Model Selection and Stability in k-means Clustering

Ohad Shamir † **and Naftali Tishby**†‡
† School of Computer Science and Engineering
‡ Interdisciplinary Center for Neural Computation
The Hebrew University, Jerusalem 91904, Israel
{ohadsh,tishby}@cs.huji.ac.il
Eligible for the Mark Fulk Award

Abstract

Clustering Stability methods are a family of widely used model selection techniques applied in data clustering. Their unifying theme is that an appropriate model should result in a clustering which is robust with respect to various kinds of perturbations, as measured by a suitable instability measure. Despite their relative success, not much is known theoretically on why or when they work, or even what kind of assumptions they make in choosing an 'appropriate' model. In this paper, we focus on the behavior of clustering stability using k-means clustering. Our main technical result is an exact characterization of the value to which appropriately scaled measures of instability converge, based on a sample drawn from any distribution in $\mathbb{R}^n$ satisfying mild regularity conditions. Besides resolving a theoretical obstacle which has been raised in the literature, it allows us to draw several interesting observations about what kind of assumptions are actually made when using these methods. For example, it appears that clustering stability tends to choose models based on the probability density along the cluster boundaries. This is often a reasonable approach, but might also lead to unexpected consequences. Several other issues of theoretical and practical relevance are also discussed.

1 Introduction

The important and difficult problem of model selection in data clustering has been the focus of extensive literature spanning several research communities in the natural and social sciences. Since clustering is often used as a first step in the data analysis process, the questions of what type of clusters or how many clusters are in the data can be crucial.

Unfortunately, an objective 'correct' answer to these questions seldom occurs in practice. There might be several reasonable ones, depending on the resolution at which we inspect the data, and our (usually subjective) definition of what constitutes a cluster. The ill-posedness of the model selection problem is compounded by the unsupervised nature of the data, often making it difficult to assess the compatibility of a even a single specific model.

These difficulties suggest that the model selection procedure should be carefully chosen to fit the nature of the problem at hand, and what the practitioner is trying to achieve. For this, one needs a good grasp of the *assumptions* about the clustering structure that are inherent to each such procedure. Understanding these assumptions is not always trivial for general-purpose model selection methods, which are not tied to strong generative assumptions.

An important family of model selection methods, whose popularity has grown in the past few years, is based on clustering stability. The unifying theme of these methods is that an appropriate model for the data should result in a clustering which is robust with respect to various kinds of perturbations. In other words, if we choose an appropriate clustering algorithm, and feed it with the 'correct' parameters (such as the number of clusters, the metric used, etc.), the clustering returned by the algorithm should not be overly sensitive to the exact structure of the data.

In particular, we will focus on *sample based* clustering stability methods. In such methods, the stability of a model is determined by running our clustering algorithm on different random subsets of our data, and comparing the resulting clusterings. A measure of discrepancy between these clusterings defines the instability of our model choice[1]. One can then repeat this procedure for different models, and look for the model with the least instability.

Although these methods have been shown to be rather effective in practice (cf. [10],[2],[3],[8]), little theory exists so far to explain their success, or for which cases are they best suited for. Over the past few years, a theoretical study of these methods has been initiated, in a framework where the data are assumed to be drawn from an i.i.d sample. However, a fundamental hurdle was the observation [1] that under mild conditions and for any model choice, the clustering algorithm should tend to converge to a single solution which is optimal with respect to the underlying distribution. As a result, for large enough samples, we get approximately the same clustering hypothesis based on each random subsample, and thus achieve an instability measure of approximately zero, regardless of whether the model fits the data or not (this problem was also pointed out in [7]). A solution to this dif-

[1] These kind of measures are referred to in the literature as both stability measures and as instability measures. Since they usually increase with more instability, we will refer to them as instability measures in this paper.

ficulty was proposed in [16], based on concepts from learning theory. In a nutshell, that paper showed that the important factor in discerning stability for different models is not the asymptotic values of the instability measure, but rather the *exact convergence rate* to these values. With this more refined analysis, it was argued that differences in the instability values for different model choices should usually be discernible for any sample size, no matter how large, despite the universal convergence to absolute stability. Although it provided the necessary groundwork, that paper only rigorously proved this assertion for a single specific example, as a proof-of-concept.

In this paper, we formally resolve the problem for the well known and popular k-means clustering framework, when the goal is to determine the value of k, or the number of clusters in the data. Assuming an algorithm which minimizes the k-means objective function, we consider arbitrary distributions in $\mathbb{R}^n$ satisfying certain mild regularity conditions, and show that scaling the instability measure by the square root of the sample size invariably results in a sample-size-free measure. Rather than converging to zero as the sample size increases to infinity, this scaled measure converges to a finite value which differs based on the choice of k. In this sense, our results show that clustering stability is 'meaningful' even for arbitrarily large sample sizes, at least for the k-means clustering framework.

Moreover, we are able to explicitly characterize the value to which this instability measure converges, depending on the underlying distribution and on k, the number of clusters chosen. Although this value is exact only for asymptotically large samples, it is significant for two reasons. The first is that this result can be seen as an approximation which improves as the sample size increases. The second and more profound reason is that if we are interested in discovering what assumptions are implicit in performing model selection with clustering stability, these should not be overly dependent on the sample size used. Therefore, as we look at larger samples, noisy and hard to analyze finite sample effects diminish, and what remains are the truly important characteristics, which should be relevant for *any* sample size. As a result, the analysis leads to observations about the assumptions inherent in using stability for model selection in k-means clustering that have both theoretical and practical interest.

2 Problem Setting and Notation

We refer the reader to Fig. 1 for a graphical illustration of the basic setting, and some of the notation introduced below.

Denote $\{1,\ldots,k\}$ as $[k]$. Vectors will be denoted by bold-face characters, and are by default assumed to be in column form. $\|\cdot\|$ will denote the Euclidean norm unless stated otherwise. The notation $\xrightarrow{P}$ denotes convergence in probability. $\mathcal{N}(\boldsymbol{\mu},\Sigma)$ denotes the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

Let $\mathcal{D}$ be a probability distribution on $\mathbb{R}^n$, with a bounded probability function $p(\cdot)$ which is continuous with respect to the n-dimensional Lebesgue measure. Assume furthermore that the following holds:

- $\int_{\mathbb{R}^n} p(\mathbf{x})\|\mathbf{x}\|^2 d\mathbf{x} < \infty$ (in other words, $\mathcal{D}$ has bounded variance).
- There exists a monotonically decreasing function $g(\cdot): \mathbb{R}\to\mathbb{R}$, such that $p(\mathbf{x}) \le g(\|\mathbf{x}\|)$ for all $\mathbf{x}\in\mathbb{R}^n$, and $\int_{r=0}^{\infty} r^n g(r) < \infty$.

The second requirement is needed in order to apply the main theorem of [14] (it is a slightly stronger version of condition (iv) there), and can probably be improved. Nevertheless, it is quite mild, and holds in particular for any distribution that is not heavy-tailed or has bounded support. As to the continuity requirement of $p(\cdot)$, it should be noted that our results hold even if we assume continuity solely on arbitrary open sets containing the optimal cluster borders, but we will take this stronger assumption for simplicity. In any case, discontinuous probability densities can be arbitrarily well approximated by a continuous one, so this is a mild requirement as well.

Let $\mathtt{A_k}$ denote an 'ideal' version of the standard k-means algorithm, which is given a sample $S=\{\mathbf{x}_i\}_{i=1}^m \subseteq \mathbb{R}^n$, sampled i.i.d from $\mathcal{D}$, and a required number of clusters k, and returns a set of centroids $\mathbf{c} = (\mathbf{c}_i,\ldots,\mathbf{c}_k)^\top \in \mathbb{R}^{nk}$, which minimize the objective function:

$$\hat{W}(\mathbf{c}) = \frac{1}{m}\sum_{i=1}^{m}\min_{j\in[k]}\|\mathbf{c}_j-\mathbf{x}_i\|^2.$$

We assume that the numbering of the centroids is by some uniform canonical ordering (for example, by sorting with respect to the coordinates). Let $\boldsymbol{\mu} = (\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k)^\top \in \mathbb{R}^{nk}$ be an optimal k-means solution with respect to $\mathcal{D}$, defined as a minimizer of

$$W(\mathbf{c}) = \int_{\mathbb{R}^n} p(\mathbf{x})\min_{j\in[k]}\|\mathbf{c}_j-\mathbf{x}_i\|^2 d\mathbf{x}.$$

We assume that such a minimizer exists, is unique up to permutation of the centroids, and that all centroids are distinct (for all $i\neq j$, $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$).

For some set of centroids $\mathbf{c}=(\mathbf{c}_1,\ldots,\mathbf{c}_k)$, and for each cluster centroid $\mathbf{c}_i$, we designate the interior of its corresponding cluster as $C_{\mathbf{c},i}$, defined as:

$$C_{\mathbf{c},i} = \left\{\mathbf{x}\in\mathbb{R}^n : \arg\min_{j\in[k]}\|\mathbf{c}_j-\mathbf{x}\|^2 = i\right\}.$$

From the continuity assumptions on p, we may assume that the set of points not in the interior of some cluster has zero measure with respect to p. We can therefore neglect the issue of how points along cluster boundaries are assigned.

The (scaled) instability measure we will use is defined as

$$\widehat{\text{instab}}(\mathtt{A_k}, m, \mathcal{D}) = \sqrt{m}\mathbb{E}_{S_1,S_2\sim\mathcal{D}^m}\left[d_{\mathcal{D}}(\mathtt{A_k}(S_1),\mathtt{A_k}(S_2))\right],$$

where $d_{\mathcal{D}}(\mathtt{A_k}(S_1),\mathtt{A_k}(S_2))$ is defined as

$$\Pr_{\mathbf{x}_1,\mathbf{x}_2\sim\mathcal{D}}\left(\mathtt{A_k}(S_1)(\mathbf{x}_1,\mathbf{x}_2)\neq\mathtt{A_k}(S_2)(\mathbf{x}_1,\mathbf{x}_2)\right),$$

and $\mathtt{A_k}(S)(\mathbf{x}_1,\mathbf{x}_2)$ is an indicator function of whether the instances $\mathbf{x}_1,\mathbf{x}_2$ are in the same cluster according to the clustering given by $\mathtt{A_k}(S)$. This definitions follows that of [16] and [1] (with the addition of scaling by $\sqrt{m}$), and captures

the essence of instability measures used in practice. Notice that $\mathtt{A}_\mathtt{k}(S_1)$,$\mathtt{A}_\mathtt{k}(S_2)$ and $d_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$ are all random variables, defined as functions of S_1 and S_2, which are random variables with distribution $\mathcal{D}^m$.

