Entropy Weighted Power k-Means Clustering

Saptarshi Chakraborty* Indian Statistical Institute, Kolkata, India

Debolina Paul* Indian Statistical Institute, Kolkata, India

Swagatam Das Electronics and Communication Sciences Unit, Indian Statistical Institute

Jason Xu Department of Statistical Science, Duke University

Abstract

Despite its well-known shortcomings, k-means remains one of the most widely used approaches to data clustering. Current research continues to tackle its flaws while attempting to preserve its simplicity. Recently, the *power k-means* algorithm was proposed to avoid poor local minima by annealing through a family of smoother surfaces. However, the approach lacks statistical guarantees and fails in high dimensions when many features are irrelevant. This paper addresses these issues by introducing *entropy regularization* to learn feature relevance while annealing. We prove consistency of the proposed approach and derive a scalable majorization-minimization algorithm that enjoys closed-form updates and convergence guarantees. In particular, our method retains the same computational complexity of k-means and power k-means, but yields significant improvements over both. Its merits are thoroughly assessed on a suite of real and synthetic data experiments.

1 Introduction

Clustering is a fundamental task in unsupervised learning for partitioning data into groups based on some similarity measure. Perhaps the most popular approach is k-means clustering (MacQueen, 1967): given a dataset $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^p$, $\mathcal{X}$ is to be partitioned into k mutually exclusive classes so that the variance within each cluster is minimized. The problem can be cast as minimization of the objective

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

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k\}$ denotes the set of cluster centroids, and $\|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2$ is the usual squared Euclidean distance metric.

Lloyd's algorithm (Lloyd, 1982), which iterates between assigning points to their nearest centroid and updating each centroid by averaging over its assigned points, is the most frequently used heuristic to solve the preceding minimization problem. Such greedy approaches, however, suffer from several well-documented drawbacks. Because the task is NP-hard (Aloise et al., 2009), Lloyd's algorithm and its variants seek to approximately solve the problem and are prone to stopping at poor local minima, especially as the number of clusters k and dimension p grow. Many new variants have since contributed to a vast literature on the topic, including spectral clustering (Ng et al., 2002), Bayesian (Lock and Dunson, 2013) and nonparametric methods (Kulis and Jordan, 2012), subspace clustering (Vidal, 2011), sparse clustering (Witten and Tibshirani, 2010), and convex clustering (Chi and Lange, 2015); a more comprehensive overview can be found in Jain (2010).

None of these methods have managed to supplant k-means clustering, which endures as the most widely used approach among practitioners due to its simplicity. Some work instead focuses on "drop-in" improvements of Lloyd's algorithm. The most prevalent strategy is clever seeding: k-means++ (Arthur and Vassilvitskii, 2007; Ostrovsky et al., 2012) is one such effective wrapper method in theory and practice, and proper initialization methods remain an active area of research (Celebi et al., 2013; Bachem et al., 2016). Geometric arguments have also been employed to overcome sensitivity to initialization. Zhang et al. (1999) proposed to replace the minimum function by the harmonic mean function to yield a smoother objective function landscape but retain a similar algorithm, though the strategy fails in all but very low dimensions. Xu and Lange (2019) generalized this idea

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

*Co-first authors contributed equally

Corresponding author: jason.q.xu@duke.edu

by using a sequence of successively smoother objectives via power means instead of the harmonic mean function to obtain better approximating functions in each iteration. The contribution of power k-means is algorithmic in nature—it effectively avoids local minima from an *optimization* perspective, and succeeds for large p when the data points are well-separated. However, it does not address the *statistical* challenges in high-dimensional settings and can perform as poorly as standard k-means in such cases. A meaningful similarity measure plays a key role in revealing clusters (De Amorim and Mirkin, 2012; Chakraborty and Das, 2017), but pairwise Euclidean distances become decreasingly informative as the number of features grows due to the curse of dimensionality.

On the other hand, there is a rich literature on clustering in high dimensions, but standard approaches such as subspace clustering are not scalable due to the use of an affinity matrix pertaining to norm regularization (Ji et al., 2014; Liu et al., 2012). For spectral clustering, even the creation of such a matrix quickly becomes intractable for modern, large-scale problems (Zhang et al., 2019). Toward learning effective feature representations, weighted k-means clustering (WK-means) (Huang et al., 2005), and sparse k-means (Witten and Tibshirani, 2010) has become a benchmark feature selection algorithm, where selection is achieved by imposing ℓ_1 and ℓ_2 constraints on the feature weights. Further related developments can be found in the works of Modha and Spangler (2003); Li and Yu (2006); Huang et al. (2008); De Amorim and Mirkin (2012); Jin and Wang (2016). These approaches typically lead to complex optimization problems in terms of transparency as well as computational efficiency—for instance, sparse k-means requires solving constrained sub-problems via bisection to find the necessary dual parameters λ^* in evaluating the proximal map of the ℓ_1 term. As they fail to retain the simplicity of Lloyd's algorithm for k-means, they lose appeal to practitioners. Moreover, these works on feature weighting and selection do not benefit from recent algorithmic developments mentioned above.

In this article, we propose a scalable clustering algorithm for high dimensional settings that leverages recent insights for avoiding poor local minima, performs adaptive feature weighing, and preserves the low complexity and transparency of k-means. Called Entropy Weighted Power k-means (EWP), we extend the merits of power k-means to the high-dimensional case by introducing feature weights together with entropy incentive terms. Entropy regularization is not only effective both theoretically and empirically, but leads to an elegant algorithm with closed-form solution updates. The idea is to minimize along a continuum of smooth surrogate functions that gradually approach the k-means objective, while the feature space also gradually adapts so that clustering is driven by informative features. By transferring the task onto a sequence of better-behaved optimization landscapes, the algorithm fares better against the curse of dimensionality and against adverse initialization of the cluster centroids than existing methods. The following summarizes our main contributions:

- We propose a clustering framework that automatically learns a weighted feature representation while simultaneously avoiding local minima through annealing.
- We develop a scalable Majorization-Minimization (MM) algorithm to minimize the proposed objective function.
- We establish descent and convergence properties of our method and prove the strong consistency of the global solution.
- Through an extensive empirical study on real and simulated data, we demonstrate the efficacy of our algorithm, finding that it outperforms comparable classical and state-of-the-art approaches.

