I}_{\{X_i \in C_k\}} \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(X_i, \hat{\mu}_k) \leq \lambda_\ell d(X_i, \hat{\mu}_\ell)\}} \mathrm{d}F(\lambda^K) \tag{47}$$

and

$$\phi_{2,k}(X_i) = \mathbb{I}_{\{X_i \in C_k\}} \int \mathbb{I}_{\{\forall \ell \neq k : \lambda_k d(X_i, \mu_k) \leq \lambda_\ell d(X_i, \mu_\ell)\}} \mathrm{d}F(\lambda^K). \tag{48}$$

Clearly, $E\phi_{2,k}(X_i) = \phi_k(X)$. So, for each $k = 1, \ldots, K$, the triangle inequality gives the bound

$$\left| \frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_k(X_i) - \phi_k(X) \right| \leq \left| \frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_k(X_i) - \frac{1}{n} \sum_{i=1}^{n} \phi_{1,k}(X_i) \right| \tag{49}$$

$$+ \left| \frac{1}{n} \sum_{i=1}^{n} \phi_{1,k}(X_i) - \frac{1}{n} \sum_{i=1}^{n} \phi_{2,k}(X_i) \right| \tag{50}$$

$$+ \left| \frac{1}{n} \sum_{i=1}^{n} \phi_{2,k}(X_i) - E\phi_{2,k}(X) \right|, \tag{51}$$

in which it is easy to see that Eq. (51) $\to 0$ by the law of large numbers.

To show Eq. (49) goes to zero, note it is

$$\left| \frac{1}{n} \sum_{i=1}^{n} \int \mathbb{I}_{\lambda_k d(X_i, \hat{\mu}_k) \leq \min_{\ell \neq k} \lambda_\ell d(X_i, \hat{\mu}_\ell)} \left(\mathbb{I}_{X_i \in C_k} - \mathbb{I}_{X_i \in \hat{C}_k} \right) \mathrm{d}F(\lambda^K) \right|$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{I}_{X_i \in C_k \setminus \hat{C}_k} + \mathbb{I}_{X_i \in \hat{C}_k \setminus C_k} \right). \tag{52}$$

To bound the sum of indicator functions, recall

$$C_k = \{x \mid d(x, \mu_k) \leq \min_{\ell \neq k} d(x, \mu_\ell)\}$$

and

$$\hat{C}_k = \{x \mid d(x, \hat{\mu}_k) \leq \min_{\ell \neq k} d(x, \hat{\mu}_\ell)\}$$

and note that the triangle inequality on the metric d gives, for any ℓ,

$$|d(x, \hat{\mu}_\ell) - d(x, \mu_\ell)| \leq d(\hat{\mu}_\ell, \mu_\ell) \to 0$$

in probability. So, for any x and any ℓ, $d(x, \hat{\mu}_\ell) \to d(x, \mu_\ell)$ in probability and in particular we have for any k

$$d(x, \hat{\mu}_k) \to d(x, \mu_k) \quad \text{and} \quad \min_{\ell \neq k} d(x, \hat{\mu}_k) \to \min_{\ell \neq k} d(x, \mu_\ell)$$

in probability. Taken together this means that for all k, $\hat{C}_k \to C_k$, in the sense of Eq. (10).

Now, let $\epsilon > 0$ and define the upper and lower approximations to C_k,

$$C_{k,\epsilon}^{+} = \left\{ x \mid d(x, \mu_k) - \epsilon \leq \min_{\ell \neq k} d(x, \mu_\ell) + \epsilon \right\}$$

and

$$C_{k,\epsilon}^{-} = \left\{ x \mid d(x, \mu_k) + \epsilon \leq \min_{\ell \neq k} d(x, \mu_\ell) - \epsilon \right\}$$

so that $C_{k,\epsilon}^{-} \subset C_k \subset C_{k,\epsilon}^{+}$. Therefore, letting

$$U = \{\hat{C}_k \subset C_{k,\epsilon}^{+}\} \quad \text{and} \quad V = \{\hat{C}_k \supset C_{k,\epsilon}^{-}\}$$

gives that $P(U), P(V) \to 1$ if $n \to \infty$ and then $\epsilon \to 0$.

