Power k-Means Clustering

Jason Xu [1] **Kenneth Lange** [2]

Abstract

Clustering is a fundamental task in unsupervised machine learning. Lloyd's 1957 algorithm for k-means clustering remains one of the most widely used due to its speed and simplicity, but the greedy approach is sensitive to initialization and often falls short at a poor solution. This paper explores an alternative to Lloyd's algorithm that retains its simplicity and mitigates its tendency to get trapped by local minima. Called *power k-means*, our method embeds the k-means problem in a continuous class of similar, better behaved problems with fewer local minima. Power k-means anneals its way toward the solution of ordinary k-means by way of majorization-minimization (MM), sharing the appealing descent property and low complexity of Lloyd's algorithm. Further, our method complements widely used seeding strategies, reaping marked improvements when used together as demonstrated on a suite of simulated and real data examples.

1. Introduction

Clustering is a foundational task in unsupervised learning and data analysis, and plays a key role in countless applications. Its purpose is to partition n objects into k similarity classes based on a measure of similarity between pairs of objects. Recent advances based on spectral formulations (Ng et al., 2002), Bayesian and nonparametric approaches (Heller & Ghahramani, 2005; Kulis & Jordan, 2012), message passing (Frey & Dueck, 2007), subspace clustering (Vidal, 2011), and continuous optimization (Chi & Lange, 2015; Shah & Koltun, 2017) continue to contribute to a vast literature. A more complete overview is provided by Mirkin (1998), Jain (2010), and Everitt et al. (2011). After decades of innovation, none of these advances have managed to displace *k-means clustering* and Lloyd's algorithm for implementing it (Steinhaus, 1956; Lloyd, 1982). Lloyd's algorithm alternates between two steps of membership reassignment and cluster recentering. Unfortunately, it is NP-hard to optimally partition n points in d-dimensional Euclidean space into k sets (Aloise et al., 2009; Dasgupta & Freund, 2009). Lloyd's algorithm suffers well-documented shortcomings such as sensitivity to initialization and deterioration in high dimensions. Nonetheless, it persists because of its speed and simplicity. In settings appropriate for k-means, many competing algorithms underperform or offer only marginal improvements while incurring higher computational cost, additional hyperparameters, or more opaque objectives.

In this paper, we propose a generalization of Lloyd's algorithm that a) makes it more robust to initialization, b) enhances its performance in high dimensions, and c) retains its speed and simplicity. *Power k-means* embeds the k-means problem in a continuum of better behaved problems. These smoothed intermediate problems have flatter objective functions that tend to guide clustering toward the global minimum of the k-means objective. Our method enjoys the same $\mathcal{O}(nkd)$ per-iteration complexity as Lloyd's algorithm. Steady and stable progress along the solution path is ensured by a majorization-minimization algorithm (Lange, 2016), which guarantees a decrease in the k-means objective at each step.

Our method can be considered an extension of the k-harmonic means (KHM) algorithm of Zhang and colleagues (Zhang et al., 1999). We give a proof of the descent property for both the KHM algorithm and our extension of it. In contrast to KHM, which implicitly replaces the k-means objective by a proxy, our algorithm ultimately seeks to optimize the same measure of quality. Instead, we simply anneal our way to its minimum. In targeting the original k-means objective, the considerable and growing body of theory relevant to analyzing k-means applies to understanding our method. These include recent developments such as new fast learning rates (Dinh et al., 2016), empirical risk bounds (Bachem et al., 2017), and statistical guarantees (Lu & Zhou, 2016). Power k-means addresses the algorithmic drawbacks of Lloyd's algorithm by proposing *internal* improvements from an optimization perspective. On the other hand, external "wrapper" methods such as well-designed initializations for k-means (Arthur & Vassilvitskii, 2007; Celebi et al.,

[1]Department of Statistical Science, Duke University [2]Departments of Biomathematics, Statistics, and Human Genetics, UCLA. Correspondence to: Jason Xu <jason.q.xu@duke.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

2013) have been successful strategies addressing some of the same issues. Our approach is nicely complementary to this line of work. As we will see in empirical studies, seeding methods can be immediately applied together with our algorithm, furthering the advantages it confers.

Before continuing, we establish some notation. Vectors and matrices appear in boldface type. The components of a vector $\boldsymbol{v}$ are written as v_i and the entries of a matrix $\boldsymbol{A}$ as a_{ij}, with ith column $\boldsymbol{a}_i$. A sequence of vectors $\boldsymbol{v}_m$ has components $v_{m,i}$, and a sequence of matrices $\boldsymbol{A}_m$ has entries $a_{m,ij}$.

1.1. Center-based Clustering

Center-based methods encode each cluster by its center and iteratively refine the center estimates and assignments of points to clusters. Lloyd's algorithm for k-means, expectation maximization (EM) for Gaussian mixture models, and fuzzy k-means are examples of center-based methods (Jain, 2010). Let $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ denote the data matrix, $\boldsymbol{\Theta} \in \mathbb{R}^{d \times k}$ the center matrix, and $\mathcal{C}_j$ the membership set for cluster j. The k-means objective is

$$f_{-\infty}(\boldsymbol{\Theta}) = \sum_{j=1}^{k} \sum_{i \in \mathcal{C}_j} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2 \quad (1)$$

$$= \sum_{i=1}^{n} \min_{1 \le j \le k} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2. \quad (2)$$

At each iteration, Lloyd's algorithm assigns each data point $\boldsymbol{x}_i$ to the cluster $\mathcal{C}_j$ minimizing the Euclidean distance $\|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|$. It then redefines the center $\boldsymbol{\theta}_j$ by averaging the $\boldsymbol{x}_i$ in cluster $\mathcal{C}_j$. Although Lloyd's algorithm is guaranteed to converge to a local minimum, it is notoriously sensitive to its starting point. Arthur & Vassilvitskii (2006) exhibit a super-polynomial worst case running time for Lloyd's algorithm and demonstrate the empirical and theoretical advantages of the now standard seeding scheme k-means++ (Arthur & Vassilvitskii, 2007; Ostrovsky et al., 2006). Clever seeding remains an active area of research (Celebi et al., 2013). For instance, recent work accelerates k-means++ sampling using Markov chain Monte Carlo (Bachem et al., 2016).