Any choice of cluster centroids $\mathbf{c}$ induces a Voronoi partition on $\mathbb{R}^n$. We will denote $F_{\mathbf{c},i,a}$, for $i \neq a$, as the boundary face between clusters i and a. Namely, the points in $\mathbb{R}^n$ whose two closest cluster centroids are $\mathbf{c}_i$ and $\mathbf{c}_a$, and are equidistant from them:

$$F_{\mathbf{c},i,a} = \left\{ \mathbf{x} \in \mathbb{R}^n \middle| \arg\min_{j \in [k]} \|\mathbf{c}_j - \mathbf{x}\|^2 = \{i, a\} \right\}.$$

Assuming c_i,c_j are distinct, $F_{\mathbf{c},i,j}$ is a (possibly empty) subset of the hyperplane $H_{\mathbf{c},i,j}$, defined as:

$$H_{\mathbf{c},i,j} = \left\{ \mathbf{x} \in \mathbb{R}^n : \left(\mathbf{x} - \frac{\mathbf{c}_i + \mathbf{c}_j}{2} \right)^\top \cdot (\mathbf{c}_1 - \mathbf{c}_2) = 0 \right\}.$$

In our discussion, we use integrals with respect to both the n-dimensional Lebesgue measure, as well as the $(n-1)$-dimensional Lebesgue measure. The type of integral we are using should be clear from context, depending on the set over which we are integrating. For example, integrals over some C_i are of the first type, while integrals over some $F_{\mathbf{c},i,j}$ are of the second type.

We now define certain matrices which we will need later on. Let I_n be the identity matrix of dimensions $n \times n$. Let Γ be a $kn \times kn$ matrix, composed of $k \times k$ blocks $\Gamma_{i,j}$ for $i, j \in [k]$. Each block $\Gamma_{i,j}$ is defined as

$$\Gamma_{i,j} = 2 \left[\int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x}) d\mathbf{x} \right] I_n - 2 \sum_{a \neq i} \frac{\int_{F_{\boldsymbol{\mu},i,a}} p(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^\top d\mathbf{x}}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_a\|}$$

if $i = j$, and

$$\Gamma_{i,j} = \frac{2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|} \int_{F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_j)^\top d\mathbf{x}$$

if $i \neq j$.

We will use the same block notation later for its inverse Γ^{-1}. This matrix is the Hessian of the mapping $W(\mathbf{c})$ defined above [2] at $\mathbf{c} = \boldsymbol{\mu}$. The existence of these integrals follows from the assumptions on $p(\cdot)$ (see the proof of the main theorem in [14]). We assume that the matrix Γ is positive definite. This is in fact an almost redundant requirement, since the optimality of $\boldsymbol{\mu}$ entails that Γ is always positive semidefinite. Therefore, cases where Γ is not positive definite correspond to singularities which are apparently pathological (for more discussion on this, see [15]).

Let V be a $kn \times kn$ matrix, composed of k diagonal blocks V_i of size $n \times n$ for $i \in [k]$ (all other elements of V are zero). V_i is defined as:

$$V_i = 4 \int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^\top d\mathbf{x}.$$

[2] This assertion is proven in [14]. The definition of Γ there differs from ours in one of the signs, apparently due to a small error in that paper [13].

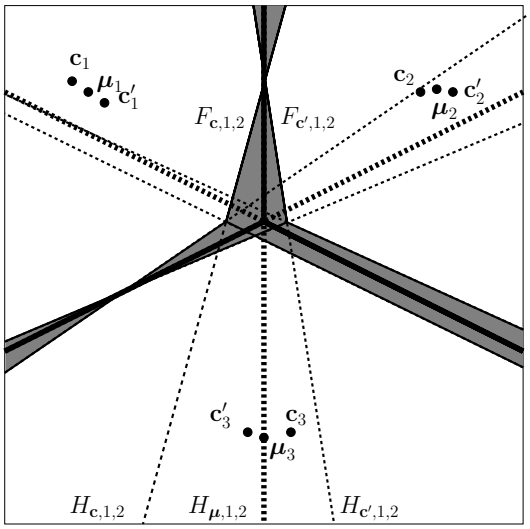

Figure 1: An illustrative drawing of the setting and notation used. Thicker lines represent the optimal k-means clustering partition (for $k = 3$ clusters) with respect to the underlying distribution. Clustering two independent random samples gives us two random centroid sets $\mathbf{c}$ and $\mathbf{c}'$. These induce two different Voronoi partitions of $\mathbb{R}^n$, and the instability measure is essentially related to the probability mass in the area which switches between clusters, when we compare these two partitions (gray area).

Each such block is essentially the per-cluster covariance matrix, with respect to the optimal clustering.

3 Main Results

In this section, we present the main results of our paper, and discuss observations that might be drawn from it about the use of clustering stability in the k-means framework. We then more briefly examine certain practical considerations related to estimating the instability measure from empirical data. All the detailed proofs are presented in Sec. 4.

3.1 Central Result and Discussion of Consequences

Our central result is the following theorem, which characterizes the exact value to which $\widehat{\mathrm{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D}, m)$ converges for any appropriate underlying distribution $\mathcal{D}$.

Theorem 1. *In the setting assumed above, we have that* $\widehat{\mathrm{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D}, m)$ *converges in probability (as* $m \to \infty$*) to:*

$$\frac{4}{\sqrt{\pi}} \sum_{1 \leq i < j \leq k} \left[\left(\int_{C_{\boldsymbol{\mu},i} \cup C_{\boldsymbol{\mu},j}} p(\mathbf{x}) d\mathbf{x} \right) \times \left(\int_{F_{\boldsymbol{\mu},i,j}} p(\mathbf{x}) \frac{\Psi(\mathbf{x}, i, j)}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|} d\mathbf{x} \right) \right], \quad (1)$$

where $\Psi(\mathbf{x}, i, j)$ *is defined as*

$$\left\| \begin{pmatrix} V_i^{1/2} & 0 \\ 0 & V_j^{1/2} \end{pmatrix} \begin{pmatrix} (\Gamma^{-1})_{i,i} & (\Gamma^{-1})_{i,j} \\ (\Gamma^{-1})_{j,i} & (\Gamma^{-1})_{j,j} \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}_i - \mathbf{x} \\ \mathbf{x} - \boldsymbol{\mu}_j \end{pmatrix} \right\|$$

(the existence of the matrix square roots and inverses follows from V and Γ being positive semidefinite and positive definite respectively). Moreover, $\sqrt{m}\ d_{\mathcal{D}}(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2))$ converges in distribution to that of:

$$2\sqrt{2} \sum_{1 \le i < j \le k} \left[\left(\int_{C_{\boldsymbol{\mu},i} \cup C_{\boldsymbol{\mu},j}} p(\mathbf{x}) d\mathbf{x} \right) \times \left(\int_{F_{\boldsymbol{\mu},i,j}} p(\mathbf{x}) \frac{\left| \begin{pmatrix} \boldsymbol{\mu}_i - \mathbf{x} \\ \mathbf{x} - \boldsymbol{\mu}_j \end{pmatrix}^\top \begin{pmatrix} \mathbf{c}_i - \boldsymbol{\mu}_i \\ \mathbf{c}_j - \boldsymbol{\mu}_j \end{pmatrix} \right|}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|} d\mathbf{x} \right) \right], \quad (2)$$

where

$$\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_k)^\top \sim \mathcal{N}(\boldsymbol{\mu}, \Gamma^{-1} V \Gamma^{-1}).$$

According to Thm. 1, for any distribution satisfying the necessary conditions, the instability measure (which is scaled by $\sqrt{m}$) converges to a constant value, which depends on the underlying distribution and the number of clusters k. This more or less solves the problem posed in [1] for k-means, and validates the conjecture posed in [16]. Namely, clustering stability remains in principle a meaningful model selection criterion, no matter how large the sample is and how close the unscaled instability measures are to 0.

In addition, the result is not very sensitive to the exact definition of the clustering instability measure. For example, $d_{\mathcal{D}}(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2))$ could have been defined as the probability mass which switches between clustering (under some optimal assignment of each cluster in $\mathtt{A_k}(S_1)$ to those of $\mathtt{A_k}(S_2)$), as was done for example in [8], leading to very similar formulas, for which the observations we make below still hold.

Thm. 1 provides a formula for the limit of the scaled stability measure. Although one can always calculate it for specific cases, it is of much more interest to draw general conclusions about the governing factors influencing its value. These factors essentially determine what is considered the 'correct' model, with a low value for the instability measure. Therefore, analyzing these factors can explain what assumptions correspond to the use of clustering stability, at least in the k-means setting. To avoid losing sight of the forest for the trees, our observations at this stage will be of a more qualitative rather than a rigorously quantitative nature.

First and foremost, the instability measure is affected by the integral of the probability density along the cluster boundaries. This is by far the most important factor influencing the instability measure. Indeed, when the density at the boundaries is exactly 0, we get an instability measure of 0. Although this density is multiplied by $\Psi(\mathbf{x}, i, j)$, note that $\Psi(\mathbf{x}, i, j)$ actually becomes 'nicer' when the boundary density is lower (since Γ^{-1} approaches a diagonal matrix with entries proportional to the inverse volumes of the clusters, hence having well-controlled eigenvalues assuming reasonably balanced clusters). Therefore, we can expect low instability even when the boundary density is low but not exactly 0.

This observation also means that higher values of k (the number of clusters) tend to lead to more instability, even if the density along the cluster boundaries remains the same. This is because we are integrating along all the boundaries, so many clusters imply a larger area over which we integrate. This gives a mathematically precise explanation to a well known experimental phenomenon, in which stability measures tend to have larger values for higher k, even in hierarchical clustering settings where more than one value of k is acceptable. When the 'correct' model has a boundary density which is on average several times lower than in competing models, this should overcome the effect of a larger or smaller integration area. However, when this is not the case, normalization procedures are called for, as in [8] (more on this later).

A secondary factor influencing the stability value is the behavior of $\Psi(\mathbf{x}, i, j)/\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$. This in turn is heavily influenced by the spectral radius of Γ, as we will explain shortly. More specifically, a large spectral radius of Γ, representing a 'locally shallow' optimal solution, results in more instability. It is interesting to note that shallow, ill-defined minima in terms of the objective function are often a sign of a mismatch between the model and the data. In fact, this observation is utilized by a different family of clustering stability methods, based on dataset perturbations ([4] for example). These methods in essence test the local robustness of the solution, which under smoothness assumptions can be related to the Hessian at that point. Moreover, the spectral radius of Γ is again related to the density along the cluster boundaries - a low density there implies a smaller spectral radius, hence lower instability.

