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

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. Despite their relative success, not much is known theoretically on why or when do they work, or even what kind of assumptions they make in choosing an 'appropriate' model. Moreover, recent theoretical work has shown that they might 'break down' for large enough samples. 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 distribution to which suitably scaled measures of instability converge, based on a sample drawn from any distribution in $\mathbb{R}^n$ satisfying mild regularity conditions. From this, we can show that clustering stability does not 'break down' even for arbitrarily large samples, in the k-means framework that we study. Moreover, it allows us to identify the factors which influence the behavior of clustering stability for any sample size. This leads to some interesting preliminary observations about what kind of assumptions are made when using these methods. While often reasonable, these assumptions might also lead to unexpected consequences.

1 Introduction

The important and difficult problem of model selection in data clustering has been the focus of an 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.

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 clustering stability methods which compare the discrepancy or 'distance' between clusterings of different random subsets of our data. These methods seek a 'stable' model, in the sense that the value of such distance measures should tend to be small.

Although these methods have been shown to be rather effective in practice (cf. [2],[4],[7],[9]), 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 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, clustering stability might 'break down' for large enough samples, since we get approximately the same clustering hypothesis based on each random subsample, and thus achieve stability regardless of whether the model fits the data or not (this problem was also pointed out in [6]). A possible solution to this difficulty was proposed in [15]. In a nutshell, that paper showed that the important factor in the way these clustering stability methods work may not be the asymptotic stability of the model, but rather *how fast exactly does it converge to this stability*. With this more refined analysis, it was argued that differences in the stability of different models 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 toy example, as a proof-of-concept.

In this paper, we formally investigate the application of clustering stability to 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 analyze the behavior of the clustering distance measure, scaled by the square root of the sample size. Rather than converging to zero in probability as the sample size increases to infinity, this scaled measure converges to a non-degenerate distribution which depends on the choice of k. From this we can show that clustering stability does not 'break down' even for arbitrarily large samples, in the sense described earlier, at least for the k-means framework that we study.

The asymptotic distribution is also interesting for two additional reasons. The first is that it 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 fundamental 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 fundamental characteristics, which should be relevant for *any* sample size. As a result, the analysis leads to some preliminary observations about the factors influencing clustering stability in k-means, of 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. $\|\cdot\|$ will denote the Euclidean norm unless stated otherwise. $\mathcal{N}(\boldsymbol{\mu},\Sigma)$ denotes the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

We will use the stochastic order notation $O_p(\cdot)$ and $o_p(\cdot)$ (cf. [18]). Let $\{X_m\}$ and $\{Y_m\}$ be sequences of random vectors, defined on the same probability space. We write $X_m = O_p(Y_m)$ to mean that for each $\epsilon > 0$ there exists a real number M such that $\Pr(\|X_m\| \geq M\|Y_m\|) < \epsilon$ if m is large enough. We write $X_m = o_p(Y_m)$ to mean that $\Pr(\|X_m\| \geq \epsilon\|Y_m\|) \to 0$ for each $\epsilon > 0$. Notice that $\{Y_m\}$ may also be non-random. For example, $X_m = o_p(1)$ means that $X_m \to 0$ in probability.

Let $\mathcal{D}$ be a probability distribution on $\mathbb{R}^n$, with a bounded probability density function $p(\cdot)$ which is continuous as a function on $\mathbb{R}^n$. Assume that the following two regularity conditions hold:

- $\int_{\mathbb{R}^n} p(\mathbf{x})\|\mathbf{x}\|^2 d\mathbf{x} < \infty$ (in words, $\mathcal{D}$ has bounded variance).
- There exists a bounded, monotonically decreasing function $g(\cdot) : \mathbb{R} \to \mathbb{R}$, such that $p(\mathbf{x}) \leq g(\|\mathbf{x}\|)$ 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 [13] (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 in some neighborhood of the optimal cluster boundaries, but we will take this stronger assumption for simplicity.

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 $\mathbf{c} = (\mathbf{c}_1,\ldots,\mathbf{c}_k) \in \mathbb{R}^{nk}$, which are a global minimum of the objective function:

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

Let $\boldsymbol{\mu} = (\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k) \in \mathbb{R}^{nk}$ 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 $i \neq j$, $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$). To avoid ambiguities involving permutation of the centroids, we assume that the numbering of the centroids is by some uniform canonical ordering (for example, by sorting with respect to the coordinates).

For some set of centroids $\mathbf{c} = (\mathbf{c}_1,\ldots,\mathbf{c}_k)$, and for each cluster centroid $\mathbf{c}_i$, we denote 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) distance between two clusterings $\mathtt{A_k}(S_1)$ and $\mathtt{A_k}(S_2)$, where S_1, S_2 are samples of size m, is defined as:

$$d_{\mathcal{D}}^m(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2)) := \sqrt{m} \Pr_{\mathbf{x}_1,\mathbf{x}_2\sim\mathcal{D}} \big(\mathtt{A_k}(S_1)(\mathbf{x}_1,\mathbf{x}_2) \neq \mathtt{A_k}(S_2)(\mathbf{x}_1,\mathbf{x}_2)\big),$$

where $\mathtt{A_k}(S)(\mathbf{x}_1,\mathbf{x}_2)$ 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 definition follows that of [1] and [15], with the additional scaling by $\sqrt{m}$ (the 'correct' scaling factor as will become evident later on). A typical way to measure instability in practice is to cluster independent subsamples of the data, and empirically estimate the distance between the resulting clusterings. Thus, understanding the behavior of $d_{\mathcal{D}}^m(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2))$ (over drawing and clustering independent samples) is of much interest in analyzing the behavior of clustering stability.

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

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

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 the context, depending on the set over which we are integrating. For example, integrals over some $C_{\mathbf{c},i}$ are of the first type, while integrals over some $F_{\mathbf{c},i,j}$ are of the second type.

Let Γ be the $kn \times kn$ matrix, which is the Hessian of the mapping $W(\cdot)$ at the optimal solution $\boldsymbol{\mu}$. This matrix is composed of $k \times k$ blocks $\Gamma_{i,j}$ for $i, j \in [k]$. Each block $\Gamma_{i,j}$ can be shown to be equal to[1]

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

if $i = j$, and for $i \neq j$ it is defined as

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

We will use the same block notation later for its inverse Γ^{-1}. The existence of these integrals can be shown to follow from the assumptions on $p(\cdot)$. 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 [14]).

Let V be a $kn \times kn$ matrix, which represents (up to a constant) the covariance matrix of $\mathcal{D}$ with respect to each cluster, assuming the optimal clustering induced by $\boldsymbol{\mu}$. More specifically, V is composed of k diagonal blocks V_i of size $n \times n$ for $i \in [k]$ (all other elements of V are zero), where

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

We shall assume that $V_i \neq 0$ for any i.

3 Main Results

In this section, we present the main results of our paper, and discuss observations that might be drawn from them about the use of clustering stability in the k-means framework. All the detailed proofs are presented in Sec. 4.

3.1 Statement of Technical Results

Our main technical result is the following theorem, which characterizes the exact distribution to which $d^m_{\mathcal{D}}(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2))$ converges for any appropriate underlying distribution $\mathcal{D}$, and its expected value.

