Noname manuscript No.
(will be inserted by the editor)

Stability and Model Selection in *k*-means Clustering

Ohad Shamir · **Naftali Tishby**

the date of receipt and acceptance should be inserted later

Abstract Clustering Stability methods are a family of widely used model selection techniques for 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, at least for the *k*-means framework. Moreover, it allows us to identify the factors which eventually determine the behavior of clustering stability. This leads to some basic observations about what kind of assumptions are made when using these methods. While often reasonable, these assumptions might also lead to unexpected consequences.

Keywords Clustering, Model Selection, Stability, Statistical Learning Theory

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.

Ohad Shamir
School of Computer Science and Engineering, The Hebrew University, Jerusalem 91904, Israel
E-mail: ohadsh@cs.huji.ac.il

Naftali Tishby
School of Computer Science and Engineering, and Interdisciplinary Center for Neural Computation, The Hebrew University, Jerusalem 91904, Israel
E-mail: tishby@cs.huji.ac.il

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 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 specific generative assumptions.

In the past few years, an increasingly popular family of such model selection methods are those 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. (Ben-Hur et al. 2002; Dudoit and Fridlyand 2002; Lange et al. 2004; Levine and Domany 2001; Smolkin and Ghosh 2003; Bertoni and Valentini 2007)), 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 (Ben-David et al. 2006, 2007) 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 (Krieger and Green 1999)). It is important to emphasize that this is not just a theoretical issue. If the scenario above indeed occurs, it implies that there exists some sample size, which depends on the underlying distribution and hence may be hard to compute, beyond which we should not trust the results of clustering stability methods.

A possible solution to this difficulty was proposed in (Shamir and Tishby 2007). In a nutshell, that paper showed that even when all considered models eventually become completely stable, the *relative* stability of each model compared to the other models can sometimes be reliably discerned - even when the sample size increases to infinity. With this more refined analysis, it was argued that there may be no upper limit to the sample size for which clustering stability remains meaningful. 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. 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 even for asymptotically large samples, clustering stability does not become meaningless, in the sense described earlier, at least for the k-means framework that we study. While the preliminary version of this paper (Shamir and Tishby 2008a) assumed an ideal algorithm, which finds the global optimum of the k-means objective function, here we extend our results to the actual algorithm used in practice, which might return a sub-optimal solution. Also, we note that using different tools, some of the results presented here can be extended to more general families of clustering frameworks beyond k-means (Shamir and Tishby 2008b).

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, sample-size-specific effects diminish, and what remains are the more fundamental characteristics of the method. As a result, the analysis leads to some basic observations about the factors influencing clustering stability for the k-means framework, which may be of theoretical and practical interest.

The paper is organized as follows. In Sec. 2, we introduce the problem setting and the notation we shall use. The notation is also summarized in table 1. In Sec. 3, we formally present the results which characterize the asymptotic behavior of clustering stability in the k-means framework. We build on these results in Sec. 4, where we discuss the factors influencing this behavior, and how do they affect what is considered as a 'stable' or 'unstable' model by clustering stability methods. These observations are illustrated with some simple examples in Sec. 5. In Sec. 6, we give a negative result about the convergence rates of clustering stability estimators to their asymptotic distribution. Almost all the proofs in the paper are concentrated in Sec. 7, except for the proof of one of the lemmas, which is placed in an appendix due to its length and it being conceptually separate from the other results. We end with conclusions in Sec. 8.

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. A list of the notation used may be found in table 1.

Denote $\{1,\ldots,k\}$ as $[k]$. Vectors will be denoted by bold-face characters. $\|\cdot\|$ will denote the Euclidean norm unless stated otherwise. $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Let $\mathscr{D}$ be a probability distribution on $\mathbb{R}^n$, with a bounded probability density function $p(\cdot)$, which is defined everywhere, and 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, $\mathscr{D}$ 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 purely for technical reasons 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 assumption, it should be noted that our results hold even if we assume continuity solely in some open neighborhood of the limit cluster boundaries to which our clustering algorithm converges (to be formally defined shortly). However, since this somewhat complicates the analysis without leading to novel insights, we will take this stronger assumption for simplicity.

Let $\mathtt{A}_\mathtt{k}(\cdot)$ denote the (possibly randomized) 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 $\mathscr{D}$, and a required number of clusters k, and returns a set of centroids $\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_k) \in \mathbb{R}^{nk}$. These will usually be thought of as random variables, dependent on the randomness of the sample. Recall that the k-means algorithm attempts to minimize the objective function

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

via alternating steps of associating each instance to its nearest centroid, and then repositioning the centroids at the center of mass of their respective clusters (for further discussion of the algorithm and its properties, see for instance (Duda et al. 2001; Hartigan 1975; Steinley 2006)). This procedure is not guaranteed, in general, to find the global minimum of $\hat{W}(\mathbf{c})$.

In a statistical setting, $\hat{W}(\mathbf{c})$ can be seen as an empirical approximation of the objective function with respect to the underlying distribution, defined as

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

As discussed earlier, we focus in this paper on the setting where the clustering algorithm converges to a single solution as the sample size goes to infinity. Therefore, we shall assume that as the sample size m increases, the centroids returned by the algorithm converges in probability to a single fixed solution $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) \in \mathbb{R}^{nk}$ (up to permutation of the centroids), with centroids which lie at the center of mass of the clusters with respect to the underlying distribution:

$$\forall i \in [k] \;\; \boldsymbol{\mu}_i = \frac{\int_{\mathbf{x} \in C_{\boldsymbol{\mu},i}} \mathbf{x} p(\mathbf{x}) d\mathbf{x}}{\int_{\mathbf{x} \in C_{\boldsymbol{\mu},i}} p(\mathbf{x}) d\mathbf{x}}.$$

For simplicity, we will also assume that all these 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), such that this numbering does not change for sufficiently small perturbations of $\boldsymbol{\mu}$.

The basic idea of clustering instability is to measure distances between clusterings, based on different samples from our data. More formally, we define the (scaled) distance between two clusterings $\mathtt{A}_\mathtt{k}(S_1)$ and $\mathtt{A}_\mathtt{k}(S_2)$, where S_1, S_2 are samples of size m, as $\sqrt{m}$ times the probability that a randomly sampled instance from $\mathscr{D}$ will belong to different clusters in $\mathtt{A}_\mathtt{k}(S_1)$ and $\mathtt{A}_\mathtt{k}(S_2)$. Formally,

$$d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2)) := \sqrt{m} \Pr_{\mathbf{x} \in \mathscr{D}} (\mathbf{x} \in \mathtt{A}_\mathtt{k}(S_1)_j, \mathbf{x} \in \mathtt{A}_\mathtt{k}(S_2)_{j'}, j \neq j'). \tag{3}$$

This definition is similar to what clustering stability methods attempt to estimate in practice, by computing the proportion of the data which switches between clusters when the clusterings $\mathtt{A}_\mathtt{k}(S_1)$ and $\mathtt{A}_\mathtt{k}(S_2)$ are compared. The main difference is the additional scaling by $\sqrt{m}$ (the 'correct' scaling factor as will become evident later on). This is usually performed by clustering independent subsamples of the data, and empirically estimating the

distance between the resulting clusterings. The average distance is taken to be the measure of the model instability. Thus, understanding the behavior of $d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$ (over drawing and clustering independent samples) is of much interest in analyzing the behavior of clustering stability.

Any choice of cluster centroids $\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_k)$ induces a Voronoi partition on $\mathbb{R}^n$. For each cluster centroid $\mathbf{c}_i$, we denote the interior of its corresponding cluster as $C_{\mathbf{c},i}$, defined as

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

Also, 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:

$$F_{\mathbf{c},i,j} := \left\{ \mathbf{x} \in \mathbb{R}^n : \arg\min_{a \in [k]} \|\mathbf{c}_a - \mathbf{x}\|^2 = \{i, j\} \right\}. \tag{4}$$

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

$$H_{\mathbf{c},i,j} := \left\{ \mathbf{x} \in \mathbb{R}^n : \left(\mathbf{x} - \frac{\mathbf{c}_i + \mathbf{c}_j}{2} \right)^\top \cdot (\mathbf{c}_1 - \mathbf{c}_2) = 0 \right\}. \tag{5}$$

In the paper, 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 use 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.

The remainder of this section formally defines two matrices which prove to play an important role in how clustering stability behaves. The first matrix, of size $kn \times kn$, is the Hessian of the mapping $W(\cdot)$ at the limit solution $\boldsymbol{\mu}$, which we shall denote as Γ. This matrix is composed of $k \times k$ blocks $\Gamma_{i,j}$ for $i, j \in [k]$, each such block being of size $n \times n$. Each block $\Gamma_{i,j}$ can be shown to be equal to[1]

$$\Gamma_{i,j} := \left(\int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x}) d\mathbf{x} \right) I_n - \sum_{a \neq i} \frac{2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_a\|} \int_{F_{\boldsymbol{\mu},i,a}} p(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^\top d\mathbf{x}, \tag{6}$$

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

$$\Gamma_{i,j} := \frac{2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|} \int_{F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_j)^\top d\mathbf{x}. \tag{7}$$

We will use the same block notation later for its inverse Γ^{-1}. We assume that the matrix Γ is positive definite. This is a mild requirement, because if $\boldsymbol{\mu}$ is a locally optimal solution then Γ is always positive semidefinite. Cases where Γ is not strictly positive definite correspond to singularities which are often pathological (for more discussion on this, see (Radchenko 2004)).

The second matrix we shall need, denoted as V, is equal (up to a constant of 4) to the covariance matrix of $\mathscr{D}$ with respect to each cluster, assuming the optimal clustering induced

[1] This is proven in (Pollard 1982). The definition of Γ there differs from ours in one of the signs, apparently due to a small error in that paper (Pollard, personal communication).

