On the Reliability of Clustering Stability in the Large Sample Regime

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

Abstract

Clustering stability is an increasingly popular family of methods for performing model selection in data clustering. The basic idea is that the chosen model should be stable under perturbation or resampling of the data. Despite being reasonably effective in practice, these methods are not well understood theoretically, and present some difficulties. In particular, when the data is assumed to be sampled from an underlying distribution, the solutions returned by the clustering algorithm will usually become more and more stable as the sample size increases. This raises a potentially serious practical difficulty with these methods, because it means there might be some hard-to-compute sample size, beyond which clustering stability estimators 'break down' and become unreliable in detecting the most stable model.

In this paper, we provide a set of general sufficient conditions, which ensure the reliability of clustering stability estimators in the large sample regime. In contrast to previous work, which concentrated on specific toy distributions or specific idealized clustering frameworks, here we make no such assumptions. We then exemplify how these conditions apply to several important families of clustering algorithms, such as maximum likelihood clustering, certain types of kernel clustering, and centroid-based clustering with any Bregman divergence. In addition, we explicitly derive the non-trivial asymptotic behavior of these estimators, for any framework satisfying our conditions. This may help us understand what is considered a 'stable' model by these estimators, at least for large enough samples.

1 Introduction

Clustering stability is an increasingly popular approach for performing model selection in data clustering, such as determining the number of clusters in the data ([4],[6],[13],[8], [5]). The basic idea is that a 'correct' model for the data is a stable model, in the sense that perturbing the dataset in some manner should not change the clustering too much. Suitably quantified and estimated, the amount of instability can be compared for different models, in order to choose the most appropriate one. In this paper, we will focus on sampling based estimators, where the perturbation is performed by resampling the data. These estimators work by drawing and clustering different random subsets of the data, and estimating the dissimilarity of the different clusterings induced on the data space.

Despite being relatively successful in practice, these methods are still not well understood theoretically. An important and non-trivial difficulty with these methods is the following observation, made and rigorously analyzed in [3], [2] and also pointed out in [7]. If we assume that the data is sampled from some underlying distribution, and the clustering algorithm works by optimizing some sort of objective function, then as the sample size increases, the algorithm will often converge to a single solution, which is a global or local optimum with respect to the underlying distribution, no matter which model we choose. Therefore, for large enough samples we might get a stable solution regardless of the chosen model. As a result, it is quite possible that there exists some sample size (intimately related to the particular underlying data distribution and hence difficult to compute), beyond which stability estimators will 'break down' and become unreliable. Namely, the estimated amount of instability will be close to zero for any model chosen, with differences between the models depending just on random and meaningless sampling artifacts.

A possible solution to this problem was introduced in [11], where an analysis of a suitably scaled version of the instability estimators indicated that they might actually be able to discern the stability of different models, no matter how large is the sample and how stable the models become. However, this was proven to hold only on a single toy example, as a proof of concept. In [12], this was extended to any continuous distribution on $\mathbb{R}^n$ using an 'ideal' k-means algorithm (capable of returning a global minimizer of the k-means objective function). However, these results still do not give us *general* sufficient conditions for the reliability of clustering stability estimators in the large sample regime.

In this paper, we present such a set of conditions, without tying ourselves to a particular clustering framework. The main condition is the existence of a *central limit theorem* for the clustering framework, in an appropriately defined sense. We then exemplify how this is fulfilled in several important and practical families of clustering algorithms, such as maximum likelihood clustering, certain types of kernel clustering, and centroid-based clustering with any kind of Bregman divergence. As a useful byproduct, we are able to explicitly characterize the asymptotic value to which these stability estimators, suitably scaled, converge, for any framework satisfying our conditions. A similar characterization was given in [12] for the k-means framework, but here the result is much more general, and the analytical expression is simpler. We expect this to lead to useful insights on the assumptions inherent in using clustering stability estimators for model selection.

Due to lack of space, the full proofs of our theorems are presented in a separate supplementary.

