ESAIM: PS 16 (2012) 436–452
DOI: 10.1051/ps/2012013

ESAIM: Probability and Statistics
www.esaim-ps.org

HOW THE INITIALIZATION AFFECTS THE STABILITY OF THE *k*-MEANS ALGORITHM *

Sébastien Bubeck[1], Marina Meilă[2] and Ulrike von Luxburg[3]

Abstract. We investigate the role of the initialization for the stability of the k-means clustering algorithm. As opposed to other papers, we consider the actual k-means algorithm (also known as Lloyd algorithm). In particular we leverage on the property that this algorithm can get stuck in local optima of the k-means objective function. We are interested in the actual clustering, not only in the costs of the solution. We analyze when different initializations lead to the same local optimum, and when they lead to different local optima. This enables us to prove that it is reasonable to select the number of clusters based on stability scores.

Mathematics Subject Classification. 62F12.

Received June 5, 2011. Revised May 2, 2012.

1. Introduction

Stability is a popular tool for model selection in clustering, in particular to select the number k of clusters. The general idea is that the best parameter k for a given data set is the one which leads to the "most stable" clustering results. While model selection based on clustering stability is widely used in practice, its behavior is still not well-understood from a theoretical point of view. A recent line of papers discusses clustering stability with respect to the k-means criterion in an idealized setting [2–4, 11–13]. It is assumed that one has access to an ideal algorithm which can globally optimize the k-means criterion. For this perfect algorithm, results on stability are proved in the limit of the sample size n tending to infinity. However, none of these results applies to the k-means algorithm (also known as Lloyd algorithm) as used in practice: they do not take into account the problem of getting stuck in local optima. In our current paper we try to overcome this shortcoming. We study the stability of the actual k-means algorithm rather than the idealized one.

Our analysis theoretically confirms the following intuition. Assume the data set has K well-separated clusters, and assume that k-means is initialized with $K' \geq K$ initial centers. We conjecture that when there is at least one initial center in each of the underlying clusters, then *the initial centers tend to stay in the clusters they had been placed in.*

Consequently, the final clustering result is essentially determined by the *number* of initial centers in each of the true clusters. In this paper we are primarily interested in this combinatorial arrangement of centers within

Keywords and phrases. Clustering, k-means, stability, model selection.

* *Marina Meilă was partly supported by NSF award IIS-031339.*

[1] Centre de Recerca Matemàtica Barcelona, Spain. sbubeck@crm.cat

[2] University of Washington, Department of Statistics, Seattle, U.S.A. mmp@stat.washington.edu

[3] Max Planck Institute for Biological Cybernetics, Tübingen, Germany. ulrike.luxburg@tuebingen.mpg.de

Article published by EDP Sciences

© EDP Sciences, SMAI 2012

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

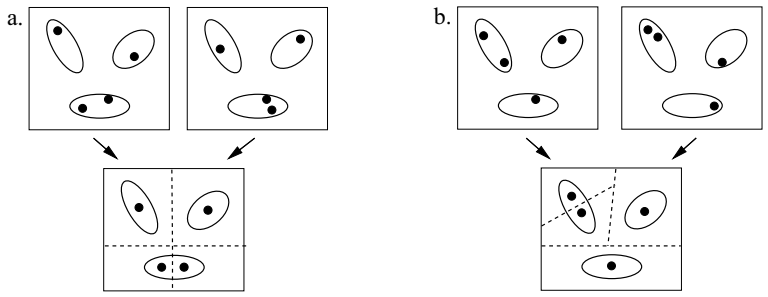


FIGURE 1. Different initial configurations and the corresponding outcomes of the k-means algorithm. (a) The two boxes in the top row depict a data set with three clusters and four initial centers. Both boxes show different realizations of the same initial configuration. As can be seen in the bottom, both initializations lead to the same k-means clustering. (b) Here the initial configuration is different from the one in (a), which leads to a different k-means clustering.

clusters, which we call a *configuration*. More precisely a configuration corresponds to a list of integers, one for each true cluster, which represent the number of centers within this cluster. We call *initial configuration* the configuration resulting from the initialization of the k-means algorithm, see Figure 1 for an illustration. Using the observation above, if one uses an initialization scheme which has the desired property of placing at least one center in each cluster with high probability, then the following will hold: if $K' = K$, we have one center per cluster, with high probability. The configuration will remain the same during the course of the algorithm. If $K' > K$, different configurations can occur. Since different configurations lead to different clusterings we obtain significantly different final clusterings depending on the random initialization, in other words we observe *instability (w.r.t initialization)*.

Note that our argument does not imply stability or instability for $K' < K$. As we have less initial centers than clusters, for any initialization scheme there will be some clusters with no initial center. In this setting centers do move between clusters, and this cannot be analyzed without looking at the actual positions of the centers. Actually, as can be seen from examples, in this case one can have either stability or instability.

The main point of our paper is that the arguments above can explain why the parameter k selected by stability based model selection is often the true number of clusters, under the assumption that the data set consists of well separated clusters and one uses an appropriate initialization scheme.

Even though the arguments above are very intuitive, even individual parts of our conjecture turn out to be surprisingly hard. In this paper we only go a first step towards a complete proof, considering mixtures of Gaussians in one dimension. For a mixture of two Gaussians ($K = 2$) we prove that the k-means algorithm is stable for $K' = 2$ and unstable for $K' = 3$. The proof technique is based on our configuration arguments outlined above. We also provide some preliminary results to study the general case, that is when the data space is $\mathbb{R}^d$ and we do not make any parametric assumption on the probability distribution. Then we have a closer look at initialization schemes for k-means, when $K' \geq K$. Is there an initialization scheme that will place at least one center in each true cluster w.h.p? Clearly, the naive method of sampling K' centers from the data set does not satisfy this property except for very small K. We study a standard but not naive initialization scheme and prove that it has the desirable property we were looking for.

Of course there exist numerous other papers which study theoretical properties of the actual k-means algorithm. However, these papers are usually concerned with the *value* of the k-means objective function at the final solution, not with the *position* of the final centers. As far as we know, our paper is the first one which analyzes the "regions of attractions" of the different local optima of the actual k-means algorithm and derives results on the stability of the k-means clustering itself.

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

2. Notation and assumptions

In the following we assume that we are given a set of n data points $X_1, \ldots, X_n \in \mathbb{R}^d$ which have been drawn i.i.d. according to some underlying distribution $\mathbb{P}$. For a center vector $c = (c_1, \ldots, c_{K'})$ with $c_i \in \mathbb{R}^d$ we denote the cluster induced by center c_k with $\mathcal{C}_k(c)$. The number of points in this cluster is denoted $N_k(c)$. The clustering algorithm we study in this paper is the standard k-means algorithm. We denote the initial centers by $c_1^{\langle 0\rangle}, \ldots, c_{K'}^{\langle 0\rangle}$ with $c_i^{\langle 0\rangle} \in \mathbb{R}^d$, and the centers after step t of the algorithm as $c_1^{\langle t\rangle}, \ldots, c_{K'}^{\langle t\rangle}$. By K we denote the true number of clusters (which will be clear from the context), by K' the number of clusters constructed by the k-means algorithm. It attempts to minimize the k-means objective function

$$W_n : \mathbb{R}^{dK'} \to \mathbb{R}, W_n(c_1, \ldots, c_{K'}) = \frac{1}{2} \sum_{i=1}^{n} \min_{k=1,..,K'} ||c_k - X_i||^2.$$

We now restate the k-means algorithm:
`Input:` $X_1, \ldots, X_n \in \mathbb{R}^d$, $K' \in \mathbb{N}$
`Initialize the centers` $c_1^{\langle 0\rangle}, \ldots, c_{K'}^{\langle 0\rangle} \in \mathbb{R}^d$
`Repeat until convergence:`
`1. Assign data points to closest centers.`
`2. Re-adjust cluster means:`

$$c_k^{\langle t+1\rangle} = \frac{1}{N_k(c^{\langle t\rangle})} \sum_{i:\, X_i \in \mathcal{C}_k(c^{\langle t\rangle})_k} X_i. \tag{2.1}$$

`Output:` $c = (c_1^{\langle \text{final}\rangle}, \ldots, c_{K'}^{\langle \text{final}\rangle})$.

Traditionally, the instability of a clustering algorithm is defined as the mean (with respect to the random sampling of data points) minimal matching distance between two clusterings obtained on two different set of data points. For the actual k-means algorithm, a second random process is the random initialization (which has not been taken into account in previous literature). Here we additionally have to take the expectation over the random initialization when computing the stability of an algorithm. In this paper we will derive qualitative rather than quantitative results on stability, thus we omit more detailed formulas.