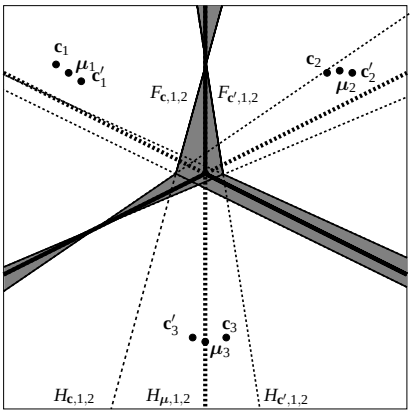

Fig. 1 An illustrative drawing of the setting and notation used. Thicker lines represent the optimal k-means clustering partition (for $k = 3$ clusters) with respect to the underlying distribution. Clustering two independent random samples gives us two random centroid sets $\mathbf{c}$ and $\mathbf{c}'$.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

$[k]$	$\{1,\ldots,k\}$.
$\mathcal{N}(\boldsymbol{\mu},\Sigma)$	Multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ.
$\mathscr{D}$,$p(\cdot)$	Underlying probability distribution and corresponding density function.
$S = (\mathbf{x}_1,\ldots,\mathbf{x}_m)$	Sample of size m drawn i.i.d from $\mathscr{D}$.
$\mathtt{A}_\mathtt{k}(\cdot)$	The standard k-means algorithm.
$\widehat{W}(\cdot)$	The k-means objective function w.r.t. an empirical sample (see Eq. (1)).
$W(\cdot)$	The k-means objective function w.r.t. the underlying distribution $\mathscr{D}$ (see Eq. (2)).
$\mathbf{c} = (\mathbf{c}_1,\ldots,\mathbf{c}_k)$	Cluster centroids, returned by the k-means algorithm based on a random sample.
$\boldsymbol{\mu} = (\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k)$	Limit cluster centroids to which the k-means algorithm converges in probability.
$d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1),\mathtt{A}_\mathtt{k}(S_2))$	Scaled stability measure, based on samples S_1,S_2 of size m (see Eq. (3)).
$\overline{\mathrm{instab}}(\mathtt{A}_\mathtt{k},\mathscr{D})$	Expected value of the limit distribution of $d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1),\mathtt{A}_\mathtt{k}(S_2))$ (see Eq. (9)).
$C_{\mathbf{c},i}$	The cluster associated with centroid $\mathbf{c}_i$.
$F_{\mathbf{c},i,j}$	The boundary between the clusters associated with centroids $\mathbf{c}_i,\mathbf{c}_j$ (see Eq. (4)).
$H_{\mathbf{c},i,j}$	The infinite hyperplane containing the cluster boundary $F_{\mathbf{c},i,j}$ (see Eq. (5)).
Γ	Hessian of $W(\cdot)$ at $\boldsymbol{\mu}$ (see Eq. (6),Eq. (7)).
V	Per-cluster covariance matrix of $\mathscr{D}$ with respect to clustering $\boldsymbol{\mu}$ (see Eq. (8)).

Table 1 Table of Notation

by $\boldsymbol{\mu}$. More specifically, V is 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), where

$$V_i := 4 \int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^\top d\mathbf{x}. \tag{8}$$

We shall assume that V_i has full rank for any i.

3 Asymptotic Behavior of Clustering Stability

In this section, we formally characterize the asymptotic behavior of clustering stability, and discuss some immediate consequences. The detailed proofs are presented in Sec. 7.

At a technical level, our main result is the following theorem, which characterizes the exact distribution to which $d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1),\mathtt{A}_\mathtt{k}(S_2))$ converges for any appropriate underlying distribution $\mathscr{D}$, as well as its expected value.

Theorem 1 *Assume $\mathscr{D}$ 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 $\mathtt{A}_\mathtt{k}(\cdot)$ be the k-means algorithm, and assume that the returned set of centroids $\mathbf{c}=(\mathbf{c}_1,\ldots,\mathbf{c}_k)$, based on i.i.d samples from $\mathscr{D}$, converge in probability to some set of k distinct centroids $\boldsymbol{\mu}=(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k)$ which are a local optimum of $W(\cdot)$. Furthermore, assume that Γ is invertible and that V_i has full rank for any $i\in[k]$. Then we have that $d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1),\mathtt{A}_\mathtt{k}(S_2))$ converges in distribution to that of*

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

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

The expected value of this distribution, denoted as $\widehat{instab}(\mathtt{A}_\mathtt{k},\mathscr{D})$, is equal to

$$\frac{2}{\sqrt{\pi}}\sum_{1\le i<j\le k}\int_{F_{\boldsymbol{\mu},i,j}}p(\mathbf{x})\frac{\Psi(\mathbf{x},i,j)}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}d\mathbf{x},\tag{9}$$

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

$$\left\|\begin{pmatrix}V_i^{1/2}&0\\0&V_j^{1/2}\end{pmatrix}\begin{pmatrix}(\Gamma^{-1})_{i,i}&(\Gamma^{-1})_{i,j}\\(\Gamma^{-1})_{j,i}&(\Gamma^{-1})_{j,j}\end{pmatrix}\begin{pmatrix}\boldsymbol{\mu}_i-\mathbf{x}\\\mathbf{x}-\boldsymbol{\mu}_j\end{pmatrix}\right\|.\tag{10}$$

All the integrals can be shown to exist by the assumptions on $p(\cdot)$.

The asymptotic distribution and $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k},\mathscr{D})$ allows us to characterize 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, $\hat{\theta}_{k,3m}$, as follows: Given a sample of size $3m$, split it randomly into 3 disjoint subsets S_1,S_2,S_3 each of size m. Estimate $d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1),\mathtt{A}_\mathtt{k}(S_2))/\sqrt{m}$ by computing*

$$\frac{1}{m}\sum_{\mathbf{x}\in S_3}\mathbf{1}\Big(\mathbf{x}\in\mathtt{A}_\mathtt{k}(S_1)_j,\mathbf{x}\in\mathtt{A}_\mathtt{k}(S_2)_{j'},j\neq j'\Big).$$

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

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

where the probability is over a sample of size $3m$ used for both estimators, and $o(1)$ converges to 0 as $m\to\infty$. This bound is understood to signify $o(1)$ if $R=\infty$.

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 $\widehat{\text{instab}}(\mathtt{A}_{k_u}, \mathscr{D})$ and $\widehat{\text{instab}}(\mathtt{A}_{k_s}, \mathscr{D})$ 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 $\widehat{\text{instab}}(\mathtt{A}_{k_s}, \mathscr{D})$ and $\widehat{\text{instab}}(\mathtt{A}_{k_u}, \mathscr{D})$. 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 $\widehat{\text{instab}}(\mathtt{A}_k, \mathscr{D})$. We emphasize that this 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.

4 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 eventually depends on $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathscr{D})$.

Thm. 1 provides an explicit formula for $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathscr{D})$. 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 $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathscr{D})$. Therefore, understanding 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. A full analysis of these factors and their inter-relationships is a complex endeavor in itself, but several basic observations can be obtained in a relatively straightforward manner. Some simple examples illustrating expected and unexpected consequences of these observations will be provided in the following section.

We will base these observations on two sets of rough but conceptually simpler upper and lower bounds on $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathscr{D})$. These bounds are presented in Thm. 3 and Corollary 1 which follow, and highlight different aspects of this quantity. Since our main focus in this section is clarity rather than generality, we will allow ourselves to assume that the probability distribution $\mathscr{D}$ is supported in the unit ball of Euclidean space[2]. For the same reason, we have made no particular effort to make the bounds tight.

Theorem 3 *$\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathscr{D})$ is upper bounded by*

$$\left(4\sqrt{\frac{2}{\pi}}\frac{\sqrt{\lambda_{\max}(V)}}{\alpha\lambda_{\min}(\Gamma)}\right)\int_{\cup_{i,j}F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})d\mathbf{x}$$

where $\alpha := \min_{i\neq j}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$, and $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the smallest and largest eigenvalues of a matrix A.

[2] Relaxing or removing this assumption will only affect multiplicative constants, which might depend on the regularity conditions we have imposed on the probability density function $p(\cdot)$.

Also, $\widehat{instab}(\mathtt{A_k}, \mathscr{D})$ *is lower bounded by*

$$\left(\sqrt{\frac{2}{\pi}}\frac{\sqrt{\lambda_{\min}(V)}}{\lambda_{\max}(\Gamma)}\right)\int_{\cup_{i,j}F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})d\mathbf{x}$$

Corollary 1 $\widehat{instab}(\mathtt{A_k}, \mathscr{D})$ *is upper bounded by*

$$\left(\frac{13\sqrt{\overline{Vol}}}{\alpha\underline{Vol}-16knP}\right)\int_{\cup_{i,j}F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})d\mathbf{x},$$

where $\overline{Vol} := \max_i \int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x})d\mathbf{x}$ *denotes the largest cluster with respect to the clustering induced by* $\boldsymbol{\mu}$ *and the underlying distribution,* $\underline{Vol} := \min_i \int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x})d\mathbf{x}$ *denotes the smallest such cluster,* $P := \sup_{\cup_{i,j}F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})$ *denotes an upper bound on the probability density along the limit cluster boundaries, and* $\alpha := \min_{i\neq j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$ *is a lower bound on the distance between any two limit centroids.*

Also, $\widehat{instab}(\mathtt{A_k}, \mathscr{D})$ *is lower bounded by*

$$\left(\frac{\sqrt{\lambda_{\min}(V)}}{2\left(\overline{Vol}+16knP/\alpha\right)}\right)\int_{\cup_{i,j}F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})d\mathbf{x}.$$

We will start by considering Corollary 1. The first thing to notice is that the integral density along the cluster boundaries, $\int_{\cup_{i,j}F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})d\mathbf{x}$, seems to play an important role in determining the instability of a model. According to the upper bound, if the density along the cluster boundaries is zero, we get that $\widehat{\text{instab}}(\mathtt{A_k}, \mathscr{D}) = 0$, and thus any such model will be asymptotically considered as the most stable one. Moreover, the same bound implies that $\widehat{\text{instab}}(\mathtt{A_k}, \mathscr{D})$ will tend to be small even if the density along the cluster boundaries is small but not exactly zero. This means that clustering stability will tend to consider models with lower density along the cluster boundaries as more 'appropriate'.