2 Problem Setting and Notation

We assume that objects to be clustered belong to a measurable subset $\mathcal{X}$ of $\mathbb{R}^n$, which for simplicity shall be assumed to be compact (i.e. closed and bounded). We assume there exists a distribution $\mathcal{D}$ with a density function $p(\cdot)$ on $\mathcal{X}$, from which we sample our data. Clustering is performed on such samples by an algorithm $\mathtt{A_k}$, parameterized by the number of clusters k, and in general corresponding to the clustering model whose stability we wish to estimate.

We assume that the algorithm returns a measurable function $f_{\hat{\boldsymbol{\theta}}} : \mathcal{X} \mapsto \mathbb{R}^k$, $f_{\hat{\boldsymbol{\theta}}} = (f_{\hat{\boldsymbol{\theta}},1}, \ldots, f_{\hat{\boldsymbol{\theta}},k})$, which is parameterized by some parameter space Θ to which $\hat{\boldsymbol{\theta}}$ belongs. This function measures the amount of 'association' or 'dis-association' of each instance in $\mathcal{X}$ to each of the k clusters. For example, in centroid based clustering, this function can be the distance of each instance from each of the centroids, and Θ is a subset of $\mathbb{R}^{nk}$, representing the concatenation of the k centroids determining the clustering. In mixture model estimation, $f_{\hat{\boldsymbol{\theta}}}(\cdot)$ can be the probability of an instance to belong to each of the mixture elements, and Θ represents the parameters of the mixture model. For simplicity, we will assume unless stated otherwise that $f_{\hat{\boldsymbol{\theta}}}$ is a measure of association. To prevent ambiguity, we assume that clusters are numbered according to some uniform canonical ordering (such as sorting with respect to the centroid coordinates in centroid based clustering).

In this paper, we shall assume for simplicity that the parameter space Θ can be framed as an open subset of some finite dimensional Euclidean space, with the regular Euclidean norm. This is not really essential for proving the sufficiency of our conditions. However, some of our sufficient conditions become less obvious in an infinite dimensional, non-parametric settings, and might fail to hold in some cases (see Subsec. 4.2 for further details).

Let $C_{\hat{\boldsymbol{\theta}},i}$ be the 'i-th cluster' with respect to $f_{\hat{\boldsymbol{\theta}}}(\cdot)$. Namely, the subset of $\mathcal{X}$ which is most associated with cluster i:

$$C_{\hat{\boldsymbol{\theta}},i} := \left\{ \mathbf{x} \in \mathcal{X} : \underset{a \in \{1,\ldots,k\}}{\operatorname{argmax}} f_{\hat{\boldsymbol{\theta}},a}(\mathbf{x}) = i \right\}.$$

For any two clusters $i \neq j$, define $F_{\hat{\boldsymbol{\theta}},i,j}$ to be the boundary between the two clusters, with respect to $\hat{\boldsymbol{\theta}}$. Formally:

$$F_{\hat{\boldsymbol{\theta}},i,j} := \left\{ \mathbf{x} \in \mathcal{X} : \underset{a \in \{1,\ldots,k\}}{\operatorname{argmax}} f_{\hat{\boldsymbol{\theta}},a}(\mathbf{x}) = \{i,j\} \right\}.$$

$f_{\hat{\boldsymbol{\theta}}}(\cdot)$, as an output of our clustering algorithm, is a random function based on the randomness of the sample used by the algorithm. In understanding the asymptotic behavior of this function, we will need to use the concept of a *Gaussian process* $G(\cdot)$ indexed by $\mathcal{X}$. In our context, this refers to a collection of random vectors in $\mathbb{R}^k$ ($G(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$), defined on the same probability space, such that any finite subset of them has a joint multivariate Gaussian distribution. In particular, $G(\mathbf{x})$ for any $\mathbf{x}$ is a Gaussian random vector in $\mathbb{R}^k$.