The rest of the paper is organized as follows. After reviewing some necessary background, Section 2.1 formulates the Entropy Weighted Power k-means (EWP) objective and provides high-level intuition. Next, an MM algorithm to solve the resulting optimization problem is derived in Section 2.2. Section 3 establishes the theoretical properties of the EWP clustering. Detailed experiments on both real and simulated datasets are presented in Section 4, followed by a discussion of our contributions in Section 5.

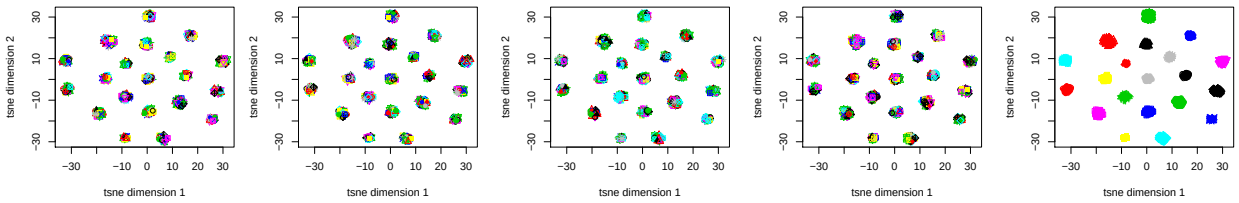
Majorization-minimization The principle of MM has become increasingly popular for large-scale optimization in statistical learning (Mairal, 2015; Lange, 2016). Rather than minimizing an objective of interest f directly, an MM algorithm successively minimizes a sequence of simpler *surrogate functions* $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)$ that *majorize* the original objective $f(\boldsymbol{\theta})$ at the current estimate $\boldsymbol{\theta}_m$. Majorization requires two conditions: 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 iterates of the MM algorithm are defined by the rule

$$\boldsymbol{\theta}_{m+1} := \arg\min_{\boldsymbol{\theta}} \; g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m), \tag{2}$$

which immediately implies the descent property

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

That is, a decrease in g results in a decrease in f. Note that $g(\boldsymbol{\theta}_{m+1} \mid \boldsymbol{\theta}_m) \leq g(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}_m)$ does not require



(a) k-means (b) WK-means (c) Power k-means (d) Sparse k-means (e) EWP

Figure 1: Peer methods fail to cluster in 100 dimensions with 5 effective features on illustrative example, while the proposed method achieves perfect separation. Solutions are visualized using `t-SNE`.

$\boldsymbol{\theta}_{m+1}$ to minimize g exactly, so that any descent step in g suffices. The MM principle offers a general prescription for transferring a difficult optimization task onto a sequence of simpler problems (Lange et al., 2000), and includes the well-known EM algorithm for maximum likelihood estimation under missing data as a special case (Becker et al., 1997).

Power k-means Zhang et al. (1999) attempt to reduce the sensitivity to initialization of k-means by minimizing the criterion

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

Known as k-harmonic means, the method replaces the min appearing in (1) by the harmonic average to yield a smoother optimization landscape, an effective approach in low dimensions. Recently, power k-means clustering extends this idea to work in higher dimensions where (3) is no longer a good proxy for (1). Instead of considering only the closest centroid or the harmonic average, the *power mean* between each point and all k centroids provides a family of successively smoother optimization landscapes. The power mean of a vector $\boldsymbol{y}$ is defined $M_s(\boldsymbol{y}) = \left(\frac{1}{k}\sum_{i=1}^{k} y_i^s\right)^{1/s}$. Within this class, $s > 1$ corresponds to the usual ℓ_s-norm of $\boldsymbol{y}$, $s = 1$ to the arithmetic mean, and $s = -1$ to the harmonic mean. Power means enjoy several properties that translate to algorithmic merits and are useful for establishing theoretical guarantees. They are monotonic, homogeneous, and differentiable with gradient

$$\frac{\partial}{\partial y_j}M_s(\boldsymbol{y}) \quad = \quad \left(\frac{1}{k}\sum_{i=1}^{k} y_i^s\right)^{\frac{1}{s}-1}\frac{1}{k}y_j^{s-1}, \tag{4}$$

and satisfy the limits

$$\lim_{s\to-\infty} M_s(\boldsymbol{y}) = \min\{y_1,\ldots,y_k\} \tag{5a}$$

$$\lim_{s\to\infty} M_s(\boldsymbol{y}) = \max\{y_1,\ldots,y_k\}. \tag{5b}$$

Further, the well-known power mean inequality $M_s(\boldsymbol{y}) \le M_t(\boldsymbol{y})$ for any $s \le t$ holds (Steele, 2004). The power k-means objective function for a given power s is given by the formula

$$f_s(\Theta) = \sum_{i=1}^{n} M_s(\|\boldsymbol{x}_i-\boldsymbol{\theta}_1\|^2,\ldots,\|\boldsymbol{x}_i-\boldsymbol{\theta}_k\|^2). \tag{6}$$

The algorithm then seeks to minimize f_s iteratively while sending $s \to -\infty$. Doing so, the objective approaches $f_{-\infty}(\Theta)$ due to (5), coinciding with the original k-means objective and retaining its interpretation as minimizing within-cluster variance. The intermediate surfaces provide better optimization landscapes that exhibit fewer poor local optima than (1). Each minimization step is carried out via MM; see Xu and Lange (2019) for details.

2 Entropy Weighted Power k-means

A Motivating Example We begin by considering a synthetic dataset with $k = 20$ clusters, $n = 1000$ points, and $p = 100$. Of the 100 features, only 5 are relevant for distinguishing clusters, while the others are sampled from a standard normal distribution (further details are described later in Simulation 2 of Section 4.1). We compare standard k-means, WK-means, power k-means, and sparse k-means with our proposed method; sparse k-means is tuned using the gap statistic described in the original paper Witten and Tibshirani (2010) as implemented in the `R` package `sparcl`. Figure 1 displays the solutions in a t-distributed Stochastic Neighbourhood Embedding (`t-SNE`) (Maaten and Hinton, 2008) for easy visualization in two dimensions. It is evident that our EWP algorithm, formulated below, yields perfect recovery while the peer algorithms fail to do so. This transparent example serves to illustrate the need for an approach that simultaneously avoids poor local solutions while accommodating high dimensionality.

