TWO-MODE PROBABILISTIC DISTANCE CLUSTERING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

YAĞMUR CANER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
INDUSTRIAL ENGINEERING

JULY 2021

Approval of the thesis:

TWO-MODE PROBABILISTIC DISTANCE CLUSTERING

submitted by **YAĞMUR CANER** in partial fulfillment of the requirements for the degree of **Master of Science in Industrial Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Esra Karasakal
Head of Department, **Industrial Engineering** _____

Prof. Dr. Cem İyigün
Supervisor, **Industrial Engineering, METU** _____

Examining Committee Members:

Prof. Dr. Sinan Gürel
Industrial Engineering, METU _____

Prof. Dr. Cem İyigün
Industrial Engineering, METU _____

Prof. Dr. Pınar Karagöz
Computer Engineering, METU _____

Assist. Prof. Dr. Sakine Batun
Industrial Engineering, METU _____

Assist. Prof. Dr. Fatma Yerlikaya Özkurt
Industrial Engineering, Atılım University _____

Date: 29.07.2021

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Yağmur Caner

Signature :

ABSTRACT

TWO-MODE PROBABILISTIC DISTANCE CLUSTERING

Caner, Yağmur
M.S., Department of Industrial Engineering
Supervisor: Prof. Dr. Cem İyigün

July 2021, 112 pages

Probabilistic Distance Clustering (PDC) is a soft clustering technique constructed around some axioms. It is a center-based approach and assigns each data point to multiple clusters with a membership probability. The PDC is applicable for one-mode data sets, where each data points' quantitative or qualitative values over each feature are stored.

This study focuses on PDC and consists of two main contributions. Firstly, the relevance of PDC to some other probabilistic models in the literature is examined. We show that PDC method and its axioms explain models from marketing, location theory, and unsupervised learning. Secondly, this thesis proposes two original solution methods for the soft Two-Mode Clustering (TMC) problem. Two-mode clustering is a technique to cluster two-mode data, representing a linkage between two sets of data points. A comprehensive computational study is conducted on continuous, noisy, and binary data sets. The use of membership probabilities for decision-making is also discussed. This study will be the pioneer soft assignment approach for two-mode clustering literature.

Keywords: clustering, probabilistic clustering, one-mode data, location theory, two-mode data clustering

ÖZ

ÇİFT MODLU OLASILIKSAL MESAFE KÜMELEMESİ

Caner, Yağmur
Yüksek Lisans, Endüstri Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Cem İyigün

Temmuz 2021 , 112 sayfa

Olasılıksal Mesafe Kümelemesi (PDC), bazı aksiyomlar etrafında oluşturulmuş yumuşak bir kümeleme tekniğidir. Merkez tabanlı bir yaklaşımdır ve her bir veri noktasını üyelik olasılığı ile birden çok kümeye atar. PDC, her veri noktasının her bir özellik üzerindeki nicel veya nitel değerlerini saklayan tek modlu veri kümeleri için geçerlidir.

Bu çalışma PDC'ye odaklanır ve iki ana katkıdan oluşur. İlk olarak, PDC'nin literatürdeki diğer bazı olasılıksal modellerle ilgisi incelenmiştir. PDC yönteminin ve aksiyomlarının pazarlama, konum teorisi ve gözetimsiz öğrenme modellerini açıkladığını gösterilmiştir. İkinci olarak, bu tez yumuşak Çift Modlu Kümeleme (TMC) problemi için iki orijinal çözüm yöntemi önermektedir. Çift modlu kümeleme, iki veri noktası kümesi arasındaki bağlantıyı temsil eden çift modlu verileri kümeleme tekniğidir. Sürekli, gürültülü ve ikili veri kümeleri üzerinde kapsamlı bir hesaplama deneyi yürütülmüştür. Üyelik olasılıklarının karar verme için kullanımı da tartışılmaktadır. Bu çalışma, çift modlu kümeleme literatürü için öncü yumuşak atama yaklaşımı olacaktır.

Anahtar Kelimeler: kümeleme, olasılıksal kümeleme, tek modlu veri, konum teorisi, çift modlu veri kümeleme

To my youth...

ACKNOWLEDGMENTS

First of all, I would like to express my endless thanks to my precious and dear advisor, Cem İyigun. He is more than a supervisor to me, a mentor who touched my life, and a great teammate. I would not have been successful without his patience and endless support during this challenging process.

I would also like to thank the members of the thesis evaluation committee Prof. Dr. Sinan Gürel, Prof. Dr. Pınar Karagöz, Assist. Prof. Dr. Sakine Batun and Assist. Prof. Dr. Fatma Yerlikaya Özkurt for their valuable comments and feedback on this study.

I cannot thank my lovely companion Tuna Berk Kaya enough, who is always by my side. Whenever my motivation fell, he gave me strength. In academic life, I think that we constantly feed each other both intellectually and spiritually. Thank you so much for always being positive and for making this process fun.

I want to thank my dear friends Nazlı Dolu Hastürk and Umur Hastürk for always being there and rushing to my aid whenever I was in trouble. I want to thank Zeynep Makasçı, my dear friend and life coach, for introducing and reconciling me to myself. I learned a lot from you, big-hearted people.

Also, I want to thank my dear colleagues Ceyhan Şahin, Dilay Özkan, and Can Er for always listening and comforting me. Moreover, I will never forget our beautiful conversations with Erdinç Durak, my colleague and trouble partner.

I would like to thank the faculty members of the Middle East Technical University Industrial Engineering Department for educating me and bringing me to the level to write this thesis.

Finally, I would like to thank my beloved family for raising me to this day and always being there for me.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Clustering is a data mining technique to identify the group of unlabelled examples of a data set that are similar and dissimilar to each other according to a predefined distance metric or similarity measure. Clustering assumes that the data set has a cluster substructure. Otherwise, the whole data set forms one cluster, or each instance becomes a cluster itself. In either case, the resulting structure cannot yield useful information.

Clustering methods are divided into two main categories, which are partitional and hierarchical. In hierarchical clustering, clusters are formed by a similarity measure without a cluster center or an objective function. Partitional clusters are represented by a cluster center (prototype), and they are formed by minimizing a defined objective, which is a function of distance from each example to cluster centers. Therefore, a partitional clustering problem consists of two subproblems: finding cluster centers and assignments of data points (examples) to these centers. Assignments can be hard (crisp) or soft. For the hard assignments, each instance (entity, data point) of the data set is assigned to exactly one cluster, which leads to disjoint clusters. On the other hand, the soft assignment method assigns each entity to clusters with some membership degree (probability).

Properties of the data set are needed to be considered while deciding on the clustering method. A rectangular data set can be *one-mode* or *two-mode*. One-mode data sets have entities (data points, objects) in rows and features in columns. They store each entity's quantitative (continuous or binary) or qualitative (categorical) values over each feature. On the other hand, a two-mode data set contains two distinct sets of entities in rows and columns. It stores a degree of linkage, dependency, frequency, or

trend of each row and column entity pair.

In this study, our focus is Probabilistic Distance Clustering (PDC) [9], which is a soft clustering approach for one-mode data sets. The contribution of this work is two-fold. Firstly, we studied PDC within the framework of the one-mode clustering model. The PDC is a model constructed around some principles. We discuss the relationship of PDC with various problems from different contexts. We study Huff Model from marketing, K-Harmonic Means and Fuzzy c-Means algorithms from clustering, and Gravity p-Median Model from location literature. We reveal that those problems originated from PDC principles. The other contribution of this study is the development of two novel solution approaches for the soft Two-Mode Clustering (TMC) problem. This work will be the pioneer soft assignment approach for two-mode clustering literature. We drive the theoretical properties of two-mode clustering problems and formulate the soft TMC by following the PDC principles.

The outline of this study is as follows. In Chapter 2, background information and literature survey on clustering is provided. In Chapter 3, Probabilistic Distance Clustering (PDC), which is a one-mode soft clustering approach is introduced. Its relation with some problems in various contexts will be discussed. The two-mode clustering concept is introduced in Chapter 4, and the development of two novel soft partitioning approaches for two-mode data sets is explained. Chapter 5 compares these algorithms with an experimental study. Moreover, two applications on a hypothetical example and a part-machine group technology problem are provided in that chapter. Finally, conclusions of the study and future research directions are shared in Chapter 6.

CHAPTER 2

A BACKGROUND ON THE CLUSTERING

Clustering is a technique that draws inferences by dividing a set of unlabelled instances (objects, entities) into subsets, which are called *clusters*. The essential purpose of clustering is to identify a meaningful structure, underlying explanatory mechanism or patterns, generative characteristics, and clusters in a collection of data points. Clusters are formed by assigning most similar objects to the same groups and dissimilar ones to the separate groups as much as possible. In other words, maximum compactness within the clusters and maximum separation between the clusters are desired. Thus, there are two general stages of clustering. The first one is defining a proximity measure between instances to evaluate similarities or dissimilarities of the objects. The second stage is the selection of the objective that quantifies the overall proximity of the entities.

Clustering is a well-studied problem in a variety of disciplines. In the literature, it is used for a wide range of problems such as image processing, object recognition, bioinformatics, business analytics, data mining. Some other application areas can be found in literature surveys of [1], [2], [3], and [4].

In the literature, there are a variety of classifications of clustering. For this study, we construct a taxonomy of clustering problems given in Figure 2.1.

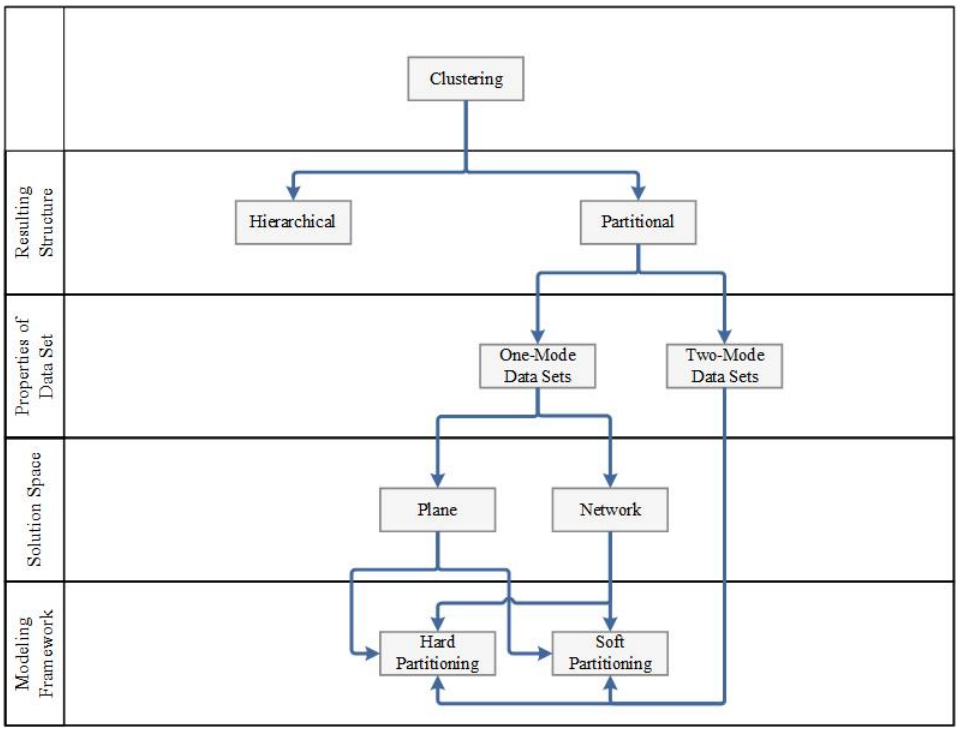


Figure 2.1: A taxonomy of clustering.