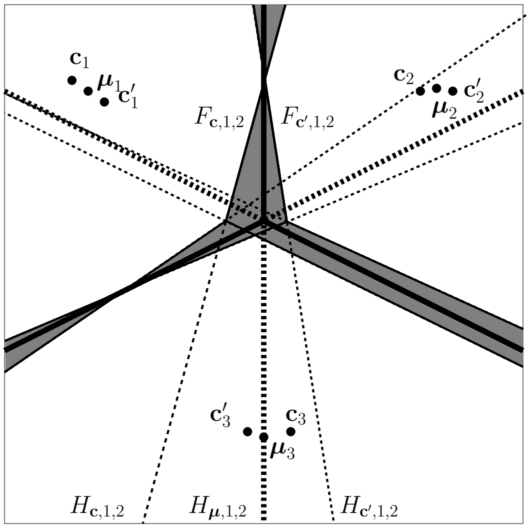

Figure 1: An illustrative drawing of the setting and notation used. Thicker lines represent the optimal k-means clustering partition (for $k = 3$ clusters) with respect to the underlying distribution. Clustering two independent random samples gives us two random centroid sets $\mathbf{c}$ and $\mathbf{c}'$.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

Theorem 1. *Assume $\mathcal{D}$ has a bounded probability density function $p(\cdot)$, which is continuous as a function on $\mathbb{R}^n$ and fulfills the two regularity conditions specified in Sec. 2. Let $\mathtt{A_k}$ be an algorithm which returns a global minimizer $\mathbf{c}$ of $\hat{W}(\cdot)$ for any k of interest, and assume that $\mathbf{c}$ converges in probability to some set of k distinct centroids $\boldsymbol{\mu}$ which are the unique global minimizer of $W(\cdot)$. Furthermore, assume that Γ is invertible and that $V_i \neq 0$ for any $i \in [k]$. Then we have that $d^m_{\mathcal{D}}(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2))$ converges in distribution to that of*

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

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

Denoting the expected value of this distribution as $\widehat{instab}(\mathtt{A_k}, \mathcal{D})$, we have that it is equal to

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

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

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

[1] This is proven in [13]. The definition of Γ there differs from ours in one of the signs, apparently due to a small error in that paper [12].

All the integrals can be shown to exist by the assumptions on $p(\cdot)$. It should be emphasized that $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$ is *not* necessarily the same as $\lim_{m\to\infty} \mathbb{E}d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$. This is because our convergence result does not necessarily imply convergence of expectations. Thus, formally speaking, the result above does not deal directly with the limit of $\mathbb{E}d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$, which has been used in [1],[15] as the theoretical definition of clustering stability. However, it turns out that for our purposes this is not too significant. It seems to be the asymptotic distribution and $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$, rather than the asymptotic expectation, which determine the asymptotic behavior of clustering stability.

The following theorem exemplifies this on a simple empirical estimator of clustering stability. The main difference between the following estimator and those proposed in the literature is that it measures the distance between just a single pair of clusterings from a pair of independent samples, rather than averaging over several pairs based on subsampling the data. This just makes our result stronger, because these kind of bootstrap procedures should only increase the reliability of the estimator, whereas here we are interested in a 'lower bound' on reliability.

Theorem 2. *Define a clustering stability estimator, $\hat{\theta}_{k,4m}$, 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 $d^m_\mathcal{D}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))/\sqrt{m}$ by computing*

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

where $(x_1,..,x_m)$ is a random permutation of S_3. For any distribution $\mathcal{D}$ satisfying the conditions of Thm. 1, assume that for some two values of k, $k_s \neq k_u$, the ratio of $\widehat{\text{instab}}(\mathtt{A}_{k_u}, \mathcal{D})$ and $\widehat{\text{instab}}(\mathtt{A}_{k_s}, \mathcal{D})$ (as defined in Thm. 1) is $\infty > R > 3$. Then we have that:

$$\Pr\left(\hat{\theta}_{k_s,4m} \geq \hat{\theta}_{k_u,4m}\right) \leq \frac{0.3 + 3\log(R)}{R} + o(1),$$

where the probability is over a sample of size $4m$ used for both estimators, and $o(1)$ converges to 0 as $m \to \infty$.

The theorem implies the following: Suppose we are considering two possible values for k, designated as k_s and k_u, such that the ratio between $\widehat{\text{instab}}(\mathtt{A}_{k_u}, \mathcal{D})$ and $\widehat{\text{instab}}(\mathtt{A}_{k_s}, \mathcal{D})$ is some reasonably large constant (one can think of it as a relatively unstable model corresponding to k_u, vs. a relatively stable model corresponding to k_s). Then the probability of *not* empirically detecting k_s as the most stable model has an upper bound which actually decreases with the sample size, converging to a constant value dependent on the ratio of $\widehat{\text{instab}}(\mathtt{A}_{k_s}, \mathcal{D})$ and $\widehat{\text{instab}}(\mathtt{A}_{k_u}, \mathcal{D})$. In this sense, according to the bound, clustering stability does not 'break down' in the large sample regime, and the asymptotic reliability of its empirical estimation is determined by $\widehat{\text{instab}}(\mathtt{A}_k, \mathcal{D})$. We emphasize that the theorem deals with the reliability of detecting the most stable model, not whether a stable model is really a 'good' model in any other sense.

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 $\widehat{\text{instab}}(\mathtt{A}_{k_s}, \mathcal{D}) = 0$, while $\widehat{\text{instab}}(\mathtt{A}_{k_u}, \mathcal{D}) > 0$ (corresponding to $R = \infty$), it is easy to show that the probability of detecting k_s as the most stable model converges to 1 as $m \to \infty$.

3.2 Factors Influencing Stability of Clustering Models

According to Thm. 1, for any distribution satisfying the necessary conditions, the distance between clusterings (after scaling by $\sqrt{m}$) converges to a generally non-degenerate distribution, which depends on the underlying distribution and the number of clusters k. As Thm. 2 shows, this implies that clustering stability does not 'break down' in the large sample regime, and its choice of the most 'appropriate' value of k seems to depend essentially on $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$.

Thm. 1 provides an explicit formula for $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$. Although one can always calculate it for specific cases, it is of much more interest to try and understand what are the governing factors influencing its value. These factors eventually determine what is considered by clustering stability as the 'correct' model, with a low value for $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$. Therefore, analyzing these factors can explain what sample-size-free assumptions correspond to the use of clustering stability, at least in the k-means setting that we study. Since a rigorous analysis is a complex endeavor in itself, we will limit ourselves to some preliminary and non-formal observations, which should be taken as such.

According to Thm. 1, the value of $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$ is asymptotically determined by three factors:

- The probability density along the cluster boundaries.
- The Hessian Γ of the objective function $W(\cdot)$ at $\boldsymbol{\mu}$.
- The variance V and mass of the clusters with respect to the underlying distribution.

A fourth factor appearing in the formula is $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$, but this can be seen simply as a normalization term, eliminating the dependence on the norm of $\mathbf{x}$.

The probability density along the cluster boundaries seems to play an interesting role. For example, when the density at the boundaries is exactly 0, we get that $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D}) = 0$. 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 of the mass of the clusters, hence having well-controlled eigenvalues assuming reasonably balanced clusters). Therefore, we might expect low instability even when the boundary density is low but not exactly 0.

As to the Hessian Γ, an exact analysis of its influence on $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$ is problematic in the general case, but a useful rough characterization is the spectrum of Γ. If all the eigenvalues of Γ^{-1} are 'large', then we might expect $\Psi(\mathbf{x}, i, j)/\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$ to be relatively large as well, leading to a higher value for $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$. On the other hand, small eigenvalues might lead to lower values of $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$. Thus, we see that a small spectral radius of the Hessian Γ, representing a 'locally shallow' optimal solution, may result 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, and therefore clustering stability seems to be doing a good thing on that regard.

When will the spectral radius of Γ be small, contributing to instability? 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]), it seems that a small spectral radius might correspond to larger values of k, as well as high density along the cluster boundaries. A similar analysis for V seems to indicate that high cluster variance increases instability as well.

