K-GROUPS: A GENERALIZATION OF K-MEANS BY ENERGY DISTANCE

Songzi Li

A Dissertation

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2015

Committee:

Maria L. Rizzo, Advisor

Christopher M. Rump,
Graduate Faculty Representative

Hanfeng Chen

Wei Ning

ABSTRACT

Maria L. Rizzo, Advisor

We propose two distribution-based clustering algorithms called K-groups. Our algorithms group the observations in one cluster if they are from a common distribution. Energy distance is a non-negative measure of the distance between distributions that is based on Euclidean distances between random observations, which is zero if and only if the distributions are identical. We use energy distance to measure the statistical distance between two clusters, and search for the best partition which maximizes the total between clusters energy distance. To implement our algorithms, we apply a version of Hartigan and Wong's moving one point idea, and generalize this idea to moving any m points. We also prove that K-groups is a generalization of the K-means algorithm. K-means is a limiting case of the K-groups generalization, with common objective function and updating formula in that case.

K-means is one of the well-known clustering algorithms. From previous research, it is known that K-means has several disadvantages. K-means performs poorly when clusters are skewed or overlapping. K-means can not handle categorical data. K-means can not be applied when dimension exceeds sample size. Our K-groups methods provide a practical and effective solution to these problems.

Simulation studies on the performance of clustering algorithms for univariate and multivariate mixture distributions are presented. Four validation indices (diagonal, Kappa, Rand and corrected Rand) are reported for each example in the simulation study. Results of the empirical studies show that both K-groups algorithms perform as well as K-means when clusters are well-separated and spherically shaped, but K-groups algorithms perform better than K-means when clusters are skewed or overlapping. K-groups algorithms are more robust than K-means with respect to outliers. Results are presented for three multivariate data sets, wine cultivars, dermatology diseases and oncology cases. In real data examples, both K-groups algorithms perform better than K-means in each case.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Maria Rizzo, for her generous help and valuable advice throughout this research. I am also grateful to have her constant encouragement and support during my early career in a non-academic field. I also want to express my appreciation to other members of my committee, Professor Hanfeng Chen, Professor Wei Ning, and Professor Christopher M. Rump, for their valued time and advice.

I want to extend my gratitude to all my professors in the Department of Mathematics and Statistics, and the Department of Applied Statistics and Operations Research, for their guidance and help during my four-year study at Bowling Green.

Finally, I wish to express my deepest gratitude to my family for always being by my side. I wish to thank my parents for their never-ending love, support and care. My special thanks goes to my fiancée Sanna for her love and sacrifice.

Songzi Li

Bowling Green, Ohio

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Cluster analysis is one of the core topics of data mining. Clustering is a fundamental tool in an unsupervised study, which is used to group similar objects together without using external information such as class labels. Cluster analysis plays an important role in a wide variety of application domains such as astronomy, psychology, market research, and bioinformatics. Although many good algorithms are available, no single algorithm is known to be optimal for all clustering methods, and no method is known to be best for all problems.

In this dissertation, we will focus on the most widely-applied clustering algorithm, K-means. From previous research, the disadvantages of K-means are as follows:

- K-means typically performs poorly when the data are skewed.
- K-means cannot handle categorical data.
- K-means performs poorly when clusters are overlapping.
- K-means does not perform well when data has white noise and outliers.
- K-means is not valid for data with infinite first moment.
- K-means cannot be applied when dimension exceeds sample size.

A new algorithm, K-groups, is proposed and applied to the classical clustering problem in this dissertation. We use energy distance to measure the statistical distance between two clusters, and to search for the best partition that maximizes the total between clusters energy distance. To

implement our algorithm, we adapt the moving one point idea of Hartigan and Wong's K-means algorithm, and generalize this idea to moving any m points. The simulation and real data results show that K-groups overcomes some disadvantages of the K-means algorithm.

1.1 Clustering Algorithms

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 (Pearson, 1894). Since then there has been extensive research about clustering algorithms. Milligan (1996) had pointed out that the difficulties of cluster analysis lie in the following three aspects:

- Clustering is essentially an inexhaustible combinatorial problem;
- There exist no widely accepted theories for clustering;
- The definition of a cluster is a bit "arbitrary," which is determined by the data characteristics and understanding of users.

These three aspects illustrate well why there are so many different clustering algorithms, and each algorithm may have its own advantages in a given situation.

Prototype-Based Algorithm. This kind of algorithm learns a prototype for each cluster, and creates clusters by grouping data objects around the prototypes. The most widely applied prototype-based algorithm are K-means (MacQueen, 1967) and Fuzzy c-Means (FCM) (Bezdek, 1981). Both K-means and Fuzzy c-Means assign the cluster mean as the prototype, and the clusters tend to be globular. Another kind of prototype-based algorithm, Self-Organizing Map (SOM) (Kohonen, 1990) uses a neighborhood function to preserve the topological properties of data objects. McLachlan and Basford (1988) used a probability function to identify the prototype, and the unknown parameters usually are estimated by the Maximum Likelihood Estimation (MLE) method.

Density-Based Algorithm. A cluster in this algorithm is a dense region of data objects that is surrounded by regions of low densities. They are widely used when data are irregular or inter-

twined, or when noise and outliers are present. DBSCAN (Ester, Kriegel, Sander, and Xu, 1996) and DENCLUE (Hinneburg and Keim, 1998) are two representative density-based algorithms. DBSCAN divides data objects into core points, border points and noise based on the Euclidean density. DENCLUE defines a probability density function based on the kernel function of each data object, then finds the cluster by detecting the variance of densities.

Graph-Based Algorithm. This kind of algorithm treats data objects as nodes, and the distance between two objects as the weight of the edge connecting the two nodes. The data can be represented as a graph, and a connected subgraph makes a cluster. Agglomerative hierarchical clustering algorithm (AHC) (Tan, Steinbach, and Kumar, 2006) is representative of graph-based algorithms. AHC merge the nearest two nodes/groups in one round until all nodes are connected. Jarvis and Patrick (1973) proposed a graph-based algorithm, Jarvis-Patrick algorithm (JP) which defines the shared nearest-neighbors for each data object, and then sparsifies the graph to get clusters.

Hybird Algorithms Hybrid algorithms use two or more clustering algorithms in combination. This approach can overcome the shortcomings of single clustering algorithms. Chameleon (Karypis, Han, and Kumar, 1999) is a typical hybrid algorithm, which firstly uses a graph-based algorithm, and then employs a special AHC to get final clusters.

1.2 Cluster Validity

Assessment of cluster validity is a necessary but difficult task in cluster analysis. Usually it is defined as giving objective evaluations to cluster results in a quantitative way (Jain and Dubes, 1988). The importance of cluster validity is that almost every clustering algorithm will find clusters, even if the data set has no natural cluster structure. That is why one needs a cluster validation measure to access how good is the clustering. There are two types of validation measures, external indices and internal indices.

External Indices. The external indices measure the extent to which the clustering structure discovered by a clustering algorithm matches some given external structure, e.g. the structure defined by the class labels. Well-known external measures include the Rand index (R) (Rand, 1971),

adjusted Rand index (Rand, 1971), Jaccard coefficient (J) (Jaccard, 1912), Folks and Mallows index (FM) (Jain and Dubes, 1988), Entropy measure (E) (Zhao and Karypis, 2004), and Variation of Information (VI) (Meilă, 2003).

Rand index, Jaccoard coefficient, and Folks and Mallows index are statistics-based measures. These kind of measure focuses on examining the group membership of each object pair, which can be quantified by comparing two matrices: the Ideal Cluster Similarity Matrix (ICuSM) and the Ideal Class Similarity Matrix (ICaSM) (Tan et al., 2006). If two objects i, j are in the same cluster, then ICuSM ij-entry equals 1, otherwise equals 0. The ICaSM is defined based on the class labels which are given. Let $f_{00}\,(f_{11})$ denote the number of entries that have $0\,(1)$ in the corresponding position of the upper triangular matrices of ICuSM and ICaSM. Let f_{01} and f_{10} denote the numbers of entry pairs that have different values in the corresponding positions of the upper triangular matrices of ICuSM and ICaSM. The validation measures R, J, and FM can be defined as

$$R = \frac{f_{00} + f_{11}}{f_{00} + f_{11} + f_{01} + f_{10}},$$

$$J = \frac{f_{11}}{f_{11} + f_{01} + f_{10}},$$

$$FM = \frac{f_{11}}{\sqrt{(f_{10} + f_{11})(f_{01} + f_{11})}}.$$

Entropy measure and Variation of Information measure are examples of information-theoretic measures, which are typically designed based on the concepts of information theory. Entropy measure assumes that the clustering quality is higher if the entropy of data objects in each cluster is smaller. Let p_{ij} represent the proportion of objects in cluster j that are from class i, and suppose that the size of cluster j is n_j and $n = \sum_j n_j$. Then the entropy of the data in cluster j is

$$E_j = -\sum_i p_{ij} \log p_{ij},$$

and the Entropy Measure index is $E = \sum_j (n_j/n) E_j$.

Internal Indices. Internal indices measure the goodness of a clustering structure without ex-

ternal information. They only make latent assumptions on the formation of cluster structures, and have higher computational cost than the external indices.

When the purpose is only to assess clustering algorithms and the class labels are available, we prefer to use external measures since they are easier to compute.

The main results of the dissertation are organized as follows. In Chapter 2, we introduce the detail of Hartigan and Wong's K-means algorithm, properties of energy distance and applications of energy distance. K-groups by first variation algorithm (moving one point per iteration) is developed in Chapter 3. K-groups by second variation algorithm (moving two points per iteration) and updating formula for m^{th} variation (moving m points per iteration) are proposed in Chapter 4. In Chapters 5 and 6 we compare the K-groups algorithms with K-means on univariate and multivariate simulated data. In Chapter 7, the K-groups algorithms are compared with K-means on real data sets.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we will introduce the K-means clustering algorithm, energy distance and applications of energy distance.

2.1 Introduction of K-means

As we mentioned before, K-means is a prototype-based algorithm. K-means clustering uses the cluster mean as the centroid, and assigns the observations to the cluster with nearest centroid. The clustering process of K-means is as follow.

- K initial centroids are selected (cluster centroid is the mean of the points in that cluster), where K is specified by the user and indicates the desired number of clusters.
- Every point in the data is assigned to the closest centroid, and each collection of points assigned to a centroid forms a cluster.
- The centroid of each cluster is then updated based on the points assigned to that cluster. This process is repeated until no point changes clusters.

Notation Let $D = \{x_1, \ldots, x_n\} \subset R^m$ be the data set to be clustered. Let $P = \{\pi_1, \ldots, \pi_K\}$ denote a partition of D, where K is the number of clusters set by the user. Then π_i is the i-th cluster and $D = \cup_i \pi_i$, where $\pi_i \cap \pi_j = \emptyset$ if $i \neq j$. The symbol ω_x is the frequency for observation x. Let n_k be the number of data objects assigned to cluster π_k, and $c_k = \sum_{x \in \pi_k} (\omega_x / n_k) x$ represent 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 objects x and y. The K-means clustering problem is equivalent to the optimization task of finding

$$\min_{c_k, 1 \leq k \leq K} \sum_{k=1}^{K} \sum_{x \in \pi_k} \omega_x d(x, c_k). \tag{2.1}$$

The research on K-means clustering focuses on three aspects: computation, search optimization, and distance design.

2.1.1 Computation

The K-means clustering problem is equivalent to searching for a global minimum, which is computationally difficult (NP-hard). The standard algorithm was proposed by Stuart LIoyd in 1957 as a technique for pulse-code modulation, though it was not published outside of Bell Labs until 1982 (Lloyd, 1982). A more efficient version was proposed and published in Fortran by Hartigan and Wong (1979). We will discuss the details of Hartigan-Wong K-means algorithm in Section 2.2.

2.1.2 Search Optimization

One weakness of K-means is that the standard algorithm may converge to a local minimum or even a saddle point. Dhillon, Guan, and Kogan (2002) proposed a "first variation" strategy for spherical K-means. Steinbach, Karypis, and Kumar (2000) proposed a simple bisecting scheme for K-means clustering, which selects and divides a cluster into two sub-clusters in each iteration .

2.1.3 Distance Design

The distance function is one of the important factors that influence the performance of K-means. The most commonly used distance functions are the Euclidean quadratic function and K-L divergence. A distance function, $d(\cdot, \cdot)$ should fulfil the *metric inequality*.

$$d(i, j) + d(i, m) \geq d(j, m) \tag{2.2}$$

for pairs of individuals $(i, j), (i, m)$, and (j, m). The following table gives the three most widely applied K-means distance functions, where $x, y \in R^n$, $a, b \in (S_1^{n-1})_+$, and $(S_1^{n-1})_+ = \{a : a \in R^n, a[j] \geq 0, \sum_{j=1}^n a[j] = 1\}$. The Kullback-Leibler divergence is a widely applied distance function for count data.

Table 2.1: Distance Functions

Distance Name	Distance Function
Quadratic Distance	$d(x, y) = \|\|x - y\|\|^2$
Spherical Distance	$d(x, y) = x^t y$
Kullback-Leibler Divergence	$KL(a, b) = \sum_{i=1}^n a[i] \log \frac{a[i]}{b[i]}$

According to Table 2.1, Quadratic Distance and Kullback-Leibler are *metric* since they satisfy $d(x, x) = 0$, and $KL(a, a) = 0$, but the Spherical Distance does not have this property. A dissimilarity distance function is not necessarily a *metric*.

Banerjee, Merugu, Dhillon, and Ghosh (2005) studied the generalization problem of K-means clustering by using Bregman distance (also called "Bregman divergence") (Bregman, 1967). Let $\psi : R^n : \rightarrow (-\infty, +\infty]$ be a closed proper convex function. Suppose that ψ is continuously differentiable on domain of ψ. The Bregman distance $D_\psi : \text{domain}(\psi) \times \text{domain}(\psi) \rightarrow R_+$ is defined as

$$D_\psi(x, y) = \psi(x) - \psi(y) - \bigtriangledown\psi(y)(x - y), \tag{2.3}$$

where $\bigtriangledown\psi$ is the gradient of ψ. In general, Bregman distance is not *symmetric* and does not satisfy the *metric inequality*. Different choices of convex function $\psi(x)$ correspond to different Bregman distances.

Table 2.2: Convex Functions of Bregman Distance

Distance Name	Kernel Function
Quadratic Distance	$\psi(x) = \|\|x\|\|^2$
Kullback-Leibler Divergence	$\psi(x) = \sum_{i=1}^n a[i] \log\ a[i] - a[i]$

There is no one distance function that dominates others in clustering. Squared Euclidean distance often performs well when the data are normally distributed, while KL-divergence has higher clustering quality on some text data. We will introduce a new distance function for clustering, energy distance, in Section 2.3.

2.2 Hartigan and Wong's K-means Algorithm

In this section, we focus on the most widely applied Hartigan and Wong K-means algorithm. As we know, the aim of K-means algorithm is to divide N points in M dimensions into K clusters so that the within-cluster variance is minimized. The computational complexity of searching for a global optimum is NP-hard (Van Leeuwen, 1990), so it is not practical to require that the solution has minimal sum of squares against all partitions, except when M, N are small and $K = 2$. Hartigan proposed a K-means clustering algorithm in 1975 and gave a more efficient version in 1979 with M. A. Wong (Hartigan and Wong, 1979). The general procedure is to search for a K-partition with locally optimal within-cluster sum of squares by iteratively moving points from one cluster to another. In order to make the algorithm easier to understand, Hartigan and Wong's notation is used below to describe the 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 L is denoted by $NC(L)$. Euclidean distance between point $I = 1, 2, ..., N$ and cluster L is denoted as $D(I, L)$. $IC1(I)$ and $IC2(I)$ represent the closest and second closest cluster centers for observation I. Summarized from Hartigan and Wong (1979), with notation corresponding to the published Fortran Code, Hartigan and Wong's algorithm is the following.

Step 1. For each observation I, find its closest and second closest cluster centers, $IC1(I)$ and $IC2(I)$. Assign point I to cluster $IC1(I)$.

Step 2. Update the cluster centers to be the averages of observations contained within them.

Step 3. Initially, all clusters belong to the live set.

Step 4. This is the quick-transfer (QTRAN) stage: For each observation $I(I = 1, ..., N)$, let $L1$

be the cluster with center $IC1(I)$ and $L2$ be the cluster with center $IC2(I)$. Compute the values

$$R1 = [NC(L1) * D(I, L1)^2]/[NC(L1) - 1]$$

and

$$R2 = [NC(L2) * D(I, L2)^2]/[NC(L2) + 1].$$

If $R1$ is less than $R2$, observation I remains in cluster $L1$. Otherwise, switch $IC1(I)$ and $IC2(I)$ and update the centers of clusters $L1$ and $L2$.

Step 5. This is the optimal-transfer (OPTRA) stage: For each observation I $(I = 1, ..., N)$, if cluster L $(L = 1, ..., K)$ is updated in the last quick-transfer(QTRAN) stage, it belongs to the live set in this stage. Otherwise, at each step, it is not in the live set if it has not been updated in the last M optimal-transfer steps. Let point I be in cluster $L1$. If $L1$ is in the live set, do Step 5a; otherwise, do Step 5b.

Step 5a. Compute the minimum

$$R2 = [NC(L) * D(I, L)^2]/[NC(L) + 1]$$

for all clusters L $(L \neq L1, L = 1, 2, ..., K)$. Let $L2$ be the cluster with the smallest $R2$. If this value is greater than or equal to

$$R1 = [NC(L1) * D(I, L1)^2]/[NC(L1) - 1]$$

no reallocation is necessary; otherwise reallocate observation I to $L2$.

Step 5b. This step is the same as Step 5a, except that the minimum $R2$ is computed only over clusters in the live set.

Step 6. Stop if the live set is empty. Otherwise, go to Step 4 after one pass through the data set.

Step 7. If no transfer takes place in the last N steps, go to Step 5. Otherwise, go to Step 4.

The Hartigan and Wong K-means algorithm uses two passes (quick-transfer and optimal-

transfer) to search for the local optimum. This method can reduce the computational cost. Generally, in order to find the best cluster for each observation, we need to compute distances between the observation and each cluster, and find minimum of them. One needs to compute K times for each observation. However, in the Hartigan and Wong algorithm, one only needs to compute distances between each observation and each cluster in the live set, and the number of clusters in the live set usually is less than K. So the Hartigan and Wong K-means algorithm has less computational burden than LIoyd's algorithm (Lloyd, 1982)

2.3 Energy Distance

There are many types of distances between statistical objects. Cramér (1928) proposed an L_2 distance. If F is the cumulative distribution function (cdf) of a random variable and F_n is the empirical cdf, then the L_2 distance between F and F_n is

$$\int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dx. \tag{2.4}$$

This L_2 distance is not distribution free. If we want to apply this distance to a goodness-of-fit test, then the critical values depend on F. A variant of Cramér distance is Cramér–von Mises–Smirnov distance which replaces dx with $dF(x)$ (Pollaczek-Geiringer, 1928):

$$\int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dF(x). \tag{2.5}$$

Neither Cramér distance nor Cramér–von Mises–Smirnov distance are rotation invariant when the sample comes from a d-dimensional space, where $d > 1$.

G. J. Székely proposed *Energy Distance* in 1986 (Székely, Alpár, and Unger, 1986). 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 2.3.1. *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 $F(x)\,\text{and}\,G(x)$ 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. Székely (2000) gave the following proposition .

Proposition 2.1. *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.

When $d = 1$, $C_1 = \pi$, the energy distance is

$$\mathcal{E}(X,Y) = \frac{1}{\pi}\int_R \frac{|\hat{f}(t)-\hat{g}(t)|^2}{|t|^2}dt. \tag{2.6}$$

According to the Parseval-Plancherel formula, if $f(x)$ and $g(x)$ are the densities of X and Y, respectively, then

$$2\pi\int_{-\infty}^{\infty}(f(x)-g(x))^2dx = \int_{-\infty}^{\infty}|\hat{f}(t)-\hat{g}(t)|^2dt.$$

Since the Fourier transform of the cdf $F(x) = \int_{-\infty}^{x} f(u)du$ is $\hat{f}(t)/(it)$, where $i = \sqrt{-1}$, we have

$$2\pi \int_{-\infty}^{\infty} (F(x) - G(x))^2 dx = \int_{-\infty}^{\infty} \frac{|\hat{f}(t) - \hat{g}(t)|^2}{t^2} dt. \tag{2.7}$$

Based on Equation (2.6) and Equation (2.7), when $d = 1$, we have

$$\mathcal{E}(X, Y) = 2 \int_{-\infty}^{\infty} (F(x) - G(x))^2 dx.$$

Thus the energy distance generalizes the Cramér's L_2 distance. Since many important distributions do not have finite expected values, we need the following generalization of Proposition 2.1.

Proposition 2.2. *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,

$$\mathcal{E}^2(X, Y) = 2|E(X) - E(Y)|^2. \tag{2.8}$$

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 (2.8). An overview of the application of energy distance will be presented in section 2.4.

2.4 Application of Energy Distance

The applications of energy distance include:

- Consistent one-sample goodness-of-fit test (Székely and Rizzo, 2005b; Rizzo, 2009).
- Consistent multi-sample tests of equality of distribution (Székely and Rizzo, 2004; Rizzo, 2002; Baringhaus and Franz, 2004).
- Hierarchical clustering algorithm (Székely and Rizzo, 2005a) that extends and generalizes the Ward's minimum variance algorithm.
- Distance components (DISCO) (Rizzo and Székely, 2010), a nonparametric extension of analysis of variance for structured data.

Advantages of energy distance include:

- Energy distance is closely related to Cramér's L_2 distance.
- Energy distance is very easy to compute and simulate.
- Energy tests are rigid motion invariant and applicable for data in arbitrary dimension.
- The energy tests are effective compared to classical tests of homogeneity and normality, especially in higher dimensions.
- Tests based on energy distance are consistent and require no distributional assumptions other than finite first moments.

Since our research focuses on cluster analysis, we only introduce a few applications of energy distance related to this topic.

2.4.1 Testing for Equal Distributions

The two-sample energy statistic for independent random samples $\mathbf{X} = \{X_1, ..., X_{n_1}\}$ and $\mathbf{Y} = \{Y_1, ..., Y_{n_2}\}$ is

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

The weighted two-sample statistic

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

can be applied for testing homogeneity (equality of distributions of X and Y). The statistic $T_{X,Y}$ tends to infinity stochastically as n_1 and n_2 increase if and only if the null distribution of homogeneity does not hold. Under the null hypothesis, the limiting distribution of $T_{X,Y}$ is a quadratic form of iid standard normal random variables. Rizzo (2002) applied a nonparametric boostrap to obtain a distribution free test procedure. Under the null hypothesis, $X_1, ...X_{n_1}, Y_1, ...Y_{n_2}$ are iid, with common distribution function $F_1 = F_2 = F$. If the significance level is α, resample from $X_1, ...X_{n_1}, Y_1, ...Y_{n_2}$ a suitable number B of random samples. For each pooled sample $X_1^b, ...X_{n_1}^b, Y_1^b, ...Y_{n_2}^b, b = 1, ..., B$, compute ${T_{n_1,n_2}}^b$. Then the boostrap estimate of $P(T_{n_1,n_2} \leq t)$ is $\frac{1}{B}\sum_{b=1}^{B} I({T_{n_1,n_2}}^b) \leq t)$, where $I(\cdot)$ is the indicator function. We reject the null hypothesis if the observed T_{n_1,n_2} exceeds $100(1-\alpha)\%$ of the replicated ${T_{n_1,n_2}}^b$.

2.4.2 A Nonparametric Extension of ANOVA

A multi-sample test of equal distributions is a type of generalization of the hypothesis of equal means. Similar to the ANOVA decomposition of variance, we can get a decomposition of distances called distance components (DISCO) (Rizzo and Székely, 2010). To simplify subsequent notation,

for two samples $A = a_1, a_2, ..., a_{n_1}$ and $B = b_1, b_2, ..., b_{n_1}$

$$G^{\alpha}(A, B) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{m=1}^{n_2} |a_i - b_m|^{\alpha}, \tag{2.9}$$

where $0 < \alpha < 2$. $G^{\alpha}(A, B)$ is a version of the Gini mean distance statistic, and the constant $\frac{n_1 n_2}{n_1 + n_2}$ is half the harmonic mean of the sample sizes (David, 1968). Suppose that $A_1, ..., A_k$ are the k samples of size $n_1, ..., n_k$, respectively, and $N = \sum_{i=1}^{k} n_i$. The total dispersion of the k samples is

$$T^{\alpha}(A_1, ...A_k) = \frac{N}{2} G^{\alpha}(\cup_{i=1}^{k} A_i, \cup_{i=1}^{k} A_i). \tag{2.10}$$

The within-sample dispersion is defined by

$$W^{\alpha}(A_1, ...A_k) = \sum_{j=1}^{k} \frac{n_j}{2} G^{\alpha}(A_j, A_j). \tag{2.11}$$

The between-sample energy statistic is defined by

$$B^{\alpha}(A_1, ...A_k) = \sum_{1 \le i < j \le k} \left\{ \frac{n_i n_j}{2N} \left[2G^{\alpha}(A_i, A_j) - G^{\alpha}(A_i, A_i) - G^{\alpha}(A_j, A_j) \right] \right\}. \tag{2.12}$$

For all $0 < \alpha < 2$ we have the decomposition $T^{\alpha} = W^{\alpha} + B^{\alpha}$, where both W^{α} and B^{α} are nonnegative. Rizzo and Székely (2010) prove that

$$B^{\alpha}(A_1, ...A_k) = \sum_{1 \le j \le i \le k} \frac{n_i + n_j}{2N} \left[\frac{n_i n_j}{n_i + n_j} \mathcal{E}^{\alpha}_{n_i, n_j}(A_j, A_i) \right]. \tag{2.13}$$

When $0 < \alpha < 2$, the statistic (2.13) determines a statistically consistent test of the hypothesis that the distributions are identical. When $\alpha = 2$ the statistic (2.13) equals zero if the means of the distributions are identical. Actually, when $\alpha = 2$ and dimension is 1, the statistic B^2 is the sample sum of squared error (Rizzo and Székely, 2010). The decomposition $T^2 = W^2 + B^2$ is the ANOVA decomposition for the one factor model. The most common ANOVA F test and T test require the normal assumption and equal variances. However, the distance components test is valid

under the mild assumption that finite α-moments exist. In this sense, distance components is more general than parametric ANOVA F tests.

2.4.3 Hierarchical Clustering

Energy distance can be applied in hierarchical cluster analysis. In agglomerative hierarchical clustering algorithms, Székely and Rizzo (2005a) computed the energy distance between clusters and merge clusters with minimum energy distance at each step; each observation is an individual cluster at initial step. As a member of the general class of hierarchical clustering algorithms including Ward's minimum variance (Ward, 1963), the algorithm is uniquely determined by its recursive formula for updating energy distances between clusters.

Suppose at the current step in the hierarchical clustering, the disjoint clusters π_i and π_j would be merged. Then the energy distance between the new cluster $\pi_i \cup \pi_j$ and a disjoint cluster π_k is given by the following formula,

$$\begin{aligned}\mathcal{E}^{\alpha}_{n_i,n_j}(\pi_i \cup \pi_j, \pi_k) =& \frac{n_i + n_k}{n_i + n_j + n_k}\mathcal{E}^{\alpha}_{n_i,n_k}(\pi_i, \pi_k) + \frac{n_j + n_k}{n_i + n_j + n_k}\mathcal{E}^{\alpha}_{n_j,n_k}(\pi_j, \pi_k) \\ & - \frac{n_k}{n_i + n_j + n_k}\mathcal{E}^{\alpha}_{n_i,n_j}(\pi_i, \pi_j). \end{aligned} \tag{2.14}$$

Based on formula (2.14), one can compute the energy distance recursively. When $\alpha = 2$, the formula (2.14) is the updating formula for Ward's minimum variance method (Ward, 1963). Thus the energy hierarchical clustering generalizes the Ward's minimum variance method.