The first level of distinction is the resulting structure of the clustering approach, which may be hierarchical and partitional clusters. The methods result in these types are named *hierarchical clustering* and *partitional clustering*, respectively.

2.1 Hierarchical Clustering

In hierarchical clustering, data sets are divided into nested subsets. The generated clusters have a hierarchical structure, which is interpreted with dendrograms. A dendrogram is a tree-like structure with some layers that contain different subclusters. There are two forms of hierarchical clustering methods that are *divisive* and *agglomerative*.

The divisive method starts with one comprehensive cluster, which contains all instances of the data set. Then, it breaks up the cluster that generates two subclusters with maximum separation until a termination condition is satisfied. Therefore, this method is also known as the *top-down approach*. On the other hand, the agglomer-

ative method assigns each object into separate clusters. Later, it merges two clusters with minimum separation such that the resulting group has maximum compactness until a termination criterion is met. Thus, it is also called the *bottom-up approach*.

Either in divisive or agglomerative methods, when the clusters are split or merged with more than two members, defining distance between those clusters is an issue. There are three linkage criteria to evaluate proximity between clusters. *Single linkage* criteria measure the proximity between two groups by selecting the two closest members of them. *Complete linkage* selects one member from each cluster and calculates the distance between them. The farthest distance is defined as the distance between clusters. *Average linkage* uses the average distance of paired members of all elements that belong to different clusters.

Although hierarchical clustering is a widespread technique, there are some criticisms about the method. Firstly, it is hard to find the termination criteria. Secondly, generally, hierarchical clustering methods do not consider an instance again when it is assigned, which may result in misclassification. Moreover, they are computationally expensive. Thus, for the larger data sets, classical hierarchical clustering algorithms are improved. The upgraded methods are summarized in the literature survey of [2].

2.2 Partitional Clustering

In contrast to the hierarchical approach, partitional clustering assigns objects of the data set to K clusters without a hierarchical structure. The assignments are obtained based on a defined objective function. Partitional clustering techniques can be divided into two categories in terms of data set properties: *one-mode data set partitioning* and *two-mode data set partitioning*.

One-mode partitioning methods are deployed for data sets, usually data matrices with entities (cases, persons, objects) in rows and attributes (variables, features) in columns. Thus, elements (entries) of these data sets refer to the values of attributes for each data point. On the other hand, there may be entities in both rows and columns in some data sets. In that case, the elements may refer to values of an indicator that relates the row entities with column entities. An indicator may represent dependen-

cies, trends, or linkage between row and column entity pairs. A data set having this structure is called *two-mode data set*.

In Section 2.2.1, we introduce the properties of the one-mode data set partitioning problem and solution approaches. Detailed information and literature review of *two-mode clustering* problems are discussed in Chapter 4.

2.2.1 One-Mode Data Set Partitioning

The most common problems and solution approaches of the literature are about the one-mode clustering problems. We discuss the one-mode data set partitioning on the plane and network solution space in the subsequent parts.

Partitional Clustering on Plane

When the solution space is plane, by optimizing a specified objective function and iteratively enhancing the quality of the partitions, partitional clustering algorithms try to uncover the groupings existent in the data. To choose the prototype points (cluster centers) that represent each cluster, these algorithms typically require particular user settings. They are also known as *prototype-based* clustering methods because of this [3]. Partitional clustering problems require two main steps: determining the locations of cluster centers and assignments of the data points to that centers. Partitioning problems are classified by their assignment type as follows:

Hard (Crisp) Partitioning: Each data point must be assigned to exactly one cluster that leads to disjoint clusters.

Soft Partitioning: A data point can be assigned to multiple clusters according to their membership probabilities.

Among hard partitioning approaches, the most extensively used one is the K-means clustering algorithm introduced by MacQueen [5]. The algorithm initially starts with K representative (prototype) points. Then, instances are assigned to their closest centroids according to a predefined proximity measure. Based on preceding assignments,

centroids are updated. The last two steps are repeated iteratively until the centroids do not change anymore. K-means model uses squared Euclidean norm (L_2 norm) in the objective function. This objective is also known as *Sum of Squared Errors (SSE)*. Minimizers of the SSE objective with respect to centroids yield mean values of cluster members. K-median method is similar to the K-means algorithm. However, its representative points are cluster medians instead of means, and it minimizes the sum of absolute distances (L_1 norm) between data points and cluster medians. In the hard partitioning category, there is also the K-medoids algorithm. Instead of defining the representative points by a function of data set instances (as in K-means or K-median), the K-medoids algorithm uses K actual data points as representatives of clusters. Its objective function is the sum of absolute error criterion (L_1 norm). The most known K-medoids approach is the Partitioning Around Medoids (PAM) algorithm presented by Kaufman and Rousseeuw [6]. PAM minimizes the objective function by iteratively switching all non-mediod points with medoids until the convergence criterion is met. Since K-median and K-medoids methods employ the L_1 norm, they are more robust to outliers.

The soft partitioning approaches use probabilities as membership functions and assigns the data points to clusters with the membership probabilities. Thus, memberships take a value between 0 and 1. The Fuzzy c-Means (FCM) [7], K-Harmonic Means (KHM) [8], and Probabilistic Distance Clustering (PDC) [9] algorithms are in this category. Detailed explanations and comparisons of those approaches can be seen in Chapter 3. There are also probabilistic model-based clustering techniques that may yield soft assignments. These methods rely on the assumption that data are coming from an underlying probability distribution. Thus, Bayesian Theorem is usually used to derive the theoretical background of these algorithms. The most conventional probabilistic model-based clustering method is the Expectation-Maximization (EM) algorithm [10].

Partitional Clustering on Network

A network is a set of interconnected objects (named nodes or vertices) with edges (or links) connecting the nodes. In some problem environments, finding the set of

similar nodes may be needed. In that case, a network clustering approach would be necessary to detect hidden structures in networks. There are many approaches to partition networks [3]. However, these studies do not deal with finding a prototype point on the network.

Finding representative points (or hub locations) on the networks may be essential for some disciplines such as logistics, marketing, and epidemiology [11].

A center-based network clustering approach is the p-Median Model presented by Hakimi [12]. The purpose of the p-Median Model is to select p facilities on a network such that the total demand weighted distance from demand nodes to selected facilities is minimized. In the clustering concept, facilities and demand nodes can be represented as cluster centers and data points, respectively. The distance is defined as the length of the shortest path between a data point and cluster center.

Drezner et al. [13] developed a more realistic approach compared to the p-Median Model by relaxing the assumption that its closest facility serves each demand point. They follow a soft assignment method, in which demand may be probabilistically distributed across facilities. This approach is known as the Gravity p-Median (GPM) model in the literature. The details of this method are given in Chapter 3.

CHAPTER 3

PROBABILISTIC DISTANCE CLUSTERING AND ITS RELATION TO SOME PROBLEMS

As stated in Chapter 2, Probabilistic D-Clustering (PDC) [9] is a soft clustering algorithm for one-mode data sets. In this chapter, we discuss the relationship of the PDC with various problems in different contexts. We examine the Huff Model [14] in marketing, the Fuzzy c-Means Algorithm (FCM) [7], and the K-Harmonic Means Algorithm (KHM) [8] for clustering, and the Gravity p-Median Model (GPM) [13] as a location problem. We reveal that those problems originated from PDC principles.

In Sections 3.1-3.4, Huff Model, Fuzzy c-Means Algorithm, K-Harmonic Means Algorithm, and Gravity p-Median Model are introduced. Probabilistic D-Clustering Algorithm and its fundamental properties are given in Section 3.5. Under Section 3.6, we discuss the relation of PDC principles with the problems in Sections 3.1-3.4, and how PDC principles explain those problems.

3.1 Huff Model

In 1964, Huff presented a model to estimate a trading area [14]. Huff states that the main focus while estimating a trading area should be on the consumer. Therefore, his model focuses on the consumer preferences on different firm options. The model questions how likely a specific consumer prefers a firm among all defined alternatives.

Huff model expresses a consumer preference probability. According to expression, a consumer i at origin $\mathbf{x}_i$ visits a shopping center k with a probability of $p_k(\mathbf{x}_i)$ that is

proportional to the center's floor area and inversely proportional to distance or time traveled to this shopping center

$$p_k(\mathbf{x}_i) = \frac{\frac{S_k}{d_k(\mathbf{x}_i)^\theta}}{\sum_{l=1}^{K} \frac{S_l}{d_l(\mathbf{x}_i)^\theta}}, \tag{3.1}$$

where

$p_k(\mathbf{x}_i)$: the probability of a customer at the origin $\mathbf{x}_i$ visits a shopping center k,
S_k: the size of the shopping center k,
$d_k(\mathbf{x}_i)$: a measure of the accessibility of shopping center k from the origin $\mathbf{x}_i$,
θ: a parameter which is to be estimated empirically.

Later on, the Huff Model is generalized in the literature by substituting some multiplicative utility functions instead of shopping centers' floor sizes [15].

By the formal definition, the Huff Model has been used in various marketing decisions. Moreover, applications show that the model proposes successful results empirically. Although the model is conceptually sensible and empirically successful, the hidden mathematical logic has not been discussed.

3.2 Fuzzy c-Means Algorithm

Bezdek et al. presented a fuzzy clustering algorithm in 1984 [7]. Authors criticize hard (nonfuzzy) partitioning in terms of its inability to measure similarities between members of a specific cluster. They originate Zadeh's (1965) [16] idea of a *membership function* that represents the similarity between a data point with all clusters.

The Fuzzy c-Means approach (FCM) defines the membership function as follows

$$p_k(\mathbf{x}_i) \in [0,1], \quad \text{where} \quad \sum_{k=1}^{K} p_k(\mathbf{x}_i) = 1, \quad i = 1, ..., N. \tag{3.2}$$

Equation (3.2) means a data point i belongs to cluster k with a membership value between 0 and 1. Thus, $p_k(\mathbf{x}_i)$ is interpreted as the grades of membership of the data point $\mathbf{x}_i$ for the K subsets of data set $\mathbf{X}$.

The objective of the FCM is to minimize the least square error

$$Obj_{FCM} = \sum_{i=1}^{N} \sum_{k=1}^{K} p_k(\mathbf{x}_i)^m \parallel \mathbf{x}_i - \mathbf{c}_k \parallel_A^2, \tag{3.3}$$

where

$p_k(\mathbf{x}_i)$: grade of membership of the data point $\mathbf{x}_i$ to cluster center $\mathbf{c}_k$,
m: weighting exponent, $1 \leq m < \infty$,
$|| \cdot ||_A$: induced A-norm on $\mathbb{R}^n$.

When the weighting exponent $m = 1$, partitioning becomes hard, and all cluster centers $\mathbf{c}_k$ are located on the geometric centroids of the data points $\mathbf{x}_i$. On the other hand, as $m \to \infty$, all clusters become equally likely; thus, memberships $p_k(\mathbf{x}_i)$ approach to $1/K$. Thus, increasing m is represented as degrading membership towards the fuzziest state by authors.

A-norm can be Euclidean, Diagonal, or Mahalanobis. When the norm is selected as Euclidean $d_k(\mathbf{x}_i) = \parallel \mathbf{x}_i - \mathbf{c}_k \parallel_I$, where I is the identity matrix, and $m > 1$, the minimizers of objective in (3.3) for cluster centers are

$$\mathbf{c}_k = \frac{\sum_{i=1}^{N} p_k(\mathbf{x}_i)^m \mathbf{x}_i}{\sum_{i=1}^{N} p_k(\mathbf{x}_i)^m}, \quad k = 1, ..., K, \tag{3.4}$$

where

$$p_k(\mathbf{x}_i) = \left(\sum_{l=1}^{K} \left(\frac{d_k(\mathbf{x}_i)}{d_l(\mathbf{x}_i)} \right)^{2/(m-1)} \right)^{-1}. \tag{3.5}$$

3.3 K-Harmonic Means Algorithm

K-Harmonic Means Clustering Algorithm (KHM) [8] was proposed by Zhang et al. The authors criticize dependency on initialization of the well-known K-Means (KM) [5] and Expectation-Maximization Algorithm (EM) [10]. Therefore, they propose a new center-based algorithm that uses harmonic averages of distances between each data point $\mathbf{x}_i$ and cluster center $\mathbf{c}_k$ as an objective function. In 2000, Zhang proposed a generalized version of KHM by using the θ^{th} power of the Euclidean distance in the objective function (KHM_θ) [17].