These observation also imply that clustering instability might tend to be larger for higher values of k. As k becomes larger, $\widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})$ is the result of integrating over a larger area (all cluster boundaries), and the Hessian Γ might tend to have a smaller spectral radius, especially if the boundaries have high density. This is somewhat compensated in the formula by the mass and variance of each cluster becoming smaller, but these seem to scale down more slowly than the cluster boundaries area (and number) scaling up, especially in high dimensions. This matches a well known experimental phenomenon, in which clusterings tend to be less stable for higher k, even in hierarchical clustering settings where more than one value of k is acceptable. When the 'correct' model has a very low boundary density and nice structure compared to competing models, this might overcome the general tendency of instability to increase with k. However, when this is not the case, normalization procedures might be called for, as in [7].

3.3 Examples

To illustrate some of the observations from the previous subsection, we empirically evaluated the instability measure on a few simple toy examples, where everything is well controlled and easy to analyze. The results are displayed in Fig. 2. We emphasize that these are just simple illustrations of possible expected and unexpected characteristics of clustering stability in some very limited cases, which can be gleaned from the theoretical results above, and are not meant to represent more realistic or higher dimensional settings.

First of all, the average value of $d_{\mathcal{D}}^m(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2))$ tends to converge to a constant value, which differs based on the choice of the model order k, and clustering stability does not seem to 'break down' as sample size increases. The three leftmost plots demonstrate how, for these particular examples, the density along the cluster boundaries seem to play an important role in determining $\widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})$. 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 boundaries between them. At some point, when the density along the cluster boundaries for $k = 3$ becomes large enough, $k = 2$ 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 boundary densities in $k = 3$ is larger than the density at the boundary 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 how the clusters look like). However, it can indicate that in a hierarchical clustering setting, clustering stability might prefer high levels in the hierarchy, which may or may not be what we want.

3.4 Convergence Rates

After establishing the asymptotic distribution of the clustering distance measures for k-means clustering, a reasonable next step is exploring what kind of guarantees can be made on the convergence rate to this asymptotic limit. 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 integer m_0, there exists a distribution $\mathcal{D}$ such that $d_{\mathcal{D}}^m(\mathtt{A}_3(S_1), \mathtt{A}_3(S_2))$ converges in probability to 0 as $m \to \infty$, but $\Pr(d_{\mathcal{D}}^m(\mathtt{A}_3(S_1), \mathtt{A}_3(S_2)) > \sqrt{m}/4)$ is at least $1/3$ 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, we briefly sketch its outline:

1. Using the central limit theorem for k-means due to Pollard [13], 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).

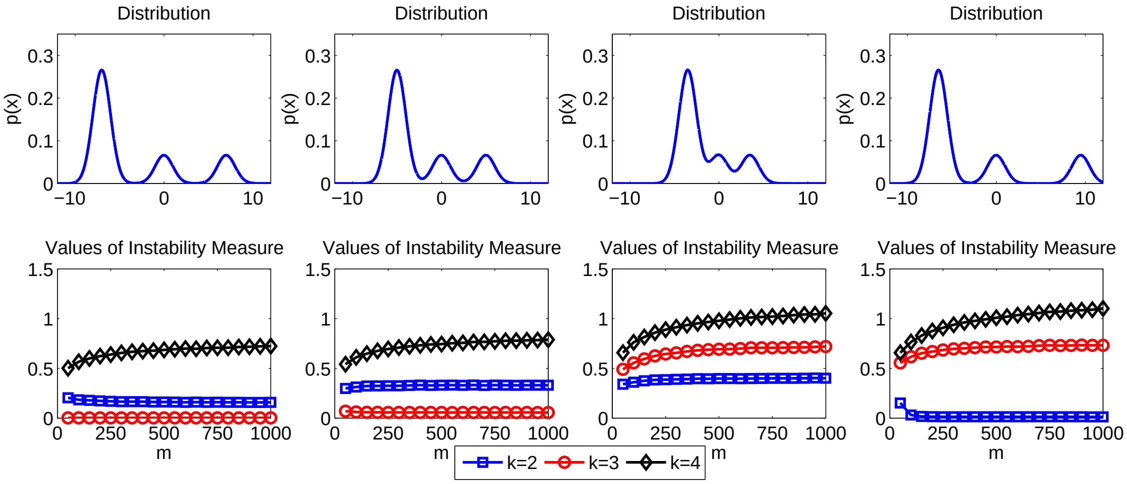

Figure 2: Illustrative examples of the behavior of clustering stability. In each column, the upper plot is the underlying distribution we sample from (a mixture of unit variance Gaussians on $\mathbb{R}$), while the lower plot is an empirical average of $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1),\mathtt{A}_\mathtt{k}(S_2))$ over 1000 trials, for different sample sizes m.

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). This allows us to formulate the asymptotic distribution of $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1),\mathtt{A}_\mathtt{k}(S_2))$, and its expected value.

For convenience, we shall use $\boldsymbol{\epsilon}=(\boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_k)$ to denote the random element $\mathbf{c}-\boldsymbol{\mu}$.

Lemma 1. *Under the notation and assumptions of the theorem, $\sqrt{m}\boldsymbol{\epsilon}=\sqrt{m}(\mathbf{c}-\boldsymbol{\mu})$ converges in distribution to $\mathbf{v}$, where $\mathbf{v}\sim\mathcal{N}\left(\mathbf{0},\Gamma^{-1}V\Gamma^{-1}\right)$. As a result, $\|\boldsymbol{\epsilon}\|=O_p(1/\sqrt{m})$.*

This lemma is a straightforward consequence of the main theorem in [13]. Notice that it 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 $\mathbf{c}_i$, all centroids are distinct, $F_{\mathbf{c},i,j}$ is non-orthogonal to $F_{\boldsymbol{\mu},i,j}$, and $\|\boldsymbol{\epsilon}\|$ is arbitrarily small. We shall tacitly use these assumptions in the remainder of the proof.

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 if $\|\boldsymbol{\epsilon}\|$ is smaller than some positive constant which depends only on $\boldsymbol{\mu}$, $\ell(\mathbf{x},\mathbf{c}_i,\mathbf{c}_j)$ can be rewritten as

$$\frac{1}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}\begin{pmatrix}\boldsymbol{\mu}_i-\mathbf{x}\\ \mathbf{x}-\boldsymbol{\mu}_j\end{pmatrix}^\top\begin{pmatrix}\boldsymbol{\epsilon}_i\\ \boldsymbol{\epsilon}_j\end{pmatrix}+O((\|\mathbf{x}\|+1)\|\boldsymbol{\epsilon}\|^2).$$

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 will separate the expression in the definition of $\ell(\mathbf{x},\mathbf{c}_i,\mathbf{c}_j)$ into 2 components and analyze them separately. We have that:

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

Notice that the first summand is exactly 0 (by definition of $\mathbf{x}$ as lying on $F_{\boldsymbol{\mu},i,j}$), and can therefore be dropped. After expanding and simplifying, we get that the above is equal to

$$(\boldsymbol{\mu}_i-\mathbf{x})\cdot\boldsymbol{\epsilon}_i-(\boldsymbol{\mu}_j-\mathbf{x})\cdot\boldsymbol{\epsilon}_j+O(\|\boldsymbol{\epsilon}\|^2)\qquad(1)$$

As to the second component in the definition of $\ell(\mathbf{x},\mathbf{c}_i,\mathbf{c}_j)$, we have that

$$\begin{aligned}&\frac{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)\cdot(\mathbf{c}_i-\mathbf{c}_j)}=\frac{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|^2+(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)\cdot(\boldsymbol{\epsilon}_i-\boldsymbol{\epsilon}_j)}\\&=\frac{1}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|\left(1+\frac{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)\cdot(\boldsymbol{\epsilon}_i-\boldsymbol{\epsilon}_j)}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|^2}\right)}\\&=\frac{1}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|\left(1+O(\|\boldsymbol{\epsilon}\|)\right)}\\&=\frac{1}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}\left(1-\frac{O(\|\boldsymbol{\epsilon}\|)}{1+O(\|\boldsymbol{\epsilon}\|)}\right)=\frac{1+O(\|\boldsymbol{\epsilon}\|)}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|},\qquad(2)\end{aligned}$$

assuming $\|\boldsymbol{\epsilon}\|$ to be small enough. Multiplying Eq. (1) and Eq. (2) gives us the expression in the lemma. □

In order to calculate the asymptotic distribution of $d^m_{\mathcal{D}}(\mathtt{A}_{\mathtt{k}}(S_1),\mathtt{A}_{\mathtt{k}}(S_2))$, we need 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 Fig. 1). 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 write the distribution of this probability mass explicitly. The purpose of the next two lemmas is to derive a more tractable, asymptotically exact approximation for each such wedge, which depends only on the values of $p(\cdot)$ along the boundary $F_{\boldsymbol{\mu},i,j}$.