To justify the assertions made above, we begin by pointing out that since we are integrating over the boundary between clusters i, j, then $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$ acts as a normalization term, eliminating the dependence on the norm of $\mathbf{x}$ (or the Euclidean distance between the boundary and the cluster centroids). Hence, the behavior of this factor depends mostly on the algebraic properties of sub-matrices of the Hessian inverse Γ^{-1}, and the matrix square root of V. An exact analysis is problematic in the general case, but a useful general characterization is the spectrum of these two matrices. If all the eigenvalues of Γ^{-1}, V are 'large', then we would expect $\Psi(\mathbf{x}, i, j)/\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$ to be relatively large as well, leading to a higher value for the stability measure. On the other hand, small eigenvalues will lead to lower values for the instability measure. By inspecting the formula for Γ, and assuming all clusters have equal sizes, we see that the diagonal elements of Γ are at most $2/k$, and can become smaller if the density along the boundary points is larger. Since the main diagonal majorizes the spectrum of the symmetric matrix Γ (cf. [5]), smaller elements in the main diagonal roughly imply smaller eigenvalues for Γ, hence larger eigenvalues for Γ^{-1}, and thus, in general, greater instability.

3.2 Examples

To illustrate some of the conclusions from the previous subsection, we empirically evaluated the instability measure on a few toy examples, where everything is well controlled and easy to analyze. The results are displayed in Fig. 2.

First of all, simple inspection shows that all the instability measures tend to converge to a constant value, which differs based on the choice of the model order k. The three leftmost plots clearly demonstrate how the densities along

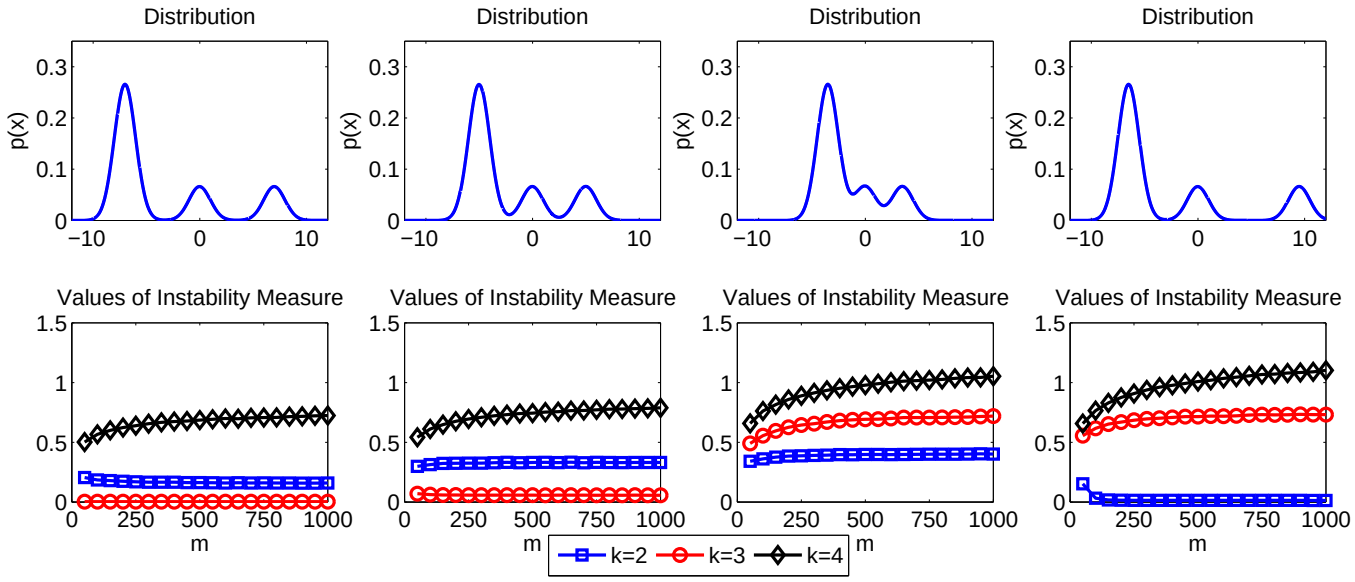

Figure 2: Illustrative examples of the behavior of clustering stability. In each column, the upper plot is the underlying distribution we sample from (a mixture of unit variance Gaussians on $\mathbb{R}$), while the lower plot is an empirically simulated value of the instability measure, for different sample sizes m.

the cluster boundaries play a crucial role in determining the instability measure. For each distribution, $k = 3$ emerges as the most stable model, since the boundaries between the clusters with $k = 3$ have low density. However, $k = 3$ becomes less stable as the Gaussians get closer to each other, leading to higher densities in the borders between them. At some point, the density at the single border point of $k = 2$ becomes smaller than the sum of the densities at the two border points of $k = 3$, and as a result $k = 2$ becomes more stable than $k = 3$.

A different manifestation of this behavior can be seen in the rightmost plot, which simulates a hierarchical clustering setting. In this case, all three Gaussians are separated, but one of them is relatively more separated than the other two. As before, $k = 4$ is less stable than $k = 3$ and $k = 2$, but now $k = 2$ is the most stable model. This is primarily because the sum of the border densities in $k = 3$ is larger than the density at the border point for $k = 2$. Deciding on $k = 2$ as the number of clusters in the data is not unreasonable (recall that clustering stability makes no explicit generative assumption on what the clusters look like), but it does indicate that using clustering stability for hierarchical clustering, when one is interested in discovering more than the topmost level in the hierarchy, must be approached with caution.

3.3 Bootstrapping in Clustering Stability

In [16], a simple and näive estimator for clustering stability, $\hat{\theta}_{k,4m}$, is used, defined as follows: Given a sample of size $4m$, split it randomly into 3 disjoint subsets S_1,S_2,S_3 of size m,m and $2m$ respectively. Estimate $d_{\mathcal{D}}(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2))$ by computing

$$\frac{1}{\sqrt{m}} \sum_{x_i, x_{m+i} \in S_3} \mathbf{1}\Big(\mathtt{A_k}(S_1)(x_i, x_{m+i}) \neq \mathtt{A_k}(S_2)(x_i, x_{m+i})\Big),$$

where $(x_1, .., x_m)$ is a random permutation of S_3, and return this value as an estimate of $\widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D}, m)$. In other words, compared to the definition of $\widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D}, m)$, we replace the expectation over instance pairs by an empirical set of m pairs, and replace the expectation over sample pairs by a single pair. Needless to say, in practice one will not use a single realization of subsample pairs, but rather average over a larger number of subsample pairs, using bootstrapping techniques. Such estimation processes should yield much better results. Nevertheless, as in most bootstrapping techniques, the number of splits is usually too small and involves too much reuse of the data to be formally justified by concentration of measure arguments.

In [16], it is shown that for a certain setting, even the näive estimator $\hat{\theta}_{k,4m}$ is enough to detect the most stable model order with high probability. Our goal is to explore whether this can be extended to the general k-means setting, at least for large enough samples where we can assume that the instability measure is well approximated by the results in Thm. 1.

For simplicity, let us focus on two fixed values k_t and k_f of k, such that the ratio between $\widehat{\text{instab}}(\mathtt{A}_{k_f}, \mathcal{D}, m)$ and $\widehat{\text{instab}}(\mathtt{A}_{k_t}, \mathcal{D}, m)$ is some reasonably large constant (one can think of it as a 'true' model corresponding to k_t, vs. a 'false' model corresponding to k_f). A single subsample pair can give us a single realization of the random variable $d_{\mathcal{D}}(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2))$. This value depends on the probability mass which switches between clusterings, and hence must again be estimated empirically from the sample at hand. If it chanced that the realization of $d_{\mathcal{D}}(\mathtt{A}_{k_t}(S_1), \mathtt{A}_{k_t}(S_2))$ and $d_{\mathcal{D}}(\mathtt{A}_{k_f}(S_1), \mathtt{A}_{k_f}(S_2))$ differ by a constant factor independent of the sample size, then using Thm. 2 in [16], we are assured that the smaller of the two can be determined empir-

ically with exponentially high probability. However, all we know is that the *expectations* of $d_{\mathcal{D}}(\mathtt{A}_{k_t}(S_1), \mathtt{A}_{k_t}(S_2))$ and $d_{\mathcal{D}}(\mathtt{A}_{k_f}(S_1), \mathtt{A}_{k_f}(S_2))$ (namely, the instability measures) differ by a constant factor. Depending on the distribution of $d_{\mathcal{D}}$, this does not preclude the possibility that most of the time, the ratio between specific realizations will not reflect the ratio between the expectations. However, using the results of Thm. 1 on the distribution of $d_{\mathcal{D}}(\mathtt{A}_{k_f}(S_1), \mathtt{A}_{k_f}(S_2))$, we can prove that this is never the case. Specifically, we can show the following:

Theorem 2. *For any distribution $\mathcal{D}$ satisfying the conditions of Thm. 1, assume that for some two values k_t, k_f, the ratio of $\widehat{instab}(\mathtt{A}_{k_f}, \mathcal{D}, m)$ and $\widehat{instab}(\mathtt{A}_{k_t}\mathcal{D}, m)$ converges to $\infty > R > 3$. Then for large enough m, we have that:*

$$\Pr\left(\hat{\theta}_{k_t,4m} \geq \hat{\theta}_{k_f,4m}\right) \leq \frac{0.3 + 3\log(R)}{R}.$$

We note that our proof actually produces an entire range of bounds, which provides a trade off for the minimality requirement on R with the tightness in terms of the constants. See the proof for further details. Also, if $\widehat{\text{instab}}(\mathtt{A}_{k_t}\mathcal{D}, m)$ is zero, while $\widehat{\text{instab}}(\mathtt{A}_{k_t}\mathcal{D}, m) > 0$ (corresponding to $R = \infty$, which might happen if the density along the cluster boundaries is exactly zero for k_t), it is easy to show that the probability of detecting k_t as the most stable model is essentially 1.

This theorem implies we can use just a single pair of subsamples to estimate the instability measures for k_t and k_f, and will still detect that k_t is more stable than k_f with a probability which scales almost linearly with the ratio between the true instability measures. Of course, repeating this procedure several times using bootstrapping techniques will only improve the probability of a correct detection.

3.4 Convergence Rates

After establishing the asymptotic value of clustering stability for k-means clustering, a reasonable next step is exploring what kind of guarantees can be made on the convergence rate of our instability measure to this asymptotic limit. In other words, given a sample of given size, what is the deviation between our stability measure, and its asymptotic value as specified in Thm. 1? As a first step, we establish the following negative result, which demonstrates that without additional assumptions, no universal guarantees can be given on the convergence rate. The theorem refers to the case $k = 3$, but the proof idea can easily be extended to other values of k.