Harmonic average of K numbers is defined as

$$HA(n1, n2, ..., n_K) = \frac{K}{\sum_{k=1}^{K} \frac{1}{n_k}} \tag{3.6}$$

The objective function of KHM_θ is

$$Obj_{KHM_\theta} = \sum_{i=1}^{N} HA(d_k(\mathbf{x}_i)^\theta |\ k = 1, ..., K) = \sum_{i=1}^{N} \frac{K}{\sum_{k=1}^{K} \frac{1}{d_k(\mathbf{x}_i)^\theta}}, \tag{3.7}$$

where $d_k(\mathbf{x}_i)^\theta$ is the θ^{th} power of the Euclidean distance from data point $\mathbf{x}_i$ to cluster center $\mathbf{c}_k$.

Partial derivatives with respect to cluster centers $\mathbf{c}_k$, $k = 1, ..., K$, are taken and set to zero in order to derive the center update procedure of the algorithm as follows

$$\frac{\delta Obj_{KHM_\theta}}{\delta \mathbf{c}_k} = \frac{\delta \left(\sum_{i=1}^{N} \frac{K}{\sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^\theta}} \right)}{\delta \mathbf{c}_k} = \mathbf{0},$$

which yields the following centers

$$\mathbf{c}_k = \frac{\sum_{i=1}^{N} \frac{\mathbf{x}_i}{d_k(\mathbf{x}_i)^{\theta+2} \left(\sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^\theta} \right)^2}}{\sum_{i=1}^{N} \frac{1}{d_k(\mathbf{x}_i)^{\theta+2} \left(\sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^\theta} \right)^2}}, \quad k = 1, ..., K. \tag{3.8}$$

To compare KM, EM, and KHM_θ, Zhang derives a unified form of center update procedures of these algorithms as

$$\mathbf{c}_k = \frac{\sum_{i=1}^{N} p_k(\mathbf{x}_i) w(\mathbf{x}_i) \mathbf{x}_i}{\sum_{i=1}^{N} p_k(\mathbf{x}_i) w(\mathbf{x}_i)}, \quad p_k(\mathbf{x}_i) \geq 0, \quad \sum_{l=1}^{K} p_k(\mathbf{x}_i) = 1, \quad and \quad w(\mathbf{x}_i) > 0, \tag{3.9}$$

where

$p_k(\mathbf{x}_i)$: the probability of a given data point $\mathbf{x}_i$ is assigned to cluster center $\mathbf{c}_k$,

$w(\mathbf{x}_i)$: weight of the data point $\mathbf{x}_i$ in the following iterations.

The author states that in KM, $p_k(\mathbf{x}_i)$ is binary as data points are assigned to its closest cluster center and $w(\mathbf{x}_i) = 1$ for all data points. In EM, the membership probability $p_k(\mathbf{x}_i)$ comes from Bayes' rule, and since it is normalized, $w(\mathbf{x}_i) = 1$ for all data points.

To obtain membership probability and weight function for KHM_θ that satisfies (3.9), Zhang reintroduces the center update procedure (3.8) as follows

$$\mathbf{c}_k = \frac{\sum_{i=1}^{N} \frac{\mathbf{x}_i}{d_k(\mathbf{x}_i)^{\theta+2} \left(\sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^\theta} \right)^2}}{\sum_{i=1}^{N} \frac{1}{d_k(\mathbf{x}_i)^{\theta+2} \left(\sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^\theta} \right)^2}} = \frac{\sum_{i=1}^{N} \frac{\frac{1}{d_k(\mathbf{x}_i)^{\theta+2}}}{\sum_{l=1}^{K} \frac{1}{d_k(\mathbf{x}_i)^{\theta+2}}} * \frac{\sum_{l=1}^{K} \frac{1}{d_k(\mathbf{x}_i)^{\theta+2}}}{\left(\sum_{l=1}^{K} \frac{1}{d_k(\mathbf{x}_i)^\theta} \right)^2} * \mathbf{x}_i}{\sum_{i=1}^{N} \frac{\frac{1}{d_k(\mathbf{x}_i)^{\theta+2}}}{\sum_{l=1}^{K} \frac{1}{d_k(\mathbf{x}_i)^{\theta+2}}} * \frac{\sum_{l=1}^{K} \frac{1}{d_k(\mathbf{x}_i)^{\theta+2}}}{\left(\sum_{l=1}^{K} \frac{1}{d_k(\mathbf{x}_i)^\theta} \right)^2}}, \tag{3.10}$$

which gives

$$p_k(\mathbf{x}_i) = \frac{\frac{1}{d_k(\mathbf{x}_i)^{\theta+2}}}{\sum_{l=1}^{K} \frac{1}{d_k(\mathbf{x}_i)^{\theta+2}}} \quad and \quad w(\mathbf{x}_i) = \frac{\sum_{l=1}^{K} \frac{1}{d_k(\mathbf{x}_i)^{\theta+2}}}{\left(\sum_{l=1}^{K} \frac{1}{d_k(\mathbf{x}_i)^\theta} \right)^2}. \tag{3.11}$$

Although $p_k(\mathbf{x}_i)$ and $w(\mathbf{x}_i)$ satisfy (3.9), the derivation procedure in (3.10) is questionable as a different representation may be derived for membership probability and weight that satisfies (3.9). See Section 3.6.3.

3.4 Gravity p-Median Model

A probabilistic perspective was applied to the classical p-Median Problem [12] by Drezner et al. in 2006 [13]. The assumption that every customer visits the facility closest to it of the p-Median Problem (PM) is relaxed. In the Gravity p-Median Model (GPM), given the locations of facilities, a customer i at the origin $\mathbf{x}_i$ prefers a specific facility k with a probability of $p_k(\mathbf{x}_i)$.

The objective function of the GPM is to minimize the total distance traveled by the customers to their probabilistically selected facilities

$$Obj_{GPM} = \sum_{i=1}^{N} \sum_{k=1}^{K} w(\mathbf{x}_i) d_k(\mathbf{x}_i) p_k(\mathbf{x}_i), \tag{3.12}$$

where

$w(\mathbf{x}_i)$: demand at origin $\mathbf{x}_i$,
$p_k(\mathbf{x}_i)$: the probability of the customers i at origin $\mathbf{x}_i$ prefers facility k,
$d(\mathbf{x}_i)$: distance between customer i at origin $\mathbf{x}_i$ and facility k at origin $\mathbf{c}_k$.

The probability definition of the Huff Model (3.1) is used with the following representation

$$p_k(\mathbf{x}_i) = \frac{u_k(d_k(\mathbf{x}_i))}{\sum_{l=1}^{K} u_l(d_l(\mathbf{x}_i))}, \tag{3.13}$$

where

$u_k(d_k(\mathbf{x}_i))$: the distance dependent utility of a facility k for a demand point $\mathbf{x}_i$.

The utility function $u_k(d_k(\mathbf{x}_i))$ can be any distance decay function such as $d_k(\mathbf{x}_i)^{-\theta}$ or $e^{-\theta d_k(\mathbf{x}_i)}$.

Let the $\{k = 1, ..., K\}$ is the set of facilities, the optimization problem of GPM is

$$\begin{aligned} \underset{p_k(\mathbf{x}_i)}{\text{minimize}} \quad & \sum_{i=1}^{N} \sum_{k=1}^{K} w(\mathbf{x}_i) d_k(\mathbf{x}_i) p_k(\mathbf{x}_i) \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i) = 1, \quad i = 1, \ldots, N, \\ & p_k(\mathbf{x}_i) \geq 0, \quad i = 1, \ldots, N \quad and \quad k = 1, \ldots, K. \end{aligned} \tag{3.14}$$

Optimality conditions of the GPM problem in (3.14) do not yield the probabilities in (3.13). Thus, the GPM problem in (3.14) should be revised as explained in Section 3.6.4. In addition, the authors restrict facility locations with customer origins. This assumption makes the problem combinatorial. That is why they could work with the objective function in (3.12) by applying the steepest descent and tabu search heuristic approaches. Analysis of the GPM when facilities can be located anywhere on the

network is remarked as future work. If the objective function in (3.12) is reconstructed as in Section 3.6.4, restrictions on facility locations can be relaxed.

3.5 Probabilistic D-Clustering

Israel and Iyigun [9] presented a new algorithm for probabilistic clustering in 2008. This algorithm proposes a generalization of the Weiszfeld method to multiple centers in the location theory. The advantages of the proposed method are stated as its simplicity and speed. Moreover, the algorithm is insensitive to outliers.

The PDC is based on a root principle that assumes the membership probability of a data point is inversely proportional to the distance from the cluster center as follows

Principle 1. *For each data point $\boldsymbol{x}_i$ of data set $\boldsymbol{X}$, and each cluster $\boldsymbol{c}_k$,*

$$p_k(\mathbf{x}_i)d_k(\mathbf{x}_i) = D(\mathbf{x}_i), \tag{3.15}$$

where

$p_k(\mathbf{x}_i)$: the membership probability of given the data point i at the origin $\mathbf{x}_i$ is assigned to cluster center k at the origin $\mathbf{c}_k$,

$d_k(\mathbf{x}_i) = d(\mathbf{x}_i, \mathbf{c}_k)$: distance between data point $\mathbf{x}_i$ and cluster center $\mathbf{c}_k$,

$D(\mathbf{x}_i)$: a constant, depending on $\mathbf{x}_i$.

Principle 1 in (3.15) means that given the cluster centers, a data point $\mathbf{x}_i$ is more likely to be assigned to its closer cluster centers.

3.5.1 The Membership Probabilities

With Principle 1 (3.15), and the fact that probabilities add up to one, the following theorem is obtained.

Theorem 1. *Principle 1 yields following membership probabilities for a data point $\boldsymbol{x}_i$*

$$p_k(\boldsymbol{x}_i) = \frac{\dfrac{1}{d_k(\boldsymbol{x}_i)}}{\sum\limits_{l=1}^{K} \dfrac{1}{d_l(\boldsymbol{x}_i)}}, \quad k = 1, ..., K. \tag{3.16}$$

Proof. Using (3.15) the following can be written for two cluster centers l, and k

$$p_l(\mathbf{x}_i) = \frac{p_k(\mathbf{x}_i)d_k(\mathbf{x}_i)}{d_l(\mathbf{x}_i)}. \tag{3.17}$$

Together with (3.17) and the fact that probabilities add to one, we have

$$\begin{aligned}
\sum_{l=1}^{K} \frac{p_k(\mathbf{x}_i)d_k(\mathbf{x}_i)}{d_l(\mathbf{x}_i)} &= 1, \\
p_k(\mathbf{x}_i)d_k(\mathbf{x}_i) \sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)} &= 1, \\
p_k(\mathbf{x}_i) \sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)} &= \frac{1}{d_k(\mathbf{x}_i)}, \\
p_k(\mathbf{x}_i) &= \frac{\dfrac{1}{d_k(\mathbf{x}_i)}}{\sum\limits_{l=1}^{K} \dfrac{1}{d_l(\mathbf{x}_i)}},
\end{aligned}$$

proving (3.16). □

3.5.2 The Joint Distance Function

Recall that in Principle 1 (3.15), $D(\mathbf{x}_i)$ is a function of $\mathbf{x}_i$.

