Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

K GROUPS: A GENERALIZATION OF K-MEANS

By Songzi Li, Maria Rizzo

Bowling Green State University * *and Bowling Green State University*†

In this paper we propose a new class of distribution-based clustering algorithms called K-groups, which assigns observation to the same cluster if they follow the identical statistical distribution. We propose two different K-groups algorithms: K-groups by first variation and K-groups by second variation, and generate Hartigan and Wong's idea of moving one point to moving $m(m > 1)$ points. For univariate data, we proved that Hartigan and Wong's K-means algorithm is a special case of K-groups by first variation. The simulation results for univariate and multivariate cases show that both K-groups algorithms perform as well as Hartigan and Wong's K-means algorithm when clusters are well-separated and normally distributed. Both K-groups algorithms perform better than K-means when data does not have finite first moment or data has strong skewness and heavy tails. For non–spherical clusters, both K-groups algorithms perform better than K-means in high dimension and K-groups by first variation is consistent as dimension increases. Results of clustering on three real data examples show that both K-groups algorithms perform better than K-means.

Introduction. Cluster analysis is one of the core topics of data mining and has a lot of application domains such as astronomy, psychology, market research and bioinformatics. Clustering is a fundamental tool in unsupervised study which is used to group similar objects together without using external information such as class label. In general, there are two purposes for using cluster analysis: understanding and utility [7]. Understanding of cluster analysis means to find groups of objects that share common characteristics, and utility of cluster analysis attempts to abstract the representative objects from objects in the same groups. The earliest research on cluster analysis can be traced back to 1894, when Karl Pearson used the moment matching method to determine the mixture parameters of two single-variable components [8]. There are many different clustering algorithms, and each algorithm has its own advantages in the certain situation. In this paper we will focus on K-means which is the most popular clustering algorithm.

MSC 2010 subject classifications: Primary K-means, K-groups; secondary First variation, Second variation

K-means. K-means is a prototype-based algorithm which uses cluster mean as the centroid, and assigns the observation to the cluster with nearest centroid. Let $D = \{x_1, \ldots, x_n\} \subset R^m$ is the data set to be clustered. $P = \{\pi_1, \ldots, \pi_K\}$ be a partition of D, where K is the number of cluster set by user. so we have $\cup_i \pi_i = D$, and $\pi_i \cap \pi_j = \emptyset$ if $i \neq j$. The symbol ω_x is the weight of x, n_k is the number of data objects assigned to cluster π_k, $c_k = \sum_{x \in \pi_k} \frac{\omega_x x}{n_k}$ represents the centroid of cluster π_k, $1 \leq k \leq K$. The function $d(x, y)$ is a distance-like function to compute the "dissimilarity" between data object x and y. So the K-means clustering is equivalent to minimization problem.

$$min_{c_k, 1 \leq k \leq K} \sum_{k=1}^{K} \sum_{x \in \pi_k} \omega_x d(x, c_k) \tag{0.1}$$

K-means algorithm is equivalent to search a global minimum problem which is computationally difficult (NP-hard). The standard algorithm was proposed by Stuart LIoyd in 1957 [5]. A more efficient version was proposed and published in Fortran by Hartigan and Wong in 1979 [3]. The distance-like function is one of important factors that influence the performance of K-means. The most common used distance functions are Euclidean quadratic distance, spherical distance, and Kullback-Leibler Divergence [12]. In this paper, we will use a new kind of distance function *Energy Distance*.

Energy Distance. G. J. Székely proposed *Energy Distance* in 1986 [10]. Energy distance is a statistical distance between observations. The concept is based on the notion of Newton's gravitational potential energy, which is a function of the distance between two bodies in a gravitational space.

Definition 0.1. Energy Distance. *The energy distance between the d-dimensional independent random variables X and Y is defined as*

$$\mathcal{E}(X, Y) = 2E|X - Y|_d - E|X - X'|_d - E|Y - Y'|_d,$$

where $E|X|_d < \infty$, $E|Y|_d < \infty$, X' is an independent and identically distributed (iid) copy of X, and Y' is an iid copy of Y.

Let $F(x)$ and $G(x)$ be the cumulative distribution functions, and $\hat{f}(t)$ and $\hat{g}(t)$ be the characteristic functions of independent random variables X and Y, respectively. [9] gave the following proposition .

PROPOSITION 0.1. *If the d-dimensional random variables X and Y are independent with $E|X|_d + E|Y|_d < \infty$, and $\hat{f}, \hat{g}$ denote the their respective characteristic functions, then the energy distance between independent random variables X and Y is*

$$2E|X-Y|_d - E|X-X'|_d - E|Y-Y'|_d = \frac{1}{C_d}\int_{R^d} \frac{|\hat{f}(t)-\hat{g}(t)|^2}{|t|^{d+1}}dt,$$

where

$$C_d = \frac{\pi^{\frac{d+1}{2}}}{\Gamma(\frac{d+1}{2})},$$

and $\Gamma(.)$ is the complete gamma function. Thus, $\mathcal{E}(X,Y) \geq 0$ with equality to zero if and only if X and Y are identically distributed.

Székely proved energy distance is a generalization of Cramér's L_2 distance [10] and gave a generalization of Proposition 0.1.

PROPOSITION 0.2. *Let X and Y be independent d-dimensional random variables with characteristic functions $\hat{f}, \hat{g}$, and $E|X|^\alpha < \infty$, $E|Y|^\alpha < \infty$ for some $0 < \alpha < 2$. If X' is an iid copy of X, and Y' is an iid copy of y, then the energy distance between random variables X and Y is defined as*

$$\mathcal{E}^\alpha(X,Y) = 2E|X-Y|^\alpha - E|X-X'|^\alpha - E|Y-Y'|^\alpha = \frac{1}{C(d,\alpha)}\int_{R^d} \frac{|\hat{f}(t)-\hat{g}(t)|^2}{|t|^{d+\alpha}}dt,$$

where $0 < \alpha < 2$, and

$$C(d,\alpha) = 2\pi^{\frac{d}{2}}\frac{\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma(\frac{d+\alpha}{2})}$$

Note that when $\alpha = 2$, the expression

$$2E|X-Y|^\alpha - E|X-X'|^\alpha - E|Y-Y'|^\alpha$$

measures the distance between means,

$$\mathcal{E}^2(X,Y) = 2|E(X)-E(Y)|^2. \tag{0.2}$$

For all $0 < \alpha < 2$, we have $\mathcal{E}^\alpha(X,Y) \geq 0$ with equality to zero if and only if X and Y are identically distributed; this characterization does not hold for $\alpha = 2$ since we have equality to zero when ever $E(X) = E(Y)$ in (0.2).