Apart from initialization schemes, several geometrically inspired efforts to address the sensitivity of Lloyd's algorithm have been made. Notably, the k-harmonic means (KHM) algorithm (Zhang et al., 1999) replaces the k-means criterion by

$$f_{-1}(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \Big(\frac{1}{k} \sum_{j=1}^{k} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^{-2} \Big)^{-1}. \quad (3)$$

The harmonic mean provides a smooth proxy to the $\min$ function and leads to a simple iterative procedure that has proven more robust to initialization in many examples. Its extension KHM$_p$ replaces $\|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^{-2}$ by $\|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^{-p}$ in criterion (3). Careful choice of the tuning parameter p can improve performance (Zhang, 2001). Further attempts to enhance KHM range from gravitational search and simulated annealing (Güngör & Ünler, 2007; Yin et al., 2011) to hybrids using differential evolution and particle swarms (Tian et al., 2009; Yang et al., 2009). These heuristics quickly become complicated, and their effectiveness varies by case. It is unclear when these alternative criteria are preferable to the well-studied k-means criterion. Empirical studies suggest that the benefits of KHM are confined to low dimensions (Zhang, 2001; Hamerly & Elkan, 2002). Teboulle (2007) provides an elegant formalism unifying several center-based clustering algorithms by reformulating k-means exactly as a smooth problem via support functions (Rockafellar, 1970), and recovers several known soft clustering methods including KHM through approximate smoothing via asymptotic nonlinear means. Our work complements the theoretical insights provided by this continuous optimization framework. We will make the case that the family of power means provides a nearly ideal approximation of k-means.

1.2. Generalized and Power Means

For any positive integer k and any continuous, strictly monotone function $g(y)$, one can define a generalized mean or Kolmogorov mean (de Carvalho, 2016) through the formula

$$M_g(\boldsymbol{y}) = g^{-1}\left[\frac{g(y_1) + \cdots + g(y_k)}{k} \right]. \quad (4)$$

One can check that $M_g(\boldsymbol{y})$ is continuous, symmetric, and monotonic in its arguments. It also satisfies the identities $M_g(y, \ldots, y) = y$ and $M_g(\mu, \ldots, \mu, y_{r+1}, \ldots, y_k) = M_g(y_1, \ldots, y_k)$, where $\mu = M_g(y_1, \ldots, y_r)$. Kolmogorov showed that any function satisfying these properties takes the form (4) (Kolmogorov & Castelnuovo, 1930). The choice $g(y) = y^s$ on the domain $(0, \infty)$ yields the family of Hölder or *power means*. We will abbreviate the mean $M_{y^s}(\boldsymbol{y})$ by $M_s(\boldsymbol{y})$. Within the class of power means, $s > 1$ corresponds to the usual ℓ_s-norm of $\boldsymbol{y}$, $s = 1$ to the arithmetic mean, and $s = -1$ to the harmonic mean. The geometric mean $\sqrt[k]{y_1 \cdots y_k}$ can be viewed as the special case $s = 0$ after taking limits.

In the family of power means $M_s(\boldsymbol{y})$, an easy calculation yields the gradient

$$\frac{\partial}{\partial y_j} M_s(\boldsymbol{y}) = \Big(\frac{1}{k} \sum_{i=1}^{k} y_i^s \Big)^{\frac{1}{s}-1} \frac{1}{k} y_j^{s-1}. \quad (5)$$

This formula shows that $M_s(\boldsymbol{y})$ is strictly increasing in each of its entries. The class of power means enjoys other properties relevant to our subsequent analysis. In addition to the previous identities, power means are homogenous in

the sense that $M_s(c\boldsymbol{y}) = cM_s(\boldsymbol{y})$ for all $c > 0$ and $\boldsymbol{y}$. This is the only family of generalized means with this property (Hardy et al., 1952). By continuity, one can extend the domain of $M_s(\boldsymbol{y})$ to boundary points where one or more $y_i = 0$; in particular, $M_s(\boldsymbol{0}) = 0$. The power mean family also satisfies the limits

$$\begin{aligned} \lim_{s\to\infty} M_s(\boldsymbol{y}) &= \max\{y_1, \ldots, y_k\}, \\ \lim_{s\to-\infty} M_s(\boldsymbol{y}) &= \min\{y_1, \ldots, y_k\}, \end{aligned} \tag{6}$$

and obeys the well-known power mean inequality $M_s(\boldsymbol{y}) \leq M_t(\boldsymbol{y})$ for any $s \leq t$ (Steele, 2004).

1.3. Majorization-minimization

The majorization-minimization (MM) principle (Lange et al., 2000) provides a generic recipe for converting hard optimization problems (non-convex or non-smooth) into a sequence of simpler problems. MM algorithms have become increasingly popular for large-scale optimization in statistics and machine learning (Mairal, 2015; Lange, 2016). Recent applications include stochastic optimization (Bietti & Mairal, 2017), regression with constraints (Xu et al., 2017), and clustering under missing data (Chi et al., 2016). All EM algorithms for maximum likelihood estimation are instances of MM (Becker et al., 1997). Given that Lloyd's algorithm can be interpreted as an EM algorithm for a Gaussian mixture model (GMM) with vanishing variances or as a variational EM approximation with isotropic GMMs (Forster & Lücke, 2018), it is natural that the broader MM principle underpins our own method.

An MM algorithm successively minimizes a sequence of surrogate functions $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$ majorizing the objective function $f(\boldsymbol{\theta})$ at the current iterate $\boldsymbol{\theta}_m$. The notion of majorization entails tangency $g(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}_m) = f(\boldsymbol{\theta}_m)$ at the current iterate and domination $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m) \geq f(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. The update rule

$$\boldsymbol{\theta}_{m+1} := \arg\min_{\boldsymbol{\theta}} \; g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$$

implies the descent property

$$f(\boldsymbol{\theta}_{m+1}) \leq g(\boldsymbol{\theta}_{m+1} \mid \boldsymbol{\theta}_m) \leq g(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}_m) = f(\boldsymbol{\theta}_m). \tag{7}$$

Note that minimizing g is not strictly necessary: the weaker condition $g(\boldsymbol{\theta}_{m+1} \mid \boldsymbol{\theta}_m) \leq g(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}_m)$ also decreases $f(\boldsymbol{\theta})$. Maximizing a function can be accomplished by an analogous combination of sequential minorization and maximization.

2. The Power k-Means Algorithm

We will define the power k-means objective function for given power s by the formula

$$f_s(\Theta) \;=\; \sum_{i=1}^{n} M_s(\|\boldsymbol{x}_i - \boldsymbol{\theta}_1\|^2, \ldots, \|\boldsymbol{x}_i - \boldsymbol{\theta}_k\|^2) \tag{8}$$

consistent with our previous notations $f_{-\infty}(\Theta)$ and $f_{-1}(\Theta)$ for the k-means and k-harmonic means objectives. The limiting relation $\lim_{s\to-\infty} f_s(\Theta) = f_{-\infty}(\Theta)$ follows from (6) and suggests that we systematically decrease $f_s(\Theta)$ while gradually sending s to $-\infty$. To this end, the MM framework can be brought to bear in deriving a descent scheme. We begin by examining the Hessian matrix of $M_s(\boldsymbol{y})$, with entries

$$\begin{aligned} \frac{\partial^2}{\partial y_j \partial y_l} M_s(\boldsymbol{y}) \;=\; & \Big(\frac{1}{k}\sum_{i=1}^{k} y_i^s\Big)^{\frac{1}{s}-2} \frac{1}{k^2} s\Big(\frac{1}{s}-1\Big) y_j^{s-1} y_l^{s-1} \\ & + \mathbf{1}_{\{j=l\}} \Big(\frac{1}{k}\sum_{i=1}^{k} y_i^s\Big)^{\frac{1}{s}-1} \frac{1}{k}(s-1) y_j^{s-2}. \end{aligned}$$