In the following we restrict our attention to the simple setting where the underlying distribution is a mixture of Gaussians on $\mathbb{R}$ and we have access to an infinite amount of data from $\mathbb{P}$. In particular, instead of estimating means empirically when calculating the new centers of a k-means step we assume access to the true means. In this case, the update step of the k-means algorithm can be written as

$$c_k^{\langle t+1\rangle} = \frac{\int_{\mathcal{C}_k(c^{\langle t\rangle})} x f(x) \mathrm{d}x}{\int_{\mathcal{C}_k(c^{\langle t\rangle})} f(x) \mathrm{d}x}$$

where f is the density of the probability distribution $\mathbb{P}$. Results in the finite data case can be derived by the help of concentrations inequalities. However, as this introduces heavy notation and our focus lies on the random initialization rather than the random drawing of data points we skip the details. To further set up notation we denote $\varphi_{\mu,\sigma}$ the pdf of a Gaussian distribution with mean μ and variance σ. We also denote $f(x) = \sum_{k=1}^{K} w_k \varphi_{\mu_k,\sigma}$ where K is the number of Gaussians, the weights w_k are positive and sum to one, the means $\mu_{1:K} = (\mu_1, \ldots, \mu_K)$ are ordered, $\mu_1 \leq \ldots \leq \mu_K$. The minimum separation between two Gaussians is denoted by $\Delta = \min_k(\mu_{k+1} - \mu_k)$. For the standard normal distribution we denote the pdf as φ and the cdf as Φ.

3. The level sets approach

In this section we want to prove that if we run the k-means algorithm with $K' = 2$ and $K' = 3$ on a mixture of two Gaussians, then the resulting clustering depends exclusively on the initial configuration. More precisely

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

if we initialize the algorithm such that each cluster gets at least one center and the initial centers are "close enough" to the true cluster means, then during the course of the algorithm the initial centers do not leave the cluster they had been placed in. This implies stability for $K' = 2$ since there is only one possible configuration satisfying this constraint. On the other hand for $K' = 3$ we have two possible configurations, and thus instability will occur.

The following function plays an important role in our analysis:

$$H : \mathbb{R}^2 \to \mathbb{R},\ H(x, y) = x\Phi(-x + y) - \varphi(-x + y).$$

Straightforward computations show that for any μ, σ, α and h one has

$$\int_{-\infty}^{h} (x - \mu + \alpha)\varphi_{\mu,\sigma}(x)\mathrm{d}x = \sigma H\left(\frac{\alpha}{\sigma}, \frac{h + \alpha - \mu}{\sigma}\right). \tag{3.1}$$

We describe necessary and sufficient conditions to obtain stability results for particular "regions" in terms of the level sets of H.

3.1. Stability in the case of two initial centers

We consider the square $S_a = [\mu_1 - a, \mu_1 + a] \times [\mu_2 - a, \mu_2 + a]$ in $\mathbb{R}^2$. The region S_a is called *a stable region* if

$$c^{\langle 0\rangle} \in S_a \Rightarrow c^{\langle 1\rangle} \in S_a. \tag{3.2}$$

Proposition 3.1 (stable region for $K' = 2$)**.** *Equation* (3.2) *is true if and only if the following four inequalities are satisfied:*

$$\bullet\ w_1 H\left(\frac{a}{\sigma}, \frac{\Delta}{2\sigma}\right) + w_2 H\left(\frac{a + \Delta}{\sigma}, \frac{\Delta}{2\sigma}\right) \geq 0 \tag{3.3}$$

$$\bullet\ w_1 H\left(-\frac{a}{\sigma}, \frac{\Delta}{2\sigma}\right) + w_2 H\left(\frac{-a + \Delta}{\sigma}, \frac{\Delta}{2\sigma}\right) \leq 0 \tag{3.4}$$

$$\bullet\ w_1 H\left(\frac{a - \Delta}{\sigma}, -\frac{\Delta}{2\sigma}\right) + w_2 H\left(\frac{a}{\sigma}, \frac{\Delta}{2\sigma}\right) \geq 0 \tag{3.5}$$

$$\bullet\ w_1 H\left(\frac{-a - \Delta}{\sigma}, -\frac{\Delta}{2\sigma}\right) + w_2 H\left(-\frac{a}{\sigma}, -\frac{\Delta}{2\sigma}\right) \leq 0. \tag{3.6}$$

Proof. Similar to the Proof of Proposition 3.3, see below. □

This proposition gives necessary and sufficient conditions for the stability of k-means in the case $K' = 2$. In the following corollary we show an example of the kind of result we can derive from Proposition 3.1. Note that the parameters a and Δ only appear relative to σ. This allows us to consider an arbitrary σ.

Corollary 3.2 (stability for $K' = 2$)**.** *Assume that* $\min(w_1, w_2) = 0.2$ *and* $\Delta = 7\sigma$*. Assume that we have an initialization scheme satisfying:*

- *with probability at least* $1 - \delta$ *we have one initial center within* 2.5σ *of* μ_1 *and one within* 2.5σ *of* μ_2*.*

Then k-means is stable in the sense that with probability at least $1 - \delta$ *it converges to a solution with one center within* 2.5σ *of* μ_1 *and one within* 2.5σ *of* μ_2*.*

Proof. We simply check numerically that for $a = 2.5\sigma, \Delta = 7\sigma$ and $w_1 = 0.2$ (we also check $w_2 = 0.2$) equations (3.3)–(3.6) are true. Then by Proposition 3.1 we know that S_a is a stable region which implies the result. □

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

3.2. Instability in the case of 3 centers

The case of 3 centers gets more intricate. Consider the prism $T_{a,b,\varepsilon}$ and its symmetric version $\mathrm{sym}(T_{a,b,\varepsilon})$ in $\mathbb{R}^3$:

$$T_{a,b,\varepsilon} = \{c \in \mathbb{R}^3 : c_1 \leq c_2 \leq c_3, c \in [\mu_1 - a, \mu_1 + a - \varepsilon] \times [\mu_1 - a + \varepsilon, \mu_1 + a] \times [\mu_2 - b, \mu_2 + b]\}$$
$$\mathrm{sym}(T_{a,b,\varepsilon}) = \{c \in \mathbb{R}^3 : c_1 \leq c_2 \leq c_3, c \in [\mu_1 - b, \mu_1 + b] \times [\mu_2 - a, \mu_2 + a - \varepsilon] \times [\mu_2 - a + \varepsilon, \mu_2 + a]\}.$$

If we have an initialization scheme such that each cluster gets at least one center and the initial centers are close enough to the true cluster means, then we initialize either in $T_{a,b,\varepsilon}$ or $\mathrm{sym}(T_{a,b,\varepsilon})$. Thus, if these regions are stable in the following sense:

$$c^{\langle 0 \rangle} \in T_{a,b,\varepsilon} \Rightarrow c^{\langle 1 \rangle} \in T_{a,b,\varepsilon} \tag{3.7}$$

then the global k-means algorithm will be unstable, leading either to a clustering in $T_{a,b,\varepsilon}$ or $\mathrm{sym}(T_{a,b,\varepsilon})$. Expressed in the terms used in the introduction, the algorithm will be initialized with different configurations and thus be unstable.