The two-sample energy statistic corresponding to energy distance $\mathcal{E}^\alpha(\mathbf{X}, \mathbf{Y})$, for independent random samples $\mathbf{X} = X_1, ..., X_{n_1}$ and $\mathbf{Y} = Y_1, ..., Y_{n_2}$ is

$$\begin{aligned}\mathcal{E}^\alpha_{n_1,n_2}(\mathbf{X}, \mathbf{Y}) &= \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n-2}|X_i - Y_m|^\alpha - \\ &\quad \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|X_i - X_j|^\alpha - \frac{1}{n_2^2}\sum_{l=1}^{n_2}\sum_{m=1}^{n_2}|Y_l - Y_m|^\alpha,\end{aligned}$$

where $\alpha \in (0, 2)$. The weighted two-sample statistic

$$T_{X,Y} = (\frac{n_1 n_2}{n_1 + n_2})\mathcal{E}_{n_1,n_2}(\mathbf{X}, \mathbf{Y})$$

can be applied for testing homogeneity [11] (equality of distributions of X and Y).

K-groups. K-means usually uses quadratic distance as a distance-like function to compute the dissimilarity between the data object and the prototype prespecified, and minimize the variance within the clusters. In this paper, we use a weighted two-sample energy statistic $T_{X,Y}$ as the statistical function to measure the dissimilarity between the clusters, and use the K-means algorithm given by Hartigan and Wong in 1979 [3] to implement our algorithm. Generally, our method belongs to the class of Distribution-Based Algorithms. This kind of algorithms takes a cluster as a dense region of data objects that is surrounded by regions of low densities. They are often employed when the clusters are irregular or intertwined, or when noise and outliers are present. Since the energy distance measures the similarity between two sets rather than the similarity between the object and prototype, we name our method *K-groups*.

We define dispersion between two sets A, B as

$$G^\alpha(A, B) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n_2}|a_i - b_m|^\alpha$$

where $0 < \alpha \leq 2$, and n_1, n_2 are the sample sizes for sets A, B. Let $P = \{\pi_1, ..., \pi_k\}$ be a partition of observations, where k is the number of clusters, prespecified. We define the total dispersion of the observed response by

$$T^\alpha(\pi_1, ...\pi_k) = \frac{N}{2}G^\alpha(\cup_{i=1}^k \pi_i, \cup_{i=1}^k \pi_i),$$

where N is the total number of observations. The within-groups dispersion is defined by

$$W^{\alpha}(\pi_1, ...\pi_k) = \sum_{j=1}^{k} \frac{n_j}{2} G^{\alpha}(\pi_j, \pi_j),$$

where n_j is the sample size for cluster π_j. The between-sample dispersion is

$$B^{\alpha}(\pi_1, ...\pi_k) = \sum_{1 \le i < j \le k} \{\frac{n_i n_j}{2N}(2G^{\alpha}(\pi_i, \pi_j) - G^{\alpha}(\pi_i, \pi_i) - G^{\alpha}(\pi_j, \pi_j))\}.$$

when $0 \le \alpha \le 2$ we have the decomposition

$$T^{\alpha}(\pi_1, ...\pi_k) = W^{\alpha}(\pi_1, ...\pi_k) + B^{\alpha}(\pi_1, ...\pi_k),$$

and both $W^{\alpha}(\pi_1, ...\pi_k)$ and $B^{\alpha}(\pi_1, ...\pi_k)$ are nonnegative. Because to maximize between-sample dispersion $B^{\alpha}(\pi_1, ...\pi_k)$, with $T^{\alpha}(\pi_1, ...\pi_k)$ constant, it is equivalent to minimize $W^{\alpha}(\pi_1, ...\pi_k)$. So our purpose is to find the best partitions which minimize the W^{α}. So the objective function for the K-Groups is

$$min_{\pi_1,...,\pi_k} \sum_{j=1}^{k} \frac{n_j}{2} G^{\alpha}(\pi_j, \pi_j) = min_{\pi_1,...,\pi_k} W^{\alpha}(\pi_1, ...\pi_k). \tag{0.3}$$

First Variation Algorithm. Motivated by the Hartigan and Wong's idea, we search for a K-partition with locally optimal W^{α} by moving points form one cluster to another. We call this reallocation step First Variation.

DEFINITION 0.2. *A first variation of a partition P is a partition P' obtained from P by removing a single point $\mathbf{a}$ from a cluster π_i of P and assigning this point to an existing cluster π_j of P.*

Let π_1 and π_2 are two different clusters in partition $P = \pi_1, ..., \pi_k$, and point $\mathbf{a} \in \pi_i$. cluster π_1^- represents cluster π_1 after removing point $\mathbf{a}$, and cluster π_2^+ represents cluster π_2 after adding point $\mathbf{a}$. Let n_1 and n_2 are the size of cluster π_1 and π_2. The dispersion of cluster π_1 and π_2 are

$$G^{\alpha}(\pi_1, \pi_1) = \frac{1}{2n_1} \sum_{i}^{n_1} \sum_{j}^{n_1} |x_i^1 - x_j^1|^{\alpha},$$

$$G^{\alpha}(\pi_2, \pi_2) = \frac{1}{2n_2} \sum_{i}^{n_2} \sum_{j}^{n_2} |x_i^2 - x_j^2|^{\alpha},$$

where $x_i^1 \in \pi_1, i = 1, ..., n_1$ and $x_i^2 \in \pi_2, i = 1, ..., n_2$. The dispersion of cluster π_1^- and π_2^+ are

$$G^\alpha(\pi_1^-, \pi_1^-) = \frac{1}{2 \cdot (n_1 - 1)} \sum_i^{n_1-1} \sum_j^{n_1-1} |x_i^{-1} - x_j^{-1}|^\alpha,$$

$$G^\alpha(\pi_2^+, \pi_2^+) = \frac{1}{2 \cdot (n_2 + 1)} \sum_i^{n_2+1} \sum_j^{n_2+1} |x_i^{+2} - x_j^{+2}|^\alpha,$$

where $x_i^{-1} \in \pi_1^-, i = 1, ..., n_1 - 1$ and $x_i^{+2} \in \pi_2^+, i = 1, ..., n_2 + 1$. The two-sample energy statistics between point $\mathbf{a}$ with cluster π_1 and π_2 are

$$\xi^\alpha(\mathbf{a}, \pi_1) = \frac{2}{n_1} \sum_i^{n_1} |x_i^1 - \mathbf{a}|^\alpha - \frac{1}{n_1^2} \sum_i^{n_1} \sum_j^{n_1} |x_i^1 - x_j^1|^\alpha, \tag{0.4}$$

$$\xi^\alpha(\mathbf{a}, \pi_2) = \frac{2}{n_2} \sum_i^{n_2} |x_i^2 - \mathbf{a}|^\alpha - \frac{1}{n_2^2} \sum_i^{n_2} \sum_j^{n_2} |x_i^2 - x_j^2|^\alpha. \tag{0.5}$$

Firstly, we compute the $G^\alpha(\pi_1, \pi_1) - G^\alpha(\pi_1^-, \pi_1^-)$.