A typical clustering stability method relies on a measure of ’distance’ between clusterings derived from two independent samples. The prototypical measure we shall focus on is the mass of $\mathcal{D}$ which switches between clusters, when we compare different clusterings derived from independent samples. In order to treat all the different clustering frameworks in a unified manner, we assume that for soft clustering, this mass is measured after assigning each element in $\mathcal{X}$ to its most likely cluster. Formally, let S_1 and S_2 be two independent samples of size m, drawn i.i.d from $\mathcal{D}$, to which we apply the clustering algorithm $\mathtt{A_k}$. Assume that $\mathtt{A_k}(S_1)$ returns $f_{\hat{\boldsymbol{\theta}}}(\cdot)$, and $\mathtt{A_k}(S_2)$ returns $f_{\hat{\boldsymbol{\theta}}'}(\cdot)$ for some $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}' \in \Theta$. Then the distance between the two clusterings is defined as

$$d_{\mathcal{D}}(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2)) := \Pr_{\mathbf{x}\sim\mathcal{D}} \left(\underset{i}{\operatorname{argmax}} f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x}) \neq \underset{i}{\operatorname{argmax}} f_{\hat{\boldsymbol{\theta}}',i}(\mathbf{x}) \right).$$

We note that this definition differs a little from the one used in some previous literature on the topic (such as [3],[11]), since it is easier to work with[1]. Assuming the underlying data distribution is not unusually symmetric, our clustering algorithm will converge in probability to some fixed optimal solution $\boldsymbol{\theta}_0$ as the sample size m increases [3]. Therefore, both $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}'$ in the equation above will converge to the same $\boldsymbol{\theta}_0$. As a result, the random variable $d_{\mathcal{D}}(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2))$ will tend to become smaller and smaller, converging in probability to 0 as $m \to \infty$. As discussed in the introduction, if this convergence to zero behavior occurs for different model choices, then it might ultimately become impossible to reliably distinguish between the models in terms of their stability. This is because the distance measure as defined above will always be very close to zero, and it is possible that relative differences in estimates of these measures will depend more on random sampling artifacts than any true underlying characteristic of the model. In this paper, we wish to investigate when will the reliability of clustering stability estimators be maintained, despite this convergence to zero phenomenon.

3 Sufficient Conditions for the Reliability of Clustering Stability Estimators

In this section, we will present a set of conditions on the clustering framework, and prove that whenever these conditions are fulfilled, clustering stability estimators remain reliable in the large sample regime, in an appropriately defined sense. These essentially consist of a consistency and central limit condition for $\hat{\boldsymbol{\theta}}$ (the solution returned by the algorithm based on a random sample), plus some regularity conditions to prevent various pathologies (plausibly not the tightest possible). These regularity conditions are usually a simple consequence of the specific clustering framework that we are dealing with, plus the central limit condition. In the next section, we will see examples for well known clustering frameworks which seem to satisfy these conditions in general.

To avoid making things too complex, the regularity conditions will be presented in a somewhat informal manner, sometimes in a way which is a bit stronger than what is really needed. A fully rigorous and precise formulation of the regularity conditions may be found in the supplementary material to this paper.

[1] In previous work, the clustering distance measure was defined as the probability over a *pair* of instances drawn independently from $\mathcal{D}$, that this pair is in the same cluster under one clustering, and in different clusters under the other clustering. The two definitions are very closely related and in fact inter-convertible in some cases.

Conditions. *The following conditions shall be assumed to hold:*

1. ***Consistency Condition:*** *$\hat{\boldsymbol{\theta}}$ converges in probability (over drawing and clustering a sample of size m, $m \to \infty$) to some $\boldsymbol{\theta}_0 \in \Theta$. Furthermore, the association of clusters to indices $\{1, \ldots, k\}$ is constant in some neighborhood of $\boldsymbol{\theta}_0$.*

2. ***Central Limit Condition:*** *$\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a multivariate zero mean Gaussian random vector Z.*

3. ***Regularity Conditions****:*

   (a) *$f_{\boldsymbol{\theta}}(\mathbf{x})$* ***is Sufficiently Smooth:*** *The function $f_{\boldsymbol{\theta}}(\mathbf{x})$ is sufficiently smooth with respect to $\mathbf{x}$ and $\boldsymbol{\theta}$.*

   (b) ***Limit Cluster Boundaries are Reasonably Nice:*** *For any i, j, $F_{\boldsymbol{\theta}_0,i,j}$ is a differentiable hyper-surface (not necessarily a hyperplane), with a neighborhood in which the underlying density function $p(\cdot)$ is continuous. Moreover, the gradient of cluster association, $\nabla(f_{\boldsymbol{\theta}_0,i}(\cdot) - f_{\boldsymbol{\theta}_0,j}(\cdot))$, has positive magnitude everywhere on $F_{\boldsymbol{\theta}_0,i,j}$.*