Proposition 3.3 (stable region for $K' = 3$)**.** *Equation* (3.7) *is true if and only if all the following inequalities are satisfied:*

$$\bullet w_1 H\left(\frac{a}{\sigma}, \frac{\varepsilon}{2\sigma}\right) + w_2 H\left(\frac{a+\Delta}{\sigma}, \frac{\varepsilon}{2\sigma}\right) \geq 0 \tag{3.8}$$

$$\bullet w_1 H\left(\frac{-a+\varepsilon}{\sigma}, \frac{\varepsilon}{2\sigma}\right) + w_2 H\left(\frac{-a+\Delta+\varepsilon}{\sigma}, \frac{\varepsilon}{2\sigma}\right) \leq 0 \tag{3.9}$$

$$\bullet w_1 H\left(\frac{a-\varepsilon}{\sigma}, \frac{a-b+\Delta-\varepsilon}{2\sigma}\right) + w_2 H\left(\frac{a-\varepsilon+\Delta}{\sigma}, \frac{a-b+\Delta-\varepsilon}{2\sigma}\right)$$
$$\geq w_1 H\left(\frac{a-\varepsilon}{\sigma}, -\frac{\varepsilon}{2\sigma}\right) + w_2 H\left(\frac{a-\varepsilon+\Delta}{\sigma}, -\frac{\varepsilon}{2\sigma}\right) \tag{3.10}$$

$$\bullet w_1 H\left(-\frac{a}{\sigma}, \frac{b-a+\Delta}{2\sigma}\right) + w_2 H\left(\frac{-a+\Delta}{\sigma}, \frac{b-a+\Delta}{2\sigma}\right) \leq w_1 H\left(-\frac{a}{\sigma}, -\frac{\varepsilon}{2\sigma}\right) + w_2 H\left(\frac{-a+\Delta}{\sigma}, -\frac{\varepsilon}{2\sigma}\right) \tag{3.11}$$

$$\bullet w_1 H\left(\frac{b-\Delta}{\sigma}, \frac{b-a-\Delta+\varepsilon}{2\sigma}\right) + w_2 H\left(\frac{b-\Delta}{\sigma}, \frac{b-a-\Delta+\varepsilon}{2\sigma}\right) \leq b/\sigma - w_1\Delta/\sigma \tag{3.12}$$

$$\bullet w_1 H\left(\frac{-b-\Delta}{\sigma}, \frac{a-b-\Delta}{2\sigma}\right) + w_2 H\left(-\frac{b}{\sigma}, \frac{a-b-\Delta}{2\sigma}\right) \geq -b/\sigma - w_1\Delta/\sigma. \tag{3.13}$$

Proof. Sketch

Let $c^{\langle 0 \rangle} \in T_{a,b,\varepsilon}$. Note that the k-means algorithm in one dimension does not change the orders of centers, hence $c_1^{\langle 1 \rangle} \leq c_2^{\langle 1 \rangle} \leq c_3^{\langle 1 \rangle}$. By the definition of $T_{a,b,\varepsilon}$, to prove that after the first step of k-means the centers $c^{\langle 1 \rangle}$ are still in $T_{a,b,\varepsilon}$ we have to check six constraints. Due to space constraints, we only show how to prove that the first constraint $c_1^{\langle 1 \rangle} \geq \mu_1 - a$ is equivalent to equation (3.8). The other conditions can be treated similarly.

The update step of the k-means algorithm on the underlying distribution readjusts the centers to the actual cluster means:

$$c_1^{\langle 1 \rangle} = \frac{1}{\int_{-\infty}^{\frac{c_1^{\langle 0 \rangle} + c_2^{\langle 0 \rangle}}{2}} f(x)} \int_{-\infty}^{\frac{c_1^{\langle 0 \rangle} + c_2^{\langle 0 \rangle}}{2}} x f(x).$$

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

Thus, $c_1^{\langle 1\rangle} \geq \mu_1 - a$ is equivalent to

$$\int_{-\infty}^{\frac{c_1^{\langle 0\rangle}+c_2^{\langle 0\rangle}}{2}} (x-\mu_1+a)f(x) \geq 0.$$

Moreover, the function $h \mapsto \int_{-\infty}^{h}(x-\mu_1+a)f(x)$ is nondecreasing for $h \in [\mu_1 - a, +\infty)$. Since $c^{\langle 0\rangle} \in T_{a,b,\varepsilon}$ we know that $(c_1^{\langle 0\rangle} + c_2^{\langle 0\rangle})/2 \geq \mu_1 - a + \varepsilon/2$ and thus the statement $\forall c^{\langle 0\rangle} \in T_{a,b,\varepsilon}, c_1^1 \geq \mu_1 - a$ is equivalent to

$$\int_{-\infty}^{\mu_1-a+\varepsilon/2} (x-\mu_1+a)f(x) \geq 0.$$

We can now apply equation (3.1) with the following decomposition to get equation (3.8):

$$\int_{-\infty}^{\mu_1-a+\varepsilon/2} (x-\mu_1+a)f(x) = w_1 \int_{-\infty}^{\mu_1-a+\varepsilon/2} (x-\mu_1+a)\varphi_{\mu_1,\sigma} + w_2 \int_{-\infty}^{\mu_1-a+\varepsilon/2} (x-\mu_2+\Delta+a)\varphi_{\mu_2,\sigma}. \quad \square$$

A simple symmetry argument allows us to treat the stability of the symmetric prism.

Proposition 3.4. *If $T_{a,b,\varepsilon}$ is stable for the pdf $f(x) = w_1\varphi_{\mu_1,\sigma} + w_2\varphi_{\mu_2,\sigma}$ and $\tilde{f}(x) = w_2\varphi_{\mu_1,\sigma} + w_1\varphi_{\mu_2,\sigma}$, then the same holds for* $\mathrm{sym}(T_{a,b,\varepsilon})$.

Proof. The k-means algorithm is invariant with respect to translation of the real axis as well as to changes in its orientation. Hence if $T_{a,b,\varepsilon}$ is stable under f (resp. $\tilde{f}$), so is $\mathrm{sym}(T_{a,b,\varepsilon})$ under $\tilde{f}(x) = w_2\varphi_{\mu_1,\sigma} + w_1\varphi_{\mu_2,\sigma}$ (resp. f). $\square$

Corollary 3.5 (instability for $K' = 3$)**.** *Assume that* $\min(w_1, w_2) = 0.2$ *and* $\Delta = 14.5\sigma$. *Assume that we have an initialization scheme satisfying:*

- *with probability at least* $(1-\delta)/2$ *we have 2 initial centers within* 2.5σ *of* μ_1 *and 1 initial center within* 2.5σ *of* μ_2*;*
- *with probability at least* $(1-\delta)/2$ *we have 1 initial centers within* 2.5σ *of* μ_1 *and 2 initial centers within* 2.5σ *of* μ_2.

Then k-means is unstable: with probability $(1-\delta)/2$ *it will converge to a solution with two centers within* 3.5σ *of* μ_1 *and with probability* $(1-\delta)/2$ *to a solution with two centers within* 3.5σ *of* μ_2.

Proof. We simply check numerically that for $a = 3.5\sigma$, $b = 2.5\sigma$, $\varepsilon = \sigma$, $\Delta = 14.5\sigma$ and $w_1 = 0.2$ (we also check $w_2 = 0.2$) equations (3.8)–(3.13) are true. Then by Propositions 3.3 and 3.4 we know that $T_{3.5\sigma,2.5\sigma,\sigma}$ and its symmetric $\mathrm{sym}(T_{3.5\sigma,2.5\sigma,\sigma})$ are stable regions which implies the result. $\square$

4. Towards more general results: the geometryof the solution space of k-means

In the section above we proved by a level set approach that in a very simple setting, if we initialize the k-means algorithm "close enough" to the true cluster centers, then the initial centers do not move between clusters. However we would like to obtain this result in a more general setting. We believe that to achieve this goal in a systematic way one has to understand the structure of the solution space of k-means. We identify the solution space with the space $\mathbb{R}^{dK'}$ by representing a set of K' centers $c_1, \ldots, c_{K'} \in \mathbb{R}^d$ as a point c in the space $\mathbb{R}^{dK'}$. Our goal in this section is to understand the "shape" of the k-means objective function on this space. Secondly, we want to understand how the k-means algorithm operates on this space. That is, what can we say about the "trajectory" of the k-means algorithm from the initial point to the final solution? For simplicity, we

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

state some of the results in this section only for the case where the data space is one dimensional. They also hold in $\mathbb{R}^d$, but are more nasty to write up.

First of all, we want to compute the derivatives of W_n with respect to the individual centers. The result can also be found in the literature, see for example Lemma 4.10 in [7]. For the convenience of the reader we also present here the simple proof of this result.