2.1 Problem Formulation

Let $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n \in \mathbb{R}^p$ denote the n data points, and $\Theta_{k\times p} = [\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k]^\top$ denote the matrix whose rows contain the cluster centroids. We introduce a feature

relevance vector $\boldsymbol{w} \in \mathbb{R}^p$ where w_l contains the weight of the l-th feature, and require these weights to satisfy the constraints

$$\sum_{l=1}^{p} w_l = 1; \qquad w_l \geq 0 \text{ for all } l = 1, \ldots, p. \tag{C}$$

The EWP objective for a given s is now given by

$$f_s(\Theta, \boldsymbol{w}) = \sum_{i=1}^{n} M_s(\|\boldsymbol{x} - \boldsymbol{\theta}_1\|_{\boldsymbol{w}}^2, \ldots, \|\boldsymbol{x} - \boldsymbol{\theta}_k\|_{\boldsymbol{w}}^2) + \lambda \sum_{l=1}^{p} w_l \log w_l, \tag{7}$$

where the *weighted* norm $\|\boldsymbol{y}\|_{\boldsymbol{w}}^2 = \sum_{l=1}^{p} w_l y_l^2$ now appears as arguments to the power mean M_s. The final term is the negative entropy of $\boldsymbol{w}$ (Jing et al., 2007). This entropy incentive is minimized when $w_l = 1/p$ for all $l = 1, \ldots, p$; in this case, equation (7) is equal to the power k-means objective, which in turn equals the k-means objective when $s \to -\infty$ (and coincides with KHM for $s = -1$). EWP thus generalizes these approaches, while newly allowing features to be adaptively weighed throughout the clustering algorithm. Moreover, we will see in Section 2.2 that entropy incentives are an ideal choice of regularizer in that they lead to closed form updates for $\boldsymbol{w}$ and $\boldsymbol{\theta}$ within an iterative algorithm.

Intuition and the curse of dimensionality Power k-means combats the curse of dimensionality by providing smoothed *objective functions* that remain appropriate as dimension increases. Indeed, in practice the value of s at convergence of power k-means becomes lower as the dimension increases, explaining its outperformance over k-harmonic means (Zhang et al., 1999): f_{-1} deteriorates as a reasonable approximation of $f_{-\infty}$. However even if poor solutions are successfully avoided from the algorithmic perspective, the curse of dimensionality still affects the *arguments* to the objective. Minimizing within-cluster variance becomes less meaningful as pairwise Euclidean distances become uninformative in high dimensions (Aggarwal et al., 2001). It is therefore desirable to reduce the effective dimension in which distances are computed.

While the entropy incentive term does not zero out variables, it weighs the dimensions according to how useful they are in driving clustering. When the data live in a high-dimensional space yet only a small number of features are relevant towards clustering, the optimal solution to our objective (7) assigns non-negligible weights to only those few relevant features, while benefiting from annealing through the weighted power mean surfaces.

2.2 Optimization

To optimize the EWP objective, we develop an MM algorithm (Lange, 2016) for sequentially minimizing (7). As shown in Xu and Lange (2019), $M_s(\boldsymbol{y})$ is concave if $s < 1$; in particular, it lies below its tangent plane. This observation provides the following inequality: denoting $\boldsymbol{y}_m$ the estimate of a variable $\boldsymbol{y}$ at iteration m,

$$M_s(\boldsymbol{y}) \leq M_s(\boldsymbol{y}_m) + \nabla_{\boldsymbol{y}} M_s(\boldsymbol{y}_m)^\top (\boldsymbol{y} - \boldsymbol{y}_m) \tag{8}$$

Substituting $\|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|_{\boldsymbol{w}}^2$ for y_j and $\|\boldsymbol{x}_i - \boldsymbol{\theta}_{mj}\|_{\boldsymbol{w}_m}^2$ for y_{mj} in equation (8) and summing over all i, we obtain

$$f_s(\boldsymbol{\Theta}, \boldsymbol{w}) \leq f_s(\boldsymbol{\Theta}_m, \boldsymbol{w}_m) - \sum_{i=1}^{n}\sum_{j=1}^{k} \phi_{ij}^{(m)} \|\boldsymbol{x}_i - \boldsymbol{\theta}_{mj}\|_{\boldsymbol{w}_m}^2 - \lambda \sum_{l=1}^{p} (w_{m,l} \log w_{m,l} - w_l \log w_l) + \sum_{i=1}^{n}\sum_{j=1}^{k} \phi_{ij}^{(m)} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|_{\boldsymbol{w}}^2.$$

Here the derivative expressions (4) provide the values of the constants

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

The right-hand side of the inequality above serves as a *surrogate function* majorizing $f_s(\boldsymbol{\Theta}, \boldsymbol{w})$ at the current estimate $\boldsymbol{\Theta}_m$. Minimizing this surrogate amounts to minimizing the expression

$$\sum_{i=1}^{n}\sum_{j=1}^{k} \phi_{ij}^{(m)} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|_{\boldsymbol{w}}^2 + \lambda \sum_{l=1}^{p} w_l \log w_l \tag{9}$$

subject to the constraints (C). This problem admits closed form solutions (detailed in the Supplement):

$$\boldsymbol{\theta}_{m+1,j} = \frac{\sum_{i=1}^{n} \phi_{ij}^{(m)} \boldsymbol{x}_i}{\sum_{i=1}^{n} \phi_{ij}^{(m)}} \tag{10}$$

$$w_{m+1,l} = \frac{\exp\left\{-\frac{\sum_{i=1}^{n}\sum_{j=1}^{k} \phi_{ij}^{(m)} (x_{il} - \theta_{jl})^2}{\lambda}\right\}}{\sum_{t=1}^{p} \exp\left\{-\frac{\sum_{i=1}^{n}\sum_{j=1}^{k} \phi_{ij}^{(m)} (x_{it} - \theta_{jt})^2}{\lambda}\right\}}. \tag{11}$$