   (c) ***Intersections and Edges of Cluster Boundaries are Relatively Negligible:*** *For any limit cluster boundary $F_{\boldsymbol{\theta}_0,i,j}$, and small enough $\epsilon > 0$, most of the volume in an ϵ-neighborhood around $F_{\boldsymbol{\theta}_0,i,j}$ is bounded away from the boundary's edges or other cluster boundaries.*

   (d) ***Minimal Parametric Stability:*** *The position of each cluster boundary $F_{\hat{\boldsymbol{\theta}},i,j}$ depend in a sufficiently smooth manner on $\hat{\boldsymbol{\theta}}$, at least in some small neighborhood of $\boldsymbol{\theta}_0$.*

We note that the conditions allow us to characterize the asymptotic Gaussian distribution of $\sqrt{m}(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) - f_{\boldsymbol{\theta}_0}(\mathbf{x}))$ for any $\mathbf{x} \in \mathcal{X}$, as formalized in the following proposition.

Proposition 1. *Conditions 2 and 3a above imply that $\sqrt{m}(f_{\hat{\boldsymbol{\theta}}}(\cdot) - f_{\boldsymbol{\theta}_0}(\cdot))$ converges in distribution to a Gaussian process $G(\cdot) = (G_1(\cdot), \ldots, G_k(\cdot))$ on $\mathbb{R}^k$, indexed by $\mathcal{X}$, of the form*

$$G(\mathbf{x}) := \big(\frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{x})\big)^{\top} Z.$$

In order to prove that stability doesn't 'break down' in the large sample regime, we have to model how instability estimators work. The basic building block is straightforward: cluster two independent samples and estimate the distance $d_{\mathcal{D}}(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2))$ between them. Since $\mathcal{D}$ is unknown, we cannot compute $d_{\mathcal{D}}(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2))$ directly, since we don't know what is the exact mass of $\mathcal{D}$ which switched between clusters. However, we can estimate it by using another sample, and calculating the percentage of sample points which switched between clusters. In practice, this entire process is repeated a number of times on different random subsets of the data, with the averaged result returned as the estimate of instability for the model chosen.

Our central result, stated in the theorem below, implies that these kind of estimators do not 'break down' in the large sample regime, in an appropriately defined sense.

Theorem 1. *Define the clustering stability estimator $\hat{\eta}^k_{m,q}$ as follows. Given an i.i.d sample of size at least $m(2q+1)$, split it randomly into q disjoint subsample pairs $\{S^1_i, S^2_i\}_{i=1}^q$ of size m each, and an additional sample S^3 of size m. Cluster all pairs $\{S^1_i, S^2_i\}_{i=1}^q$ with algorithm $\mathtt{A_k}$, resulting in a set of function pairs $\{f_{\hat{\boldsymbol{\theta}}_i}(\cdot), f_{\hat{\boldsymbol{\theta}}'_i}(\cdot)\}_{i=1}^q$. Return the estimate:*

$$\frac{1}{q}\sum_{i=1}^{q}\frac{1}{m}\sum_{\mathbf{x}\in S^3}\mathbf{1}\left(\operatorname*{argmax}_{j} f_{\hat{\boldsymbol{\theta}}_i,j}(\mathbf{x}) \neq \operatorname*{argmax}_{j} f_{\hat{\boldsymbol{\theta}}'_i,j}(\mathbf{x})\right).$$

If $f_{\hat{\boldsymbol{\theta}}}(\cdot)$ is a measure of dis-association rather than association, argmax *should be replaced above by* argmin. *With the set of conditions defined earlier, we have that for any $\epsilon > 0$,*

$$\lim_{m\to\infty} \Pr\left(\left|\sqrt{m}\,\hat{\eta}^k_{m,q} - \widehat{instab}(\mathtt{A_k}, \mathcal{D})\right| > \epsilon\right) = o(q^{-1/2}),$$

where

$$\widehat{instab}(\mathtt{A_k}, \mathcal{D}) = \frac{2}{\sqrt{\pi}}\sum_{1\le i<j\le k}\int_{F_{\boldsymbol{\theta}_0,i,j}} \frac{p(\mathbf{x})\sqrt{Var(G_i(\mathbf{x}) - G_j(\mathbf{x}))}}{\|\nabla(f_{\boldsymbol{\theta}_0,i}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,j}(\mathbf{x}))\|}d\mathbf{x}.$$