Thus, the quadratic form generated by the Hessian satisfies along direction $\boldsymbol{v}$

$$\begin{aligned} & \frac{k^2}{\left(\frac{1}{k}\sum_{i=1}^{k} y_i^s\right)^{\frac{1}{s}-2}} \boldsymbol{v}^t d^2 M_s(\boldsymbol{y}) \boldsymbol{v} \;= \\ & \quad (1-s)\left[\Big(\sum_{i=1}^{k} y_i^{s-1} v_i\Big)^2 - \Big(\sum_{i=1}^{k} y_i^s\Big)\Big(\sum_{i=1}^{k} y_i^{s-2} v_i^2\Big)\right]. \end{aligned}$$

The second factor on the right is always nonpositive by the Cauchy-Schwarz inequality, while the factor $1 - s$ is nonnegative if and only if $s \leq 1$. Hence, $M_s(\boldsymbol{y})$ is concave whenever $s \leq 1$ and convex otherwise. Because $M_g(y, \ldots, y) = y$, $M_s(\boldsymbol{y})$ is neither strictly concave nor strictly convex. Concavity and convexity do not carry over to the composite function $f_s(\Theta)$.

To derive an MM algorithm, we exploit the concavity of $M_s(\boldsymbol{y})$ for $s < 1$. Concavity supplies the linear majorization

$$M_s(\boldsymbol{y}) \;\leq\; M_s(\boldsymbol{y}_m) + \sum_{j=1}^{k} \frac{\partial}{\partial y_j} M_s(\boldsymbol{y}_m)(y_j - y_{m,j}) \tag{9}$$

at any anchor point $\boldsymbol{y}_m$. The required partial derivatives appear in equation (5). If we substitute $\|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2$ for y_j and $\|\boldsymbol{x}_i - \boldsymbol{\theta}_{m,j}\|^2$ for $y_{m,j}$ and sum inequality (9) over i, the majorization

$$f_s(\boldsymbol{\Theta}) \;\leq\; \overbrace{f_s(\boldsymbol{\Theta}_m) - \sum_{i=1}^{n}\sum_{j=1}^{k} w_{m,ij}\|\boldsymbol{x}_i - \boldsymbol{\theta}_{m,j}\|^2}^{c_m}$$

$$+\sum_{i=1}^{n}\sum_{j=1}^{k} w_{m,ij}\|\boldsymbol{x}_i-\boldsymbol{\theta}_j\|^2;$$

$$w_{m,ij} = \frac{\frac{1}{k}\|\boldsymbol{x}_i-\boldsymbol{\theta}_{m,j}\|^{2(s-1)}}{\left(\frac{1}{k}\sum_{l=1}^{k}\|\boldsymbol{x}_i-\boldsymbol{\theta}_{m,l}\|^{2s}\right)^{(1-\frac{1}{s})}}$$

follows. The constant c_m does not depend on $\boldsymbol{\Theta}$ and is thus irrelevant in minimizing this surrogate function. To minimize the right-hand side, we solve its stationarity equation:

$$\begin{aligned}\mathbf{0} &= -2\sum_{i=1}^{n} w_{m,ij}(\boldsymbol{x}_i-\boldsymbol{\theta}_j)\\ \boldsymbol{\theta}_{m+1,j} &= \frac{1}{\sum_{i=1}^{n} w_{m,ij}}\sum_{i=1}^{n} w_{m,ij}\boldsymbol{x}_i.\end{aligned}$$

This results in a straightforward iterative procedure, and is guaranteed to decrease $f_s(\boldsymbol{\theta})$ regardless of the choice of s. By analogy to Lloyd's algorithm, each step alternates between updating the weights $w_{m,ij}$ and recomputing the centers. The MM updates fall within the convex hull of the data points. The overall procedure, summarized in Steps 3 and 4 of Algorithm 1, has the same per-iteration complexity $\mathcal{O}(nkd)$ as Lloyd's algorithm. One can check that taking $s=-1$ exactly recovers the KHM iterates (Zhang et al., 1999), and for s fixed, the updates coincide with the SKM steps suggested by Teboulle (2007) without approximate smoothing parameter. In contrast, the KHM$_p$ algorithm for $p \neq 2$ does not operate on the class of objectives studied here, nor is it consistent with Tebuolle's formalism, discussed further in Section 3. Annealing enters in Step 5 of the algorithm, discussed below.

Algorithm 1 Power k-means algorithm pseudocode

1: Initialize $s_0 < 0$ and $\boldsymbol{\Theta}_0$, input data $\boldsymbol{X} \in \mathbb{R}^{d\times n}$, and set constant $\eta > 1$:
2: **repeat**
3: $\quad w_{m,ij} \leftarrow \left(\sum_{l=1}^{k}\|\boldsymbol{x}_i-\boldsymbol{\theta}_{m,l}\|^{2s_m}\right)^{\frac{1}{s_m}-1}\|\boldsymbol{x}_i-\boldsymbol{\theta}_{m,j}\|^{2(s_m-1)}$
4: $\quad \boldsymbol{\theta}_{m+1,j} \leftarrow \left(\sum_{i=1}^{n} w_{m,ij}\right)^{-1}\sum_{i=1}^{n} w_{m,ij}\boldsymbol{x}_i$
5: $\quad s_{m+1} \leftarrow \eta \cdot s_m$ (optional)
6: **until** convergence

2.1. Annealing

As s tends to $-\infty$, the power means surfaces $f_s(\boldsymbol{\Theta})$ provide a family of progressively rougher landscapes. Figure 1 illustrates how more and more local valleys appear as s tends to $-\infty$. In the other extreme when $s=1$, one can show analytically that all optimal centers $\boldsymbol{\theta}_j$ collapse to the sample mean $\bar{\boldsymbol{X}}$. In contrast to the heuristic annealing methods applied with KHM (Güngör & Ünler, 2007), moving along a sequence of power mean objectives $f_s(\boldsymbol{\Theta})$ automatically provides a form of annealing, much like Gibbs distributions naturally lend themselves to deterministic annealing with EM algorithms (Ueda & Nakano, 1998). Such behavior is intuitively desirable, guiding the solution path toward a global minimizer as it threads its way across the landscapes. This intuition is supported by empirical studies in Section 4.

In general, altering the objective function at each step of an MM algorithm does not guarantee the descent property and the advantages that follow from it. Fortunately, this is not the case for the proposed algorithm.