CHAPTER 3

K-GROUPS BY FIRST VARIATION

In this chapter, we propose the K-groups by first variation algorithm. K-means uses quadratic distance to compute the dissimilarity between the data object and the prespecified prototype, and minimizes the within-cluster variance. We use a weighted two-sample energy statistic, $T^{\alpha}_{X,Y}$, as the statistical distance function to measure the dissimilarity between two clusters. Since the energy distance measures the dissimilarity between two sets rather than the similarity between the objects and prototypes, we name our method *K-groups*.

Based on Section 2.4.2, let $P = \{\pi_1, ..., \pi_k\}$ be a partition of observations with sizes $n_1, n_2, ...n_k$, 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 $N = \sum_1^k n_i$ is the total number of observations. The within-group dispersion is defined by

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

The between-group dispersion is

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

For all $0 < \alpha < 2$, we have the decomposition

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

where both $W^{\alpha}(\pi_1, ...\pi_k)$ and $B^{\alpha}(\pi_1, ...\pi_k)$ are nonnegative. Our purpose is to find the best partition which minimizes W^{α}. Hence, the K-groups clustering problem is equivalent to the optimization task of finding

$$\min_{\pi_1,...,\pi_k} \sum_{j=1}^{k} \frac{n_j}{2} G^{\alpha}(\pi_j, \pi_j) = \min_{\pi_1,...,\pi_k} W^{\alpha}(\pi_1, ...\pi_k). \tag{3.1}$$

3.1 First Variation

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

Definition 3.1.1. A first variation of a partition P is a partition P' obtained from P by removing a single point **a** from a cluster π_i of P and assigning this point to an existing cluster π_j of P, where $i \neq j$.

Let π_1 and π_2 be two different clusters in partition $P = \{\pi_1, ..., \pi_k\}$, and point $\mathbf{a} \in \pi_1$. Cluster π_1^- represents cluster π_1 after removing point **a**, and cluster π_2^+ represents cluster π_2 after adding point **a**. Let n_1 and n_2 be the sizes of clusters π_1 and π_2 before moving point **a**. The dispersions of cluster π_1 and π_2 are

$$\frac{n_1}{2} G^{\alpha}(\pi_1, \pi_1) = \frac{1}{2n_1} \sum_{i}^{n_1} \sum_{j}^{n_1} |x_i^1 - x_j^1|^{\alpha},$$
$$\frac{n_2}{2} G^{\alpha}(\pi_2, \pi_2) = \frac{1}{2n_2} \sum_{i}^{n_2} \sum_{j}^{n_2} |x_i^2 - x_j^2|^{\alpha},$$

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

$$\frac{n_1 - 1}{2} G^\alpha(\pi_1^-, \pi_1^-) = \frac{1}{2 \cdot (n_1 - 1)} \sum_i^{n_1 - 1} \sum_j^{n_1 - 1} |x_i^{-1} - x_j^{-1}|^\alpha,$$
$$\frac{n_2 + 1}{2} G^\alpha(\pi_2^+, \pi_2^+) = \frac{1}{2 \cdot (n_2 + 1)} \sum_i^{n_2 + 1} \sum_j^{n_2 + 1} |x_i^{+2} - x_j^{+2}|^\alpha,$$

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 **a** and clusters π_1 and π_2 are

$$\xi^\alpha(\mathbf{a}, \pi_1) = \frac{2}{n_1} \sum_i^{n_1} |x_i^1 - \mathbf{a}|^\alpha - \frac{1}{n_1^2} \sum_i^{n_1} \sum_j^{n_1} |x_i^1 - x_j^1|^\alpha, \tag{3.2}$$

$$\xi^\alpha(\mathbf{a}, \pi_2) = \frac{2}{n_2} \sum_i^{n_2} |x_i^2 - \mathbf{a}|^\alpha - \frac{1}{n_2^2} \sum_i^{n_2} \sum_j^{n_2} |x_i^2 - x_j^2|^\alpha. \tag{3.3}$$

First, we compute $\frac{n_1}{2} G^\alpha(\pi_1, \pi_1) - \frac{n_1 - 1}{2} G^\alpha(\pi_1^-, \pi_1^-)$:

$$\begin{aligned}
&\frac{n_1}{2} G^\alpha(\pi_1, \pi_1) - \frac{n_1 - 1}{2} G^\alpha(\pi_1^-, \pi_1^-) \\
&\quad = \frac{1}{2 \cdot n_1} \sum_i^{n_1} \sum_j^{n_1} |x_i^1 - x_j^1|^\alpha - \frac{1}{2 \cdot (n_1 - 1)} \sum_i^{n_1 - 1} \sum_j^{n_1 - 1} |x_i^{-1} - x_j^{-1}|^\alpha \\
&\quad = \frac{1}{2 \cdot n_1} \sum_i^{n_1} \sum_j^{n_1} |x_i^1 - x_j^1|^\alpha - \frac{1}{2 \cdot (n_1 - 1)} \left[\sum_i^{n_1} \sum_j^{n_1} |x_i^1 - x_j^1|^\alpha \right. \\
&\qquad \left. - 2 \sum_i^{n_1} |x_i^1 - \mathbf{a}|^\alpha \right] \\
&\quad = \frac{1}{n_1 - 1} \sum_i^{n_1} |x_i^1 - \mathbf{a}|^\alpha - \frac{1}{2 \cdot n_1 (n_1 - 1)} \sum_i^{n_1} \sum_j^{n_1} |x_i^1 - x_j^1|^\alpha.
\end{aligned} \tag{3.4}$$

Multiply equation (3.2) times $\frac{n_1}{2(n_1-1)}$ to obtain

$$\begin{aligned}\frac{n_1}{2(n_1-1)}\xi^\alpha(\mathbf{a},\pi_1) =& \frac{1}{n_1-1}\sum_i^{n_1}|x_i^1-\mathbf{a}|^\alpha \\ &- \frac{1}{2\cdot n_1(n_1-1)}\sum_i^{n_1}\sum_j^{n_1}|x_i^1-x_j^1|^\alpha. \end{aligned} \tag{3.5}$$

Subtract (3.5) from (3.4) to obtain

$$\frac{n_1}{2}G^\alpha(\pi_1,\pi_1)-\frac{n_1-1}{2}G^\alpha(\pi_1^-,\pi_1^-)=\frac{n_1}{2(n_1-1)}\xi^\alpha(\mathbf{a},\pi_1). \tag{3.6}$$

Next, we compute $\frac{n_2+1}{2}G^\alpha(\pi_2^+,\pi_2^+)-\frac{n_2}{2}G^\alpha(\pi_2,\pi_2)$:

$$\begin{aligned}&\frac{n_2+1}{2}G^\alpha(\pi_2^+,\pi_2^+)-\frac{n_2}{2}G^\alpha(\pi_2,\pi_2) \\ &\quad=\frac{1}{2\cdot(n_2+1)}\sum_i^{n_2+1}\sum_j^{n_2+1}|x_i^{+2}-x_j^{+2}|^\alpha-\frac{1}{2\cdot n_2}\sum_i^{n_2}\sum_j^{n_2}|x_i^2-x_j^2|^\alpha \\ &\quad=\frac{1}{2\cdot(n_2+1)}\left[\sum_i^{n_2}\sum_j^{n_2}|x_i^2-x_j^2|^\alpha+2\sum_i^{n_2}|x_i^2-\mathbf{a}|^\alpha\right] \\ &\qquad-\frac{1}{2\cdot n_2}\sum_i^{n_2}\sum_j^{n_2}|x_i^2-x_j^2|^\alpha \\ &\quad=\frac{1}{n_2+1}\sum_i^{n_2}|x_i^2-\mathbf{a}|^\alpha-\frac{1}{2\cdot n_2(n_2+1)}\sum_i^{n_2}\sum_j^{n_2}|x_i^2-x_j^2|^\alpha. \end{aligned} \tag{3.7}$$

Equation (3.3) times $\frac{n_2}{2(n_2+1)}$ is

$$\begin{aligned}\frac{n_2}{2(n_1+1)}\xi^\alpha(\mathbf{a},\pi_2) =& \frac{1}{n_2+1}\sum_i^{n_2}|x_i^2-\mathbf{a}|^\alpha \\ &- \frac{1}{2\cdot n_2(n_2+1)}\sum_i^{n_2}\sum_j^{n_2}|x_i^2-x_j^2|^\alpha. \end{aligned} \tag{3.8}$$

Subtract equation (3.8) from equation (3.7) to obtain

$$\frac{n_2+1}{2}G^{\alpha}(\pi_2^+,\pi_2^+)-\frac{n_2}{2}G^{\alpha}(\pi_2,\pi_2)=\frac{n_2}{2(n_2+1)}\xi^{\alpha}(\mathbf{a},\pi_2). \tag{3.9}$$

Subtract equation (3.9) from equation (3.6) to obtain

$$\begin{aligned}\frac{n_1}{2}G^{\alpha}(\pi_1,\pi_1)+\frac{n_2}{2}G^{\alpha}(\pi_2,\pi_2)-\frac{n_1-1}{2}G^{\alpha}(\pi_1^-,\pi_1^-)-\frac{n_2+1}{2}G^{\alpha}(\pi_2^+,\pi_2^+)\\ =\frac{n_1}{2(n_1-1)}\xi^{\alpha}(\mathbf{a},\pi_1)-\frac{n_2}{2(n_2+1)}\xi^{\alpha}(\mathbf{a},\pi_2).\end{aligned} \tag{3.10}$$

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

Theorem 3.1. *Suppose that $P=\{\pi_1,\pi_2,...\pi_k\}$ is a partition and first variation $P'=\{\pi_1^-,\pi_2^+,...,\pi_k\}$ is obtained from P by moving a point $\boldsymbol{a}\in\pi_1$ to cluster π_2. Then*

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

Similar to the Hartigan and Wong K-means algorithm, we move 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 Theorem 3.1 above we propose the K-groups by first variation algorithm to search for an optimal partition with respect to energy distance between clusters.

3.2 K-groups Algorithm by First Variation

We seek a local solution such that no movement of a point from one cluster to another will reduce the within-cluster sum of dispersion $W^{\alpha}(P)$. Here notation is consistent with Section 3.1, but different from the notation of Hartigan and Wong algorithm in Section 2.2.

Notation Let N be the total sample size of observations, M be the dimension of the sample,

and K be the number of clusters, 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 and cluster π_i is denoted by $\xi^\alpha(I, \pi_i)$. The K-groups by first varaition algorithm is the following:

Step 1. Each point $I\,(I = 1, ..., N)$ is randomly assigned to cluster $\pi_i\,(i = 1, ..., K)$. Let $\pi(I)$ represent the cluster containing I, and $n(\pi(I))$ represent 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 the minimum

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

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

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

In the case $\alpha = 2$, for univariate data, we can prove that the K-groups by first variation algorithm and Hartigan and Wong K-means algorithm have the same objective function and updating formula.

Proposition 3.2. *When $\alpha = 2$, for univariate data,*

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

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.

$$\begin{aligned}
\frac{n_i}{2}G^2(\pi_i,\pi_i) &= \frac{1}{2n_i}\sum_{l=1}^{n_i}\sum_{m=1}^{n_i}|x_l - x_m|^2 \\
&= \frac{1}{2n_i}\sum_{l=1}^{n_i}\sum_{m=1}^{n_i}(x_l^2 - 2x_l x_m + x_m^2) \\
&= \frac{1}{2n_i}\left[n_i\sum_{l=1}^{n_i}x_l^2 - 2\sum_{l=1}^{n_i}\sum_{m=1}^{n_i}x_l x_m + n_i\sum_{m=1}^{n_i}x_m^2\right] \\
&= \frac{1}{2n_i}\left[2n_i\sum_{l=1}^{n_i}x_l^2 - 2\sum_{l=1}^{n_i}\sum_{m=1}^{n_i}x_l x_m\right] \\
&= \frac{1}{2n_i}\left[2n_i\sum_{l=1}^{n_i}x_l^2 - 2n_i^2 c_i^2\right] \\
&= \sum_{l=1}^{n_i}x_l^2 - n_i c_i^2. \qquad (3.11)
\end{aligned}$$

□

Proposition 3.3. *For univariate data,*

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

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.

$$\begin{aligned}
\sum_{x_j\in\pi_i}(x_j - c_i)^2 &= \sum_{j=1}^{n_i}\left(x_j^2 - 2\cdot x_j\cdot c_i + c_i^2\right) \\
&= \sum_{j=1}^{n_i}x_j^2 - 2\sum_{j=1}^{n_i}x_j c_i + \sum_{j=1}^{n-i}c_i^2 \\
&= \sum_{j=1}^{n_i}x_j^2 - 2n_i c_i^2 + n_i c_i^2 \\
&= \sum_{j=1}^{n_i}x_j^2 - n_i c_i^2. \qquad (3.12)
\end{aligned}$$

□

Since the objective for K-means is

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

and the objective for K-groups is

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

based on Proposition (3.2) and Proposition (3.3) 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 $i = 1, ..., k$. Hence K-groups and K-means have the same objective function for univariate data when $\alpha = 2$. We have the following theorem.

Theorem 3.4. *When $\alpha = 2$, for univariate data the K-groups algorithm and Hartigan and Wong K-means algorithm have the same objective function.*

Next, we prove that the updating formulas of K-groups by first variation and Hartigan and Wong K-means algorithm, are formally the same for univariate data when $\alpha = 2$.

Proposition 3.5. *Suppose point I belongs to cluster L, and the sample size of L is n. Then*

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

Proof. We only need to prove that $\frac{1}{2}\xi^2(I, L) = D(I, L)^2$. We have

$$\xi^2(I, L) = \frac{2}{n} \sum_{i=1}^{n} |I - x_i|^2 - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|^2, \tag{3.13}$$

and

$$D(I,L)^2 = (I - \frac{\sum_{i=1}^{n} x_i}{n})^2. \tag{3.14}$$

We can simplify equation (3.13) as follows:

$$\begin{aligned}
\xi^2(I,L) &= \frac{2}{n}\sum_{i=1}^{n}|I - x_i|^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|^2 \\
&= \frac{2}{n}\left(nI^2 - 2I\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}x_i^2\right) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i^2 - 2x_ix_j + x_j^2\right) \\
&= \left(2I^2 - 4I\frac{\sum_{i=1}^{n}x_i^2}{n} + 2\frac{\sum_{i=1}^{n}x_i^2}{n}\right) - \frac{1}{n^2}\left(2n\sum_{i=1}^{n}x_i^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n}x_ix_j\right) \\
&= \left(2I^2 - 4I\frac{\sum_{i=1}^{n}x_i^2}{n} + 2\frac{\sum_{i=1}^{n}x_i^2}{n}\right) - \left(2\frac{\sum_{i=1}^{n}x_i^2}{n} - 2\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}x_ix_j}{n^2}\right) \\
&= 2\left(I^2 - 2I\frac{\sum_{i=1}^{n}x_i}{n} + \frac{2\sum_{i=1}^{n}\sum_{j=1}^{n}x_ix_j}{n^2}\right) \\
&= 2\left(I - \frac{\sum_{i=1}^{n}x_i}{n}\right)^2.
\end{aligned} \tag{3.15}$$

Based on equation (3.15) and equation (3.14), we have

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

□

By similar steps, we have

$$\frac{n}{2(n+1)}\xi^2(I,L) = \frac{n \cdot D(I,L)^2}{n+1}. \tag{3.16}$$

Based on Proposition (3.5) and equation (3.16), updating formulas of K-groups by first variation and Hartigan and Wong K-means algorithm are the same for univariate data when $\alpha = 2$.

According to the results above, K-means is a special case of K-groups by first variation for univariate data when $\alpha = 2$. Based on the properties of energy statistics in Section 2.3, we know

that when $0 < \alpha < 2$, the energy distance $\xi^{\alpha}(X, Y) = 0$ if and only if random variables X and Y follow the same statistical distribution. However, when $\alpha = 2$ we have $\xi^{\alpha}(X, Y) = 0$ when $E(X) = E(Y)$. In spite of the fact that the objective functions and updating formulas for K-groups by first variation and K-means are equivalent when $\alpha = 2$, they are clearly different clustering methods. Note that K-means separates clusters with different means, while K-groups separates clusters that differ in distribution when $\alpha < 2$. Hence K-groups by first variation generalizes the Hartigan and Wong's K-means algorithm.

CHAPTER 4

K-GROUPS BY SECOND VARIATION

The purpose 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 this section, we introduce the second variation method and generalize to m^{th} variation where $m > 1$.

Definition 4.0.1. A second variation of a partition P is a partition $P^{(2)}$ obtained from P by removing two points $\{a_1, a_2\}$ from a cluster π_i of P and assigning these points to an existing cluster π_j of P, $i \neq j$.

Definition 4.0.2. 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 , $i \neq j$.

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 Székely and Rizzo (2004), 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 the m^{th} variation. 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.

To illustrate, we generate two random samples with sizes m and 100, and compute the energy distance between these two random samples with different value of $m = 1, 2, ..., 100$. Figure 4.1 shows how the energy distance changes when we increase the value of m.

Based on Figure 4.1(b), we can see empirically how energy distance tends to infinity stochastically as m increases, if two random samples are generated from different statistical distributions. When two random samples follow the same statistical distribution, as in Figure 4.1(a), this is not the case.

The idea of moving more than one point idea does not work very well for the K-means algorithm. The quadratic distance measures the difference between the means of random samples. To illustrate, we generate two random samples with size m and 100, then compute the quadratic distance between these two random samples with different value of $m = 1, 2, ..., 100$. Figure 4.2 shows how the quadratic distance changes when we increase the value of m.

Based on Figures 4.2(a) and 4.2(b), when the means are equal, no matter whether two random samples are generated from the same statistical distribution or not, the quadratic distances for different values of m are bounded. The quadratic distance can not distinguish between two samples when they have the same mean. However, when two random samples have different means, as in Figures 4.3(a) and 4.3(b) quadratic distance increases as value of m increases.

4.1 Second Variation

In this section, we show that energy statistics have a nice updating formula for the second variation. We use $\{a_1, a_2\}$ to represent two points. Let π_1^- represent the cluster after removing points

$\{a_1, a_2\}$, and cluster π_2^+ denote the cluster after adding those two points. Partition $P^{(2)}$ is the second variation of partition P. The two-sample energy statistic between points $\{a_1, a_2\}$ and cluster π_1 and π_2 are by definition

$$\begin{aligned}\xi^\alpha(\{a_1, a_2\}, \pi_1) =& \frac{2}{2n_1}\sum_i^{n_1}\sum_j^{2}|x_i^1 - a_j|^\alpha - \frac{1}{n_1^2}\sum_i^{n_1}\sum_j^{n_1}|x_i^1 - x_j^1|^\alpha \\ &- \frac{1}{4}\sum_i^{2}\sum_j^{2}|a_i - a_j|^\alpha, \end{aligned} \tag{4.1}$$

and

$$\begin{aligned}\xi^\alpha(\{a_1, a_2\}, \pi_2) =& \frac{2}{2n_2}\sum_i^{n_2}\sum_j^{2}|x_i^2 - a_j|^\alpha - \frac{1}{n_2^2}\sum_i^{n_2}\sum_j^{n_2}|x_i^2 - x_j^2|^\alpha \\ &- \frac{1}{4}\sum_i^{2}\sum_j^{2}|a_i - a_j|^\alpha. \end{aligned} \tag{4.2}$$

First, we compute the difference between $\frac{n_1}{2}G^\alpha(\pi_1, \pi_1) - \frac{n_1-2}{2}G^\alpha(\pi_1^-, \pi_1^-)$, as

$$\begin{aligned}&\frac{n_1}{2}G^\alpha(\pi_1, \pi_1) - \frac{n_1-2}{2}G^\alpha(\pi_1^-, \pi_1^-) \\ &\quad= \frac{1}{2\cdot n_1}\sum_i^{n_1}\sum_j^{n_1}|x_i^1 - x_j^1|^\alpha - \frac{1}{2\cdot(n_1-2)}\sum_i^{n_1-2}\sum_j^{n_1-2}|x_i^{-1} - x_j^{-1}|^\alpha \\ &\quad= \frac{1}{2\cdot n_1}\sum_i^{n_1}\sum_j^{n_1}|x_i^1 - x_j^1|^\alpha - \frac{1}{2\cdot(n_1-2)}\left[\sum_i^{n_1}\sum_j^{n_1}|x_i^1 - x_j^1|^\alpha \right. \\ &\qquad \left. -2\sum_i^{n_1}\sum_j^{2}|x_i^1 - a_j|^\alpha + \sum_i^{2}\sum_j^{2}|a_i - a_j|^\alpha\right] \end{aligned}$$

$$\begin{aligned}&= \frac{1}{n_1-2}\sum_i^{n_1}\sum_j^{2}|x_i^1 - a_j|^\alpha - \frac{2}{2\cdot n_1(n_1-2)}\sum_i^{n_1}\sum_j^{n_1}|x_i^1 - x_j^1|^\alpha \\ &\quad - \frac{1}{2\cdot(n_1-2)}\sum_i^{2}\sum_j^{2}|a_i - a_j|^\alpha. \end{aligned} \tag{4.3}$$

Multiply $\frac{2n_1}{2(n_1-2)}$ times equation (4.1) to obtain

$$\begin{aligned}\frac{2n_1}{2(n_1-2)}\xi^{\alpha}(\{a_1,a_2\},\pi_1) =& \frac{1}{n_1-2}\sum_{i}^{n_1}\sum_{j}^{2}|x_i^1-a_j|^{\alpha} \\ &- \frac{2}{2\cdot n_1(n_1-2)}\sum_{i}^{n_1}\sum_{j}^{n_1}|x_i^1-x_j^1|^{\alpha} \\ &- \frac{n_1}{4\cdot(n_1-2)}\sum_{i}^{2}\sum_{j}^{2}|a_i-a_j|^{\alpha}. \end{aligned} \tag{4.4}$$

Subtract equation (4.4) from equation (4.3) to obtain

$$\frac{n_1}{2}G^{\alpha}(\pi_1,\pi_1)-\frac{n_1-2}{2}G^{\alpha}(\pi_1^-,\pi_1^-)=\frac{2n_1}{2(n_1-2)}\xi^{\alpha}(\{a_1,a_2\},\pi_1)+\frac{1}{4}\sum_{i}^{2}\sum_{j}^{2}|a_i-a_j|^{\alpha}. \tag{4.5}$$

Then we compute $\frac{n_2+2}{2}G^{\alpha}(\pi_2^+,\pi_2^+)-\frac{n_2}{2}G^{\alpha}(\pi_2,\pi_2)$, as

$$\begin{aligned}&\frac{n_2+2}{2}G^{\alpha}(\pi_2^+,\pi_2^+)-\frac{n_2}{2}G^{\alpha}(\pi_2,\pi_2) \\ &\quad=\frac{1}{2\cdot(n_2+2)}\sum_{i}^{n_2+2}\sum_{j}^{n_2+2}|x_i^{+2}-x_j^{+2}|^{\alpha}-\frac{1}{2\cdot n_2}\sum_{i}^{n_2}\sum_{j}^{n_2}|x_i^2-x_j^2|^{\alpha} \\ &\quad=\frac{1}{2\cdot(n_2+2)}\left[\sum_{i}^{n_2}\sum_{j}^{n_2}|x_i^2-x_j^2|^{\alpha}+2\sum_{i}^{n_2}\sum_{j}^{2}|x_i^2-a_j|^{\alpha}\right. \\ &\qquad\left.+\sum_{i}^{2}\sum_{j}^{2}|a_i-a_j|^{\alpha}\right]-\frac{1}{2\cdot n_2}\sum_{i}^{n_2}\sum_{j}^{n_2}|x_i^2-x_j^2|^{\alpha} \\ &\quad=\frac{1}{n_2+2}\sum_{i}^{n_2}\sum_{j}^{2}|x_i^2-a_j|^{\alpha}-\frac{2}{2\cdot n_2(n_1+2)}\sum_{i}^{n_2}\sum_{j}^{n_2}|x_i^2-x_j^2|^{\alpha} \\ &\qquad+\frac{1}{2\cdot(n_2+2)}\sum_{i}^{2}\sum_{j}^{2}|a_i-a_j|^{\alpha}\}. \end{aligned} \tag{4.6}$$

Multiply $\frac{2n_2}{2(n_2+2)}$ times equation(4.2) to get

$$\begin{aligned}\frac{2n_2}{2(n_2+2)}\xi^{\alpha}(\{a_1,a_2\},\pi_2) =& \frac{2}{n_2+2}\sum_{i}^{n_2}\sum_{j}^{2}|x_i^2-a_j|^{\alpha} \\ &- \frac{2}{2n_2(n_2+2)}\sum_{i}^{n_2}\sum_{j}^{n_2}|x_i^2-x_j^2|^{\alpha} \\ &- \frac{n_2}{4\cdot(n_2+2)}\sum_{i}^{2}\sum_{j}^{2}|a_i-a_j|^{\alpha}. \end{aligned} \tag{4.7}$$

Subtract equation (4.7) from equation (4.6) to get

$$\frac{n_2+2}{2}G^{\alpha}(\pi_2^+,\pi_2^+)-\frac{n_2}{2}G^{\alpha}(\pi_2,\pi_2)=\frac{2n_2}{2(n_2+1)}\xi^{\alpha}(\{a_1,a_2\},\pi_2)+\frac{1}{4}\sum_{i}^{2}\sum_{j}^{2}|a_i-a_j|^{\alpha}. \tag{4.8}$$

Finally, subtract equation (4.8) from equation (4.5) to obtain

$$\begin{aligned}&\frac{n_1}{2}G^{\alpha}(\pi_1,\pi_1)+\frac{n_2}{2}G^{\alpha}(\pi_2,\pi_2)-\frac{n_1-2}{2}G^{\alpha}(\pi_1^-,\pi_1^-)-\frac{n_2+2}{2}G^{\alpha}(\pi_2^+,\pi_2^+) \\ &\qquad =\frac{2n_1}{2(n_1-2)}\xi^{\alpha}(\{a_1,a_2\},\pi_1)-\frac{2n_2}{2(n_2+2)}\xi^{\alpha}(\{a_1,a_2\},\pi_2). \end{aligned} \tag{4.9}$$

Theorem 4.1. *Suppose $P=\{\pi_1,\pi_2,...\pi_k\}$ is a partition, and $P^{(2)}=\{\pi_1^-,\pi_2^+,...,\pi_k\}$ is a second variation of P by moving points $\{a_1,a_2\}$ from cluster π_1 to π_2. Then*

$$W^{\alpha}(P)-W^{\alpha}(P^{(2)})=\frac{n_1}{(n_1-2)}\xi^{\alpha}(\{a_1,a_2\},\pi_1)-\frac{n_2}{(n_2+2)}\xi^{\alpha}(\{a_1,a_2\},\pi_2).$$

Similar to K-groups by first variation, we move points $\{a_1,a_2\}$ from cluster π_1 to π_2 if

$$\frac{n_1}{(n_1-2)}\xi^{\alpha}(\{a_1,a_2\},\pi_1)-\frac{n_2}{(n_2+2)}\xi^{\alpha}(\{a_1,a_2\},\pi_2)$$

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

Proposition 4.2. *Let $\{a_1,a_2\}$ be two different observations in cluster π_1. Cluster π_2 is a member*

of partition $P = \{\pi_1, \pi_2, ...\pi_k\}$ *and* $\pi_1 \neq \pi_2$. *Let* $n_1 \text{ and } n_2$ *be the sizes of cluster* π_1 *and* π_2. *Then*

$$\begin{aligned}
&\frac{2n_1}{n_1-2}\xi^\alpha(\{a_1,a_2\},\pi_1) - \frac{2n_2}{n_2+2}\xi^\alpha(\{a_1,a_2\},\pi_2)\\
=&\frac{n_1-1}{n_1-2}\frac{n_1}{n_1-1}\xi^\alpha(a_1,\pi_1) - \frac{n_2+1}{n_2+2}\frac{n_2}{n_2+1}\xi^\alpha(a_1,\pi_2)\\
&+\frac{n_1-1}{n_1-2}\frac{n_1}{n_1-1}\xi^\alpha(a_2,\pi_1) - \frac{n_2+1}{n_2+2}\frac{n_2}{n_2+1}\xi^\alpha(a_2,\pi_2)\\
&-\frac{n_1+n_2}{(n_1-2)(n_2+2)}\sum_{i=1}^{2}\sum_{j=1}^{2}|a_i-a_j|^\alpha.
\end{aligned}$$