The MM steps result in updates similar to that of the Lloyd's updates (Lloyd, 1982) in the sense that each step alternates between updating ϕ_{ij}'s and updating $\boldsymbol{\Theta}$ and $\boldsymbol{w}$. These updates are summarised in Algorithm 1; though there are three steps rather than two, the overall per-iteration complexity of this algorithm is the same as that of k-means (and power k-means) at $\mathcal{O}(nkp)$ (Lloyd, 1982). We require the tuning parameter $\lambda > 0$ to be specified, typically chosen via cross-validation detailed in Section 4.1. It should be

noted that the initial value s_0 and the constant η do not require careful tuning: we fix them at $s_0 = -1$ and $\eta = 1.05$ across *all* real and simulated settings considered in this paper.

Algorithm 1: Entropy Weighted Power k-means Algorithm (EWP)

Data: $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\lambda > 0$, $\eta > 1$
Result: $\boldsymbol{\Theta}$
initialize $s_0 < 0$ and $\boldsymbol{\Theta}_0$
repeat:

$$\phi_{ij}^{(m)} \leftarrow \frac{1}{k}\|\boldsymbol{x}_i - \boldsymbol{\theta}_{m,j}\|_{\boldsymbol{w}_m}^{2(s_m-1)} \left(\frac{1}{k}\sum_{j=1}^{k}\|\boldsymbol{x}_i - \boldsymbol{\theta}_{m,j}\|_{\boldsymbol{w}_m}^{2s_m}\right)^{(\frac{1}{s_m}-1)}$$

$$\boldsymbol{\theta}_{m+1,j} \leftarrow \frac{\sum_{i=1}^{n}\phi_{ij}^{(m)}\boldsymbol{x}_i}{\sum_{i=1}^{n}\phi_{ij}^{(m)}}$$

$$w_{m+1,l} \leftarrow \frac{\exp\left\{-\frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\phi_{ij}^{(m)}(x_{il}-\theta_{jl})^2}{\lambda}\right\}}{\sum_{t=1}^{p}\exp\left\{-\frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\phi_{ij}^{(m)}(x_{it}-\theta_{jt})^2}{\lambda}\right\}}$$

$$s_{m+1} \leftarrow \eta s_m$$

until convergence

3 Theoretical Properties

We note that all iterates $\boldsymbol{\theta}_m$ in Algorithm 1 are defined within the convex hull of the data, all weight updates lie within $[0,1]$, and the procedure enjoys convergence guarantees as an MM algorithm (Lange, 2016). Before we state and prove the main result of this section on strong consistency, we present results characterizing the sequence of minimizers. Proven in the Supplement, Theorems 1 and 2 show that the minimizers of surfaces f_s always lie in the convex hull of the data C^k, and converge uniformly to the minimizer of $f_{-\infty}$.

Theorem 1. *Let $s \leq 1$ also let $(\boldsymbol{\Theta}_{n,s}, \boldsymbol{w}_{n,s})$ be minimizer of $f_s(\boldsymbol{\Theta}, \boldsymbol{w})$. Then we have $\boldsymbol{\Theta}_{n,s} \in C^k$.*

Theorem 2. *For any decreasing sequence $\{s_m\}_{m=1}^{\infty}$ such that $s_1 \leq 1$ and $s_m \to -\infty$, $f_{s_m}(\boldsymbol{\Theta}, \boldsymbol{w})$ converges uniformly to $f_{-\infty}(\boldsymbol{\Theta}, \boldsymbol{w})$ on $C^k \times [0,1]^p$.*

Strong consistency is a fundamental requirement of any "good" estimator in the statistical sense: as the number of data points grows, one should be able to recover true parameters with arbitrary precision Terada (2014, 2015); Chakraborty and Das (2019). The proof of our main result builds upon the core argument for k-means consistency by Pollard (1981), and extends the argument through novel arguments involving uniform convergence of the family of annealing functions.

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ be independently and identically distributed from distribution P with support on a compact set $C \subset \mathbb{R}^p$. For notational convenience, we write $\mathcal{M}_s(\boldsymbol{x}, \boldsymbol{\Theta}, \boldsymbol{w})$ for $M_s(\|\boldsymbol{x} - \boldsymbol{\theta}_1\|_{\boldsymbol{w}}, \ldots, \|\boldsymbol{x} - \boldsymbol{\theta}_1\|_{\boldsymbol{w}})$. We consider the following minimization problem

$$\min_{\boldsymbol{\Theta}, \boldsymbol{w}} \frac{1}{n}\sum_{i=1}^{n}\mathcal{M}_s(\boldsymbol{x}_i, \boldsymbol{\Theta}, \boldsymbol{w}) + \lambda\sum_{l=1}^{p} w_l \log w_l,$$

which is nothing but a scaled version of equation (7). Intuitively, as $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n}\mathcal{M}_s(\boldsymbol{x}_i, \boldsymbol{\Theta}, \boldsymbol{w})$ is very close to $\int \mathcal{M}_s(\boldsymbol{x}, \boldsymbol{\Theta}, \boldsymbol{w})dP$ almost surely by appealing to the Strong Law of Large Numbers (SLLN). Together with (5), as $n \to \infty$ and $s \to -\infty$ we expect

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{M}_s(\boldsymbol{x}_i, \boldsymbol{\Theta}, \boldsymbol{w}) + \lambda\sum_{l=1}^{p} w_l \log w_l \tag{12}$$

to be in close proximity of

$$\int \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\boldsymbol{x} - \boldsymbol{\theta}\|_{\boldsymbol{w}} dP + +\lambda\sum_{l=1}^{p} w_l \log w_l, \tag{13}$$

so that minimizers of (12) should be very close to the minimizers of (13) under certain regularity conditions.