The theorem implies that if we take q to be large enough and see what happens to $\sqrt{m}\ \hat{\eta}^k_{m,q}$ in the large sample regime, it will return estimates which are reliably close to $\widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})$. If $\widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})$ differs for different model choices for $\mathtt{A_k}$, then we will be able to discern the more 'stable' model with high probability over the sampling process, simply by comparing the value of $\hat{\eta}^k_{m,q}$ for each model. This is despite the fact that without scaling by the square root of the sample size, $\hat{\eta}^k_{m,q}$ converges to zero in probability.

The theorem gives us more than that: it explicitly characterizes to what value our clustering instability estimator, after scaling, converges[2]. It should be emphasized that this value is not very dependent on the exact form of our estimator, at least in terms of the main factors appearing there. In a nutshell, the asymptotic instability of a clustering model is simply the integral, across all limit cluster boundaries, of the product of three quantities:

- The underlying probability density $p(\cdot)$ along the cluster boundaries. Therefore, high density along the cluster boundaries contributes to more instability.
- The variance of the Gaussian process characterizing the asymptotic fluctuations of the random function $f_{\hat{\boldsymbol{\theta}},i}(\cdot) - f_{\hat{\boldsymbol{\theta}},j}(\cdot)$. By definition, the association of a point $\mathbf{x}$ to cluster i or cluster j under the limit clustering $f_{\boldsymbol{\theta}_0}$ is determined by the sign of $f_{\boldsymbol{\theta}_0,i}(\mathbf{x}) - f_{\boldsymbol{\theta}_0,j}(\mathbf{x})$. Thus, the larger are these fluctuations, the larger are the fluctuations of the cluster boundaries, which contributes to more instability.
- The inverse of the gradient of the fixed limit function $f_{\boldsymbol{\theta}_0}(\cdot)$ across the cluster boundaries. A large gradient for $f_{\boldsymbol{\theta}_0,i}(\cdot) - f_{\boldsymbol{\theta}_0,j}(\cdot)$ at the boundary points implies that its value changes abruptly as we cross the cluster boundary. In this case, random fluctuations of $\hat{\boldsymbol{\theta}}$ around $\boldsymbol{\theta}_0$ will correspond to relatively smaller fluctuations of the cluster boundary, and this contributes to less instability.

For example, if for some model choice, the probability density $p(\mathbf{x})$ is exactly zero at the boundaries, then $\widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})$ equals zero. Namely, the clustering stability estimator asymptotically considers a model with zero density at the boundaries as the most stable possible. Also, as k increases, integration is performed over a larger area, and as a result we might get more instability, a fact which has been noted empirically [8]. It should be emphasized that the three quantities described above are not independent, and it is possible that a change in one quantity will be offset by a reciprocal change in another quantity. Thus, an exact analysis is non-trivial, and the observations above should be viewed as preliminary.

4 When do the Conditions Hold?

In this section, we investigate when does the set of conditions from subsection 3 hold. Recall that these conditions can be divided into 3 parts: a consistency condition, a central limit condition, and some regularity conditions to ensure that the framework is well behaved.

The most basic requirement for the problem we are dealing with to be meaningful at all is consistency. Namely, we assume that as the sample size increases to infinity, the clustering returned by the algorithm will converge to some limit clustering, and the distance between independent clusterings converge to zero in probability. If this does not happen, the problem of clustering stability potentially 'breaking down' in the large sample regime is irrelevant.

On the other hand, the central limit condition is the most non-trivial, and cannot simply be assumed to hold. Therefore, the focus of this subsection will be on exemplifying clustering frameworks where the central limit condition holds. As to the other regularity conditions, they are relatively mild, and are mostly a simple consequence of the specific clustering framework that we are considering, plus the central limit condition. Due to lack of space, we will not further discuss them here.