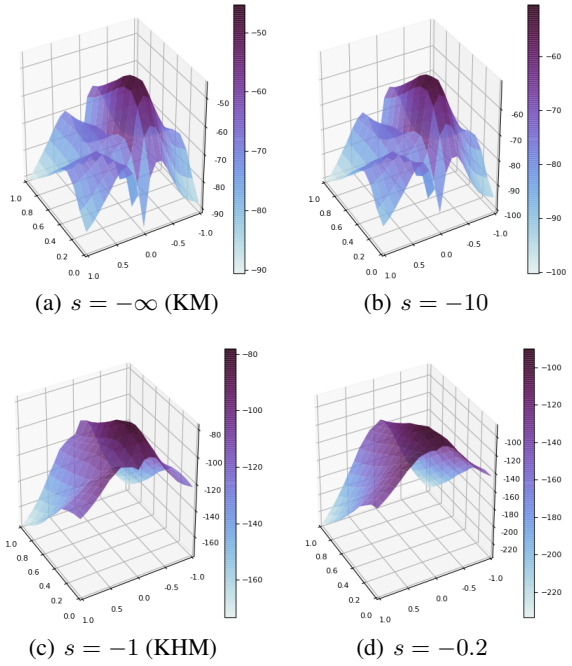
Proposition 2.1. *For any decreasing sequence $s_m \leq 1$, the iterates $\boldsymbol{\Theta}_m$ produced by the MM updates (Alg. 1) generates a decreasing sequence of objective values $f_{s_m}(\boldsymbol{\Theta}_m)$ bounded below by 0. As a consequence, the sequence of objective values converges.*

Proof. The result follows immediately by combining the MM inequalities (7) with the power mean inequality □

Proposition 2.1 allows one to freely choose a schedule for decreasing s. In practice, the multiplicative schedule indicated in Algorithm 1 works well. As we show in Section 4, a default rule $s_{m+1} = \eta s_m$ with $\eta = 1.05$ and $s_0 < 0$ is successful across synthetic and real datasets from multiple domains of varying size n and dimension d. A sensitivity analysis to η reveals that this conclusion is stable to reasonable perturbations of η. In contrast, the initial value s_0 does affect performance, but we will see in Section 4 that this parameter does not require careful tuning; a broad range of s_0 values lead to marked improvements over k-means.

We now understand KHM as an attempt to optimize one member $f_{-1}(\boldsymbol{\Theta})$ of an entire family of objectives; this strategy can be interpreted as early stopping along our solution path. Comparison of Figures 1(a) and 1(c) illustrate why KHM is more robust to initialization than Lloyd's algorithm. Although the global minimizers of $f_{-1}(\boldsymbol{\Theta})$ and $f_{-\infty}(\boldsymbol{\Theta})$ are similar, $f_{-1}(\boldsymbol{\Theta})$ features fewer local minima that may trap the algorithm. This phenomenon is partly explained by the intuition originally motivating KHM; namely, the harmonic average behaves similarly to the $\min$ function in its sensitivity to small inputs. However, notice that off the diagonal, $\frac{\partial}{\partial y_j}\min \boldsymbol{y}$ is 1 if y_j is minimal and 0 otherwise. Examining the partial derivatives (5) when $s=-1$ reveals that the shape of $M_s(\boldsymbol{y})$ may differ substantially from $\min \boldsymbol{y}$ when many y_i have similar values, suggesting that $f_{-1}(\boldsymbol{\Theta})$ is a poor proxy for $f_{-\infty}(\boldsymbol{\Theta})$ in some regimes. Practitioners have indeed found that as d increases and there is little difference in distances between pairs of sample points, KHM suffers from the curse of dimensionality (Kriegel et al., 2009). In contrast, as $s \to -\infty$, the derivatives (5)

tend to $\frac{\partial}{\partial z_j} \min \boldsymbol{y}$ where the latter is defined. Though harder to visualize in high dimensions, we will see in Section 4 that power k-means retains the benefits of annealing as dimension d increases, remaining successful in settings where both KHM and Lloyd's algorithm deteriorate.



(a) $s = -\infty$ (KM) (b) $s = -10$

(c) $s = -1$ (KHM) (d) $s = -0.2$

Figure 1. Cross-sections of the k-means and power means objective surfaces for varying powers s. Here $-f_s(\boldsymbol{\Theta})$ is plotted for $n = 100$ simulated data points from $k = 3$ clusters in dimension $d = 1$. Two cluster centers vary along the axes, holding the third fixed at its true value.

3. Properties

We have seen that the centers $\boldsymbol{\theta}_j$ stay within the convex hull C of the data points at each iteration. The following propositions show that this a productive strategy for finding global minima of the objectives. By symmetry, there are at least $k!$ equivalent global minimizers.

Proposition 3.1. *For any $s \leq 1$, the global minimum of $f_s(\boldsymbol{\Theta})$ lies in the Cartesian product C^k.*

Proof. Let $P_C(\boldsymbol{\theta})$ denote the Euclidean projection of $\boldsymbol{\theta}$ onto C. For any $\boldsymbol{v} \in C$, the obtuse angle condition $[\boldsymbol{\theta} - P_C(\boldsymbol{\theta})]^t[\boldsymbol{v} - P_C(\boldsymbol{\theta})] \leq 0$ holds. Since $\boldsymbol{x}_i \in C$, it follows that

$$\begin{aligned}\|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2 &= \|\boldsymbol{x}_i - P_C(\boldsymbol{\theta}_j)\|^2 \\ &\quad + 2[\boldsymbol{x}_i - P_C(\boldsymbol{\theta}_j)]^t[P_C(\boldsymbol{\theta}_j) - \boldsymbol{\theta}_j] + \|P_C(\boldsymbol{\theta}_j) - \boldsymbol{\theta}_j\|^2 \\ &\geq \|\boldsymbol{x}_i - P_C(\boldsymbol{\theta}_j)\|^2 + \|P_C(\boldsymbol{\theta}_j) - \boldsymbol{\theta}_j\|^2.\end{aligned}$$

Thus, $\|\boldsymbol{x}_i - P_C(\boldsymbol{\theta}_j)\|^2 < \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2$ whenever $\boldsymbol{\theta}_j \notin C$. Given that $M_s(\boldsymbol{y})$ is strictly increasing in each of its arguments, one can strictly decrease each term $M_s(\|\boldsymbol{x}_i - \boldsymbol{\theta}_1\|^2, \ldots, \|\boldsymbol{x}_i - \boldsymbol{\theta}_k\|^2)$ contributing to $f_s(\boldsymbol{\Theta})$ by substituting $P_C(\boldsymbol{\theta}_j)$ for $\boldsymbol{\theta}_j$. Furthermore, because C^k is compact and $f_s(\boldsymbol{\Theta})$ is continuous, $f_s(\boldsymbol{\Theta})$ attains its its minimum on C^k. □