A second observation that can be made is that when faced with two different choices of k, both of which with low density along the boundaries, both the upper and lower bounds in Corollary 1 will often tend to be larger for the bigger value of k. To see this, notice first that the cluster boundary area, $\cup_{i,j}F_{\boldsymbol{\mu},i,j}$, increases with the number of clusters. Also, if the clusters are reasonably balanced, we should expect $\overline{\text{Vol}}$ to scale down inversely with k, whereas $\sqrt{\lambda_{\min}(V)}$ scales down at a slower rate, especially in high dimensions. If P is small enough to make knP relatively negligible, these factors imply that the bounds in Corollary 1 will tend to be larger for the bigger value of k. To give a concrete and very simple example, consider a uniform distribution on the cube $[-1/\sqrt{n}, 1/\sqrt{n}]^n$, with $k_1 = 2^n$ and $k_2 = 2^{2n}$ (the example can be easily generalized). The optimal clustering for each k is a uniform grid partition of the cube (intuitively, we slice the cube once along each dimension for $k_1 = 2^n$, and 3 times for $k_2 = 2^{2n}$). To correspond to the regime of low density along the boundaries, suppose we slightly modify the distribution, by making the probability density at thin slivers around the optimal cluster boundaries to be very small. Obviously, this does not materially change the optimal clustering. Comparing how the elements in the bounds change as we move from k_1 to k_2, we have that $\overline{\text{Vol}}$ and $\underline{\text{Vol}}$ decrease by 2^n, α decreases by 2, $\sqrt{\lambda_{\min}(V)}$ decreases by $2\sqrt{2}$, and $\int_{\cup_{i,j}F_{\boldsymbol{\mu},i,j}} p(\mathbf{x})d\mathbf{x}$ increases by 2. As a result, we get that the upper bound in Corollary 1 increases by approximately $2^{n/2+2}$, and the lower bound increases by approximately $2^{n-1/2}$.

This observation matches a 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, for example, a very low boundary density and nice structure compared to all the competing models, this might overcome any inherent tendency of instability to increase with k. However, when this is not the case, normalization procedures might be called for, as in (Lange et al. 2004). Although one can argue that this phenomenon is exacerbated by finite sample effects (since the same sample size m is used to measure the clustering stability for different values of k), we see here that it relies on factors which do not depend on the sample size, and thus will not be resolved simply by scaling the sample size with k.

Turning to Thm. 3 allows us to see in which direction is clustering instability affected by the local geometry of the limit clustering in the solution space. Specifically, recall that Γ is the Hessian of the objective function at the limit clustering $\boldsymbol{\mu}$, and thus describes the local geometry of the objective function around that point. If $\boldsymbol{\mu}$ represents a shallow, ill-defined local optimum of the objective function, then we might expect the eigenvalues of Γ to be small. From Thm. 3, we see that this will tend to make $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathscr{D})$ larger. For the same reason, a deep and well-defined local optimum will tend to make $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathscr{D})$ smaller. Thus, clustering stability seems to take into account and penalize shallow and ill-defined local optimum in terms of the objective function, which is indeed often a sign of a mismatch between the model and the data.

Finally, it is important to emphasize that most of the observations above are concerned with tendencies, and have no pretensions to universality, in the sense that they apply for every possible clustering setting. In particular, as was recently pointed out and studied in (Ben-David and von Luxburg 2008), there definitely exist situations where the density along the cluster boundaries is *not* positively correlated with the model instability. Thus, these observations should be seen as aids in understanding what kind of assumptions clustering stability methods tend to make in choosing the most 'appropriate' model, rather than as universal assertions about their behavior.

5 Examples

To illustrate some of the observations from the previous section, we empirically evaluated the instability measure on a few simple toy examples, where everything is well controlled and easy to analyze. These examples consist of various mixtures of Gaussians, where the k-means algorithm (with 10 random initializations) was used as a basis to estimate the model stability for different values of k. 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 be a real empirical study of clustering stability.

First of all, we see that in all cases considered, the average value of $d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$ 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. If we would have eliminated the scaling by the square root of the sample size in the definition of $d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$, then we would have graphs which converge to zero for all values of k, but the ratio between them would have remained more or less constant.

The three leftmost columns demonstrate how, for these particular examples, the density along the cluster boundaries seem to play an important role in determining $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathscr{D})$.

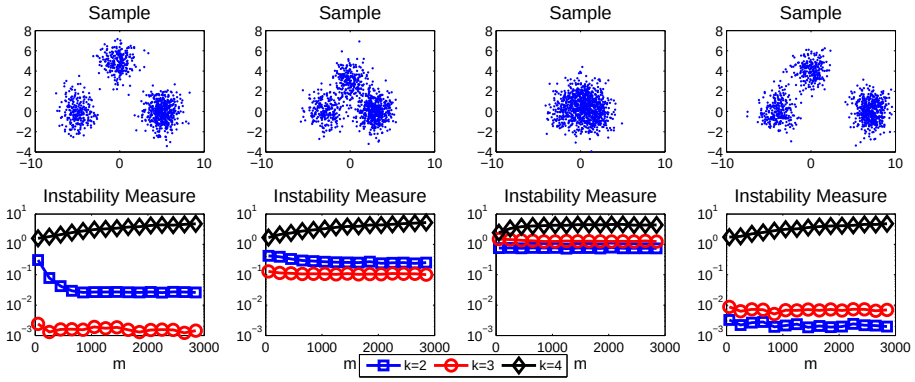

Fig. 2 Illustrative examples of the behavior of clustering stability. In each column, the upper plot is a representative sample from the underlying distribution (in all cases, a mixture of unit variance Gaussians in $\mathbb{R}^2$), while the lower plot is the average value of $d_{\mathscr{D}}^m(\mathtt{A}_k(S_1), \mathtt{A}_k(S_2))$ (empirically averaged over 1000 trials), for different sample sizes m.

In the two leftmost columns, $k = 3$ emerges as the most stable model, as 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, with higher densities in the boundaries between them. At some point, when the Gaussians become close enough, $k = 2$ becomes more stable than $k = 3$.

A different and more unexpected 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. 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 higher levels of the hierarchy, which may or may not be what we want.

6 A Negative Result on 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. For simplicity, we will also assume that we use an 'ideal' k-means algorithm which actually finds the global minimum of the objective function given a sample. The setting which we use to prove the theorem is simple enough so that the real k-means algorithm can be expected to have a similar behavior.

Theorem 4 *For any positive integer m_0, there exists a distribution $\mathscr{D}$ such that $d_{\mathscr{D}}^m(\mathtt{A}_3(S_1), \mathtt{A}_3(S_2))$ converges in probability to 0 as $m \to \infty$, but $\Pr(d_{\mathscr{D}}^m(\mathtt{A}_3(S_1), \mathtt{A}_3(S_2)) > \sqrt{m}/4)$ is at least $1/3$ for some $m \geq m_0$.*

The intuition behind the theorem is that for a suitably designed distribution, an arbitrarily large sample might be needed for the empirically derived clustering $\mathbf{c}$ to get 'close' to the limit clustering $\boldsymbol{\mu}$. As a result, for that setting and sample size, the central limit asymptotic behavior that we have analysed will be a poor approximation. However, it should be emphasized that the setting used in the theorem is highly artificial, and not necessarily typical of real-world clustering problems. Therefore, finding sufficient and empirically verifiable conditions which do allow finite sample guarantees is of much interest.

7 Proofs

7.1 Proof of Thm. 1

Before embarking on the proof, we briefly sketch its outline:

1. Using tools from the statistical theory of Z-estimators, we characterize the asymptotic Gaussian distribution of the cluster centroids $\mathbf{c}$, in terms of the underlying distribution $\mathscr{D}$ (Lemma 1). This result reproves the central limit theorem for k-means due to Pollard (Pollard 1982), but without requiring an algorithm capable of finding the global optimum of the k-means objective function.
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 $\mathscr{D}$, 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 $d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1),\mathtt{A}_\mathtt{k}(S_2))$, and its expected value.

For convenience, we shall use $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_k)$ to denote the random element $\mathbf{c}-\boldsymbol{\mu}$. Also, we will use the stochastic order notation $O_p(\cdot)$ and $o_p(\cdot)$ (cf. (van der Vaart and Wellner 1996)). 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 $\varepsilon > 0$ there exists a real number M such that $\Pr(\|X_m\| \geq M\|Y_m\|) < \varepsilon$ if m is large enough. We write $X_m = o_p(Y_m)$ to mean that $\Pr(\|X_m\| \geq \varepsilon\|Y_m\|) \to 0$ for each $\varepsilon > 0$. Notice that $\{Y_m\}$ may also be non-random. For example, $X_m = o_p(1)$ means that $X_m \to 0$ in probability.

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* $\mathbf{v} \sim \mathscr{N}\left(\mathbf{0}, \Gamma^{-1}V\Gamma^{-1}\right)$*. As a result,* $\|\boldsymbol{\epsilon}\| = O_p(1/\sqrt{m})$.

Since proving the lemma requires specific tools and additional notation which we will not need later on, we present the proof separately in the appendix. Notice that the lemma 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 multiplicative components and analyze them separately. We have that:

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

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

$$(\boldsymbol{\mu}_i - \mathbf{x}) \cdot \boldsymbol{\epsilon}_i - (\boldsymbol{\mu}_j - \mathbf{x}) \cdot \boldsymbol{\epsilon}_j + O(\|\boldsymbol{\epsilon}\|^2) \tag{11}$$

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

$$\begin{aligned}
&\frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \cdot (\mathbf{c}_i - \mathbf{c}_j)} = \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2 + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \cdot (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j)} \\
&= \frac{1}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| \left(1 + \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \cdot (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j)}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\right)} = \frac{1}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| (1 + O(\|\boldsymbol{\epsilon}\|))} \\
&= \frac{1}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|} \left(1 - \frac{O(\|\boldsymbol{\epsilon}\|)}{1 + O(\|\boldsymbol{\epsilon}\|)}\right) = \frac{1 + O(\|\boldsymbol{\epsilon}\|)}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|},
\end{aligned} \tag{12}$$

assuming $\|\boldsymbol{\epsilon}\|$ to be small enough. Multiplying Eq. (11) and Eq. (12) gives us the expression in the lemma. □

In order to calculate the asymptotic distribution of $d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$, we need to characterize the distribution of the probability mass of $\mathscr{D}$ 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 $H_{\boldsymbol{\mu},i,j}$, let $F \subseteq H_{\boldsymbol{\mu},i,j}$ be some finite intersection of half-spaces. 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*