We can now bound Eq. (52) by noting

$$0 \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \in C_k \setminus \hat{C}_k} \mathbb{I}_V + \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \in C_k \setminus \hat{C}_k} \mathbb{I}_{V^c}$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \in \hat{C}_k \setminus C_k} \mathbb{I}_U + \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \in \hat{C}_k \setminus C_k} \mathbb{I}_{U^c}$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \in C_k \setminus C_{k,\epsilon}^{-}} + \mathbb{I}_{V^c} + \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \in C_{k,\epsilon} \setminus C_k} + \mathbb{I}_{U^c}.$$

So, taking the expectation on both sides gives that the expectation of Eq. (49) is bounded from below by zero and from above by

$$P(X \in C_k \setminus C_{k,\epsilon}^{-}) + \eta_1 + P(V^c)$$

$$+ \eta_2 + P(X \in C_{k,\epsilon}^{+} \setminus C_k) + \eta_3 + P(U^c) + \eta_4,$$

where $\eta_1, \ldots, \eta_4 > 0$ give (upper) bounds on the convergence of the indicator functions to their probabilities. As L^1 convergence implies convergence in probability, if we let $n \to \infty$, and then let $\epsilon \to \infty$ then we can let the η_j's go to zero to complete this part of the argument. That is, with probability at least $1 - \xi$, for pre-assigned $\xi > 0$, there is an N so large that $n > N$ ensures Eq. (49) has an upper bound that goes to zero, that is, Eq. 49 $\to 0$ in probability.

It remains to bound Eq. (50). Parallel to the sets C_k, $\hat{C}_k$, $C_{k,\epsilon}^{+}$, and $C_{k,\epsilon}^{-}$ define the sets

$$G_k(X) = \left\{ \lambda_k d(X, \mu_k) \leq \min_{\ell \neq k} \lambda_\ell d(X, \mu_\ell) \right\}$$

$$\hat{G}_k(X) = \left\{ \lambda_k d(X, \hat{\mu}_k) \leq \min_{\ell \neq k} \lambda_\ell d(X, \hat{\mu}_\ell) \right\}$$

$$G_{k,\epsilon}^{+}(X) = \left\{ \lambda_k d(X, \hat{\mu}_k) - \epsilon \leq \min_{\ell \neq k} \lambda_\ell (d(X, \hat{\mu}_\ell) + \epsilon) \right\}$$

$$G_{k,\epsilon}^{-}(X) = \left\{ \lambda_k d(X, \hat{\mu}_k) + \epsilon \leq \min_{\ell \neq k} \lambda_\ell (d(X, \hat{\mu}_\ell) - \epsilon) \right\}.$$

Also, write

$$W_{k,i,\epsilon} = \{X_i \in C_k \mid |d(X_i, \hat{\mu}_k) - d(X_i, \mu_k)| < \epsilon\}$$

and let $B(b) = \{\lambda_1^K \mid \forall k : 1/b \leq \lambda_k \leq b\}$ so that as $b \to \infty$ $F(B(b)^c) \to 0$.

Now, since all the summands are bounded by one, we have that Eq. (50) is

$$0 \leq \left| \frac{1}{n} \sum_{i=1}^{n} \phi_{1,k}(X_i) - \frac{1}{n} \sum_{i=1}^{n} \phi_{2,k}(X_i) \right|$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \in C_k} \int \left| \mathbb{I}_{\hat{G}_k(X_i)} - \mathbb{I}_{G_k(X_i)} \right| \mathbb{I}_{B(b)} \mathrm{d}F(\lambda_1^K) + F(B(b)^c)$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \in C_k} \int \left(\mathbb{I}_{\hat{G}_k(X_i) \setminus G_k(X_i)} + \mathbb{I}_{G_k(X_i) \setminus \hat{G}_k(X_i)} \right)$$

$$\times \mathbb{I}_{B(b)} \mathrm{d}F(\lambda_1^K) + F(B(b)^c)$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{X_i\in C_k}\mathbb{I}_{W_{k,i,\epsilon}}\int \mathbb{I}_{G^+_{k,\epsilon}(X_i)\setminus G_k(X_i)}\mathbb{I}_{B(b)}\mathrm{d}F(\lambda_1^K) \tag{53}$$

$$+\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{X_i\in C_k}\mathbb{I}_{W^c_{k,i,\epsilon}}\int \mathbb{I}_{G^+_{k,\epsilon}(X_i)\setminus G_k(X_i)}\mathbb{I}_{B(b)}\mathrm{d}F(\lambda_1^K) \tag{54}$$

$$+\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{X_i\in C_k}\mathbb{I}_{W_{k,i,\epsilon}}\int \mathbb{I}_{G_k(X_i)\setminus G^-_{k,\epsilon}(X_i)}\mathbb{I}_{B(b)}\mathrm{d}F(\lambda_1^K) \tag{55}$$

$$+\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{X_i\in C_k}\mathbb{I}_{W^c_{k,i,\epsilon}}\int \mathbb{I}_{G_k(X_i)\setminus G^-_{k,\epsilon}(X_i)}\mathbb{I}_{B(b)}\mathrm{d}F(\lambda_1^K)+F(B(b)^c). \tag{56}$$