Theorem 2. *Principle 1 yields the constant $D(\boldsymbol{x}_i)$ as*

$$D(\boldsymbol{x}_i) = \frac{1}{\sum\limits_{k=1}^{K} \dfrac{1}{d_k(\boldsymbol{x}_i)}}. \tag{3.18}$$

Proof. Using (3.15) the following can be written for cluster k

$$p_k(\mathbf{x}_i) = \frac{D(\mathbf{x}_i)}{d_k(\mathbf{x}_i)}, \quad k = 1, ..., K. \tag{3.19}$$

With (3.19) and the fact that probabilities add to one, we have

$$\begin{aligned}
\sum_{k=1}^{K} \frac{D(\mathbf{x}_i)}{d_k(\mathbf{x}_i)} &= 1, \\
D(\mathbf{x}_i) \sum_{k=1}^{K} \frac{1}{d_k(\mathbf{x}_i)} &= 1, \\
D(\mathbf{x}_i) &= \frac{1}{\sum\limits_{k=1}^{K} \dfrac{1}{d_k(\mathbf{x}_i)}},
\end{aligned}$$

proving (3.18) □

The constant $D(\mathbf{x}_i)$ is defined as the joint distance function (JDF) of $\mathbf{x}_i$ and equal to the harmonic averages of distances between each data point $\mathbf{x}_i$ and cluster center $\mathbf{c}_k$ over number of clusters K as

$$D(\mathbf{x}_i) = \frac{1}{\sum_{k=1}^{K} \frac{1}{d_k(\mathbf{x}_i)}} = \frac{\frac{K}{\sum_{k=1}^{K} \frac{1}{d_k(\mathbf{x}_i)}}}{K} = \frac{HA(d_k(\mathbf{x}_i)|\ k = 1, ..., K)}{K}.$$

Then, the joint distance function of the whole data set $\mathbf{X}$ is the sum of constant $D(\mathbf{x}_i)$ over all data points $\{\mathbf{x}_i \mid i = 1, ..., N\}$ as follows

$$\sum_{i=1}^{N} D(\mathbf{x}_i) = \sum_{i=1}^{N} \frac{HA(d_k(\mathbf{x}_i)|\ k = 1, ..., K)}{K} \tag{3.20}$$

$$= \frac{1}{K} \sum_{i=1}^{N} HA(d_k(\mathbf{x}_i)|\ k = 1, ..., K). \tag{3.21}$$

3.5.3 The Optimization Problem of the Method

As the nature of the clustering problem, the minimum total distance between all data points and the cluster centers is desired. In the probabilistic assignment, one can easily guess that the expected total distance between all data points and the cluster centers should be minimized. Therefore, one may expect the following optimization problem

$$\begin{aligned} \underset{p_k(\mathbf{x}_i),\, \mathbf{c}_k}{\text{minimize}} \quad & \sum_{i=1}^{N} \sum_{k=1}^{K} d_k(\mathbf{x}_i) p_k(\mathbf{x}_i) \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i) = 1, \quad i = 1, \ldots, N, \\ & p_k(\mathbf{x}_i) \geq 0, \quad i = 1, \ldots, N \quad \textit{and} \quad k = 1, \ldots, K. \end{aligned} \tag{3.22}$$

However, in that case a solution may not exist as explained in Proposition 1.

Proposition 1. *Optimality conditions of the problem in (3.22) may not yield a solution for* $N \geq 3$.

Proof. For simplicity consider the case $K = 2$, and centers $\mathbf{c}_k$ are given then the problem (3.22) becomes

$$\begin{aligned} \underset{p_1(\mathbf{x}_i),\, p_2(\mathbf{x}_i)}{\text{minimize}} \quad & \sum_{i=1}^{N} d_1(\mathbf{x}_i)p_1(\mathbf{x}_i) + d_2(\mathbf{x}_i)p_2(\mathbf{x}_i) \\ \text{subject to} \quad & p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) = 1, \quad i = 1, \ldots, N, \\ & p_1(\mathbf{x}_i), p_2(\mathbf{x}_i) \geq 0, \quad i = 1, \ldots, N. \end{aligned} \tag{3.23}$$

The Lagrangian of the problem (3.23) is

$$\mathcal{L}(p_1(\mathbf{x}_i), p_2(\mathbf{x}_i), \lambda_i) = \left(\sum_{i=1}^{N} d_1(\mathbf{x}_i)p_1(\mathbf{x}_i) + d_2(\mathbf{x}_i)p_2(\mathbf{x}_i) \right) - \lambda_i(p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) - 1)$$

$$\frac{\partial \mathcal{L}}{\partial p_1(\mathbf{x}_i)} = d_1(\mathbf{x}_i) - \lambda_i = 0, \quad i = 1, ..., N \tag{3.24}$$

$$\frac{\partial \mathcal{L}}{\partial p_2(\mathbf{x}_i)} = d_2(\mathbf{x}_i) - \lambda_i = 0, \quad i = 1, ..., N \tag{3.25}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) - 1 = 0, \quad i = 1, ..., N$$

Equations (3.24-3.25) yield

$$d_1(\mathbf{x}_i) = d_2(\mathbf{x}_i), \quad i = 1, ..., N, \tag{3.26}$$

which gives alternative solutions for cluster centers that are lying on a line formed by equal distances from data points. Thus, when $N \geq 3$, even a solution may not exist. □

Teboulle [18], and Ben-Israel et al. [9] prove that the *smoothed* version of the classical clustering problem, $min\{d_1, d_2\}$, uses the squares of probabilities in the objective. The optimization problem is then

$$\begin{aligned} \underset{p_k(\mathbf{x}_i),\, \mathbf{c}_k}{\text{minimize}} \quad & \sum_{i=1}^{N} \sum_{k=1}^{K} d_k(\mathbf{x}_i)p_k(\mathbf{x}_i)^2 \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i) = 1, \quad i = 1, \ldots, N, \\ & p_k(\mathbf{x}_i) \geq 0, \quad i = 1, \ldots, N \quad and \quad k = 1, \ldots, K. \end{aligned} \tag{3.27}$$

In the case centers are known, the objective of the problem (3.27) becomes function of membership probabilities as follows

$$\begin{aligned} \underset{p_k(\mathbf{x}_i)}{\text{minimize}} \quad & \sum_{i=1}^{N}\sum_{k=1}^{K} d_k(\mathbf{x}_i)p_k(\mathbf{x}_i)^2 \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i) = 1, \quad i = 1,\ldots,N, \\ & p_k(\mathbf{x}_i) \geq 0, \quad i = 1,\ldots,N \quad and \quad k = 1,\ldots,K \end{aligned} \tag{3.28}$$

The problem (3.28) is a convex constraint optimization model so that its optimality conditions can be found Lagrangian method as in Proposition 1. Optimality conditions of (3.28) give Principle 1 in (3.15).

3.5.4 Centers

In the case of probabilities are given, the objective function in (3.27) can be written as function of cluster centers

$$Obj_{PDC} = \sum_{i=1}^{N}\sum_{k=1}^{K} d(\mathbf{x}_i, \mathbf{c}_k)p_k(\mathbf{x}_i)^2.$$

Theorem 3. *Let the distance function $d(\boldsymbol{x}_i, \boldsymbol{c}_k)$ be Euclidean so that*

$$Obj_{PDC} = \sum_{i=1}^{N}\sum_{k=1}^{K} \| \boldsymbol{x}_i - \boldsymbol{c}_k \| \; p_k(\boldsymbol{x}_i)^2, \tag{3.29}$$

and assume $\boldsymbol{c}_k$ and $\boldsymbol{x}_i$ are not equal for any data point i. Then the minimizers $\boldsymbol{c}_k$ are obtained by

$$\boldsymbol{c}_k = \frac{\sum\limits_{i=1}^{N} u_k(\boldsymbol{x}_i)\boldsymbol{x}_i}{\sum\limits_{i=1}^{N} u_k(\boldsymbol{x}_i)}, \quad k = 1, ..., K, \tag{3.30}$$

where

$$u_k(\boldsymbol{x}_i) = \frac{p_k(\boldsymbol{x}_i)^2}{d(\boldsymbol{x}_i, \boldsymbol{c}_k)}. \tag{3.31}$$

Proof. The gradient of $d(\mathbf{x}_i, \mathbf{c}_k) = \| \mathbf{x}_i - \mathbf{c}_k \|$ with respect to $\mathbf{c}_k$ is

$$\nabla_{\mathbf{c}_k} \| \mathbf{x}_i - \mathbf{c}_k \| = -\frac{\mathbf{x}_i - \mathbf{c}_k}{\| \mathbf{x}_i - \mathbf{c}_k \|} = -\frac{\mathbf{x}_i - \mathbf{c}_k}{d(\mathbf{x}_i, \mathbf{c}_k)}.$$

Therefore, the gradient of (3.29) with respect to $\mathbf{c}_k$ is

$$\nabla_{\mathbf{c}_k} Obj_{PDC} = -\sum_{i=1}^{N} \frac{\mathbf{x}_i - \mathbf{c}_k}{d(\mathbf{x}_i, \mathbf{c}_k)} p_k(\mathbf{x}_i)^2$$

Setting the gradient equal to zero

$$\begin{aligned}
-\sum_{i=1}^{N} \frac{\mathbf{x}_i - \mathbf{c}_k}{d(\mathbf{x}_i, \mathbf{c}_k)} p_k(\mathbf{x}_i)^2 &= \mathbf{0}, \\
\sum_{i=1}^{N} \frac{\mathbf{x}_i - \mathbf{c}_k}{d(\mathbf{x}_i, \mathbf{c}_k)} p_k(\mathbf{x}_i)^2 &= \mathbf{0}, \\
\sum_{i=1}^{N} \frac{p_k(\mathbf{x}_i)^2}{d(\mathbf{x}_i, \mathbf{c}_k)} \mathbf{x}_i - \sum_{i=1}^{N} \frac{p_k(\mathbf{x}_i)^2}{d(\mathbf{x}_i, \mathbf{c}_k)} \mathbf{c}_k &= \mathbf{0}, \\
\sum_{i=1}^{N} \frac{p_k(\mathbf{x}_i)^2}{d(\mathbf{x}_i, \mathbf{c}_k)} \mathbf{x}_i - \mathbf{c}_k \sum_{i=1}^{N} \frac{p_k(\mathbf{x}_i)^2}{d(\mathbf{x}_i, \mathbf{c}_k)} &= \mathbf{0},
\end{aligned}$$

gives

$$\mathbf{c}_k = \frac{\sum_{i=1}^{N} \frac{p_k(\mathbf{x}_i)^2}{d(\mathbf{x}_i, \mathbf{c}_k)} \mathbf{x}_i}{\sum_{i=1}^{N} \frac{p_k(\mathbf{x}_i)^2}{d(\mathbf{x}_i, \mathbf{c}_k)}},$$

proving (3.30) and (3.31). □

3.5.5 Generalized Principles

Iyigun and Ben-Israel [19] generalize Principle 1. In this section, generalized versions of PDC are presented. The membership probabilities, joint distance function, objective function, and center update procedure are summarized for each principle in Table 3.1-3.3.

Exponents of membership probability and distance function can be used in PDC. Generalization of Principle 1 is given as Principle 2.

Principle 2. *For each data point $\boldsymbol{x}_i \in \boldsymbol{X}$, and each cluster center $\boldsymbol{c}_k$,*

$$p_k(\mathbf{x}_i)^{\alpha} d_k(\mathbf{x}_i)^{\beta} = C(\mathbf{x}_i), \tag{3.32}$$

where α, β are positive and $C(\boldsymbol{x}_i)$ is constant depending on data point $\boldsymbol{x}_i$.