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

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

$$\int_{F' \cap B} |\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| \sup_{\xi_{\mathbf{y}} \in [\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}), \tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})]} |p(\mathbf{y},\xi_{\mathbf{y}}) - p(\mathbf{y},0)| d\mathbf{y}. \qquad (14)$$

This integral can be upper bounded by

$$\sup_{\mathbf{y} \in F' \cap B} |\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| \sup_{\xi_{\mathbf{y}} \in [\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}), \tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})]} |p(\mathbf{y},\xi_{\mathbf{y}}) - p(\mathbf{y},0)| \int_{F' \cap B} 1 d\mathbf{y}. \qquad (15)$$

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

$$\sup_{\mathbf{y}\in F'\cap B} |\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})| = O(\|\boldsymbol{\epsilon}\| + \|\boldsymbol{\epsilon}\|^2), \tag{16}$$

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. (16) 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

$$\sup_{\mathbf{y}\in F'\cap B} \sup_{\xi_{\mathbf{y}}\in[\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}),\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})]} |p(\mathbf{y},\xi) - p(\mathbf{y},0)| = o_p(1). \tag{17}$$

Substituting Eq. (16) and Eq. (17) into Eq. (15), and using the fact that $\|\boldsymbol{\epsilon}\| = O_p(1/\sqrt{m})$, we get that the expression in Eq. (15) (and hence Eq. (14)) 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 $\mathscr{D}$ 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*

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

so that

$$t(\mathbf{c},\mathbf{c}',i,j) = \int_{Q(\mathbf{c},\mathbf{c}',i,j)} p(\mathbf{x})d\mathbf{x}. \tag{18}$$

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 $\mathscr{D}$ 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. (18) 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 $|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y}) - \tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})|$.

Also, let $F_{\min}(\boldsymbol{\mu},\mathbf{c},\mathbf{c}',i,j)$ be the subset of $F_{\max}(\boldsymbol{\mu},\mathbf{c},\mathbf{c}',i,j)$, such that any point $\mathbf{x}=(\mathbf{y},0)$ in it 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 exactly $|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})-\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})|$. Since it is formed from intersections of half-spaces, it is measurable and we can perform integration with respect to it.

For notational convenience, we will drop most of the parameters from now on, as they should be clear from the context. 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 definitions, by Fubini's theorem, we have that:

$$\int_{F'_{\max}}\left|\int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})}p(\mathbf{y},\xi)d\xi\right|d\mathbf{y}\geq\int_{Q}p(\mathbf{x})d\mathbf{x}\geq\int_{F'_{\min}}\left|\int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})}p(\mathbf{y},\xi)d\xi\right|d\mathbf{y}, \tag{19}$$

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. (19) 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. (19) 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

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

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

$$\int_{F'_{\max}\cap B^c}\left|\int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})} p(\mathbf{y},\xi)d\xi\right| d\mathbf{y} = O_p\left(h(R)/\sqrt{m}\right), \tag{20}$$

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 $|l(\mathbf{x},\mathbf{c}_i,\mathbf{c}'_j)|\le a(\|\mathbf{x}\|+1)(\|\boldsymbol{\epsilon}\|+\|\boldsymbol{\epsilon}'\|)/\sqrt{m}$ for some constant $a>0$, a very similar analysis reveals that

$$\int_{F'\cap B^c} p(\mathbf{y},0)|l(\mathbf{x},\mathbf{c}_i,\mathbf{c}'_j)|d\mathbf{y} = O_p\left(h(R)/\sqrt{m}\right). \tag{21}$$

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

$$\int_{F'_{\max}\cap B}\left|\int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})} p(\mathbf{y},\xi)d\xi\right| d\mathbf{y} = \int_{F'_{\max}\cap B}|\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})-\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}+o_p(1/\sqrt{m}). \tag{22}$$

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

$$\left|\int_{F'_{\max}\cap B}|\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})-\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}-\int_{F'\cap B}|\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})-\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}\right| = o_p(1/\sqrt{m}). \tag{23}$$

The l.h.s can be upper bounded by

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

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

$$\int_{(F'_{\max}\triangle F')\cap B}|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y} = o_p(1/\sqrt{m}). \tag{24}$$

By Lemma 2, inside the bounded domain of B, we have that $|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})|\le a\|\boldsymbol{\epsilon}\|$ 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

$$\int_{(F'_{\max}\triangle F')\cap B}|\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}\le a\|\boldsymbol{\epsilon}\|\int_{(F'_{\max}\triangle F')\cap B}1d\mathbf{y}, \tag{25}$$

Note that $\int_{(F'_{\max}\triangle F')\cap B}1d\mathbf{y}$ 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

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

Combining this with Eq. (25), and the fact that $\|\boldsymbol{\epsilon}\|=O_p(1/\sqrt{m})$, justifies Eq. (24), and hence Eq. (23). Combining Eq. (20), Eq. (22) and Eq. (23), we get that

$$\int_{F'_{\max}}\left|\int_{\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})}^{\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})} p(\mathbf{y},\xi)d\xi\right| d\mathbf{y} = \int_{F'\cap B}|\tilde{\ell}_{\boldsymbol{\epsilon}'}(\mathbf{y})-\tilde{\ell}_{\boldsymbol{\epsilon}}(\mathbf{y})|p(\mathbf{y},0)d\mathbf{y}+o_p(1/\sqrt{m})+O_p(h(R)/\sqrt{m}). \tag{26}$$

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. (26) 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. (21), 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. (20) and Eq. (21)). 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. (26) can be rewritten as

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

This gives us an equivalent formulation of the upper bound in Eq. (19). As discussed immediately after Eq. (19), 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. By definition, $d^m_{\mathscr{D}}(\mathtt{A}_k(S_1),\mathtt{A}_k(S_2))$ is equal to

$$\sum_{1\le i<j\le k} \sqrt{m}t(\mathbf{c},\mathbf{c}',i,j). \tag{27}$$

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

$$\int_{F_{\boldsymbol{\mu},i,j}} \frac{\sqrt{m}p(\mathbf{x})}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|} \left| \begin{pmatrix} \boldsymbol{\mu}_i-\mathbf{x} \\ \mathbf{x}-\boldsymbol{\mu}_j \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{\epsilon}_i-\boldsymbol{\epsilon}'_i \\ \boldsymbol{\epsilon}_j-\boldsymbol{\epsilon}'_j \end{pmatrix} \right| d\mathbf{x} + o_p(1). \tag{28}$$

By the continuous mapping theorem (van der Vaart and Wellner 1996), we have that $\sqrt{m}(\boldsymbol{\epsilon}_i-\boldsymbol{\epsilon}'_i,\boldsymbol{\epsilon}_j-\boldsymbol{\epsilon}'_j)^\top$ converges in distribution to $(\mathbf{v}_i-\mathbf{v}'_i,\mathbf{v}_j-\mathbf{v}'_j)^\top$, where $\mathbf{v},\mathbf{v}'$ are two independent copies of the random variable defined in Lemma 1. By standard results on the distribution of the difference of independent and identically distributed Gaussian random variables, this distribution is equal to that of $\sqrt{2}(\mathbf{v}_i,\mathbf{v}_j)^\top$. Moreover, it is not difficult to show that Eq. (28), ignoring the remainder term, is a continuous function of $\sqrt{m}(\boldsymbol{\epsilon}_i-\boldsymbol{\epsilon}'_i,\boldsymbol{\epsilon}_j-\boldsymbol{\epsilon}'_j)^\top$. 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

$$\int_{F_{\boldsymbol{\mu},i,j}} \frac{\sqrt{2}p(\mathbf{x})}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|} \left| \begin{pmatrix} \boldsymbol{\mu}_i-\mathbf{x} \\ \mathbf{x}-\boldsymbol{\mu}_j \end{pmatrix}^\top \begin{pmatrix} \mathbf{v}_i \\ \mathbf{v}_j \end{pmatrix} \right| d\mathbf{x}. \tag{29}$$

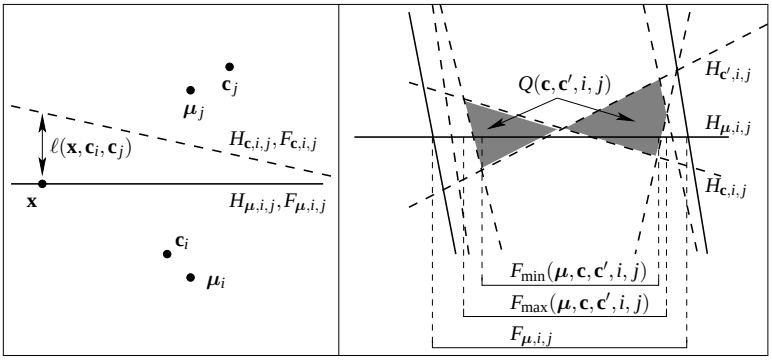

Fig. 3 An illustrative drawing of some of the notation and geometrical constructs used in the proof of Thm. 1. Solid lines represent cluster boundaries with respect to the optimal cluster centroids $\boldsymbol{\mu}$, while dashed lines represent cluster boundaries with respect to cluster centroids $\mathbf{c}$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

Substituting Eq. (29) into Eq. (27), 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. This is equal to For notational convenience.

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

By Fubini's theorem, this is equal to:

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

For notational convenience, denote $\Sigma=\Gamma^{-1}V\Gamma^{-1}$ as the covariance matrix of $\mathbf{v}$. The expression inside the expectation above is normally distributed, as a linear transformation of a normal random vector. Using standard results on the distribution of such transformations, 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

$$\frac{2}{\sqrt{\pi}}\sum_{1\le i<j\le k}\int_{F_{\boldsymbol{\mu},i,j}}\frac{p(\mathbf{x})}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}\sqrt{\begin{pmatrix}\boldsymbol{\mu}_i-\mathbf{x}\\ \mathbf{x}-\boldsymbol{\mu}_j\end{pmatrix}^\top\begin{pmatrix}\Sigma_{i,i} & \Sigma_{i,j}\\ \Sigma_{j,i} & \Sigma_{j,j}\end{pmatrix}\begin{pmatrix}\boldsymbol{\mu}_i-\mathbf{x}\\ \mathbf{x}-\boldsymbol{\mu}_j\end{pmatrix}}d\mathbf{x}.$$

The final form of $\widehat{\mathrm{instab}}(\mathtt{A}_\mathtt{k},\mathscr{D})$ is achieved by rewriting Σ as $(V^{1/2}\Gamma^{-1})^\top(V^{1/2}\Gamma^{-1})$, and simplifying.