Proposition 4.1 (derivatives of k-means)**.** *Given a finite data set $X_1, \ldots, X_n \in \mathbb{R}$. For $k, l \in \{1, \ldots, K'\}$ and $i \in \{1, \ldots, n\}$ consider the hyperplanes in $\mathbb{R}^{K'}$ defined by*

$$H_{k,l,i} := \{c \in \mathbb{R}^{K'} : X_i = (c_k + c_l)/2\}, \; H_{k,l} = \{c \in \mathbb{R}^{K'} : c_k = c_l\}.$$

Define the set $H := \cup_{k,l=1}^{K'} (H_{k,l} \cup \cup_{i=1}^{n} H_{k,l,i})$. Then we have:

1. *W_n is differentiable on $\mathbb{R}^{K'} \backslash H$ with partial derivatives*

$$\frac{\partial W_n(c)}{\partial c_k} = \sum_{i:\, X_i \in \mathcal{C}_k} (c_k - X_i).$$

2. *The second partial derivatives of W_n on $\mathbb{R}^{K'} \backslash H$ are*

$$\frac{\partial W_n(c)}{\partial c_k \partial c_l} = 0 \qquad \text{and} \qquad \frac{\partial W_n(c)}{\partial c_k \partial c_k} = N_k. \tag{4.1}$$

3. *The third derivatives of W_n on $\mathbb{R}^{K'} \backslash H$ all vanish.*

Proof. First of all, note that the sets $H_{k,l,i} \cup H_{k,l}$ contain the center vectors for which there exists a data point X_i which lies on the boundary of two centers c_k and c_l. Now let us look at the first derivative. We compute it by foot:

$$\frac{\partial W_n(c)}{\partial c_k} = \lim_{h \to 0} \frac{1}{h} (W_n(c_1, \ldots, c_K) - W_n(c_1, \ldots, c_k + h, \ldots, c_K)).$$

When $c \notin H$ we know that no data point lies on the boundary between two cluster centers. Thus, if h is small enough, the assignment of data points to cluster centers does not change if we replace c_k by $c_k + h$. With this property, the expression above is trivial to compute and yields the first derivative, the other derivatives follow similarly. □

A straightforward consequence is as follows:

Proposition 4.2 (k-means does Newton iterations)**.** *The update step of re-adjusting the clustering mean in the k-means algorithms corresponds exactly to a step of a Newton optimization.*

Proof. This proposition follows directly from Proposition 4.1, the definition of the Newton iteration on W_n and the definition of the k-means update step. This fact has also been stated (less rigorously and without proof) in [5]. □

Together, the two propositions show an interesting picture. We have seen in Proposition 4.1 that the k-means objective function W_n is differentiable on $\mathbb{R}^{K'} \backslash H$. This means that the space $\mathbb{R}^{K'}$ is separated into many cells with hyperplane boundaries $H_{k,l,i}$. By construction, the cells are convex (as they are intersections of half-spaces). Our finding means that each data set $X_1, \ldots, X_n$ induces a partitioning of this solution space into convex cells. To avoid confusion, at this point we would like to stress again that we are not looking at a fixed clustering solution on the data space (which can be described by cells with hyperplane boundaries, too), but at the space of all center vectors c. It is easy to see that all centers c within one cell correspond to exactly one clustering of the data

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

points (*i.e.*, one specific partition of the data into K' subsets). As it is well known that the k-means algorithm never visits a clustering twice, we can conclude that each cell is visited at most once by the algorithm.Within each cell, W_n is quadratic (as the third derivatives vanish). Moreover, we know that k-means behaves as the Newton iteration. On a quadratic function, the Newton optimization jumps in one step to the minimum of the function. This means that if k-means enters a cell that contains a local optimum of the k-means objective function, then the next step of k-means jumps to this local optimum and stops.

Now let us look more closely at the trajectories of the k-means algorithm. [15] inspired us to derive the following property.

Proposition 4.3 (trajectories of k-means)**.** *Let $c^{\langle t\rangle}$ and $c^{\langle t+1\rangle}$ be two consecutive solutions visited by the k-means algorithm. Consider the line connecting those two solutions in $\mathbb{R}^{K'}$, and let $c^\alpha = (1-\alpha)c^{\langle t\rangle} + \alpha c^{\langle t+1\rangle}$ be a point on this line (for some $\alpha \in [0,1]$). Then $W_n(c^\alpha) \leq W_n(c^{\langle t\rangle})$.*

Proof. The following inequalities hold true:

$$\begin{aligned} W_n(c^\alpha) &= \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k(c^\alpha)} ||X_i - c_k^\alpha||^2 \\ &\leq \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k(c^t)} ||X_i - c_k^\alpha||^2 \\ &\leq \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k(c^t)} \alpha||X_i - c_k^{\langle t\rangle}||^2 + (1-\alpha)||X_i - c_k^{\langle t+1\rangle}||^2 \\ &\leq \alpha W_n(c^{\langle t\rangle}) + (1-\alpha)W_n(c^{\langle t\rangle}) = W_n(c^{\langle t\rangle}). \end{aligned}$$

For the first and third inequality we used the fact that assigning points in $\mathcal{C}_k(c)$ to the center c_k is the best thing to do to minimize W_n. For the second inequality we used that $x \to ||x||^2$ is convex. □

We believe that the properties of the k-means objective function and the algorithm are the key to prove more general stability results. However, there is still an important piece missing, as we are going to explain now. Since k-means performs Newton iterations on W_n, one could expect to get information on the trajectories in the configuration space by using a Taylor expansion of W_n. However, as we have seen above, each step of the k-means algorithm crosses one of the hyperplanes $H_{k,l,i}$ on which W_n is non-differentiable. Hence, a direct Taylor expansion approach on W_n cannot work. On the other hand, surprisingly one can prove that the limit objective function $W := \lim \frac{1}{n}W_n$ is almost surely a continuously differentiable function on $\mathbb{R}^{K'}$ (we omit the proof in this paper). Thus one may hope that one could first study the behavior of the algorithm for W, and then apply concentration inequalities to carry over the results to W_n. Unfortunately, here we face another problem: one can prove that in the limit case, a step of the k-means algorithm is *not* a Newton iteration on W.

Proposition 4.3 directly evokes a scheme to design stable regions. Assume that we can find two regions $A \subset B \subset \mathbb{R}^{K'}$ of full rank (*i.e.*, with non-empty interior) and such that

$$\max_{x\in\partial A} W_n(x) \leq \min_{x\in\partial B} W_n(x). \tag{4.2}$$

Then, if we initialize in A we know that we will converge to a configuration in B. This approach sounds very promising. However, we found that it was impossible to satisfy both equation (4.2) and the constraint that A has to be "big enough" so that we initialize in A with high probability.

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

Algorithm PRUNED MINDIAM

Input: $w_{\min}$, number of centers K'

1. Initialize with L random points $c_{1:L}^{\langle 0\rangle}$, L computed by (5.4).
2. Run one step of k-means, that is:
   (a) to each center $c_j^{\langle 0\rangle}$ assign region $\mathcal{C}_j^0$, $j = 1 : L$;
   (b) calculate $c_{1:L}^{\langle 1\rangle}$ as the centers of mass of regions $\mathcal{C}_{1:L}^0$.
3. Remove all centers $c_j^{\langle 1\rangle}$ for which $P[\mathcal{C}_j^1] \leq p_0$, where p_0 is given by (5.4). We are left with $c_{j'}^{\langle 1\rangle}$, $j' = 1 : L'$.
4. Choose K' of the remaining centers by the MINDIAM heuristic:
   (a) select one center at random;
   (b) repeat until K' centroids are selected:
   select the centroid $c_q^{\langle 1\rangle}$ that maximizes the minimum distance to the already selected centroids.

Output: the K' selected centroids $c_k^{\langle 1\rangle}$, $k = 1 : K'$.

FIGURE 2. The PRUNED MINDIAM initialization.

Finally, we would like to elaborate on a few more complications towards more general results:

- On a high level, we want to prove that if K' is slightly larger than the true K, then k-means is unstable. On the other hand, if K' gets close to the number n of data points, we trivially have stability again. Hence, there is some kind of "turning point" where the algorithm is most unstable. It will be quite a challenge to work out how to determine this turning point.
- Moreover, even if we have so many data points that the above problem is unlikely to occur, our analysis breaks down if K' gets too large. The reason is that if K' is much bigger than K, then we cannot guarantee any more that initial centers will be in stable regions. Just the opposite will happen: at some point we will have outliers as initial centers, and then the behavior of the algorithm becomes rather unpredictable.
- Finally, consider the case of $K' < K$. As we have already mentioned in the introduction, in this case it is not necessarily the case that different initial configurations lead to different clusterings. Hence, a general statement on (in)stability is not possible in this case. This also means that the tempting conjecture "the true K has minimal stability" is not necessarily true.

5. AN INITIALIZATION ALGORITHM AND ITS ANALYSIS

We have seen that one can prove results on clustering stability for k-means if we use a "good" initialization scheme which tends to place initial centers in different Gaussians. We now show that an established initialization algorithm, the PRUNED MINDIAM initialization described in Figure 2 has this property, *i.e.* it has the effect of placing the initial centroids in disjoint, bounded neighborhoods of the means $\mu_{1:K}$. This often rediscovered algorithm is credited to [8]. In [6] it was analyzed it in the context of the EM algorithm. Later [14] used it in experimental evaluations of EM, and it was found to have a significant advantage w.r.t more naive initialization methods in some cases. While this and other initializations have been extensively studied in conjunction with EM, we are not aware of any studies of PRUNED MINDIAM for k-means.