Theorem 3. *For any positive numbers M and m_0, there exists a distribution $\mathcal{D}$ such that $\widehat{instab}(\mathtt{A}_3, \mathcal{D}, m) \to 0$ as $m \to \infty$, but $\widehat{instab}(\mathtt{A}_3, \mathcal{D}, m) \geq M$ for some $m \geq m_0$.*

The theorem does not imply that the asymptotic convergence rate is arbitrarily bad. In fact, a complicated second-order analysis (omitted from this paper due to lack of space), seems to indicate a uniform power-law convergence rate for any distribution satisfying the conditions of Thm. 1, as well as a few other conditions such as Lipschitz-continuity and bounded third moment. However, the exact constants in this power law can be arbitrarily bad, depending on various characteristics of the distribution. Finding sufficient and empirically verifiable conditions which provide finite sample guarantees is therefore of much interest.

4 Proofs

4.1 Proof of Thm. 1

Before embarking on the proof itself, we briefly sketch the idea behind the proof:

1. Using a result due to Pollard [14], we can characterize the asymptotic Gaussian distribution of the cluster centroids $\mathbf{c}$, in terms of the underlying distribution $\mathcal{D}$ (Lemma 1).

2. The cluster boundaries are determined by the positions of the centroids. Hence, we can derive the asymptotic distribution of these boundaries. In particular, for every boundary $F_{\mathbf{c},i,j}$, we characterize the asymptotic distribution of the pointwise Euclidean distance between two realizations of this boundary, over drawing and clustering two independent samples. This distance is defined relative to a projection on the hyperplane $H_{\boldsymbol{\mu},i,j}$ (Lemma 2).

3. We show that the probability mass of $\mathcal{D}$, which switches between clusters i and j over the two independent clusterings, has an asymptotic distribution definable by an integral involving the distance function above, and the values of $p(\cdot)$ on $F_{\boldsymbol{\mu},i,j}$ (Lemma 3 and Lemma 4).

4. This allows us to characterize the asymptotic distribution of $d_{\mathcal{D}}(\mathtt{A}_{\mathtt{k}}(S_1), \mathtt{A}_{\mathtt{k}}(S_2))$, and hence the asymptotic value of $\widehat{\text{instab}}(\mathtt{A}_{\mathtt{k}}, \mathcal{D}, m)$ (Thm. 1).

Some of the steps above are not direct, but are rather probabilistic approximations, which hold with arbitrarily high probability for arbitrarily high precision as $m \to \infty$. This kind of convergence in probability implies convergence in distribution. Since convergence in distribution is transitive and holds under continuous mappings (cf. [6]), we indeed get the distribution to which $\sqrt{m}d_{\mathcal{D}}(\mathtt{A}_{\mathtt{k}}(S_1), \mathtt{A}_{\mathtt{k}}(S_2))$ converges; hence its expected value is the one to which our instability measure, $\widehat{\text{instab}}(\mathtt{A}_{\mathtt{k}}, \mathcal{D}, m)$, converges. To prevent ambiguity due to asymptotic convergence to 0 for most quantities we are dealing with, we scale them by $\sqrt{m}$ (the 'correct' scale factor in this case, as will shortly become apparent).

Lemma 1. *Under the notation and assumptions above, $\sqrt{m}(\mathbf{c} - \boldsymbol{\mu})$ converges in distribution to $\mathcal{N}\left(\mathbf{0}, \Gamma^{-1}V\Gamma^{-1}\right)$.*

This lemma is an immediate consequence of the main theorem in [14].

According to Lemma 1, $\|\mathbf{c} - \boldsymbol{\mu}\|$ converges in probability to 0 as $m \to \infty$. This and the asymptotically Gaussian distribution allows us to assume that for large enough values of m, with arbitrarily high probability and for any $i, j \in [k], i \neq j$, the nearest centroid to $\boldsymbol{\mu}_i$ is $\mathbf{c}_i$, and $F_{\mathbf{c},i,j}$ is non-orthogonal to $F_{\boldsymbol{\mu},i,j}$ with probability 1.

Lemma 2. *For some $i,j \in [k], i \neq j$, assume that $F_{\boldsymbol{\mu},i,j} \neq \emptyset$. For any $\mathbf{x} \in H_{\boldsymbol{\mu},i,j}$, define the function:*

$$\ell(\mathbf{x},\mathbf{c}_i,\mathbf{c}_j) = \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| \left(\frac{\mathbf{c}_i+\mathbf{c}_j}{2} - \mathbf{x}\right)\cdot(\mathbf{c}_i - \mathbf{c}_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\cdot(\mathbf{c}_i - \mathbf{c}_j)}.$$

Then for any fixed $\mathbf{x}$, $\sqrt{m}\ell(\mathbf{x},\mathbf{c}_i,\mathbf{c}_j)$ converges in distribution to that of

$$\frac{\sqrt{m}}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}\begin{pmatrix}\boldsymbol{\mu}_i - \mathbf{x}\\ \mathbf{x} - \boldsymbol{\mu}_j\end{pmatrix}^\top \begin{pmatrix}\mathbf{c}_i - \boldsymbol{\mu}_i\\ \mathbf{c}_j - \boldsymbol{\mu}_j\end{pmatrix}$$

Considering the projection of $H_{\mathbf{c},i,j}$ to $H_{\boldsymbol{\mu},i,j}$, we have that $\ell(\mathbf{x},\mathbf{c}_i,\mathbf{c}_j)$ is the signed Euclidean distance of $\mathbf{x}$ from the point on $H_{\mathbf{c},i,j}$ which projects to it (see the left half of Fig. 3). This is because $\ell(\mathbf{x},\mathbf{c}_i,\mathbf{c}_j)$ must satisfy the equation:

$$\left(\left(\mathbf{x} + \ell(\mathbf{x},\mathbf{c}_i,\mathbf{c}_j)\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}\right) - \frac{\mathbf{c}_i+\mathbf{c}_j}{2}\right)\cdot(\mathbf{c}_i - \mathbf{c}_j) = 0,$$

Proof. We have that

$$\begin{aligned}
&\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|\ell(\mathbf{x},\mathbf{c}_i,\mathbf{c}_j)\\
&= \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2\left(\frac{\mathbf{c}_i+\mathbf{c}_j}{2} - \mathbf{x}\right)\cdot(\mathbf{c}_i - \mathbf{c}_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\cdot(\mathbf{c}_i - \mathbf{c}_j)}\\
&\approx \left(\frac{\mathbf{c}_i+\mathbf{c}_j}{2} - \mathbf{x}\right)\cdot(\mathbf{c}_i - \mathbf{c}_j).
\end{aligned}$$

The asymptotically exact approximation follows from the asymptotic convergence in probability of $\mathbf{c}_i, \mathbf{c}_j$ to $\boldsymbol{\mu}_i, \boldsymbol{\mu}_j$ respectively. Therefore the denominator is approximately $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2$ for large enough values of m, and the terms cancel out.

We now calculate the asymptotic distribution of $\left(\frac{\mathbf{c}_i+\mathbf{c}_j}{2} - \mathbf{x}\right)\cdot(\mathbf{c}_i - \mathbf{c}_j)$ for some fixed $\mathbf{x}$. Let $\mathbf{c}_i = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i$ and $\mathbf{c}_j = \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_j$. We have that:

$$\begin{aligned}
&\left(\frac{\mathbf{c}_i+\mathbf{c}_j}{2} - \mathbf{x}\right)\cdot(\mathbf{c}_i - \mathbf{c}_j)\\
&= \left(\frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_j+\boldsymbol{\epsilon}_i+\boldsymbol{\epsilon}_j}{2} - \mathbf{x}\right)\cdot\big((\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j)\big)\\
&= \left(\frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_j}{2} - \mathbf{x}\right)\cdot(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\\
&\quad + \left(\frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_j}{2} - \mathbf{x}\right)\cdot(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j)\\
&\quad + \left(\frac{\boldsymbol{\epsilon}_i+\boldsymbol{\epsilon}_j}{2}\right)\cdot(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{\|\boldsymbol{\epsilon}_i\|^2 - \|\boldsymbol{\epsilon}_j\|^2}{2}
\end{aligned}$$

Notice that the first summand is exactly 0 (by definition of $\mathbf{x}$ as lying on $F_{\boldsymbol{\mu},i,j}$), while the last summand is asymptotically negligible compared to the other summands (since $\boldsymbol{\epsilon}_i$ has a Gaussian distribution with variance $\Theta(1/m)$). As a result, we can approximate the expression above as:

$$\begin{aligned}
&\approx \left(\frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_j}{2} - \mathbf{x}\right)\cdot(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j) + \left(\frac{\boldsymbol{\epsilon}_i+\boldsymbol{\epsilon}_j}{2}\right)\cdot(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\\
&= (\boldsymbol{\mu}_i - \mathbf{x})\cdot\boldsymbol{\epsilon}_i - (\boldsymbol{\mu}_j - \mathbf{x})\cdot\boldsymbol{\epsilon}_j\\
&= (\boldsymbol{\mu}_i - \mathbf{x})\cdot(\mathbf{c}_i - \boldsymbol{\mu}_i) - (\boldsymbol{\mu}_j - \mathbf{x})\cdot(\mathbf{c}_j - \boldsymbol{\mu}_j).
\end{aligned}$$

The result in the lemma follows. □

In order to calculate the asymptotic distribution of $d_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$, we want to characterize the distribution of the probability mass of $\mathcal{D}$ in the 'wedges' created between two boundaries for clusters i,j, based on two independent samples (see right half of Fig. 3). For any two given boundaries, calculating the probability mass requires integration of the underlying density function $p(\cdot)$ over these wedges, making it very hard to explicitly derive the distribution of this probability mass. The purpose of the next two lemmas is to derive a more tractable approximation which is asymptotically exact, and depends only on the the values of $p(\cdot)$ along the $F_{\boldsymbol{\mu},i,j}$. This asymptotic approximation holds for each $F_{\boldsymbol{\mu},i,j}$ separately, but since there is only a fixed finite number of such boundaries, we have that this approximation asymptotically holds simultaneously for all the boundaries. We begin with an auxiliary lemma, required for the main Lemma 4 which follows.