We begin with an auxiliary lemma, required for the main Lemma 4 which follows. To state these lemmas, we will need some additional notation. For some $H_{\boldsymbol{\mu},i,j}$, fix some (possibly unbounded) polytope $F \subseteq H_{\boldsymbol{\mu},i,j}$. For notational convenience, we shall assume w.l.o.g that $H_{\boldsymbol{\mu},i,j}$ 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 regularity conditions on $p(\cdot)$ will still hold). Also, denote $F' = \{\mathbf{y} \in \mathbb{R}^{n-1} : (\mathbf{y},0) \in F\}$, which is simply the $n-1$ dimensional representation of F on the hyperplane. Finally, for ease of notation, denote $\ell((\mathbf{y},0),\mathbf{c}_i,\mathbf{c}_j)$ for any $\mathbf{y} \in F'$ as $\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})$, where $\boldsymbol{\epsilon} = \mathbf{c} - \boldsymbol{\mu}$.

Lemma 3. *Let $\boldsymbol{\epsilon},\boldsymbol{\epsilon}'$ be two independent copies of $\mathbf{c}-\boldsymbol{\mu}$, each induced by clustering an independent sample of size m. Let $B = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq R\}$ be a ball of radius R centered at the origin. Then we have that*

$$\left|\int_{F'\cap B}\Big|\int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})} p(\mathbf{y},\xi)d\xi\Big|d\mathbf{y} - \int_{F'\cap B}\Big|\int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})} p(\mathbf{y},0)d\xi\Big|d\mathbf{y}\right| = o_p(1/\sqrt{m}), \quad (3)$$

where the constants implicit in the r.h.s depend on R.

Proof. Since $p(\cdot)$ is a non-negative function, we can rewrite the expression in the lemma as

$$\left|\int_{F'\cap B}\int_{\min\{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})\}}^{\max\{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})\}} p(\mathbf{y},\xi)d\xi d\mathbf{y} - \int_{F'\cap B}\int_{\min\{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})}^{\max\{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})\}} p(\mathbf{y},0)d\xi d\mathbf{y}\right|,$$

or

$$\left|\int_{F'\cap B}\int_{\min\{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})\}}^{\max\{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})\}} p(\mathbf{y},\xi) - p(\mathbf{y},0)d\xi d\mathbf{y}\right|.$$

By the integral mean value theorem, since $p(\cdot)$ is continuous, we have that the expression above is equal to:

$$\left|\int_{F'\cap B}|\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y}) - \tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})|(p(\mathbf{y},\xi_{\mathbf{y}}) - p(\mathbf{y},0))d\mathbf{y}\right|,$$

where $\xi_{\mathbf{y}}$ is between the minimum and maximum of $\{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})\}$. For simplicity of notation, we will write $\xi_{\mathbf{y}} \in [\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})]$.

The expression above is upper bounded in turn by:

$$\int_{F'\cap B}(|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| + |\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})|)\sup_{\xi_{\mathbf{y}}\in[\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})]}|p(\mathbf{y},\xi_{\mathbf{y}}) - p(\mathbf{y},0)|d\mathbf{y},$$

assuming the integral exists. Since $\boldsymbol{\epsilon},\boldsymbol{\epsilon}'$ have the same distribution, it is enough to show existence and analyze the convergence to zero in probability for

$$\int_{F'\cap B}|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})|\sup_{\xi_{\mathbf{y}}\in[\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})]}|p(\mathbf{y},\xi_{\mathbf{y}}) - p(\mathbf{y},0)|d\mathbf{y}. \quad (4)$$

This integral can be upper bounded by

$$\sup_{\mathbf{y}\in F'\cap B}|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})|\sup_{\xi_{\mathbf{y}}\in[\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})]}|p(\mathbf{y},\xi_{\mathbf{y}}) - p(\mathbf{y},0)|\int_{F'\cap B}1d\mathbf{y}. \quad (5)$$

Since B is bounded, we have according to Lemma 2 that if $\|\boldsymbol{\epsilon}\|$ is small enough,

$$\sup_{\mathbf{y}\in F'\cap B}|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| = O(\|\boldsymbol{\epsilon}\| + \|\boldsymbol{\epsilon}\|^2), \quad (6)$$

and a similar equation holds for $\tilde{\ell}_{\boldsymbol{\epsilon}'}(\cdot)$ with $\boldsymbol{\epsilon}$ replaced by $\boldsymbol{\epsilon}'$ in the r.h.s. To make the equations less cumbersome, we will ignore the higher order term $\|\boldsymbol{\epsilon}\|^2$, since $\boldsymbol{\epsilon}$ converges to 0 in probability anyway by Lemma 1 (it is straightforward to verify that the analysis below still holds). From Eq. (6) and the sentence which follows, we have that $\sup_{\mathbf{y}\in F'\cap B,\xi_{\mathbf{y}}\in[\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})]}\xi_{\mathbf{y}} = O(\|\boldsymbol{\epsilon}\|)$. Since $\|\boldsymbol{\epsilon}\|$ converges to zero in probability, this implies that $\xi_{\mathbf{y}}$ converges to zero in probability, uniformly for any $\mathbf{y} \in F' \cap B$. Moreover, $p(\cdot)$ is uniformly continuous in the compact domain B, and thus $p(\mathbf{y},\xi_{\mathbf{y}})$ converges uniformly in probability to $p(\mathbf{y},0)$. As a result, we have that

$$\sup_{\mathbf{y}\in F'\cap B}\sup_{\xi_{\mathbf{y}}\in[\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})]}|p(\mathbf{y},\xi) - p(\mathbf{y},0)| = o_p(1). \quad (7)$$

Substituting Eq. (6) and Eq. (7) into Eq. (5), and using the fact that $\|\boldsymbol{\epsilon}\| = O_p(1/\sqrt{m})$, we get that the expression in Eq. (5) (and hence Eq. (4)) is $o_p(1/\sqrt{m})$ 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, 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*

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

so that

$$t(\mathbf{c},\mathbf{c}',i,j) = \int_{Q(\mathbf{c},\mathbf{c}',i,j)} p(\mathbf{x})d\mathbf{x}. \quad (8)$$

Then $t(\mathbf{c},\mathbf{c}',i,j)$ is distributed as

$$\int_{F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})|l(\mathbf{x},\mathbf{c}_i,\mathbf{c}'_j)|d\mathbf{x} + o_p(1/\sqrt{m}),$$

where $l(\mathbf{x},\mathbf{c}_i,\mathbf{c}'_j)$ is distributed as

$$\frac{1}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}\begin{pmatrix}\boldsymbol{\mu}_i - \mathbf{x}\\ \mathbf{x} - \boldsymbol{\mu}_j\end{pmatrix}^{\top}\begin{pmatrix}\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}'_i\\ \boldsymbol{\epsilon}_j - \boldsymbol{\epsilon}'_j\end{pmatrix}.$$

Proof. The right half of Fig. 3 should help to 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 $F_{\mathbf{c}',i,j}$. A potential 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. Our goal is to upper and lower bound the integral in Eq. (8) by expressions which are identical up to $o_p(1/\sqrt{m})$ terms, giving us the desired result.