Proof. The two-sample energy statistics between points a_1, a_2 and clusters π_1, π_2 are

$$\xi^\alpha(a_1,\pi_1) = \frac{2}{n_1}\sum_{j=1}^{n_1}|a_1-x_j^1|^\alpha - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i^1-x_j^1|^\alpha, \tag{4.10}$$

$$\xi^\alpha(a_2,\pi_1) = \frac{2}{n_1}\sum_{j=1}^{n_1}|a_2-x_j^1|^\alpha - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i^1-x_j^1|^\alpha, \tag{4.11}$$

$$\xi^\alpha(a_1,\pi_2) = \frac{2}{n_2}\sum_{j=1}^{n_2}|a_1-x_j^2|^\alpha - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}|x_i^2-x_j^2|^\alpha, \tag{4.12}$$

and

$$\xi^\alpha(a_2,\pi_2) = \frac{2}{n_2}\sum_{j=1}^{n_2}|a_2-x_j^2|^\alpha - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}|x_i^2-x_j^2|^\alpha. \tag{4.13}$$

The two-sample energy statistics between the set $\{a_1, a_2\}$ and clusters π_1, π_2 are

$$\begin{aligned}
\xi^\alpha(\{a_1,a_2\},\pi_1) =&\frac{2}{2n_1}\sum_{i=1}^{2}\sum_{j=1}^{n_1}|a_i-x_j^1|^\alpha - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i^1-x_j^1|^\alpha\\
&-\frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}|a_i-a_j|^\alpha,
\end{aligned} \tag{4.14}$$

and

$$\begin{aligned}\xi^{\alpha}(\{a_1,a_2\},\pi_2) =& \frac{2}{2n_2}\sum_{i=1}^{2}\sum_{j=1}^{n_2}|a_i-x_j^2|^{\alpha}-\frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}|x_i^2-x_j^2|^{\alpha}\\ &-\frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}|a_i-a_j|^{\alpha}. \end{aligned} \tag{4.15}$$

Subtract $\frac{2n_2}{n_2+2}\xi^{\alpha}(\{a_1,a_2\},\pi_2)$ from $\frac{2n_1}{n_1-2}\xi^{\alpha}(\{a_1,a_2\},\pi_1)$ to obtain

$$\begin{aligned}&\frac{2n_1}{n_1-2}\xi^{\alpha}(\{a_1,a_2\},\pi_1)-\frac{2n_2}{n_2+2}\xi^{\alpha}(\{a_1,a_2\},\pi_2)\\ &=\frac{2n_1}{n_1-2}\left[\frac{2}{2n_1}\sum_{i=1}^{2}\sum_{j=1}^{n_1}|a_i-x_j^1|^{\alpha}-\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i^1-x_j^1|^{\alpha}-\frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}|a_i-a_j|^{\alpha}\right]\\ &\quad-\frac{2n_2}{n_2+2}\left[\frac{2}{2n_2}\sum_{i=1}^{2}\sum_{j=1}^{n_2}|a_i-x_j^2|^{\alpha}-\frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}|x_i^2-x_j^2|^{\alpha}-\frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}|a_i-a_j|^{\alpha}\right]\\ &=\frac{2}{n_1-2}\sum_{i=1}^{2}\sum_{j=1}^{n_1}|a_i-x_j^1|^{\alpha}-\frac{2}{n_1(n_1-2)}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i^1-x_j^1|^{\alpha}\\ &\quad-\frac{2}{n_2+2}\sum_{i=1}^{2}\sum_{j=1}^{n_2}|a_i-x_j^2|^{\alpha}+\frac{2}{n_2(n_2+2)}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}|x_i^2-x_j^2|^{\alpha}\\ &\quad+\frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}|a_i-a_j|^{\alpha}\left(\frac{n_2}{n_2+2}-\frac{n_1}{n_1-2}\right)\\ &=\left[\frac{2}{n_1-2}\sum_{j=1}^{n_1}|a_1-x_j^1|^{\alpha}-\frac{1}{n_1(n_1-2)}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i^1-x_j^1|^{\alpha}\right]\\ &\quad-\left[\frac{2}{n_2+2}\sum_{j=1}^{n_2}|a_1-x_j^2|^{\alpha}-\frac{1}{n_2(n_2+2)}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}|x_i^2-x_j^2|^{\alpha}\right]\\ &\quad+\left[\frac{2}{n_1-2}\sum_{j=1}^{n_1}|a_2-x_j^1|^{\alpha}-\frac{1}{n_1(n_1-2)}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i^1-x_j^1|^{\alpha}\right]\\ &\quad-\left[\frac{2}{n_2+2}\sum_{j=1}^{n_2}|a_2-x_j^2|^{\alpha}-\frac{1}{n_2(n_2+2)}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}|x_i^2-x_j^2|^{\alpha}\right]\\ &\quad-\frac{n_1+n_2}{(n_1-2)(n_2+2)}\sum_{i=1}^{2}\sum_{j=1}^{2}|a_i-a_j|^{\alpha}. \end{aligned} \tag{4.16}$$

Then according to equations (4.10), (4.11), (4.12), and (4.13), equation (4.16) can be written as

$$\begin{aligned}
&\frac{2n_1}{n_1-2}\xi^\alpha(\{a_1,a_2\},\pi_1)-\frac{2n_2}{n_2+2}\xi^\alpha(\{a_1,a_2\},\pi_2)\\
=&\frac{n_1-1}{n_1-2}\frac{n_1}{n_1-1}\xi^\alpha(a_1,\pi_1)-\frac{n_2+1}{n_2+2}\frac{n_2}{n_2+1}\xi^\alpha(a_1,\pi_2)+\\
&\frac{n_1-1}{n_1-2}\frac{n_1}{n_1-1}\xi^\alpha(a_2,\pi_1)-\frac{n_2+1}{n_2+2}\frac{n_2}{n_2+1}\xi^\alpha(a_2,\pi_2)-\\
&\frac{n_1+n_2}{(n_1-2)(n_2+2)}\sum_{i=1}^{2}\sum_{j=1}^{2}|a_i-a_j|^\alpha. \qquad (4.17)
\end{aligned}$$

□

Based on the updating formula of K-groups by first variation algorithm 3.10, we assign point a_1 to π_2 if

$$\frac{n_1}{n_1-1}\xi^\alpha(a_1,\pi_1)-\frac{n_2}{n_2+1}\xi^\alpha(a_1,\pi_2),$$

is positive; otherwise we keep point a_1 in cluster π_1. Since $\frac{n_1-1}{n_1-2}>1$, $\frac{n_2+1}{n_2+2}<1$, $\xi^\alpha(a_1,\pi_1)>0$, and $\xi^\alpha(a_1,\pi_2)>0$, it follows that

$$\frac{n_1}{n_1-2}\xi^\alpha(a_1,\pi_1)-\frac{n_2}{n_2+2}\xi^\alpha(a_1,\pi_2)>\frac{n_1}{n_1-1}\xi^\alpha(a_1,\pi_1)-\frac{n_2}{n_2+1}\xi^\alpha(a_1,\pi_2).$$

Suppose we need to assign points a_1 and a_2 to cluster π_2. Then

$$\frac{n_1}{n_1-2}\xi^\alpha(a_1,\pi_1)-\frac{n_2}{n_2+2}\xi^\alpha(a_1,\pi_2)+\frac{n_1}{n_1-2}\xi^\alpha(a_2,\pi_1)-\frac{n_2}{n_2+2}\xi^\alpha(a_2,\pi_2)>0.$$

Since $-\frac{n_1+n_2}{(n_1-2)(n_2+2)}\sum_{i=1}^{2}\sum_{j=1}^{2}|a_i-a_j|^\alpha$ is negative, we can not determine the sign of

$$\frac{2n_1}{n_1-2}\xi^\alpha(\{a_1,a_2\},\pi_1)-\frac{2n_2}{n_2+2}\xi^\alpha(\{a_1,a_2\},\pi_2). \qquad (4.18)$$

If $n_1,n_2\gg 1$, then we have $\frac{n_1-1}{n_1-2}\approx 1$, $\frac{n_2+1}{n_2+2}\approx 1$. Assume that we need to keep points a_1 and a_2 in the same cluster. Then based on formula (4.17) we should keep the pair $\{a_1,a_2\}$ in cluster π_1. The other situations will be harder to determine the sign of equation (4.9). Figure 4.4 and Table 4.1

Table 4.1: Compare Moving One Point with Moving Pair

	Energy Distance to π_1	Energy Distance to π_2
a_1	0.548	0.554
a_2	0.498	0.314
a_3	1.400	1.935
$\{a_2, a_3\}$	0.776	1.116
$\{a_1, a_2\}$	0.747	0.665

will illustrate those situations. If a_1, a_2 and a_3 are points in π_2, there are several possible results for individual moves and pair moves. Based on the Table 4.1, point a_1 should be assigned to π_1, point a_2 should be assigned to π_2, and point a_3 should be assigned to π_1. However, the pair $\{a_1, a_2\}$ should be assigned to π_2 and pair $\{a_2, a_3\}$ should be assigned to π_1. We can see that moving pairs of points is not equivalent to two moves of individual point.

4.2 m^{th} variation

Now we want to generalize to m^{th} variation where m is any integer. 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

$$\begin{aligned}\xi^{\alpha}(\{a_1, ...a_m\}, \pi_1) =& \frac{2}{m \cdot n_1} \sum_{i}^{n_1} \sum_{j}^{m} |x_i^1 - a_j|^{\alpha} - \frac{1}{m^2} \sum_{i}^{m} \sum_{j}^{m} |a_i - a_j|^{\alpha} \\ & - \frac{1}{n_1^2} \sum_{i}^{n_1} \sum_{j}^{n_1} |x_i^1 - x_j^1|^{\alpha}, \end{aligned} \tag{4.19}$$

and

$$\xi^{\alpha}(\{a_1,...a_m\},\pi_2) = \frac{2}{m\cdot n_2}\sum_i^{n_2}\sum_j^{m}|x_i^2-a_j|^{\alpha} - \frac{1}{m^2}\sum_i^{m}\sum_j^{m}|a_i-a_j|^{\alpha} - \frac{1}{n_2^2}\sum_i^{n_2}\sum_j^{n_2}|x_i^2-x_j^2|^{\alpha}. \tag{4.20}$$

Similar to the derivation of first variation and second 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

$$\begin{aligned}
&\frac{n_1}{2}G^{\alpha}(\pi_1,\pi_1) - \frac{n_1-m}{2}G^{\alpha}(\pi_1^-,\pi_1^-)\\
&\quad= \frac{1}{2\cdot n_1}\sum_i^{n_1}\sum_j^{n_1}|x_i^1-x_j^1|^{\alpha} - \frac{1}{2\cdot(n_1-m)}\sum_i^{n_1-m}\sum_j^{n_1-m}|x_i^{-1}-x_j^{-1}|^{\alpha}\\
&\quad= \frac{1}{2\cdot n_1}\sum_i^{n_1}\sum_j^{n_1}|x_i^1-x_j^1|^{\alpha} - \frac{1}{2\cdot(n_1-m)}\left[\sum_i^{n_1}\sum_j^{n_1}|x_i^1-x_j^1|^{\alpha}\right.\\
&\qquad\left. -2\sum_i^{n_1}\sum_j^{m}|x_i^1-a_j|^{\alpha} + \sum_i^{m}\sum_j^{m}|a_i-a_j|^{\alpha}\right]\\
&\quad= \frac{1}{n_1-m}\sum_i^{m}\sum_j^{n_1}|x_i^1-a_j|^{\alpha} - \frac{1}{2\cdot(n_1-m)}\sum_i^{m}\sum_j^{m}|a_i-a_j|^{\alpha}\\
&\qquad - \frac{m}{2\cdot n_1(n_1-m)}\sum_i^{n_1}\sum_j^{n_1}|x_i^1-x_j^1|^{\alpha}.
\end{aligned} \tag{4.21}$$

Multiply $\frac{m\cdot n_1}{2(n_1-m)}$ times equation(4.19) to get

$$\begin{aligned}
\frac{mn_1}{2(n_1-m)}\xi^{\alpha}(\{a_1,...a_m\},\pi_1) =& \frac{1}{n_1-m}\sum_i^{n_1}\sum_j^{m}|x_i^1-a_j|^{\alpha} -\\
& \frac{n_1}{2m(n_1-m)}\sum_i^{m}\sum_j^{m}|a_i-a_j|^{\alpha} -\\
& \frac{m}{2n_1(n_1-m)}\sum_i^{n_1}\sum_j^{n_1}|x_i^1-x_j^1|^{\alpha}.
\end{aligned} \tag{4.22}$$

Subtract equation (4.22) from equation (4.21) to obtain

$$\frac{n_1}{2}G^\alpha(\pi_1,\pi_1)-\frac{n_1-m}{2}G^\alpha(\pi_1^-,\pi_1^-) \\ =\frac{mn_1}{2(n_1-m)}\xi^\alpha(\{a_1,...a_m\},\pi_1)+\frac{1}{2m}\sum_i^m\sum_j^m|a_i-a_j|^\alpha. \tag{4.23}$$

Based on a similar derivation, we can show that

$$\frac{n_2+m}{2}G^\alpha(\pi_2^+,\pi_2^+)-\frac{n_2}{2}G^\alpha(\pi_2,\pi_2) \\ =\frac{mn_2}{2(n_2+m)}\xi^\alpha(\{a_1,...a_m\},\pi_2)+\frac{1}{2m}\sum_i^m\sum_j^m|a_i-a_j|^\alpha. \tag{4.24}$$

Subtract equation (4.24) from equation (4.23) to get

$$\frac{n_1}{2}G^\alpha(\pi_1,\pi_1)+\frac{n_2}{2}G^\alpha(\pi_2,\pi_2)-\frac{n_1-m}{2}G^\alpha(\pi_1^-,\pi_1^-)-\frac{n_2+m}{2}G^\alpha(\pi_2^+,\pi_2^+) \\ =\frac{mn_1}{2(n_1-m)}\xi^\alpha(\{a_1,...a_m\},\pi_1)-\frac{mn_2}{2(n_2+m)}\xi^\alpha(\{a_1,...a_m\},\pi_2). \tag{4.25}$$

Theorem 4.3. *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.

Proposition 4.4. *Let $A=\{a_1,a_2,...,a_m\}$ be m different observations of cluster π_1, cluster π_2 be a member of partition $P=\{\pi_1,\pi_2,...\pi_k\}$ and $\pi_1\neq\pi_2$. Let n_1 and n_2 be the sizes of clusters π_1*

and π_2, set S_1 be a nonempty subset of A with size m_1 $(1 < m_1 < n_1)$, and the complementary set $S_2 = A \backslash S_1$ with size $m_2 = m - m_1$. Let a_i^1 $(i = 1, ..., m_1)$ belong to S_1, and a_j^2 $(j = 1, ..., m_2)$ belong to S_2. Then

$$\begin{aligned}
&\frac{mn_1}{n_1 - m}\xi^\alpha(A, \pi_1) - \frac{mn_2}{n_2 + m}\xi^\alpha(A, \pi_2) \\
&\quad = \frac{n_1 - m_1}{n_1 - m}\frac{n_1 m_1}{n_1 - m_1}\xi^\alpha(S1, \pi_1) - \frac{n_2 + m_1}{n_2 + m}\frac{n_2 m_1}{n_2 + m_1}\xi^\alpha(S1, \pi_2) \\
&\quad\quad + \frac{n_1 - m_2}{n_1 - m}\frac{n_1 m_2}{n_1 - m_2}\xi^\alpha(S2, \pi_1) - \frac{n_2 + m_2}{n_2 + m}\frac{n_2 m_2}{n_2 + m_2}\xi^\alpha(S2, \pi_2) \\
&\quad\quad - \frac{m(n_1 + n_2)}{(n_1 - m)(n_2 + m)}\left[T^\alpha(S1, S2) - W^\alpha(S1, S2)\right].
\end{aligned}$$

Proof. The two-sample energy statistics between sets S_1, S_2 and clusters π_1, π_2 are

$$\begin{aligned}
\xi^\alpha(S_1, \pi_1) =& \frac{2}{n_1 m_1}\sum_{i=1}^{m_1}\sum_{j=1}^{n_1}|a_i^1 - x_j^1|^\alpha - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i^1 - x_j^1|^\alpha \\
&- \frac{1}{m_1^2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_1}|a_i^1 - a_j^1|^\alpha,
\end{aligned} \tag{4.26}$$

$$\begin{aligned}
\xi^\alpha(S_1, \pi_2) =& \frac{2}{n_2 m_1}\sum_{i=1}^{m_1}\sum_{j=1}^{n_2}|a_i^1 - x_j^2|^\alpha - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}|x_i^2 - x_j^2|^\alpha \\
&- \frac{1}{m_2^2}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}|a_i^1 - a_j^1|^\alpha,
\end{aligned} \tag{4.27}$$

$$\begin{aligned}
\xi^\alpha(S_2, \pi_1) =& \frac{2}{n_1 m_2}\sum_{i=1}^{m_2}\sum_{j=1}^{n_1}|a_i^2 - x_j^1|^\alpha - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i^1 - x_j^1|^\alpha \\
&- \frac{1}{m_2^2}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}|a_i^2 - a_j^2|^\alpha,
\end{aligned} \tag{4.28}$$

$$\xi^{\alpha}(S_2,\pi_2) = \frac{2}{n_2 m_2}\sum_{i=1}^{m_2}\sum_{j=1}^{n_2}|a_i^2 - x_j^2|^{\alpha} - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}|x_i^2 - x_j^2|^{\alpha} - \frac{1}{m_2^2}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}|a_i^2 - a_j^2|^{\alpha}. \tag{4.29}$$

The two-sample energy statistics between set A and clusters π_1, π_2 are

$$\xi^{\alpha}(A,\pi_1) = \frac{2}{m n_1}\sum_{i=1}^{m}\sum_{j=1}^{n_1}|a_i - x_j^1|^{\alpha} - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i^1 - x_j^1|^{\alpha} - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|a_i - a_j|^{\alpha}, \tag{4.30}$$

$$\xi^{\alpha}(A,\pi_2) = \frac{2}{m n_2}\sum_{i=1}^{m}\sum_{j=1}^{n_2}|a_i - x_j^2|^{\alpha} - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}|x_i^2 - x_j^2|^{\alpha} - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|a_i - a_j|^{\alpha}. \tag{4.31}$$

Subtract $\frac{mn_2}{n_2+m}\xi^{\alpha}(A,\pi_2)$ from $\frac{mn_1}{n_1-m}\xi^{\alpha}(A,\pi_1)$ to get

$$\begin{aligned}
&\frac{mn_1}{n_1-m}\xi^{\alpha}(A,\pi_1) - \frac{mn_2}{m+n_2}\xi^{\alpha}(A,\pi_2)\\
&\quad= \frac{mn_1}{n_1-m}\left[\frac{2}{mn_1}\sum_{i=1}^{m}\sum_{j=1}^{n_1}|a_i - x_j^1|^{\alpha} - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i^1 - x_j^1|^{\alpha} - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|a_i - a_j|^{\alpha}\right]\\
&\qquad - \frac{mn_2}{n_2+m}\left[\frac{2}{mn_2}\sum_{i=1}^{m}\sum_{j=1}^{n_2}|a_i - x_j^2|^{\alpha} - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}|x_i^2 - x_j^2|^{\alpha} - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|a_i - a_j|^{\alpha}\right]\\
&\quad= \frac{2}{n_1-m}\left[\sum_{i=1}^{m_1}\sum_{j=1}^{n_1}|a_i^1 - x_j^1|^{\alpha} + \sum_{i=1}^{m_2}\sum_{j=1}^{n_1}|a_i^2 - x_j^1|^{\alpha}\right]\\
&\qquad - \left[\frac{m_1}{n_1(n_1-m)}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i^1 - x_j^1|^{\alpha} + \frac{m_2}{n_1(n_1-m)}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i^1 - x_j^1|^{\alpha}\right]
\end{aligned}$$

$$
\begin{aligned}
&- \left[\frac{n_1}{m_1(n_1-m)} \sum_{i=1}^{m_1}\sum_{j=1}^{m_1} |a_i^1 - a_j^1|^\alpha + \frac{n_1}{m_2(n_1-m)} \sum_{i=1}^{m_2}\sum_{j=1}^{m_2} |a_i^2 - a_j^2|^\alpha \right] \\
&- \frac{2}{n_2+m} \left[\sum_{i=1}^{m_1}\sum_{j=1}^{n_2} |a_i^1 - x_j^2|^\alpha + \sum_{i=1}^{m_2}\sum_{j=1}^{n_2} |a_i^2 - x_j^2|^\alpha \right] \\
&+ \left[\frac{m_1}{n_2(n_2+m)} \sum_{i=1}^{n_2}\sum_{j=1}^{n_2} |x_i^2 - x_j^2|^\alpha + \frac{m_2}{n_2(n_2+m)} \sum_{i=1}^{n_2}\sum_{j=1}^{n_2} |x_i^2 - x_j^2|^\alpha \right] \\
&+ \left[\frac{n_2}{m_1(n_2+m)} \sum_{i=1}^{m_1}\sum_{j=1}^{m_1} |a_i^1 - a_j^1|^\alpha + \frac{n_2}{m_2(n_2+m)} \sum_{i=1}^{m_2}\sum_{j=1}^{m_2} |a_i^2 - a_j^2|^\alpha \right] \\
&- \frac{n_1}{n_1-m} \left[\frac{1}{m} \sum_{i=1}^{m}\sum_{j=1}^{m} |a_i - a_j|^\alpha - \frac{1}{m_1} \sum_{i=1}^{m_1}\sum_{j=1}^{m_1} |a_i^1 - a_j^1|^\alpha - \frac{1}{m_2} \sum_{i=1}^{m_2}\sum_{j=1}^{m_2} |a_i^2 - a_j^2|^\alpha \right] \\
&+ \frac{n_2}{n_2+m} \left[\frac{1}{m} \sum_{i=1}^{m}\sum_{j=1}^{m} |a_i - a_j|^\alpha - \frac{1}{m_1} \sum_{i=1}^{m_1}\sum_{j=1}^{m_1} |a_i^1 - a_j^1|^\alpha - \frac{1}{m_2} \sum_{i=1}^{m_2}\sum_{j=1}^{m_2} |a_i^2 - a_j^2|^\alpha \right] \\
= &\left[\frac{2}{n_1-m} \sum_{i=1}^{m_1}\sum_{j=1}^{n_1} |a_i^1 - x_j^1|^\alpha - \frac{m_1}{n_1(n_1-m)} \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} |x_i^1 - x_j^1|^\alpha \right. \\
&\left. - \frac{n_1}{m_1(n_1-m)} \sum_{i=1}^{m_1}\sum_{j=1}^{m_1} |a_i^1 - a_j^1|^\alpha \right] \\
&- \left[\frac{2}{n_2+m} \sum_{i=1}^{m_1}\sum_{j=1}^{n_2} |a_i^1 - x_j^2|^\alpha - \frac{m_1}{n_2(n_2+m)} \sum_{i=1}^{n_2}\sum_{j=1}^{n_2} |x_i^2 - x_j^2|^\alpha \right. \\
&\left. - \frac{n_2}{m_1(n_2+m)} \sum_{i=1}^{m_1}\sum_{j=1}^{m_1} |a_i^1 - a_j^1|^\alpha \right] \\
&+ \left[\frac{2}{n_1-m} \sum_{i=1}^{m_2}\sum_{j=1}^{n_1} |a_i^2 - x_j^1|^\alpha - \frac{m_2}{n_1(n_1-m)} \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} |x_i^1 - x_j^1|^\alpha \right. \\
&\left. - \frac{n_1}{m_2(n_1-m)} \sum_{i=1}^{m_2}\sum_{j=1}^{m_2} |a_i^2 - a_j^2|^\alpha \right] \\
&- \left[\frac{2}{n_2+m} \sum_{i=1}^{m_2}\sum_{j=1}^{n_2} |a_i^2 - x_j^2|^\alpha - \frac{m_2}{n_2(n_2+m)} \sum_{i=1}^{n_2}\sum_{j=1}^{n_2} |x_i^2 - x_j^2|^\alpha \right. \\
&\left. - \frac{n_2}{m_2(n_2+m)} \sum_{i=1}^{m_2}\sum_{j=1}^{m_2} |a_i^2 - a_j^2|^\alpha \right]
\end{aligned}
$$

$$- \frac{m(n_1+n_2)}{(n_1-m)(n_2+m)} \left[\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{m} |a_i - a_j|^\alpha - \frac{1}{m_1} \sum_{i=1}^{m_1} \sum_{j=1}^{m_1} |a_i^1 - a_j^1|^\alpha \right.$$
$$\left. - \frac{1}{m_2} \sum_{i=1}^{m_2} \sum_{j=1}^{m_2} |a_i^2 - a_j^2|^\alpha \right] \tag{4.32}$$

According to equation (4.26), (4.27), (4.28), and (4.29), equation (4.32) can be written as

$$\begin{aligned} &\frac{mn_1}{n_1-m}\xi^\alpha(A,\pi_1) - \frac{mn_2}{n_2+m}\xi^\alpha(A,\pi_2) \\ &\quad = \frac{n_1-m_1}{n_1-m}\frac{n_1m_1}{n_1-m_1}\xi^\alpha(S1,\pi_1) - \frac{n_2+m_1}{n_2+m}\frac{n_2m_1}{n_2+m_1}\xi^\alpha(S1,\pi_2) \\ &\qquad + \frac{n_1-m_2}{n_1-m}\frac{n_1m_2}{n_1-m_2}\xi^\alpha(S2,\pi_1) - \frac{n_2+m_2}{n_2+m}\frac{n_2m_2}{n_2+m_2}\xi^\alpha(S2,\pi_2) \\ &\qquad - \frac{m(n_1+n_2)}{(n_1-m)(n_2+m)}\left[T^\alpha(S1,S2) - W^\alpha(S1,S2)\right] \end{aligned} \tag{4.33}$$

□

Based on Section 2.4.2, we have

$$T^\alpha(S1,S2) = W^\alpha(S1,S2) + B^\alpha(S1,S2).$$

Both $W^\alpha(S1,S2)$ and $B^\alpha(S1,S2)$ are nonnegative, so the last part of equation (4.33)

$$\frac{m(n_1+n_2)}{(n_1-m)(n_2+m)}\left[T^\alpha(S1,S2) - W^\alpha(S1,S2)\right]$$

is positive. Since $\frac{n_1-m_1}{n_1-m} > 1$ and $\frac{n_2+m_2}{n_2+m} < 1$, thus

$$\frac{n_1m_1}{n_1-m}\xi^\alpha(S_1,\pi_1) - \frac{n_2m_1}{n_2+m}\xi^\alpha(S_1,\pi_2) > \frac{n_1m_1}{n_1-m_1}\xi^\alpha(S_1,\pi_1) - \frac{n_2m_1}{n_2+m_1}\xi^\alpha(S_1,\pi_2).$$

Similar to the result of Proposition 4.2, one cannot determine the sign of

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

Thus moving m points is not equivalent to moving m_1 and m_2 points respectively, where $m_1 + m_2 = m$.