$$\begin{aligned}
&G^\alpha(\pi_1, \pi_1) - G^\alpha(\pi_1^-, \pi_1^-) \\
&= \frac{1}{2 \cdot n_1} \sum_i^{n_1} \sum_j^{n_1} |x_i^1 - x_j^1|^\alpha - \frac{1}{2 \cdot (n_1 - 1)} \sum_i^{n_1-1} \sum_j^{n_1-k} |x_i^{-1} - x_j^{-1}|^\alpha \\
&= \frac{1}{2 \cdot n_1} \sum_i^{n_1} \sum_j^{n_1} |x_i^1 - x_j^1|^\alpha - \frac{1}{2 \cdot (n_1 - 1)} \{ \sum_i^{n_1} \sum_j^{n_1} |x_i^1 - x_j^1|^\alpha \\
&- 2 \sum_i^{n_1} |x_i^1 - \mathbf{a}|^\alpha \} \\
&= \frac{1}{n_1 - 1} \sum_i^{n_1} |x_i^1 - \mathbf{a}|^\alpha - \frac{1}{2 \cdot n_1 (n_1 - 1)} \sum_i^{n_1} \sum_j^{n_1} |x_i^1 - x_j^1|^\alpha.
\end{aligned} \tag{0.6}$$

Times $\frac{n_1}{2(n_1-1)}$ of equation (0.4), then we have

$$\begin{aligned}
\frac{n_1}{2(n_1 - 1)} \xi^\alpha(\mathbf{a}, \pi_1) &= \frac{1}{n_1 - 1} \sum_i^{n_1} |x_i^1 - a|^\alpha \\
&- \frac{1}{2 \cdot n_1 (n_1 - 1)} \sum_i^{n_1} \sum_j^{n_1} |x_i^1 - x_j^1|^\alpha.
\end{aligned} \tag{0.7}$$

Equation (0.6) minus equation (0.7), then we have

$$G^\alpha(\pi_1,\pi_1)-G^\alpha(\pi_1^-,\pi_1^-)=\frac{n_1}{2(n_1-1)}\xi^\alpha(\mathbf{a},\pi_1). \tag{0.8}$$

Then we compute $G^\alpha(\pi_2^+,\pi_2^+)-G^\alpha(\pi_2,\pi_2)$.

$$\begin{aligned}
&G^\alpha(\pi_2^+,\pi_2^+)-G^\alpha(\pi_2,\pi_2)\\
&=\frac{1}{2\cdot(n_2+1)}\sum_i^{n_2+1}\sum_j^{n_2+1}|x_i^{+2}-x_j^{+2}|^\alpha-\frac{1}{2\cdot n_2}\sum_i^{n_2}\sum_j^{n_2}|x_i^2-x_j^2|^\alpha\\
&=\frac{1}{2\cdot(n_2+1)}\{\sum_i^{n_2}\sum_j^{n_2}|x_i^2-x_j^2|^\alpha+2\sum_i^{n_2}|x_i^2-\mathbf{a}|^\alpha\}\\
&-\frac{1}{2\cdot n_2}\sum_i^{n_2}\sum_j^{n_2}|x_i^2-x_j^2|^\alpha\\
&=\frac{1}{n_2+1}\sum_i^{n_2}|x_i^2-a|^\alpha-\frac{1}{2\cdot n_2(n_1+1)}\sum_i^{n_2}\sum_j^{n_2}|x_i^2-x_j^2|^\alpha. \qquad (0.9)
\end{aligned}$$

Times $\frac{n_2}{2(n_1+1)}$ of equation (0.5), then we have

$$\begin{aligned}
&\frac{n_2}{2(n_1+1)}\xi^\alpha(\mathbf{a},\pi_2)=\frac{1}{n_2+1}\sum_i^{n_2}|x_i^2-\mathbf{a}|^\alpha-\\
&\frac{1}{2\cdot n_2(n_2+1)}\sum_i^{n_2}\sum_j^{n_2}|x_i^2-x_j^2|^\alpha. \qquad (0.10)
\end{aligned}$$

Equation (0.9) minus equation (0.10), then we have

$$G^\alpha(\pi_2^+,\pi_2^+)-G^\alpha(\pi_2,\pi_2)=\frac{n_2}{2(n_2+1)}\xi^\alpha(\mathbf{a},\pi_2). \tag{0.11}$$

Equation (0.8)minus equation (0.11), then we have

$$\begin{aligned}
&G^\alpha(\pi_1,\pi_1)+G^\alpha(\pi_2,\pi_2)-G^\alpha(\pi_1^-,\pi_1^-)-G^\alpha(\pi_2^+,\pi_2^+)\\
&=\frac{n_1}{2(n_1-1)}\xi^\alpha(\mathbf{a},\pi_1)-\frac{n_2}{2(n_2+1)}\xi^\alpha(\mathbf{a},\pi_2). \qquad (0.12)
\end{aligned}$$

Based on the derivation above, we have the following theorem.

THEOREM 0.1. *Suppose that $P=\{\pi_1,\pi_2,...\pi_k\}$ is a partition, and $P'=\{\pi_1^-,\pi_2^+,...,\pi_k\}$ is P first variation, so we have*

$$W^\alpha(P)-W^\alpha(P')=\frac{n_1}{2(n_1-1)}\xi^\alpha(\boldsymbol{a},\pi_1)-\frac{n_2}{2(n_2+1)}\xi^\alpha(\boldsymbol{a},\pi_2).$$

Similar as the Hartigan and Wong K-means algorithm, we remove point $\mathbf{a}$ from cluster π_1 to π_2 If

$$\frac{n_1}{2(n_1-1)}\xi^\alpha(\mathbf{a},\pi_1) - \frac{n_2}{2(n_2+1)}\xi^\alpha(\mathbf{a},\pi_2)$$

is positive. Otherwise we keep the point $\mathbf{a}$ in cluster π_1. Based on the computation above we propose the K-Groups algorithm.

Notation Let N be the total sample size of observations, M be the dimension of the sample, and K be the clusters number prespecified. The number of points in cluster $\pi_i (i = 1, ..., K)$ is denoted by $n_i, (i = 1, ..., K)$. The two-sample energy statistic between point I to cluster π_i is denoted as $\xi^\alpha(I, \pi_i)$. The K-Groups algorithm is the following

Step 1. For each point $I, (I = 1, ..., N)$ randomly assign to cluster $\pi_i, i = 1, ..., K$. let $\pi(I)$ represent the cluster where I belongs, and $n(\pi(I))$ represents the size of cluster $\pi(I)$.

Step 2. For each point $I, (I = 1, ..., N)$, Compute

$$E1 = \frac{n(\pi(I))}{2(n(\pi(I))-1)}\xi^\alpha(I, \pi(I))$$

and minimum of the quantity

$$E2 = min\left[\frac{n(\pi_i)}{2(n(\pi_i)+1)}\xi^\alpha(I, \pi_i)\right]$$

for all clusters $\pi_i, \pi \neq \pi(I)$. If $E1$ is less than $E2$, observation I remains in cluster $\pi(I)$. Otherwise, move the point I to cluster π, and update the cluster $\pi(I)$ and π.

Step 3. Stop if there is no relocation in the last N steps.