We make three necessary conceptual assumptions. Firstly to ensure that K is well-defined we assume that the mixture weights are bounded below by a known weight $w_{\min}$.

Assumption 5.1. $w_k \geq w_{\min}$ for all k.

We also require to know a lower bound Δ and an upper bound $\Delta_{\max}$ on the separation between two Gaussians, and we assume that these separations are "sufficiently large". In addition, later we shall make several technical assumptions related to a parameter τ used in the proofs, which also amount to conditions on the separation. These assumptions shall be made precise later.

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

Theorem 5.2 (PRUNED MINDIAM initialization)**.** *Let $f = \sum_1^K w_k \varphi_{\mu_k,1}$ be a mixture of K Gaussians with centers $\mu_{1:K}$, $\mu_k \leq \mu_{k+1}$, and unit variance. Let $\tau \in (0, 0.5)$, $\delta_{\text{miss}} > 0$, δ_{impure} defined in Proposition 5.7. If we run Algorithm* PRUNED MINDIAM *with any $2 \leq K' \leq 1/w_{\min}$, then, subject to Assumptions 5.1–5.13 (specified later), with probability $1 - 2\delta_{\text{miss}} - \delta_{\text{impure}}$ over the initialization there exist K disjoint intervals $\tilde{A}_k$, specified in Section 5.4, one for each true mean μ_k, so that all K' centers $c_{k'}^{\langle 1\rangle}$ are contained in $\bigcup_k \tilde{A}_k$ and*

$$\text{if } K' = K, \text{ each } \tilde{A}_k \text{ will contain exactly one center } c_{k'}^{\langle 1\rangle}, \tag{5.1}$$

$$\text{if } K' < K, \text{ each } \tilde{A}_k \text{ will contain at most one center } c_{k'}^{\langle 1\rangle}, \tag{5.2}$$

$$\text{if } K' > K, \text{ each } \tilde{A}_k \text{ will contain at least one center } c_{k'}^{\langle 1\rangle}. \tag{5.3}$$

The idea to prove this result is to show that the following statements hold with high probability. By selecting L preliminary centers in step 1 of PRUNED MINDIAM, each of the Gaussians obtains at least one center (Sect. 5.1). After steps 2a, 2b we obtain "large" clusters (mass $> p_0$) and "small" ones (mass $\leq p_0$). A cluster can also be "pure" (respectively "impure") if most of its mass comes from a single Gaussian (respectively from several Gaussians). Step 3 removes all "small" cluster centers, but (and this is a crucial step of our argument) w.h.p it will also remove all "impure" cluster centers (Sect. 5.2). The remaining clusters are "pure" and "large"; we show (Sect. 5.3) that each of their centers is reasonably close to some Gaussian mean μ_k. Hence, if the Gaussians are well separated, the selection of final centers $c_q^{\langle 1\rangle}$ in step 4 "cycles through different Gaussians" before visiting a particular Gaussian for the second time (Sect. 5.4). The rest of this section outlines these steps in more details.

5.1. Step 1 of PRUNED MINDIAM. Picking the initial centroids $c^{\langle 0\rangle}$

We need to pick a number of initial centers L large enough that each Gaussian has at least 1 center w.h.p. We formalize this here and find a value for L that ensures the probability of this event is at least $1 - \delta_{\text{miss}}$, where δ_{miss} is a tolerance of our choice. Another event that must be avoided for a "good" initialization is that all centroids $c_j^{\langle 0\rangle}$ belonging to a Gaussian end up with initial clusters $\mathcal{C}_j^0$ that have probability less than p_0. If this happens, then after thresholding, the respective Gaussian is left with no representative centroid, *i.e.* it is "missed". We set the tolerance for this event to $\delta_{\text{thresh}} = \delta_{\text{miss}}$. Let $t = 2\Phi(-\Delta/2)$ the *tail probability* of a cluster and A_k the symmetric neighborhood of μ_k that has $\varphi_{\mu_k,1}(A_k) = 1 - t$.

Proposition 5.3. *If we choose*

$$L \geq \left(\ln \frac{1}{\delta_{\text{miss}} w_{\min}}\right) \Big/ \left((1-t) w_{\min}\right) \quad \text{and} \quad p_0 = \frac{1}{eL} \tag{5.4}$$

then the probability over all random samplings of centroids $c_{1:L}^{\langle 0\rangle}$ that at least one centroid $c_j^{\langle 0\rangle}$ with assigned mass $P[\mathcal{C}_j^0] \geq p_0$ can be found in each A_k, $k = 1 : K$, is greater or equal to $1 - 2\delta_{\text{miss}}$.

The proof of this result is complicated but standard fare (*e.g.* Chernoff bounds) and is therefore omitted.

After steps 1, 2a and 2b of PRUNED MINDIAM are performed, we obtain centers $c_{1:L}^{\langle 1\rangle}$ situated at the centers of mass of their respective clusters $\mathcal{C}_{1:L}^1$. Removing the centers of small clusters follows. We now describe a beneficial effect of this step.

5.2. Step 3 of PRUNED MINDIAM. Thresholding removes impure clusters

We introduce the concept of *purity* of a cluster, which is related to the ratio of points from a certain Gaussian w.r.t to the total probability mass of the cluster. Denote P_k the probability distribution induced by the k-th Gaussian $\varphi_{\mu_k,1}$.

Definition 5.4. A cluster $\mathcal{C}$ is $(1-\tau)$-*pure* if most of its points come from a single Gaussian, *i.e.* if $w_k P_k[\mathcal{C}] \geq (1-\tau) P[\mathcal{C}]$, with $\tau < 1/2$ being a positive constant. A cluster which is not $(1-\tau)$-pure is τ-*impure* (or simply *impure*).

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

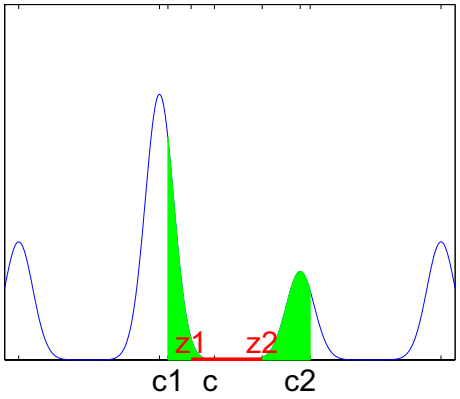


FIGURE 3. Concrete example of a large impure cluster $[z1, z2]$; $c1, c, c2$ represent the consecutive cluster centers $c_{j-1}^{\langle 0\rangle}, c_j^{\langle 0\rangle}, c_{j+1}^{\langle 0\rangle}$. We demonstrate that if $P[z1, z2] > p_0$ then the interval $[c1, c2]$ (which is twice its length) must have mass $> p_1 \gg p_0$. If L is large enough, having such a large interval contain a single c_j is improbable. Numerical values: mixture with $\Delta = 10, w_{\min} = 0.15$, impurity $\tau([z_1, z_2]) = 0.07$, $P[z1, z2] = 0.097$, $P[c1, c2] = 0.24$; using $\delta_{\text{miss}} = 0.02, \tau = 0.015$ one gets $L = 38$, $p_0 = 0.095 < P[z1, z2]$, $p_1 = 0.0105 < P[c1, c2]$, $\delta_{\text{impure}} = 0.016 \gg (1 - P[c1, c2])^{L-1} = 0.00003$.

The values of τ that we consider useful are of the order $0.001-0.02$ and, as it will appear shortly, $\tau < w_{\min}/2$. The purity of a cluster helps in the following way: if a cluster is pure, then it can be "tied" to one of the Gaussians. Moreover, its properties (like center of mass) will be dictated by the Gaussian to which it is tied, with the other Gaussians' influence being limited; Section 5.3 exploits this idea.

But there will also be clusters that are impure, and so they cannot be tied to any Gaussian. Their properties will be harder to analyze, and one expects their behavior to be less predictable. Luckily, impure clusters are very likely small. As we show now, the chance of having an impure cluster with mass larger than p_0 is bounded by a δ_{impure} which we are willing to tolerate.

Because of limited space, we leave out the long and complex rigourous proofs of this result, and give here just the main ideas. Let $\mathcal{C}_j = [z_1, z_2]$ be a τ-impure cluster, with $P[\mathcal{C}_j] \geq p_0$, c_j the centroid that generates $\mathcal{C}_j$ (not necessarily at its center of mass) and c_{j-1}, c_{j+1} the centroids of the adjacent clusters (not necessarily centers of mass). As one can show, even though an impure cluster contains some probability mass from each Gaussian, in most of this section we only need consider the two Gaussians which are direct neighbors of $\mathcal{C}$. Let us denote the parameters of these (consecutive) Gaussians by $\mu_{1,2}, w_{1,2}$.