Proposition 3.2. *For any decreasing sequence $s_m \leq 1$ tending to $-\infty$, the sequence $f_{s_m}(\boldsymbol{\Theta})$ converges uniformly to $f_{-\infty}(\boldsymbol{\Theta})$ on C^k. Furthermore, $\lim_{m\to\infty} \min_{\boldsymbol{\Theta}} f_{s_m}(\boldsymbol{\Theta}) = \min_{\boldsymbol{\Theta}} f_{-\infty}(\boldsymbol{\Theta})$.*

Proof. For any $\boldsymbol{\Theta}$, the limit (6) and the power mean inequality imply that $f_{s_m}(\boldsymbol{\Theta})$ converges monotonically to the continuous function $f_{-\infty}(\boldsymbol{\Theta})$. In view of Dini's theorem, monotonicity ensures that this convergence is in fact uniform on the compact set C^k. As for the convergence of the minimum values, Proposition 3.1 allows us to confine our attention to C^k. For any $\epsilon > 0$, uniform convergence entails the existence of an integer M such that whenever $m \geq M$,

$$\sup_{\boldsymbol{\Theta} \in C^k} |f_{-\infty}(\boldsymbol{\Theta}) - f_{s_m}(\boldsymbol{\Theta})| < \epsilon.$$

For any such m and $\boldsymbol{\Theta}_m \in \operatorname{argmin} f_{s_m}(\boldsymbol{\Theta})$, we have

$$\begin{aligned}\min_{\boldsymbol{\Theta} \in C^k} f_{-\infty}(\boldsymbol{\Theta}) &\leq f_{-\infty}(\boldsymbol{\Theta}_m) \leq f_{s_m}(\boldsymbol{\Theta}_m) + \epsilon \\ &\leq f_{s_m}(\boldsymbol{\Theta}) + \epsilon \leq f_{-\infty}(\boldsymbol{\Theta}) + 2\epsilon\end{aligned}$$

for all $\boldsymbol{\Theta} \in C^k$. Taking $\boldsymbol{\Theta} \in \arg\min f_{-\infty}(\boldsymbol{\Theta})$ and sending $\epsilon \to 0$ establish the claim. □

Before proceeding, we make several remarks regarding convergence. First, the MM iterates $\boldsymbol{\Theta}_m$ need not minimize the objectives $f_{s_m}(\boldsymbol{\Theta})$. Second, while compactness implies the existence of convergent subsequences of $\boldsymbol{\Theta}_m$, their limits do not necessarily minimize $f_{-\infty}(\boldsymbol{\Theta})$ globally. Third, Algorithm 1 terminates at some finite value of s^* in practice. If we continue iterating at s^* until $f_{s^*}(\boldsymbol{\Theta})$ stabilizes, then any limit points of the MM algorithm are stationary points of $f_{s^*}(\boldsymbol{\Theta})$ (Teboulle, 2007; Lange, 2016). Alternatively, if we switch to Lloyd's algorithm when we reach s^*, then a finite number of further iterates will converge to a local minimum. In our experience, the difference in solutions between simply stopping at s^* or continuing with either of these alternatives is negligible. In contrast, the KHM$_p$ algorithm does not possess the descent property. We demonstrate this in the Supplement through a new interpretation of KHM$_p$ as an approximate Newton's method.

3.1. Membership and weighing functions

A number of studies have analyzed and designed algorithms based on *membership functions* $h_j(\boldsymbol{x}_i \mid \boldsymbol{\Theta}) \geq 0$ and *weighing functions* $\alpha(\boldsymbol{\Theta} \mid \boldsymbol{x}_i) > 0$ (Kearns et al., 1998; Medasani

et al., 1998; MacKay, 2003). The former define the proportion of point $\boldsymbol{x}_i$ assigned to cluster j. The latter define the influence of point $\boldsymbol{x}_i$ on subsequent center updates. For Lloyd's algorithm, $h_j(\boldsymbol{x}_i \mid \boldsymbol{\Theta}) = 1$ if $\boldsymbol{x}_i$ is closest to center j and 0 otherwise. For EM, $h_j(\boldsymbol{x}_i \mid \boldsymbol{\Theta})$ is a posterior probability under a Gaussian mixture model (Hamerly & Elkan, 2002). In contrast to Lloyd's algorithm, EM, and fuzzy clustering (Bezdek et al., 1984) which all have constant $\alpha(\boldsymbol{\Theta} \mid \boldsymbol{x}_i)$, KHM$_p$ features a dynamic weighing function (Zhang, 2001) with

$$h_j(\boldsymbol{x}_i \mid \boldsymbol{\Theta}) = \frac{\|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^{-(p+2)}}{\sum_{l=1}^k \|\boldsymbol{x}_i - \boldsymbol{\theta}_l\|^{-(p+2)}} \tag{10}$$

$$\alpha(\boldsymbol{\Theta} \mid \boldsymbol{x}_i) = \frac{\sum_{l=1}^k \|\boldsymbol{x}_i - \boldsymbol{\theta}_l\|^{-(p+2)}}{\left(\sum_{l=1}^k \|\boldsymbol{x}_i - \boldsymbol{\theta}_l\|^{-p}\right)^2}. \tag{11}$$

When $p > 2$, a point near its closest center $\boldsymbol{\theta}_j$ is down-weighted since $\alpha(\boldsymbol{\Theta} \mid \boldsymbol{x}_i) = \mathcal{O}(\|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^{p-2})$, a dynamic weighing that Zhang (2001) calls a form of boosting. Several authors have remarked that this analogy is yet to be made precise (Freund & Schapire, 1997; Hamerly & Elkan, 2002), and the same observation also reveals increasing sensitivity to outliers as the tuning parameter p increases. Hamerly & Elkan (2002) test hybrid algorithms by methodically swapping membership and weighting functions among several center-based algorithms, reporting empirical advantages under those in (10). Algorithm 1 entails the following membership and weighting functions:

$$h_j(\boldsymbol{x}_i \mid \boldsymbol{\Theta}) = \frac{\|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^{2(s-1)}}{\sum_{l=1}^k \|\boldsymbol{x}_i - \boldsymbol{\theta}_l\|^{2(s-1)}},$$

$$\alpha(\boldsymbol{\Theta} \mid \boldsymbol{x}_i) = \frac{\sum_{l=1}^k \|\boldsymbol{x}_i - \boldsymbol{\theta}_l\|^{2(s-1)}}{\left(\sum_{l=1}^k \|\boldsymbol{x}_i - \boldsymbol{\theta}_l\|^{2s}\right)^{(1-\frac{1}{s})}}.$$

These functions coincide with those under KHM$_p$ when $2p = -s$. Hence, the empirical strengths explored by Hamerly & Elkan (2002) carry over to power k-means clustering. Notice that in power k-means clustering, $\lim_{s\to-\infty} \alpha(\boldsymbol{\Theta} \mid \boldsymbol{x}_i) = \mathcal{O}(1)$, as expected since our formulation approaches k-means in the limit. Power k-means therefore benefits from dynamic weighing as $f_s(\boldsymbol{\Theta})$ gradually transitions to $f_{-\infty}(\boldsymbol{\Theta})$, yet the transition automatically stabilizes, requiring no tuning parameter p to trade off desirable "boosting" behavior against increasing sensitivity to outliers.