Table 3.1: Properties of Principle 2.

Membership Probability	$p_k(\mathbf{x}_i) = \dfrac{\dfrac{1}{d_k(\mathbf{x}_i)^{\beta/\alpha}}}{\sum_{l=1}^{K} \dfrac{1}{d_l(\mathbf{x}_i)^{\beta/\alpha}}}, \quad k = 1, ..., K$	(3.33)
Joint Distance Function	$C(\mathbf{x}_i) = \left[\dfrac{HA(d_k(\mathbf{x}_i)^{\beta/\alpha} \mid k = 1, ..., K)}{K} \right]^{\alpha}$	(3.34)
Objective Function	$Obj_{PDC} = \sum_{i=1}^{N} \sum_{k=1}^{K} d_k(\mathbf{x}_i)^{\beta} p_k(\mathbf{x}_i)^{\alpha+1}$	(3.35)
Centers	$\mathbf{c}_k = \dfrac{\sum_{i=1}^{N} d_k(\mathbf{x}_i)^{\beta-2} p_k(\mathbf{x}_i)^{\alpha+1} \mathbf{x}_i}{\sum_{i=1}^{N} d_k(\mathbf{x}_i)^{\beta-2} p_k(\mathbf{x}_i)^{\alpha+1}}$	(3.36)

Ben-Israel and Iyigun [9] state that any distance decay function can be used in Principle 1 in (3.15). They examine the exponential function as Principle 3.

Principle 3. *For each data point $\boldsymbol{x}_i \in \boldsymbol{X}$, and each cluster center $\boldsymbol{c}_k$,*

$$p_k(\mathbf{x}_i) e^{d_k(\mathbf{x}_i)} = E(\mathbf{x}_i), \tag{3.37}$$

where $E(\boldsymbol{x}_i)$ is constant depending on data point $\boldsymbol{x}_i$.

Table 3.2: Properties of Principle 3.

Membership Probability	$p_k(\mathbf{x}_i) = \dfrac{\dfrac{1}{e^{d_k(\mathbf{x}_i)}}}{\sum_{l=1}^{K} \dfrac{1}{e^{d_l(\mathbf{x}_i)}}}, \quad k = 1, ..., K$	(3.38)
Joint Distance Function	$E(\mathbf{x}_i) = \dfrac{HA(e^{d_k(\mathbf{x}_i)} \mid k = 1, ..., K)}{K}$	(3.39)
Objective Function	$Obj_{PDC} = \sum_{i=1}^{N} \sum_{k=1}^{K} e^{d_k(\mathbf{x}_i)} p_k(\mathbf{x}_i)^2$	(3.40)
Centers	$\mathbf{c}_k = \dfrac{\sum_{i=1}^{N} d_k(\mathbf{x}_i)^{-1} p_k(\mathbf{x}_i)^2 e^{d_k(\mathbf{x}_i)} \mathbf{x}_i}{\sum_{i=1}^{N} d_k(\mathbf{x}_i)^{-1} p_k(\mathbf{x}_i)^2 e^{d_k(\mathbf{x}_i)}}$	(3.41)

In the first three principles, cluster sizes are assumed to be equal. Iyigun and Ben-Israel[19] consider the case of cluster sizes are distinct and adjust Principle 1 for cluster size as Principle 4.

Principle 4. *For each data point $\boldsymbol{x}_i \in \boldsymbol{X}$, and each cluster center $\boldsymbol{c}_k$,*

$$\frac{p_k(\mathbf{x}_i) d_k(\mathbf{x}_i)}{q_k} = B(\mathbf{x}_i), \tag{3.42}$$

where q_k is the size of cluster k and $B(\boldsymbol{x}_i)$ is constant depending on data point $\boldsymbol{x}_i$.

Table 3.3: Properties of Principle 4.

Membership Probability	$p_k(\mathbf{x}_i) = \dfrac{\dfrac{q_k}{d_k(\mathbf{x}_i)}}{\sum_{l=1}^{K} \dfrac{q_l}{d_l(\mathbf{x}_i)}}, \quad k = 1, ..., K$	(3.43)
Joint Distance Function	$B(\mathbf{x}_i) = \dfrac{HA(d_k(\mathbf{x}_i)/q_k \mid k = 1, ..., K)}{K}$	(3.44)
Objective Function	$Obj_{PDC} = \sum_{i=1}^{N} \sum_{k=1}^{K} \dfrac{d_k(\mathbf{x}_i) p_k(\mathbf{x}_i)^2}{q_k}$	(3.45)
Centers	$\mathbf{c}_k = \dfrac{\sum_{i=1}^{N} d_k(\mathbf{x}_i)^{-1} p_k(\mathbf{x}_i)^2 \mathbf{x}_i}{\sum_{i=1}^{N} d_k(\mathbf{x}_i)^{-1} p_k(\mathbf{x}_i)^2}$	(3.46)

3.6 A Unified View on Probabilistic Methods

In this section, the structure of models explained in Sections 3.1-3.4 is discussed. Their constructions are criticized and revisited from the perspective of PDC principles (Section 3.5). Although those models are from different research topics, namely clustering, marketing, and location models, PDC principles can explain their concepts and theories as discussed in Sections 3.6.1-3.6.4.

3.6.1 Principles of Huff Model

In Section 3.1, Huff Model is introduced. We criticize that the model is conceptually meaningful but lacks theoretical explanation. Therefore, in this section, we reexamine the Huff Model probability expression under the umbrella of PDC principles.

Proposition 2. *Principle 2 in (3.32), and Principle 4 in (3.42) explain Huff Model's*

probability definition.

Proof. Huff Model's probability expression in (3.1) can be written as

$$p_k(\mathbf{x}_i)\frac{d_k(\mathbf{x}_i)^\theta}{S_k} = \frac{1}{\sum_{l=1}^{K}\frac{S_l}{d_l(\mathbf{x}_i)^\theta}}. \tag{3.47}$$

The right-hand side of (3.47) is the reciprocal of the sum of store areas over distances between customer i and stores l. Since in the Huff model, locations of shopping centers $\mathbf{c}_l$ are predefined, the right-hand side of (3.47) is a constant depending on the location of customer $\mathbf{x}_i$. In particular, for $K = 2$, we can write

$$\begin{aligned}
&\frac{p_1(\mathbf{x}_i)d_1(\mathbf{x}_i)^\theta}{S_1} = \frac{p_2(\mathbf{x}_i)d_2(\mathbf{x}_i)^\theta}{S_2},\\
&p_1(\mathbf{x}_i)\frac{d_1(\mathbf{x}_i)^\theta}{S_1} = (1-p_1(\mathbf{x}_i))\frac{d_2(\mathbf{x}_i)^\theta}{S_2},\\
&p_1(\mathbf{x}_i)\left(\frac{d_1(\mathbf{x}_i)^\theta}{S_1}+\frac{d_2(\mathbf{x}_i)^\theta}{S_2}\right) = \frac{d_2(\mathbf{x}_i)^\theta}{S_2},\\
&p_1(\mathbf{x}_i) = \frac{\frac{d_2(\mathbf{x}_i)^\theta}{S_2}}{\left(\frac{d_1(\mathbf{x}_i)^\theta}{S_1}+\frac{d_2(\mathbf{x}_i)^\theta}{S_2}\right)},\\
&p_1(\mathbf{x}_i) = \frac{\frac{d_2(\mathbf{x}_i)^\theta}{S_2}}{\left(\frac{d_1(\mathbf{x}_i)^\theta}{S_1}+\frac{d_2(\mathbf{x}_i)^\theta}{S_2}\right)}\frac{\frac{S_1S_2}{d_1(\mathbf{x}_i)^\theta d_2(\mathbf{x}_i)^\theta}}{\frac{S_1S_2}{d_1(\mathbf{x}_i)^\theta d_2(\mathbf{x}_i)^\theta}},
\end{aligned}$$

which yields the following probability

$$p_1(\mathbf{x}_i) = \frac{\frac{S_1}{d_1(\mathbf{x}_i)^\theta}}{\frac{S_1}{d_1(\mathbf{x}_i)^\theta}+\frac{S_2}{d_2(\mathbf{x}_i)^\theta}},$$

and similarly,

$$p_2(\mathbf{x}_i) = \frac{\frac{S_2}{d_2(\mathbf{x}_i)^\theta}}{\frac{S_1}{d_1(\mathbf{x}_i)^\theta}+\frac{S_2}{d_2(\mathbf{x}_i)^\theta}},$$

proving Proposition 2, where $\alpha = 1$, $\beta = \theta$ in Principle 2 (Table 3.1), and $q_k = S_k$ in Principle 4 (Table 3.3). □

In a particular case of $\alpha = 1$, $\beta = 1$, and store sizes S_k are equal for all stores, the Huff Model follows Principle 1 (3.15).

Moreover, the probability definition of the Huff Model is the optimal solution of following optimization problem

$$\begin{aligned} \underset{p_k(\mathbf{x}_i)}{\text{minimize}} \quad & \sum_{i=1}^{N}\sum_{k=1}^{K} \frac{d_k(\mathbf{x}_i)^\theta p_k(\mathbf{x}_i)^2}{S_k} \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i) = 1, \quad i = 1, \ldots, N, \\ & p_k(\mathbf{x}_i) \geq 0, \quad i = 1, \ldots, N \quad and \quad k = 1, \ldots, K, \end{aligned}$$

where the floor sizes of the facilities S_k and origins of shopping centers $\mathbf{c}_k$ are known.

3.6.2 Principles of Fuzzy c-Means

The Fuzzy c-Means Algorithm (FCM) is explained in Section 3.2. Under this section, we relate the principles of PDC with FCM by Proposition 3.

Proposition 3. *FCM follows Principle 2 (3.32).*

Proof. Consider the objective (3.3), center update procedure (3.4), and membership probabilities (3.2) of FCM, when we substitute the constant m with $\alpha + 1$, the objective function of FCM (3.3) becomes

$$Obj_{FCM} = \sum_{i=1}^{N}\sum_{k=1}^{K} p_k(\mathbf{x}_i)^{\alpha+1} d_k(\mathbf{x}_i)^2, \tag{3.48}$$

and the centers (3.4) are

$$\mathbf{c}_k = \frac{\sum_{i=1}^{N} p_k(\mathbf{x}_i)^{\alpha+1}\mathbf{x}_i}{\sum_{i=1}^{N} p_k(\mathbf{x}_i)^{\alpha+1}}, \quad k = 1, ..., K, \tag{3.49}$$

where

$$p_k(\mathbf{x}_i) = \left(\sum_{l=1}^{K} \left(\frac{d_k(\mathbf{x}_i)}{d_l(\mathbf{x}_i)} \right)^{2/\alpha} \right)^{-1} = \frac{1}{\sum_{l=1}^{K} \left(\frac{d_k(\mathbf{x}_i)}{d_l(\mathbf{x}_i)} \right)^{2/\alpha}}$$

$$= \frac{1}{d_k(\mathbf{x}_i)^{2/\alpha} \sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^{2/\alpha}}} = \frac{\frac{1}{d_k(\mathbf{x}_i)^{2/\alpha}}}{\sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^{2/\alpha}}}. \quad (3.50)$$

The objective function (3.48), centers (3.49), and membership probabilities (3.50) are the same with (3.35), (3.36), and (3.33) respectively for $\beta = 2$. Thus, FCM follows Principle 2 for $\alpha = m - 1$ and $\beta = 2$. Note that the condition of m is greater than 1 ensures α to be positive. □

3.6.3 Principles of K-Harmonic Means

In Section 3.3, the generalized K-Harmonic Means Algorithm (KHM_θ) is described, and the algorithm's membership probability and weight definition are questioned. In this section, we investigate the relation of the PDC principles with the KHM_θ. From the PDC perspective, the probability and weight definition of KHM_θ are revised.