We can proof the K-groups algorithm and Hartigan and Wong K-means algorithm have the same objective function when let $\alpha = 2$.

PROPOSITION 0.3. *When $\alpha = 2$,*

$$\frac{n_i}{2}G^\alpha(\pi_i, \pi_i) = \sum_{l=1}^{n_i} x_l^2 - n_i c_i^2$$

where $c_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_j$, and $x_j \in \pi_i, j = 1, ...n_i$

PROOF.

$$\begin{aligned}
\frac{n_i}{2}G^2(\pi_i,\pi_i) &= \frac{1}{2n_i}\sum_{l=1}^{n_i}\sum_{m=1}^{n_i}|x_l - x_m|^2 \\
&= \frac{1}{2n_i}\sum_{l=1}^{n_i}\sum_{m=1}^{n_i}(x_l^2 - 2x_l x_m + x_m^2) \\
&= \frac{1}{2n_i}[n_i\sum_{l=1}^{n_i}x_l^2 - 2\sum_{l=1}^{n_i}\sum_{m=1}^{n_i}x_l x_m + n_i\sum_{m=1}^{n_i}x_m^2] \\
&= \frac{1}{2n_i}[2n_i\sum_{l=1}^{n_i}x_l^2 - 2\sum_{l=1}^{n_i}\sum_{m=1}^{n_i}x_l x_m] \\
&= \frac{1}{2n_i}[2n_i\sum_{l=1}^{n_i}x_l^2 - 2n_i^2 c_i^2] \\
&= \sum_{l=1}^{n_i}x_l^2 - n_i c_i^2. \qquad (0.13)
\end{aligned}$$

□

PROPOSITION 0.4.

$$\sum_{x_j\in\pi_i}(x_j - c_i)^2 = \sum_{l=1}^{n_i}x_l^2 - n_i c_i^2$$

where $c_i = \frac{1}{n_i}\sum_{j=1}^{n_i}x_j$, *and* $x_j \in \pi_i, j = 1, ...n_i$.

PROOF.

$$\begin{aligned}
\sum_{x_j\in\pi_i}(x_j - c_i)^2 &= \sum_{j=1}^{n_i}(x_j^2 - 2\cdot x_j\cdot c_i + c_i^2) \\
&= \sum_{j=1}^{n_i}x_j^2 - 2\sum_{j=1}^{n_i}x_j c_i + \sum_{j=1}^{n-i}c_i^2 \\
&= \sum_{j=1}^{n_i}x_j^2 - 2n_i c_i^2 + n_i c_i^2 \\
&= \sum_{j=1}^{n_i}x_j^2 - n_i c_i^2. \qquad (0.14)
\end{aligned}$$

□

Since the objective function for K-means is

$$min_{c_i,1\leq i\leq k}\sum_{i=1}^{k}\sum_{x_j\in\pi_i}(x_j-c_i)^2,$$

the objective function for K-Groups is

$$min_{\pi_1,...,\pi_k}\sum_{i=1}^{k}\frac{n_i}{2}G^{\alpha}(\pi_i,\pi_i),$$

Based on the Proposition (0.3) and Proposition (0.4) we have

$$\sum_{x_j\in\pi_i}(x_j-c_i)^2=\frac{n_i}{2}G^2(\pi_i,\pi_i).$$

for all $i=1,...,k$. So the K-groups and K-means have the same objective function. We have the following theorem

THEOREM 0.2. *When $\alpha=2$, the K-groups algorithm and Hartigan and Wong K-means algorithm have the same objective function.*

Then we proof the update formula of K-groups and Hartigan and Wong K-means algorithm are same when $\alpha=2$.

PROPOSITION 0.5. *Suppose point I belongs in cluster L, the sample size of L is n, then we have*

$$\frac{n}{2(n-1)}\xi^2(I,L)=\frac{n\cdot D(I,L)^2}{n-1}$$

PROOF. we only need to proof $\frac{1}{2}\xi^2(I,L)=D(I,L)^2$.

$$\xi^2(I,L)=\frac{2}{n}\sum_{i=1}^{n}|I-x_i|^2-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i-x_j|^2, \tag{0.15}$$

$$D(I,L)^2=(I-\frac{\sum_{i=1}^{n}x_i}{n})^2. \tag{0.16}$$

We can simplify the equation (0.15) as follows

$$\begin{aligned}
\xi^2(I,L) &= \frac{2}{n}\sum_{i=1}^{n}|I-x_i|^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i-x_j|^2 \\
&= \frac{2}{n}(nI^2 - 2I\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}x_i^2) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i^2 - 2x_ix_j + x_j^2) \\
&= (2I^2 - 4I\frac{\sum_{i=1}^{n}x_i^2}{n} + 2\frac{\sum_{i=1}^{n}x_i^2}{n}) - \frac{1}{n^2}(2n\sum_{i=1}^{n}x_i^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n}x_ix_j) \\
&= (2I^2 - 4I\frac{\sum_{i=1}^{n}x_i^2}{n} + 2\frac{\sum_{i=1}^{n}x_i^2}{n}) - (2\frac{\sum_{i=1}^{n}x_i^2}{n} - 2\frac{2\sum_{i=1}^{n}\sum_{j=1}^{n}x_ix_j}{n^2}) \\
&= 2(I^2 - 2I\frac{\sum_{i=1}^{n}x_i}{n} + \frac{2\sum_{i=1}^{n}\sum_{j=1}^{n}x_ix_j}{n^2}) \\
&= 2(I - \frac{\sum_{i=1}^{n}x_i}{n})^2,
\end{aligned} \tag{0.17}$$

Base on the equation(0.17) we have

$$\frac{1}{2}\xi^2(I,L) = D(I,L)^2.$$

□

With the similar calculation we have

$$\frac{n}{2(n+1)}\xi^2(I,L) = \frac{n * D(I,L)^2}{n+1}. \tag{0.18}$$

Base on Proposition (0.5) and equation(0.18), update formula of K-groups and Hartigan and Wong K-means algorithm are same when $\alpha = 2$.

According the results above, K-means is a special case of K-groups when $\alpha = 2$. However, according to the properties of energy statistics, we know that when $0 < \alpha < 2$ the energy distance $\xi^\alpha(X,Y) = 0$ if and only if random variable X and Y follow the same statistical distribution. However, when $\alpha = 2$ we have $\xi^\alpha(X,Y) = 0$ when ever $E(X) = E(Y)$. In spite of the objective function and update formula for K-Groups and K-means are equivalent when $\alpha = 2$, they are still different clustering method.

Second Variation Algorithm. The objective of K-groups is to find a global minimum of within-cluster sum of dispersion. However, in most cases we can only get the local minimum by first variation method. Usually

in order to solve this problem, one can try different initial random starts, and choose the best result with minimum within-cluster dispersion. In order to search the global optimization, we are trying to move more than one observations. The reasons why we want to move more than one point are the following.