4.3 K-groups Algorithm by Second Variation

In this section, we propose a new algorithm based on the updating formula of second variation. According to Figure 4.1(b), the energy distance between two clusters will tend to infinity stochastically as m increases only if two clusters are generated from different distributions; the energy distance is more sensitive to the difference between distributions if we move more points. 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. For second variation, 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. 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

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

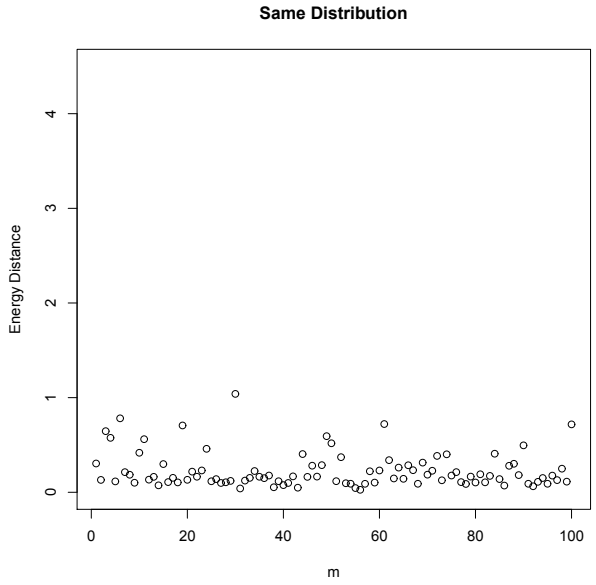
and the minimum

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

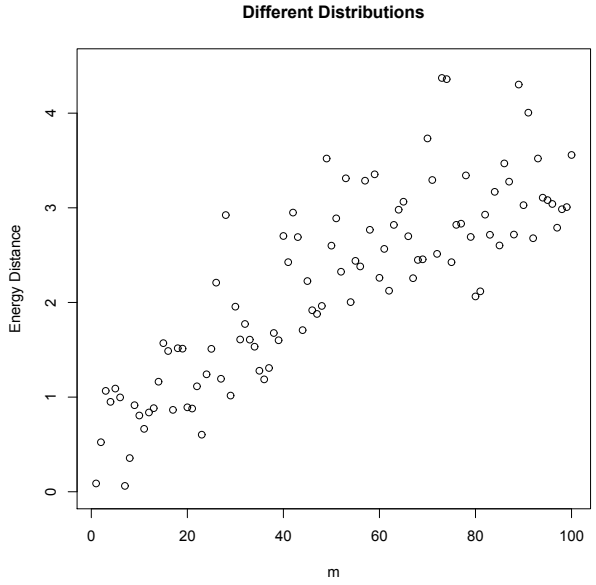
for all clusters π_i, where $\pi_i \neq \pi(II)$. If E_1 is less than E_2, pair II remains in cluster $\pi(II)$; otherwise, move the pair II to cluster π_i with minimum value of E_2, and update the cluster $\pi(II)$ and π_i.

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.

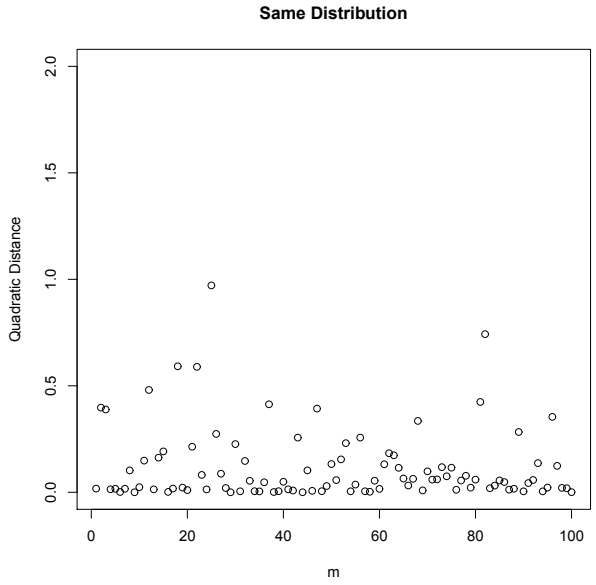


(a)

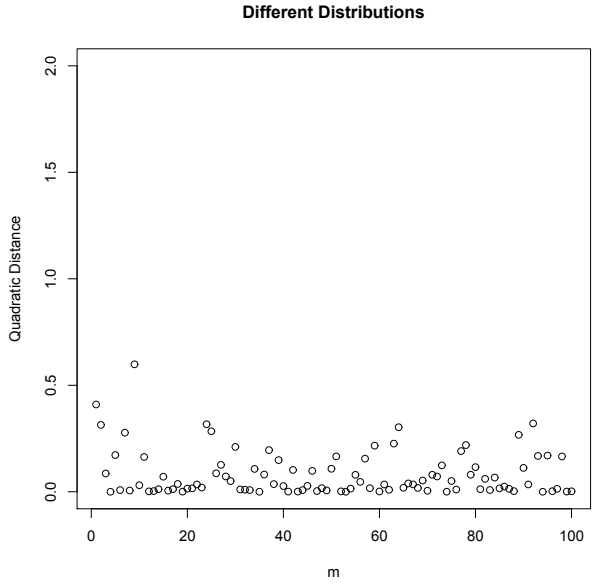


(b)

Figure 4.1: Graph (a) represents the energy distance if both random samples are generated from the same statistical distribution $U(0,1)$. Graph (b) represents the energy distance if two random samples are generated from the different statistical distributions $U(0,1)$ and $U(0.3,0.7)$.

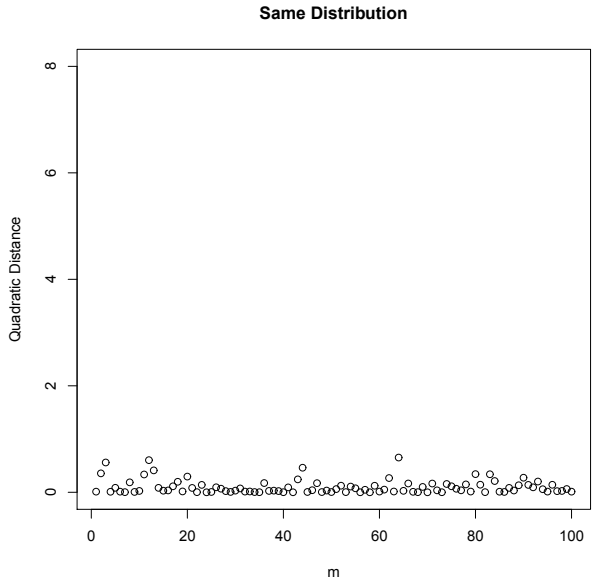


(a)

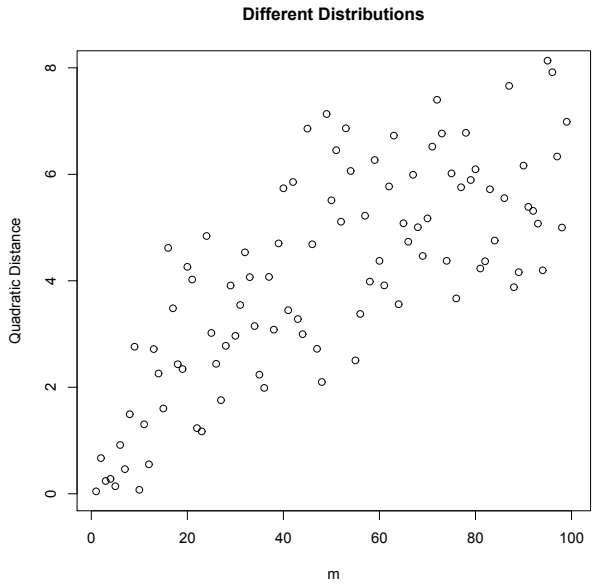


(b)

Figure 4.2: Graph (a) represents the quadratic distance if both random samples are generated from the same statistical distribution $U(0,1)$. Graph (b) represents the quadratic distance if two random samples are generated from different statistical distributions $U(0,1)$ and $U(0.3,0.7)$.



(a)



(b)

Figure 4.3: Graph (a) represents the quadratic distance if both random samples are generated from the same statistical distribution $U(0,1)$. Graph (b) represents the quadratic distance if two random samples are generated from different statistical distributions $U(0,1)$ and $U(0.3,1.3)$.

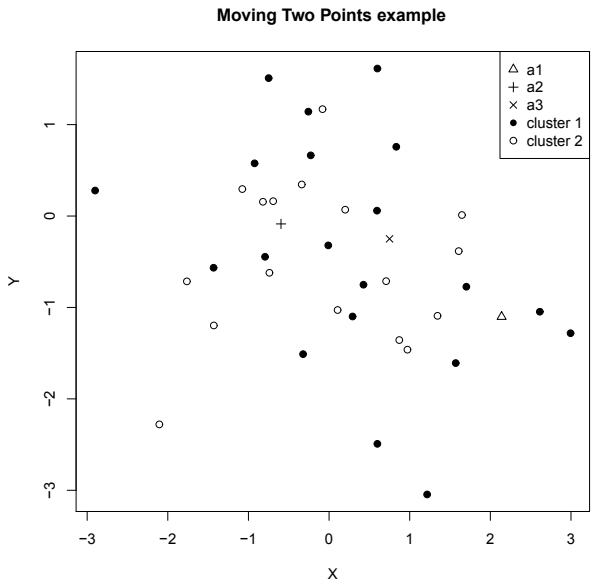


Figure 4.4: Cluster 1 (black points) are generated from a multi-normal distribution with mean $(0, 0)$ and diagonal covariance matrix with diagonal $(2, 2)$. Cluster 2 (white points) are generated from a multi-normal distribution with mean $(0, 0)$ and diagonal covariance matrix with diagonal $(1, 1)$. Points a_1, a_2 and a_3 are three different observations from cluster 2.

CHAPTER 5

UNIVARIATE SIMULATION STUDY

From the previous K-means research, we know that K-means is the most widely applied clustering algorithm for the non-model based unsupervised clustering problem when the number of clusters is prespecified. The advantages of K-means include the following:

- K-means performs well when the data is normally distributed.
- K-means performs very well when the clusters are well separated from each other.
- K-means has a fast algorithm and it is easy to understand.

However, K-means also has several disadvantages.

- K-means performs poorly when the data are skewed.
- K-means cannot handle categorical data, because mean is not a good estimate of center for artificial coding of nominal data.
- K-means performs poorly when clusters are overlapping.
- K-means does not perform well when data has white noise and outliers.
- K-means is not valid for data with infinite first moment.
- K-means cannot be applied when dimension exceeds sample size.

As we introduced in Chapter 3, K-groups is a distribution-based algorithm which assigns observations to the same cluster if they follow the identical statistical distribution. This kind of algorithm

will perform better when clusters are irregular, intertwined, or when noise and outliers are present. In this chapter, we compare K-groups and K-means on different simulated data sets.

5.1 Simulation Design

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 $n = 100, 200, 400, \text{ and } 800$, and for each size n, we calculate average and standard error for validation indices diagonal (Diag), Kappa, Rand, and corrected Rand (cRand) based on $B = 1000$ 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 $E|X - Y|^\alpha$. Table 5.1 displays the mixture distributions for balanced cluster sizes, Table 5.2 summarises the mixture distributions for unbalanced cluster sizes, and Table 5.3 lists the mixture distributions for different values of α. 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* (Li and Rizzo, 2015).

5.1.1 Symmetric Distribution

The standard normal is the most common symmetric distribution, and it is known that K-means performs well on well-separated clusters which are normally distributed. We want to know how K-groups performs in this situation.

The normal mixtures with mixing parameter p are simulated as follows:

1. Generate U from a uniform $(0, 1)$ distribution.

2. If $U < p$, generate X from a normal(μ_1, σ_1^2) distribution; otherwise generate X from a normal(μ_2, σ_2^2) distribution. The normal mixture distributions will be denoted by $p\,\mathrm{N}\,(\mu_1, \sigma_1^2) + (1 - p)\,\mathrm{N}\,(\mu_2, \sigma_2^2)$.

Student's T distribution is a well-known symmetric distribution with heavy tails. The Student's

T mixtures with mixing parameter p are simulated as follows:

1. Generate U from a uniform $(0, 1)$ distribution.

2. If $U < p$, generate X from T(v) distribution; otherwise generate X from a T$(v) + \mu$ distribution. Here μ is the location parameter, and v is the degrees of freedom. The Student's T mixture distributions will be denoted by $p\,\mathrm{T}\,(v) + (1-p)(\,\mathrm{T}\,(v) + \mu)$.

The standard logistic distribution is another symmetric distribution with heavy tails. We use logistic distributions to simulate clusters with equal means but different variances. The logistic mixtures with mixing parameter p are simulated as follows:

1. Generate U from a uniform $(0, 1)$ distribution.

2. If $U < p$, generate X from logistic(μ_1, s_1) distribution; otherwise generate X from a logistic(μ_2, s_2) distribution. Here μ is a location parameter, and s is a scale parameter. The logistic mixture distributions will be denoted by $p\,\mathrm{Logistic}(\mu_1, s_1) + (1-p)\,\mathrm{Logistic}\,(\mu_2, s_2)$.

We have seen that K-means uses the cluster mean as a prototype. However, for some distributions, the first moment does not exist. We want to compare how K-means and K-groups perform in this situation. The standard Cauchy distribution is a well known example of a symmetric distribution with infinite first moment. Cauchy mixtures with mixing parameter p are simulated as follows:

1. Generate U from a uniform $0, 1$ distribution.

2. If $U < p$, generate X from Cauchy(θ_1, γ_1) distribution; otherwise generate X from a Cauchy(θ_2, γ_2) distribution. Here θ is the location parameter, and γ is the scale parameter. The Cauchy mixture distributions will be denoted by $p\,\mathrm{Cauchy}(\theta_1, \gamma_1) + (1-p)\,\mathrm{Cauchy}\,(\theta_2, \gamma_2)$.

5.1.2 Skewed Distribution

We simulated skewed mixture distributions to compare the performance of K-means, K-groups by first variation, and K-groups by second variation. The simulation design is the same as for symmetric distributions.

A Weibull distribution with appropriate parameters has mild skewness and exponential tails. The Weibull mixtures with mixing parameter p are simulated as follows:

1. Generate U from a uniform $(0, 1)$ distribution.

2. If $U < p$, generate X from Weibull(λ_1, k_1) distribution; otherwise generate X from a Weibull$(\lambda_2, k_2) + \mu$ distribution. Here λ is the scale parameter, k is the shape parameter, and μ is the location parameter. The Weibull mixture distributions will be denoted by $p\,\mathrm{Weibull}\,(\lambda_1, k_1) + (1-p)(\,\mathrm{Weibull}\,(\lambda_2, k_2) + \mu)$.

A Beta distribution with appropriate parameters has moderate skewness. The Beta mixtures with mixing parameter p are simulated as follows:

1. Generate U from a uniform $(0, 1)$ distributions.

2. If $U < p$, generate X from Beta(a_1, b_1) distribution; otherwise generate X from a Beta$(a_2, b_2)+\mu$ distribution. Here μ is the location parameter. The Beta mixture distributions will be denoted by $p\,\mathrm{Beta}\,(a_1, b_1) + (1-p)(\,\mathrm{Beta}\,(a_2, b_2) + \mu)$.

A chi-square distribution with appropriate parameter has large skewness and heavy tails. The chi-square mixtures with mixing parameter p are simulated as follows:

1. Generate U from a uniform $(0, 1)$ distribution.

2. If $U < p$, generate X from $\chi^2(k)$ distribution; otherwise generate X from a $\chi^2(k) + \mu$ distribution. Here k is the degrees of freedom, and μ is the location parameter. The chi-square mixture distributions will be denoted by $p\,\chi^2(k) + (1-p)(\,\chi^2(k) + \mu)$.

The lognormal distribution is another distribution that has large skewness and heavy tails. The

lognormal mixtures with mixing parameter p are simulated as follows:

1. Generate U from a uniform $(0, 1)$ distribbution.

2. If $U < p$, generate X from lognormal(μ_1, σ_1^2) distribution; otherwise generate X from a lognormal(μ_2, σ_2^2) distribution. Here μ is the mean, and σ is the standard deviation. The lognormal mixture distributions will be denoted by $p\,\mathrm{Lognormal}\,(\mu_1, \sigma_1^2) + (1-p)\,\mathrm{Lognormal}\,(\mu_2, \sigma_2^2)$.

5.1.3 Unbalanced Clusters

We use two mixture distributions, normal mixtures, and lognormal mixtures to compare how both K-groups algorithms and K-means perform when clusters have unbalanced sizes. Each mixture distribution is simulated with sample size $n = 200$ with different mixing parameters $p = 0.1, 0.2, ..., 0.9$. The average validation indices are computed based on $B = 1000$ iterations. Table 5.2 displays all the mixture distributions we use for unbalanced clusters.

5.1.4 Alpha Effect

We use two mixture distributions, normal mixtures, and Cauchy mixtures to compare how both K-groups algorithms and K-means perform with different values of α. Each mixture distribution is simulated with sample size 200 and result compared at different $\alpha = 0.2, 0.4, ..., 2$. The average validation indices are computed based on $B = 1000$ iterations. Table 5.3 shows all the mixture distributions we use for α effect.

5.2 Simulation Result

This section presents the empirical results comparing the validation indices for K-means, K-groups by first variation, and K-groups by second variation in the univariate cases. For each iteration validation measures are computed for each method. The average and standard error of validation indices are reported for each method over the $B = 1000$ samples.

Normal mixture distribution results

Table 5.4, Table 5.5, Table 5.6, and Table 5.7 summarize the simulation results for the normal mixtures $0.5\,\mathrm{N}\,(0,1)+0.5\,\mathrm{N}\,(3,1), 0.5\,\mathrm{N}\,(0,1)+0.5\,\mathrm{N}\,(2,1), 0.5\,\mathrm{N}\,(0,1)+0.5\,\mathrm{N}\,(1,1)$, and $0.5\,\mathrm{N}\,(0,1)+0.5\,\mathrm{N}\,(0,3)$ respectively. Table 5.4 shows the result for the normal mixture distribution in which two clusters are well separated. Based on the the average Rand and cRand indices for different sizes $n = 100, 200, 400, \text{ and } 800$, all three algorithms perform well. When we increase the value of n, the pairwise differences in performance among three algorithms are decreasing. Table 5.5 and Table 5.6 are normal mixture distributions with overlapping clusters. The average Rand and cRand indices of Table 5.5 and Table 5.6 are consistently lower than Table 5.4. The average Rand and cRand indices of the three algorithms are very close for all simulation sizes in Table 5.5 and Table 5.6. Table 5.7 summarizes results for the normal mixture with identical means but different variances. In this case, the average Rand indices of the three algorithms are very close for all sizes n.

Figure 5.1 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.

Student T mixture distribution results

Table 5.8, Table 5.9, and Table 5.10 summarize the simulation results for the Student's T mixtures $0.5\,\mathrm{T}\,(1)+0.5(\,\mathrm{T}\,(4)+3)$, $0.5\,\mathrm{T}\,(4)+0.5(\,\mathrm{T}\,(4)+2)$, and $0.5\,\mathrm{T}\,(1)+0.5(\,\mathrm{T}\,(4)+1)$ respectively. The average Rand and cRand indices in all these three tables are nearly the same for all three algorithms.

Figure 5.2 displays the simulation results for Student T mixtures $0.5\,\mathrm{T}\,(4)+0.5(\,\mathrm{T}\,(4)+d)$, where $d = 0.2, 0.4, ..., 3$. The average cRand indices of these three algorithms are almost the same for each value of d. Thus, the results suggest that both K-groups algorithms and K-means have similar performance when the cluster have slightly heavy tails.

Logistc mixture distribution results

Table 5.11 displays the simulation results of logistic mixture $0.5\,\text{Logistic}\,(0,1) + 0.5\,\text{Logistic}\,(0,4)$. In this situation, two clusters have exactly the same mean. The average Rand and cRand indices of both K-groups algorithms are lower than K-means. Since logistic distribution put large weight on the mean, so it is reasonable that the K-means has better performance in this situation.

Cauchy mixture distribution results

Table 5.12, Table 5.13, and Table 5.14 display the simulation results of Cauchy mixtures $0.5\,\text{Cauchy}\,(0,1)+0.5\,\text{Cauchy}\,(3,1)$, $0.5\,\text{Cauchy}\,(0,1)+0.5\,\text{Cauchy}\,(2,1)$, and $0.5\,\text{Cauchy}\,(0,1)$ $+0.5\,\text{Cauchy}\,(1,1)$. The clusters generated from a Cauchy distribution will have outliers. The average Rand and cRand indices for both K-groups algorithms dominate the average Rand and cRand indices for K-means for all sizes $n = 100, 200, 400, \text{ and } 800$ in these three tables.

Figure 5.3 displays the simulation results for Cauchy mixture $0.5\,\text{Cauchy}(0,1) + 0.5\text{Cauchy}$ $(d,1)$, 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.

Weibull mixture distribution results

Table 5.15, Table 5.16, and Table 5.17 summarize the simulation results for Weibull mixtures $0.5\,\text{Weibull}\,(1,5)+0.5(\,\text{Weibull}\,(1,5)+2)$, $0.5\,\text{Weibull}\,(1,5)+0.5(\,\text{Weibull}\,(1,5)+1)$, and $0.5\,\text{Weibull}\,(1,5) + 0.5(\,\text{Weibull}\,(1,5) + 0.5)$. The average Rand and cRand indices of the three algorithms are very close for all sizes n.

Figure 5.4 shows the simulation results for Weibull mixtures $0.5\,\text{Weibull}(1,5) + 0.5(\text{Weibull}$ $(1,5) + d)$, where $d = 0.2, 0.4, ..., 3$. The average cRand indices of K-groups by first variation and K-means are almost the same for each value of d . The performance of K-groups by second variation is slightly worse than the other two methods. Thus, the results suggest that both K-groups algorithms and K-means perform similarly when clusters are slightly skewed.

Beta mixture distribution results

Table 5.18 and Table 5.19 show the simulation results for Beta mixtures $0.5\,\mathrm{Beta}\,(2,1)+0.5\,\mathrm{Beta}\,((2,1)+2)$, and $0.5\,\mathrm{Beta}\,(2,1)+0.5(\,\mathrm{Beta}\,(2,1)+0.5)$. Based on Table.5.18, K-groups by first variation and K-means have perfect clustering performance with average Rand and cRand indices 1. K-groups by second variation obtains slightly lower average Rand and cRand indices which are around 0.99. In Table 5.19, the average Rand and cRand indices of the three algorithms are very close when $n = 100\,\mathrm{and}\,200$. When $n = 400\,\mathrm{and}\,800$, both K-groups algorithms have slightly better performances than K-means.

Figure 5.5 shows the simulation results for beta mixtures $0.5\,\mathrm{Beta}\,(2,1)+0.5(\,\mathrm{Beta}\,(2,1)+d)$, where $d = 0.2, 0.4, ...3$. The average cRand indices of the three algorithms are almost the same at each value of d. Thus, the results suggest that both K-groups algorithms and K-means have almost the same performance when clusters are moderately skewed.

Chi-square mixture distribution results

Table 5.20, Table 5.21, and Table 5.22 summarize the simulation results for Chi-square mixtures with 10 degrees of freedom $0.5\chi^2(10)+0.5(\chi^2(10)+30), 0.5\chi^2(10)+0.5(\chi^2(10)+10)$, and $0.5\chi^2(10)+0.5(\,\chi^2(10)+5)$. In Table 5.20, The average Rand and cRand indices of the three algorithms are close. In Table 5.21 and Table 5.22 the average Rand and cRand indices of both K-groups algorithms are slightly higher than K-means.

Table 5.23, Table 5.24, and Table 5.25 summarize the simulation results for chi-square mixtures with 1 degrees of freedom $0.5\chi^2(1)+0.5(\chi^2(1)+8)$,$0.5\chi^2(1)+0.5(\chi^2(1)+3)$, and $0.5\chi^2(1)+0.5(\chi^2(1)+1)$. The average Rand and cRand indices of both K-groups algorithms and K-means are very close in Table 5.23 and Table 5.24. The average Rand and cRand indices of both K-groups algorithms are higher than K-means in Table 5.25.

Figure 5.6 displays the simulation results for chi-square mixtures $0.5\chi^2(10)+0.5(\chi^2(10)+d)$, where $d = 0.5, 1, ..., 10$. The average cRand indices of both K-groups algorithms are slightly

higher than K-means when $d < 10$. When $d \geq 10$, the average cRand indices of K-groups by first variation is almost the same as K-means, while K-groups by second variation has slightly lower average cRand indices than the other two algorithms. Thus, the results suggest that both K-groups algorithms and K-means have similar performance when clusters are moderately skewed and heavily tailed.

Figure 5.7 displays the simulation results for chi-square mixtures $0.5\chi^2(1) + 0.5(\chi^2(1) + d)$, where $d = 0.5, 1..., 8$. The average cRand indices of both K-groups algorithms are higher than K-means when $d < 5$. When $d > 5$, the average cRand indices of the three algorithms are almost the same. Thus, the results suggest that both K-groups algorithms have better performance than K-means when clusters are strongly skewed, heavily tailed, and overlapped.

Lognormal mixture distribution results

Table 5.26, Table 5.27, and Table 5.28 summarize simulation results for lognormal mixtures $0.5\,\mathrm{Lognormal}\,(0,1) + 0.5\,\mathrm{Lognormal}\,(10,1)$, $0.5\,\mathrm{Lognormal}\,(0,1) + 0.5\,\mathrm{Lognormal}\,(3,1)$, and $0.5\,\mathrm{Lognormal}\,(0,1) + 0.5\,\mathrm{Lognormal}\,(1,1)$. The average Rand and cRand indices of both K-groups algorithms are consistently higher than K-means for all simulation sizes n.

Figure 5.8 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 performance than K-means when clusters are strongly skewed, heavy tailed.

Unbalanced clusters results

Figure 5.9 shows the results of normal mixtures $p\,\mathrm{N}\,(0,1) + (1-p)\,\mathrm{N}\,(3,1)$ where $p = 0.1, 0.2, ..., 0.9$. All three algorithms reach the maximum average cRand indices at $p = 0.5$, and minimum average cRand indices at $p = 0.1$ and 0.9. Thus, the results suggest that both K-groups algorithms and K-means algorithm have best clustering performance when we have balanced clusters.

Figure 5.10 shows the results of lognormal mixtures $p\,\mathrm{Lognormal}\,(0,1)+(1-p)\,\mathrm{Lognormal}\,(3,1)$ The average cRand indices increase as p increases.