Lemma 3. *For some $H_{\boldsymbol{\mu},i,j}$, fix some (possibly unbounded) polytope $F \subseteq H_{\boldsymbol{\mu},i,j}$. For notational convenience, assume $H_{\boldsymbol{\mu},i,j}$ is aligned with the axes, in the sense that for all $\mathbf{x} \in H_{\boldsymbol{\mu},i,j}$, its last coordinate is 0 (it can be easily shown that the conditions on $p(\cdot)$ would still hold). Also, denote $F' = \{\mathbf{y} \in \mathbb{R}^{n-1} : (\mathbf{y},0) \in F\}$, and for any $\mathbf{y} \in F'$, denote $\ell'(\mathbf{y}) = \ell((\mathbf{y},0),\mathbf{c}_i,\mathbf{c}_j)$. Let $\ell'_1(\mathbf{y})$ and $\ell'_2(\mathbf{y})$ be two independent copies of $\ell'(\mathbf{y})$ (which is a random variable over the random draw of the sample S). Then we have that as $m \to \infty$,*

$$\begin{aligned}
\sqrt{m}\Bigg|\int_{F'}\Big|\int_{\ell'_1(\mathbf{y})}^{\ell'_2(\mathbf{y})} p(\mathbf{y},x)dx\Big|d\mathbf{y}\\
- \int_{F'}\Big|\int_{\ell'_1(\mathbf{y})}^{\ell'_2(\mathbf{y})} p(\mathbf{y},0)dx\Big|d\mathbf{y}\Bigg| \xrightarrow{P} 0. \quad (3)
\end{aligned}$$

Proof. By the integral mean value theorem, since $p(\cdot)$ is continuous, we have that Eq. (3) is equal to:

$$\begin{aligned}
\sqrt{m}\Bigg|\int_{F'}\Big[&\big|\ell'_2(\mathbf{y}) - \ell'_1(\mathbf{y})\big|p(\mathbf{y},\xi_\mathbf{y})\\
&-\big|\ell'_2(\mathbf{y}) - \ell'_1(\mathbf{y})\big|p(\mathbf{y},0)\Big]d\mathbf{y}\Bigg|,
\end{aligned}$$

where $\xi_\mathbf{y} \in [\ell'_1(\mathbf{y}),\ell'_2(\mathbf{y})]$ (or $[\ell'_2(\mathbf{y}),\ell'_1(\mathbf{y})]$ if $\ell'_1(\mathbf{y}) > \ell'_2(\mathbf{y})$, but for notational convenience we stick with this formulation). This is upper bounded in term by:

$$\sqrt{m}\int_{F'}(|\ell'_1(\mathbf{y})| + |\ell'_2(\mathbf{y})|)\sup_{\xi\in[\ell'_1(\mathbf{y}),\ell'_2(\mathbf{y})]}|p(\mathbf{y},\xi) - p(\mathbf{y},0)|d\mathbf{y},$$

assuming the integral exists. Since $\ell'_1(\mathbf{y})$ and $\ell'_2(\mathbf{y})$ have the same distribution, then it is enough to show existence and convergence to zero in probability for:

$$\sqrt{m}\int_{F'}|\ell'_1(\mathbf{y})|\sup_{\xi\in[\ell'_1(\mathbf{y}),\ell'_2(\mathbf{y})]}|p(\mathbf{y},\xi) - p(\mathbf{y},0)|d\mathbf{y}. \quad (4)$$

Using Lemma 2, we know that for any fixed $\mathbf{y}$, $\sqrt{m}\ell'_1(\mathbf{y})$ is normally distributed, and its variance is

$$\begin{aligned}
&\frac{1}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\begin{pmatrix}\boldsymbol{\mu}_i - (\mathbf{y},0)\\ (\mathbf{y},0) - \boldsymbol{\mu}_j\end{pmatrix}^\top\\
&\times\begin{pmatrix}(\Gamma^{-1}V\Gamma^{-1})_{i,i} & (\Gamma^{-1}V\Gamma^{-1})_{i,j}\\ (\Gamma^{-1}V\Gamma^{-1})_{j,i} & (\Gamma^{-1}V\Gamma^{-1})_{j,j}\end{pmatrix}\begin{pmatrix}\boldsymbol{\mu}_i - (\mathbf{y},0)\\ (\mathbf{y},0) - \boldsymbol{\mu}_j\end{pmatrix}
\end{aligned}$$

(see corollary 3.3.3 in [18]). Therefore, it is possible to upper bound the standard deviation of $\sqrt{m}\ell_1'(\mathbf{y})$ by $\sqrt{a\|\mathbf{y}\|^2+b}$ for some positive constants a, b. From this, the probability that $\sqrt{m}|\ell_1'(\mathbf{y})| \leq c\sqrt{a\|\mathbf{y}\|^2+b}$ for some positive constant c converges exponentially fast (with c) to 1. Note that since ℓ_1' is a deterministic function of $\mathbf{y}$, this probabilistic bound holds uniformly for all $\mathbf{y}$. The same argument holds for ℓ_2'. As a result, for any desired confidence parameter $\delta > 0$, there is a positive constant c_δ such that Eq. (4) is upper bounded with a probability of at least $1-\delta$ by:

$$\int_{F'} c_\delta \sqrt{a\|\mathbf{y}\|^2+b} \times \sup_{|\xi| \leq c_\delta \sqrt{a\|\mathbf{y}\|^2+b}/\sqrt{m}} |p(\mathbf{y},\xi) - p(\mathbf{y},0)| d\mathbf{y}. \quad (5)$$

By continuity of $p(\cdot)$, the function over which we integrate is continuous in $\mathbf{y}$ and hence locally integrable. If we replace $1/\sqrt{m}$ by a non-negative continuous parameter d, it is easily seen that the existence of this integral for some value of d implies its existence for any smaller value of d, and the integral is a continuous function of d. Also, this integral exists and is exactly 0 when $d = 0$. Therefore, it suffices to show that the integral exists for *some* value of d (or $1/\sqrt{m}$), and continuity arguments ensure that the expression converges to 0 as $d \to 0$ (namely $m \to \infty$). In particular, choose $m = \lceil c_\delta^2 a \rceil$. Then it is enough to show the existence of:

$$\int_{F'} c_\delta \sqrt{a\|\mathbf{y}\|^2+b} \sup_{|\xi| \leq \sqrt{\|\mathbf{y}\|^2+b/a}} |p(\mathbf{y},\xi) - p(\mathbf{y},0)| d\mathbf{y}. \quad (6)$$

The restriction of this integral for any bounded subset of F' (or all of F' if it is bounded) obviously exists. Otherwise, define some large enough ball B of radius R around the origin, such that for any $(\mathbf{y},0) \in B^c$, there exists a constant d such that $c_\delta\sqrt{a\|\mathbf{y}\|^2+b} \leq d\|\mathbf{y}\|$. We have that the restriction of the integral in Eq. (6) to B exists. Outside B, we have that

$$\begin{aligned}
&\int_{\mathbf{y}\in F', (\mathbf{y},0)\in B^c} c_\delta \sqrt{a\|\mathbf{y}\|^2+b} \sup_{|\xi| \leq \sqrt{\|\mathbf{y}\|^2+b/a}} |p(\mathbf{y},\xi) - p(\mathbf{y},0)| d\mathbf{y} \\
&\leq d \int_{\mathbf{y}\in F', (\mathbf{y},0)\in B^c} \|\mathbf{y}\| \sup_{\xi\in\mathbb{R}} |p(\mathbf{y},\xi) - p(\mathbf{y},0)| d\mathbf{y} \\
&\leq d \int_{\mathbf{y}\in F', (\mathbf{y},0)\in B^c} \|\mathbf{y}\| \sup_{\xi\in\mathbb{R}} (p(\mathbf{y},\xi) + p(\mathbf{y},0)) d\mathbf{y} \\
&\leq 2d \int_{\mathbf{y}\in F', (\mathbf{y},0)\in B^c} \|\mathbf{y}\| g(\|\mathbf{y}\|) \\
&\leq 2d \int_{r=R}^{\infty} r g(r) * e r^{n-1},
\end{aligned}$$

where $g(\|\mathbf{y}\|)$ is the dominating function on $p(\cdot)$ assumed to exist (see section 2), and e is the surface area of a $n-2$ sphere. The existence of the last integral follows from the definition of $g(\cdot)$, and it can be made arbitrarily small by selecting the radius R of the ball B to be large enough.

What we have essentially shown is that for any fixed confidence parameter $\delta > 0$ and for any m, the expression in Eq. (3) is upper bounded with a probability at least $1-\delta$ by Eq. (5), where the constant c_δ depends on δ. Eq. (5) converges to 0 as $m \to \infty$. Therefore, for any fixed $\epsilon, \delta > 0$, we can choose m large enough so that with a probability of at least $1-\delta$, the expression in Eq. (3) is smaller than ϵ. As a result, Eq. (3) converges in probability to 0 as required. □

Lemma 4. *For some non-empty $F_{\boldsymbol{\mu},i,j}$, let $t(\mathbf{c},\mathbf{c}',i,j)$ be a random variable over $\mathcal{D}^{2m}$, defined as the probability mass of $\mathcal{D}$ which switches between clusters i,j with respect to the two clusterings defined by $\mathbf{c},\mathbf{c}'$, induced by independently sampling and clustering a pair of samples S_1, S_2 each of size m. More formally, define the set-valued random variable*

$$\begin{aligned}
B(\mathbf{c},\mathbf{c}',i,j) = \{\mathbf{x} \in \mathbb{R}^n : (\mathbf{x} \in C_{\mathbf{c},i} \wedge \mathbf{x} \in C_{\mathbf{c}',j}) \\
\vee (\mathbf{x} \in C_{\mathbf{c}',i} \wedge \mathbf{x} \in C_{\mathbf{c},j})\} \cup F_{\mathbf{c},i,j} \cup F_{\mathbf{c}',i,j},
\end{aligned}$$

so that

$$t(\mathbf{c},\mathbf{c}',i,j) = \int_{B(\mathbf{c},\mathbf{c}',i,j)} p(\mathbf{x}) d\mathbf{x}. \quad (7)$$

Then $\sqrt{m}t(\mathbf{c},\mathbf{c}',i,j)$ converges in distribution to

$$\sqrt{m} \int_{\mathbf{x}\in F_{\boldsymbol{\mu},i,j}} p(\mathbf{x}) |l(\mathbf{x},\boldsymbol{\mu}_i,\boldsymbol{\mu}_j)| d\mathbf{x},$$

where $\sqrt{m}l(\mathbf{x},\boldsymbol{\mu}_i,\boldsymbol{\mu}_j)$ is distributed as

$$\frac{\sqrt{2m}}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|} \begin{pmatrix} \boldsymbol{\mu}_i - \mathbf{x} \\ \mathbf{x} - \boldsymbol{\mu}_j \end{pmatrix}^\top \begin{pmatrix} \mathbf{c}_i - \boldsymbol{\mu}_i \\ \mathbf{c}_j - \boldsymbol{\mu}_j \end{pmatrix}.$$

Proof. See the right half of Fig. 3 to help clarify the notation and the intuition of the following proof. Intuitively, the probability mass which switches between clusters i and j over the two samples is the probability mass of $\mathcal{D}$ lying 'between' $F_{\mathbf{c},i,j}$ and $F_{\mathbf{c}',i,j}$. The first problem is that this probability mass is also affected by the positions of other neighboring boundaries. However, the fluctuations of these additional boundaries decrease as $m \to \infty$, and their effect on the probability mass in question becomes negligible, as we show below. Our goal is to upper and lower bound the integral in Eq. (7) (scaled by $\sqrt{m}$) by approximations which converge in probability to each other.