Now, to see Eq. (53) $\to 0$, note that

$$0\leq (53)\leq \int \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{X_i\in C_k}\mathbb{I}_{W_{k,i,\epsilon}}\mathbb{I}_{G^+_{k,\epsilon}(X_i)\setminus G_k(X_i)}\mathbb{I}_{B(b)}\mathrm{d}F(\lambda_1^K).$$

So, taking expectations gives that

$$0\leq E\int \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{X_i\in C_k}\mathbb{I}_{W_{k,i,\epsilon}}\mathbb{I}_{G^+_{k,\epsilon}(X_i)\setminus G_k(X_i)}\mathbb{I}_{B(b)}\mathrm{d}F(\lambda_1^K)$$
$$\leq \int\left(E\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{X_i\in C_k}\mathbb{I}_{W_{k,i,\epsilon}}\mathbb{I}_{G^+_{k,\epsilon}(X_i)\setminus G_k(X_i)}\mathbb{I}_{B(b)}\right)\mathrm{d}F(\lambda_1^K).$$

For each fixed λ_1^K the integrand (in parentheses) goes to zero and is bounded by one. By the bounded convergence theorem, the asymptotic upper bound for the expectation of Eq. (53) is zero as $n\to\infty$ and $\epsilon\to 0$, that is, Eq. (53) converges to zero in L^1 which implies it converges to zero in probability as well.

To see Eq. 54 $\to 0$ we use a similar argument. Observe that

$$0\leq \text{Eq.}(54)\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{W^c_{k,i,\epsilon}}.$$

As with Eq. (53), we obtain L^1 convergence of the upper bound. Thus,

$$E\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{W^c_{k,i,\epsilon}}=P(W^c_{k,1,\epsilon})\to 0,$$

as $\epsilon\to 0$. (Note that the X_i's are IID so we have arbitrarily chosen $i=1$ for the event in the probability for convenience.) This gives that Eq. 54 $\to 0$ in probability.

We can show Eq. 55 $\to 0$ in probability by a technique similar to that used for Eq. (53) and we can show the first term in Eq. (56) $\to 0$ in probability by a technique similar to that used for Eq. (54).

APPENDIX B. PROOF OF PROPOSITION 2

We can write

$$\phi_k=\int\left(\prod_{\ell\neq k}\mathbb{I}_{\{\lambda_k d_k\leq\lambda_\ell d_\ell\}}f(\lambda_\ell|\theta)\right)f(\lambda_k|\theta)\mathrm{d}\lambda^K$$
$$=\int_1^\infty\left(\prod_{\ell\neq k}\int_1^\infty\theta\mathrm{e}^{-\theta(\lambda_\ell-1)}\int_0^\infty\delta((d_\ell\lambda_\ell-d_k\lambda_k)-t_\ell)dt_\ell d\lambda_\ell\right)$$
$$\times f(\lambda_k|\theta)\mathrm{d}\lambda_k$$

where $\delta(\cdot)$ is the Dirac delta function, defined so that $\int_a^b\delta(x-t)\mathrm{d}t=1$ if $a\leq x\leq b$ and zero otherwise. This means,

$$\mathbb{I}_{a\leq b}=\int_0^\infty\delta(b-a-t)dt\quad\text{and}\quad f(x)=\int f(t)\delta(x-t)\mathrm{d}t,$$

see ref. 27. So, by Fubini's theorem we can interchange the order of integration to integrate over the λ_ℓ's first. We get that

$$\phi_k^{\mathrm{LE}}=\int_0^\infty\left(\prod_{\ell\neq k}\int_0^\infty\frac{\theta}{d_\ell}\mathrm{e}^{-\theta\left((t_\ell+\lambda_k d_k)/d_\ell\right)-1}\mathbb{I}_{\left((t_\ell+\lambda_k d_k)/d_\ell\right)\geq 1}dt_\ell\right)$$
$$\times f(\lambda_k|\theta)\mathrm{d}\lambda_k$$
$$=\int_0^\infty\left(\prod_{\ell\neq k}\exp\left(-\theta\left(\frac{d_k}{d_\ell}\lambda_k-1\right)-\frac{\theta}{d_\ell}t_\ell\right)\Big|_{t_\ell=\max(d_\ell-\lambda_k d_k,0)}\right)$$
$$\times f(\lambda_k|\theta)\mathrm{d}\lambda_k$$
$$=\int_0^\infty\prod_{\ell\neq k}\begin{cases}\exp\left(-\theta\left(\frac{d_k}{d_\ell}\lambda_k-1\right)\right) & \lambda_k\geq d_\ell/d_k\\ 1 & \text{otherwise}\end{cases}f(\lambda_k|\theta)\mathrm{d}\lambda_k. \tag{57}$$