To formalize this intuition, let $\boldsymbol{\Theta}^*$, $\boldsymbol{w}^*$ be minimizers of

$$\Phi(\boldsymbol{\Theta}, \boldsymbol{w}) = \int \min_{1\leq j\leq k} \|\boldsymbol{x} - \boldsymbol{\theta}_j\|_{\boldsymbol{w}}^2 dP + \lambda\sum_{l=1}^{p} w_l \log w,$$

and define $\boldsymbol{\Theta}_{n,s}$, $\boldsymbol{w}_{n,s}$ as the minimizers of

$$\int \mathcal{M}_s(\boldsymbol{x}, \boldsymbol{\Theta}, \boldsymbol{w})dP_n + \lambda\sum_{l=1}^{p} w_l \log w_l,$$

where P_n is the empirical measure. We will show that $\boldsymbol{\Theta}_{n,s} \xrightarrow{a.s.} \boldsymbol{\Theta}^*$ and $\boldsymbol{w}_{n,s} \xrightarrow{a.s.} \boldsymbol{w}^*$ as $n \to \infty$ and $s \to -\infty$ under the following identifiability assumption:

A1 For any neighbourhood N of $(\boldsymbol{\Theta}^*, \boldsymbol{w}^*)$, there exists $\eta > 0$ such that if $(\boldsymbol{\Theta}, \boldsymbol{w}) \notin N$ implies that $\Phi(\boldsymbol{\Theta}, \boldsymbol{w}) > \Phi(\boldsymbol{\Theta}^*, \boldsymbol{w}^*) + \eta$.

Theorem 3 establishes a uniform SLLN, which plays a key role in the proof of the main result (Theorem 4).

Theorem 3. *(SLLN) Fix $s \leq 1$. Let $\mathcal{G}$ denote the family of functions $g_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{x}) = \mathcal{M}_s(\boldsymbol{x}, \boldsymbol{\Theta}, \boldsymbol{w})$. Then $\sup_{g\in\mathcal{G}} |\int g dP_n - \int g dP| \to 0$ a.s. $[P]$.*

Proof sketch: The main idea of the proof is to find a finite function class $\mathcal{G}_\epsilon$, for all $\epsilon > 0$, such that for all $g \in \mathcal{G}$, there exists functions $\bar{g}, \dot{g} \in \mathcal{G}_\epsilon$ such that $\dot{g} \leq g \leq \bar{g}$ and $\int(\bar{g} - \dot{g})dP < \epsilon$. This can be achieved by appealing to the compactness (and hence total boundedness) of $C^k \times [0,1]^p$. One might take

$$\mathcal{G}_\epsilon = \{\phi(M_s(\|\boldsymbol{x} - \boldsymbol{\theta}_1'\|_{\boldsymbol{w}'}^2, \ldots, \|\boldsymbol{x} - \boldsymbol{\theta}_k'\|_{\boldsymbol{w}'}^2) \pm \epsilon/2) : \boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_k' \in C_{\delta_1} \text{ and } \boldsymbol{w}' \in W_{\delta_2}\},$$

where $\phi(x) = \max\{x, 0\}$. A rigorous proof of the theorem can be found in the Supplement.

We are now ready to establish the main consistency result, stated and proven below.

Theorem 4. *Under the condition A1,* $\boldsymbol{\Theta}_{n,s} \xrightarrow{a.s.} \boldsymbol{\Theta}^*$ *and* $\boldsymbol{w}_{n,s} \xrightarrow{a.s.} \boldsymbol{w}^*$ *as* $n \to \infty$ *and* $s \to -\infty$.

Proof sketch: Note that it is enough to show that $(\boldsymbol{\theta}_{n,s}, \boldsymbol{w}_{n,s})$ eventually lies in any neighbourhood of $(\boldsymbol{\Theta}^*, \boldsymbol{w}^*)$. By assumption A1, it suffices to establish that $\Phi(\boldsymbol{\theta}_{n,s}, \boldsymbol{w}_{n,s}) \leq \Phi(\boldsymbol{\Theta}^*, \boldsymbol{w}^*) + \eta$, eventually. For notational convenience, we write $\alpha(\boldsymbol{w}) = \lambda \sum_{l=1}^{p} w_l \log w_l$. Now observe that $\Phi(\boldsymbol{\Theta}_{n,s}, \boldsymbol{w}_{n,s}) - \Phi(\boldsymbol{\Theta}^*, \boldsymbol{w}^*) = \xi_1 + \xi_2 + \xi_3$, where,

$$\xi_1 = \Phi(\boldsymbol{\Theta}_{n,s}, \boldsymbol{w}_{n,s}) - \int \mathcal{M}_s(\boldsymbol{x}, \boldsymbol{\Theta}_{n,s}, \boldsymbol{w}_{n,s}) dP - \lambda\alpha(\boldsymbol{w}_{n,s}),$$

$$\xi_2 = \int \mathcal{M}_s(\boldsymbol{x}, \boldsymbol{\Theta}_{n,s}, \boldsymbol{w}_{n,s}) dP - \int \mathcal{M}_s(\boldsymbol{x}, \boldsymbol{\Theta}_{n,s}, \boldsymbol{w}_{n,s}) dP_n,$$

$$\xi_3 = \int \mathcal{M}_s(\boldsymbol{x}, \boldsymbol{\Theta}_{n,s}, \boldsymbol{w}_{n,s}) dP_n + \lambda\alpha(\boldsymbol{w}_{n,s}) - \Phi(\boldsymbol{\Theta}^*, \boldsymbol{w}^*).$$