Define $F_{\max}(\boldsymbol{\mu},\mathbf{c},\mathbf{c}',i,j) \in H_{\boldsymbol{\mu},i,j}$ as the projection of $B(\mathbf{c},\mathbf{c}',i,j)$ on $H_{\boldsymbol{\mu},i,j}$. Define $F_{\min}(\boldsymbol{\mu},\mathbf{c},\mathbf{c}',i,j)$ as the projection of a subset of $B(\mathbf{c},\mathbf{c}',i,j)$, such that for any point $\mathbf{x}$ in it, the width of $B(\mathbf{c},\mathbf{c}',i,j)$ relative to $H_{\boldsymbol{\mu},i,j}$ is equal to $|\ell_2'(\mathbf{y}) - \ell_1'(\mathbf{y})|$. For notational convenience, we drop most of the parameters from now on. As in Lemma 3, we assume that H is aligned with the axes, such that for any $\mathbf{x} \in H$, its last coordinate is 0. Let $F'_{\max}$ and $F'_{\min}$ be the $n-1$ dimensional projections of $F_{\max}$ and $F_{\min}$, by removing this last zero coordinate. Then we have that:

$$\begin{aligned}
&\int_{F'_{\max}} \left| \int_{\ell_1'(\mathbf{y})}^{\ell_2'(\mathbf{y})} p(\mathbf{y},x) dx \right| d\mathbf{y} \geq \int_{B(\mathbf{c},\mathbf{c}',i,j)} p(\mathbf{z}) d\mathbf{z} \\
&\geq \int_{F'_{\min}} \left| \int_{\ell_1'(\mathbf{y})}^{\ell_2'(\mathbf{y})} p(\mathbf{y},x) dx \right| d\mathbf{y}. \quad (8)
\end{aligned}$$

By Lemma 3, we have that

$$\sqrt{m}\left|\int_{F'_{\max}}\Big|\int_{\ell'_1(\mathbf{y})}^{\ell'_2(\mathbf{y})}p(\mathbf{y},x)dx\Big|d\mathbf{y}-\int_{F'_{\max}}|\ell'_2(\mathbf{y})-\ell'_1(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}\right|\xrightarrow{P}0 \quad (9)$$

and

$$\sqrt{m}\left|\int_{F'_{\min}}\Big|\int_{\ell'_1(\mathbf{y})}^{\ell'_2(\mathbf{y})}p(\mathbf{y},x)dx\Big|d\mathbf{y}-\int_{F'_{\min}}|\ell'_2(\mathbf{y})-\ell'_1(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}\right|\xrightarrow{P}0. \quad (10)$$

Also, we can show that as $m\to\infty$,

$$\sqrt{m}\left|\int_{F'_{\max}}|\ell'_2(\mathbf{y})-\ell'_1(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}-\int_{F'}|\ell'_2(\mathbf{y})-\ell'_1(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}\right|\xrightarrow{P}0 \quad (11)$$

and

$$\sqrt{m}\left|\int_{F'}|\ell'_2(\mathbf{y})-\ell'_1(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}-\int_{F'_{\min}}|\ell'_2(\mathbf{y})-\ell'_1(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}\right|\xrightarrow{P}0. \quad (12)$$

We explain the proof for Eq. (11), the second is proven in a similar manner. We have that:

$$\begin{aligned}&\sqrt{m}\left|\int_{F'_{\max}}|\ell'_2(\mathbf{y})-\ell'_1(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}-\int_{F'}|\ell'_2(\mathbf{y})-\ell'_1(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}\right|\\&\le\int_{F'_{\max}\triangle F'}\sqrt{m}|\ell'_2(\mathbf{y})-\ell'_1(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}\\&=\int_{F'_{\max}\triangle F'}\sqrt{m}(|\ell'_2(\mathbf{y})|+|\ell'_1(\mathbf{y})|)p(\mathbf{y},0)d\mathbf{y}.\end{aligned}$$

The argument is very similar to the proof of Lemma 3, and we refer the reader to it for the specifics. Based on the results of Lemma 2, for any desired confidence parameter δ, it holds with a probability of at least $1-\delta$, that $\sqrt{m}|\ell'_1(\mathbf{y})|$ and $\sqrt{m}|\ell'_2(\mathbf{y})|$ are upper bounded (uniformly for all $\mathbf{y}$) by $c_\delta\sqrt{a\|\mathbf{y}\|^2+b}$, where a and b are fixed positive constants, and c_δ is a parameter dependent on δ. Repeating the arguments of Lemma 3, it follows that the integral above exists for any possible realization of $F_{\max}\triangle F$, and since the $\sqrt{m}$ terms cancel out, the integrand is dependent on m only via c_δ, which is fixed for a given confidence parameter δ. When $\mathbf{c}=\mathbf{c}'=\boldsymbol{\mu}$, we have that $F'_{\max}=F'$, and the integral above is 0. Since the integral is a continuous function of $\mathbf{c},\mathbf{c}'$, and they converge in probability to $\boldsymbol{\mu}$, we have that as $m\to\infty$, the integral converges in probability to zero.

Combining Eq. (8),Eq. (9),Eq. (10),Eq. (11),Eq. (12), we have that $\sqrt{m}\,t(\mathbf{c},\mathbf{c}',i,j)$ converges in distribution to that of:

$$\sqrt{m}\int_{F_{\boldsymbol{\mu},i,j}}p(\mathbf{x})|\ell(\mathbf{x},\mathbf{c}_i,\mathbf{c}_j)-\ell(\mathbf{x},\mathbf{c}'_i,\mathbf{c}'_j)|d\mathbf{x}. \quad (13)$$

It now remains to explicitly give the asymptotic distribution of (Eq. (13)). The distribution of $\sqrt{m}\,\ell(\mathbf{x},\mathbf{c}_i,\mathbf{c}_j)$ and $\sqrt{m}\,\ell(\mathbf{x},\mathbf{c}'_i,\mathbf{c}'_j)$ are known from Lemma 2, and they are independent Gaussians (since S_1 and S_2 were drawn independently). Using corollary 3.3.4 in [18], we have that $\sqrt{m}(\ell(\mathbf{x},\mathbf{c}_i,\mathbf{c}_j)-\ell(\mathbf{x},\mathbf{c}'_i,\mathbf{c}'_j))$ converges in distribution to that of:

$$\frac{\sqrt{2m}}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}\begin{pmatrix}\boldsymbol{\mu}_i-\mathbf{x}\\ \mathbf{x}-\boldsymbol{\mu}_j\end{pmatrix}^\top\begin{pmatrix}\mathbf{c}_i-\boldsymbol{\mu}_i\\ \mathbf{c}_j-\boldsymbol{\mu}_j\end{pmatrix} \quad (14)$$

The expression inside the absolute value in (13) above can be seen as a deterministic continuous function of $\ell(\mathbf{x},\mathbf{c}_i,\mathbf{c}_j)-\ell(\mathbf{x},\mathbf{c}'_i,\mathbf{c}'_j)$, which converges to the distribution defined in (14). Therefore, the expression in (13) converges in distribution to:

$$\sqrt{m}\int_{F_{\boldsymbol{\mu},i,j}}p(\mathbf{x})|l(\mathbf{x},\boldsymbol{\mu}_i,\boldsymbol{\mu}_j)|d\mathbf{x},$$

where $\sqrt{m}l(\mathbf{x},\boldsymbol{\mu}_i,\boldsymbol{\mu}_j)$ is distributed as in Eq. (14). □

We now turn to proving Thm. 1. Let $t(\mathbf{c},\mathbf{c}',i,j)$ be as defined in Lemma 4. Let $\widehat{C}_{\mathbf{c},\mathbf{c}',i}$ denote the set of points in $\mathbb{R}^n$ which remain in the same cluster i for both clusterings defined by $\mathbf{c},\mathbf{c}'$. Then the probability that a pair of instance drawn from $\mathcal{D}$ are in the same cluster under one clustering, and in different clusters under a second clustering, is distributed as

$$2\sum_{1\le i<j\le k}\left(\int_{\widehat{C}_{\mathbf{c},\mathbf{c}',i}\cup\widehat{C}_{\mathbf{c},\mathbf{c}',j}}p(\mathbf{x})d\mathbf{x}\right)t(\mathbf{c},\mathbf{c}',i,j).$$

In our setting, for any $i\in[k]$, we have that as $m\to\infty$,

$$\left(\int_{\widehat{C}_{\mathbf{c},\mathbf{c}',i}}p(\mathbf{x})d\mathbf{x}\right)\xrightarrow{P}\left(\int_{C_{\boldsymbol{\mu},i}}p(\mathbf{x})d\mathbf{x}\right),$$

and therefore the distribution above, times $\sqrt{m}$ converges to:

$$2\sqrt{m}\sum_{1\le i<j\le k}\left(\int_{C_{\boldsymbol{\mu},i}\cup C_{\boldsymbol{\mu},j}}p(\mathbf{x})d\mathbf{x}\right)t(\mathbf{c},\mathbf{c}',i,j).$$

In Lemma 4, we calculated the distribution to which $\sqrt{m}t(\mathbf{c},\mathbf{c}',i,j)$ converges. Plugging this result into the expression above, we get the desired distribution to which $d_{\mathcal{D}}(\mathtt{A}_k(S_1),\mathtt{A}_k(S_2))$ converges.