We will present three theorems, which demonstrate that the central limit condition holds on some important families of clustering frameworks. These families do not cover all interesting clustering

[2] It is reassuring to note that the formula for $\widehat{\text{instab}}(\mathtt{A_k}, \mathcal{D})$ can be shown to be a strict generalization of the one appearing in [12], which applied to k-means only, after taking care of the slightly different definition of $d_{\mathcal{D}}(\mathtt{A_k}(S_1), \mathtt{A_k}(S_2))$ there.

algorithms, but should be sufficient evidence that clustering stability provably does not 'break down' in many situations.

A convenient framework in which to derive such central limit results is the statistical theory 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 $\Psi_m(\theta) = \sum_{i=1}^{m}(\theta - x_i)$. A more general and relevant example is when one attempts to find the 'maximum likelihood' parameters of a probabilistic model, by searching for a solution which zeroes the derivative of the likelihood function on the data. For a full formal treatment of Z-estimators, see [14]. Proving central limit theorems for Z-estimators is a well studied topic. Therefore, if we manage to define our clustering algorithms as Z-estimators, we can apply known results on such estimators to get central limit theorems.

4.1 Bregman Divergence Clustering

The first case we shall consider is centroid clustering based on Bregman divergences. This family of clustering algorithms has received growing attention in the past few years (see for example [1] and references therein). The idea is to perform centroid-based clustering (where clusters are represented by centroids, and each data instance is associated with its 'nearest' centroid), but instead of using just the regular squared Euclidean distance, we use any kind of Bregman divergence. A Bregman divergence $d_\Upsilon(\cdot,\cdot)$ between two vectors in $\mathbb{R}^n$ is defined as

$$d_\Upsilon(\mathbf{x}, \mathbf{y}) := \Upsilon(\mathbf{x}) - \Upsilon(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\Upsilon(\mathbf{y})\rangle,$$

where $\Upsilon(\cdot)$ is any differentiable, strictly convex real function defined on a closed convex set in $\mathbb{R}^n$. A large variety of distances and divergence measures are Bregman divergences. These include squared Euclidean distance, Mahalanobis distances, information theoretic divergences such as KL-divergence and I-divergence, and quite a few others (see [1] for further details).

In the framework that we consider, the goal of clustering is to find a set of centroids in $\mathbb{R}^n$, such that the average divergence between each instance and its closest centroid is minimized. Namely, given a sample $\mathbf{x}_1, \ldots, \mathbf{x}_m$, we seek a set of k centroids $\mathbf{c}_1, \ldots, \mathbf{c}_k$ which minimizes the objective function

$$\frac{1}{m}\sum_{i=1}^{m} \min_j d_\Upsilon(\mathbf{x}_i, \mathbf{c}_j). \tag{1}$$

The parameter space Θ of possible clusterings can be thought of as a subset of $\mathbb{R}^{nk}$, such that any $\boldsymbol{\theta} \in \Theta$ is simply the concatenation of the k centroids in $\mathbb{R}^n$, by some canonical ordering. We will use the notation $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$, so that $\boldsymbol{\theta}_i \in \mathbb{R}^n$ is the centroid corresponding to the i-th cluster.

We assume that our clustering algorithm attempts to minimize Eq. (1). For concreteness, the algorithm may use an iterative scheme similar to k-means, as in [1]: in each step, points are assigned to the nearest centroid (with respect to the Bregman divergence used), and the centers are updated so as to minimize the average divergence between them and the instances in the cluster. *We will not need to assume that a globally optimal solution is found*: convergence in probability to a locally optimal solution (with respect to the underlying distribution, as the sample size increases to infinity) will suffice.