3.2. Non-Euclidean distances

Our exposition focuses on Euclidean distances and considers experimental settings ideal for Lloyd's algorithm. It is worth mentioning that the power means framework can be broadened to accommodate alternative notions of distance. For instance, one can substitute $\|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|_1$ instead of $\|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2$ for y_j in majorization (9). The resulting MM update then reduces to weighted medians. In Gaussian mixture models, the Mahalanobis distance $(\boldsymbol{x}_i - \boldsymbol{\theta}_j)^t \boldsymbol{\Omega}^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta}_j)$ leads to explicit updates for both the centers $\boldsymbol{\theta}_j$ and the shared covariance matrix $\boldsymbol{\Omega}$. Other measures such as ϕ-divergences or Bregman divergences may be desirable, for instance with exponential family mixtures (Banerjee et al., 2005; Lucic et al., 2016). Here, we would expect the power means framework to confer similar benefits over Bregman hard clustering (Banerjee et al., 2005) just as it improves k-means under Euclidean distances. Derivative expressions applicable to such extensions are readily available (Teboulle, 2007).

4. Results and Performance

In our simulation study, we generated $n = 2500$ multivariate standard normal samples from $k = 50$ clusters, whose centers $\boldsymbol{\theta}_j$ were initialized randomly in the hypercube with entries $\theta_{lj} \sim r \cdot \text{Unif}(0, 1)$ for $l = 1, \ldots, d$. When $d = 2$, this experiment is the same as that considered in the original KHM paper of Zhang (2001); a similar setting was investigated by Pelleg & Moore (1999) and recreated by Hamerly & Elkan (2002). Although these studies feature various between-cluster variances controlled by the scale factor r, they all generate data from well-separated Gaussians. Modeled after these benchmarks, our simulations represent ideal conditions for Lloyd's algorithm, enabling a generous, conservative comparison. In each dataset, we randomly draw $r \sim \text{Unif}(30, 60)$ and repeat for various dimensions d. All algorithms were run until relative change in objective fell below $\epsilon = 10^{-6}/\sqrt{d}$.

We compare Lloyd's algorithm, KHM, and power k-means with initial power s_0 under matched initial centers, seeded using k-means++. Table 1 reports performance in terms of the quality ratio

$$\sqrt{f_{-\infty}(\hat{\boldsymbol{\Theta}}) / f_{-\infty}(\hat{\boldsymbol{\Theta}}_{\text{optimal}})} \tag{12}$$

considered by Zhang (2001); Hamerly & Elkan (2002) over 50 synthetic datasets. Though (12) is natural and clearly relevant from an optimization perspective, lower values of $f_{-\infty}(\boldsymbol{\Theta})$ do not directly translate into better clusterings. Solutions are therefore also evaluated using the variation of information (VI), an information-theoretic distance that is agnostic to label permutations (Meilă, 2007). Average VI between outputs and the true clusters is reported in Table 2. Best performers for each d are marked with asterisks and boldfaced. We observe that KHM is competitive with Lloyd's algorithm in low dimensions, consistent with previous findings (Zhang et al., 1999; Zhang, 2001; Hamerly & Elkan, 2002); both break down as d grows. The difference in cluster quality is especially apparent in terms of VI. Figure 2 provides a visual comparison of power means and

Table 1. Average Root k-Means Quality Ratio, with (Standard Errors) Below

	$d = 2$	$d = 5$	$d = 10$	$d = 20$	$d = 50$	$d = 100$	$d = 200$
Lloyd	1.036	1.236	1.363	1.411	1.476	1.492	1.481
	(0.018)	(0.057)	(0.122)	(0.130)	(0.123)	(0.110)	(0.123)
KHM	1.044	1.290	1.473	1.504	1.556	1.586	1.556
	(0.029)	(0.063)	(0.115)	(0.135)	(0.154)	(0.135)	(0.147)
$s_0 = -1.0$	***1.029**	***1.164**	1.185	1.221	1.178	1.181	1.149
	(0.018)	(0.047)	(0.065)	(0.079)	(0.059)	(0.067)	(0.060)
$s_0 = -3.0$	1.030	1.187	***1.155**	***1.110**	***1.044**	***1.054**	***1.059**
	(0.017)	(0.043)	(0.058)	(0.064)	(0.055)	(0.059)	(0.056)
$s_0 = -9.0$	1.032	1.220	1.293	1.296	1.192	1.086	1.069
	(0.018)	(0.054)	(0.10)	(0.104)	(0.088)	(0.070)	(0.077)
$s_0 = -18.0$	1.034	1.228	1.328	1.370	1.351	1.254	1.203
	(0.018)	(0.056)	(0.118)	(0.116)	(0.107)	(0.117)	(0.104)

Table 2. Average Variation of Information, with (Standard Errors) Below

	$d = 2$	$d = 5$	$d = 10$	$d = 20$	$d = 50$	$d = 100$	$d = 200$
Lloyd	0.637	0.261	0.234	0.223	0.199	0.206	0.183
	(0.160)	(0.055)	(0.077)	(0.055)	(0.057)	(0.059)	(0.044)
KHM	0.651	0.328	0.339	0.319	0.263	0.280	0.231
	(0.153)	(0.086)	(0.086)	(0.074)	(0.072)	(0.074)	(0.052)
$s_0 = -1.0$	***0.593**	***0.199**	0.133	0.136	0.084	0.087	0.069
	(0.134)	(0.046)	(0.046)	(0.054)	(0.034)	(0.035)	(0.037)
$s_0 = -3.0$	0.593	0.226	***0.111**	***0.069**	***0.022**	***0.027**	***0.026**
	(0.139)	(0.054)	(0.044)	(0.039)	(0.029)	(0.029)	(0.026)
$s_0 = -9.0$	0.608	0.252	0.199	0.169	0.078	0.036	***0.026**
	(0.143)	(0.053)	(0.074)	(0.055)	(0.038)	(0.031)	(0.027)
$s_0 = -18.0$	0.615	0.259	0.218	0.208	0.140	0.101	0.077
	(0.152)	(0.056)	(0.078)	(0.057)	(0.048)	(0.049)	(0.040)

Lloyd's algorithm when $d = 2$. Though KHM preceded the work of Arthur & Vassilvitskii (2007), it is noteworthy that KHM performs slightly worse than Lloyd's even in the plane under k-means++ seeding. Remarkably, power k-means performs best in *all* settings and under *all* choices of s_0 over a broad range. Though this suggests that s_0 can be chosen quite freely rather than requiring careful tuning, we see that choosing s_0 judiciously only further improves performance. As proof of concept, the bottom rows of each table verify what we would expect intuitively. Namely, power k-means behaves more like Lloyd's as the initial power s_0 decreases, though in practice there is no clear reason to initialize at a very low s_0.