7.2 Proof of Thm. 2

The proof is composed of several lemmas. The key insight is that the asymptotic distribution of $d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1),\mathtt{A}_\mathtt{k}(S_2))$, perhaps surprisingly, turns out to be a certain non-standard seminorm of a Gaussian random vector. Using theorems on seminorms of Gaussian measures allows us

to bound the probability of $d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1),\mathtt{A}_\mathtt{k}(S_2))$ 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 $d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1),\mathtt{A}_\mathtt{k}(S_2))$ is equal to that of $s(\mathbf{v})$, where $\mathbf{v}\sim\mathscr{N}(0,\Gamma^{-1}V\Gamma^{-1})$ and $s(\cdot)$ is a continuous seminorm 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 $d^m_{\mathscr{D}}(\mathtt{A}_\mathtt{k}(S_1),\mathtt{A}_\mathtt{k}(S_2))$ is equal to

$$\sqrt{2}\sum_{1\le i<j\le k}\int_{F_{\boldsymbol{\mu},i,j}}\frac{p(\mathbf{x})}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}\left|\begin{pmatrix}\boldsymbol{\mu}_i-\mathbf{x}\\ \mathbf{x}-\boldsymbol{\mu}_j\end{pmatrix}^\top\begin{pmatrix}\mathbf{v}_i\\ \mathbf{v}_j\end{pmatrix}\right|d\mathbf{x}, \tag{30}$$

where $\mathbf{v}$ is as defined in the lemma. It is quite straightforward to verify that the expression is indeed a seminorm on $\mathbf{v}$: homogeneity and the triangle inequality are immediate, and it is clear that Eq. (30) is always non-negative. As to continuity, fix some arbitrary $\mathbf{v}\in\mathbb{R}^n$, and let $\mathbf{v}+\boldsymbol{\epsilon}$ be a small perturbation of $\mathbf{v}$. Then we have that

$$\begin{aligned}
s(\mathbf{v}+\boldsymbol{\epsilon}) &= \sum_{1\le i<j\le k}\int_{F_{\boldsymbol{\mu},i,j}}\frac{\sqrt{2}p(\mathbf{x})}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}\left|\begin{pmatrix}\boldsymbol{\mu}_i-\mathbf{x}\\ \mathbf{x}-\boldsymbol{\mu}_j\end{pmatrix}^\top\begin{pmatrix}\mathbf{v}_i+\boldsymbol{\epsilon}_i\\ \mathbf{v}_j+\boldsymbol{\epsilon}_j\end{pmatrix}\right|d\mathbf{x}\\
&\le \sum_{1\le i<j\le k}\int_{F_{\boldsymbol{\mu},i,j}}\frac{\sqrt{2}p(\mathbf{x})}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}\left|\begin{pmatrix}\boldsymbol{\mu}_i-\mathbf{x}\\ \mathbf{x}-\boldsymbol{\mu}_j\end{pmatrix}^\top\begin{pmatrix}\mathbf{v}_i\\ \mathbf{v}_j\end{pmatrix}\right|d\mathbf{x}\\
&\quad+\sum_{1\le i<j\le k}\int_{F_{\boldsymbol{\mu},i,j}}\frac{\sqrt{2}p(\mathbf{x})}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}\left|\begin{pmatrix}\boldsymbol{\mu}_i-\mathbf{x}\\ \mathbf{x}-\boldsymbol{\mu}_j\end{pmatrix}^\top\begin{pmatrix}\boldsymbol{\epsilon}_i\\ \boldsymbol{\epsilon}_j\end{pmatrix}\right|d\mathbf{x}\\
&\le s(\mathbf{v})+\sum_{1\le i<j\le k}\int_{F_{\boldsymbol{\mu},i,j}}\frac{\sqrt{2}p(\mathbf{x})}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}\left\|\begin{pmatrix}\boldsymbol{\mu}_i-\mathbf{x}\\ \mathbf{x}-\boldsymbol{\mu}_j\end{pmatrix}^\top\right\|\|\boldsymbol{\epsilon}\|d\mathbf{x}\\
&\le s(\mathbf{v})+\|\boldsymbol{\epsilon}\|\sum_{1\le i<j\le k}\int_{F_{\boldsymbol{\mu},i,j}}\frac{\sqrt{2}p(\mathbf{x})}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|}(\|\boldsymbol{\mu}\|+2\|\mathbf{x}\|)d\mathbf{x},
\end{aligned}$$

which by the regularity conditions on $p(\cdot)$, is upper bounded by $s(\mathbf{v})+C\|\boldsymbol{\varepsilon}\|$ for some constant C. Therefore, we get that $s(\mathbf{v}+\boldsymbol{\epsilon})-s(\mathbf{v})\le C\|\boldsymbol{\varepsilon}\|$ for any $\mathbf{v},\boldsymbol{\epsilon}$. By an appropriate substitution, this immediately implies that $s(\mathbf{v})-s(\mathbf{v}+\boldsymbol{\epsilon})\le C\|\boldsymbol{\varepsilon}\|$ as well. Therefore, for any $\mathbf{v}$, $|s(\mathbf{v}+\boldsymbol{\epsilon})-s(\mathbf{v})|\le C\|\boldsymbol{\varepsilon}\|$, which converges to zero as $\boldsymbol{\varepsilon}\to 0$, hence $s(\cdot)$ is indeed continuous. □

Lemma 6 *Let $\mathbf{v}$ be a normally distributed random vector in $\mathbb{R}^n$, whose covariance matrix has full rank. Let $s(\cdot)$ be a seminorm on $\mathbb{R}^n$ which is not 0 by identity, 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 $M,\boldsymbol{\varepsilon}$ such that $Mb_\theta>1$ and $\boldsymbol{\varepsilon}a_\theta<1$, it holds that

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

and

$$\Pr(s(\mathbf{v})<\boldsymbol{\varepsilon}\mathbb{E}[s(\mathbf{v})])\le erf(erf^{-1}(\theta)a_\theta\boldsymbol{\varepsilon}).$$

Proof To prove the lemma, we will need two auxiliary results from the literature on Gaussian measures. For completeness, we present these two results below, in the form in which they apply to our setting. The first theorem, due to Borel, may be found as theorem III.3 in (Milman and Schechtman 1986). The second theorem is a direct implication of theorem 1 in (Latała and Oleszkiewicz 1999). A small note about notation: for any $A \subseteq \mathbb{R}^n$, and any scalar $t > 0$, we let $tA := \{\mathbf{x} : \mathbf{x}/t \in A\}$.

Theorem 5 (Borel) *Let* $\mathbf{v}$ *be a zero mean Gaussian random vector, and let* $A \subseteq \mathbb{R}^n$ *be a symmetric convex set such that* $\Pr(\mathbf{v} \in A) = \theta > 1/2$. *Then for any* $t > 1$,

$$\Pr(\mathbf{v} \notin tA) \leq \theta \left(\frac{1-\theta}{\theta} \right)^{(1+t)/2}.$$

Theorem 6 (Latała and Oleszkiewicz) *Let* $\mathbf{v}$ *be a zero mean Gaussian random vector, and let* $A \subseteq \mathbb{R}^n$ *be a symmetric convex closed set. For any* $x > 0$, *let* $n(x)$ *be the probability of a standard normal random variable to lie in* $[-x,x]$. *Let* $a > 0$ *be such that* $\Pr(\mathbf{v} \in A) = n(a)$. *Then for any* $0 \leq t \leq 1$,

$$\Pr(\mathbf{v} \in tA) \leq n(ta).$$

In the proof, we will apply the two theorems above on (closed) balls around the origin with respect to $s(\cdot)$, namely sets of the form $\{\mathbf{x} \in \mathbb{R}^n : s(\mathbf{x}) \leq a\}$ for some $a > 0$. The fact that these are symmetric and convex sets is immediate from the standard definition of a seminorm. In Lemma 5, we have also shown that $s(\cdot)$ is a continuous function from $\mathbb{R}^n$ to $\mathbb{R}$. Since the sets we are considering are pre-images, under the continuous function $s(\cdot)$, of closed sets of the form $[0,a] \subseteq \mathbb{R}$, we have that they are closed as well. This justifies our use of the two theorems above.

We now turn to the proof itself. Since $s(\mathbf{v})$ is a seminorm of a Gaussian random vector, its distribution function $F(t) = \Pr(s(\mathbf{v}) \leq t)$ is absolutely continuous, except possibly at the single point $\inf\{t \geq 0 | F(t) > 0\}$ (see for example (Hoeffman-Jørgensen et al. 1979)). Also, we assume that $\mathbf{v}$ is non-degenerate and $s(\cdot)$ is not identically zero, therefore $\Pr(s(\mathbf{v}) \leq t)$ is arbitrarily small for small enough $t > 0$. As a result, for any $\theta \in (1/2, 1)$, there is a corresponding positive parameter med_θ such that

$$\Pr(s(\mathbf{v}) \leq \mathrm{med}_\theta) = \theta.$$

Notice that the set $\{\mathbf{v} : s(\mathbf{v}) \leq \mathrm{med}_\theta\}$ is exactly a closed ball of radius med_θ around the origin, with respect to $s(\cdot)$. Applying Thm. 5 and Thm. 6, we get that

$$\Pr(s(\mathbf{v}) > M\mathrm{med}_\theta) \leq \theta \left(\frac{1-\theta}{\theta} \right)^{(1+M)/2} \tag{31}$$

$$\Pr(s(\mathbf{v}) \leq \varepsilon\mathrm{med}_\theta) \leq \mathrm{erf}(\mathrm{erf}^{-1}(\theta)\varepsilon). \tag{32}$$

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

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

Using Eq. (31), 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 $\mathbb{E}[s(\mathbf{v})] \leq \text{med}_\theta a_\theta$, where a_θ is as defined in the lemma.