For the purpose of the proof, we are looking here at the situation after step 2a, thus the centroids $c_{j-1,j,j+1}$ should be $c_{j-1,j,j+1}^{\langle 0\rangle}$, but we renounce this convention temporarily to keep the notation light. We want to bound the probability of cluster $\mathcal{C}_j$ being impure and large. Note that step 2b of the PRUNED MINDIAM does not affect either of these properties, as it only acts on the centers.

A simple observation is the following. Since $z_1 = \frac{c_{j-1}+c_j}{2}$ and $z_2 = \frac{c_{j+1}+c_j}{2}$ we have $c_{j+1} - c_{j-1} = 2(z_2 - z_1) = 2\Delta z$. The idea is to show that if an impure region has probability larger than p_0, then the interval $[c_{j-1}, c_{j+1}]$ has probability at least p_1, significantly larger than p_0. On the other hand, the probability of sampling from P a single center $\mathcal{C}_j$ out of a total of L in an interval of length $2\Delta z$ is $P[c_{j-1}, c_{j+1}](1-P[c_{j-1}, c_{j+1}])^{L-1} < (1-p_1)^{L-1}$.

If p_1 and L are large enough, then $(1-p_1)^{L-1} \stackrel{\text{def}}{=} \delta_{\text{impure}}$ will be vanishingly small. We proceed in two steps: first we find the minimum length Δz_0 of a cluster $\mathcal{C}_j$ which is impure and large. Then, we find a lower bound p_1 on the probabability of any interval $[c, c + 2\Delta z_0]$ under the mixture distribution. The following assumption ensures that the purity $1 - \tau$ is attainable in each Gaussian.

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

Assumption 5.5. Let $\gamma_{k,k'}(x) = \frac{w_{k'}\varphi_{\mu_{k'},1}(x)}{w_k\varphi_{\mu_k,1}(x)}$ (a local purity measure). Then

$$\sum_{k'\neq k} \gamma_{k,k'}\left(\Phi^{-1}\left(\frac{1}{2}+\frac{(1-\tau)p_0}{2w_{\min}}\right)\right) \leq \frac{\tau}{1-\tau}.$$

The next assumption ensures that $\Delta z_0 > 0$, *i.e.* it is an informative bound.

Assumption 5.6. $d\left(\frac{\tau p_0}{w_{\min}}\right) < \frac{1}{2}\Delta$.

Proposition 5.7 (impure clusters are small w.h.p)**.** *Let w_1, w_2 be the mixture weights of two consecutive Gaussians and define $\Delta z_0 = \Delta - d\left(\frac{\tau p_0}{w_1}\right) - d\left(\frac{\tau p_0}{w_2}\right)$,*

$$p_1 = w_1\Phi\left(\frac{\Delta-2\Delta z_0}{2} - \frac{\ln\frac{w_1}{w_2}}{\Delta-2\Delta z_0}\right) + w_2\Phi\left(\frac{\Delta-2\Delta z_0}{2} - \frac{\ln\frac{w_2}{w_1}}{\Delta-2\Delta z_0}\right)$$

and $\delta_{\text{impure}} = (1-p_1)^{L-1}$. Let $\mathcal{C}_j^0, j = 1,\ldots,L$ be the regions associated with $c_{1:L}^{\langle 0\rangle}$ after step 2a of the PRUNED MINDIAM *algorithm. If Assumptions 5.1–5.6 hold, then the probability that there exists $j \in \{1,\ldots,L\}$ so that $P[\mathcal{C}_j^0] \geq p_0$ and $w_1P_1[\mathcal{C}_j^0] \geq \tau P[\mathcal{C}_j^0]$, $w_2P_2[\mathcal{C}_j^0] \geq \tau P[\mathcal{C}_j^0]$ is at most δ_{impure}. This probability is over the random initialization of the centroids $c_{1:L}^{\langle 0\rangle}$.*

To apply this proposition without knowing the values of w_1, w_2 one needs to minimize the bound p_1 over the range $w_1, w_2 > w_{\min}$, $w_2 + w_1 \leq 1-(K-2)w_{\min}$. This minimum can be obtained numerically if the other quantities are known.

We also stress that because of the two-step approach, first minimizing Δz_0, then $P[c, c+2\Delta z_0]$, the bound δ_{impure} obtained is not tight and could be significantly improved.

5.3. The $(1-\tau)$-pure cluster

Now we focus on the clusters that have $P[\mathcal{C}] > p_0$ and are $(1-\tau)$-pure. By Proposition 5.7, w.h.p their centroids are the only ones which survive the thresholding in step 3 of the PRUNED MINDIAM algorithm. In this section we will find bounds on the distance $|c_j^{\langle 1\rangle} - \mu_k|$ between $\mathcal{C}_j$'s center of mass and the mean of "its" Gaussian.

We start by listing some useful properties of the standard Gaussian. Denote by $r(x)$ the center of mass of $[x,\infty)$ under the truncated standard Gaussian, and by $d(t)$ the solution of $1-\Phi(d) = t$, with $0 < t < 1$. Intuitively, $d(t)$ is the cutoff location for a tail probability of t. Note that any interval whose probability under the standard normal exceeds t must intersect $[-d(t), d(t)]$. Let $a > 0$ (in the following a as to be thought as a small positive constant).

Proposition 5.8. *(i) $r(x)$ is convex, positive and increasing for $x \geq 0$. (ii) For $w \in [2a,\infty)$ the function $d(a/w)$ is convex, positive and increasing w.r.t w, and $r(d(a/w))$ is also convex, positive and increasing.*

Proposition 5.9. *Let $\mathcal{C} = [z_1, z_2]$ be an interval (with z_1, z_2 possibly infinite), c its center of mass under the normal distribution $\varphi_{\mu,1}$ and $P[\mathcal{C}]$ its probability under the same distribution. If $1/2 \geq P[\mathcal{C}] \geq p$, then $|c-\mu| \leq r(d(p))$ and $\min\{|z_1-\mu|, |z_2-\mu|\} \leq d(p) = -\Phi^{-1}(p)$.*

The proofs are straightforward and omitted. Define now $w_{\max} = 1-(K-1)w_{\min}$ the maximum possible cluster size in the mixture and

$$R(w) = r\left[-\Phi^{-1}\left(\frac{(1-\tau)p_0}{w}\right)\right], \quad \tilde{R}(w_1,w_2) = -\Phi^{-1}\left[\frac{\tau w_1}{(1-\tau)w_2} + \Phi(d(\frac{(1-\tau)p_0}{w_1} - \Delta)\right].$$

In the next proposition, we will want to assume that $\tilde{R} \geq 0$. The following assumption is sufficient for this purpose.

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

Assumption 5.10. $\frac{\tau}{w_{\min}} \leq \frac{1}{2} - \Phi(-\Delta/2)$

Proposition 5.11 (The $(1-\tau)$-pure cluster)**.** *Let cluster $\mathcal{C} = [z_1, z_2]$ with $z_2 > \mu_k$, $P[\mathcal{C}] \geq p_0$ and $w_k P_k[\mathcal{C}] \geq (1-\tau)P[\mathcal{C}]$ for some k, with τ satisfying Assumptions 5.5 and 5.10. Let c, c_k denote the center of mass of $\mathcal{C}$ under P, P_k respectively. Then*

$$|c_k - \mu_k| \leq R(w_k) \tag{5.5}$$

and, whenever $k < K$

$$z_2 - \mu_k \leq -\tilde{R}(w_k, w_{k+1}) \leq -\tilde{R}(w_{\max}, w_{\min}). \tag{5.6}$$

Proposition 5.12 (corollary)**.** *If $c_k > \mu_k$ and $k < K$ then*

$$\begin{aligned} c - \mu_k &\leq (1-\tau)R(w_k) + \tau(\Delta - \tilde{R}(w_k, w_{k+1})) && (5.7)\\ &\leq (1-\tau)R(w_{\max}) + \tau(\Delta - \tilde{R}(w_{\max}, w_{\min})) && (5.8)\\ &\leq (1-\tau)R(w_{\max}) + \tau\Delta && (5.9) \end{aligned}$$

else

$$\mu_k - c \leq R(w_k) \leq R(w_{\max}) \quad c - \mu_k \leq \tau(\Delta - \tilde{R}(w_k, w_{k+1})). \tag{5.10}$$

By symmetry, a similar statement involving $\mu_{k-1}, w_{k-1}, \mu_k, w_k$ and c holds when $z_2 > \mu_k$ is replaced by $z_1 < \mu_k$. With it we have essentially shown that an almost pure cluster which is not small cannot be too far from its Gaussian center μ_k.