Alpha effect results

Figure 5.11 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 5.12 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 5.1: Univariate Mixture Distributions

Distribution	Skewness	Kurtosis
$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1)$	Symmetric	Light tails
$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(2,1)$	Symmetric	Light tails
$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(1,1)$	Symmetric	Light tails
$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(0,3)$	Symmetric	Light tails
$0.5\,\mathrm{T}\,(4) + 0.5(\,\mathrm{T}\,(4) + 3)$	Symmetric	Heavy tails
$0.5\,\mathrm{T}\,(4) + 0.5(\,\mathrm{T}\,(4) + 2)$	Symmetric	Heavy tails
$0.5\,\mathrm{T}\,(4) + 0.5(\,\mathrm{T}\,(4) + 1)$	Symmetric	Heavy tails
$0.5\,\mathrm{Logistic}\,(0,1) + 0.5\,\mathrm{Logistic}\,(0,4)$	Symmetric	Heavy tails
$0.5\,\mathrm{Cauchy}\,(0,1) + 0.5\,\mathrm{Cauchy}\,(3,1)$	Symmetric	Heavy tails
$0.5\,\mathrm{Cauchy}\,(0,1) + 0.5\,\mathrm{Cauchy}\,(2,1)$	Symmetric	Heavy tails
$0.5\,\mathrm{Cauchy}\,(0,1) + 0.5\,\mathrm{Cauchy}\,(1,1)$	Symmetric	Heavy tails
$0.5\,\mathrm{Weibull}\,(1,5) + 0.5(\,\mathrm{Weibull}\,(1,5) + 2)$	Mild skewness	Nearly normal tails
$0.5\,\mathrm{Weibull}\,(1,5) + 0.5(\,\mathrm{Weibull}\,(1,5) + 1)$	Mild skewness	Nearly normal tails
$0.5\,\mathrm{Weibull}\,(1,5) + 0.5(\,\mathrm{Weibull}\,(1,5) + 0.5)$	Mild skewness	Nearly normal tails
$0.5\,\mathrm{Beta}\,(2,1) + 0.5(\,\mathrm{Beta}\,(2,1) + 2)$	Moderate skewness	Light tails
$0.5\,\mathrm{Beta}\,(2,1) + 0.5(\,\mathrm{Beta}\,(2,1) + 0.5)$	Moderate skewness	Light tails
$0.5\chi^2_{10} + 0.5(\chi^2_{10} + 30)$	Moderate skewness	Heavy tails
$0.5\chi^2_{10} + 0.5(\chi^2_{10} + 10)$	Moderate skewness	Heavy tails
$0.5\chi^2_{10} + 0.5(\chi^2_{10} + 5)$	Moderate skewness	Heavy tails
$0.5\chi^2_1 + 0.5(\chi^2_1 + 8)$	Large skewness	Heavy tails
$0.5\chi^2_1 + 0.5(\chi^2_1 + 3)$	Large skewness	Heavy tails
$0.5\chi^2_1 + 0.5(\chi^2_1 + 1)$	Large skewness	Heavy tails
$0.5\,\mathrm{Lognormal}\,(0,1) + 0.5\,\mathrm{Lognormal}\,(10,1)$	Large skewness	Heavy tails
$0.5\,\mathrm{Lognormal}\,(0,1) + 0.5\,\mathrm{Lognormal}\,(3,1)$	Large skewness	Heavy tails
$0.5\,\mathrm{Lognormal}\,(0,1) + 0.5\,\mathrm{Lognormal}\,(1,1)$	Large skewness	Heavy tails

Table 5.2: Unbalanced Mixture Distributions

p value	Mixture Normal	Mixture Lognormal
0.1	$0.1\,\mathrm{N}\,(0,1) + 0.9\,\mathrm{N}\,(3,1)$	$0.1\,\mathrm{Lognormal}\,(0,1) + 0.9\,\mathrm{Lognormal}\,(3,1)$
0.2	$0.2\,\mathrm{N}\,(0,1) + 0.8\,\mathrm{N}\,(3,1)$	$0.2\,\mathrm{Lognormal}\,(0,1) + 0.8\,\mathrm{Lognormal}\,(3,1)$
0.3	$0.3\,\mathrm{N}\,(0,1) + 0.7\,\mathrm{N}\,(3,1)$	$0.3\,\mathrm{Lognormal}\,(0,1) + 0.7\,\mathrm{Lognormal}\,(3,1)$
0.4	$0.4\,\mathrm{N}\,(0,1) + 0.6\,\mathrm{N}\,(3,1)$	$0.4\,\mathrm{Lognormal}\,(0,1) + 0.6\,\mathrm{Lognormal}\,(3,1)$
0.5	$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1)$	$0.5\,\mathrm{Lognormal}\,(0,1) + 0.5\,\mathrm{Lognormal}\,(3,1)$
0.6	$0.6\,\mathrm{N}\,(0,1) + 0.4\,\mathrm{N}\,(3,1)$	$0.6\,\mathrm{Lognormal}\,(0,1) + 0.4\,\mathrm{Lognormal}\,(3,1)$
0.7	$0.7\,\mathrm{N}\,(0,1) + 0.3\,\mathrm{N}\,(3,1)$	$0.7\,\mathrm{Lognormal}\,(0,1) + 0.3\,\mathrm{Lognormal}\,(3,1)$
0.8	$0.8\,\mathrm{N}\,(0,1) + 0.2\,\mathrm{N}\,(3,1)$	$0.8\,\mathrm{Lognormal}\,(0,1) + 0.2\,\mathrm{Lognormal}\,(3,1)$
0.9	$0.9\,\mathrm{N}\,(0,1) + 0.1\,\mathrm{N}\,(3,1)$	$0.9\,\mathrm{Lognormal}\,(0,1) + 0.1\,\mathrm{Lognormal}\,(3,1)$

Table 5.3: Alpha Effect

α value	Normal Mixture	Cauchy Mixture
0.2	$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1)$	$0.5\,\mathrm{Cauchy}\,(0,1) + 0.5\,\mathrm{Cauchy}\,(3,1)$
0.4	$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1)$	$0.5\,\mathrm{Cauchy}\,(0,1) + 0.5\,\mathrm{Cauchy}\,(3,1)$
0.6	$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1)$	$0.5\,\mathrm{Cauchy}\,(0,1) + 0.5\,\mathrm{Cauchy}\,(3,1)$
0.8	$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1)$	$0.5\,\mathrm{Cauchy}\,(0,1) + 0.5\,\mathrm{Cauchy}\,(3,1)$
0.1	$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1)$	$0.5\,\mathrm{Cauchy}\,(0,1) + 0.5\,\mathrm{Cauchy}\,(3,1)$
1.2	$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1)$	$0.5\,\mathrm{Cauchy}\,(0,1) + 0.5\,\mathrm{Cauchy}\,(3,1)$
1.4	$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1)$	$0.5\,\mathrm{Cauchy}\,(0,1) + 0.5\,\mathrm{Cauchy}\,(3,1)$
1.6	$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1)$	$0.5\,\mathrm{Cauchy}\,(0,1) + 0.5\,\mathrm{Cauchy}\,(3,1)$
1.8	$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1)$	$0.5\,\mathrm{Cauchy}\,(0,1) + 0.5\,\mathrm{Cauchy}\,(3,1)$
2.0	$0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1)$	$0.5\,\mathrm{Cauchy}\,(0,1) + 0.5\,\mathrm{Cauchy}\,(3,1)$

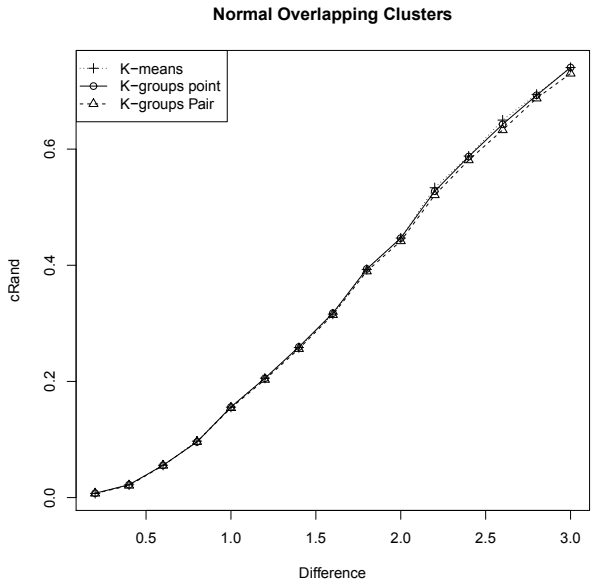


Figure 5.1: Overlapping effect for normal mixture distributions, $n = 200,\ B = 500$

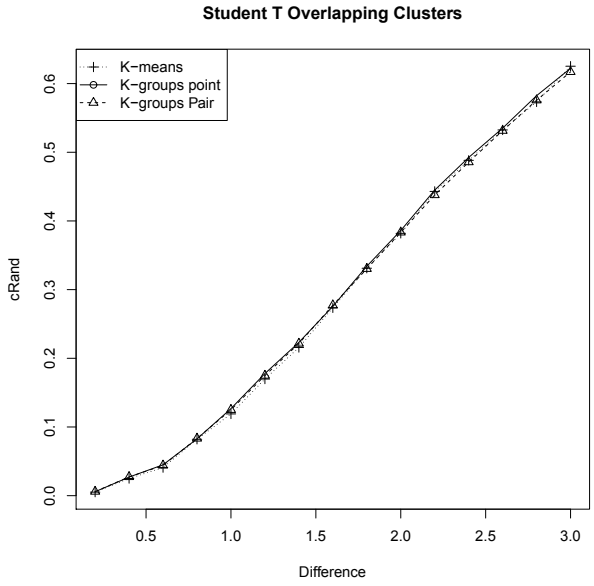


Figure 5.2: Overlapping effect for Student T mixture distributions, $n = 200,\ B = 500$

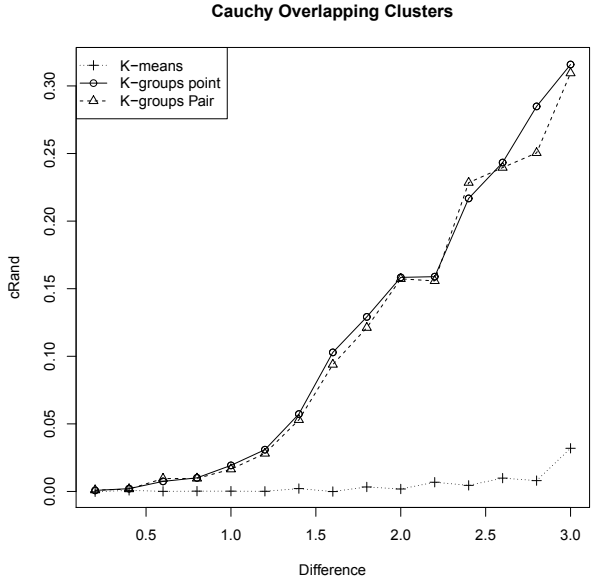


Figure 5.3: Overlapping effect for Cauchy mixture distributions, $n = 200,\ B = 500$

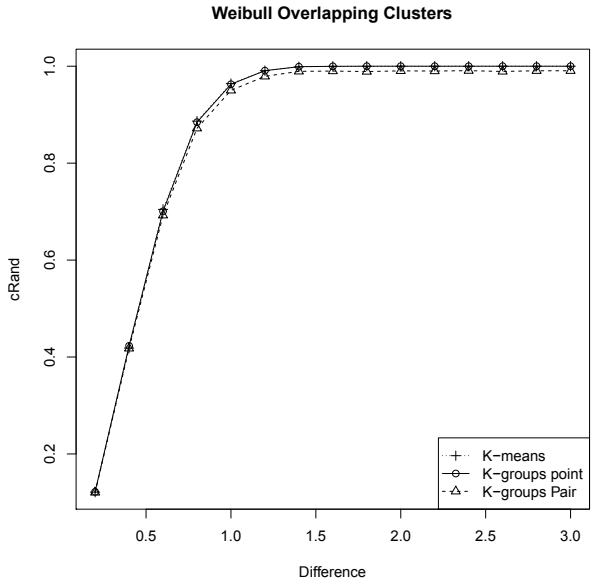


Figure 5.4: Overlapping effect for Weibull mixture distributions, $n = 200,\ B = 500$

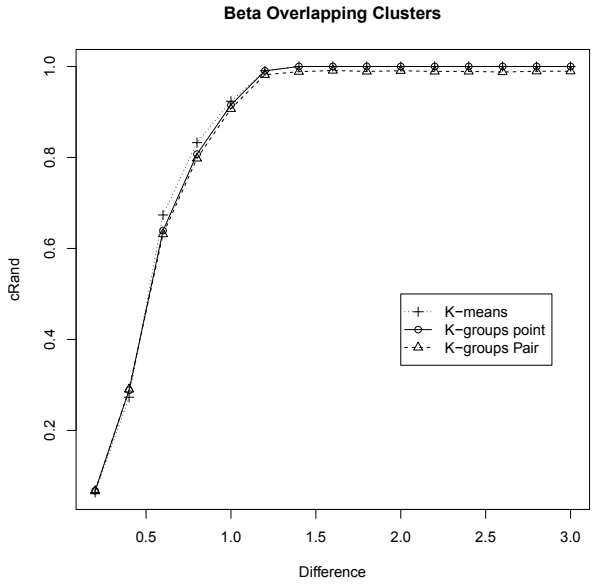


Figure 5.5: Overlapping effect for beta mixture distributions, $n = 200,\ B = 500$

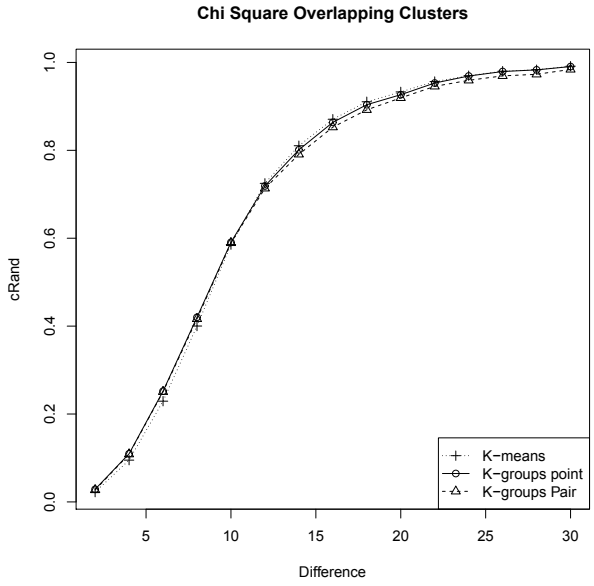


Figure 5.6: Overlapping effect for chi-square mixture distributions, $v = 10, n = 200, B = 500$

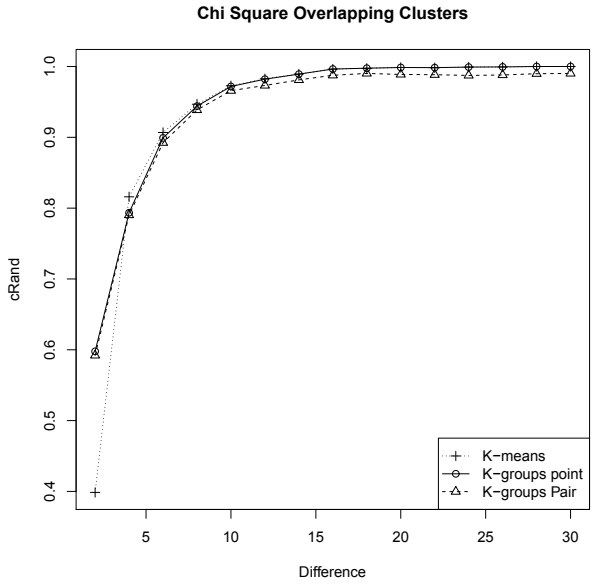


Figure 5.7: Overlapping effect for chi-square mixture distributions, $v = 1,\ n = 200,\ B = 500$

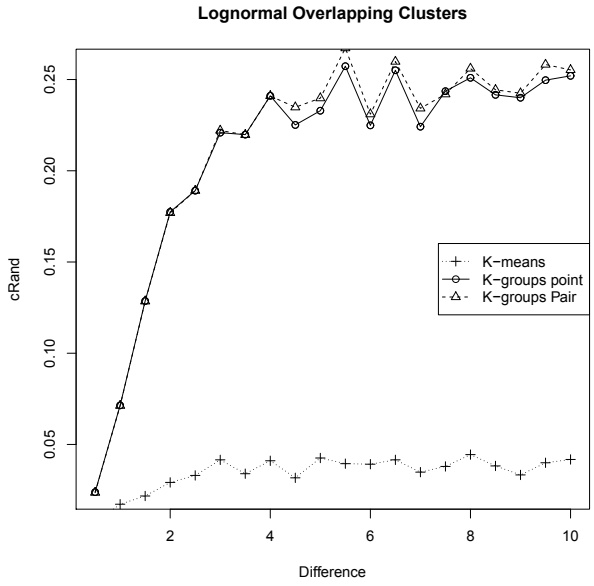


Figure 5.8: Overlapping effect for lognormal mixture distributions, $n = 200,\ B = 500$

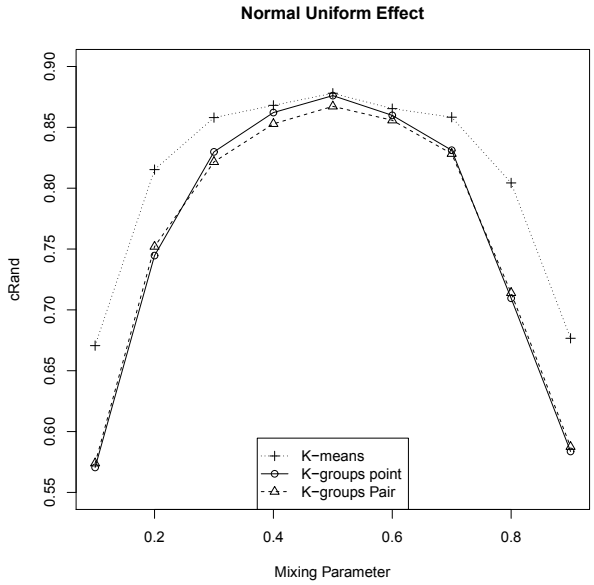


Figure 5.9: Uniform effect for normal mixture distributions, $n = 200,\ B = 1000$

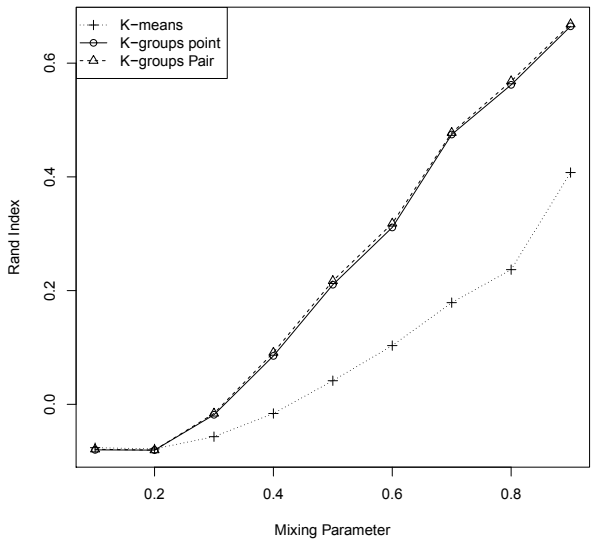


Figure 5.10: Uniform effect for lognormal mixture distributions, $n = 200,\ B = 1000$

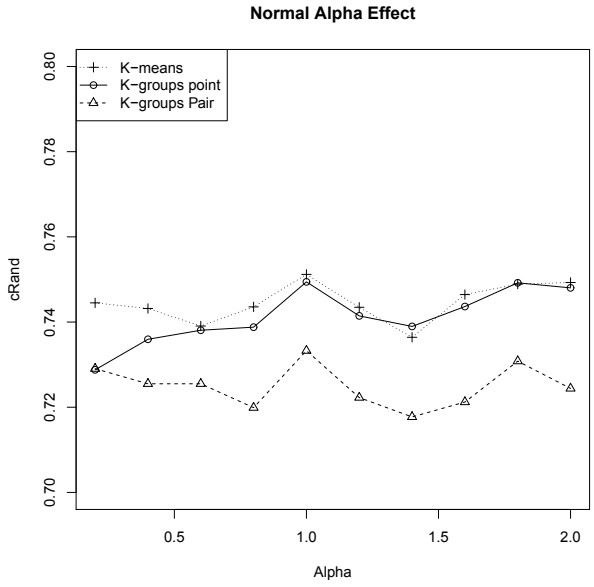


Figure 5.11: α effect for normal mixture distributions, $n = 200,\ B = 1000$

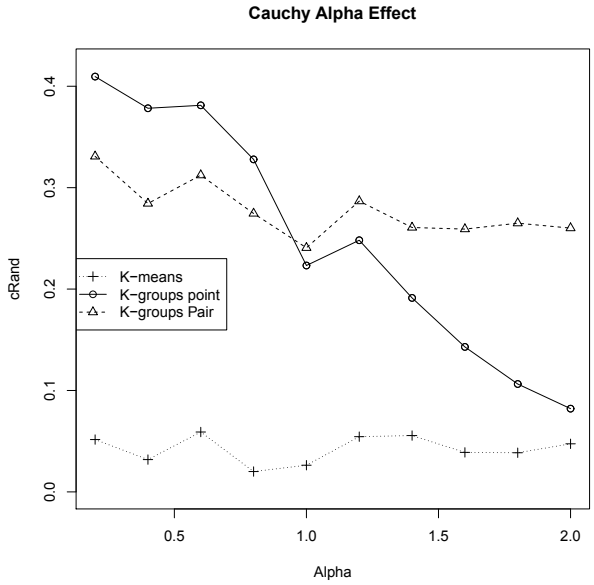


Figure 5.12: α effect for Cauchy mixture distributions, $n = 200,\ B = 1000$

Table 5.4: Normal Mixture Distribution $0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(3,1), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.9350	(0.0023)	0.9338	(0.0022)	0.9292	(0.0023)
	Kappa	0.8690	(0.0046)	0.8667	(0.0045)	0.8573	(0.0046)
	Rand	0.8783	(0.0040)	0.8761	(0.0039)	0.8681	(0.0039)
	Crand	0.7565	(0.0080)	0.7522	(0.0078)	0.7363	(0.0078)
$n = 200$	Diag	0.9313	(0.0013)	0.9308	(0.0013)	0.9291	(0.0013)
	Kappa	0.8621	(0.0025)	0.8612	(0.0025)	0.8577	(0.0025)
	Rand	0.8720	(0.0021)	0.8712	(0.0021)	0.8682	(0.0021)
	Crand	0.7440	(0.0043)	0.7425	(0.0043)	0.7365	(0.0043)
$n = 400$	Diag	0.9304	(0.0006)	0.9305	(0.0005)	0.9295	(0.0006)
	Kappa	0.8607	(0.0012)	0.8608	(0.0011)	0.8589	(0.0012)
	Rand	0.8705	(0.0010)	0.8706	(0.0009)	0.8689	(0.0010)
	Crand	0.7410	(0.0021)	0.7412	(0.0019)	0.7379	(0.0021)
$n = 800$	Diag	0.9313	(0.0003)	0.9311	(0.0002)	0.9308	(0.0003)
	Kappa	0.8625	(0.0006)	0.8620	(0.0006)	0.8615	(0.0006)
	Rand	0.8720	(0.0005)	0.8716	(0.0004)	0.8712	(0.0005)
	Crand	0.7441	(0.0011)	0.7433	(0.0009)	0.7425	(0.0011)

Table 5.5: Normal Mixture Distribution $0.5\,\mathrm{N}(0,1) + 0.5\,\mathrm{N}(2,1), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.8378	(0.0037)	0.8371	(0.0037)	0.8338	(0.0038)
	Kappa	0.6742	(0.0074)	0.6731	(0.0074)	0.6667	(0.0076)
	Rand	0.7282	(0.0050)	0.7273	(0.0051)	0.7231	(0.0051)
	Crand	0.4565	(0.0101)	0.4547	(0.0100)	0.4462	(0.0102)
$n = 200$	Diag	0.8395	(0.0018)	0.8392	(0.0018)	0.8371	(0.0018)
	Kappa	0.6785	(0.0037)	0.6779	(0.0037)	0.6738	(0.0037)
	Rand	0.7305	(0.0024)	0.7301	(0.0025)	0.7273	(0.0025)
	Crand	0.4611	(0.0049)	0.4602	(0.0050)	0.4547	(0.0050)
$n = 400$	Diag	0.8395	(0.0009)	0.8390	(0.0009)	0.7753	(0.0009)
	Kappa	0.6787	(0.0018)	0.6777	(0.0018)	0.5504	(0.0018)
	Rand	0.7306	(0.0012)	0.7299	(0.0012)	0.6514	(0.0010)
	Crand	0.4613	(0.0025)	0.4598	(0.0025)	0.3029	(0.0021)
$n = 800$	Diag	0.8430	(0.0003)	0.8434	(0.0003)	0.8432	(0.0003)
	Kappa	0.6858	(0.0007)	0.6867	(0.0007)	0.6862	(0.0007)
	Rand	0.7352	(0.0004)	0.7357	(0.0004)	0.7355	(0.0004)
	Crand	0.4704	(0.0009)	0.4715	(0.0009)	0.4710	(0.0009)

Table 5.6: Normal Mixture Distribution $0.5\,\mathrm{N}(0,1) + 0.5\,\mathrm{N}(1,1), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.6903	(0.0046)	0.6902	(0.0047)	0.6887	(0.0047)
	Kappa	0.3801	(0.0092)	0.3801	(0.0093)	0.3774	(0.0092)
	Rand	0.5725	(0.0035)	0.5725	(0.0036)	0.5713	(0.0035)
	Crand	0.1450	(0.0071)	0.1451	(0.0072)	0.1427	(0.0071)
$n = 200$	Diag	0.6896	(0.0024)	0.6895	(0.0023)	0.6889	(0.0023)
	Kappa	0.3792	(0.0047)	0.3789	(0.0047)	0.3780	(0.0047)
	Rand	0.5721	(0.0018)	0.5719	(0.0018)	0.5715	(0.0018)
	Crand	0.1442	(0.0036)	0.1439	(0.0036)	0.1431	(0.0036)
$n = 400$	Diag	0.6906	(0.0011)	0.6903	(0.0011)	0.6902	(0.0011)
	Kappa	0.3812	(0.0023)	0.3808	(0.0023)	0.3804	(0.0023)
	Rand	0.5725	(0.0008)	0.5725	(0.0008)	0.5724	(0.0009)
	Crand	0.1450	(0.0017)	0.1450	(0.0017)	0.1486	(0.0018)
$n = 800$	Diag	0.6906	(0.0006)	0.6900	(0.0006)	0.6911	(0.0006)
	Kappa	0.3812	(0.0012)	0.3800	(0.0012)	0.3822	(0.0012)
	Rand	0.5727	(0.0004)	0.5722	(0.0004)	0.5731	(0.0004)
	Crand	0.1454	(0.0009)	0.1445	(0.0009)	0.1462	(0.0009)

Table 5.7: Normal Mixture Distribution $0.5\,\mathrm{N}\,(0,1) + 0.5\,\mathrm{N}\,(0,3), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.5917	(0.0057)	0.5705	(0.0049)	0.5714	(0.0050)
	Kappa	0.1818	(0.0109)	0.1408	(0.0098)	0.1424	(0.0099)
	Rand	0.5186	(0.0025)	0.5099	(0.0019)	0.5104	(0.0019)
	Crand	0.0373	(0.0049)	0.0199	(0.0038)	0.0209	(0.0039)
$n = 200$	Diag	0.5737	(0.0033)	0.5543	(0.0028)	0.5536	(0.0028)
	Kappa	0.1466	(0.0066)	0.1082	(0.0055)	0.1069	(0.0056)
	Rand	0.5130	(0.0011)	0.5066	(0.0008)	0.5065	(0.0008)
	Crand	0.0261	(0.0023)	0.0131	(0.0016)	0.0130	(0.0017)
$n = 400$	Diag	0.5590	(0.0020)	0.5411	(0.0015)	0.5410	(0.0015)
	Kappa	0.1178	(0.0039)	0.0822	(0.0030)	0.0823	(0.0030)
	Rand	0.5089	(0.0005)	0.5039	(0.0003)	0.5039	(0.0003)
	Crand	0.0179	(0.0011)	0.0079	(0.0007)	0.0079	(0.0007)
$n = 800$	Diag	0.5396	(0.0010)	0.5291	(0.0007)	0.5303	(0.0007)
	Kappa	0.0815	(0.0020)	0.0587	(0.0014)	0.0610	(0.0015)
	Rand	0.5041	(0.0002)	0.5087	(0.0001)	0.5020	(0.0001)
	Crand	0.0083	(0.0004)	0.0037	(0.0002)	0.0041	(0.0002)

Table 5.8: Student's T Mixture Distribution $0.5\,\mathrm{T}\,(4) + 0.5(\,\mathrm{T}\,(4) + 3), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.8936	(0.0033)	0.8927	(0.0032)	0.8894	(0.0031)
	Kappa	0.7859	(0.0067)	0.7844	(0.0064)	0.7776	(0.0062)
	Rand	0.8102	(0.0050)	0.8087	(0.0050)	0.8032	(0.0048)
	Crand	0.6205	(0.0101)	0.6173	(0.0101)	0.6065	(0.0097)
$n = 200$	Diag	0.8956	(0.0017)	0.8957	(0.0015)	0.8936	(0.0015)
	Kappa	0.7903	(0.0036)	0.7909	(0.0031)	0.7866	(0.0031)
	Rand	0.8133	(0.0025)	0.8132	(0.0025)	0.8100	(0.0024)
	Crand	0.6266	(0.0051)	0.6265	(0.0051)	0.6200	(0.0049)
$n = 400$	Diag	0.8960	(0.0007)	0.8962	(0.0007)	0.8950	(0.0007)
	Kappa	0.7916	(0.0015)	0.7921	(0.0015)	0.7897	(0.0015)
	Rand	0.8136	(0.0012)	0.8139	(0.0012)	0.8121	(0.0012)
	Crand	0.6273	(0.0024)	0.6279	(0.0024)	0.6243	(0.0025)
$n = 800$	Diag	0.8955	(0.0005)	0.8955	(0.0005)	0.8947	(0.0005)
	Kappa	0.7909	(0.0010)	0.7909	(0.0010)	0.7894	(0.0010)
	Rand	0.8130	(0.0008)	0.8130	(0.0008)	0.8118	(0.0008)
	Crand	0.6260	(0.0016)	0.6260	(0.0016)	0.6236	(0.0016)