- We want to move from the local optimum obtained by the first variation.
- Based on the result of [11], if two samples follow different distributions, the weighted two-sample energy statistic

$$T_{X,Y} = \left(\frac{n_1 n_2}{n_1 + n_2}\right) \mathcal{E}_{n_1,n_2}(\mathbf{X}, \mathbf{Y})$$

  will approach infinity stochastically as N tends to infinity and neither $\frac{n_1}{N}$ nor $\frac{n_2}{N}$ goes to zero, where N denotes the total data size.
- Energy statistics admit a nice updating formula for moving more than one observations. We will show later that the difference of within-cluster sum of dispersion equals the difference of weighted two-sample energy statistics if we move any $m\,(m > 1)$ points from cluster to cluster.

DEFINITION 0.3. *A m^{th} variation of a partition P is a partition $P^{(m)}$ obtained from P by removing m points $\{a_1, a_2, ..., a_m\}$ from a cluster π_i of P and assigning these points to an existing cluster π_j of P , $i \neq j$.*

We want to move m points $\{a_1, a_2, ..., a_m\}$ from cluster π_1 to another cluster π_2. Cluster π_1^- represents cluster π_1 after removing m points $\{a_1, a_2, ..., a_m\}$, and cluster π_2^+ represents cluster π_2 after adding those m points. Let n_1 and n_2 be the sizes of π_1 and π_2 before moving m points. The derivation of updating formula for the m^{th} variation is similar to the second variation. The two-sample energy statistic between the m points $\{a_1, a_2, ..., a_m\}$ and clusters π_1, π_2 are by definition

$$\begin{aligned}\xi^{\alpha}(\{a_1, ...a_m\}, \pi_1) =& \frac{2}{m \cdot n_1} \sum_{i}^{n_1} \sum_{j}^{m} |x_i^1 - a_j|^{\alpha} - \frac{1}{m^2} \sum_{i}^{m} \sum_{j}^{m} |a_i - a_j|^{\alpha} \\ & - \frac{1}{n_1^2} \sum_{i}^{n_1} \sum_{j}^{n_1} |x_i^1 - x_j^1|^{\alpha}, \end{aligned} \tag{0.19}$$

and

$$\xi^{\alpha}(\{a_1,...a_m\},\pi_2) =\frac{2}{m\cdot n_2}\sum_{i}^{n_2}\sum_{j}^{m}|x_i^2-a_j|^{\alpha}-\frac{1}{m^2}\sum_{i}^{m}\sum_{j}^{m}|a_i-a_j|^{\alpha}$$
$$-\frac{1}{n_2^2}\sum_{i}^{n_2}\sum_{j}^{n_2}|x_i^2-x_j^2|^{\alpha}. \tag{0.20}$$

Similar to the derivation of first variation, we compute $\frac{n_1}{2}G^{\alpha}(\pi_1,\pi_1)-\frac{n_1-m}{2}G^{\alpha}(\pi_1^-,\pi_1^-)$ and $\frac{n_2+m}{2}G^{\alpha}(\pi_2^+,\pi_2^+)-\frac{n_2}{2}G^{\alpha}(\pi_2,\pi_2)$, as

$$\begin{aligned}
&\frac{n_1}{2}G^{\alpha}(\pi_1,\pi_1)-\frac{n_1-m}{2}G^{\alpha}(\pi_1^-,\pi_1^-)\\
=&\frac{1}{2\cdot n_1}\sum_{i}^{n_1}\sum_{j}^{n_1}|x_i^1-x_j^1|^{\alpha}-\frac{1}{2\cdot(n_1-m)}\sum_{i}^{n_1-m}\sum_{j}^{n_1-m}|x_i^{-1}-x_j^{-1}|^{\alpha}\\
=&\frac{1}{2\cdot n_1}\sum_{i}^{n_1}\sum_{j}^{n_1}|x_i^1-x_j^1|^{\alpha}-\frac{1}{2\cdot(n_1-m)}\Bigg[\sum_{i}^{n_1}\sum_{j}^{n_1}|x_i^1-x_j^1|^{\alpha}\\
&-2\sum_{i}^{n_1}\sum_{j}^{m}|x_i^1-a_j|^{\alpha}+\sum_{i}^{m}\sum_{j}^{m}|a_i-a_j|^{\alpha}\Bigg]\\
=&\frac{1}{n_1-m}\sum_{i}^{m}\sum_{j}^{n_1}|x_i^1-a_j|^{\alpha}-\frac{1}{2\cdot(n_1-m)}\sum_{i}^{m}\sum_{j}^{m}|a_i-a_j|^{\alpha}\\
&-\frac{m}{2\cdot n_1(n_1-m)}\sum_{i}^{n_1}\sum_{j}^{n_1}|x_i^1-x_j^1|^{\alpha}.
\end{aligned} \tag{0.21}$$

Multiply $\frac{m\cdot n_1}{2(n_1-m)}$ times equation(0.19) to get

$$\begin{aligned}
\frac{mn_1}{2(n_1-m)}\xi^{\alpha}(\{a_1,...a_m\},\pi_1) =&\frac{1}{n_1-m}\sum_{i}^{n_1}\sum_{j}^{m}|x_i^1-a_j|^{\alpha}-\\
&\frac{n_1}{2m(n_1-m)}\sum_{i}^{m}\sum_{j}^{m}|a_i-a_j|^{\alpha}-\\
&\frac{m}{2n_1(n_1-m)}\sum_{i}^{n_1}\sum_{j}^{n_1}|x_i^1-x_j^1|^{\alpha}.
\end{aligned} \tag{0.22}$$

Subtract equation (0.22) from equation (0.21) to obtain

$$\begin{aligned}&\frac{n_1}{2}G^\alpha(\pi_1,\pi_1)-\frac{n_1-m}{2}G^\alpha(\pi_1^-,\pi_1^-)\\&=\frac{mn_1}{2(n_1-m)}\xi^\alpha(\{a_1,...a_m\},\pi_1)+\frac{1}{2m}\sum_i^m\sum_j^m|a_i-a_j|^\alpha.\end{aligned}\tag{0.23}$$

Based on a similar derivation, we can show that

$$\begin{aligned}&\frac{n_2+m}{2}G^\alpha(\pi_2^+,\pi_2^+)-\frac{n_2}{2}G^\alpha(\pi_2,\pi_2)\\&=\frac{mn_2}{2(n_2+m)}\xi^\alpha(\{a_1,...a_m\},\pi_2)+\frac{1}{2m}\sum_i^m\sum_j^m|a_i-a_j|^\alpha.\end{aligned}\tag{0.24}$$

Subtract equation (0.24) from equation (0.23) to get

$$\begin{aligned}&\frac{n_1}{2}G^\alpha(\pi_1,\pi_1)+\frac{n_2}{2}G^\alpha(\pi_2,\pi_2)-\frac{n_1-m}{2}G^\alpha(\pi_1^-,\pi_1^-)-\frac{n_2+m}{2}G^\alpha(\pi_2^+,\pi_2^+)\\&=\frac{mn_1}{2(n_1-m)}\xi^\alpha(\{a_1,...a_m\},\pi_1)-\frac{mn_2}{2(n_2+m)}\xi^\alpha(\{a_1,...a_m\},\pi_2).\end{aligned}\tag{0.25}$$