Proposition 4. *KHM_θ follows Principle 2 in (3.32).*

Proof. Center update procedure of KHM_θ (3.8) can be rearranged as

$$\mathbf{c}_k = \frac{\sum_{i=1}^{N} \frac{\mathbf{x}_i}{d_k(\mathbf{x}_i)^{\theta+2} \left(\sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^\theta} \right)^2}}{\sum_{i=1}^{N} \frac{1}{d_k(\mathbf{x}_i)^{\theta+2} \left(\sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^\theta} \right)^2}} = \frac{\sum_{i=1}^{N} d_k(\mathbf{x}_i)^{\theta-2} \frac{\frac{1}{d_k(\mathbf{x}_i)^\theta}}{\sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^\theta}} \frac{\frac{1}{d_k(\mathbf{x}_i)^\theta}}{\sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^\theta}} \mathbf{x}_i}{\sum_{i=1}^{N} d_k(\mathbf{x}_i)^{\theta-2} \frac{\frac{1}{d_k(\mathbf{x}_i)^\theta}}{\sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^\theta}} \frac{\frac{1}{d_k(\mathbf{x}_i)^\theta}}{\sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^\theta}}}. \quad (3.51)$$

When $\beta = \theta$ and $\alpha = 1$, the term $\frac{1}{d_k(\mathbf{x}_i)^\theta} \Big/ \sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^\theta}$ in (3.51) equals to Principle

2 probabilities, given in (3.33). Thus, (3.51) can be expressed as

$$\mathbf{c}_k = \frac{\sum_{i=1}^{N} d_k(\mathbf{x}_i)^{\theta-2} p_k(\mathbf{x}_i)^2 \mathbf{x}_i}{\sum_{i=1}^{N} d_k(\mathbf{x}_i)^{\theta-2} p_k(\mathbf{x}_i)^2}, \tag{3.52}$$

which is the center update procedure of Principle 2 in (3.36) for $\beta = \theta$ and $\alpha = 1$. $\square$

If one wants to represent centers (3.51) in the form of (3.9), the meaningful expression would be

$$p_k(\mathbf{x}_i) = \frac{\frac{1}{d_k(\mathbf{x}_i)^\theta}}{\sum_{l=1}^{K} \frac{1}{d_k(\mathbf{x}_i)^\theta}} \quad and \quad w(\mathbf{x}_i) = d_k(\mathbf{x}_i)^{\theta-2} \frac{\frac{1}{d_k(\mathbf{x}_i)^\theta}}{\sum_{l=1}^{K} \frac{1}{d_l(\mathbf{x}_i)^\theta}}. \tag{3.53}$$

Moreover, in this particular case of $\beta = \theta$ and $\alpha = 1$, and assuming that optimal probabilities are given, the objective of Principle 2 (3.35) will be equal to the JDF of the whole data set as

$$Obj_{PDF} = \sum_{i=1}^{N} C(\mathbf{x}_i) = \sum_{i=1}^{N} \frac{HA(d_k(\mathbf{x}_i)^\theta |\ k = 1, ..., K)}{K}, \tag{3.54}$$

which is equal to the objective function of KHM_θ (3.7) over the constant K. Therefore, optimizing the objective of KHM_θ and PDC are the same.

3.6.4 Principles of Gravity p-Median Model

We explain the Gravity p-Median Model (GPM) in Section 3.4. We note that probabilities in (3.13) are not optimal solutions to the GPM problem in (3.14). That claim will be proven in this section by Proposition 5. The link between PDC and GPM will be shown. With this linkage, the optimization problem will be revised, and the utility expression and attractiveness of facilities considered by Drezner et al. [13] will be discussed.

Proposition 5. *The probabilities in (3.13) are not optimal solutions of the Gravity p-Median Problem in (3.14).*

Proof. For simplicity consider the case of $K = 2$, the optimization problem in (3.14) is then

$$\begin{aligned} \underset{p_1(\mathbf{x}_i),\, p_2(\mathbf{x}_i)}{\text{minimize}} \quad & \sum_{i=1}^{N} w(\mathbf{x}_i) d_1(\mathbf{x}_i) p_1(\mathbf{x}_i) + w(\mathbf{x}_i) d_2(\mathbf{x}_i) p_2(\mathbf{x}_i) \\ \text{subject to} \quad & p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) = 1, \quad i = 1, \ldots, N, \\ & p_1(\mathbf{x}_i), p_2(\mathbf{x}_i) \geq 0, \quad i = 1, \ldots, N. \end{aligned} \tag{3.55}$$

The Lagrangian of the problem (3.55) is

$$\mathcal{L}(p_1(\mathbf{x}_i), p_2(\mathbf{x}_i), \lambda_i) = \left(\sum_{i=1}^{N} w(\mathbf{x}_i) d_1(\mathbf{x}_i) p_1(\mathbf{x}_i) + w(\mathbf{x}_i) d_2(\mathbf{x}_i) p_2(\mathbf{x}_i) \right) - \lambda_i (p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) - 1)$$

$$\frac{\partial \mathcal{L}}{\partial p_1(\mathbf{x}_i)} = w(\mathbf{x}_i) d_1(\mathbf{x}_i) - \lambda_i = 0, \quad i = 1, ..., N \tag{3.56}$$

$$\frac{\partial \mathcal{L}}{\partial p_2(\mathbf{x}_i)} = w(\mathbf{x}_i) d_2(\mathbf{x}_i) - \lambda_i = 0, \quad i = 1, ..., N \tag{3.57}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) - 1 = 0, \quad i = 1, ..., N$$

Equations (3.56-3.57) yield

$$d_1(\mathbf{x}_i) = d_2(\mathbf{x}_i), \quad i = 1, ..., N, \tag{3.58}$$

which gives alternative solutions for cluster centers that are lying on a line formed by equal distances from data points. Thus, when $N \geq 3$, even a solution may not exist. In addition, the solution is independent of the probability definition in (3.13). □

As stated in Section 3.5.3 , Teboulle [18] and Ben-Israel et al. [9] prove that the classical clustering problem, $min\{d_1, d_2, ..., d_K\}$, is smoothed by the squares of probabilities in objective. Since the Gravity p-Median Model relaxes the assumption that every customer visits the facility closest to him/her of standard p-Median Model (PM) objective, which is

$$Obj_{PM} = \sum_{i=1}^{N} w(\mathbf{x}_i) min\{d_1, d_2, ..., d_K\}, \tag{3.59}$$

the objective function of (3.59) should have been smoothed as follows

$$\begin{aligned} \underset{p_k(\mathbf{x}_i)}{\text{minimize}} \quad & \sum_{i=1}^{N}\sum_{k=1}^{K} w(\mathbf{x}_i)d_k(\mathbf{x}_i)p_k(\mathbf{x}_i)^2 \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i) = 1, \quad i = 1,\ldots,N, \\ & p_k(\mathbf{x}_i) \geq 0, \quad i = 1,\ldots,N \quad \textit{and} \quad k = 1,\ldots,K. \end{aligned} \tag{3.60}$$

The revised GPM in (3.60) is related to the PDC principles as explained in the following proposition

Proposition 6. *When the utility function $u_k(d_k(\boldsymbol{x}_i)) = d_k(\boldsymbol{x}_i)^{-\theta}$, and $\theta = 1$, the revised Gravity p-Median Problem in (3.60) follows Principle 1 in (3.15). Thus, the probabilities in (3.13) are optimality conditions of the revised GPM in (3.60).*

Proof. For simplicity consider the case of $K = 2$, the optimization problem in (3.60) is then

$$\begin{aligned} \underset{p_1(\mathbf{x}_i),\, p_2(\mathbf{x}_i)}{\text{minimize}} \quad & \sum_{i=1}^{N} w(\mathbf{x}_i)d_1(\mathbf{x}_i)p_1(\mathbf{x}_i)^2 + w(\mathbf{x}_i)d_2(\mathbf{x}_i)p_2(\mathbf{x}_i)^2 \\ \text{subject to} \quad & p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) = 1, \quad i = 1,\ldots,N, \\ & p_1(\mathbf{x}_i), p_2(\mathbf{x}_i) \geq 0, \quad i = 1,\ldots,N. \end{aligned} \tag{3.61}$$

The Lagrangian of the problem (3.61) is

$$\begin{aligned} \mathcal{L}(p_1(\mathbf{x}_i), p_2(\mathbf{x}_i), \lambda_i) = & \left(\sum_{i=1}^{N} w(\mathbf{x}_i)d_1(\mathbf{x}_i)p_1(\mathbf{x}_i)^2 + w(\mathbf{x}_i)d_2(\mathbf{x}_i)p_2(\mathbf{x}_i)^2 \right) \\ & - \lambda_i(p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) - 1) \end{aligned}$$

$$\frac{\partial \mathcal{L}}{\partial p_1(\mathbf{x}_i)} = w(\mathbf{x}_i)d_1(\mathbf{x}_i)p_1(\mathbf{x}_i) - \lambda_i = 0, \quad i = 1, ..., N \tag{3.62}$$

$$\frac{\partial \mathcal{L}}{\partial p_2(\mathbf{x}_i)} = w(\mathbf{x}_i)d_2(\mathbf{x}_i)p_2(\mathbf{x}_i) - \lambda_i = 0, \quad i = 1, ..., N \tag{3.63}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) - 1 = 0, \quad i = 1, ..., N$$

Equations (3.62)-(3.63) yield

$$p_1(\mathbf{x}_i)d_1(\mathbf{x}_i) = p_2(\mathbf{x}_i)d_2(\mathbf{x}_i), \quad i = 1, ..., N, \tag{3.64}$$

which gives the Principle 1 in (3.15), and probabilities (3.13) for $u_k(d_k(\mathbf{x}_i)) = d_k(\mathbf{x}_i)^{-\theta}$, where $\theta = 1$. □

A Note on the Utility Expression

In the original GPM, Drezner et al. [13] use the same objective function in (3.12), independent from the utility definition of probabilities (3.13). Even the objective (3.12) is smoothed correctly as in (3.60), this causes inconsistency between optimality conditions of (3.60) and probabilities (3.13). We benefit from the PDC principles to correct this inconsistency as in Proposition 7 and show that utility expression should have been considered in the objective function.

Proposition 7. *For $u_k(d_k(\boldsymbol{x}_i)) = d_k(\boldsymbol{x}_i)^{-\theta}$, GPM follows Principle 2 of PDC (3.32), and the original GPM (3.14) should be revised as follows*

$$\begin{aligned} \underset{p_k(\boldsymbol{x}_i)}{\text{minimize}} \quad & \sum_{i=1}^{N}\sum_{k=1}^{K} w(\boldsymbol{x}_i) d_k(\boldsymbol{x}_i)^{\theta} p_k(\boldsymbol{x}_i)^2 \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(\boldsymbol{x}_i) = 1, \quad i = 1, \dots, N, \\ & p_k(\boldsymbol{x}_i) \geq 0, \quad i = 1, \dots, N \quad \textit{and} \quad k = 1, \dots, K, \end{aligned} \tag{3.65}$$

since

$$p_k(\boldsymbol{x}_i) = \frac{u_k(d_k(\boldsymbol{x}_i))}{\sum\limits_{l=1}^{K} u_l(d_l(\boldsymbol{x}_i))} = \frac{\dfrac{1}{d_k(\boldsymbol{x}_i)^{\theta}}}{\sum\limits_{l=1}^{K} \dfrac{1}{d_l(\boldsymbol{x}_i)^{\theta}}}, \tag{3.66}$$

is the optimal solution of the optimization model in (3.65).