Now, we break the domain of integration into subdomains so we can use use the fact that many of the terms in the product are one for some subdomains. If we sort the terms by increasing d_ℓ, we can handle the conditional terms in the product by breaking the integration up into $K+1$ possibly empty intervals, doing the integration separately on each interval. The properties of ψ now give that the boundaries of these regions are given by

$$A_m=\begin{cases}d_{\psi(m)}/d_k & m\in\{1,\ldots,K\}\\ \infty & m=K+1.\end{cases}$$

The intervals to integrate over are then

$$(0,A_1]\cap[1,\infty),(A_1,A_2]\cap[1,\infty),\ldots(A_K,A_{K+1}=\infty]\cap[1,\infty).$$

However, as $A_m=d_{\psi(m)}/d_k\leq 1$ for $m<\psi^{-1}(k)$, the first $\psi^{-1}(k)$ of these intervals are void. Thus 57 becomes

$$\phi_k^{\mathrm{LE}}=\sum_{j=\psi^{-1}(k)}^{K}\int_{A_k}^{A_{k+1}}\theta\mathrm{e}^{-\theta(\lambda_j-1)}\prod_{\substack{m=1\\ \psi(m)\neq k}}^{j}\mathrm{e}^{-\theta(A_m^{-1}\lambda_k-1)}\mathrm{d}\lambda_k$$
$$=\sum_{j=\psi^{-1}(k)}^{K}\int_{A_k}^{A_{k+1}}\theta\exp\left(-\theta\sum_{m=1}^{j}(A_m^{-1}\lambda_k-1)\right)\mathrm{d}\lambda_k,$$

where we have used the fact that $A_{\psi^{-1}(k)}=1$ to simplify the expression. The last integral can now be easily evaluated. We get

$$\phi_k^{\mathrm{LE}}=\sum_{j=\psi^{-1}(k)}^{K}\theta\mathrm{e}^{\theta j}\int_{A_k}^{A_{k+1}}\exp\left(-\theta\lambda_j\sum_{m=1}^{j}A_m^{-1}\right)\mathrm{d}\lambda_k$$
$$=\sum_{j=\psi^{-1}(k)}^{K}\left[\frac{\mathrm{e}^{\theta j}}{\sum_{m=1}^{j}A_m^{-1}}\exp\left(\theta\lambda_j\sum_{m=1}^{j}A_m^{-1}\right)\Big|_{\lambda_k=A_{j+1}}^{\lambda_k=A_j}\right.$$

If we collect similar terms and use the given definitions of C and D, the last expression becomes

$$\phi_k^{\mathrm{LE}} = \frac{\theta}{d_j} \sum_{j=\psi^{-1}(k)}^{K} \frac{C_k - C_{k+1}}{B_k}.$$

As the 'C terms' may be quite close, we re-express ϕ_k to help get better numerical stability by avoiding working with the difference of two similar numbers. Thus,

$$\phi_k^{\mathrm{LE}} = \frac{\theta}{d_k}\left(\frac{C_{\psi^{-1}(k)}}{B_{\psi^{-1}(k)}} - \frac{C_{\psi^{-1}(k)+1}}{B_{\psi^{-1}(k)}} + \frac{C_{\psi^{-1}(k)+1}}{B_{\psi^{-1}(k)+1}} - \frac{C_{\psi^{-1}(k)+2}}{B_{\psi^{-1}(k)+1}} + \cdots + \frac{C_K}{B_K}\right), \tag{58}$$

where the arguments of ϕ_k^{LE}, B, and C are understood. However, terms B_{k-1} and B_k are likely to be close together so for greater numerical stability $B_{j-1}^{-1} - B_j^{-1}$ can be expressed as

$$\frac{1}{B_{k-1}} - \frac{1}{B_k} = \frac{1}{B_{k-1}} - \left(B_{k-1} + \frac{\theta}{d_{\psi(k)}}\right)^{-1} = \left(B_{k-1}\left(B_{k-1}\frac{d_{\psi(k)}}{\theta} + 1\right)\right)^{-1}.$$