The first step will be to cast our clustering algorithm as a Z-estimator, using a generalization of an idea from [9]. For this, define for any $i \in \{1, \ldots, k\}$ the following function from $\Theta \times \mathbb{R}^n$ to $\mathbb{R}^n$:

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

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

$$\Psi_m^i(\boldsymbol{\theta}) := \frac{1}{m}\sum_{j=1}^{m} \Delta_i(\boldsymbol{\theta}, \mathbf{x}_j) \quad , \quad \Psi^i(\boldsymbol{\theta}) := \int_{\mathcal{X}} \Delta_i(\boldsymbol{\theta}, \mathbf{x}) p(\mathbf{x}) d\mathbf{x}.$$

The key insight is that given an empirical sample of size m, our Bregman Clustering algorithm always returns a solution of $\hat{\boldsymbol{\theta}}$ such that $\Psi_m(\hat{\boldsymbol{\theta}}) = 0$. This is a consequence of the fact, proven as Proposition 1 in [1], that the optimum location for a centroid in any Bregman Clustering framework lies at the center of mass of its cluster, and our algorithm indeed returns a set of centroids which fulfill this condition. It can be easily verified that such a solution zeros $\Psi_m(\cdot)$. Thus, our algorithm can indeed be viewed as a Z-estimator, and it is possible to prove a central limit behavior.

Theorem 2. *Let* $\mathtt{A_k}$ *be a centroid-based clustering algorithm using any Bregman divergence, which returns a solution* $\hat{\boldsymbol{\theta}}$ *such that* $\Psi_m(\hat{\boldsymbol{\theta}}) = 0$. *As* $m \to \infty$, *assume that* $\hat{\boldsymbol{\theta}}$ *converges in probability to a fixed* $\boldsymbol{\theta}_0 \in \Theta$, *such that* $\Psi(\boldsymbol{\theta}_0) = 0$, *and* $\Psi(\cdot)$ *has a non-singular derivative at* $\boldsymbol{\theta}_0$ *with a continuous inverse. Then* $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ *converges in distribution to a zero mean Gaussian random vector.*

We note that the form of the derivative of $\Psi(\cdot)$ at $\boldsymbol{\theta}_0$ depends on the specific clustering algorithm that we are considering. For example, an explicit calculation for the k-means framework has been carried out in [9], and the analysis can be generalized to other frameworks.

4.2 Kernel Clustering

The second example we shall examine is kernel clustering. In particular, we will focus on the kernel k-means algorithm [10]. Recall that kernel k-means can be seen as implicitly transforming our data points into a high or even infinite dimensional reproducing kernel Hilbert space $\mathcal{H}$, and performing regular k-means in that space: points are assigned to their nearest centroids in $\mathcal{H}$ (with respect to the norm induced by the inner product), and the centroids are updated to represent the mean of the points in each cluster. These operations are performed via a kernel function, which corresponds to the inner product operation in $\mathcal{H}$.

In this case, we can think of Θ as $\mathcal{H}^k$, where $\mathcal{H}$ is the reproducing kernel Hilbert space associated with the kernel. Thus, each element in Θ represent a set of k centroids in $\mathcal{H}$. Kernel k-means always returns a solution where the centroids are at the center of mass of their respective clusters. Thus, to cast this algorithm as a Z-estimator, we can simply reuse the functions Ψ, Ψ_m defined in Subsec. 4.1, only this time these are functions from $\mathcal{H}^k$ to $\mathcal{H}^k$. Formally, let $\phi : \mathcal{X} \mapsto \mathcal{H}$ be the feature map from the data space to the reproducing kernel Hilbert space induced by the kernel, and define $\Delta_i(\boldsymbol{\theta}, \mathbf{x}), \Psi_m(\boldsymbol{\theta}), \Psi(\boldsymbol{\theta})$ exactly as in Subsec. 4.1, replacing $\mathbf{x}$ everywhere with $\phi(\mathbf{x})$.

Following the framework of this paper, we will assume in the theorem below that Θ (and hence $\mathcal{H}$) is a subset of a finite dimensional Euclidean space, corresponding for example to polynomial kernels. We note however that the tools we use allow us to prove, in some cases, a similar theorem for infinite dimensional $\mathcal{H}$ as well. Unfortunately, these tools are not applicable in some important infinite-dimensional kernel clustering frameworks, such as those employing universal kernels (see the full proof for further details). Once again, we do not need to assume that the algorithm returns a globally optimal solution, only that it consistently converges to some local optimum.