Table 5.9: Student's T Mixture Distribution $0.5\,\mathrm{T}\,(4) + 0.5(\,\mathrm{T}\,(4) + 2), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.8067	(0.0044)	0.8098	(0.0040)	0.8071	(0.0039)
	Kappa	0.6116	(0.0088)	0.6185	(0.0079)	0.6128	(0.0078)
	Rand	0.6890	(0.0051)	0.6921	(0.0049)	0.6887	(0.0048)
	Crand	0.3780	(0.0103)	0.3843	(0.0098)	0.3773	(0.0097)
$n = 200$	Diag	0.8081	(0.0021)	0.8094	(0.0020)	0.8080	(0.0020)
	Kappa	0.6155	(0.0042)	0.6184	(0.0041)	0.6155	(0.0041)
	Rand	0.6901	(0.0025)	0.6917	(0.0025)	0.6899	(0.0025)
	Crand	0.3802	(0.0051)	0.3835	(0.0051)	0.3799	(0.0051)
$n = 400$	Diag	0.8113	(0.0009)	0.8122	(0.0009)	0.8112	(0.0009)
	Kappa	0.6223	(0.0019)	0.6240	(0.0019)	0.6221	(0.0019)
	Rand	0.6939	(0.0012)	0.6949	(0.0012)	0.6937	(0.0011)
	Crand	0.3878	(0.0024)	0.3898	(0.0024)	0.3874	(0.0023)
$n = 800$	Diag	0.8140	(0.0005)	0.8152	(0.0004)	0.8143	(0.0004)
	Kappa	0.6279	(0.0009)	0.6304	(0.0009)	0.6285	(0.0009)
	Rand	0.6970	(0.0006)	0.6985	(0.0006)	0.6973	(0.0005)
	Crand	0.3941	(0.0012)	0.3970	(0.0012)	0.3947	(0.0012)

Table 5.10: Student's T Mixture Distribution $0.5\,\mathrm{T}\,(4) + 0.5(\,\mathrm{T}\,(4) + 1), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.6672	(0.0050)	0.6748	(0.0046)	0.6734	(0.0045)
	Kappa	0.3327	(0.0099)	0.3488	(0.0091)	0.3457	(0.0089)
	Rand	0.5565	(0.0032)	0.5610	(0.0032)	0.5598	(0.0030)
	Crand	0.1130	(0.0064)	0.1220	(0.0064)	0.1197	(0.0061)
$n = 200$	Diag	0.6742	(0.0025)	0.6780	(0.0023)	0.6775	(0.0023)
	Kappa	0.3482	(0.0050)	0.3559	(0.0046)	0.3550	(0.0047)
	Rand	0.5611	(0.0017)	0.5633	(0.0016)	0.5630	(0.0016)
	Crand	0.1223	(0.0035)	0.0126	(0.0033)	0.1261	(0.0033)
$n = 400$	Diag	0.6757	(0.0012)	0.6766	(0.0011)	0.6760	(0.0011)
	Kappa	0.3515	(0.0025)	0.3533	(0.0023)	0.3522	(0.0023)
	Rand	0.5617	(0.0008)	0.5623	(0.0008)	0.5619	(0.0008)
	Crand	0.1235	(0.0018)	0.1247	(0.0016)	0.1239	(0.0016)
$n = 800$	Diag	0.6775	(0.0006)	0.6786	(0.0005)	0.6787	(0.0004)
	Kappa	0.3544	(0.0012)	0.3568	(0.0011)	0.3572	(0.0011)
	Rand	0.5630	(0.0004)	0.5637	(0.0004)	0.5637	(0.0004)
	Crand	0.1260	(0.0009)	0.1275	(0.0008)	0.1274	(0.0008)

Table 5.11: Logistic Mixture Distribution $0.5\,\text{Logistic}\,(0,1) + 0.5\,\text{Logistic}\,(0,4), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.6291	(0.0054)	0.6078	(0.0062)	0.6064	(0.0062)
	Kappa	0.2546	(0.0099)	0.2156	(0.0120)	0.2119	(0.0119)
	Rand	0.5346	(0.0028)	0.5262	(0.0029)	0.5256	(0.0029)
	Crand	0.0694	(0.0056)	0.0052	(0.0058)	0.0512	(0.0058)
$n = 200$	Diag	0.6281	(0.0031)	0.5943	(0.0038)	0.5944	(0.0038)
	Kappa	0.2563	(0.0057)	0.1897	(0.0077)	0.1896	(0.0075)
	Rand	0.5344	(0.0015)	0.5214	(0.0016)	0.5212	(0.0015)
	Crand	0.0688	(0.0030)	0.0428	(0.0032)	0.0425	(0.0031)
$n = 400$	Diag	0.6632	(0.0014)	0.5831	(0.0026)	0.5821	(0.0020)
	Kappa	0.3274	(0.0025)	0.1646	(0.0052)	0.1626	(0.0037)
	Rand	0.5539	(0.0009)	0.5182	(0.0009)	0.5174	(0.0009)
	Crand	0.1077	(0.0019)	0.0364	(0.0019)	0.0349	(0.0019)
$n = 800$	Diag	0.6348	(0.0007)	0.5803	(0.0013)	0.5830	(0.0010)
	Kappa	0.2722	(0.0012)	0.1626	(0.0028)	0.1662	(0.0019)
	Rand	0.5367	(0.0004)	0.5152	(0.0004)	0.5163	(0.0004)
	Crand	0.0734	(0.0008)	0.0305	(0.0008)	0.0326	(0.0009)

Table 5.12: Cauchy Mixture Distribution $0.5\,\text{Cauchy}\,(0,1) + 0.5, \text{Cauchy}\,(3,1), \alpha = 0.5$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.5694	(0.0090)	0.8044	(0.0053)	0.7599	(0.0105)
	Kappa	0.0764	(0.0194)	0.6061	(0.0110)	0.5084	(0.0023)
	Rand	0.5210	(0.0062)	0.6879	(0.0056)	0.6540	(0.0083)
	Crand	0.0424	(0.0123)	0.3758	(0.0112)	0.3083	(0.0165)
$n = 200$	Diag	0.5444	(0.0049)	0.8107	(0.0020)	0.7835	(0.0059)
	Kappa	0.0390	(0.0103)	0.6204	(0.0041)	0.5612	(0.0012)
	Rand	0.5113	(0.0032)	0.6933	(0.0024)	0.6732	(0.0045)
	Crand	0.0225	(0.0065)	0.3866	(0.0049)	0.3464	(0.0090)
$n = 400$	Diag	0.5276	(0.0024)	0.8120	(0.0009)	0.8045	(0.0020)
	Kappa	0.0162	(0.0048)	0.6235	(0.0019)	0.6076	(0.0043)
	Rand	0.5049	(0.0015)	0.6947	(0.0012)	0.6879	(0.0017)
	Crand	0.0097	(0.0030)	0.3895	(0.0024)	0.3759	(0.0034)
$n = 800$	Diag	0.5150	(0.0012)	0.8192	(0.0005)	0.8138	(0.0010)
	Kappa	0.5166	(0.0024)	0.6382	(0.0009)	0.6272	(0.0021)
	Rand	0.5001	(0.0008)	0.7038	(0.0006)	0.6968	(0.0008)
	Crand	0.0045	(0.0015)	0.4076	(0.0012)	0.3937	(0.0017)

Table 5.13: Cauchy Mixture Distribution $0.5, \text{Cauchy}\,(0,1) + 0.5, \text{Cauchy}\,(2,1), \alpha = 0.5$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.5470	(0.0052)	0.7255	(0.0069)	0.6751	(0.0100)
	Kappa	0.0280	(0.0103)	0.4464	(0.0145)	0.3277	(0.0226)
	Rand	0.5049	(0.0026)	0.6074	(0.0053)	0.5773	(0.0064)
	Crand	0.0102	(0.0051)	0.2148	(0.0107)	0.0017	(0.0128)
$n = 200$	Diag	0.5296	(0.0019)	0.7420	(0.0029)	0.6946	(0.0065)
	Kappa	0.0073	(0.0029)	0.4827	(0.0060)	0.3772	(0.0143)
	Rand	0.5007	(0.0006)	0.6187	(0.0025)	0.5907	(0.0039)
	Crand	0.0052	(0.0012)	0.2375	(0.0051)	0.1814	(0.0078)
$n = 400$	Diag	0.5223	(0.0009)	0.7517	(0.0015)	0.7065	(0.0040)
	Kappa	0.0028	(0.0014)	0.5027	(0.0030)	0.4067	(0.0089)
	Rand	0.5002	(0.0003)	0.6267	(0.0012)	0.5982	(0.0023)
	Crand	0.0026	(0.0006)	0.2534	(0.0025)	0.1964	(0.0044)
$n = 800$	Diag	0.5114	(0.0004)	0.7466	(0.0007)	0.7310	(0.0020)
	Kappa	0.0009	(0.0007)	0.4932	(0.0015)	0.4620	(0.0040)
	Rand	0.4997	(0.0001)	0.6214	(0.0006)	0.6123	(0.0011)
	Crand	0.0011	(0.0002)	0.2429	(0.0012)	0.2248	(0.0022)

Table 5.14: Cauchy Mixture Distribution $0.5, \text{Cauchy}\,(0,1) + 0.5, \text{Cauchy}\,(1,1), \alpha = 0.5$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.5434	(0.0032)	0.6058	(0.0065)	0.5761	(0.0059)
	Kappa	0.0137	(0.0045)	0.2015	(0.0138)	0.1200	(0.0137)
	Rand	0.5009	(0.0007)	0.5262	(0.0032)	0.5138	(0.0025)
	Crand	0.0011	(0.0011)	0.0523	(0.0065)	0.0282	(0.0049)
$n = 200$	Diag	0.5273	(0.0015)	0.6115	(0.0038)	0.5757	(0.0041)
	Kappa	0.0040	(0.0012)	0.2216	(0.0079)	0.1296	(0.0095)
	Rand	0.4999	(0.0003)	0.5285	(0.0017)	0.5161	(0.0015)
	Crand	0.0006	(0.0001)	0.0569	(0.0034)	0.0324	(0.0031)
$n = 400$	Diag	0.5196	(0.0007)	0.6288	(0.0020)	0.5765	(0.0020)
	Kappa	0.9317	(0.0006)	0.2571	(0.0040)	0.1353	(0.0049)
	Rand	0.4999	(0.0001)	0.5353	(0.0008)	0.5173	(0.0007)
	Crand	0.0003	(< 0.0001)	0.0706	(0.0017)	0.0343	(0.0016)
$n = 800$	Diag	0.5151	(0.0003)	0.6373	(0.0010)	0.5635	(0.0010)
	Kappa	-0.0005	(0.0003)	0.2732	(0.0020)	0.187	(0.0024)
	Rand	0.5000	(< 0.0001)	0.5380	(0.0004)	0.5153	(0.0003)
	Crand	< -0.0001	(< 0.0001)	0.0761	(0.0008)	0.0310	(0.0008)

Table 5.15: Weibull Mixture Distribution $0.5\,\text{Weibull}\,(1,5) + 0.5(\,\text{Weibull}\,(1,5) + 2), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	1	(0)	1	(0)	0.9949	(0.0005)
	Kappa	1	(0)	1	(0)	0.9898	(0.0010)
	Rand	1	(0)	1	(0)	0.9899	(0.0010)
	Crand	1	(0)	1	(0)	0.9799	(0.0020)
$n = 200$	Diag	1	(0)	1	(0)	0.9973	(0.0001)
	Kappa	1	(0)	1	(0)	0.9946	(0.0003)
	Rand	1	(0)	1	(0)	0.9946	(0.0003)
	Crand	1	(0)	1	(0)	0.9893	(0.0007)
$n = 400$	Diag	1	(0)	1	(0)	0.9988	(< 0.0001)
	Kappa	1	(0)	1	(0)	0.9976	(0.0001)
	Rand	1	(0)	1	(0)	0.9977	(0.0001)
	Crand	1	(0)	1	(0)	0.9953	(0.0002)
$n = 800$	Diag	1	(0)	1	(0)	0.9996	(< 0.0001)
	Kappa	1	(0)	1	(0)	0.9992	(< 0.0001)
	Rand	1	(0)	1	(0)	0.9992	(< 0.0001)
	Crand	1	(0)	1	(0)	0.9985	(< 0.0001)

Table 5.16: Weibull Mixture Distribution $0.5\,\text{Weibull}\,(1,5) + 0.5(\,\text{Weibull}\,(1,5) + 1), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.9915	(0.0009)	0.9911	(0.0009)	0.9856	(0.0011)
	Kappa	0.9828	(0.0018)	0.9820	(0.0019)	0.9708	(0.0022)
	Rand	0.9832	(0.0018)	0.9823	(0.0018)	0.9715	(0.0021)
	Crand	0.9664	(0.0036)	0.9647	(0.0037)	0.9431	(0.0042)
$n = 200$	Diag	0.9917	(0.0004)	0.9915	(0.0005)	0.9885	(0.0005)
	Kappa	0.9833	(0.0009)	0.9829	(0.0010)	0.9769	(0.0010)
	Rand	0.9835	(0.0009)	0.9832	(0.0010)	0.9773	(0.0010)
	Crand	0.9671	(0.0018)	0.9664	(0.0020)	0.9547	(0.0019)
$n = 400$	Diag	0.9920	(0.0002)	0.9917	(0.0002)	0.9903	(0.0002)
	Kappa	0.9839	(0.0004)	0.9835	(0.0005)	0.9805	(0.0005)
	Rand	0.9841	(0.0004)	0.9836	(0.0005)	0.9808	(0.0005)
	Crand	0.9682	(0.0008)	0.9673	(0.0009)	0.9616	(0.0009)
$n = 800$	Diag	0.9915	(0.0001)	0.9913	(0.0001)	0.9905	(0.0001)
	Kappa	0.9831	(0.0002)	0.9827	(0.0002)	0.9809	(0.0002)
	Rand	0.9832	(0.0002)	0.9829	(0.0002)	0.9811	(0.0002)
	Crand	0.9665	(0.0005)	0.9658	(0.0005)	0.9623	(0.0005)

Table 5.17: Weibull Mixture Distribution $0.5\,\mathrm{Weibull}\,(1,5) + 0.5(\,\mathrm{Weibull}\,(1,5) + 0.5), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.8798	(0.0033)	0.8789	(0.0033)	0.8753	(0.0034)
	Kappa	0.7584	(0.0066)	0.7569	(0.0065)	0.7497	(0.0068)
	Rand	0.7886	(0.0050)	0.7872	(0.0050)	0.7819	(0.0051)
	Crand	0.5772	(0.0100)	0.5744	(0.0100)	0.5637	(0.0102)
$n = 200$	Diag	0.8794	(0.0016)	0.8788	(0.0016)	0.8775	(0.0016)
	Kappa	0.7580	(0.0032)	0.7570	(0.0032)	0.7545	(0.0033)
	Rand	0.7879	(0.0024)	0.7870	(0.0025)	0.7851	(0.0025)
	Crand	0.5758	(0.0049)	0.5741	(0.0050)	0.5702	(0.0050)
$n = 400$	Diag	0.8804	(0.0008)	0.8808	(0.0008)	0.8792	(0.0008)
	Kappa	0.7605	(0.0016)	0.0016	(0.0016)	0.7583	(0.0016)
	Rand	0.7894	(0.0012)	0.0012	(0.0012)	0.7876	(0.0012)
	Crand	0.5789	(0.0025)	0.0024	(0.0024)	0.5752	(0.0024)
$n = 800$	Diag	0.8805	(0.0003)	0.8804	(0.0003)	0.8806	(0.0003)
	Kappa	0.7609	(0.0007)	0.7607	(0.0007)	0.7612	(0.0007)
	Rand	0.7895	(0.0005)	0.7894	(0.0005)	0.7897	(0.0005)
	Crand	0.5791	(0.0010)	0.5788	(0.0011)	0.5795	(0.0010)

Table 5.18: Beta Mixture Distribution $0.5\,\mathrm{Beta}\,(2,1) + 0.5(\,\mathrm{Beta}\,(2,1) + 2), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	1	(0)	1	(0)	0.9951	(0.0005)
	Kappa	1	(0)	1	(0)	0.9901	(0.0010)
	Rand	1	(0)	1	(0)	0.9902	(0.0010)
	Crand	1	(0)	1	(0)	0.9805	(0.0020)
$n = 200$	Diag	1	(0)	1	(0)	0.9972	(0.0002)
	Kappa	1	(0)	1	(0)	0.9945	(0.0005)
	Rand	1	(0)	1	(0)	0.9945	(0.0005)
	Crand	1	(0)	1	(0)	0.9891	(0.0010)
$n = 400$	Diag	1	(0)	1	(0)	0.9987	(0.0001)
	Kappa	1	(0)	1	(0)	0.9974	(0.0002)
	Rand	1	(0)	1	(0)	0.9974	(0.0002)
	Crand	1	(0)	1	(0)	0.9949	(0.0005)
$n = 800$	Diag	1	(0)	1	(0)	0.9993	(< 0.0001)
	Kappa	1	(0)	1	(0)	0.9986	(0.0001)
	Rand	1	(0)	1	(0)	0.9986	(0.0001)
	Crand	1	(0)	1	(0)	0.9972	(0.0002)

Table 5.19: Beta Mixture Distribution $0.5\,\mathrm{Beta}\,(2,1) + 0.5(\,\mathrm{Beta}\,(2,1) + 0.5), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.8387	(0.0045)	0.8437	(0.0045)	0.8398	(0.0046)
	Kappa	0.6773	(0.0088)	0.6883	(0.0087)	0.6808	(0.0089)
	Rand	0.7308	(0.0059)	0.7377	(0.0061)	0.7326	(0.0061)
	Crand	0.4616	(0.0119)	0.4756	(0.0122)	0.4653	(0.0123)
$n = 200$	Diag	0.8410	(0.0026)	0.8493	(0.0024)	0.8462	(0.0024)
	Kappa	0.6820	(0.0051)	0.6988	(0.0048)	0.6927	(0.0049)
	Rand	0.7340	(0.0034)	0.7451	(0.0033)	0.7409	(0.0033)
	Crand	0.4680	(0.0069)	0.4903	(0.0066)	0.4818	(0.0067)
$n = 400$	Diag	0.8406	(0.0013)	0.8539	(0.0012)	0.8521	(0.0013)
	Kappa	0.6813	(0.0027)	0.7083	(0.0024)	0.7045	(0.0025)
	Rand	0.7329	(0.0018)	0.7511	(0.0017)	0.7487	(0.0018)
	Crand	0.4658	(0.0037)	0.5024	(0.0034)	0.4975	(0.0036)
$n = 800$	Diag	0.8458	(0.0007)	0.8624	(0.0007)	0.8611	(0.0007)
	Kappa	0.6913	(0.0014)	0.7247	(0.0014)	0.7222	(0.0014)
	Rand	0.7396	(0.0010)	0.7632	(0.0010)	0.7614	(0.0010)
	Crand	0.4793	(0.0021)	0.5264	(0.0021)	0.5229	(0.0021)

Table 5.20: Chi-square Mixture Distribution $0.5\chi^2_{10} + 0.5(\chi^2_{10} + 30), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.9969	(0.0005)	0.9969	(0.0005)	0.9927	(0.0007)
	Kappa	0.9938	(0.0011)	0.9937	(0.0011)	0.9852	(0.0015)
	Rand	0.9939	(0.0011)	0.9938	(0.0011)	0.9854	(0.0015)
	Crand	0.9878	(0.0022)	0.9877	(0.0022)	0.9709	(0.0030)
$n = 200$	Diag	0.9971	(0.0002)	0.9970	(0.0002)	0.9952	(0.0003)
	Kappa	0.9942	(0.0005)	0.9941	(0.0005)	0.9903	(0.0006)
	Rand	0.9943	(0.0005)	0.9941	(0.0005)	0.9904	(0.0006)
	Crand	0.9886	(0.0010)	0.9883	(0.0011)	0.9808	(0.0013)
$n = 400$	Diag	0.9972	(0.0001)	0.9972	(0.0001)	0.9961	(0.0001)
	Kappa	0.9945	(0.0002)	0.9944	(0.0002)	0.9923	(0.0003)
	Rand	0.9946	(0.0002)	0.9945	(0.0002)	0.9923	(0.0003)
	Crand	0.9892	(0.0005)	0.9890	(0.0005)	0.9847	(0.0006)
$n = 800$	Diag	0.9974	(< 0.0001)	0.9972	(< 0.0001)	0.9969	(< 0.0001)
	Kappa	0.9948	(0.0001)	0.9944	(0.0001)	0.9938	(0.0001)
	Rand	0.9948	(0.0001)	0.9945	(0.0001)	0.9938	(0.0001)
	Crand	0.9897	(0.0003)	0.9890	(0.0002)	0.9877	(0.0003)

Table 5.21: Chi-square Mixture Distribution $0.5\chi^2_{10} + 0.5(\chi^2_{10} + 10), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.8772	(0.0037)	0.8803	(0.0036)	0.8763	(0.0036)
	Kappa	0.7530	(0.0075)	0.7596	(0.0072)	0.7516	(0.0072)
	Rand	0.7853	(0.0056)	0.7899	(0.0055)	0.7838	(0.0054)
	Crand	0.5705	(0.0112)	0.5798	(0.0111)	0.5676	(0.0109)
$n = 200$	Diag	0.8809	(0.0017)	0.8835	(0.0017)	0.8819	(0.0017)
	Kappa	0.7612	(0.0034)	0.7667	(0.0033)	0.7634	(0.0033)
	Rand	0.7904	(0.0026)	0.7944	(0.0026)	0.7919	(0.0025)
	Crand	0.5808	(0.0052)	0.5888	(0.0052)	0.5839	(0.0051)
$n = 400$	Diag	0.8804	(0.0008)	0.8843	(0.0008)	0.8829	(0.0008)
	Kappa	0.7605	(0.0017)	0.7685	(0.0016)	0.7657	(0.0016)
	Rand	0.7895	(0.0013)	0.7955	(0.0013)	0.7933	(0.0012)
	Crand	0.5791	(0.0027)	0.5911	(0.0025)	0.5867	(0.005)
$n = 800$	Diag	0.8843	(0.0004)	0.8850	(0.0004)	0.8846	(0.0004)
	Kappa	0.7689	(0.0008)	0.7704	(0.0008)	0.7695	(0.0008)
	Rand	0.7954	(0.0006)	0.7965	(0.0006)	0.7958	(0.0006)
	Crand	0.5909	(0.0012)	0.5930	(0.0012)	0.5916	(0.0011)

Table 5.22: Chi-Square Mixture Distribution $0.5\chi^2_{10} + 0.5(\chi^2_{10} + 5), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.6933	(0.0052)	0.7079	(0.0050)	0.7066	(0.0049)
	Kappa	0.3871	(0.0102)	0.4163	(0.0097)	0.4135	(0.0098)
	Rand	0.5761	(0.0041)	0.5873	(0.0041)	0.5862	(0.0041)
	Crand	0.1522	(0.0083)	0.1747	(0.0083)	0.1725	(0.0082)
$n = 200$	Diag	0.6959	(0.0026)	0.7103	(0.0024)	0.7090	(0.0025)
	Kappa	0.3918	(0.0050)	0.4209	(0.0048)	0.4184	(0.0049)
	Rand	0.5773	(0.0020)	0.5888	(0.0021)	0.5879	(0.0021)
	Crand	0.1547	(0.0041)	0.1777	(0.0042)	0.1758	(0.0042)
$n = 400$	Diag	0.6963	(0.0013)	0.7109	(0.0012)	0.7096	(0.0012)
	Kappa	0.3934	(0.0025)	0.4221	(0.0024)	0.4196	(0.0024)
	Rand	0.5775	(0.0010)	0.5891	(0.0010)	0.5880	(0.0010)
	Crand	0.1550	(0.0021)	0.1783	(0.0021)	0.1761	(0.0021)
$n = 800$	Diag	0.6947	(0.0006)	0.7100	(0.0007)	0.7080	(0.0006)
	Kappa	0.3878	(0.0012)	0.4197	(0.0014)	0.4160	(0.0013)
	Rand	0.5759	(0.0004)	0.5885	(0.0005)	0.5867	(0.0057)
	Crand	0.1518	(0.0009)	0.1770	(0.0011)	0.1735	(0.0011)

Table 5.23: Chi-square Mixture distribution $0.5\chi_1^2 + 0.5(\chi_1^2 + 8), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.9863	(0.0011)	0.9856	(0.0012)	0.9816	(0.0013)
	Kappa	0.9723	(0.0023)	0.9709	(0.0024)	0.9630	(0.0027)
	Rand	0.9730	(0.0022)	0.9716	(0.0023)	0.9640	(0.0026)
	Crand	0.9460	(0.0045)	0.9433	(0.0047)	0.9280	(0.0052)
$n = 200$	Diag	0.9869	(0.0005)	0.9862	(0.0006)	0.9841	(0.0006)
	Kappa	0.9738	(0.0011)	0.9723	(0.0012)	0.9682	(0.0013)
	Rand	0.9742	(0.0011)	0.9728	(0.0012)	0.9688	(0.0013)
	Crand	0.9485	(0.0022)	0.9457	(0.0024)	0.9377	(0.0026)
$n = 400$	Diag	0.9866	(0.0003)	0.9860	(0.0003)	0.9849	(0.0003)
	Kappa	0.9732	(0.0006)	0.9720	(0.0006)	0.9698	(0.0006)
	Rand	0.9736	(0.0006)	0.9725	(0.0006)	0.9703	(0.0006)
	Crand	0.9473	(0.0011)	0.9450	(0.0012)	0.9407	(0.0012)
$n = 800$	Diag	0.9868	(0.0001)	0.9861	(0.0001)	0.9860	(0.0001)
	Kappa	0.9737	(0.0003)	0.9723	(0.0003)	0.9719	(0.0003)
	Rand	0.9740	(0.0003)	0.9727	(0.0003)	0.9723	(0.0003)
	Crand	0.9481	(0.0005)	0.9455	(0.0006)	0.9447	(0.0005)

Table 5.24: Chi-square Mixture Distribution $0.5\chi_1^2 + 0.5(\chi_1^2 + 3), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.9263	(0.0052)	0.9232	(0.0028)	0.9193	(0.0029)
	Kappa	0.8516	(0.0107)	0.8456	(0.0056)	0.8377	(0.0058)
	Rand	0.8678	(0.0067)	0.8585	(0.0047)	0.8519	(0.0048)
	Crand	0.7357	(0.0135)	0.7171	(0.0095)	0.7039	(0.0096)
$n = 200$	Diag	0.9309	(0.0023)	0.9236	(0.0014)	0.9217	(0.0014)
	Kappa	0.8616	(0.0045)	0.8466	(0.0027)	0.8430	(0.0028)
	Rand	0.8730	(0.0031)	0.8589	(0.0023)	0.8559	(0.0024)
	Crand	0.7460	(0.0063)	0.7179	(0.0047)	0.7118	(0.0047)
$n = 400$	Diag	0.9338	(0.0012)	0.9234	(0.0007)	0.9221	(0.0007)
	Kappa	0.8675	(0.0022)	0.8467	(0.0014)	0.8441	(0.0014)
	Rand	0.8764	(0.0015)	0.8585	(0.0012)	0.8564	(0.0012)
	Crand	0.7528	(0.0031)	0.7171	(0.0024)	0.7129	(0.0024)
$n = 800$	Diag	0.9362	(0.0006)	0.9260	(0.0003)	0.9255	(0.0003)
	Kappa	0.8721	(0.0011)	0.8516	(0.0006)	0.8505	(0.0007)
	Rand	0.8805	(0.0007)	0.8630	(0.0006)	0.8621	(0.0006)
	Crand	0.7611	(0.0015)	0.7260	(0.0010)	0.7541	(0.0011)