The Supplement includes further comparisons on the BIRCH ($n = 100\,000, d = 2$) and MNIST ($n = 60\,000, d = 784$) benchmark datasets, as well as additional results in terms of adjusted Rand index (ARI) and under uniform random initializations. We advocate VI over the popular ARI because the latter is not a metric. Nonetheless, these various measures indicate the same trends across settings, revealing marked improvements under power k-means and reinforcing the finding that our method systematically reaches better solutions and is more stable across initial guesses. All simulations were implemented in Julia (Bezanson et al., 2017) and conducted on a standard Macbook laptop. To give a sense of runtimes absent a detailed comparison, the power k-means algorithm typically converges in around 40 iterations on MNIST, or roughly 20 seconds, under half of what Lloyd's algorithm requires (with more details in the Supplement). Recent work by Shah & Koltun (2017) shows that Lloyd's algorithm outperforms more complicated methods such as affinity propagation (Frey & Dueck, 2007) to cluster the MNIST data, which requires roughly 40 hours on a more powerful machine.

4.1. Protein data

We now analyze protein expression data collected in a murine study of trisomy 21, more commonly known as Down syndrome (Ahmed et al., 2015; Higuera et al., 2015). The dataset contains 1080 expression level measurements of 77 proteins that signal in the nuclear fraction of the cortex. The mice can be classified into eight classes (trisomic or not,

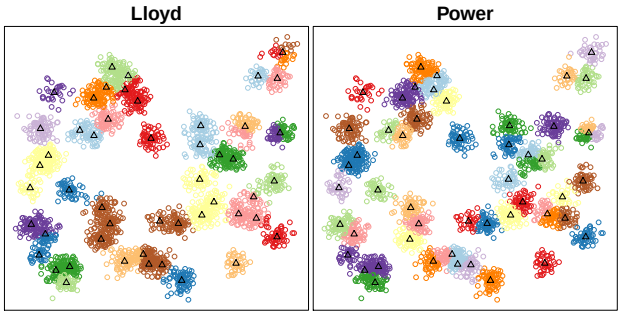


Figure 2. Visualization of solutions obtained using Lloyd's algorithm compared to power k-means. Optimal centers (via running Lloyd's algorithm initialized at true centers) are depicted as triangles. Even in two dimensions, Lloyd's algorithm fails on several visibly well-separated clusters that power means correctly identifies. Tables 1 and 2 show in detail that this difference becomes more pronounced as dimension increases.

conditioned to learn or not, received drug treatment or not). Higuera et al. (2015) employ self-organizing maps (SOM) to identify subsets of proteins that contribute to learning. These authors cluster the trisomic and control mice using separate SOMs and assess quality via quantization error $f_{-\infty}(\hat{\boldsymbol{\Theta}})$, number of mixed class nodes (assigned mice from more than one class), and total observations assigned to mixed nodes. These measures are displayed in the columns of Table 3. Comparison to the best results obtained by Higuera et al. indicates that while SOM outperforms k-means++, power k-means is superior to both under the same measures. As scientific conclusions about critical protein responses depend fundamentally on the clustering step, power k-means presents the preferred alternative in terms of quality. While SOMs additionally provide a low-dimensional visualization of the data (Higuera et al., 2015), various methods such as t-SNE (Maaten & Hinton, 2008) can be straightforwardly applied if planar representations are desired. Additional plots and further details of this case study appear in the Supplement.

Table 3. Protein expression level clustering quality, mouse trisomy learning study (Higuera et al., 2015)

	Error	Mixed nodes	Total mixed
Control mice			
SOM	0.579	8	110
Power	**0.570**	**7**	**92**
k-means++	0.592	11	164
Trisometric mice			
SOM	0.698	5	84
Power	**0.693**	**4**	**70**
k-means++	0.718	9	152

5. Discussion

We present *power k-means*, a novel algorithm for the classical task of k-means clustering. Based on incrementally reducing a sequence of power means objectives, our new method retains the simplicity and low time-complexity of Lloyd's algorithm. Capitalizing on the majorization-minimization principle and the classical mathematical theory behind power means, we derive several nice properties of our algorithm that translate to practical merits. Power k-means naturally benefits from annealing as the underlying objective tends to the k-means objective, providing reduced sensitivity to initialization and substantially improved performance in high dimensions. The method generalizes and extends k-harmonic means clustering, rescuing a good idea from its limitations to low-dimensional problems.

Though power k-means requires an initial power s_0 to be specified, it does not require careful tuning. We have shown that even under ideal assumptions for Lloyd's algorithm, power k-means outperforms when s_0 is carelessly chosen from a broad range. Therefore, as a drop-in replacement for Lloyd's algorithm, improved performance can be expected, and these gains can be further maximized by tuning s_0 if desired. Although we focus on the Euclidean case, we have noted extensions via alternatives to the squared Euclidean distance, and a closer look at these modified versions is warranted. Algorithms for k-means are not only themselves widely used for clustering, but are useful for dimension reduction and as subroutines or initializations in more complex methods. Because power k-means copes better as dimension increases, our contributions broaden the extent to which such strategies are effective. Further theoretical studies are a fruitful avenue for future research. Convergence analyses may provide insights into designing optimal annealing schedules and more rigorous guidance for best choice of s_0. In particular, a closer look at consistency and statistical rates are warranted. While we have seen that power k-means can benefit from k-means++ initialization, other wrapper methods such as geometric acceleration approaches (Elkan, 2003; Ryšavỳ & Hamerly, 2016) may further improve performance and scalability. These directions remain open and ripe for further research.

Acknowledgements

We thank the anonymous referees for their insightful and constructive comments. Kenneth Lange was supported by grants from the National Human Genome Research Institute (HG006139) and the National Institute of General Medical Sciences (GM053275).

References

Ahmed, M. M., Dhanasekaran, A. R., Block, A., Tong, S., Costa, A. C., Stasko, M., and Gardiner, K. J. Protein dynamics associated with failed and rescued learning in the Ts65Dn mouse model of down syndrome. *PloS one*, 10(3):e0119491, 2015.

Aloise, D., Deshpande, A., Hansen, P., and Popat, P. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

Arthur, D. and Vassilvitskii, S. How slow is the k-means method? In *Proceedings of the Twenty-second Annual Symposium on Computational geometry*, pp. 144–153. ACM, 2006.