We need to show that each of the ξ_i's, $i = 1, 2, 3$, can be made smaller than $\eta/3$ as $n \to \infty$ and $s \to -\infty$. ξ_1 can be made smaller than $\eta/3$ by appealing to Theorem 2. ξ_2 can also be made smaller than $\eta/3$ by appealing to Theorem 3. To show that $\xi_3 < \eta/3$, we note that

$$\xi_3 \leq \int \mathcal{M}_s(\boldsymbol{x}, \boldsymbol{\Theta}^*, \boldsymbol{w}^*) dP_n - \int \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}^*} \|\boldsymbol{x} - \boldsymbol{\theta}\|_{\boldsymbol{w}^*} dP$$

since it minimizes $\int \mathcal{M}_s(\boldsymbol{x}, \boldsymbol{\Theta}, \boldsymbol{w}) dP_n + \lambda\alpha(\boldsymbol{w})$ globally. The first term can me made close to to $\int \mathcal{M}_s(\boldsymbol{x}, \boldsymbol{\Theta}^*, \boldsymbol{w}^*) dP$ as $n \to \infty$, while the term $\mathcal{M}_s(\boldsymbol{x}, \boldsymbol{\Theta}^*, \boldsymbol{w}^*) dP$ becomes arbitrarily close to $\int \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}^*} \|\boldsymbol{x} - \boldsymbol{\theta}\|_{\boldsymbol{w}^*} dP$ as $s \to -\infty$. Details of the complete proof can be found in the Supplement.

4 Empirical Performance

We examine the performance of EWP on a variety of simulated and real datasets compared to classical and state-of-the-art peer algorithms. For evaluation purposes, we use the Normalized Mutual Information (NMI) (Vinh et al., 2010) and Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) between the ground-truth partition and the partition obtained by each algorithm. A value of 1 indicates perfect clustering and a value of 0 indicates arbitrary labels. As our algorithm is meant to perform as a drop-in replacement to k-means, we focus comparisons to Lloyd's classic algorithm (Lloyd, 1982), WK-means (Huang et al., 2005), Power k-means (Xu and Lange, 2019) and sparse k-means (Witten and Tibshirani, 2010). It should be noted that sparse k-means already entails higher computational complexity, and we do not exhaustively consider alternate methods which require orders of magnitude higher complexity. In all cases, each algorithm is initiated with the same set of randomly chosen centroids.

Codes: https://github.com/DebolinaPaul/EWP

Table 1: Average NMI values, Simulation 1.

	$d = 5$	$d = 10$	$d = 20$	$d = 50$	$d = 100$
k-means	0.3913	0.3701	0.3674	0.3629	0.3517
WK-means	0.5144	0.50446	0.5050	0.5026	0.5029
Power k-means	0.3924	0.3873	0.3722	0.3967	0.3871
Sparse k-means	0.3679	0.3677	0.3668	0.3675	0.3637
EWP	**0.9641**	**0.9217**	**0.9139**	**0.9465**	**0.9082**

4.1 Synthetic Experiments

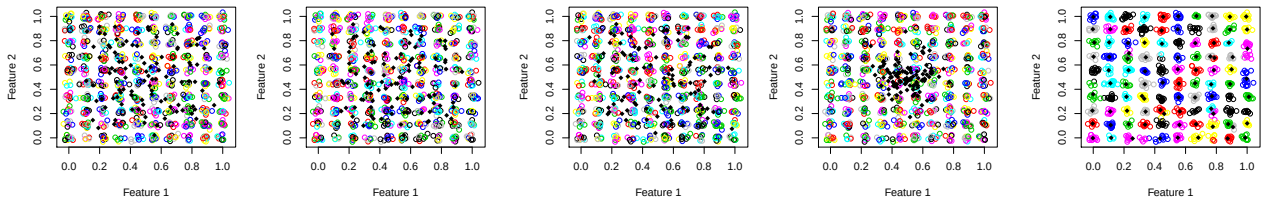
We now consider a suite of simulation studies to validate the proposed EWP algorithm.

Simulation 1 The first experiment assesses performance as the dimension and number of uninformative features grows. We generate $n = 1000$ observations with $k = 100$ clusters. Each observation has $p = d + 2$ many features as d varies between 5 and 100. The first two features reveal cluster structure, while the remaining d variables are uninformative, generated independently from a $Unif(0, 2)$ distribution. True centroids are spaced uniformly on a grid with $\theta_m = \frac{m-1}{10}$, and $x_{ij} \sim \frac{1}{10} \sum_{m=1}^{10} \mathcal{N}(\theta_m, 0.15)$. Despite the simple data generating setup, clustering is difficult due to the low signal to noise ratio in this setting.

We report the average NMI values between the ground-truth partition and the partition obtained by each of the algorithms over 20 trials in Table 1. Details and standard deviations appear in Table S1 of the Supplement. The best performing algorithm in each column appears in bold, and the best solutions for $d = 20$ are plotted in Figure 2. The benefits using EWP are visually stark, and Table 1 verifies in detail that EWP outperforms the classical k-means algorithm, as well as the state-of-the-art sparse-k-means and the power k-means algorithms. The same trends are conveyed under other evaluation metrics such as adjusted Rand index (ARI); see the Supplement. The inability of k-means as well as power k-means to properly learn the feature weights results in poor performance of these algorithms. On the other hand, although WK-means and sparse k-means can select features successfully, they fail from the optimization perspective when k is large and there are many local minima to trap the algorithm.

Simulation 2 We next examine the effect of k on the performance, taking $n = 100 \cdot k$ and $p = 100$ while k varies from 20 to 500. The matrix $\Theta_{k \times p}$, whose rows contain the cluster centroids, is generated as follows.

1. Select 5 relevant features $l_1, \ldots, l_5$ at random.



(a) k-means (b) WK-means (c) Power k-means (d) Sparse k-means (e) EWP

Figure 2: Solutions obtained by the peer algorithms for an example dataset with $k = 100$ and $d = 20$ in Simulation 1 (4.1). The obtained cluster centroids appear as black diamonds in the figure.