Table 5.25: Chi-square Mixture Distribution $0.5\chi_1^2 + 0.5(\chi_1^2 + 1), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.5659	(0.0045)	0.6232	(0.0086)	0.6181	(0.0079)
	Kappa	0.1070	(0.0097)	0.2431	(0.0172)	0.2343	(0.0158)
	Rand	0.5078	(0.0015)	0.5405	(0.0065)	0.5358	(0.0056)
	Crand	0.0157	(0.0030)	0.0811	(0.0131)	0.0717	(0.0112)
$n = 200$	Diag	0.5569	(0.0025)	0.6031	(0.0040)	0.6074	(0.0044)
	Kappa	0.1021	(0.0050)	0.2062	(0.0078)	0.2148	(0.0087)
	Rand	0.5065	(0.0006)	0.5254	(0.0026)	0.5287	(0.0030)
	Crand	0.0131	(0.0013)	0.0509	(0.0052)	0.0575	(0.0061)
$n = 400$	Diag	0.5525	(0.0013)	0.5965	(0.0020)	0.5927	(0.0015)
	Kappa	0.0994	(0.0024)	0.1921	(0.0039)	0.1840	(0.0027)
	Rand	0.5057	(0.0003)	0.5207	(0.0013)	0.5178	(0.0006)
	Crand	0.0114	(0.0006)	0.0415	(0.0026)	0.0356	(0.0012)
$n = 800$	Diag	0.5482	(0.0007)	0.5863	(0.0009)	0.5863	(0.0008)
	Kappa	0.0935	(0.0012)	0.1717	(0.0015)	0.1717	(0.0014)
	Rand	0.5049	(0.0001)	0.5155	(0.0003)	0.5154	(0.0003)
	Crand	0.0098	(0.0003)	0.0310	(0.0006)	0.0309	(0.0006)

Table 5.26: Lognormal Mixture Distribution $0.5\text{Lognormal}(0, 1) + 0.5\text{Lognormal}(10, 1), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.6070	(0.0070)	0.7510	(0.0079)	0.7505	(0.0079)
	Kappa	0.1936	(0.0148)	0.5037	(0.0148)	0.5030	(0.0149)
	Rand	0.5279	(0.0036)	0.6349	(0.0079)	0.6347	(0.0078)
	Crand	0.0558	(0.0072)	0.2699	(0.0157)	0.2694	(0.0157)
$n = 200$	Diag	0.5882	(0.0037)	0.7477	(0.0042)	0.7479	(0.0042)
	Kappa	0.1668	(0.0074)	0.4964	(0.0080)	0.4968	(0.0079)
	Rand	0.5187	(0.0015)	0.6282	(0.0042)	0.6282	(0.0041)
	Crand	0.0375	(0.0031)	0.2565	(0.0084)	0.2565	(0.0082)
$n = 400$	Diag	0.5130	(0.0021)	0.7353	(0.0022)	0.7356	(0.0022)
	Kappa	0.3730	(0.0043)	0.4733	(0.0041)	0.4738	(0.0041)
	Rand	0.4998	(0.0007)	0.6137	(0.0020)	0.6139	(0.0020)
	Crand	0.1530	(0.0014)	0.2275	(0.0041)	0.2280	(0.0041)
$n = 800$	Diag	0.5663	(0.0013)	0.7510	(0.0009)	0.7508	(0.0008)
	Kappa	0.1320	(0.0024)	0.5019	(0.0016)	0.5016	(0.0015)
	Rand	0.5109	(0.0004)	0.6268	(0.0009)	0.6265	(0.0008)
	Crand	0.0213	(0.0008)	0.2537	(0.0018)	0.2531	(0.0017)

Table 5.27: Lognormal Mixture Distribution $0.5\text{Lognormal}(0,1) + 0.5\text{Lognormal}(3,1), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.6045	(0.0067)	0.7254	(0.0074)	0.7289	(0.0075)
	Kappa	0.1912	(0.0140)	0.4537	(0.0135)	0.4605	(0.0138)
	Rand	0.5263	(0.0034)	0.6089	(0.0066)	0.6123	(0.0067)
	Crand	0.0527	(0.0069)	0.2178	(0.0133)	0.2247	(0.0136)
$n = 200$	Diag	0.5857	(0.0036)	0.7194	(0.0040)	0.7225	(0.0040)
	Kappa	0.1601	(0.0072)	0.4395	(0.0075)	0.4454	(0.0075)
	Rand	0.5175	(0.0015)	0.6009	(0.0035)	0.6036	(0.0036)
	Crand	0.0350	(0.0029)	0.2019	(0.0071)	0.2073	(0.0072)
$n = 400$	Diag	0.5763	(0.0019)	0.7241	(0.0020)	0.7214	(0.0021)
	Kappa	0.1514	(0.0037)	0.4483	(0.0038)	0.4442	(0.0039)
	Rand	0.5135	(0.0006)	0.6014	(0.0018)	0.6006	(0.0018)
	Crand	0.0272	(0.0013)	0.2031	(0.0036)	0.2012	(0.0037)
$n = 800$	Diag	0.5711	(0.0011)	0.7241	(0.0010)	0.7210	(0.0009)
	Kappa	0.1433	(0.0022)	0.4483	(0.0021)	0.4421	(0.0019)
	Rand	0.5113	(0.0003)	0.6014	(0.0009)	0.5984	(0.0008)
	Crand	0.0233	(0.0006)	0.2031	(0.0018)	0.1971	(0.0016)

Table 5.28: Lognormal Mixture Distribution $0.5\text{Lognormal}(0,1) + 0.5\text{Lognormal}(1,1), \alpha = 1$

Size	Indices	K-means		K-groups Point		K-groups Pair	
$n = 100$	Diag	0.5806	(0.0050)	0.6397	(0.0050)	0.6395	(0.0052)
	Kappa	0.1348	(0.0107)	0.2795	(0.0092)	0.2788	(0.0092)
	Rand	0.5132	(0.0020)	0.5401	(0.0030)	0.5398	(0.0029)
	Crand	0.0259	(0.0039)	0.0800	(0.0061)	0.0795	(0.0060)
$n = 200$	Diag	0.5643	(0.0027)	0.6369	(0.0027)	0.5643	(0.0027)
	Kappa	0.1178	(0.0054)	0.2741	(0.0047)	0.1178	(0.0048)
	Rand	0.5088	(0.0008)	0.5380	(0.0014)	0.5088	(0.0014)
	Crand	0.0177	(0.0016)	0.0730	(0.0029)	0.0177	(0.0029)
$n = 400$	Diag	0.5544	(0.0015)	0.6342	(0.0014)	0.6342	(0.0014)
	Kappa	0.1046	(0.0028)	0.2695	(0.0024)	0.2696	(0.0025)
	Rand	0.5066	(0.0003)	0.5365	(0.0007)	0.5366	(0.0007)
	Crand	0.0132	(0.0007)	0.0731	(0.0015)	0.0733	(0.0015)
$n = 800$	Diag	0.5445	(0.0008)	0.6305	(0.0006)	0.6302	(0.0006)
	Kappa	0.0937	(0.0014)	0.2666	(0.0011)	0.2661	(0.0012)
	Rand	0.5047	(0.0001)	0.5340	(0.0003)	0.5339	(0.0003)
	Crand	0.0096	(0.0003)	0.0682	(0.0007)	0.0679	(0.0007)

Table 5.29: Uniform Effect: Normal Mixture

Mixing Parameter	K-means	K-groups Point	K-groups Pair
$p = 0.1$	0.6689	0.6055	0.6003
$p = 0.2$	0.8058	0.7331	0.7316
$p = 0.3$	0.8733	0.8604	0.8541
$p = 0.4$	0.8673	0.8626	0.8585
$p = 0.5$	0.8703	0.8648	0.8643
$p = 0.6$	0.8723	0.8611	0.8532
$p = 0.7$	0.8376	0.8191	0.8140
$p = 0.8$	0.8017	0.7339	0.7306
$p = 0.9$	0.7001	0.5740	0.5907

Table 5.30: Uniform Effect: Logormal Mixture

Mixing Parameter	K-means	K-groups Point	K-groups Pair
$p = 0.1$	0.6507	0.5494	0.5487
$p = 0.2$	0.5693	0.5173	0.5085
$p = 0.3$	0.5200	0.5210	0.5234
$p = 0.4$	0.5074	0.5634	0.5602
$p = 0.5$	0.5170	0.6201	0.6219
$p = 0.6$	0.5857	0.6970	0.6932
$p = 0.7$	0.6626	0.7756	0.7747
$p = 0.8$	0.7515	0.8310	0.8334
$p = 0.9$	0.6777	0.5760	0.5709

Table 5.31: α Effect: Normal Mixture

Mixing Parameter	K-means	K-groups Point	K-groups Pair
$\alpha = 0.2$	0.7445	0.7328	0.7229
$\alpha = 0.4$	0.7431	0.7359	0.7255
$\alpha = 0.6$	0.7390	0.7380	0.7255
$\alpha = 0.8$	0.7435	0.7388	0.7199
$\alpha = 1.0$	0.7511	0.7494	0.7332
$\alpha = 1.2$	0.7434	0.7414	0.7222
$\alpha = 1.4$	0.7364	0.7389	0.7177
$\alpha = 1.6$	0.7464	0.7436	0.7212
$\alpha = 1.8$	0.7488	0.7492	0.7308
$\alpha = 2.0$	0.7492	0.7480	0.7244

Table 5.32: α Effect: Cauchy Mixture

Mixing Parameter	K-means	K-groups Point	K-groups Pair
$\alpha = 0.2$	0.0363	0.3816	0.3224
$\alpha = 0.4$	0.0477	0.3829	0.3376
$\alpha = 0.6$	0.0356	0.3724	0.2968
$\alpha = 0.8$	0.0703	0.3503	0.2932
$\alpha = 1.0$	0.0493	0.2793	0.2674
$\alpha = 1.2$	0.0743	0.2114	0.2822
$\alpha = 1.4$	0.0373	0.1732	0.2722
$\alpha = 1.6$	0.0470	0.1375	0.2961
$\alpha = 1.8$	0.0137	0.0730	0.2474
$\alpha = 2.0$	0.0515	0.0955	0.2550

CHAPTER 6

MULTIVARIATE SIMULATION STUDY

In this Chapter, we compare K-groups by first variation, K-groups by second variation and K-means on different multivariate mixture distributions. We want to see how the dimension affects the performance of K-means and both K-groups algorithms.

6.1 Simulation Design

Each mixture distribution is simulated with sample size $n = 200$ at different dimensions $d = 1, 2, 5, 10, 20, \text{and } 40$. We compute the average Diag, Kappa, Rand and cRand indices based on $B = 1000$ iterations. Table 6.1 displays mixture distributions we used in this Chapter.

The multivariate normal mixtures with mixing parameter p are simulated as follows:

1. Generate U from a uniform $(0, 1)$ distribution.

2. If $U < p$ generate X from a multivariate normal $\mathrm{N}(\mu_1, \Sigma_1)$; otherwise generate X from a multivariate normal$\mathrm{N}(\mu_2, \Sigma_2)$. Here μ is the mean vector and Σ is the covariance matrix. The multivariate normal mixture distributions will be denoted by $p\,\mathrm{N}\,(\mu_1, \Sigma_1) + (1-p)\,\mathrm{N}\,(\mu_2, \Sigma_2)$.

The multivariate Student T mixtures with mixing parameter p are simulated as follows:

1. Generate U from a uniform $(0, 1)$ distribution.

2. If $U < p$ generate X from a multivariate $\mathrm{T_d}(v)$; other wise generate X from a multivariate $\mathrm{T_d}(v) + \mu$. Here v is the degrees of freedom, μ is the location parameter. The multivariate Student T mixture distributions will be denoted by $p\,\mathrm{T_d}(v) + (1-p)(\mathrm{T_d}(v) + \mu)$.

The multivariate lognormal mixtures with mixing parameter p are simulated as follows:

1. Generate U from a uniform $(0, 1)$ distribution.

2. If $U < p$ generate X from a multivariate $\mathrm{N}(\mu_1, \Sigma_1)$ and take exponential transformation; otherwise generate X from a multivariate $\mathrm{N}(\mu_2, \Sigma_2)$ and take exponential transformation. Here the meaning of parameters μ and Σ are the same as multivariate normal. The multivariate lognormal mixture distributions will be denoted by $p\,\mathrm{Lognormal}(\mu_1, \Sigma_1) + (1-p)\,\mathrm{Lognormal}(\mu_2, \Sigma_2)$.

K-means performs well when we have spherical clusters. Based on the previous univariate and multivariate simulation results, K-groups algorithms also have good performance when clusters have spherical shape. We want to compare K-groups to K-means when we have cubic shaped clusters. A d-dimensional cubic is simulated as follows:

1. Generate U from a uniform $(0, 1)$ distribution.

2. For each coordinate, if $U < p$ generate X from a uniform (a_1, b_1); otherwise generate X from a uniform (a_2, b_2). Here parameters a and b are lower bound and upper bound for the uniform distribution.

3. Repeat step 2 d times. The mixture distribution will be denoted by $p\,\mathrm{Cubic}^d(a_1, b_2) + (1-p)\,\mathrm{Cubic}^d(a_2, b_2)$.

6.2 Simulation Result

This section presents the empirical results by comparing validation indices for K-means, K-groups by first variation, and K-groups by second variation in the multivariate cases. For each iteration, validation measures are computed for each method. The average validation indices are reported for each method over $B = 1000$ iterations.

Normal mixture distribution results

Table 6.2 and Table 6.3 summarize the simulation results of multivariate normal mixture dis-

tributions $0.5\,\mathrm{N_d}\,(0, I) + 0.5\,\mathrm{N_d}\,(3, I)$ and $0.5\,\mathrm{N_d}\,(0, I) + 0.5\,\mathrm{N_d}\,(0, 4I)$. For every dimension d, the average Rand and cRand indices of these three algorithms are very close in both Tables. For each algorithm, the average Rand and cRand indices for different dimension d are also very close. Thus, the results suggest that the both K-groups algorithms and K-means still perform almost the same when clusters are normally distributed in the multivariate case.

Student T mixture distribution results

Table 6.4 and Table 6.5 displays the simulation results of multivariate Student T mixture mixtures with 4 degrees of freedom $0.5\,\mathrm{T_d}\,(4) + 0.5(\,\mathrm{T_d}(4) + 3)$ and $0.5\,\mathrm{T_d}\,(4) + 0.5(\,\mathrm{T_d}\,(4) + 1)$. For every dimension d, the average Rand and cRand indices of these three algorithms are very close in both tables. For each algorithm, the average Rand and cRand indices for different dimension d are also very close. Thus, the results suggest that both K-groups algorithms and K-means have similar performance when clusters have lightly heavy tails in the multivariate case.

Table 6.6 and Table 6.7 shows simulation results of multivariate Student T mixtures with 2 degrees of freedom $0.5\,\mathrm{T_d}\,(2)+0.5(\,\mathrm{T_d}\,(2)+3)$ and $0.5\,\mathrm{T_d}\,(2)+0.5(\,\mathrm{T_d}\,(2)+1)$. For every dimension d, the average Rand and cRand indices of both K-groups algorithms are higher than K-means in both tables. For each algorithm, the average Rand and cRand indices for different dimension d are almost the same. Thus, the results suggest that both K-groups algorithms have better performances when clusters have strong heavy tails in the multivariate case.

Cubic mixture results

Table 6.8 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 6.1 displays the simulation result of cubic mixtures $0.5\,\mathrm{Cubic}^d\,(0, 1) + 0.5\,\mathrm{Cubic}^d\,(0.3, 0.7)$, 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.

Lognormal mixture distribution results

Table 6.9 and Table 6.10 displays the simulation results of multivariate lognormal distributions $0.5\,\mathrm{Lognormal}\,(0, I) + 0.5\,\mathrm{Lognormal}\,(3, I)$ and $0.5\,\mathrm{Lognormal}\,(0, I) + 0.5\,\mathrm{Lognormal}(0, 4I)$. For every dimension d, the average Rand and cRand indices of both K-groups algorithms are consistent higher than K-means. For each algorithm, the average Rand and cRand indices of different dimension d are almost the same. Thus, the results suggest that both K-groups algorithms have better performance when clusters are strongly skewed and heavy tailed in the multivariate cases.

Table 6.1: Multivariate Mixture Distributions

Distribution	Skewness	Kurtosis
$0.5\,\mathrm{N_d}\,(0, I) + 0.5\,\mathrm{N_d}\,(3, I)$	Symmetric	Normal tails
$0.5\,\mathrm{N_d}\,(0, I) + 0.5\,\mathrm{N_d}\,(0, 4I)$	Symmetric	Normal tails
$0.5\,\mathrm{T_d}\,(4) + 0.5(\,\mathrm{T_d}\,(4) + 3)$	Symmetric	Heavy tails
$0.5\,\mathrm{T_d}\,(4)^{(d)} + 0.5(\,\mathrm{T_d}\,(4) + 1)$	Symmetric	Heavy tails
$0.5\,\mathrm{T_d}\,(2)^{(d)} + 0.5(\,\mathrm{T_d}\,(2) + 3)$	Symmetric	Heavy tails
$0.5\,\mathrm{T_d}\,(2)^{(d)} + 0.5(\,\mathrm{T_d}\,(2) + 1)$	Symmetric	Heavy tails
$0.5\,\mathrm{Cubic}^d\,(0, 1) + 0.5\,\mathrm{Cubic}^d\,(0.3, 0.7)$	Symmetric	Heavy tails
$0.5\,\mathrm{Lognormal}\,(0, I) + 0.5\,\mathrm{Lognormal}\,(3, I)$	Large skewness	Heavy tails
$0.5\,\mathrm{Lognormal}\,(0, I) + 0.5\,\mathrm{Lognormal}\,(0, 4I)$	Large skewness	Heavy tails

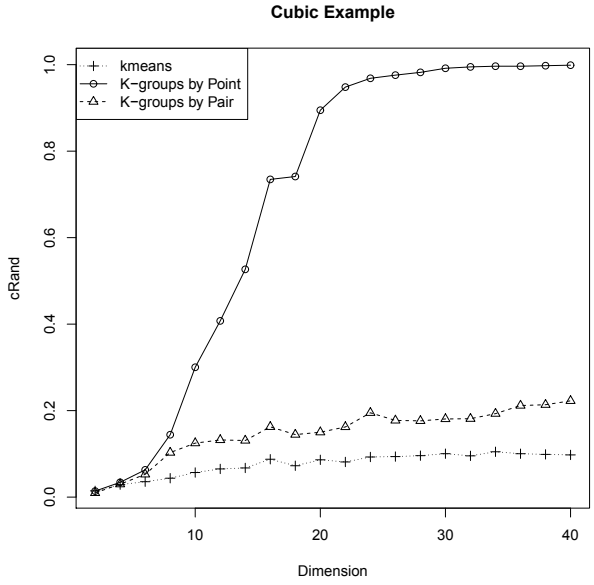


Figure 6.1: Multivariate cubic mixtures, $d = 2, 4, ..., 40,\ n = 200,\ B = 500$

Table 6.2: Normal Mixture Distribution $0.5\,\mathrm{N_d}\,(0, I) + 0.5\,\mathrm{N_d}\,(3, I), \alpha = 1$

Method	d	Diag	Kappa	Rand	cRand
K-means	1	0.9321	0.8635	0.8735	0.7469
K-groups Point	1	0.9932	0.8638	0.8736	0.7473
K-groups Pair	1	0.9303	0.8600	0.8703	0.7407
K-means	2	0.9371	0.8737	0.8820	0.7641
K-groups Point	2	0.9376	0.8748	0.8830	0.7661
K-groups Pair	2	0.9349	0.8693	0.8763	0.7565
K-means	5	0.9337	0.8669	0.8763	0.7527
K-groups Point	5	0.9936	0.8667	0.8762	0.7585
K-groups Pair	5	0.9931	0.8657	0.8753	0.7507
K-means	10	0.9285	0.8565	0.8671	0.7343
K-groups Point	10	0.9271	0.8538	0.8648	0.7296
K-groups Pair	10	0.9257	0.8510	0.8624	0.7248
K-means	20	0.9340	0.8674	0.8767	0.7534
K-groups Point	20	0.9331	0.8657	0.8751	0.7503
K-groups Pair	20	0.9293	0.8580	0.8685	0.7370
K-means	40	0.9347	0.8688	0.8780	0.7560
K-groups Point	40	0.9342	0.8678	0.8771	0.7542
K-groups Pair	40	0.9312	0.8618	0.8719	0.7438

Table 6.3: Normal Mixture Distribution $0.5\,\mathrm{N_d}\,(0, I) + 0.5\,\mathrm{N_d}\,(0, 4I), \alpha = 1$

Method	d	Diag	Kappa	Rand	cRand
K-means	1	0.5400	0.0817	0.5024	0.0048
K-groups Point	1	0.5340	0.0684	0.5012	0.0025
K-groups Pair	1	0.5361	0.0719	0.5017	0.0034
K-means	2	0.5366	0.0852	0.5028	0.0056
K-groups Point	2	0.5326	0.0783	0.5022	0.0044
K-groups Pair	2	0.5328	0.0758	0.5015	0.0031
K-means	5	0.5385	0.0704	0.5014	0.0029
K-groups Point	5	0.5378	0.0632	0.5007	0.0015
K-groups Pair	5	0.5372	0.0637	0.5008	0.0016
K-means	10	0.5398	0.0765	0.5023	0.0047
K-groups Point	10	0.5375	0.0758	0.5018	0.0037
K-groups Pair	10	0.5372	0.0752	0.5017	0.0034
K-means	20	0.5398	0.0791	0.5021	0.0043
K-groups Point	20	0.5375	0.0738	0.5017	0.0034
K-groups Pair	20	0.5355	0.0694	0.5015	0.0031
K-means	40	0.5381	0.0727	0.5021	0.0043
K-groups Point	40	0.5348	0.0679	0.5013	0.0026
K-groups Pair	40	0.5346	0.0684	0.5013	0.0026

Table 6.4: Student T Mixture Distribution $0.5\,\mathrm{T_d}\,(4) + 0.5(\,\mathrm{T_d}\,(4) + 3), \alpha = 1$

Method	d	Diag	Kappa	Rand	cRand
K-means	1	0.8965	0.7914	0.8147	0.6295
K-groups Point	1	0.8940	0.7865	0.8107	0.6215
K-groups Pair	1	0.8925	0.7832	0.8081	0.6163
K-means	2	0.9045	0.8080	0.8270	0.5867
K-groups Point	2	0.9015	0.8022	0.8220	0.5924
K-groups Pair	2	0.8950	0.7892	0.8114	0.5765
K-means	5	0.8830	0.7658	0.7933	0.5843
K-groups Point	5	0.8850	0.7698	0.7962	0.5945
K-groups Pair	5	0.8795	0.7589	0.7882	0.5809
K-means	10	0.8825	0.7640	0.7921	0.5843
K-groups Point	10	0.8860	0.7709	0.7972	0.5945
K-groups Pair	10	0.8815	0.7615	0.7904	0.5809
K-means	20	0.9000	0.7982	0.8202	0.6405
K-groups Point	20	0.8985	0.7953	0.8179	0.6358
K-groups Pair	20	0.8965	0.7914	0.8145	0.6291
K-means	40	0.8915	0.7810	0.8060	0.6120
K-groups Point	40	0.8900	0.7789	0.8043	0.6086
K-groups Pair	40	0.8855	0.7699	0.7976	0.5952

Table 6.5: Student T Mixture Distribution $0.5\,\mathrm{T_d}\,(4) + 0.5(\,\mathrm{T_d}\,(4) + 1), \alpha = 1$

Method	d	Diag	Kappa	Rand	cRand
K-means	1	0.6715	0.3443	0.5587	0.1175
K-groups Point	1	0.6750	0.3540	0.5608	0.1217
K-groups Pair	1	0.6740	0.3522	0.5597	0.1194
K-means	2	0.6675	0.3367	0.5558	0.1119
K-groups Point	2	0.6715	0.3438	0.5584	0.1169
K-groups Pair	2	0.6725	0.3447	0.5592	0.1184
K-means	5	0.6685	0.3342	0.5563	0.1128
K-groups Point	5	0.6645	0.3279	0.5538	0.1077
K-groups Pair	5	0.6670	0.3315	0.5562	0.1124
K-means	10	0.6820	0.3622	0.5649	0.1296
K-groups Point	10	0.6925	0.3842	0.5740	0.1479
K-groups Pair	10	0.6915	0.3832	0.5738	0.1474
K-means	20	0.6695	0.3331	0.5588	0.1175
K-groups Point	20	0.6810	0.3586	0.5655	0.1311
K-groups Pair	20	0.6800	0.3577	0.5646	0.1293
K-means	40	0.6730	0.3468	0.5615	0.1233
K-groups Point	40	0.6900	0.3795	0.5698	0.1397
K-groups Pair	40	0.6930	0.3868	0.5722	0.1445

Table 6.6: Student T Mixture Distribution $0.5\,\mathrm{T_d}\,(2) + 0.5(\,\mathrm{T_d}\,(2) + 3), \alpha = 1$

Method	d	Diag	Kappa	Rand	cRand
K-means	1	0.8520	0.7021	0.7478	0.4956
K-groups Point	1	0.8585	0.7154	0.7567	0.5134
K-groups Pair	1	0.8540	0.7064	0.7504	0.5008
K-means	2	0.8440	0.6880	0.7374	0.4749
K-groups Point	2	0.8527	0.7047	0.7487	0.4974
K-groups Pair	2	0.8532	0.7056	0.7497	0.4993
K-means	5	0.8487	0.6855	0.7406	0.4676
K-groups Point	5	0.8540	0.7076	0.7504	0.5008
K-groups Pair	5	0.8575	0.7031	0.7470	0.4941
K-means	10	0.8340	0.6670	0.7335	0.4673
K-groups Point	10	0.8575	0.7147	0.7555	0.5110
K-groups Pair	10	0.8575	0.7147	0.7555	0.5111
K-means	20	0.8307	0.6511	0.7308	0.4647
K-groups Point	20	0.8655	0.7302	0.7672	0.5345
K-groups Pair	20	0.8622	0.7237	0.7625	0.5250
K-means	40	0.8325	0.6571	0.7386	0.4772
K-groups Point	40	0.8677	0.7349	0.7698	0.5397
K-groups Pair	40	0.8640	0.7273	0.7643	0.5287

Table 6.7: Student T Mixture Distribution $0.5\,\mathrm{T_d}\,(2) + 0.5(\,\mathrm{T_d}\,(2) + 1), \alpha = 1$

Method	d	Diag	Kappa	Rand	cRand
K-means	1	0.5911	0.1620	0.5241	0.0486
K-groups Point	1	0.6646	0.3306	0.5542	0.1084
K-groups Pair	1	0.6660	0.3332	0.5549	0.1099
K-means	2	0.5927	0.1660	0.5233	0.0466
K-groups Point	2	0.6639	0.3280	0.5540	0.1080
K-groups Pair	2	0.6624	0.3244	0.5529	0.1057
K-means	5	0.5905	0.1641	0.5214	0.0428
K-groups Point	5	0.6582	0.3165	0.5494	0.0987
K-groups Pair	5	0.6575	0.3145	0.5489	0.0978
K-means	10	0.6005	0.1788	0.5260	0.0514
K-groups Point	10	0.6605	0.3207	0.5516	0.1032
K-groups Pair	10	0.6628	0.3246	0.5532	0.1064
K-means	20	0.5954	0.1702	0.5255	0.0509
K-groups Point	20	0.6656	0.3313	0.5547	0.1094
K-groups Pair	20	0.6639	0.3279	0.5536	0.1073
K-means	40	0.5977	0.1779	0.5250	0.0497
K-groups Point	40	0.6601	0.3198	0.5514	0.1029
K-groups Pair	40	0.6582	0.3170	0.5501	0.1003

Table 6.8: 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

Table 6.9: Lognormal Mixture Distribution $0.5\,\mathrm{Lognormal}\,(0, I) + 0.5\,\mathrm{Lognormal}\,(3, I)$,$\alpha = 1$

Method	d	Diag	Kappa	Rand	cRand
K-means	1	0.5868	0.1607	0.5176	0.0353
K-groups Point	1	0.7194	0.4381	0.6006	0.2012
K-groups Pair	1	0.7209	0.4418	0.6027	0.2054
K-means	2	0.5855	0.1529	0.5162	0.0032
K-groups Point	2	0.7209	0.4418	0.6023	0.2045
K-groups Pair	2	0.7237	0.4466	0.6039	0.2077
K-means	5	0.5861	0.1613	0.5188	0.0380
K-groups Point	5	0.7221	0.4473	0.6031	0.2065
K-groups Pair	5	0.7204	0.4432	0.6017	0.2037
K-means	10	0.5859	0.1577	0.5184	0.0356
K-groups Point	10	0.7331	0.4635	0.6131	0.2258
K-groups Pair	10	0.7349	0.4667	0.6142	0.2281
K-means	20	0.5866	0.1601	0.5176	0.0356
K-groups Point	20	0.7319	0.4610	0.6122	0.2246
K-groups Pair	20	0.7338	0.4651	0.6136	0.2275
K-means	40	0.5872	0.1663	0.5170	0.0341
K-groups Point	40	0.7137	0.4304	0.5943	0.1888
K-groups Pair	40	0.7182	0.4395	0.5992	0.1986

Table 6.10: Lognormal Mixture Distribution $0.5\mathrm{Lognormal}(0, I) + 0.5\mathrm{Lognormal}(0, 4I)$,$\alpha = 1$

Method	d	Diag	Kappa	Rand	cRand
K-means	1	0.5340	0.0232	0.5010	0.0023
K-groups Point	1	0.5535	0.0897	0.5057	0.0117
K-groups Pair	1	0.5559	0.0949	0.5065	0.0132
K-means	2	0.5329	0.0202	0.5010	0.0008
K-groups Point	2	0.5554	0.0868	0.5067	0.0126
K-groups Pair	2	0.5565	0.0889	0.5068	0.0128
K-means	5	0.5321	0.0179	0.5009	0.0012
K-groups Point	5	0.5513	0.0805	0.5052	0.0098
K-groups Pair	5	0.5526	0.0829	0.5054	0.0102
K-means	10	0.5378	0.0199	0.5017	0.0025
K-groups Point	10	0.5549	0.0843	0.5060	0.0115
K-groups Pair	10	0.5580	0.0909	0.5068	0.0131
K-means	20	0.5318	0.0190	0.5006	0.0014
K-groups Point	20	0.5529	0.0853	0.5055	0.0113
K-groups Pair	20	0.5542	0.0871	0.5056	0.0116
K-means	40	0.5378	0.0207	0.5020	0.0031
K-groups Point	40	0.5592	0.0958	0.5070	0.0133
K-groups Pair	40	0.5607	0.1008	0.5074	0.0143

CHAPTER 7

REAL DATA EXAMPLES

In this chapter, we will investigate the application of K-groups algorithm using three real data sets.