For notational convenience, let $\Sigma=\Gamma^{-1}V\Gamma^{-1}$, and

$$\varphi(\mathbf{x},i,j)=\begin{pmatrix}\boldsymbol{\mu}_i-\mathbf{x}\\ \mathbf{x}-\boldsymbol{\mu}_j\end{pmatrix}^\top\begin{pmatrix}\Sigma_{i,i}&\Sigma_{i,j}\\ \Sigma_{j,i}&\Sigma_{j,j}\end{pmatrix}\begin{pmatrix}\boldsymbol{\mu}_i-\mathbf{x}\\ \mathbf{x}-\boldsymbol{\mu}_j\end{pmatrix}.$$

Since $\widehat{\mathtt{instab}}(\mathtt{A}_k,\mathcal{D},m)=\sqrt{m}\mathbb{E}d_{\mathcal{D}}(\mathtt{A}_k(S_1),\mathtt{A}_k(S_2))$, and $d_{\mathcal{D}}(\mathtt{A}_k(S_1),\mathtt{A}_k(S_2))$ is bounded, we get that $\widehat{\mathtt{instab}}(\mathtt{A}_k,\mathcal{D},m)$ converges to $\sqrt{m}$ times the expectation of the distribution to

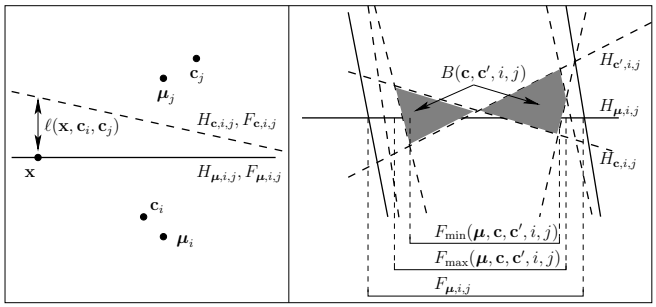

Figure 3: An illustrative drawing of some of the notation and geometrical constructs used in the proof of Thm. 1. Solid lines represent cluster boundaries with respect to the optimal cluster centroids $\boldsymbol{\mu}$, while dashed lines represent cluster boundaries with respect to cluster centroids $\mathbf{c}$ or $\mathbf{c}'$ returned by the clustering algorithm based on an empirical sample. See the text for more details.

which
$d_{\mathcal{D}}(\mathtt{A}_k(S_1), \mathtt{A}_k(S_2))$ converges. This is equal to:

$$\sqrt{m}\mathbb{E}\left[2\sum_{1\le i<j\le k}\left(\int_{C_{\boldsymbol{\mu},i}\cup C_{\boldsymbol{\mu},j}} p(\mathbf{x})d\mathbf{x}\right) \times\left(\int_{F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})\left|l(\mathbf{x},\boldsymbol{\mu}_i,\boldsymbol{\mu}_j)\right|d\mathbf{x}\right)\right]. \quad (15)$$

By Fubini's theorem, this is equal to:

$$2\sqrt{m}\left[\sum_{1\le i<j\le k}\left(\int_{C_{\boldsymbol{\mu},i}\cup C_{\boldsymbol{\mu},j}} p(\mathbf{x})d\mathbf{x}\right) \times\left(\int_{F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})\mathbb{E}\left|l(\mathbf{x},\boldsymbol{\mu}_i,\boldsymbol{\mu}_j)\right|d\mathbf{x}\right)\right]. \quad (16)$$

We now use the fact that for a univariate normally distributed variable $a \sim \mathcal{N}(\mu, \sigma^2)$, it holds that
$\mathbb{E}[|a|] = \sigma\sqrt{2/\pi}$. As a result, we get the expression

$$\frac{4}{\sqrt{\pi}}\sum_{1\le i<j\le k}\left[\left(\int_{C_{\boldsymbol{\mu},i}\cup C_{\boldsymbol{\mu},j}} p(\mathbf{x})d\mathbf{x}\right) \times\left(\int_{F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})\frac{\sqrt{\varphi(\mathbf{x},i,j)}}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}d\mathbf{x}\right)\right]. \quad (17)$$

Finally, we can write this in a simpler form. Recall that $\Sigma = \Gamma^{-1}V\Gamma^{-1}$. Both V and Γ are positive semidefinite, hence Σ can be rewritten as $(V^{1/2}\Gamma^{-1})^\top V^{1/2}\Gamma^{-1}$. Substituting into the expression for $\varphi(\mathbf{x}, i, j)$ gives us the required result.

4.2 Proof of Thm. 2

The proof is composed of several lemmas. The key insight is that the distribution of $d_{\mathcal{D}}(\mathtt{A}_k(S_1), \mathtt{A}_k(S_2))$ can be viewed as a certain non-standard norm of a Gaussian random vector. Using results on Gaussian measures in general Banach spaces allows us to uniformly bound the probability of a large deviation from the expectation.

Lemma 5. *The distribution of $d_{\mathcal{D}}(\mathtt{A}_k(S_1), \mathtt{A}_k(S_2))$ is equal to that of $\|\mathbf{v}\|_*$, where $\mathbf{v} \sim \mathcal{N}(0, \Gamma^{-1}V\Gamma^{-1})$ and $\|\mathbf{v}\|_*$ is a norm on $\mathbb{R}^{nk}$.*

Proof. Denote $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ where $\mathbf{v}_i \in \mathbb{R}^n$ for any i. By Thm. 1, the distribution of $d_{\mathcal{D}}(\mathtt{A}_k(S_1), \mathtt{A}_k(S_2))$ can be written as:

$$\sum_{1\le i<j\le k} a_{i,j}\int_{F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})\left|\begin{pmatrix}\boldsymbol{\mu}_i-\mathbf{x}\\ \mathbf{x}-\boldsymbol{\mu}_j\end{pmatrix}^\top\begin{pmatrix}\mathbf{v}_i\\ \mathbf{v}_j\end{pmatrix}\right|d\mathbf{x}, \quad (18)$$

where $\mathbf{v}$ is as defined in the lemma, and $a_{i,j}$ are certain positive constants dependent on $\mathcal{D}$. Perhaps unexpectedly, it turns out that this expression defines a norm on $\mathbf{v}$: linearity and the triangle inequality are easy to show. Also, Eq. (18) is always non-negative. Finally, Eq. (18) is zero if and only if $\mathbf{v} = \mathbf{0}$. One direction is trivial. For the other direction, note that $p(\cdot)$ must be strictly positive for some non-degenerate subset of some cluster boundary, in order that $\widehat{\text{instab}}(\mathtt{A}_k, \mathcal{D}, m)$ be positive for large enough m. Therefore, if $\mathbf{v} \neq 0$ then Eq. (18) is larger than $\mathbf{0}$. □

Lemma 6. *Let $\mathbf{x}$ be a normally distributed random vector in $\mathbb{R}^n$, let $\|\cdot\|_*$ be any norm on $\mathbb{R}^n$, and let $\theta \in (1/2, 1)$ be a free parameter. Introduce the following two parameters which depend on θ:*

$$a_\theta = 1 + \frac{2(1-\theta)}{\log\left(\frac{\theta}{1-\theta}\right)}$$

$$b_\theta = 1 - \theta + \frac{1-\exp(-(erf^{-1}(\theta))^2)}{\sqrt{\pi}erf^{-1}(\theta)}.$$

Then for any M, ϵ such that $Mb_\theta > 1$ and $\epsilon a_\theta < 1$, it holds that

$$\Pr(\|\mathbf{x}\|_* > M\mathbb{E}\|\mathbf{x}\|_*) \le \theta\left(\frac{1-\theta}{\theta}\right)^{(1+Mb_\theta)/2},$$

and

$$\Pr(\|\mathbf{x}\|_* \le \epsilon\mathbb{E}\|\mathbf{x}\|_*) \le erf(erf^{-1}(\theta)a_\theta\epsilon).$$

Proof. It is easily seen that $\|\mathbf{x}\|_*$ is continuously distributed, assuming $\mathbf{x}$ has positive variance. For any $\theta \in (1/2, 1)$, let med_θ be a nonnegative number which satisfies:

$$\Pr(\|\mathbf{x}\|_* \le \text{med}_\theta) = \theta.$$

Using two results from the literature (theorem III.3 in [12], and theorem 1 from [9]), we have that for any $M \ge 1$, and for any $\epsilon \in [0, 1]$, it holds that:

$$\Pr(\|\mathbf{x}\|_* > M\text{med}_\theta) \le \theta\left(\frac{1-\theta}{\theta}\right)^{(1+M)/2} \quad (19)$$

$$\Pr(\|\mathbf{x}\|_* \le \epsilon\text{med}_\theta) \le \text{erf}(\text{erf}^{-1}(\theta)\epsilon). \quad (20)$$

It remains to convert these bounds on the deviation from med_θ to the deviation from $\mathbb{E}\|\mathbf{x}\|_*$. To achieve this, we need

to upper and lower bound $\mathbb{E}\|\mathbf{x}\|_*/\text{med}_\theta$. By substitution of variables, we have that

$$\mathbb{E}\|\mathbf{x}\|_* = \int_0^\infty \Pr(\|\mathbf{x}\|_* > t)dt$$
$$= \text{med}_\theta \int_0^\infty \Pr(\|\mathbf{x}\|_* > M\text{med}_\theta)dM,$$

Using Eq. (19), this can be upper bounded by

$$\text{med}_\theta \left(1 + \int_1^\infty \theta \left(\frac{1-\theta}{\theta}\right)^{(1+M)/2} dM\right),$$

which after straightforward computations leads to $\mathbb{E}[\|\mathbf{x}\|_*] \leq \text{med}_\theta a_\theta$, where a_θ is as defined in the lemma.

In a similar manner, we can write:

$$\mathbb{E}\|\mathbf{x}\|_* = \int_0^\infty 1 - \Pr(\|\mathbf{x}\|_* \leq t)dt$$
$$= \text{med}_\theta \int_0^\infty 1 - \Pr(\|\mathbf{x}\|_* \leq \epsilon\text{med}_\theta)d\epsilon,$$

which is lower bounded in term, using Eq. (20), by

$$\text{med}_\theta \int_0^1 1 - \text{erf}(\text{erf}^{-1}(\theta)\epsilon)d\epsilon$$

Again by straightforward computations, we reach the conclusion that $\mathbb{E}\|\mathbf{x}\|_* \geq \text{med}_\theta b_\theta$, where b_θ is as defined in the lemma.