2. Simulate $\theta_{j,l_m} \sim Unif(0,1)$ for all $j = 1,\ldots,k$ and $m = 1,\ldots,5$.

3. Set $\theta_{j,l} = 0$ for all $l \notin \{l_1,\ldots,l_5\}$ and all j.

After obtaining Θ, x_{il} is simulated as follows:

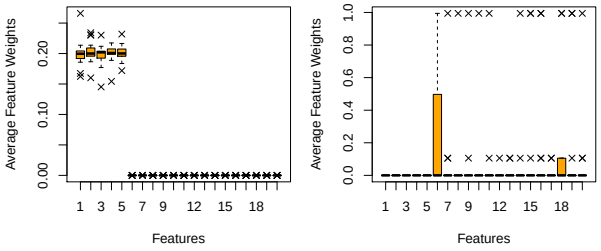
$$x_{il} \sim \mathcal{N}(0,1) \text{ if } l \notin \{l_1,\ldots,l_5\}$$
$$x_{il} \sim \frac{1}{k}\sum_{j=1}^{k}\mathcal{N}(\theta_{j,l}, 0.015) \text{ if } l \in \{l_1,\ldots,l_5\}.$$

We run each of the algorithms 20 times and report the average NMI values between the ground-truth partition an the partition obtained by each of the algorithms in Table 2; again, results in terms of ARI as well as standard errors are tabulated in the Supplement. Table 2 shows that k-means, WK-means, power k-means, and sparse k-means lead to almost the same result, while EWP outperforms the peer algorithms for each k by narrowing down the number of features while avoiding local minima.

Feature Selection We turn to examine the feature weighting properties of the EWP algorithm more closely. We take $n = 1000$, $p = 20$ and follow the same data generation procedure described in Simulation 2. For simplicity, in the first step of the simulation study, we select $l_i = i$ for $i = 1,\ldots,5$. We record the feature weights obtained by EWP and sparse k-means over 100 replicate datasets. The box-plot for these 100 optimal feature weights are shown in Figure 3 for both algorithms. The proposed method successfully assigns almost all weight to relevant features

Table 2: NMI values for Simulation 2, showing the effect of increasing number of clusters.

Algorithm	$k = 20$	$k = 100$	$k = 200$	$k = 500$
k-means	0.0674	0.2502	0.3399	0.3559
WK-means	0.0587	0.2247	0.3584	0.3678
Power k-means	0.0681	0.2785	0.3578	0.3867
Sparse k-means	0.0679	0.2490	0.6705	0.3537
EWP	**0.9987**	**0.9844**	**0.9756**	**0.9908**



(a) EWP (b) Sparse k-means

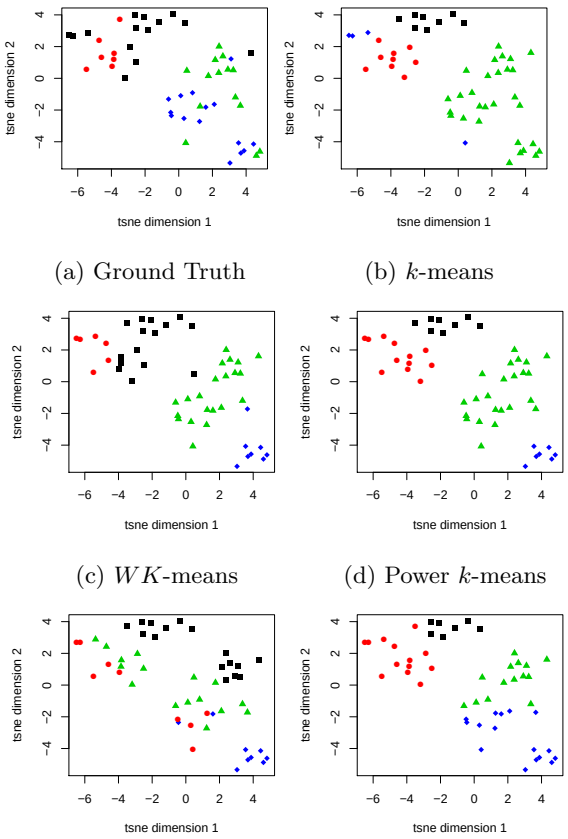
Figure 3: Boxplots show that EWP consistently identifies true features while sparse k-means fails to do so.

1 through 5, even though it does not make use of a sparsity-inducing penalty. Meanwhile, feature weights assigned by sparse k-means do not follow any clear pattern related to informative features, even though the ground truth is sparse in relevant features. The analogous plot for WK-means appears as Figure S1 in the Supplement, which shows even worse performance than sparse k-means. The study clearly illustrates the necessity of feature weighing together with annealing for successful k-means clustering in high dimensions.

4.2 Case Study and Real Data

We now assess performance on real data, beginning with a case study on Glioma. The GLIOMA dataset consists of 50 datapoints and is divided into 4 classes consisting of cancer glioblastomas (CG), noncancer glioblastomas (NG), cancer oligodendrogliomas (CO) and non-cancer oligodendrogliomas (NO). Each observation consists of 4434 features. The data were collected in the study by (Nutt et al., 2003), and are also available in (Li et al., 2018).

In our experimental studies, we compare the EWP algorithm to the four peer algorithms considered in Section 4.1. In order to visualize clustering solutions, we embed the data into the plane via `t-SNE` (Maaten and Hinton, 2008). The best partitioning obtained from each algorithm is shown in Figure 4, which makes it visually clear that clustering under EWP more closely



(a) Ground Truth (b) k-means

(c) WK-means (d) Power k-means

(e) Sparse k-means (f) EWP

Figure 4: `t-SNE` plots for the GLIOMA dataset, color-coded by the partitioning obtained at convergence by each peer algorithm.

resembles the ground truth compared to competitors. This is detailed by average NMI values as well as standard deviations in parentheses listed in Table 3. Analogous results in terms of ARI is reported in the supplement.

Table 3: Mean NMI and (standard deviation), GLIOMA data

k-means	WK-means	Power	Sparse	EWP
0.490(0.040)	0.427(0.034)	0.499(0.020)	0.108(0.001)	**0.594**(0.001)

Further real-data experiments To further validate our method in various real data scenarios, we perform a series of experiments on 10 benchmark datasets collected from the UCI machine learning repository (Dua and Graff, 2017), Keel Repository (Alcalá-Fdez et al., 2011) and ASU repository (Li et al., 2018). A brief description of these datasets can be found in Table S3 of the Supplement.