Proof of Proposition 5.11. (5.5) follows from Proposition 5.9. Now for bounding z_2, in the case $k < K$. Because $(1-\tau)P[\mathcal{C}] \leq w_k$ (the contribution of Gaussian k to cluster $\mathcal{C}$ cannot exceed all of w_k) we have $P_{k+1}[C] \leq \frac{\tau P[\mathcal{C}]}{w_{k+1}} \leq \frac{\tau w_k}{(1-\tau)w_{k+1}}$ and $P_{k+1}[C] = \Phi(z_2 - \mu_{k+1}) - \Phi(z_1 - \mu_{k+1}) \geq \Phi(z_2 - \mu_{k+1}) - \Phi(c_1 - \mu_{k+1})$ from which the first inequality in (5.6) follows. The function $\tilde{R}$ is increasing with w_k when w_{k+1} constant or $w_{k+1} = \text{constant} - w_1$, which gives the second bound. □

Proof of the corollary. First note that we can safely assume $z_1 \geq \mu_k$. If the result holds for this case, then it is easy to see that having $z_1 < \mu_k$ only brings the center of mass c closer to μ_k.

$$c = \frac{w_k P_k[C] c_k + \sum_{k' \neq k} w_{k'} P_{k'}[C] c_{k'}}{P[\mathcal{C}]} \leq (1-\tau)c_k + \tau z_2. \tag{5.11}$$

Now (5.7), (5.8) follow from Proposition 5.11. For (5.9) Assumption 5.10 assures that $\tilde{R} \geq 0$. As a consequence, this bound is convex in w_k. If $k = 1$ and $c_1 \leq \mu_1$, or $k = K$ and $c_K > \mu_K$ then the second term in the sum (5.11) pulls c_1 in the direction of μ_1 (respectively c_K in the direction of μ_K) and we can get the tighter bounds (5.10). □

In conclusion, we have shown now that if the unpruned center c "belongs" to Gaussian k, then

$$c \in \tilde{A}_k(w_k) = [\mu_k - R_\tau^-(w_k), \mu_k + R_\tau^+(w_k)]$$

whith $R_\tau^-(w_k) = (1-\tau)R(w_k) + \tau(\mu_k - \mu_{k-1})$, $R_\tau^+(w_k) = (1-\tau)R(w_k) + \tau(\mu_{k+1} - \mu_k)$, $R_\tau^-(w_1) = R(w_1)$, and $R_\tau^+(w_K) = R(w_K)$.

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

5.4. Step 4 of PRUNED MINDIAM. Selecting the centers by the MINDIAM heuristic

From Section 5.2 we know that w.h.p all centroids unpruned at this stage are $(1-\tau)$ pure. We want to ensure that after the selection in step 4 each Gaussian has at least one $c_j^{\langle 1\rangle}$ near its center. For this, it is sufficient that the regions $\tilde{A}_k$ are disjoint, *i.e.*

$$(\mu_{k+1}-\mu_k) - (R_\tau^+(w_k) + R_\tau^-(w_{k+1})) > R_\tau^-(w_k) + R_\tau^+(w_k)$$
$$(\mu_{k+1}-\mu_k) - (R_\tau^+(w_k) + R_\tau^-(w_{k+1})) > R_\tau^-(w_{k+1}) + R_\tau^+(w_{k+1})$$

for all k. Replacing $R_\tau^\pm(w_k)$ with their definitions and optimizing over all possible $w_{1:K} \geq w_{\min}$ and for all $\Delta\mu \leq \mu_{k+1}-\mu_k \leq \Delta_{\max}$ produces

$$\tilde{A}_k \;=\; [\mu_k \pm (1-\tau)R(w_{\max}) \pm \tau\Delta_{\max}]$$

and

Assumption 5.13. $(1-3\tau)\Delta - \tau\Delta_{\max} \;>\; [3R(w_{\max}) + R(w_{\min})](1-\tau)$.

6. SIMULATIONS

In this section we test our conjecture in practice and run some simulations to emphasize the different theoretical results of the previous sections. We also investigate whether it is necessary to look at the stability of k-means with respect to the random drawing of the data set. In the following when we refer to randomization we mean with respect to the initialization while the resampling corresponds to the random drawing of the data set.

Setup of the experiments. As distributions we consider mixtures of Gaussians in one, two, and ten dimensions. Each mixture consists of several, reasonably well separated clusters. We report the results on three such data sets:

- "Two dim four balanced clusters": mixture of four Gaussians in $\mathbb{R}^2$ with means (-3.3), $(0,0)$, $(3,3)$, $(3,-3)$; the covariance matrix of all clusters is diagonal with entries 0.2 and 1 on the diagonal; the mixing coefficients are uniform, that is all clusters have the same weight.
- "Two dim four imbalanced clusters": as above, but with mixing coefficients $0.1, 0.5, 0.3, 0.1$.
- "Ten dim ten clusters": mixture of ten Gaussians in $\mathbb{R}^{10}$ with means $(i,0,0,\ldots)$ for $i = 1,\ldots,10$. All Gaussians are spherical with variance 0.05 and mixing coefficients are uniform.

As clustering algorithm we use the standard k-means algorithm with the following initializations:

- Standard initialization: randomly pick K' data points.
- MINDIAM initialization, coincides with step 5 in Figure 2.
- PRUNED MINDIAM initialization, as analyzed in Section 5 (see Fig. 2)
- Deterministic initialization: K' fixed points sampled from the distribution.

For a range of parameters $K' \in \{2,\ldots,10\}$ we compute the clustering stability by the following protocols:

- Randomization, no resampling: we draw once a data set of $n = 100$ points from the distribution. Then we run the k-means algorithm with different initializations.
- Resampling, no randomization: we fix a set of deterministic starting points (by drawing them once from the underlying distribution). Then we draw 100 data sets of size $n = 100$ from the underlying distribution, and run k-means with the deterministic starting points on these data sets.
- Resampling and randomization: we combine the two previous approaches.

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

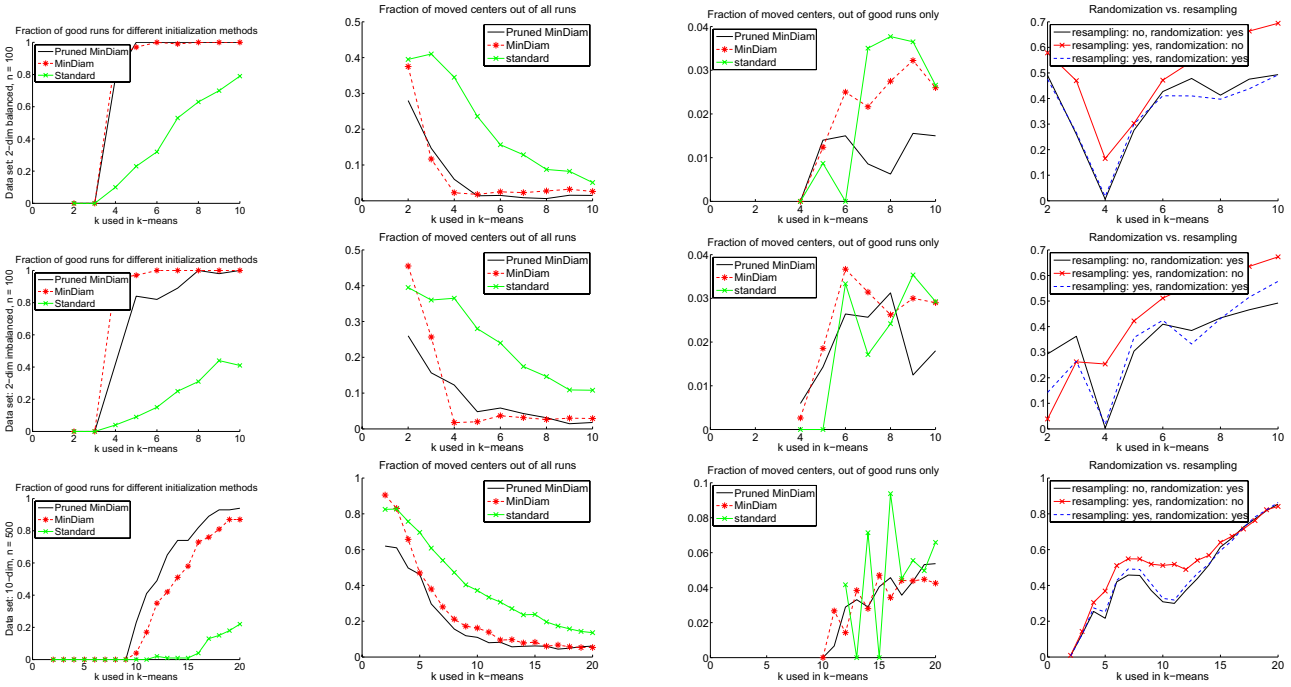


FIGURE 4. Simulation results. First row: data set "two dim four balanced clusters". Second row: data set "two dim four imbalanced clusters". Third row: data set "ten dim ten clusters" (see text for details).