Arthur, D. and Vassilvitskii, S. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035. Society for Industrial and Applied Mathematics, 2007.

Bachem, O., Lucic, M., Hassani, H., and Krause, A. Fast and provably good seedings for k-means. In *Advances in Neural Information Processing Systems*, pp. 55–63, 2016.

Bachem, O., Lucic, M., Hassani, S. H., and Krause, A. Uniform deviation bounds for k-means clustering. In *International Conference on Machine Learning*, pp. 283–291, 2017.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(Oct):1705–1749, 2005.

Becker, M. P., Yang, I., and Lange, K. EM algorithms without missing data. *Statistical Methods in Medical Research*, 6:38–54, 1997.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

Bezdek, J. C., Ehrlich, R., and Full, W. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203, 1984.

Bietti, A. and Mairal, J. Stochastic optimization with variance reduction for infinite datasets with finite sum structure. In *Advances in Neural Information Processing Systems*, pp. 1622–1632, 2017.

Celebi, M. E., Kingravi, H. A., and Vela, P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1):200–210, 2013.

Chi, E. C. and Lange, K. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015.

Chi, J. T., Chi, E. C., and Baraniuk, R. G. k-pod: A method for k-means clustering of missing data. *The American Statistician*, 70(1):91–99, 2016.

Dasgupta, S. and Freund, Y. Random projection trees for vector quantization. *IEEE Transactions on Information Theory*, 55(7):3229–3242, 2009.

de Carvalho, M. Mean, what do you mean? *The American Statistician*, 70(3):270–274, 2016.

Dinh, V. C., Ho, L. S., Nguyen, B., and Nguyen, D. Fast learning rates with heavy-tailed losses. In *Advances in Neural Information Processing Systems*, pp. 505–513, 2016.

Elkan, C. Using the triangle inequality to accelerate k-means. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 147–153, 2003.

Everitt, B., Landau, S., Leese, M., and Stahl, D. *Cluster Analysis*. Wiley, 2011.

Forster, D. and Lücke, J. Can clustering scale sublinearly with its clusters? A variational EM acceleration of GMMs and k-means. In *International Conference on Artificial Intelligence and Statistics*, pp. 124–132, 2018.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Frey, B. J. and Dueck, D. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

Güngör, Z. and Ünler, A. k-harmonic means data clustering with simulated annealing heuristic. *Applied Mathematics and Computation*, 184(2):199–209, 2007.

Hamerly, G. and Elkan, C. Alternatives to the k-means algorithm that find better clusterings. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pp. 600–607. ACM, 2002.

Hardy, G. H., Littlewood, J. E., and Pólya, G. *Inequalities*. Cambridge University Press, 1952.

Heller, K. A. and Ghahramani, Z. Bayesian hierarchical clustering. In *Proceedings of the 22nd international conference on Machine learning*, pp. 297–304. ACM, 2005.

Higuera, C., Gardiner, K. J., and Cios, K. J. Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. *PloS one*, 10(6): e0129126, 2015.

Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

Kearns, M., Mansour, Y., and Ng, A. Y. An information-theoretic analysis of hard and soft assignment methods for clustering. In *Learning in Graphical Models*, pp. 495–520. Springer, 1998.

Kolmogorov, A. N. and Castelnuovo, G. *Sur la notion de la moyenne*. G. Bardi, tip. della R. Accad. dei Lincei, 1930.

Kriegel, H.-P., Kröger, P., and Zimek, A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1, 2009.

Kulis, B. and Jordan, M. I. Revisiting k-means: new algorithms via Bayesian nonparametrics. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1131–1138. Omnipress, 2012.

Lange, K. *MM Optimization Algorithms*. SIAM, 2016.

Lange, K., Hunter, D. R., and Yang, I. Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics*, 9: 1–20, 2000.

Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

Lu, Y. and Zhou, H. H. Statistical and computational guarantees of Lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

Lucic, M., Bachem, O., and Krause, A. Strong coresets for hard and soft Bregman clustering with applications to exponential family mixtures. In *Artificial Intelligence and Statistics*, pp. 1–9, 2016.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov): 2579–2605, 2008.

MacKay, D. J. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

Mairal, J. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

Medasani, S., Kim, J., and Krishnapuram, R. An overview of membership function generation techniques for pattern recognition. *International Journal of Approximate Reasoning*, 19(3-4):391–417, 1998.

Meilă, M. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

Mirkin, B. Mathematical classification and clustering: From how to what and why. In *Classification, Data Analysis, and Data Highways*, pp. 172–181. Springer, 1998.

Ng, A. Y., Jordan, M. I., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pp. 849–856, 2002.

Ostrovsky, R., Rabani, Y., Schulman, L. J., and Swamy, C. The effectiveness of Lloyd-type methods for the k-means problem. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pp. 165–176. IEEE, 2006.

Pelleg, D. and Moore, A. Accelerating exact k-means algorithms with geometric reasoning. In *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 277–281. ACM, 1999.

Rockafellar, R. T. *Convex Analysis*, volume 28. Princeton University Press, 1970.

Ryšavỳ, P. and Hamerly, G. Geometric methods to accelerate k-means algorithms. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 324–332. SIAM, 2016.

Shah, S. A. and Koltun, V. Robust continuous clustering. *Proceedings of the National Academy of Sciences*, 114 (37):9814–9819, 2017.

Steele, J. M. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

Steinhaus, H. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1(804):801, 1956.

Teboulle, M. A unified continuous optimization framework for center-based clustering methods. *Journal of Machine Learning Research*, 8(Jan):65–102, 2007.

Tian, Y., Liu, D., and Qi, H. k-harmonic means data clustering with differential evolution. In *2009 International Conference on Future BioMedical Information Engineering (FBIE)*, 2009.

Ueda, N. and Nakano, R. Deterministic annealing EM algorithm. *Neural Networks*, 11(2):271–282, 1998.

Vidal, R. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

Xu, J., Chi, E., and Lange, K. Generalized linear model regression under distance-to-set penalties. In *Advances in Neural Information Processing Systems*, pp. 1385–1394, 2017.

Yang, F., Sun, T., and Zhang, C. An efficient hybrid data clustering method based on k-harmonic means and particle swarm optimization. *Expert Systems with Applications*, 36(6):9847–9852, 2009.

Yin, M., Hu, Y., Yang, F., Li, X., and Gu, W. A novel hybrid k-harmonic means and gravitational search algorithm approach for clustering. *Expert Systems with Applications*, 38(8):9319–9324, 2011.

Zhang, B. Generalized k-harmonic means—dynamic weighting of data in unsupervised learning. In *Proceedings of the 2001 SIAM International Conference on Data Mining*, pp. 1–13. SIAM, 2001.

Zhang, B., Hsu, M., and Dayal, U. k-harmonic means—a data clustering algorithm. *Hewlett-Packard Labs Technical Report HPL-1999-124*, 1999.