Average performances over 20 independent trials are reported in Table 4 in terms of NMI, with the analogous results in terms of ARI appearing in the Supplement. The EWP algorithm outperforms by a large margin across all instances when compared to the other peer algorithms. To determine the statistical significance of the results, we employ Wilcoxon's signed-rank test (Wasserman, 2006) at the 5% level of significance. In Table 4, an entry marked with $+$ ($\simeq$) differs from the corresponding result of EWP with statistical significance. Finally, we emphasize that our results comprise a conservative comparison in that parameters $s_0 = -1$ and $\eta = 1.05$ are fixed across *all* settings. While this demonstrates that careful tuning of these parameters is not necessary for successful clustering, performance can be further improved by doing so (Xu and Lange, 2019).

Table 4: NMI values on Benchmark Real Data

Dataset	k-means	WK	Power	Sparse	EWP
Newthyroid	0.403^{+}	0.262^{+}	0.407^{+}	0.102^{+}	**0.581**
Automobile	0.165^{+}	0.203^{+}	0.168^{+}	0.168^{+}	**0.311**
WarpAR10P	0.171+	0.233^{+}	0.201^{+}	0.185^{+}	**0.350**
WarpPIE10P	$0.240^{\simeq}$	$0.241^{\simeq}$	0.180^{+}	0.179^{+}	**0.276**
Iris	0.758^{+}	0.788^{+}	0.742^{+}	$0.814^{\simeq}$	**0.884**
Wine	0.428^{+}	0.642^{+}	0.416^{+}	0.428^{+}	**0.747**
Mammographic	0.107^{+}	0.0194^{+}	0.115^{+}	0.110^{+}	**0.405**
WDBC	0.463^{+}	0.005^{+}	0.464^{+}	0.467^{+}	**0.656**
LIBRAS	$0.553^{\simeq}$	0.339^{+}	0.461^{+}	0.254^{+}	**0.575**
Wall Robot 4	0.168^{+}	0.184^{+}	0.171^{+}	0.186^{+}	**0.234**

5 Discussion

Despite decades of advancement on k-means clustering, Lloyd's algorithm remains the most popular choice in spite of its well-known drawbacks. Extensions and variants that address these flaws fail to preserve its simplicity, scalability, and ease of use. Many of these methods still fall short at poor local optima or fail when data are high-dimensional with low signal-to-noise ratio, and few come with rigorous statistical guarantees such as consistency.

The contributions in this paper seek to fill this methodological gap, with a novel formulation that draws good intuition from classic and recent developments. With the emphasis on simplicity as a chief priority, we derive a method that can be seen as a drop-in replacement to Lloyd's classic k-means algorithm, reaping large improvements in practice even when there are a large number of clusters or features in the data. By designing the algorithm from the perspective of MM, our method is robust as a descent algorithm and achieves an ideal $\mathcal{O}(nkp)$ complexity. In contrast to some popular existing methods such as sparse k-means, the proposed approach is provably consistent.

Extending the intuition to robust measures and other divergences in place of the Euclidean distance are warranted. Further, research toward finite-sample prediction error bounds or convergence rates relating to the annealing schedule will also be fruitful avenues for future work.

References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer.

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., and Herrera, F. (2011). Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17.

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

Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.

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

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

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

Chakraborty, S. and Das, S. (2017). k-means clustering with a new divergence-based distance metric: Convergence and performance analysis. *Pattern Recognition Letters*, 100:67–73.

Chakraborty, S. and Das, S. (2019). On the strong consistency of feature-weighted k-means clustering in a nearmetric space. *STAT*, 8(1):e227.

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

De Amorim, R. C. and Mirkin, B. (2012). Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering. *Pattern Recognition*, 45(3):1061–1075.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Huang, J. Z., Ng, M. K., Rong, H., and Li, Z. (2005). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):657–668.

Huang, J. Z., Xu, J., Ng, M., and Ye, Y. (2008). Weighting method for feature selection in k-means. *Computational Methods of feature selection*, pages 193–209.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

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

Ji, P., Salzmann, M., and Li, H. (2014). Efficient dense subspace clustering. In *IEEE Winter Conference on Applications of Computer Vision*, pages 461–468. IEEE.

Jin, J. and Wang, W. (2016). Influential features PCA for high dimensional clustering. *The Annals of Statistics*, 44(6):2323–2359.

Jing, L., Ng, M. K., and Huang, J. Z. (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on knowledge and data engineering*, 19(8).

Kulis, B. and Jordan, M. I. (2012). Revisiting k-means: New algorithms via Bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*.

Lange, K. (2016). *MM optimization algorithms*, volume 147. SIAM.

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

Li, C. and Yu, J. (2006). A novel fuzzy c-means clustering algorithm. In *International Conference on Rough Sets and Knowledge Technology*, pages 510–515. Springer.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. (2018). Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6):94.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. (2012). Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Lock, E. F. and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616.

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

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

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

Modha, D. S. and Spangler, W. S. (2003). Feature weighting in k-means clustering. *Machine learning*, 52(3):217–237.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856.

Nutt, C. L., Mani, D., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M. E., Batchelor, T. T., et al. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer research*, 63(7):1602–1607.

Ostrovsky, R., Rabani, Y., Schulman, L. J., and Swamy, C. (2012). The effectiveness of Lloyd-type methods for the k-means problem. *Journal of the ACM (JACM)*, 59(6):28.

Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135–140.

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

Terada, Y. (2014). Strong consistency of reduced k-means clustering. *Scandinavian Journal of Statistics*, 41(4):913–931.

Terada, Y. (2015). Strong consistency of factorial k-means clustering. *Annals of the Institute of Statistical Mathematics*, 67(2):335–357.

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

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726.

Xu, J. and Lange, K. (2019). Power k-means clustering. In *International Conference on Machine Learning*, pages 6921–6931.

Zhang, B., Hsu, M., and Dayal, U. (1999). K-harmonic means—a data clustering algorithm. *Hewlett-Packard Labs Technical Report HPL-1999-124*, 55.

Zhang, T., Ji, P., Harandi, M., Huang, W., and Li, H. (2019). Neural collaborative subspace clustering. In *International Conference on Machine Learning*, pages 7384–7393.