As in Lemma 3, we assume that $H_{\boldsymbol{\mu},i,j}$ is aligned with the axes, such that for any $\mathbf{x} \in H_{\boldsymbol{\mu},i,j}$, its last coordinate is 0. Define $F_{\max}(\boldsymbol{\mu}, \mathbf{c}, \mathbf{c}', i, j) \subseteq H_{\boldsymbol{\mu},i,j}$ as the projection of $Q(\mathbf{c}, \mathbf{c}', i, j)$ on $H_{\boldsymbol{\mu},i,j}$. By definition of $\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}), \tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})$, any point $\mathbf{x} = (\mathbf{y}, 0)$ in $F_{\max}(\boldsymbol{\mu}, \mathbf{c}, \mathbf{c}', i, j)$ has the property that the width of $Q(\mathbf{c}, \mathbf{c}', i, j)$ relative to $H_{\boldsymbol{\mu},i,j}$ at $\mathbf{x}$ is at most $|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}) - \tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})|$.

Define $\delta F(\boldsymbol{\mu}, \mathbf{c}, \mathbf{c}', i, j) \subseteq H_{\boldsymbol{\mu},i,j}$ as the projection on $H_{\boldsymbol{\mu},i,j}$ of $\partial Q(\mathbf{c}, \mathbf{c}', i, j) \backslash (F_{\mathbf{c},i,j} \cup F_{\mathbf{c}',i,j})$, where $\partial Q(\mathbf{c}, \mathbf{c}', i, j)$ is the boundary of $Q(\mathbf{c}, \mathbf{c}', i, j)$. In words, it is the projection of the boundaries of $Q(\mathbf{c}, \mathbf{c}', i, j)$, other than $F_{\mathbf{c},i,j}, F_{\mathbf{c}',i,j}$, on $H_{\boldsymbol{\mu},i,j}$. Any point $\mathbf{x} = (\mathbf{y}, 0)$ in $\delta F(\boldsymbol{\mu}, \mathbf{c}, \mathbf{c}', i, j)$ has the property that the width of $Q(\mathbf{c}, \mathbf{c}', i, j)$, relative to $H_{\boldsymbol{\mu},i,j}$ at $\mathbf{x}$, is less than $|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}) - \tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})|$. This is because the segment of the normal to $H_{\boldsymbol{\mu},i,j}$ at $\mathbf{x}$, between $H_{\mathbf{c},i,j}$ and $H_{\mathbf{c}',i,j}$, passes through other clusters besides clusters i, j.

For notational convenience, we will drop most of the parameters from now on, as they should be clear from the context. Let $F_{\min} = F_{\max} \backslash \delta F$. By the properties of $F_{\max}, \delta F$, any point $\mathbf{x} = (\mathbf{y}, 0)$ in $F_{\min}$ has the property that the width of Q relative to $H_{\boldsymbol{\mu},i,j}$ at $\mathbf{x}$ is exactly $|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}) - \tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})|$.

Let $F'_{\max}, F'_{\min}$ and F' be the $n-1$ dimensional projections of $F_{\max}, F_{\min}$ and F respectively, by removing the last zero coordinate which we assume to characterize $H_{\boldsymbol{\mu},i,j}$. As a result of the previous discussion, by Fubini's theorem, we have that:

$$\int_{F'_{\max}} \left| \int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})} p(\mathbf{y}, \xi) d\xi \right| d\mathbf{y} \geq \int_Q p(\mathbf{x}) d\mathbf{x}$$
$$\geq \int_{F'_{\min}} \left| \int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})} p(\mathbf{y}, \xi) d\xi \right| d\mathbf{y}, \tag{9}$$

Assuming these integrals exist. Our goal will be to show that both the upper and lower bounds above are of the form

$$\int_{F_{\boldsymbol{\mu},i,j}} p(\mathbf{x}) |l(\mathbf{x}, \mathbf{c}_i, \mathbf{c}'_j)| d\mathbf{x} + o_p(1/\sqrt{m}),$$

which entails that the 'sandwiched' integral in Eq. (9) has the same form. We will prove this assertion for the upper bound only, as the proof for the lower bound is almost identical.

As in Lemma 3, we let B be a closed ball of radius R in $\mathbb{R}^n$ centered on the origin, and separately analyze the integral in the upper bound of Eq. (9) with respect to what happens inside and outside this ball.

By Lemma 2, assuming $\|\boldsymbol{\epsilon}\|$ is small enough, there exists a constant $a > 0$ dependent only on $\boldsymbol{\mu}$, such that

$$|\ell_{\boldsymbol{\epsilon}}(\mathbf{y})| \leq a(\|\mathbf{y}\| + 1)(\|\boldsymbol{\epsilon}\| + \|\boldsymbol{\epsilon}\|^2).$$

As before, to avoid making our equations too cumbersome, we shall ignore in the analysis below the higher order term $\|\boldsymbol{\epsilon}\|^2$, since $\boldsymbol{\epsilon}$ converges to 0 in probability and therefore it becomes insignificant compared to $\|\boldsymbol{\epsilon}\|$. Also, since we conveniently assume that $H_{\boldsymbol{\mu},i,j}$ passes through the origin, then any normal to a point in $H_{\boldsymbol{\mu},i,j} \cap B^c$ lies outside B. This is not critical for our analysis (in the general case, we could have simply defined B as centered on some point in $H_{\boldsymbol{\mu},i,j}$), but does simplify things a bit. With these observations, we have that

$$\begin{aligned}
&\int_{F'_{\max} \cap B^c} \left| \int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})} p(\mathbf{y}, \xi) d\xi \right| d\mathbf{y} \\
&\leq \int_{F'_{\max} \cap B^c} |\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}) - \tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y}| \sup_{\xi \in \mathbb{R}} p(\mathbf{y}, \xi) d\mathbf{y} \\
&\leq \int_{F'_{\max} \cap B^c} (|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| + |\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})|) \sup_{\xi \in \mathbb{R}} p(\mathbf{y}, \xi) d\mathbf{y} \\
&\leq a(\|\boldsymbol{\epsilon}\| + \|\boldsymbol{\epsilon}'\|) \int_{F'_{\max} \cap B^c} (\|\mathbf{y}\| + 1) \sup_{\xi \in \mathbb{R}} p(\mathbf{y}, \xi) d\mathbf{y} \\
&\leq a(\|\boldsymbol{\epsilon}\| + \|\boldsymbol{\epsilon}'\|) \int_{H_{\boldsymbol{\mu},i,j} \cap B^c} (\|\mathbf{x}\| + 1) g(\|\mathbf{x}\|) d\mathbf{x} \\
&\leq a(\|\boldsymbol{\epsilon}\| + \|\boldsymbol{\epsilon}'\|) \int_{r=R}^{\infty} (r+1) g(r) * e r^{n-1} dr,
\end{aligned}$$

where $g(\cdot)$ is the dominating function on $p(\cdot)$ assumed to exist by the regularity conditions (see section 2), and e is the surface area of an n dimensional unit sphere. By the assumptions on $g(\cdot)$ and the fact that $\|\boldsymbol{\epsilon}\|, \|\boldsymbol{\epsilon}'\| = O_p(1/\sqrt{m})$, we have that

$$\int_{F'_{\max} \cap B^c} \left| \int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})} p(\mathbf{y}, \xi) d\xi \right| d\mathbf{y} = O_p\left(h(R)/\sqrt{m}\right), \tag{10}$$

where $h(R) \to 0$ as $R \to \infty$. Notice that to reach this conclusion, we did not use any characteristics of $F'_{\max}$, beside it being a subset of $H_{\boldsymbol{\mu},i,j}$. Therefore, since $|l(\mathbf{x}, \mathbf{c}_i, \mathbf{c}'_j)| \leq a(\|\mathbf{x}\| + 1)(\|\boldsymbol{\epsilon}\| + \|\boldsymbol{\epsilon}'\|)/\sqrt{m}$ for some constant $a > 0$, a very similar analysis reveals that

$$\int_{F' \cap B^c} p(\mathbf{y}, 0) |l(\mathbf{x}, \mathbf{c}_i, \mathbf{c}'_j)| d\mathbf{y} = O_p\left(h(R)/\sqrt{m}\right). \tag{11}$$