Theorem 3. *Let* $\mathtt{A_k}$ *be a finite dimensional kernel* k*-means clustering algorithm, which returns a solution* $\hat{\boldsymbol{\theta}}$ *such that* $\Psi_m(\hat{\boldsymbol{\theta}}) = 0$. *Assume that as* $m \to \infty$, $\hat{\boldsymbol{\theta}}$ *converges in probability to a fixed* $\boldsymbol{\theta}_0 \in \Theta$, *such that* $\Psi(\boldsymbol{\theta}_0) = 0$, *and* $\Psi(\cdot)$ *has a non-singular derivative at* $\boldsymbol{\theta}_0$ *with a continuous inverse. Then* $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ *converges in distribution to a zero mean Gaussian random vector.*

4.3 Maximum Likelihood Clustering

The next theorem treats estimators which are based on a statistical modeling of the clusters, and attempt to maximize the log-likelihood of the data given the model. Namely, we assume that Θ is a parametric family of distributions, and attempt to maximize $\sum_{i=1}^{m} \log(q(\mathbf{x}|\hat{\boldsymbol{\theta}}))/m$, where $q(\mathbf{x}|\hat{\boldsymbol{\theta}})$ is the probability of the observation $\mathbf{x}$ given the model $\hat{\boldsymbol{\theta}}$. In this case, we assume that $f_{\hat{\boldsymbol{\theta}}}(\mathbf{x})$ corresponds to a vector in the k-simplex, so that $f_{\hat{\boldsymbol{\theta}},i}(\mathbf{x})$ is the probability that $\mathbf{x}$ belongs to cluster i. Θ is the parameter space of the distributional family that we are considering. For example, Θ can encode the mean, covariance matrix and relative weight of each component in a Gaussian mixture model. The asymptotic normality of maximum likelihood estimators is a classical and well-known result. Here we present a variant (not the strongest possible), which emphasizes the fact that the returned solution need not be globally optimal, as long as consistency is maintained and a locally optimal solution is returned.

Theorem 4. *Let $\mathtt{A_k}$ be a statistical model estimator, based on a parametric family of distributions Θ, where Θ is an open subset of some Euclidean space. Given an i.i.d sample $\mathbf{x}_1, \ldots, \mathbf{x}_m$, the algorithm returns a solution $\hat{\boldsymbol{\theta}} \in \Theta$ which locally maximizes the log-likelihood function with respect to the sample. Assume that as $m \to \infty$, $\hat{\boldsymbol{\theta}}$ converges in probability to a fixed local maximizer $\boldsymbol{\theta}_0 \in \Theta$ of the log-likelihood function with respect to the underlying distribution $\mathcal{D}$, namely $\int_{\mathcal{X}} \log(q(\mathbf{x}|\hat{\boldsymbol{\theta}}))p(\mathbf{x})d\mathbf{x}$. Assume that the latter is twice differentiable with respect to $\hat{\boldsymbol{\theta}}$ at some neighborhood of $\boldsymbol{\theta}_0$, with a non-singular second derivative at $\boldsymbol{\theta}_0$. Then $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a zero mean multivariate Gaussian random variable.*

5 Conclusions and Future Work

In this paper, we investigated the reliability of clustering stability estimators in the large sample regime. We provided a set of sufficient conditions, revolving around a central limit requirement, which ensure that these estimators do not ’break down’ even for arbitrarily large samples, and characterized their asymptotic behavior. In contrast to previous work, the results are general and do not assume a particular clustering framework or distribution. We also discussed several families of well-known clustering algorithms, for which the main central limit requirement holds in general. These include maximum likelihood clustering, certain types of kernel clustering, and centroid-based clustering with any Bregman divergence.

Although the central limit approach proved to be a convenient framework, it remains an open question how far it is from being *necessary* for stability estimators not to ’break down’ in the large sample regime. This question is relevant because a central limit requirement might be too strong to hold over all useful clustering frameworks. In addition, we still need to better understand the meaning of the asymptotic value of clustering instability as given in Thm. 1, beyond extreme cases such as zero density along cluster boundaries. Hopefully, this will also help to understand the behavior of clustering stability on small samples.

Acknowledgements: The authors wish to thank Yonatan Harpaz for help with the proof of Thm. 1.

References

[1] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[2] 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.

[3] 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.

[4] 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.

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

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

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

[8] 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.

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

[10] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

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

[12] 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*, 2008.

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

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