In a similar manner, we can write $\mathbb{E}[s(\mathbf{v})]$ as

$$\int_0^\infty 1 - \Pr(s(\mathbf{v}) \leq t)dt = \text{med}_\theta \int_0^\infty 1 - \Pr(s(\mathbf{v}) \leq \varepsilon \text{med}_\theta)d\varepsilon,$$

which is lower bounded in term, using Eq. (32), by

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

Again by straightforward computations, we reach the conclusion that $\mathbb{E}[s(\mathbf{v})] \geq \text{med}_\theta b_\theta$, where b_θ is as defined in the lemma.

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

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

The other bound in the lemma is derived similarly. □

We can now turn to the proof of Thm. 2. By Lemma 5, both $d^m_{\mathscr{D}}(\mathtt{A}_{\mathtt{k_s}}(S_1), \mathtt{A}_{\mathtt{k_s}}(S_2))$ and $d^m_{\mathscr{D}}(\mathtt{A}_{\mathtt{k_u}}(S_1), \mathtt{A}_{\mathtt{k_u}}(S_2))$ converge in distribution to $s(\mathbf{v}_{k_u})$ and $s(\mathbf{v}_{k_s})$, 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,

$$\begin{aligned}
&\Pr(d^m_{\mathscr{D}}(\mathtt{A}_{\mathtt{k_u}}(S_1), \mathtt{A}_{\mathtt{k_u}}(S_2)) \leq 1.1 d^m_{\mathscr{D}}(\mathtt{A}_{\mathtt{k_s}}(S_1), \mathtt{A}_{\mathtt{k_s}}(S_2))) \\
&\leq \Pr(d^m_{\mathscr{D}}(\mathtt{A}_{\mathtt{k_u}}(S_1), \mathtt{A}_{\mathtt{k_u}}(S_2)) \leq c) + \Pr(1.1 d^m_{\mathscr{D}}(\mathtt{A}_{\mathtt{k_s}}(S_1), \mathtt{A}_{\mathtt{k_s}}(S_2)) \geq c) \\
&\leq \Pr(s(\mathbf{v}_{k_u}) \leq c) + \Pr(1.1 s(\mathbf{v}_{k_s}) \geq c) + o(1). \qquad (33)
\end{aligned}$$

We will first treat the simple case where $R = \infty$, corresponding to $\widehat{\text{instab}}(\mathtt{A}_{k_s}, \mathscr{D}) = 0$ and $\widehat{\text{instab}}(\mathtt{A}_{k_u}, \mathscr{D}) > 0$. In that case, we have that $s(\mathbf{v}_{k_s}) = 0$ with probability 1, whereas $s(\mathbf{v}_{k_u}) \leq c$ with arbitrarily small probability if we pick $c > 0$ small enough. As a result, the expression above is $o(1)$ as required.

Turning now to the case of $R < \infty$, note that the combination of Lemma 5 and Lemma 6 allows us to upper bound the probability that $s(\mathbf{v}_{k_u})$ is smaller than its expectation by a factor $\varepsilon < 1$, and upper bound the probability that $s(\mathbf{v}_{k_s})$ 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 $1.1M/\varepsilon \leq R$, where R is as defined in the lemma, we get that Eq. (33) above is upper bounded by

$$\theta_1 \left(\frac{1-\theta_1}{\theta_1}\right)^{((1+M)b_{\theta_1})/2} + \text{erf}(\text{erf}^{-1}(\theta_2)a_{\theta_2}\varepsilon) + o(1) \qquad (34)$$

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 $\theta_1 = 0.9$, $\theta_2 = 0.8$, $M = 2\log(R)/(b_{\theta_1}\log(\theta_1/(1-\theta_1)))$, $\varepsilon = 1.1M/R$, and using the fact that $\text{erf}(x) \leq (2/\sqrt{\pi})x$ for any $x \geq 0$, we get that Eq. (33) is upper bounded by $(0.3 + 3\log(R))/R + o(1)$ for any $R > 3$.

Assume the event

$$d^m_{\mathscr{D}}(\mathtt{A}_{\mathtt{k_u}}(S_1),\mathtt{A}_{\mathtt{k_u}}(S_2)) > 1.1 d^m_{\mathscr{D}}(\mathtt{A}_{\mathtt{k_s}}(S_1),\mathtt{A}_{\mathtt{k_s}}(S_2)), \tag{35}$$

occurs. Recall that the quantities in Eq. (35) depend on the unknown underlying distribution $\mathscr{D}$, 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 $\hat{\theta}_{k_u,3m}$ and $\hat{\theta}_{k_s,3m}$. Thus, even if Eq. (35) occurs, it is still possible that $\hat{\theta}_{k_u,3m} \leq \hat{\theta}_{k_s,3m}$, and we wish to upper bound the probability for this occuring.

Notice that conditioned on the event in Eq. (35), $\hat{\theta}_{k_u,3m}$ and $\hat{\theta}_{k_s,3m}$ are nothing more than plug-in estimators of $d^m_{\mathscr{D}}(\mathtt{A}_{\mathtt{k_u}}(S_1),\mathtt{A}_{\mathtt{k_u}}(S_2))/\sqrt{m}$ and $d^m_{\mathscr{D}}(\mathtt{A}_{\mathtt{k_s}}(S_1),\mathtt{A}_{\mathtt{k_s}}(S_2))/\sqrt{m}$ respectively, based on an i.i.d sample of size m. Since these quantities decrease with m, a standard Hoeffding bound would not apply. However, we have by Thm. 2 in (Shamir and Tishby 2007) that concentration of measure still occurs, and the probability that $\hat{\theta}_{k_u,3m} \leq \hat{\theta}_{k_s,3m}$, conditioned on the event in Eq. (35), is $o(1)$ (namely, converges to 0 as $m \to \infty$). Therefore, the probability that Eq. (35) 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.

7.3 Proof of Thm. 3

Letting A be some real symmetric matrix, we will use $\|A\|$ to denote its operator norm.

We will start by bounding $\Psi(\mathbf{x},i,j)$ in the definition of $\widehat{\mathrm{instab}}(\mathtt{A}_{\mathtt{k}},\mathscr{D})$ (Eq. (10)). Since both V and Γ are real symmetric matrices, we have that

$$\begin{aligned}
\Psi(\mathbf{x},i,j) &= \left\| \begin{pmatrix} V_i^{1/2} & 0 \\ 0 & V_j^{1/2} \end{pmatrix} \begin{pmatrix} (\Gamma^{-1})_{i,i} & (\Gamma^{-1})_{i,j} \\ (\Gamma^{-1})_{j,i} & (\Gamma^{-1})_{j,j} \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}_i - \mathbf{x} \\ \mathbf{x} - \boldsymbol{\mu}_j \end{pmatrix} \right\| \\
&\leq \left\| \begin{pmatrix} V_i^{1/2} & 0 \\ 0 & V_j^{1/2} \end{pmatrix} \right\| \left\| \begin{pmatrix} (\Gamma^{-1})_{i,i} & (\Gamma^{-1})_{i,j} \\ (\Gamma^{-1})_{j,i} & (\Gamma^{-1})_{j,j} \end{pmatrix} \right\| \left\| \begin{pmatrix} \boldsymbol{\mu}_i - \mathbf{x} \\ \mathbf{x} - \boldsymbol{\mu}_j \end{pmatrix} \right\| \\
&= \sqrt{\lambda_{\max}(V)} \lambda_{\max}(\Gamma^{-1}) \left\| \begin{pmatrix} \boldsymbol{\mu}_i - \mathbf{x} \\ \mathbf{x} - \boldsymbol{\mu}_j \end{pmatrix} \right\| = \frac{\sqrt{\lambda_{\max}(V)}}{\lambda_{\min}(\Gamma)} \left\| \begin{pmatrix} \boldsymbol{\mu}_i - \mathbf{x} \\ \mathbf{x} - \boldsymbol{\mu}_j \end{pmatrix} \right\|.
\end{aligned}$$

Substituting this into the definition of $\widehat{\mathrm{instab}}(\mathtt{A}_{\mathtt{k}},\mathscr{D})$ in Eq. (9), we get an upper bound of the form

$$\begin{aligned}
&\frac{2}{\sqrt{\pi}} \sum_{1\leq i<j\leq k} \left[\int_{F_{\boldsymbol{\mu},i,j}} \frac{\sqrt{\lambda_{\max}(V)}}{\lambda_{\min}(\Gamma)} \frac{\left\| \begin{pmatrix} \boldsymbol{\mu}_i - \mathbf{x} \\ \mathbf{x} - \boldsymbol{\mu}_j \end{pmatrix} \right\|}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|} p(\mathbf{x}) d\mathbf{x} \right] \\
&\qquad\qquad \leq \frac{2}{\sqrt{\pi}} \sum_{1\leq i<j\leq k} \left[\int_{F_{\boldsymbol{\mu},i,j}} \frac{\sqrt{\lambda_{\max}(V)}}{\lambda_{\min}(\Gamma)} \frac{2\sqrt{2}}{\alpha} p(\mathbf{x}) d\mathbf{x} \right],
\end{aligned}$$

where $\alpha = \min_{i,j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$, and the last transition is due to the assumption that the distribution is supported in the unit ball (hence we can assume that $\|\mathbf{x}\|, \|\boldsymbol{\mu}_i\|, \|\boldsymbol{\mu}_j\|$ are all at most 1). Simplifying, we get the upper bound in the theorem.

Turning to the lower bound and repeating the same technique, we get that

$$\Psi(\mathbf{x},i,j) \geq \frac{\sqrt{\lambda_{\min}(V)}}{\lambda_{\max}(\Gamma)} \left\| \begin{pmatrix} \boldsymbol{\mu}_i - \mathbf{x} \\ \mathbf{x} - \boldsymbol{\mu}_j \end{pmatrix} \right\|.$$

Substituting this into the definition of $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathscr{D})$ in Eq. (9), we get a lower bound of the form

$$\frac{2}{\sqrt{\pi}} \sum_{1 \leq i < j \leq k} \left[\int_{F_{\boldsymbol{\mu},i,j}} \frac{\sqrt{\lambda_{\min}(V)}}{\lambda_{\max}(\Gamma)} \frac{\left\| \begin{pmatrix} \boldsymbol{\mu}_i - \mathbf{x} \\ \mathbf{x} - \boldsymbol{\mu}_j \end{pmatrix} \right\|}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|} p(\mathbf{x}) d\mathbf{x} \right].$$

Since $\|(\boldsymbol{\mu}_i - \mathbf{x}, \mathbf{x} - \boldsymbol{\mu}_j)\|$ is minimized for $\mathbf{x} = (\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)/2$, we can lower bound the expression above by

$$\frac{2}{\sqrt{\pi}} \sum_{1 \leq i < j \leq k} \left[\int_{F_{\boldsymbol{\mu},i,j}} \frac{\sqrt{\lambda_{\min}(V)}}{\lambda_{\max}(\Gamma)} \frac{1}{\sqrt{2}} p(\mathbf{x}) d\mathbf{x} \right].$$

Simplifying, we get the lower bound in the theorem.