THEOREM 0.3. *Suppose $P=\{\pi_1,\pi_2,...\pi_k\}$ is a partition, and $P^{(m)}=\{\pi_1^-,\pi_2^+,...,\pi_k\}$ is a m^{th} variation of P by moving points $\{a_1,a_2,...,a_m\}$ from cluster π_1 to π_2. Then*

$$W^\alpha(P)-W^\alpha(P^{(m)})=\frac{mn_1}{2(n_1-m)}\xi^\alpha(\{a_1,...a_m\},\pi_1)-\frac{mn_2}{2(n_2+m)}\xi^\alpha(\{a_1,...a_m\},\pi_2).$$

Similar to first variation and second variation, we assign points $\{a_1,a_2,...,a_m\}$ to cluster π_2 if

$$\frac{mn_1}{2(n_1-m)}\xi^\alpha(\{a_1,...a_m\},\pi_1)-\frac{mn_2}{2(n_2+m)}\xi^\alpha(\{a_1,...a_m\},\pi_2)$$

is positive; otherwise we keep points $\{a_1,a_2,...,a_m\}$ in cluster π_1.

According to theorem above, we can move any m points from one cluster to another cluster. However, the computation cost for moving more points is excessive. Suppose the total sample size is N, and we have K clusters, K prespecified. For K-groups by first variation algorithm, one needs to compute distance NK times in each loop. Suppose $m=2$, one needs to compute distance $\frac{KN(N-1)}{2}$ times in each loop, because there are $\frac{N(N-1)}{2}$ combinations of two points. For m^{th} variation, ones to compute distance $\frac{CN!}{m!(N-m)!}$ times in each loop. The computation cost will increase exponentially in N. Even

though we have very nice mathematical formula for moving m points, the computational cost is too expensive. Here, we let $m = 2$ and propose the algorithm by moving two points. We call this algorithm *K-groups by second variation*.

It is not practical to consider all possible combinations of two points. Since our objective is to minimize the within-group sum of dispersion, we do not need to consider all possible combinations. We pair two points together if they have the minimum energy distance, and we assume these two points should be assigned to the same cluster.

Notation Let even number N be the total sample size of observations, M be the dimension of the sample, and K be the number of clusters, K prespecified. The size of cluster π_i, $(i = 1, ..., K)$ is denoted by n_i. The two-sample energy statistic between pair II to cluster π_i is denoted by $\xi^{\alpha}(II, \pi_i)$. The K-groups algorithm by second variation is the following:

Step 1. Each pair of points $II\,(II = 1, ..., N/2)$, is randomly assigned to cluster $\pi_i\,(i = 1, ..., K)$. Let $\pi(II)$ represent the cluster containing II , and $n(\pi(II))$ represent the size of cluster $\pi(II)$.

Step 2. For each pair $II\,(II = 1, ..., N/2)$, compute

$$E_1 = \frac{n(\pi(II))}{n(\pi(II)) - 2}\xi^{\alpha}(II, \pi(II))$$

and the minimum

$$E_2 = \min\left[\frac{n(\pi_i)}{n(\pi_i) + 2}\xi^{\alpha}(II, \pi_i)\right]$$

for all clusters π_i, where $\pi_i \neq \pi(II)$. If E_1 is less than E_2, pair II remains in cluster $\pi(II)$; otherwise, move the pair II to cluster π_i with minimum value of E_2, and update the cluster $\pi(II)$ and π_i.

Step 3. Stop if there is no relocation in the last $\frac{N}{2}$ steps.

For an odd number N, we randomly take one observation out. After running the K-groups by second variation, we assign the observation to the cluster based on the updating formula of K-groups by first variation algorithm.

Simulation Results. A variety of cluster structures can be generated as mixtures of different distributions. Each of our simulated data sets was designed as mixture, where each component of the mixture corresponds to a cluster. Each mixture distribution is simulated at different sample sizes 200, we calculate average and standard error for validation indices diagonal (Diag), Kappa, Rand, and corrected Rand (cRand) based on $B = 500$

iterations. In K-groups methods, for the mixture distributions which have finite first and second moment, we use $\alpha = 1$; otherwise we use the smaller value of $\alpha = 0.5$ to have finite moments $E|X - Y|^{\alpha}$. All algorithms were implemented in R and all simulations carried out in R. The K-groups algorithms are available upon request in an R package *kgroups* [4]. We want to compare K-groups by first variation, K-groups by second variation and K-means under different cluster structure.

Figure1 displays the simulation results for normal mixture $0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(d,1)$, where $d = 0.2, 0.4, ..., 3$. The average cRand indices of the three algorithms are almost the same for each value of d. The results for symmetric normal mixtures suggest that both K-groups algorithms and K-means have similar performance when the cluster are normally distributed.

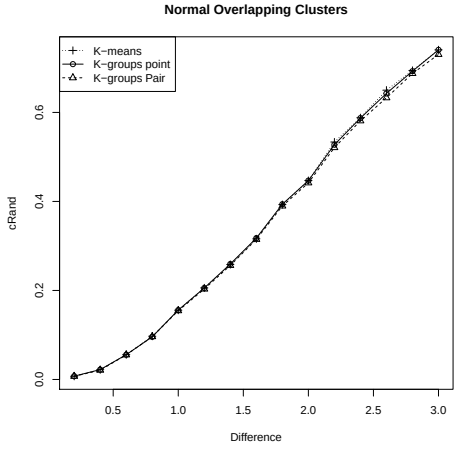

FIG 1. *Overlapping effect for normal mixture distributions,* $n = 200$, $B = 500$

Figure 2 displays the simulation results for Cauchy mixture $0.5\,\mathrm{Cauchy}(0,1) + 0.5\mathrm{Cauchy}\,(d,1)$, where $d = 0.2, 0.4, ..., 3$. The average cRand indices of both K-groups algorithms dominate the average cRand of K-means for each value of d. Thus, the results suggest that both K-groups algorithms are more robust respect to outliers and heavy tails.

Figure 3 displays the simulation results for lognormal mixtures $0.5\,\mathrm{lognormal}\,(0,1) + 0.5\,\mathrm{lognormal}\,(d,1)$ where $d = 0.5, 1, ..., 10$. The average cRand indices of both K-groups algorithms dominate the K-means for each value of d. Thus, the results suggest that the K-groups algorithms have much better perfor-

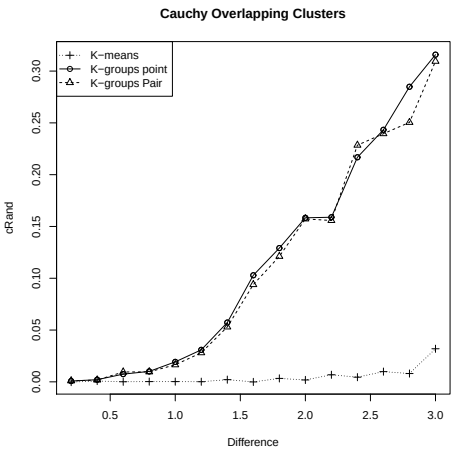

FIG 2. *Overlapping effect for Cauchy mixture distributions,* $n = 200$, $B = 500$

mance than K-means when clusters are strongly skewed, heavy tailed.