Proof. For simplicity consider the case of $K = 2$, the optimization problem in (3.65) is then

$$\begin{aligned} \underset{p_1(\mathbf{x}_i), p_2(\mathbf{x}_i)}{\text{minimize}} \quad & \sum_{i=1}^{N} w(\mathbf{x}_i) d_1(\mathbf{x}_i)^{\theta} p_1(\mathbf{x}_i)^2 + w(\mathbf{x}_i) d_2(\mathbf{x}_i)^{\theta} p_2(\mathbf{x}_i)^2 \\ \text{subject to} \quad & p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) = 1, \quad i = 1, \dots, N, \\ & p_1(\mathbf{x}_i), p_2(\mathbf{x}_i) \geq 0, \quad i = 1, \dots, N. \end{aligned} \tag{3.67}$$

The Lagrangian of the problem (3.67) is

$$\begin{aligned} \mathcal{L}(p_1(\mathbf{x}_i), p_2(\mathbf{x}_i), \lambda_i) = & \left(\sum_{i=1}^{N} w(\mathbf{x}_i) d_1(\mathbf{x}_i)^{\theta} p_1(\mathbf{x}_i)^2 + w(\mathbf{x}_i) d_2(\mathbf{x}_i)^{\theta} p_2(\mathbf{x}_i)^2 \right) \\ & - \lambda_i (p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) - 1) \end{aligned}$$

$$\frac{\partial \mathcal{L}}{\partial p_1(\mathbf{x}_i)} = w(\mathbf{x}_i) d_1(\mathbf{x}_i)^\theta p_1(\mathbf{x}_i) - \lambda_i = 0, \quad i = 1, ..., N \tag{3.68}$$

$$\frac{\partial \mathcal{L}}{\partial p_2(\mathbf{x}_i)} = w(\mathbf{x}_i) d_2(\mathbf{x}_i)^\theta p_2(\mathbf{x}_i) - \lambda_i = 0, \quad i = 1, ..., N \tag{3.69}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) - 1 = 0, \quad i = 1, ..., N$$

Equations (3.68)-(3.69) yield

$$p_1(\mathbf{x}_i) d_1(\mathbf{x}_i)^\theta = p_2(\mathbf{x}_i) d_2(\mathbf{x}_i)^\theta. \quad i = 1, ..., N, \tag{3.70}$$

Equality in (3.70) is the Principle 2 in (3.32), and probabilities in (3.66) are the memberships of the Principle 2 in (3.33), where $\beta = \theta$, and $\alpha = 1$. □

Similarly, for $u_k(d_k(\mathbf{x}_i)) = e^{-\theta d_k(\mathbf{x}_i)}$, the original GPM (3.14) would be

$$\begin{aligned} \underset{p_k(\mathbf{x}_i)}{\text{minimize}} \quad & \sum_{i=1}^{N} \sum_{k=1}^{K} w(\mathbf{x}_i) e^{-\theta d_k(\mathbf{x}_i)} p_k(\mathbf{x}_i)^2 \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i) = 1, \quad i = 1, \ldots, N, \\ & p_k(\mathbf{x}_i) \geq 0, \quad i = 1, \ldots, N \quad and \quad k = 1, \ldots, K, \end{aligned} \tag{3.71}$$

and optimality conditions of (3.71) follows Principle 3 in (3.37).

Attractiveness of Facilities

Drezner et al. [13] consider the attractiveness of facilities by multiplying utilities of the membership probability in (3.13) with attractiveness value of the facilities as follows

$$p_k(\mathbf{x}_i) = \frac{a_k u_k(d_k(\mathbf{x}_i))}{\sum_{l=1}^{K} a_l u_l(d_l(\mathbf{x}_i))}, \tag{3.72}$$

where a_k is the attractiveness of facility k.

That is conceptually sensible as (3.72) means that the probability of a customer i visits facility k is proportional to the attractiveness of the facility and inversely proportional to its distance. However, the objective function in (3.12) is not adjusted for attractiveness values, which results in a contradiction between attractiveness-included probabilities (3.72) and the optimality conditions of the model.

Attractiveness can be measured by multiplication of individual measures depends on each retail facility such as store areas, brand attraction, concept of the store etc. [15]. These measures somehow estimate the number of customers that a store can draw on. Therefore, attractiveness resembles cluster sizes in the clustering concept. This guides us to Principle 4 in (3.42), where cluster sizes are considered in the PDC method.

For $u_k(d_k(\mathbf{x}_i)) = d_k(\mathbf{x}_i)^{-\theta}$, and attractiveness is a_k, GPM probabilities in (3.72) follows Principle 2 (3.32) and Principle 4 (3.42). Thus, the GPM objective (3.12) will be revised as

$$Obj_{GPM} = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{w(\mathbf{x}_i) d_k(\mathbf{x}_i)^{\theta} p_k(\mathbf{x}_i)^2}{a_k}.$$

Similarly, when $u_k(d_k(\mathbf{x}_i)) = e^{-\theta d_k(\mathbf{x}_i)}$, and attractiveness is a_k, GPM probabilities in (3.72) follows Principle 3 (3.37) and Principle 4 (3.42). Therefore, the GPM objective (3.12) will be reconstructed as

$$Obj_{GPM} = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{w(\mathbf{x}_i) e^{\theta d_k(\mathbf{x}_i)} p_k(\mathbf{x}_i)^2}{a_k}.$$

3.7 Conclusion

Chapter 3 discusses well-known probabilistic approaches from different concepts; the Huff Model from marketing, the Fuzzy c-Means and K-Harmonic Means Algorithms from clustering and the Gravity p-Median Model from the location literature. We show that principles defined in the Probabilistic D-Clustering approach explain the other methods both conceptually and theoretically. We prove the Huff Model follows PDC principles, and the Fuzzy c-Means and K-Harmonic Means Algorithms are special cases of PDC approach. We also prove that Gravity p-Median Model has inconsistencies with its probability and objective function definitions. Using the PDC principles, the model is revised. With this contribution, instead of restricting the GPM problem solutions as combinatorial, problem-based heuristics can be developed to obtain solutions anywhere on the network.

CHAPTER 4

PROBABILISTIC DISTANCE CLUSTERING FOR TWO-MODE CLUSTERING PROBLEM

In a classical clustering setting, data are usually stored in a matrix **X**, which has two dimensions or *ways*. Carroll and Arabie [20] define the term *mode* as a particular class of entities. The distinct sets of entities indexed by the *ways* correspond to the *modes* of a matrix. Let $\mathcal{R}$ and $\mathcal{C}$ be sets of row and column entities of a data set, respectively. **One-mode two-way** data sets consist of $\mathcal{R}$ and $\mathcal{C}$, which are identical sets. Similarity, dissimilarity, or proximity matrices are examples of this type of data sets. On the other hand, for **two-mode two-way** data sets, $\mathcal{R}$ and $\mathcal{C}$ are distinct. Thus, elements x_{ij} of data sets are values of an indicator that relates entities of row and column modes. An indicator can be representative of dependencies, trends, linkage, preference, or frequency. Figure 4.1 summarizes the discussion on the one-mode and two-mode data sets.

Two-Mode Clustering is a technique to cluster row and column entities of two-mode two-way data sets to obtain compact partitions (submatrices). In other words, two-mode clustering divides $\mathcal{R}$ and $\mathcal{C}$ into subsets $\{S_k \mid k = 1, \ldots, K\}$ and $\{T_l \mid l = 1, \ldots, L\}$, respectively. The submatrices are formed by cartesian product of $\{S_k \mid k = 1, \ldots, K\}$ and $\{T_l \mid l = 1, \ldots, L\}$. With these submatrices, one can understand the relationship between the first and second modes. The question of the Two-Mode Clustering Problem is **"Which group of row entities are related with which group of column entities in what level?"**.

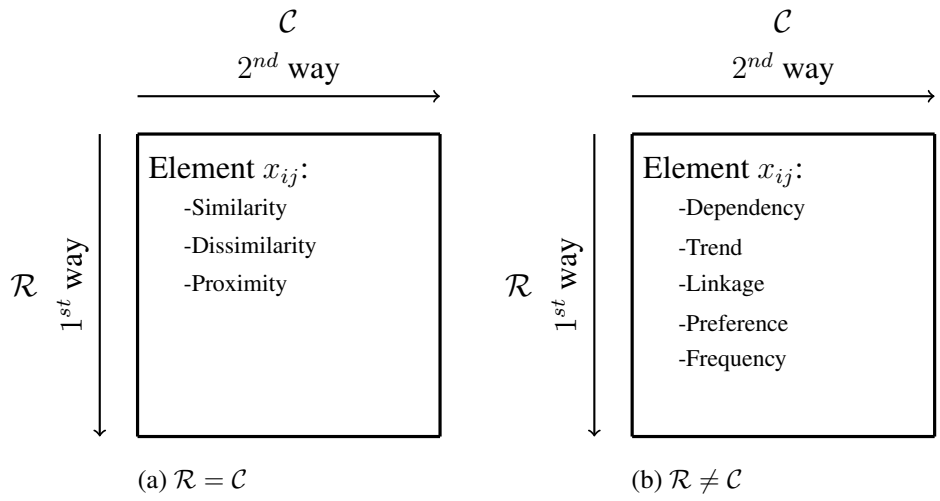


Figure 4.1: Representation of modes, (a) one-mode two-way data set, (b) two-mode two-way data set

In the literature, the term *biclustering* is used interchangeably for two-mode clustering. However, in biclustering problems, $\mathcal{R}$ refers to a set of entities, whereas $\mathcal{C}$ represents a set of features. Biclustering, clusters $\mathcal{R}$ and $\mathcal{C}$ simultaneously to find which groups of entities are explained by which group of features.

Two-Mode Clustering is used for problems from a variety of disciplines. One of them is group cell technology in a manufacturing system, which aims to form production cells with part-machine groupings [21]. With a Two-Mode Clustering method, an industrial engineer can understand which group of parts are frequently processed by which group of machines and design a manufacturing layout accordingly. In marketing, brand switching problems are solved by Two-Mode Clustering [22]. Entries of brand switching data sets are preference transitions of consumers from time t to $t+1$. A brand switching data set solution reveals that which group of goods preferred today is frequently replaced by which group of goods in the future and can be used for marketing or advertisement purposes. One other marketing application is for market segmentation to determine the subgroups of customers having specific subsets of customer pain points [23]. Two-Mode Clustering is also a useful approach in social network problems. Social proximity between two groups of people can be evaluated by a Two-Mode Clustering method [24], [25]. According to Rathipriya et al. [26],

on web usage data, by using a two-mode clustering approach, hidden browsing patterns can be found and used to identify the web user-web page interactions to be used for *e*-commerce. Another central usage area is text mining. Balbi [27] explains that Two-Mode Clustering is very useful for text mining. Instead of single words, looking at clusters of words to cluster documents is more content revealing and decreases the information complication. Another application field is the university performance study at which, by Two-Mode Clustering, performance is not measured for a single activity metric, but it is estimated in a multi-dimensional framework manner [28]. Moreover, Two-Mode Clustering is applicable for biomedical data sets, where a set of properties is linked to molecular units or patients [29]. It is also used for clinical diagnostic purposes [30] to understand which group of symptoms are more likely to be observed in which group of diseases [31].

There are three categories of Two-Mode Clustering: *partitioning*, *nested clustering*, and *overlapping clustering* [31]. In partitioning, clusters are non-empty, disjoint, and contain the complete set of modes (Figure 4.2a). In nested clustering, clusters may intersect. However, in that case, the intersection must be in the form of subsets in a hierarchical manner (Figure 4.2b). Lastly, overlapping clustering also allows intersections, but those can be non-nested (Figure 4.2c). Among these categories, the most studied one in the literature is the Two-Mode Partitioning (TMP) problem. We introduce the TMP problem and a review of solution methods in Section 4.1.