7.4 Proof of Corollary 1

Thm. 3 gave us upper and lower bounds on $\widehat{\text{instab}}(\mathtt{A}_\mathtt{k}, \mathscr{D})$ in terms of the maximal and minimal eigenvalues of V and Γ. From this, we will derive Corollary 1 by bounding these eigenvalues.

We will start by upper bounding $\lambda_{\max}(V)$. Since V is a block diagonal matrix, we have that $\lambda_{\max}(V) = \max_{i \in [k]} \lambda_{\max}(V_i)$, where V_i is block i in V (see Eq. (8)). By the definition of V_i and a straightforward application of the Cauchy-Schwartz inequality, we have that

$$\lambda_{\max}(V_i) = \max_{\mathbf{y}:\|\mathbf{y}\|=1} \mathbf{y}^\top V_i \mathbf{y} \leq 4 \left(\max_{\mathbf{x} \in C_{\boldsymbol{\mu},i}} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \right) \int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x}) d\mathbf{x} \leq 16 \int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x}) d\mathbf{x}.$$

and therefore

$$\lambda_{\max}(V) \leq 16 \max_i \int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x}) d\mathbf{x} \tag{36}$$

We now wish to lower bound $\lambda_{\min}(\Gamma)$. By the definition of Γ (Eq. (6) and Eq. (7)), it can be decomposed as the difference of two matrices J and N. J is a $kn \times kn$ diagonal matrix, composed of k segments of the form

$$\left(\int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x}) d\mathbf{x} \right) I_n,$$

I_n being the unit matrix of size $n \times n$. N is a $kn \times kn$ block matrix, composed of $k \times k$ blocks. Each block (i,j) is of the form

$$N_{i,i} := \sum_{a \neq i} \frac{2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_a\|} \int_{F_{\boldsymbol{\mu},i,a}} p(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^\top d\mathbf{x}$$

and for $i \neq j$ it is

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

By Weyl's theorem (cf. (Horn and Johnson 1985)), we have that

$$\lambda_{\min}(\Gamma) = \lambda_{\min}(J-N) \geq \lambda_{\min}(J) - \lambda_{\min}(N) \geq \lambda_{\min}(J) - \rho(N),$$

where $\rho(N)$ is the spectral radius of N. Since J is a diagonal matrix, its eigenvalues correspond to the elements on the diagonal, and therefore $\lambda_{\min}(J)$ is at least $\min_i \int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x})d\mathbf{x}$. As to $\rho(N)$, since the spectral radius lower bounds any matrix norm, we can upper bound $\rho(N)$ by, say, the maximum column sum matrix norm of N (defined as $\max_j \sum_{i=1}^{kn} |n_{i,j}|$, where $n_{i,j}$ are the entries of N). From the definition of N and the assumption of the distribution supported on the unit ball, this norm for N can be (very roughly) upper bounded by $16knP/\alpha$, where $P = \sup_{\cup_{i,j} F_{i,j}} p(\mathbf{x})$ and $\alpha = \min_{i,j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$. As a result, we get that

$$\lambda_{\min}(\Gamma) \geq \min_i \int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x})d\mathbf{x} - 16knP/\alpha.$$

Finally, we turn to upper bound $\lambda_{\max}(\Gamma)$. Again applying Weyl's theorem, we have that

$$\lambda_{\max}(\Gamma) = \lambda_{\max}(J-N) \leq \lambda_{\max}(J) + \lambda_{\max}(N) \leq \lambda_{\max}(J) + \rho(N).$$

As we have seen above, we can roughly upper bound $\rho(N)$ by $16kn\alpha$, and we can upper bound $\lambda_{\max}(J)$ by $\max_i \int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x})d\mathbf{x}$. As a result, we get that

$$\lambda_{\max}(\Gamma) \leq \max_i \int_{C_{\boldsymbol{\mu},i}} p(\mathbf{x})d\mathbf{x} + 16knP/\alpha.$$

Substituting the bounds we have derived above into Thm. 3 and simplifying gives us the corollary.

7.5 Proof of Thm. 4

To prove the theorem, we will borrow a setting discussed in (Linder 2002) for a different purpose.

Let Δ be some small positive constant (say $\Delta < 0.1$). Consider the parameterized family of distributions $\{D_\varepsilon\}$ (where $\varepsilon \in (0,1/4)$) on the real line, which assigns probability mass $(1-\varepsilon)/4$ to $x=-1$ and $x=-1-\Delta$, and $(1+\varepsilon)/4$ 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 $d^m_{\mathscr{D}_\varepsilon}(\mathtt{A}_\mathtt{k}(S_1), \mathtt{A}_\mathtt{k}(S_2))$ converges to 0 in probability, since the boundary points between the optimal clusters have zero density.

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

By Slud's inequality (see (Anthony and Bartlet 1999)), 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 $A^1_{m,\varepsilon}$ is equal to the probability of a success rate of more than half in m Bernoulli trials, whose probability of success is $(1-\varepsilon)/2$. Using the theorem above, we get after a few straightforward algebraic manipulations that

$$\Pr(A^1_{m,\varepsilon}) \geq 1 - \Phi\left(\frac{4}{\sqrt{m}} + 2\varepsilon\sqrt{m}\right). \tag{37}$$

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

$$\Pr(A^2_{m,\varepsilon}) \geq 1/2. \tag{38}$$

Finally, it is straightforward to show that $\Pr(B_{m,\varepsilon})$ is arbitrarily close to 1 uniformly for any ε, if m is large enough. Combining this with Eq. (37), Eq. (38) 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

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

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 $\mathscr{D}_\varepsilon$, then the probability that event $A^1_{m,\varepsilon'} \cap B_{m,\varepsilon}$ occurs for one sample, and $A^2_{m',\varepsilon} \cap B_{m,\varepsilon}$ 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

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

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.

8 Conclusions

In this paper, we analyzed the method of clustering stability for model order selection in the k-means framework, based on an explicit characterization of its asymptotic behavior. We concluded that this method does not 'break down' in the large sample regime, in the sense that even when the sample size goes to infinity and the model becomes stable for any choice of k, these stability estimators still tell us something meaningful, rather than just returning random noise. Based on these results, we made some observations on the factors which may tend to make a model 'stable' or 'unstable'. Such observations are particularly useful for understanding what kind of assumptions are implicitly made, when one uses the clustering stability method. These factors appear to constitute 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 section 5.

Although it is possible to extend some of the results presented here to more general clustering frameworks, beyond *k*-means (see (Shamir and Tishby 2008b)), the most obvious challenge is to extend our analysis from the asymptotic domain to the finite sample size domain. Showing that clustering stability does not 'break down' in the large sample regime might have 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. 4, additional assumptions are needed for such results.

9 Appendix - Proof of Lemma 1

As discussedd in Sec. 7, we devote this appendix to reprove Pollard's central limit theorem for *k*-means (Pollard 1982) under the weaker assumption that the algorithm we work with is non-ideal, and might return locally optimal solutions. This is achieved by using more modern tools from empirical process theory, and specifically from the area of *Z-estimators*. Intuitively, a Z-estimator is any statistical estimator, which works by trying to zero a function or a set of functions based on a sample. For example, suppose that m instances are drawn $i.i.d$ from some distribution on $\mathbb{R}$. Then the sample mean can be seen as a Z-estimator: given a sample $x_1,\ldots,x_m$, it returns a value $\hat{\theta}$ which zeros the function $\Lambda_m(\theta)=\sum_{i=1}^{m}(\theta-x_i)$. For a full formal treatment of Z-estimators, see (van der Vaart and Wellner 1996). To prove the lemma, we will apply a general central limit theorem for Z-estimators, due to van der Vaart. This result (which appears for instance as Thm. 3.3.1 in (van der Vaart and Wellner 1996)) is very general, and we will quote it below as applied to the specific setting where both the data and the hypothesis class reside in Euclidean spaces. In particular, this allows us to ignore some technical conditions which hold trivially in a finite-dimensional setting.

Theorem 7 (van der Vaart) *Let $\{\Lambda_m\}_{m=1}^{\infty}$ and Λ be a sequence of random maps and a fixed map, respectively, between a subset Θ of some Euclidean space, into some other Euclidean space. Let $(\mathbf{c}^m)_{m=1}^{\infty}$ be a sequence of random vectors, which satisfy $\Lambda_m(bc_m)=0$ for all m. Assume that as $m\to\infty$,*

$$\frac{\|\sqrt{m}(\Lambda_m(\mathbf{c}^m)-\Lambda(\mathbf{c}^m))-\sqrt{m}(\Lambda_m(\boldsymbol{\mu})-\Lambda(\boldsymbol{\mu}))\|}{1+\sqrt{m}\|\mathbf{c}-\boldsymbol{\mu}\|}\to 0 \tag{39}$$

in probability, and that the sequence $\sqrt{m}(\Lambda_m(\boldsymbol{\mu})-\Lambda(\boldsymbol{\mu}))$ converges in distribution to a vector-valued random variable Z. Furthermore, assume that $\Lambda(\cdot)$ is differentiable at $\boldsymbol{\mu}$ with an invertible derivative $\dot{\Lambda}_{\boldsymbol{\mu}}$. If $\Lambda(\boldsymbol{\mu})=0$, and $(\mathbf{c}^m)$ converges in probability to $\boldsymbol{\mu}$, then $\sqrt{m}(\mathbf{c}-\boldsymbol{\mu})$ converges in distribution to $-\dot{\Lambda}_{\boldsymbol{\mu}}^{-1}Z$,

We will apply the theorem where $\mathbf{c}^m$ is the set of centroids returned by the algorithm based on a random sample, and $\boldsymbol{\mu}$ is the limit set of clustering to which we converge. We will drop the m superscript when it is obvious from context.