Figure 4 shows the results of normal mixtures $0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1)$ with $\alpha = 0.2, 0.4, ..., 2$. The average cRand indices of K-means and K-groups by first variation are very close. When $d = 2$, K-means and K-groups by first variation have very close average cRand indices. The average cRand indices of K-groups by second variation are slight lower than the other two algorithms. Generally, for each value of α, the average cRand indices of both K-groups algorithms and K-means are very close. Thus, the results suggest that there is no α effect when clusters are normally distributed.

Figure 5 shows the results of Cauchy mixtures $0.5\,\mathrm{Cauchy}\,(0,1)+0.5\,\mathrm{Cauchy}\,(3,1)$ with $\alpha = 0.2, 0.4, ..., 2$. The average cRand indices of K-groups by first variation decrease as α increases, and when $\alpha = 2$ the average cRand indices of K-groups by first variation and K-means are very close. K-groups by second variation have more stable average cRand indices than the other two algorithms. Thus, the results suggest that there is α effect when clusters have infinite first moment.

Table 1 summarize the simulation results of multivariate cubic mixture $0.5\,\mathrm{Cubic}^d\,(0,1) + 0.5\,\mathrm{Cubic}^d\,(0.3, 0.7)$. For every dimension d, the average Rand and cRand indices of both K-groups algorithms are higher than K-means. For each algorithm, the average Rand and cRand indices increase as the dimension d increases.

Figure 6 displays the simulation result of cubic mixtures $0.5\,\mathrm{Cubic}^d\,(0,1)+$ $0.5\,\mathrm{Cubic}^d$

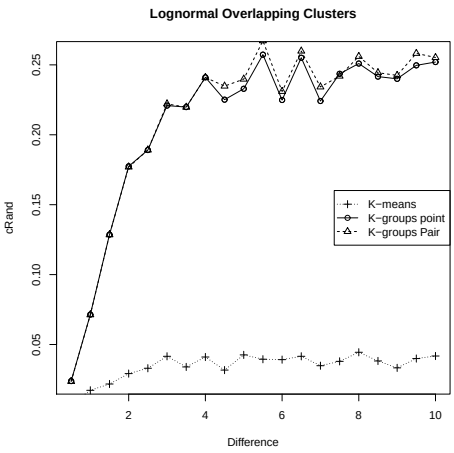

FIG 3. *Overlapping effect for lognormal mixture distributions,* $n = 200$, $B = 500$

$(0.3, 0.7)$, where $d = 1, 2, 4, ..., 40$. The average Rand and cRand indices of these three algorithms are almost the same when $d < 5$. However, the average Rand and cRand indices of both K-groups algorithms are consistently higher than K-means when $d > 5$. Furthermore, the average Rand and cRand indices of K-groups by first variation approach 1 as dimension d increases. Thus, the results suggest that K-groups by first variation algorithm has better performance than the other two algorithms when clusters are cubic shaped in the multivariate case.

Now we want to use some real data to test the K-groups and K-means algorithm.

Diagnosis of Erythemato-Squamous Diseases in Dermatology. The dermatology data analyzed is publicly available from the UCI Machine Learning Repository [1] at ftp.ics.uci.edu. The data was analyzed by [2], and contributed by Güvenir. The erythemato-squamous dieases are proriasis, seboreic dermatitis, lichen planus, pityriasis rosea, choronic dermatitis and pityriasis rubra pilaris. According to [2], diagnosis is difficult since all these diseases share the similar clinical features of erythema and scaling. Another difficulty is that a disease may show histopathological features of another disease initially, but have characteristic feature at the following stages.

The data consists of 366 objects with 34 attributes. There are 12 clinical attributes and 22 histopathological attributes. All except two take values in $0, 1, 2, 3$, where 0 indicates the feature was not present and 3 is the largest

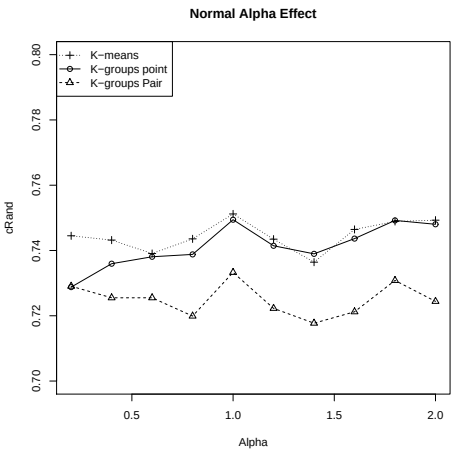

FIG 4. *α effect for normal mixture distributions, $n = 200$, $B = 1000$*

TABLE 1
Cubic Mixture $0.5\,\mathrm{Cubic}^d\,(0,1) + 0.5\,\mathrm{Cubic}^d\,(0.3, 0.7), \alpha = 1$

Method	d	Diag	Kappa	Rand	cRand
K-means	1	0.5381	0.0758	0.5021	0.0043
K-groups Point	1	0.5352	0.0710	0.5014	0.0028
K-groups Pair	1	0.5352	0.0710	0.5014	0.0028
K-means	2	0.5439	0.0877	0.5032	0.0065
K-groups Point	2	0.5440	0.0879	0.5034	0.0068
K-groups Pair	2	0.5542	0.0884	0.5034	0.0069
K-means	5	0.5536	0.1067	0.5056	0.0113
K-groups Point	5	0.5713	0.1427	0.5128	0.0257
K-groups Pair	5	0.5676	0.1355	0.5120	0.0240
K-means	10	0.5705	0.1393	0.5128	0.0257
K-groups Point	10	0.7875	0.5758	0.6923	0.3847
K-groups Pair	10	0.6647	0.3287	0.5672	0.1346
K-means	20	0.6065	0.2078	0.5274	0.0550
K-groups Point	20	0.9976	0.9951	0.9952	0.9904
K-groups Pair	20	0.7213	0.4416	0.6045	0.2090
K-means	40	0.6396	0.2794	0.5406	0.0810
K-groups Point	40	0.9999	0.9999	0.9999	0.9997
K-groups Pair	40	0.7471	0.4960	0.6228	0.2456

amount possible. The attribute of family history takes value 0 or 1, and the age of patient takes positive integer values. There are eight missing values in the age of patient. The clinical and histopathological attributes are summarized in Table 2. We standardize all the attributes to zero mean