We note for later that none of the constants implicit in the $O_p(\cdot)$ notation, other than $h(R)$, depend on R. Turning now to what happens inside the ball, we have by Lemma 3 that

$$\begin{aligned}
&\int_{F'_{\max} \cap B} \left| \int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})} p(\mathbf{y}, \xi) d\xi \right| d\mathbf{y} \\
&= \int_{F'_{\max} \cap B} |\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y}) - \tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| p(\mathbf{y}, 0) d\mathbf{y} + o_p(1/\sqrt{m}).
\end{aligned} \tag{12}$$

Leaving this equation aside for later, we will now show that

$$\left|\int_{F'_{\max}\cap B} |\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y}) - \tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| p(\mathbf{y},0) d\mathbf{y} - \int_{F'\cap B} |\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y}) - \tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| p(\mathbf{y},0) d\mathbf{y}\right| = o_p(1/\sqrt{m}). \qquad (13)$$

The l.h.s can be upper bounded by

$$\begin{aligned}&\int_{(F'_{\max}\triangle F')\cap B} |\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}) - \tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})| p(\mathbf{y},0) d\mathbf{y}\\ &\leq \int_{(F'_{\max}\triangle F')\cap B} (|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| + |\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})|) p(\mathbf{y},0) d\mathbf{y}.\end{aligned}$$

As $\boldsymbol{\epsilon}, \boldsymbol{\epsilon}'$ have the same distribution, we just need to show that

$$\int_{(F'_{\max}\triangle F')\cap B} |\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| p(\mathbf{y},0) d\mathbf{y} = o_p(1/\sqrt{m}). \qquad (14)$$

By Lemma 2, inside the bounded domain of B, we have that $|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| \leq a\|\boldsymbol{\epsilon}\|$ for some constant a dependent solely on $\boldsymbol{\mu}$ and R (as before, to avoid making the equations too cumbersome, we ignore terms involving higher powers of $\|\boldsymbol{\epsilon}\|$). Moreover, since $p(\mathbf{y},0)$ is bounded, we can absorb this bound into a and get that

$$\int_{(F'_{\max}\triangle F')\cap B} |\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| p(\mathbf{y},0) d\mathbf{y} \leq a\|\boldsymbol{\epsilon}\| \int_{(F'_{\max}\triangle F')\cap B} 1 d\mathbf{y}, \qquad (15)$$

Note that $\int_{(F'_{\max}\triangle F')\cap B} 1 d\mathbf{y}$ is a continuous function of $\boldsymbol{\epsilon}, \boldsymbol{\epsilon}'$ in some neighborhood of 0. Moreover, since $F'_{\max} = F'$ when $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}' = 0$, the integral above is 0 at $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}' = 0$. Since $\|\boldsymbol{\epsilon}\|, \|\boldsymbol{\epsilon}\|$ converge to 0 in probability, it follows that

$$\int_{(F'_{\max}\triangle F')\cap B} 1 d\mathbf{y} = o_p(1).$$

Combining this with Eq. (15), and the fact that $\|\boldsymbol{\epsilon}\| = O_p(1/\sqrt{m})$, justifies Eq. (14), and hence Eq. (13). Combining Eq. (10),Eq. (12) and Eq. (13), we get that

$$\begin{aligned}&\int_{F'_{\max}} \left|\int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})} p(\mathbf{y},\xi) d\xi\right| d\mathbf{y}\\ &= \int_{F'\cap B} |\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y}) - \tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| p(\mathbf{y},0) d\mathbf{y}\\ &+ o_p(1/\sqrt{m}) + O_p(h(R)/\sqrt{m}). \qquad (16)\end{aligned}$$

By Lemma 2, definition of $l(\mathbf{x},\mathbf{c}_i,bc'_j)$, and the fact that $\|\boldsymbol{\epsilon}\|, \|\boldsymbol{\epsilon}'\| = O_p(1/\sqrt{m})$, we have that $\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}) - \tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})$ is equal to $|l(\mathbf{x},\mathbf{c}_i,\mathbf{c}'_j)| + o_p((\|\mathbf{y}\|+1)/\sqrt{m})$. This implies that the distribution of the r.h.s of Eq. (16) is equal to

$$\int_{F'\cap B} p(\mathbf{y},0)|l(\mathbf{x},\mathbf{c}_i,\mathbf{c}'_j)| d\mathbf{y} + o_p(1/\sqrt{m}) + O_p(h(R)/\sqrt{m}).$$

By Eq. (11), this is equal in turn to

$$\int_{F'} p(\mathbf{y},0)|l(\mathbf{x},\mathbf{c}_i,\mathbf{c}'_j)| d\mathbf{y} + o_p(1/\sqrt{m}) + O_p(h(R)/\sqrt{m}).$$

We now use the fact that R can be picked arbitrarily. Notice that the first remainder term has implicit constants which depend on R, but the second remainder term depends on R only through $h(R)$ (recall the development leading to Eq. (10) and Eq. (11)). Therefore, the first remainder term converges to 0 at a rate faster than $1/\sqrt{m}$ in probability for any R, and the second remainder term can be made arbitrarily smaller than $1/\sqrt{m}$ in high probability by picking R to be large enough, since $h(R) \to 0$ as $R \to \infty$. Thus, for any $\delta > 0$, we can pick R so that the remainder terms eventually become smaller than $\delta/\sqrt{m}$ with arbitrarily high probability. As a result, we can replace the remainder terms by $o_p(1/\sqrt{m})$, with implicit constants not depending on R, and get that Eq. (16) can be rewritten as

$$\begin{aligned}&\int_{F'_{\max}} \left|\int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})} p(\mathbf{y},\xi) d\xi\right| d\mathbf{y}\\ &= \int_{F'} p(\mathbf{y},0)|l(\mathbf{x},\mathbf{c}_i,\mathbf{c}'_j)| d\mathbf{y} + o_p(1/\sqrt{m}).\end{aligned}$$

This gives us an equivalent formulation of the upper bound in Eq. (9). As discussed immediately after Eq. (9), an identical analysis can be performed for the lower bound appearing there, and this leads to the result of the lemma. □

We now turn to prove 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 by definition, $d^m_{\mathcal{D}}(\mathtt{A_k}(S_1),\mathtt{A_k}(S_2))$ is equal to

$$2\sum_{1\leq i<j\leq k}\left(\int_{\widehat{C}_{\mathbf{c},\mathbf{c}',i}\cup\widehat{C}_{\mathbf{c},\mathbf{c}',j}} p(\mathbf{x})d\mathbf{x}\right)\sqrt{m}t(\mathbf{c},\mathbf{c}',i,j). \qquad (17)$$

As a straightforward consequence of $\|\epsilon\| = O_p(1/\sqrt{m})$, we have that

$$\int_{\widehat{C}_{\mathbf{c},\mathbf{c}',i}\cup\widehat{C}_{\mathbf{c},\mathbf{c}',j}} p(\mathbf{x})d\mathbf{x} = \int_{C_{\boldsymbol{\mu},i}\cup C_{\boldsymbol{\mu},j}} p(\mathbf{x})d\mathbf{x} + o_p(1). \qquad (18)$$