Putting this back into Eq. (58) gives:

$$\phi_k^{\mathrm{LE}} = \frac{\theta}{d_k}\left(\frac{C_{\psi^{-1}(k)}}{B_{\psi^{-1}(k)}} - \sum_{j=\psi^{-1}(k)+1}^{K} C_j\left(B_{j-1}\left(B_{j-1}\frac{d_{\psi(j)}}{\theta} + 1\right)\right)^{-1}\right) = \frac{\theta}{d_k}\left(\frac{C_{\psi^{-1}(k)}}{B_{\psi^{-1}(k)}} - D_{\psi^{-1}(k)}\right).$$

REFERENCES

[1] A. Ben-Hur, A. Elisseeff, and I. Guyon, A stability method for discovering structure in clustered data, Pacif Symp Biocomput 7 (2002), 6–17.

[2] C. Giurcaneanu, and I. Tabus, Cluster structure inference based on clustering stability with applications to microarray data analysis, EURASIP J Appl Signal Process 1 (2004), 64–80.

[3] O. Abul, A. Lo, R. Alhajj, F. Polat, and K. Barker, Cluster validity analysis using subsampling, IEEE Int Conf Syst Man Cyber 2 (2003), 1435–1440.

[4] T. Lange, V. Roth, M. Braun, and J. Buhmann, Stability-based validation of clustering schemes, Neural Comp 16 (2004), 1299–1323.

[5] U. Moller, and D. Radke, A cluster validity approach based on nearest-neighbor resampling, Proc 18th Int Conf Pattern Recogn 1 (2006), 892–895.

[6] L. Hubert, and P. Arabie, Comparing partitions, J Classif 2 (1985), 193–218.

[7] M. Meila, Comparing clusterings: an information based distance, J Multivariate Anal 98 (2007), 873–895.

[8] P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J Comput Appl Math 20 (1987), 53–65.

[9] R. Tibshirani, Estimating the number of clusters in a data set via the gap statistic, J R Stat Soc Ser B 63 (2001), 411–423.

[10] T. Simpson, J. Armstrong, and A. Jarman, Merged consensus clustering to assess and improve class discovery with microarray data, BMC Bioinform 11 (2010), 590.

[11] S. Monti, P. Tamayo, J. Merisov, and T. Golub, Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data, Mach Learn 52 (2003), 91–118.

[12] H. Wang, H. Shan, and A. Bannerjee, Bayesian cluster ensembles, Stat Anal Data Mining 4 (2011), 54–70.

[13] S. Datta and S. Datta, Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes, BMC Bioinform 7 (2006), 397.

[14] Y. LeCun and C. Cortes, The MNIST database of handwritten digits, NEC Research Institute, 1998.

[15] W. Newey, and D. McFadden, Large sample estimation and hypothesis testing, In Handbook of Econometrics, Vol IV, R. Engle and D. McFadden, eds. Amsterdam, North Holland, 1994, 2111–2245.

[16] D. Pollard, Consistency of K-means clustering, Ann Stat 9 (1981), 135–140.

[17] S. Ray and B. Lindsay, The topography of multivariate normal mixtures, Ann Stat 33 (2005), 2042–2065.

[18] H. Koepke, Bayesian Cluster Validation, Master's Thesis, Dept. of Computer Science, University of British Columbia, 2008.

[19] H.-P. Kriegel, P. Kroger, and A. Zimek, Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering, ACM Trans Knowledge Disc Data 3 (2009), 1–58.

[20] W. Street, W. Wolberg, and O. Mangasarian, Nuclear feature extraction for breast tumor diagnosis, SPIE International Symposium on Electronic Imaging: Science and Technology, 1905, 1993, 861–870.

[21] A. Frank, and A. Asuncion, 2010. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Science, http://archive.ics.uci.edu/ml. Accessed January 3, 2013.

[22] K. Nakai and M. Kanehisa, Expert system for predicting protein localization sites in gram-negative bacteria, Proteins Struct Funct Bioinform 11 (1991), 95–110.

[23] H. Koepke and B. Clarke, On the limits of clustering in high dimensions via cost functions, Stat Anal Data Mining 4 (2011), 30–53.

[24] F. Murtagh, The remarkable simplicity of very high dimensional data: application to model-based clustering, J Classif 26 (2011), 249–277.

[25] F. Murtagh, Ultrametric model of mind, I: review, p-Adic Numbers Ultramet Anal Appl 4 (2012), 207–221.

[26] F. Murtagh, Ultrametric model of mind, II: application to text content analysis, p-Adic Numbers Ultramet Anal Appl 4 (2012), 193–206.

[27] G. Arfken and H. Weber, Mathematical Methods for Physicists, San Diego, Academic Press, 1995.