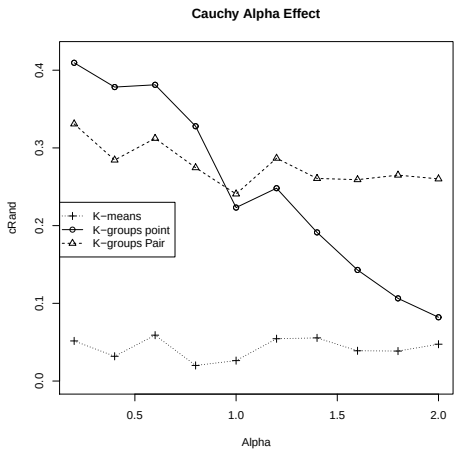

FIG 5. *α effect for Cauchy mixture distributions, $n = 200$, $B = 1000$*

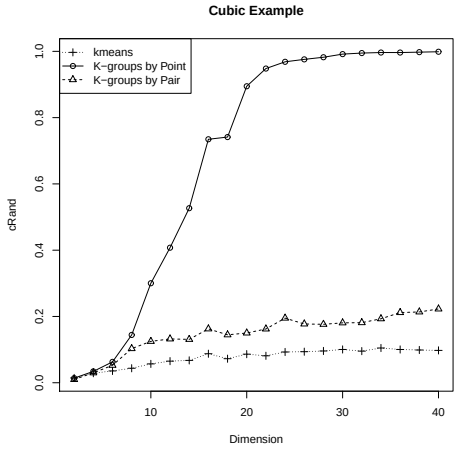

FIG 6. *Multivariate cubic mixtures, $d = 2, 4, ..., 40$, $n = 200$, $B = 500$*

and unit standard deviation and delete the observations which contain the missing values. The effective data size is 358 in the clustering analysis.

Table 3 shows the clustering result of K-means, K-groups by first variation, K-groups by second variation, and Hierarchical ξ. The maximum Rand and cRand index values 0.9740 and 0.9188 are obtained by K-groups by first variation. The Hierarchical ξ obtains the second largest Rand and cRand

TABLE 2
Dermatology Data Summary

	Clinical Attributes		Histopathological Attributes
1.	erythema	12.	melanin incontinence
2.	scaling	13.	eosinophils in the infiltrate
3.	definite borders	14.	PNL infiltrate
4.	itching	15.	fibrosis of the paillary derims
5.	koebner phenomenon	16.	exocytosis
6.	polygonal papules	17.	acanthosis
7.	follicular papules	18.	hyperkeratosis
8.	oral mucosal involvement	19.	parakeratosis
9.	knee and elbow involvement	20.	clubbing of the rete ridges
10.	scalp involvement	21.	elongation of the rete ridges
11.	family history	22.	thinning of the suprapapillary epidermis
34.	age	23.	pongiform pustule
		24.	munro microabcess
		25.	focal hyperganulosis
		26.	disapperance of the granular layer
		27.	vaculolization and damage of basal layer
		28.	spongiosis
		29.	saw-tooth appearance of retes
		30.	follicular horn plug
		31.	perifollicular parakeratosis
		32.	inflammatory mononuclear infiltrate
		33.	band-like infiltrate

index values 0.9730 and 0.9159. K-groups by second variation obtains the Rand and cRand index values 0.9543 and 0.8602. K-means obtains smallest Rand and cRand index values among those four algorithms, 0.9441 and 0.8390.

TABLE 3
Dermatology Data Results

Indices	K-means	K-groups Point	K-groups Pair	Hierarchical ξ
Diag	0.8324	0.9553	0.8910	0.9497
Kappa	0.7882	0.9440	0.8640	0.9370
Rand	0.9441	0.9740	0.9543	0.9730
cRand	0.8390	0.9188	0.8602	0.9159

Conclusion. The simulation results for univariate and multivariate cases show that both K-groups algorithms perform as well as Hartigan and Wong's K-means algorithm when clusters are well-separated and normally distributed. Both K-groups algorithms perform better than K-means when data does not have finite first moment. For data which has strong skewness and heavy tails, both K-groups algorithms perform better than K-means. For non–spherical clusters, both K-groups algorithms perform better than K-means in high dimension and K-groups by first variation is consistent as dimension increases. Results of clustering on three real data examples show that both K-groups algorithms perform better than K-means, and in some situations, K-groups by first variation performs better than Hierarchical ξ.

In summary, our proposed K-groups method can be recommended for all types of unsupervised clustering problems with pre-specified number of clusters, because performance was typically comparable to or better than K-means. K-groups has other advantages and it is a more general method. It can be applied to cluster feature vectors in arbitrary dimension and the index α can be chosen to handle very heavy tailed data with non-finite expected distances. We have developed and applied a simple updating formula analogous to Hartigan and Wong, which has been implemented in R, and the method is also easily implemented in Python, Matlab or other widely used languages.

Future research directions are as follows.

- K-means and both K-groups algorithms have computational time $O(n^2)$, where n is the total sample size. We plan to use parallel computing to cut down the computational time.
- Big data is a challenge for clustering tasks, since the computational time of traditional algorithms are too long. We plan to divide the big

data into mutually exclusive subsets with reasonable sizes and run the K-groups algorithm on each subset. Then we can use m^{th} variation formula to merge the clustering result of different subsets together.

- If we know some observations should be assigned to the same cluster, we can bunch these observations together and use m^{th} variation formula to implement the clustering tasks.
- Energy distance is a functional distance between the characteristic functions of two independent random variables X and Y. [6] extended energy–type distance to separable Hilbert space by choosing appropriate kernel functions which are negative definite. Based on similar ideas, we can extend the K-groups algorithms using different kernel functions.

Bibliography.

[1] Blake, C. and C. J. Merz (1998). {UCI} repository of machine learning databases.

[2] Güvenir, H. A., G. Demiröz, and N. Ilter (1998). Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. *Artificial Intelligence in Medicine 13*(3), 147–165.

[3] Hartigan, J. A. and M. A. Wong (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 100–108.

[4] Li, S. and M. L. Rizzo (2015). *kgroups: Cluster Analysis Based on Energy Distance*. R package version 1.0.

[5] Lloyd, S. (1982). Least squares quantization in pcm. *Information Theory, IEEE Transactions on 28*(2), 129–137.

[6] Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability 41*(5), 3284–3305.

[7] Milligan, G. W. (1996). *Clustering Validation: Results and Implications for Applied Analyses*. World Scientific.

[8] Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 71–110.

[9] Székely, G. (2000). Technical report 03-05: E-statistics: energy of statistical samples. *Department of Mathematics and Statistics, Bowling Green State University*.

[10] Székely, G. J., M. Alpár, and É. Unger (1986). *Paradoxes in Probability Theory and Mathematical Statistics*. D. Reidel Dordrecht.

[11] Székely, G. J. and M. L. Rizzo (2004). Testing for equal distributions in high dimension. *InterStat 5*.

[12] Tan, P.-N., M. Steinbach, and V. Kumar (2006). Cluster analysis: basic concepts and algorithms. *Introduction to Data Mining*, 487–568.

Address of the First and Second authors
Usually a few lines long
E-mail: lisongzi24601@gmail.com
mrizzo@bgsu.edu

Address of the Third author
Usually a few lines long
Usually a few lines long
??
??