Therefore, we have that if $Mb_\theta > 1$, then

$$\Pr(\|\mathbf{x}\|_* > M\mathbb{E}\|\mathbf{x}\|_*) \leq \Pr(\|\mathbf{x}\|_* > Mb_\theta\text{med}_\theta)$$
$$\leq \theta \left(\frac{1-\theta}{\theta}\right)^{(1+Mb_\theta)/2}.$$

The other bound in the lemma is derived similarly. □

We can now turn to the proof of Thm. 2. The combination of Lemma 5 and Lemma 6 allows us to upper bound the probability that $d_\mathcal{D}\left(\mathtt{A}_{k_f}(S1), \mathtt{A}_{k_f}(S2)\right)$ is smaller than its expectation (namely $\widehat{\text{instab}}(\mathtt{A}_{k_f}, \mathcal{D}, m)$) by a factor $\epsilon < 1$, and upper bound the probability that $d_\mathcal{D}\left(\mathtt{A}_{k_t}(S1), \mathtt{A}_{k_t}(S2)\right)$ is larger than its expectation (namely $\widehat{\text{instab}}(\mathtt{A}_{k_t}, \mathcal{D}, m)$) by some factor $M > 1$, provided that ϵ, M satisfy the conditions specified in Lemma 6. By a union bound argument, if we choose appropriate M and ϵ so that $M/\epsilon \leq R$, where R is the ratio between the stability measures, we get that

$$\Pr(d_\mathcal{D}\left(\mathtt{A}_{k_t}(S1), \mathtt{A}_{k_t}(S2)\right) \geq d_\mathcal{D}\left(\mathtt{A}_{k_f}(S1), \mathtt{A}_{k_f}(S2)\right)$$

is upper bounded by

$$\theta \left(\frac{1-\theta_1}{\theta_1}\right)^{(1+Mb_{\theta_1})/2} + \text{erf}(\text{erf}^{-1}(\theta_2)a_{\theta_2}\epsilon),$$

for any $\theta_1, \theta_2 \in (1/2, 1)$. Choosing different values for them (as well as the choice of appropriate M, ϵ) leads to different bounds, with a trade off between the tightness of the bound in terms of the constants, and minimality requirements on R (which stem from the requirements on M, ϵ by Lemma 6). The bound in the theorem is derived by choosing $\theta_1 = 0.9$, $\theta_2 = 0.8$, $M = 2\log(R)/(b_\theta \log(\theta_1/(1-\theta_1)))$, and $\epsilon = M/R$.

4.3 Proof of Thm. 3

To prove the theorem, we will borrow a setting discussed in [11] for a different purpose.

Let Δ be some small positive constant (say $\Delta < 0.1$). Consider the parameterized family of distributions $\{D_\epsilon\}$ (where $\epsilon \in (0, 1/4)$) on the real line, which assigns probability mass $(1-\epsilon)/4$ to $x = -1$ and $x = -1-\Delta$, and $(1+\epsilon)/4$ to $x = 1$ and $x = 1+\Delta$. Any such distribution satisfies the requirements of Thm. 1, except continuity. However, as discussed previously, the theorem only requires continuity in some region around the boundary points, so we may ignore this difficulty. Alternatively, we may introduce continuity by convolution with a small local smoothing operator. For any ϵ, it is easily seen that $l_{\mathcal{D}_\epsilon} = 0$, since the boundary points between the optimal clusters have zero density.

Let $A^1_{m,\epsilon}$ denote the event where for a sample of size m drawn i.i.d from $\mathcal{D}_\epsilon$, there are more instances on $\{-1-\Delta, -1\}$ than on $\{1, 1+\Delta\}$. Also, let $A^2_{m,\epsilon}$ denote the event that for a sample of size m drawn i.i.d from $\mathcal{D}_\epsilon$, there are more instances on $\{1, 1+\Delta\}$ than on $\{-1-\Delta, -1\}$. Finally, let $B_{m,\epsilon}$ denote the event that every point in $\{-1-\Delta, -1, 1, 1+\Delta\}$ receives at least one instance from the sample. Clearly, if $A^1_{m,\epsilon} \cap B_{m,\epsilon}$ occurs, then the optimal cluster centers for the sample are $\{-1-\Delta, -1, 1+\Delta'\}$ for some $\Delta' \in [0, \Delta]$, and if $A^2_{m,\epsilon} \cap B_{m,\epsilon}$ occurs, then the optimal cluster centers for the sample are $\{-1-\Delta', 1, 1+\Delta\}$ for some $\Delta' \in [0, \Delta]$.

By Thm. 2.1 in [17], for any Bernoulli random variable X such that $\mathbb{E}[X] = p \leq 1/2$, and any whole number k such that $k/m \leq 1-p$, if $X_1, \ldots, X_m$ are m i.i.d copies of X, then

$$\Pr\left(\frac{1}{m}\sum_{i=1}^m X_i \geq \frac{k}{m}\right) \geq 1 - \Phi\left(\sqrt{\frac{m}{p(1-p)}}\left(\frac{k}{m} - p\right)\right),$$

where $\Phi(\cdot)$ is the cumulative normal distribution function. The probability of the event $A^1_{m,\epsilon}$ is equal to the probability of a success rate of more than half in m Bernoulli trials, whose probability of success is $(1-\epsilon)/2$. Using the theorem above, we get after a few straightforward algebraic manipulations and relaxations that

$$\Pr(A^1_{m,\epsilon}) \geq 1 - \Phi\left(\frac{4}{\sqrt{m}} + 2\epsilon\sqrt{m}\right). \quad (21)$$

The probability of the event $A^2_{m,\epsilon}$ is equal to the probability of a success rate of less than half in m Bernoulli trials, whose probability of success is $(1-\epsilon)/2$. By a standard normal approximation argument, we have that for large enough values of m, and for any $\epsilon \in (0, 1/4)$, it holds that

$$\Pr(A^2_{m,\epsilon}) \geq 1/2. \quad (22)$$

Finally, it is straightforward to show that $\Pr(B_{m,\epsilon})$ can be made arbitrarily close to 1 uniformly for any ϵ, by choosing a large enough m. Combining this with Eq. (21), Eq. (22) and the easily proven formula $\Pr(A \cap B) \geq \Pr(A) - \Pr(B^\complement)$ for any two events A, B, we get that by choosing a large enough sample size $m' > m_0$, and an appropriate value ϵ', then

$$\Pr(A^1_{m',\epsilon'} \cap B_{m',\epsilon'}), \Pr(A^2_{m',\epsilon'} \cap B_{m',\epsilon'}) \geq 1/2 - \nu$$

for an arbitrarily small $\nu > 0$. For that choice of m', ϵ', if we draw and cluster two independent samples S_1, S_2 of size m from $\mathcal{D}_{\epsilon'}$, then the probability that event $A^1_{m',\epsilon'} \cap B_{m',\epsilon'}$ occurs for one sample, and $A^2_{m',\epsilon'} \cap B_{m',\epsilon'}$ occurs for the second sample, is at least $2(1/2 - \nu)^2$, or at least $1/3$ for a small enough ν. Note that in this case, we get the two different clusterings discussed above, and

$$d_{\mathcal{D}_{\epsilon'}}(\mathtt{A}_3(\mathtt{S}_1), \mathtt{A}_3(\mathtt{S}_2)) = \frac{1+\epsilon^2}{2} > \frac{1}{2}.$$

So with a probability of at least $1/3$ over drawing and clustering two independent samples, the distance between the clusterings is more than $1/2$. Using our definitions, this implies

$$\widehat{\mathrm{instab}}(\mathtt{A}_3, \mathcal{D}_{\epsilon'}, m') > \frac{\sqrt{m'}}{6},$$

where the r.h.s can be made larger than M by choosing m' large enough.

5 Discussion

In this paper, we analyzed the behavior of clustering stability in the k-means framework. We were able to explicitly characterize its asymptotic behavior, concluded that it is indeed meaningful even for arbitrarily large sample sizes (resolving the problem posed in [1] for k-means), and drew several interesting observation, among which is that clustering stability for large enough samples tends to choose models mainly based on their probability density along the cluster boundaries. Having a low density along the cluster boundaries appears to be a very reasonable requirement from a 'correct' model, and probably explains why clustering stability works in many situations. However, it also means that clustering stability will sometimes behave unexpectedly, for example in hierarchical clustering situations, as demonstrated in subsection 3.2. Several other aspects of clustering stability, such as the frequently employed bootstrapping techniques, were also discussed based on our new findings.

There are several directions for future research. The most obvious perhaps is achieving finite sample guarantees on the convergence of the instability measure to the value specified in Thm. 1, or at least when samples are large enough to determine the asymptotically most stable model with high probability. As demonstrated in Thm. 3, additional assumptions are needed for a convergence result. Also, it would be interesting to perform a similar analysis for other clustering methods beyond the k-means framework.

Finally, we would like to emphasize that our paper is not meant to advocate clustering stability as the 'best' method in any sense. Model selection in clustering is a very difficult problem, and no single approach is likely to succeed in all cases. However, we do believe that a better theoretical understanding of the assumptions inherent to different approaches can lead to a more informed selection of methods, from the bewildering choice currently offered in the literature. This paper is a small step in that direction.

Acknowledgements: The authors wish to thank Gideon Schechtman and Leonid Kontorovich for providing the necessary pointers for the proof of Thm. 2.

References

[1] Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, pages 5–19, 2006.

[2] Asa Ben-Hur, André Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, pages 6–17, 2002.

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

[4] Bittner et. al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406:536–540, August 2000.

[5] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[6] Alan F. Karr. *Probability*. Springer, 1993.

[7] A. Krieger and P. Green. A cautionary note on using internal cross validation to select the number of clusters. *Psychometrika*, 64(3):341–353, 1999.

[8] Tilman Lange, Volker Roth, Mikio L. Braun, and Joachim M. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 16(6):1299–1323, June 2004.

[9] RafałLatała and Krzysztof Oleszkiewicz. Gaussian measures of dilatations of convex symmetric sets. *Annals of Probability*, 27(4):1922–1938, 1999.

[10] Erel Levine and Eytan Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13(11):2573–2593, 2001.

[11] T. Linder. *Principles of nonparametric learning*, chapter 4: Learning-theoretic methods in vector quantization. Number 434 in CISM Courses and Lecture Notes (L. Gyorfi ed.). Springer-Verlag, New York, 2002.

[12] Vitali D. Milman and Gideon Schechtman. *Asymptotic Theory of Finite Dimensional Normed Spaces*. Springer, 1986.

[13] David Pollard. Personal communication.

[14] David Pollard. A central limit theorem for k-means clustering. *The Annals of Probability*, 10(4):919–926, November 1982.

[15] Peter Radchenko. *Asymptotics Under Nonstandard Conditions*. PhD thesis, Yale University, 2004.

[16] Ohad Shamir and Naftali Tishby. Cluster stability for finite samples. In *Advances in Neural Information Processing Systems 21*, 2007.

[17] E. V. Slud. Distribution inequalities for the binomial law. *The Annals of Probability*, 5(3):402–412, June 1977.

[18] Y.L. Tong. *The Multivariate Normal Distribution*. Springer, 1990.