By Lemma 4, we have that $\sqrt{m}t(\mathbf{c},\mathbf{c}',i,j)$ is of the form

$$\int_{F_{\boldsymbol{\mu},i,j}} \frac{\sqrt{m}p(\mathbf{x})}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}\left|\begin{pmatrix}\boldsymbol{\mu}_i - \mathbf{x}\\ \mathbf{x} - \boldsymbol{\mu}_j\end{pmatrix}^{\top}\begin{pmatrix}\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}'_i\\ \boldsymbol{\epsilon}_j - \boldsymbol{\epsilon}'_j\end{pmatrix}\right| d\mathbf{x} + o_p(1). \qquad (19)$$

By the continuous mapping theorem [18] and standard results on the difference of independent, identically distributed Gaussian vectors [17], we have that $\sqrt{m}(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}'_i, \boldsymbol{\epsilon}_j - \boldsymbol{\epsilon}'_j)^{\top}$ converges in distribution to $\sqrt{2}(\mathbf{v}_i,\mathbf{v}_j)^{\top}$, where $\mathbf{v}$ is as defined in Lemma 1. Moreover, it is not difficult to show that Eq. (19), ignoring the remainder term, is a continuous function of $(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}'_i, \boldsymbol{\epsilon}_j - \boldsymbol{\epsilon}'_j)^{\top}$. The idea is that it is obviously continuous with the integral restricted to some fixed ball around the origin, and the contributions outside the ball can be made arbitrarily small if the ball is large enough, by the assumptions on $p(\mathbf{x})$ (a similar argument was made in the proof of Lemma 4). Thus, by the continuous mapping theorem, $\sqrt{m}t(\mathbf{c},\mathbf{c}',i,j)$ converges in distribution to

$$\int_{F_{\boldsymbol{\mu},i,j}} \frac{\sqrt{2}p(\mathbf{x})}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}\left|\begin{pmatrix}\boldsymbol{\mu}_i - \mathbf{x}\\ \mathbf{x} - \boldsymbol{\mu}_j\end{pmatrix}^{\top}\begin{pmatrix}\mathbf{v}_i\\ \mathbf{v}_j\end{pmatrix}\right| d\mathbf{x}. \qquad (20)$$

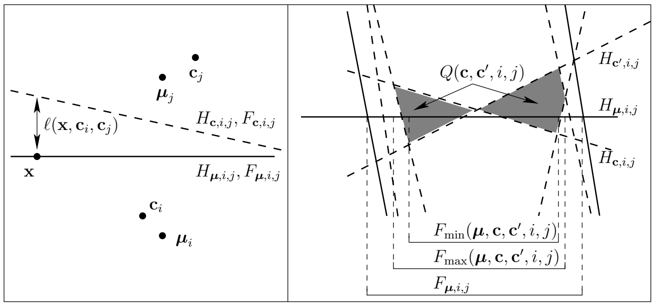
$\mathbf{c}_j$
$\boldsymbol{\mu}_j$
$\ell(\mathbf{x}, \mathbf{c}_i, \mathbf{c}_j)$
$H_{\mathbf{c},i,j}, F_{\mathbf{c},i,j}$
$\mathbf{x}$
$H_{\boldsymbol{\mu},i,j}, F_{\boldsymbol{\mu},i,j}$
$\mathbf{c}_i$
$\boldsymbol{\mu}_i$
$Q(\mathbf{c}, \mathbf{c}', i, j)$
$H_{\mathbf{c}',i,j}$
$H_{\boldsymbol{\mu},i,j}$
$H_{\mathbf{c},i,j}$
$F_{\min}(\boldsymbol{\mu}, \mathbf{c}, \mathbf{c}', i, j)$
$F_{\max}(\boldsymbol{\mu}, \mathbf{c}, \mathbf{c}', i, j)$
$F_{\boldsymbol{\mu},i,j}$

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.

Substituting Eq. (18) and Eq. (20) into Eq. (17), we get convergence in distribution to the one specified in our theorem.

The only thing remaining is to derive the expected value of this distribution. For notational convenience, let $\Sigma = \Gamma^{-1}V\Gamma^{-1}$, and

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

the expected value of the distribution is equal to:

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

By Fubini's theorem, this is equal to:

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

The expression inside the expectation is normally distributed, as a linear transformation of a normal random vector. Using standard results on the distribution of such transformations [17], and since for any univariate $a \sim \mathcal{N}(\mu, \sigma^2)$ it holds that $\mathbb{E}[|a|] = \sigma\sqrt{2/\pi}$, we can reduce the above to

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

The final form of $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$ is achieved by rewriting Σ as $(V^{1/2}\Gamma^{-1})^\top V^{1/2}\Gamma^{-1}$, substituting into the expression $\psi(\mathbf{x}, i, j)$, and simplifying.

4.2 Proof of Thm. 2

The proof is composed of several lemmas. The key insight is that the asymptotic distribution of $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$ can be viewed as a certain non-standard norm of a Gaussian random vector. Using theorems on Gaussian measures in Banach spaces allows us to bound the probability of $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$ being much larger or much smaller than its expectation, and thus bound the probability that the empirical clustering stability estimator will return deceiving results.

Lemma 5. *The asymptotic distribution of $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$ 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 $\mathbf{v}_i \in \mathbb{R}^n$. By Thm. 1, the asymptotic distribution of $d^m_{\mathcal{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$ is equal to

$$\sum_{1 \le i < j \le k} a_{i,j} \int_{F_{\boldsymbol{\mu},i,j}} p(\mathbf{x}) \left| \begin{pmatrix} \boldsymbol{\mu}_i - \mathbf{x} \\ \mathbf{x} - \boldsymbol{\mu}_j \end{pmatrix}^\top \begin{pmatrix} \mathbf{v}_i \\ \mathbf{v}_j \end{pmatrix} \right| d\mathbf{x}, \quad (21)$$

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. (21) is always non-negative. Finally, Eq. (21) 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 $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathcal{D})$ be positive (which is implied by the assumptions in the theorem). From this, it is straightforward to show that if $\mathbf{v} \neq 0$ then Eq. (21) is larger than $\mathbf{0}$. □

Lemma 6. *Let $\mathbf{v}$ be a non-degenerate normally distributed random vector in $\mathbb{R}^n$, let $\|\cdot\|_*$ be a norm on $\mathbb{R}^n$ as defined in Lemma 5, and let $\theta \in (1/2, 1)$ be a free parameter. Introduce the following two parameters which depend on θ:*

$$a_\theta = 1 + \frac{2(1-\theta)}{\log\left(\frac{\theta}{1-\theta}\right)}\ ,\ b_\theta = 1 - \theta + \frac{1 - \exp(-(\mathit{erf}^{-1}(\theta))^2)}{\sqrt{\pi}\mathit{erf}^{-1}(\theta)}.$$

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

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

and

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

Proof. The distribution of a norm of a Gaussian random vector is continuous, except possibly at 0 (cf. [3]). For any $\theta \in (1/2, 1)$, let med_θ be a positive number which satisfies:

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

Using two results from the literature on Gaussian measures in Banach spaces (theorem III.3 in [11], and theorem 1 from [8]), we have that for any $M \geq 1$, and for any $\epsilon \in [0,1]$, it holds that:

$$\Pr(\|\mathbf{v}\|_* > M\mathrm{med}_\theta) \leq \theta\left(\frac{1-\theta}{\theta}\right)^{(1+M)/2} \tag{22}$$

$$\Pr(\|\mathbf{v}\|_* \leq \epsilon\mathrm{med}_\theta) \leq \mathrm{erf}(\mathrm{erf}^{-1}(\theta)\epsilon). \tag{23}$$

It remains to convert these bounds on the deviation from med_θ to the deviation from $\mathbb{E}\|\mathbf{v}\|_*$. To achieve this, we need to upper and lower bound $\mathbb{E}\|\mathbf{v}\|_*/\mathrm{med}_\theta$. By substitution of variables, we have that $\mathbb{E}\|\mathbf{v}\|_*$ is equal to

$$\int_0^\infty \Pr(\|\mathbf{v}\|_* > t)dt = \mathrm{med}_\theta \int_0^\infty \Pr(\|\mathbf{v}\|_* > M\mathrm{med}_\theta)dM.$$