Then we compute the stability (using the instability measure of [3]) with respect to the minimal matching distance between the clusters. Each experiment was repeated 100 times, we always report the mean values over those repetitions.

Note that all experiments were also conducted with different data set sizes ($n = 50, 100, 500$), stability was computed with and without normalization (we used the normalization suggested in [9]), and the k-means algorithm was used with and without restarts. All those variations did not significantly effect the outcome, hence we omit the plots.

Results. First we evaluate the effect of the different initializations. To this end, we count how many initializations were "good initializations" in the sense that each true cluster contains at least one initial center. In all experiments we consistently observe that both the pruned and non-pruned min diameter heuristic already achieve many good runs if K' coincides with K or is only slightly larger than the true K (of course, good runs cannot occur for $K' < K$). The standard random initialization does not achieve the same performance. See Figure 4, first column.

Second, we record how often it was the case that initial cluster centers cross cluster borders. We can see in Figure 4 (second column) that this behavior is strongly correlated with the number of "good initializations". Namely, for initialization methods which achieve a high number of good initializations the fraction of centers which cross cluster borders is very low. Moreover, one can see in the third column of Figure 4 that centers usually do not cross cluster borders if the initialization was a good one. This coincides with our theoretical results.

Finally, we compare the different protocols for computing the stability: using randomization but no resampling, using resampling but no randomization, and using both randomization and resampling, *cf.* right most plots in Figure 4. In simple data sets, all three protocols have very similar performance, see for example the

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

first row in Figure 4. That is, the stability values computed on the basis of resampling behave very similarly to the ones computed on the basis of randomization, and all three methods clearly detect the correct number of clusters. Combining randomization and resampling does not give any advantage. However, on the more difficult data sets (the imbalanced one and the 10-dimensional one), we can see that resampling without randomization performs worse than the two protocols with randomization (second and third row of Fig. 4). While the two protocols using randomization have a clear local minimum (and "global on the right") around the correct number of clusters, stability based on resampling alone fails to achieve this. We never observed the opposite effect in any of our simulations (we ran many more experiments than reported in this paper). This shows, as we had hoped, that randomization plays an important role for clustering stability, and in certain settings can achieve better results than resampling alone.

Finally, in the experiments above we ran the k-means algorithm in two modes: with restarts, where the algorithm is started 50 times and only the best solution is kept; and without restarts. The results did not differ much (above we report the results without restarts). This means that in practice, for stability based parameter selection one can save computing time by simply running k-means without restarting it many times (as is usually done in practice). From our theory we had even expected that running k-means without restarts achieves better results than with restarts. We thought that many restarts diminish the effect of exploring local optima, and thus induce more stability than "is there". But the experiments did not corroborate this intuition.

7. Conclusions and outlook

Previous theoretical work on model selection based on the stability of the k-means algorithm has assumed an "ideal k-means algorithm" which always ends in the global optimum of the objective function. The focus was to explain how the random drawing of sample points influences the positions of the final centers and thus the stability of the clustering. This analysis explicitly excluded the question when and how the k-means algorithm ends in different local optima. In particular, this means that these results only have a limited relevance for the actual k-means algorithm as used in practice.

In this paper we study the actual k-means algorithm. We have shown that the initialization strongly influences the k-means clustering results. We also show that if one uses a "good" initialization scheme, then the k-means algorithm is stable if it is initialized with the correct number of centers, and unstable if it is initialized with too many centers. Even though we have only proved these results in a simple setting so far, we are convinced that the same mechanism also holds in a more general setting.

These results are a first step towards explaining why the selection of the number of clusters based on clustering stability is so successful in practice [9]. From this practical point of view, our results suggest that introducing randomness by the initialization may be sufficient for an effective model selection algorithm. Another aspect highlighted by this work is that the situations $K' < K$ and $K' > K$ may represent two distinct regimes for clustering, that require separate concepts and methods to be analyzed.

The main conceptual insight in the first part of the paper is the idea described in Section 1 that the initial configuration determines the stability or instability of the k-means algorithm. With this idea we indirectly characterize the "regions of attraction" of different local optima of the k-means objective function. To our knowledge, this is the first such characterization in the vast literature of k-means.

In the second part of the paper we study an initialization scheme for the k-means algorithm. Our intention is not to come up with a new scheme, but to show that a scheme already in use is "good" in the sense that it tends to put initial centers in different clusters. It is important to realize that such a property does not hold for the widely used uniform random initialization.

On the technical side, most of the proofs and proof ideas in this section are novel. In very broad terms, our analysis is reminiscent to that of [6]. One reason we needed new proof techniques lie partly in the fact that we analyze one-dimensional Gaussians, whose concentration properties differ qualitatively from those of high dimensional Gaussians. We loose some of the advantages high dimensionality confers. A second major difference is that k-means behaves qualitatively differently from EM whenever more than one Gaussian is involved. While

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press

EM weights a point "belonging" to a cluster by its distance to the cluster center, to the effect that far away points have a vanishing influence on a center c_j, this is not true for k-means. A far-away point can have a significative influence on the center of mass c_j, precisely because of the leverage given by the large distance. In this sense, k-means is a more brittle algorithm than EM, is less predictible and harder to analyze. In order to deal with this problem we "eliminated" impure clusters in Section 5.2. Third, while [6] is concerned with finding the correct centers when K is known, our analysis carries over to the regime when K' is too large, which is qualitatively very different of the former.

Of course many initialization schemes have been suggested and analyzed in the literature for k-means (for examples see [1, 10]). However, these papers analyze the *clustering cost* obtained with their initialization, not the positions of the initial centers.

References

[1] D. Arthur and S. Vassilvitskii, k-means++: the advantages of careful seeding, in *Proc. of SODA* (2007).
[2] S. Ben-David and U. von Luxburg, Relating clustering stability to properties of cluster boundaries, in *Proc. of COLT* (2008).
[3] S. Ben-David, U. von Luxburg and D. Pál, A sober look on clustering stability, in *Proc. of COLT* (2006).
[4] S. Ben-David, D. Pál and H.-U. Simon, Stability of k-means clustering, in *Proc. of COLT* (2007).
[5] L. Bottou and Y. Bengio, Convergence properties of the k-means algorithm, in *Proc. of NIPS* (1995).
[6] S. Dasgupta and L. Schulman, A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *J. Mach. Learn. Res.* **8** (2007) 203–226.
[7] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*. Springer (2000).
[8] D. Hochbaum and D. Shmoys, A best possible heuristic for the k-center problem. *Math. Operat. Res.* **10** (1985) 180–184.
[9] T. Lange, V. Roth, M. Braun and J. Buhmann, Stability-based validation of clustering solutions. *Neural Comput.* **16** (2004) 1299–1323.
[10] R. Ostrovsky, Y. Rabani, L.J. Schulman and C. Swamy, The effectiveness of Lloyd-type methods for the k-means problem, in *Proc. of FOCS* (2006).
[11] O. Shamir and N. Tishby, Cluster stability for finite samples, in *Proc. of NIPS* (2008).
[12] O. Shamir and N. Tishby, Model selection and stability in k-means clustering, in *Proc. of COLT* (2008).
[13] O. Shamir and N. Tishby, On the reliability of clustering stability in the large sample regime, in *Proc. of NIPS* (2008).
[14] N. Srebro, G. Shakhnarovich and S. Roweis, An investigation of computational and informational limits in Gaussian mixture clustering, in *Proc. of ICML* (2006).
[15] Z. Zhang, B. Dai and A. Tung, Estimating local optimums in EM algorithm over Gaussian mixture model, in *Proc. of ICML* (2008).

https://doi.org/10.1051/ps/2012013 Published online by Cambridge University Press