The first step will be to cast the *k*-means algorithm as a Z-estimator, using a construct which appears in (Pollard 1982). For this, define for any $i\in[k]$ the following function from $\mathbb{R}^{nk}\times\mathbb{R}^n$ to $\mathbb{R}^n$:

$$\Delta_i(\mathbf{c},\mathbf{x}):=\begin{cases}2(\mathbf{c}_i-\mathbf{x}) & \mathbf{x}\in C_{\mathbf{c},i}\\ \mathbf{0} & \text{otherwise}\end{cases}$$

The factor of 2 is not really necessary, but would be convenient later for directly citing certain results from (Pollard 1982) without the need to convert constants.

Furthermore, assuming $\mathbf{x}_1,\ldots,\mathbf{x}_m$ is a sample drawn i.i.d from $\mathscr{D}$, define the random map $\Lambda_m(\cdot) = (\Lambda_m^1(\cdot),\ldots,\Lambda_m^k(\cdot))$ and the deterministic map $\Lambda(\cdot) = (\Lambda^1(\cdot),\ldots,\Lambda^k(\cdot))$ as

$$\Lambda_m^i(\mathbf{c}) := \frac{1}{m}\sum_{j=1}^{m}\Delta_i(\mathbf{c},\mathbf{x}_j) \quad , \quad \Lambda^i(\mathbf{c}) := \int_{\mathbb{R}^n}\Delta_i(\mathbf{c},\mathbf{x})p(\mathbf{x})d\mathbf{x}.$$

for any $i \in [k]$.

The key insight is that given an empirical sample of size m, our k-means clustering algorithm always returns a solution of $\mathbf{c}$ such that $\Lambda_m(\mathbf{c}) = 0$. This is a consequence of the fact that in a k-means clustering, each centroid lies at the center of mass of its respective cluster. It can be easily verified that such a solution zeros $\Lambda_m(\cdot)$. Thus, the k-means algorithm can indeed be viewed as a certain type of Z-estimator.

With this construct in hand, we need to verify that the conditions of Thm. 7 indeed hold. Proving that $\Lambda(\mathbf{c})$ is differentiable, and deriving its form, is a purely technical exercise in multivariate calculus, which may be found as lemma C in (Pollard 1982). The resulting matrix is exactly Γ, which we have defined in Eq. (6) and Eq. (7). Showing that this is in fact the Hessian of the k-means objective function is also proven in (Pollard 1982).

Thus, the only thing really left to show is that Eq. (39) indeed holds in our case. Notice that it is enough to show that

$$\frac{\|\sqrt{m}(\Lambda_m^i(\mathbf{c}) - \Lambda^i(\mathbf{c})) - \sqrt{m}(\Lambda_m^i(\boldsymbol{\mu}) - \Lambda^i(\boldsymbol{\mu}))\|}{1+\sqrt{m}\|\mathbf{c}-\boldsymbol{\mu}\|} \to 0$$

for any $i \in [k]$. A relatively simple sufficient condition for this (implied by lemma 3.3.5 in (van der Vaart and Wellner 1996)) is the following: For any cluster $i \in [k]$, any coordinate $j \in \{1,\ldots,n\}$, set of centroids $\mathbf{c}$, and instance $\mathbf{x}$, let $\Delta_i^j(\mathbf{c},\mathbf{x})$ be the projection of $\Delta_i(\mathbf{c},\mathbf{x})$ on its j-th coordinate. Then for Eq. (39) to hold, it is sufficient to show that for some $\delta > 0$, any $i \in [k]$, and any coordinate $j \in \{1,\ldots,n\}$, the set of functions

$$\{\Delta_i^j(\mathbf{c},\cdot) - \Delta_i^j(\boldsymbol{\mu},\cdot)\}_{\|\mathbf{c}-\boldsymbol{\mu}\|<\delta}$$

is a *Donsker class*. Intuitively, a set of real functions $\{f(\cdot)\}$ (with any probability distribution $\mathscr{D}$) is called Donsker if it satisfies a uniform central limit theorem. Without getting too much into the details, this means that if we sample m elements i.i.d from $\mathscr{D}$, then $(f(\mathbf{x}_1)+\ldots+f(\mathbf{x}_m))/\sqrt{m}$ converges in distribution (as $m\to\infty$) to a Gaussian random variable, and the convergence is uniform over all $f(\cdot)$ in the set, in an appropriately defined sense.

We use the fact that if $\mathscr{F} = \{f(\cdot)\}$ and $\mathscr{G} = \{g(\cdot)\}$ are Donsker classes, then so is $\mathscr{F}\cdot\mathscr{G} = \{f(\cdot)g(\cdot)\}$, and that any subset of a Donsker class is also Donsker (see section 2.10 in (van der Vaart and Wellner 1996)). This allows us to reduce the problem to showing that for any i, j, the following two function classes, from $\mathbb{R}^n$ to $\mathbb{R}$, are Donsker:

$$\{c_{i,j} - x_j\}_{\|\mathbf{c}-\boldsymbol{\mu}\|<\delta} \quad , \quad \{\mathbf{1}_{C_{\mathbf{c},i}}(\cdot)\}_{\|\mathbf{c}-\boldsymbol{\mu}\|\leq\delta}. \tag{40}$$

The first class is composed of linear functions with bounded offsets in Euclidean space, which is well known to be Donsker. The second class is composed of indicator functions for any possible cluster in a clustering induced by $\mathbf{c}$ close enough to $\boldsymbol{\mu}$. Since each cluster is composed of a finite intersection of half-spaces, this class is known to have finite VC-dimension, and hence is also Donsker. These and related results can be found for instance in (Dudley 1999).

Thus, we have shown that for the settings assumed in our theorem, Eq. (39) holds. We now return to deal with the other ingredients required to apply Thm. 7.

Considering the asymptotic distribution of $\sqrt{m}(\Lambda_m(\boldsymbol{\mu}) - \Lambda(\boldsymbol{\mu}))$, since $\Lambda(\boldsymbol{\mu}) = 0$ by assumption, we have that for any $i \in [k]$, it is equal to

$$\left(\sqrt{m}\Lambda_m^1(\boldsymbol{\mu}), \ldots, \sqrt{m}\Lambda_m^k(\boldsymbol{\mu})\right) = \left(\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\Delta_1(\boldsymbol{\mu}, \mathbf{x}_j), \ldots, \frac{1}{\sqrt{m}}\sum_{j=1}^{m}\Delta_k(\boldsymbol{\mu}, \mathbf{x}_j)\right). \qquad (41)$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_m$ is the sample by which Λ_m is defined. The r.h.s of Eq. (41) is a sum of identically distributed, independent random variables with zero mean and bounded variance (by the assumptions on the underlying distribution), normalized by $\sqrt{m}$. As a result, by the standard central limit theorem, each $\sqrt{m}(\Lambda_m^i(\boldsymbol{\mu}) - \Lambda^i(\boldsymbol{\mu}))$ converges in distribution to a zero mean Gaussian random vector, with covariance matrix

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

Moreover, it is easily verified from the definitions that $\mathrm{Cov}(\Delta_i(\boldsymbol{\mu}, \mathbf{x}), \Delta_{i'}(\boldsymbol{\mu}, \mathbf{x})) = 0$ for any $i \neq i'$. Therefore, $\sqrt{m}(\Lambda_m - \Lambda)(\boldsymbol{\mu})$ converges in distribution to a zero mean Gaussian random vector, whose covariance matrix V is composed of k diagonal blocks $(V_1, \ldots, V_k)$, all other elements of V being zero.

Applying Thm. 7, we now get that $\sqrt{m}(\mathbf{c} - \boldsymbol{\mu})$ converges in distribution to $\Gamma^{-1}Z$, where $Z \sim \mathcal{N}(0, V)$. This asymptotic distribution can also be written as $\mathcal{N}(0, \Gamma^{-1}V\Gamma^{-1})$.

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

References

M. Anthony and P. Bartlet. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

S. Ben-David and U. von Luxburg. Relating clustering stability to properties of cluster boundaries. In *Proceedings of the Twenty First Annual Conference on Computational Learning Theory*, 2008.

S. Ben-David, U. von Luxburg, and D. Pál. A sober look at clustering stability. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, pages 5–19, 2006.

S. Ben-David, D. Pál, and H.-U. Simon. Stability of k-means clustering. In *Proceedings of the Twentieth Annual Conference on Computational Learning Theory*, pages 20–34, 2007.

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

A. Bertoni and G. Valentini. Model order selection for biomolecular data clustering. *BMC Bioinformatics*, 8 ((Suppl 2):S7), 2007.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, second edition, 2001.

R. Dudley. *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1999.

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

J. Hartigan. *Clustering Algorithms*. Wiley, 1975.

J. Hoeffman-Jørgensen, L. A. Shepp, and R. Dudley. On the lower tail of gaussian seminorms. *The Annals of Probability*, 7(2):319–342, 1979.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

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.

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

R. Latała and K. Oleszkiewicz. Gaussian measures of dilatations of convex symmetric sets. *Annals of Probability*, 27(4):1922–1938, 1999.

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

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.

V. D. Milman and G. Schechtman. *Asymptotic Theory of Finite Dimensional Normed Spaces*. Springer, 1986.

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

P. Radchenko. *Asymptotics Under Nonstandard Conditions*. PhD thesis, Yale University, 2004.

O. Shamir and N. Tishby. Cluster stability for finite samples. In *Advances in Neural Information Processing Systems 21*, 2007.

O. Shamir and N. Tishby. Model selection and stability in *k*-means clustering. In *Proceedings of the Twenty First Annual Conference on Computational Learning Theory*, 2008a.

O. Shamir and N. Tishby. On the reliability of clustering stability in the large sample regime. In *Advances in Neural Information Processing Systems 22*, 2008b.

M. Smolkin and D. Ghosh. Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, 36(4), 2003.

D. Steinley. K-means clustering: A half-century synthesis. *British journal of mathematical & statistical psychology*, 59(1):1–34, 2006.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes : With Applications to Statistics*. Springer, 1996.