Using Eq. (22), this can be upper bounded by

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

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

In a similar manner, we can write $\mathbb{E}\|\mathbf{v}\|_*$ as

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

which is lower bounded in term, using Eq. (23), by

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

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

Therefore, we have that if $Mb_\theta > 1$, then $\Pr(\|\mathbf{v}\|_* > M\mathbb{E}\|\mathbf{v}\|_*)$ is upper bounded by

$$\Pr(\|\mathbf{v}\|_* > Mb_\theta\mathrm{med}_\theta) \leq \theta\left(\frac{1-\theta}{\theta}\right)^{(1+Mb_\theta)/2}.$$

The other bound in the lemma is derived similarly. □

We can now turn to the proof of Thm. 2. By Lemma 5, both $d^m_{\mathcal{D}}(\mathtt{A}_{\mathtt{k}_\mathtt{s}}(S_1), \mathtt{A}_{\mathtt{k}_\mathtt{s}}(S_2))$ and $d^m_{\mathcal{D}}(\mathtt{A}_{\mathtt{k}_\mathtt{u}}(S_1), \mathtt{A}_{\mathtt{k}_\mathtt{u}}(S_2))$ converge in distribution to $\|\mathbf{v}_{k_u}\|_*$ and $\|\mathbf{v}_{k_s}\|_*$, where $\mathbf{v}_{k_u}, \mathbf{v}_{k_s}$ are Gaussian random variables (non-degenerate by the assumptions on Γ and V). By a union bound argument and the definition of convergence in distribution, we have that for any fixed number c,

$$\Pr(d^m_{\mathcal{D}}(\mathtt{A}_{\mathtt{k}_\mathtt{u}}(S_1), \mathtt{A}_{\mathtt{k}_\mathtt{u}}(S_2)) \leq 1.1 d^m_{\mathcal{D}}(\mathtt{A}_{\mathtt{k}_\mathtt{s}}(S_1), \mathtt{A}_{\mathtt{k}_\mathtt{s}}(S_2)))$$
$$\leq \Pr(\|\mathbf{v}_{k_u}\|_* \leq c) + \Pr(1.1\|\mathbf{v}_{k_s}\|_*) \geq c) + o(1). \tag{24}$$

The combination of Lemma 5 and Lemma 6 allows us to upper bound the probability that $\|\mathbf{v}_{k_u}\|_*$ is smaller than its expectation by a factor $\epsilon < 1$, and upper bound the probability that $\|\mathbf{v}_{k_s}\|_*$ is larger than its expectation by some factor $M > 1$, provided that ϵ, M satisfy the conditions specified in Lemma 6.

Therefore, if we choose M and ϵ so that $1.1M/\epsilon \leq R$, where R is as defined in the lemma, we get that Eq. (24) above is upper bounded by

$$\theta_1\left(\frac{1-\theta_1}{\theta_1}\right)^{((1+M)b_{\theta_1})/2} + \mathrm{erf}(\mathrm{erf}^{-1}(\theta_2)a_{\theta_2}\epsilon), \tag{25}$$

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 constants, and minimality requirements on R (which stem from the requirements on M, ϵ by Lemma 6). Choosing $\theta_1 = 0.9$, $\theta_2 = 0.8$, $M = 2\log(R)/(b_{\theta_1}\log(\theta_1/(1-\theta_1)))$, $\epsilon = 1.1M/R$, and using the fact that $\mathrm{erf}(x) \leq (2/\sqrt{\pi})x$ for any $x \geq 0$, we get that Eq. (25) is upper bounded by $(0.3 + 3\log(R))/R$ for any $R > 3$, and therefore Eq. (24) is upper bounded by $(0.3 + 3\log(R))/R + o(1)$.

Assume the event

$$d^m_{\mathcal{D}}(\mathtt{A}_{\mathtt{k}_\mathtt{u}}(S_1), \mathtt{A}_{\mathtt{k}_\mathtt{u}}(S_2)) > 1.1 d^m_{\mathcal{D}}(\mathtt{A}_{\mathtt{k}_\mathtt{s}}(S_1), \mathtt{A}_{\mathtt{k}_\mathtt{s}}(S_2)), \tag{26}$$

occurs. Recall that the quantities in Eq. (26) depend on the unknown underlying distribution $\mathcal{D}$, and therefore cannot be calculated directly. Instead, we empirically estimate these quantities (divided by $\sqrt{m}$ to be exact), as defined in the theorem statement, to get the stability estimators $\hat{\theta}_{k_u,4m}$ and $\hat{\theta}_{k_s,4m}$. Thus, even if Eq. (26) occurs, it is still possible that $\hat{\theta}_{k_u,4m} \leq \hat{\theta}_{k_s,4m}$. Luckily, by Thm. 2 in [15], the probability for this, conditioned on the event in Eq. (26) is $o(1)$ (namely, converges to 0 as $m \to \infty$). Therefore, the probability that Eq. (26) does not occur, or that it does occur but the empirical comparison of these quantities fail, is $(0.3 + 3\log(R))/R + o(1)$ as required.

4.3 Proof of Thm. 3

To prove the theorem, we will borrow a setting discussed in [10] 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 mentioned in Sec. 2, 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 $d^m_{\mathcal{D}_\epsilon}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$ converges to 0 in probability, 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 $\{-1-\Delta, -1, 1, 1+\Delta\}$ is hit by 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 [16], for any Bernoulli random variable X such that $\mathbb{E}[X] = p \leq 1/2$, and any whole number a such that $a/m \leq 1-p$, if $X_1, \ldots, X_m$ are m i.i.d copies of X, then

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

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

$$\Pr(A^1_{m,\epsilon}) \geq 1 - \Phi\left(\frac{4}{\sqrt{m}} + 2\epsilon\sqrt{m}\right). \tag{27}$$

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

$$\Pr(A^2_{m,\epsilon}) \geq 1/2. \tag{28}$$

Finally, it is straightforward to show that $\Pr(B_{m,\epsilon})$ is arbitrarily close to 1 uniformly for any ϵ, if m is large enough. Combining this with Eq. (27), Eq. (28) 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, it holds that

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

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

$$d^m_{\mathcal{D}_\epsilon}(\mathtt{A}_3(S_1), \mathtt{A}_3(S_2)) = \frac{\sqrt{m}(1+\epsilon^2)}{4} > \frac{\sqrt{m}}{4}.$$

So with a probability of at least $1/3$ over drawing and clustering two independent samples, the distance between the clusterings is more than $\sqrt{m}/4$, as required.

5 Conclusions and Future Work

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 does not 'break down' in the large sample regime, and made some preliminary observations about the factors influencing it. These factors appear to be reasonable requirements from a 'correct' model, and accords with clustering stability working successfully in many situations. However, they also imply that clustering stability might sometimes behave unexpectedly, for example in hierarchical clustering situations, as illustrated in subsection 3.3.

There are several directions for future research. The most obvious perhaps is to extend our results and observations from the asymptotic domain to the finite sample size domain. Showing that clustering stability does not 'break down' in the large sample regime has theoretical and practical relevance, but leaves open the question of why clustering stability can work well for small finite samples. One route to achieve this might be through finite sample guarantees, but as demonstrated in Thm. 3, additional assumptions are needed for such results. Also, it would be interesting to perform a similar analysis for other clustering methods beyond the k-means framework.

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] V.I. Bogachev. *Gaussian Measures*. American Mathematical Society, 1998.

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

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

[6] 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.

[7] 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.

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

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

[10] 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.

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

[12] David Pollard. Personal communication.

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

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

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

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

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

[18] Aad W. Van Der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes : With Applications to Statistics*. Springer, 1996.