7.1 Classification of Wines Cultivars

The Wine data analyzed in this application is publicly available from the UCI Machine Learning Repository (Forina, Armanino, Leardi, and Drava, 1991) at ftp.ics.uci.edu. The data was used in a comparison of classifiers in high dimensional setting.

The data are the results of a chemical analysis of wines grown in the same region in Italy, but derived from three different cultivars. The data consists of 178 instances and 13 attributes. All attributes are summarized in Table 7.1. The standard deviations of attributes Magnesium and Proline are 14.28 and 314.90, respectively, which are much larger than the standard deviations of other attributes. In order to have attributes on a common scale, we will standardize all the attributes to mean zero and unit standard deviation, and do the clustering analysis on the standardized data.

Table 7.4 shows the clustering results of K-means, K-groups by first variation, K-groups by second variation, and Hierarchical ξ. The maximum Rand and cRand index values 0.9918 and 0.9816 are obtained from K-groups by second variation. The second largest Rand and cRand index values 0.9620 and 0.9148 are obtained from K-groups by first variation. K-means also obtains very good Rand and cRand index values 0.9542 and 0.8974. The hierarchical ξ obtains the smallest Rand and cRand index value among those four algorithms.

Table 7.5, 7.6, 7.7 and Table 7.8 are agreement tables between the known partition and the partition determined by clustering algorithms. Note that K-groups by second variation correctly

Table 7.1: Wine Data Summary

Attributes	Property of attributes	Mean	Standard deviation
1.Alcohol	Numerical	13.00	0.81
2.Malic acid	Numerical	2.33	1.11
3.Ash	Numerical	2.36	0.27
4.Alcalinity of ash	Numerical	19.49	3.33
5.Magnesium	Numerical	99.74	14.28
6.Total phenols	Numerical	2.29	0.62
7.Flavanoids	Numerical	2.02	0.99
8.Nonflavanoid phenols	Numerical	0.36	0.12
9.Proanthocyanins	Numerical	1.59	0.57
10.Color intensity	Numerical	5.05	2.31
11.Hue	Numerical	0.95	0.22
12.OD280/OD315 of diluted wines	Numerical	2.61	0.70
13.Proline	Numerical	746.89	314.90

classifies all cases of cultivar I, cultivar III, and misclassifies one observation of cultivar II as cultivar I. K-groups by first variation algorithm is less successful at recovering the clusters, misclassifies two observations of cultivar II as cultivar I, and three observations of cultivar II as cultivar III. The K-means algorithm misclassifies three observations of cultivar II observations as cultivar I, and three observations of cultivar II as cultivar III. The Hierarchical ξ algorithm performs worse than the other three algorithms, misclassifies one observation of cultivar I as cultivar II, seven observations of cultivar II as cultivar III, and six observations of cultivar II as cultivar III. From the result above, we find that the majority of misclassifications occur in cultivar II.

Figure 7.1(a)–Figure 7.1(e) are the 2-D plots of true (expert labeled) classification, K-groups by first variation, K-groups by second variation, K-means, and Hierarchical ξ.

7.2 Diagnosis of Erythemato-Squamous Diseases in Dermatology

The dermatology data analyzed is publicly available from the UCI Machine Learning Repository (Blake and Merz, 1998) at ftp.ics.uci.edu. The data was analyzed by Güvenir, Demiröz, and Ilter (1998), and contributed by Güvenir. The erythemato-squamous dieases are proriasis, seboreic dermatitis, lichen planus, pityriasis rosea, choronic dermatitis and pityriasis rubra pilaris. According

to Güvenir et al. (1998), 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 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 7.2. We standardize all the attributes to zero mean and unit standard deviation and delete the observations which contain the missing values. The effective data size is 358 in the clustering analysis.

Table 7.2: 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

Table 7.9 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 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 7.10, 7.11, 7.12, and Table 7.13 are agreement tables between the known partition and the partition determined by clustering algorithms. Note that K-groups by first variation correctly classifies psoriasis, lichen planus, pityriasis rubra pilaris, and misclassifies two psoriasis as seboreic dermatitis, 12 seboreic dermatitis as pityriasis rosea, one pityriasis rosea as seboreic dermatitis, and one chronic dermatitis as pityriasis rosea. Hierarchical ξ correctly classifies psoriasis, pityriasis rubra pilaris, and misclassifies 13 seboreic dermatitis as pityriasis rosea, one lichen planus as chronic dermatitis, two seboreic pityriasis rosea as seboreic dermatitis, and two chronic dermatitis as pityriasis rosea. K-groups by second variation correctly classifies chronic dermatitis, pityriasis rubra pilaris, misclassifies one psoriasis as seboreic dermatitis, 35 seboreic dermatitis as pityriasis rosea, one lichen planus as pityriasis rosea, and two pityriasis rosea as seboreic dermatitis. K-means correctly classifies seboreic dermatitis, lichen planus, and misclassifies two psoriasis as seboreic dermatitis, 48 pityriasis rosea as seboreic dermatitis, one chronic dermatitis as seboreic dermatitis, and nine pityriasis rubra pilaris as pityriasis rosea. From the result above, most misclassifications occur between seboreic dermatitis and pityriasis rosea, which means these two groups of disease show very similar clinical features.

Figure 7.2(a)–Figure7.2(e) are the 3-D plots of true (expert labeled) classification, K-groups by first variation, K-groups by second variation, K-means and Hierarchical ξ.

7.3 Diagnosis of Breast Cancer in Oncology

The Breast cancer data analyzed is publicly available from the UCI Machine learning Repository ftp.ics.uci.edu. The data was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. Samples arrived periodically as Dr. Wolberg reported his clinical cases.

The database therefore reflects this chronological grouping of the data.

The data consists of 699 objects with 10 attributes. All other attributes except sample id take integer value in $1, ..., 10$. There are 16 missing values. All of the attributes are summarized in Table 7.3. The first attribute is sample id, which does not carry any information about the breast cancer. Thus nine attributes are used to analyze the data. Since all nine attributes have the same range, there is no need to standardize them. Cases with missing values are deleted, so the effective data size is 683 in the clustering analysis.

Table 7.3: Breast Cancer Data Summary

Attributes	Property of attributes	Mean	Standard deviation
1. sample code number	Numerical		
2. clump thickness	Numerical	4.44	2.82
3. uniformity of cell size	Numerical	3.15	3.06
4. uniformity of cell shape	Numerical	3.21	2.98
5. marginal adhesion	Numerical	2.83	2.86
6. single epithelial cell size	Numerical	3.23	2.22
7. bare nuclei	Numerical	3.21	2.15
8. bland chromatin	Numerical	3.44	2.44
9. normal nucleoli	Numerical	2.86	3.05
10. mitoses	Numerical	1.60	1.73

Table 7.14 shows the clustering result of K-means, K-groups by first variation, K-groups by second variation, and Hierarchial ξ. K-groups by first variation and K-groups by second variation perform the same, which have the largest Rand and cRand index values 0.9239 and 0.8467. The Hierarchical ξ performs slightly worse than K-groups algorithms with Rand and cRand index values 0.9212 and 0.8417. K-means obtains the smallest Rand and adjusted Rand index values 0.9132 and 0.8246 among these four algorithms.

Table 7.15, 7.16, 7.17, and Table 7.18 are agreement tables between the known partition and the partition determined by clustering algorithms. Note that K-groups by first variation correctly classifies the majority of observations, but misclassifies 14 observations of benign as malignant, and 13 observations of malignant as benign. The classification table of K-groups by second variation is exactly the same as K-groups by first variation. Hierarchical ξ misclassifies 24 observations

of benign as malignant, and only four observations of malignant as benign. K-means misclassifies nine observations of benign as malignant, 22 observations of malignant as benign.

Figure 7.3(a)–Figure 7.3(e) are the 2-D plots of true (expert labeled) classification, K-groups by first variation, K-groups by second variation, K-means and Hierarchical ξ. From the figures, one can see that even though the K-groups by first variation and K-groups by second variation have the same Rand and cRand index values, the 2-D plots of these two methods shows slight differences.

Table 7.4: Wine Data Results

Indices	K-means	K-groups Point	K-groups Pair	Hierarchical ξ
Diag	0.9662	0.9719	0.9943	0.9213
Kappa	0.9490	0.9575	0.9914	0.8817
Rand	0.9542	0.9620	0.9918	0.8980
cRand	0.8974	0.9148	0.9816	0.7711

Table 7.5: Classification of Wine Data by K-means

Class	1	2	3	Cases
Cultivar I	59	3	0	62
Cultivar II	0	65	0	65
Cultivar III	0	3	48	51
Total	59	71	48	178

Table 7.6: Classification of Wine Data by K-groups Point

Class	1	2	3	Cases
Cultivar I	59	2	0	61
Cultivar II	0	66	0	66
Cultivar III	0	3	48	51
Total	59	71	48	178

Table 7.7: Classification of Wine Data by K-groups Pair

Class	1	2	3	Cases
Cultivar I	59	1	0	60
Cultivar II	0	70	0	70
Cultivar III	0	0	48	48
Total	59	71	48	178

Table 7.8: Classification of Wine Data by Hierarchical ξ

Class	1	2	3	Cases
Cultivar I	58	7	0	65
Cultivar II	1	58	0	59
Cultivar III	0	6	48	54
Total	59	71	48	178

Table 7.9: 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

Table 7.10: Classification of Dermatology Data by K-means

Class	1	2	3	4	5	6	Cases
1. psoriasis	109	0	0	0	0	0	106
2. seboreic dermatitis	2	60	0	48	1	0	111
3. lichen planus	0	0	71	0	0	0	71
4. pityriasis rosea	0	0	0	0	0	9	9
5. chronic dermatitis	0	0	0	0	47	0	47
6. pityriasis rubra pilaris	0	0	0	0	0	11	11
Total	111	60	71	48	48	20	358

Table 7.11: Classification of Dermatology Data by K-groups Point

Class	1	2	3	4	5	6	Cases
1. psoriasis	109	0	0	0	0	0	109
2. seboreic dermatitis	2	48	0	1	0	0	51
3. lichen planus	0	0	71	0	0	0	71
4. pityriasis rosea	0	12	0	47	1	0	60
5. chronic dermatitis	0	0	0	0	47	0	47
6. pityriasis rubra pilaris	0	0	0	0	0	20	20
Total	111	62	71	48	48	20	358

Table 7.12: Classification of Dermatology Data by K-groups Pair

Class	1	2	3	4	5	6	Cases
1. psoriasis	110	0	0	0	0	0	110
2. seboreic dermatitis	1	25	0	2	0	0	28
3. lichen planus	0	0	70	0	0	0	70
4. pityriasis rosea	0	35	1	46	0	0	82
5. chronic dermatitis	0	0	0	0	48	0	48
6. pityriasis rubra pilaris	0	0	0	0	0	20	20
Total	111	60	71	48	48	20	358

Table 7.13: Classification of Dermatology Data by Hierarchical ξ

Class	1	2	3	4	5	6	Cases
1. psoriasis	111	0	0	0	0	0	111
2. seboreic dermatitis	0	47	0	2	0	0	49
3. lichen planus	0	0	70	0	0	0	70
4. pityriasis rosea	0	13	0	46	2	0	61
5. chronic dermatitis	0	0	1	0	46	0	47
6. pityriasis rubra pilaris	0	0	0	0	0	20	20
Total	111	60	71	48	48	20	358

Table 7.14: Breast Cancer Data Results

Indices	K-means	K-groups Point	K-groups Pair	Hierarchical ξ
Diag	0.9546	0.9604	0.9604	0.9590
Kappa	0.8989	0.9131	0.9131	0.9115
Rand	0.9132	0.9239	0.9239	0.9212
cRand	0.8246	0.8467	0.8467	0.8417

Table 7.15: Classification of Breast Cancer Data by K-means

Class	1	2	Cases
1.benign	435	22	457
2.malignant	9	217	226
Total	444	239	683

Table 7.16: Classification of Breast Cancer Data by K-groups by Point

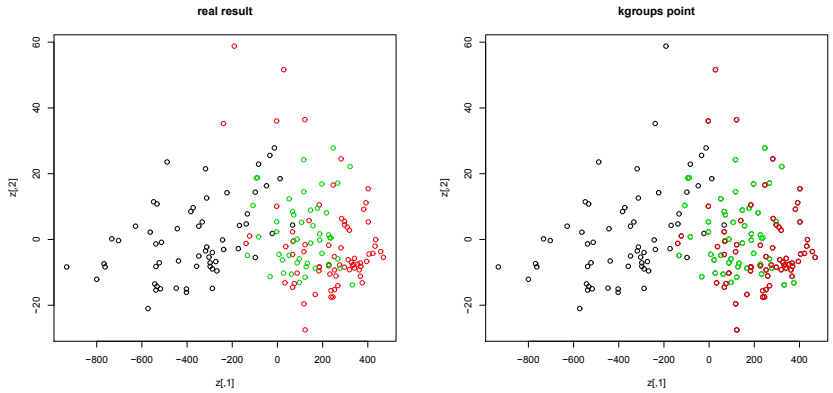
Class	1	2	Cases
1.benign	430	13	443
2.malignant	14	226	240
Total	444	239	683

Table 7.17: Classification of Breast Cancer Data by K-groups by Pair

Class	1	2	Cases
1.benign	430	13	443
2.malignant	14	226	240
Total	444	239	683

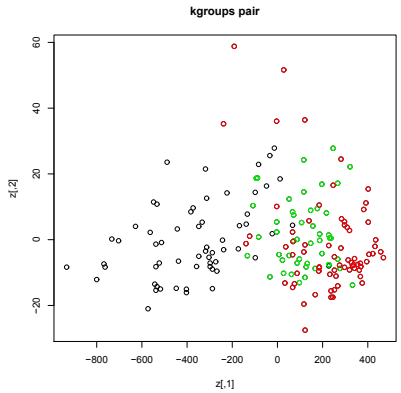
Table 7.18: Classification of Breast Cancer Data by Hierarchical ξ

Class	1	2	Cases
1.benign	420	4	424
2.malignant	24	235	259
Total	444	239	683

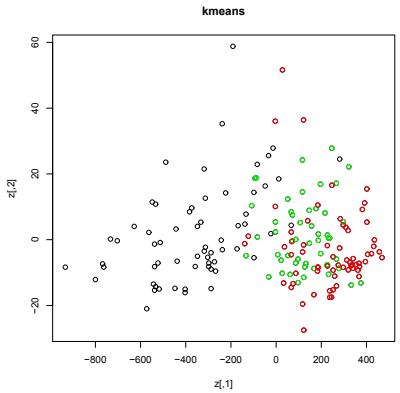


(a)

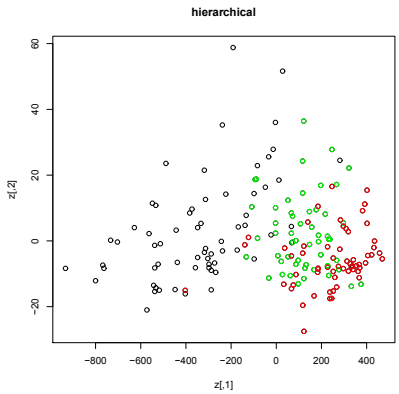
(b)



(c)

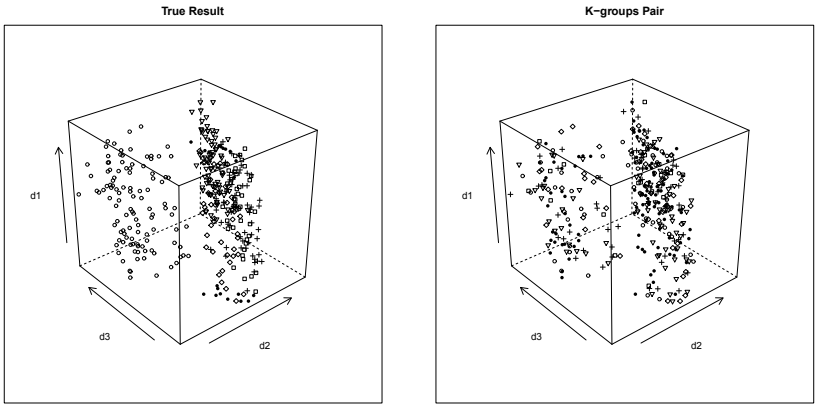


(d)



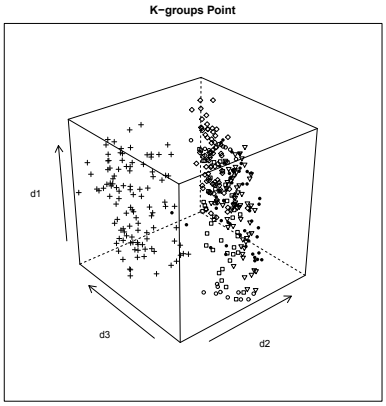
(e)

Figure 7.1: 7.1(a)–7.1(e) are wine data 2-D plots on the first two principal components axes

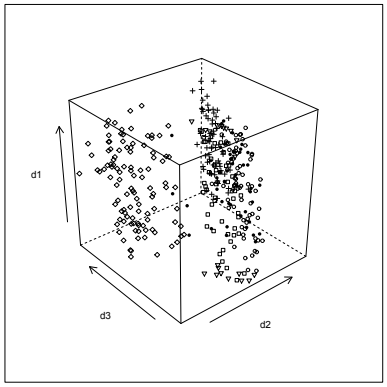


(a)

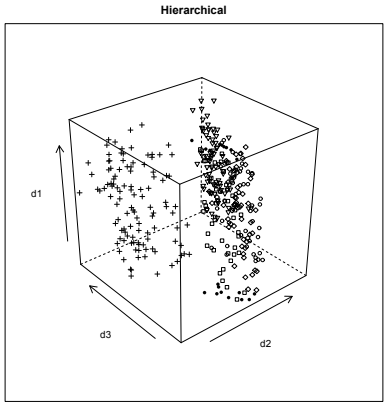
(b)



(c)

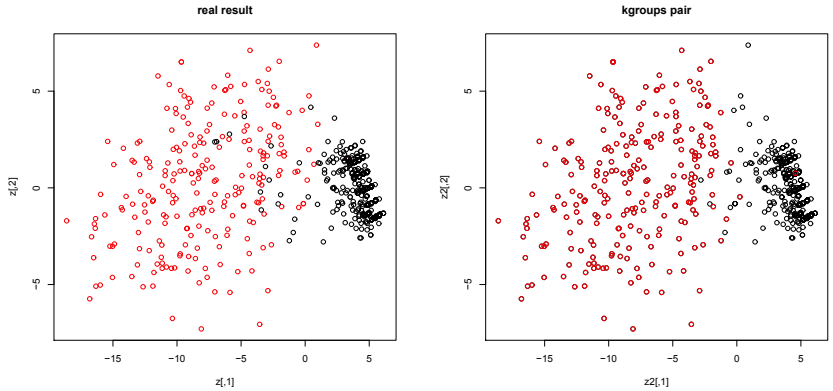


(d)



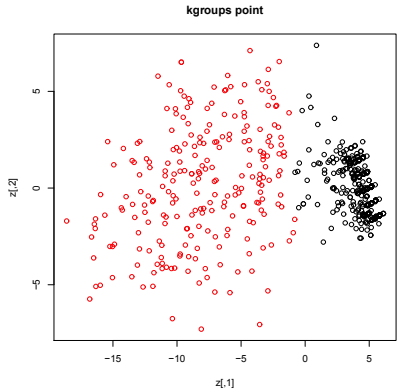
(e)

Figure 7.2: 7.2(a)–7.2(e) are dermatology data 3-D plots on the first three principal components axes

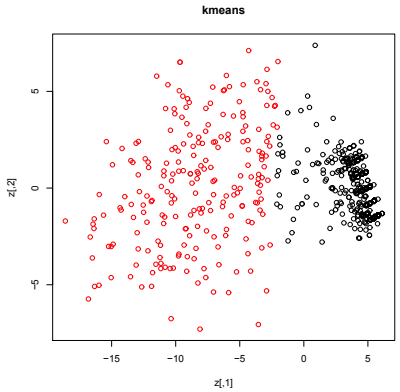


(a)

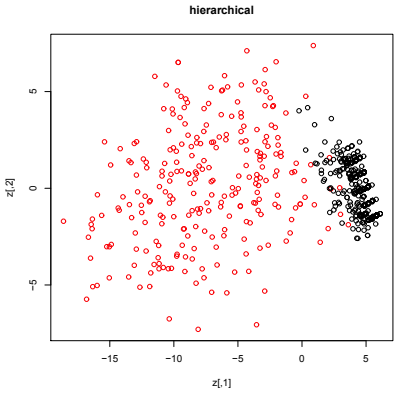
(b)



(c)



(d)



(e)

Figure 7.3: 7.3(a)–7.3(e) are breast cancer data 2-D plots on the first two principal components axes

CHAPTER 8

SUMMARY

In this dissertation, a new class of clustering algorithms, K-groups is proposed based on the energy distance. The new approach is a distribution-based clustering algorithm, which assigns observations to the same cluster if they follow the identical statistical distribution. This kind of algorithm will perform better when clusters are irregular, intertwined, or when noise and outliers are present. We propose two different K-groups algorithms: K-groups by first variation and K-groups by second variation, based on the updating formulas derived in Chapter 3 and Chapter 4, and generalize 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 the more general K-groups by first variation when $\alpha = 2$.

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. However, the unbalanced cluster sizes simulations suggest that both K-groups algorithms perform worse than K-means when clusters are normally distributed. 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 unsuper-

vised 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.

8.1 Improve computational complexity

As we discussed in Chapter 4, 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.

8.2 Apply K-groups algorithm to handle big data

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.

8.3 Apply K-groups algorithm to semi-supervised clustering problems

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.

8.4 Extension of K-groups to random variables taking value in Hilbert spaces

As we mention in Chapter 2, energy distance is a functional distance between the characteristic functions of two independent random variables X and Y. Lyons (2013) 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

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

Baringhaus, L. and C. Franz (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis 88*(1), 190–206.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers.

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

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics 7*(3), 200–217.

Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal 1928*(1), 13–74.

David, H. (1968). Miscellanea: Gini's mean difference rediscovered. *Biometrika 55*(3), 573–575.

Dhillon, I. S., Y. Guan, and J. Kogan (2002). Iterative clustering of high dimensional text data augmented by local search. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pp. 131–138. IEEE.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, Volume 96, pp. 226–231.

Forina, M., C. Armanino, R. Leardi, and G. Drava (1991). A class-modelling technique based on potential functions. *Journal of Chemometrics 5*(5), 435–453.

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.

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

Hinneburg, A. and D. A. Keim (1998). An efficient approach to clustering in large multimedia databases with noise. In *KDD*, Volume 98, pp. 58–65.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist 11*(2), 37–50.

Jain, A. K. and R. C. Dubes (1988). *Algorithms for Clustering Data*. Prentice-Hall, Inc.

Jarvis, R. A. and E. A. Patrick (1973). Clustering using a similarity measure based on shared near neighbors. *Computers, IEEE Transactions on 100*(11), 1025–1034.

Karypis, G., E.-H. Han, and V. Kumar (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer 32*(8), 68–75.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE 78*(9), 1464–1480.

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

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

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

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, pp. 281–297.

McLachlan, G. J. and K. E. Basford (1988). Mixture models. inference and applications to clustering. *Statistics: Textbooks and Monographs, New York: Dekker, 1988 1*.

Meilă, M. (2003). Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines*, pp. 173–187. Springer.

Milligan, G. W. (1996). *Clustering Validation: Results and Implications for Applied Analyses*.

World Scientific.

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

Pollaczek-Geiringer, H. (1928). Statistik seltener ereignisse. *Naturwissenschaften 16*(43), 800–807.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association 66*(336), 846–850.

Rizzo, M. L. (2002). A test of homogeneity for two multivariate populations. *Proceedings of the American Statistical Association, Physical and Engineering Sciences Section [CD-ROM], American Statistical Association, Alexandria, VA*.

Rizzo, M. L. (2009). New goodness-of-fit tests for Pareto distributions. *ASTIN Bulletin: Journal of the International Association of Actuaries 39*(2), 691–715.

Rizzo, M. L. and G. J. Székely (2010). DISCO analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics 4*(2), 1034–1055.

Steinbach, M., G. Karypis, and V. Kumar (2000). A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, Volume 400, pp. 525–526. Boston.

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

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

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

Székely, G. J. and M. L. Rizzo (2005a). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of Classification 22*(2), 151–183.

Székely, G. J. and M. L. Rizzo (2005b). A new test for multivariate normality. *Journal of Multivariate Analysis 93*(1), 58–80.

Tan, P.-N., M. Steinbach, and V. Kumar (2006). Cluster analysis: basic concepts and algorithms.

Introduction to Data Mining, 487–568.

Van Leeuwen, J. (1990). *Handbook of Theoretical Computer Science, Vol B: Formal Models and Semantics*, Volume 137. Elsevier.

Ward, Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*(301), 236–244.

Zhao, Y. and G. Karypis (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning 55*(3), 311–331.