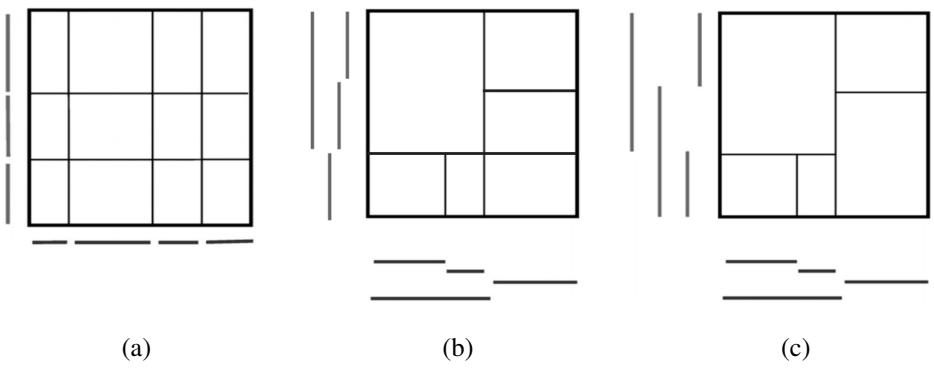
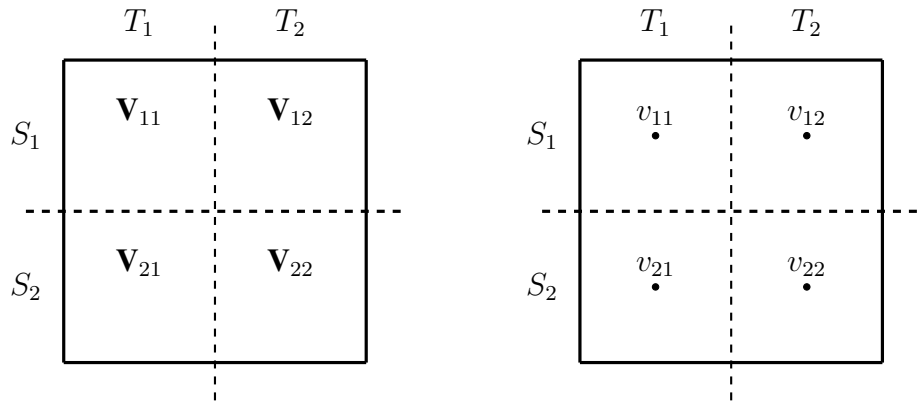


Figure 4.2: Categories of Two-Mode Clustering; (a) partitioning, (b) nested clustering, (c) overlapping clustering

4.1 Two-Mode Partitioning Problem

Let $\mathbf{X}$ be a two-mode data set with N row and M column entities. The elements of data set $\mathbf{X}$ are x_{ij} that relate the row and column modes. A row entity is a row vector denoted by $\mathbf{x}_i^r$. Similarly, a column entity is represented by column vector $\mathbf{x}_j^c$. In the Two-Mode Partitioning problem, the row and column entities are assigned into K and L clusters, respectively. The assigned row entities to a row cluster k are stored in a set, S_k. Correspondingly, T_l is a set of column entities assigned to a column cluster l. Therefore, entities in S_k and T_l form KL partitions (submatrices) denoted by $\mathbf{V}_{kl}$. Submatrix centers v_{kl} are calculated by taking averages of elements $x_{ij} \in \mathbf{V}_{kl}$. A representation of TMP for $K = L = 2$ is given in Figure 4.3.



(a) Partitions (submatrices)

(b) Partition (submatrix) centers

Figure 4.3: Two-Mode Partitioning (TMP) problem for $K = L = 2$

The objective of TMP is to minimize the total deviation between assigned elements of submatrices, $x_{ij} \in \mathbf{V}_{kl}$, and submatrix centers v_{kl}. In the literature, it is common to use squared Euclidean distance for criterion as in (4.1).

$$
\begin{aligned}
\underset{p_k(\mathbf{x}_i^r),\, q_l(\mathbf{x}_j^c),\, v_{kl}}{\text{minimize}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(\mathbf{x}_i^r) q_l(\mathbf{x}_j^c)(x_{ij} - v_{kl})^2 \\
\text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i^r) = 1, \quad i = 1, \ldots, N, \\
& \sum_{l=1}^{L} q_l(\mathbf{x}_j^c) = 1, \quad j = 1, \ldots, M, \\
& \sum_{i=1}^{N} p_k(\mathbf{x}_i^r) > 0, \quad k = 1, \ldots, K, \\
& \sum_{j=1}^{M} q_l(\mathbf{x}_j^c) > 0, \quad l = 1, \ldots, L, \\
& p_k(\mathbf{x}_i^r), q_l(\mathbf{x}_j^c) \in \{0, 1\}, \quad \forall i, j, k, l.
\end{aligned}
\tag{4.1}
$$

The decision variable $p_k(\mathbf{x}_i^r)$ is binary and takes a value of 1, if a row entity $\mathbf{x}_i^r$ is assigned to a row cluster k. Similarly, $q_l(\mathbf{x}_j^c)$ becomes 1, when a column $\mathbf{x}_j^c$ is assigned to a column cluster l. The first two constraints with the fifth one of (4.1) ensure that each row and column entity is assigned to exactly one cluster. The third and fourth constraints are for avoiding empty clusters.

Brusco and Doreian [32] compute the number of solutions for the TMP as follows

$$
\left[\frac{1}{K!}\sum_{k=0}^{K}(-1)^k\binom{K}{k}(K-k)^N\right] \times \left[\frac{1}{L!}\sum_{l=0}^{L}(-1)^l\binom{L}{l}(L-l)^M\right] \tag{4.2}
$$

Even in the smallest scale, for $N = M = 10$ and $K = L = 3$, the approximate number of possible solutions becomes 8.705×10^7.

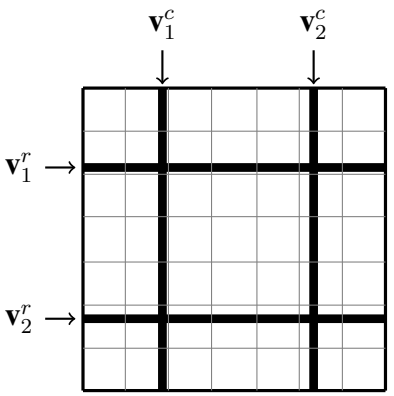
According to the overview paper of Mechelen et al. [31], the solution methodology for Two-Mode Clustering (TMC) is divided into two main classes, which are *indirect* and *direct* clustering. In indirect clustering, partitions are obtained by clustering row and column objects successively with the classical one-mode clustering methods [33]. The direct clustering partitions the first and second modes simultaneously, which will be our focus in this thesis. We categorize direct clustering methods for the Two-Mode Partitioning problem as *exact algorithms*, *problem-specific heuristics*, and *meta-heuristics*. An exact algorithm is proposed by Brusco and Doreian [32], which deploys a branch and bound procedure. But it is only applicable for small data

sets. The authors state that for a 19×19 data set and $K = L = 3$, the computation takes for less than one minute, whereas for $K = L = 6$, it takes approximately three hours. Vichi [34] modifies MacQueen's K-Means Algorithm [5] for TMP problem. This algorithm is called *Two-Mode KL-Means Partitioning*. Moreover, a Two-Mode Fuzzy c-Means Algorithm is proposed by Rosmalen et al. [35], which is an extension of the classical Fuzzy c-Means Algorithm [7] to two-mode data sets. This algorithm converges to soft partitions. However, this algorithm is only used as an interstep to Two-Mode KL-Means Partitioning, which is a hard partitioning algorithm. They also deploy an adoption of the Fuzzy Steps method proposed by Heiser and Groenen [36] that gradually lowers the fuzzy parameter in each step. Therefore, soft partitioning solutions are not discussed. Meta-heuristic approaches are also implemented to TMP problems. Trejos and Castillo perform simulated annealing [37] and tabu search [38] heuristics. An integer-coded genetic algorithm is presented by Hansohm [39]. In addition, Brusco and Doreian propose a real-coded genetic algorithm, which uses partition centers as gene expression [40]. Rosmalen et al. [35] compare the Two-Mode KL-Means Partitioning [34] , simulated annealing [37], tabu search [38], and Two-Mode Fuzzy c-Means [35] algorithms with a simulation study. According to this study, the multi start Two-Mode KL-Means Algorithm performs better compared to other approaches. However, the authors note that when the data set is hard to partition, i.e. when the submatrix structures in data set are not well-defined, multi start Two-Mode Fuzzy c-Means algorithm has the best performance.

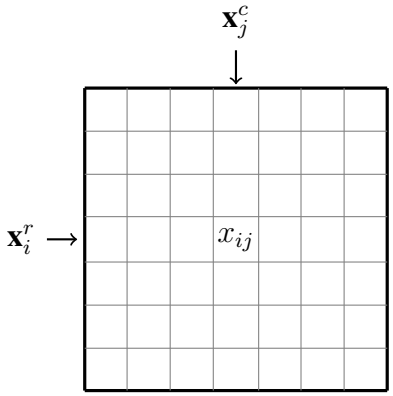
In this thesis, we aim to develop a soft clustering approach to solve TMC problems in order to provide more information to its users. A soft partitioning solution by itself can be used to interpret the relationship between modes. If it is needed the soft clustering algorithm can be easily modified to obtain crisp partitions. In addition, it may yield overlapping clusters, i.e., a row entity may be assigned to two row clusters with equal membership probabilities. Thus, our proposed approach will be useful because it will be a general solution method for both partitioning (Figure 4.2a) and overlapping (Figure 4.2c) categories of Two-Mode Clustering problems. To develop such a method, we follow the Probabilistic D-Clustering (PDC) introduced in Chapter 3. Based on PDC principles, we present two novel algorithms in the following sections.

4.2 Two-Mode Probabilistic Distance Clustering

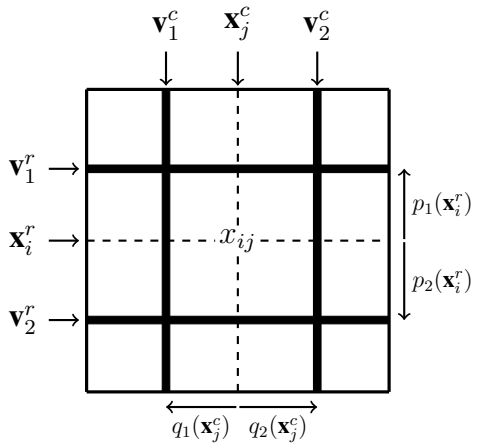
In the soft TMC, each entity may belong to more than one cluster with some probability. Therefore, $p_k(\mathbf{x}_i^r) \in [0,1]$ is the membership probability of a row entity $\mathbf{x}_i^r$ is assigned to a row cluster k. Similarly, $q_l(\mathbf{x}_j^c) \in [0,1]$ is the membership probability of a column entity $\mathbf{x}_j^c$ is assigned to a column cluster l. The center of a row cluster k, and a column cluster l are denoted by $\mathbf{v}_k^r$ and $\mathbf{v}_l^c$, respectively. In Figure 4.4a, representation of centers are given for $K = L = 2$. An element x_{ij} belongs to a row entity $\mathbf{x}_i^r$ and column entity $\mathbf{x}_j^c$ as shown in Figure 4.4b. Position of x_{ij} in data set $\mathbf{X}$ depends on positions of $\mathbf{x}_i^r$ and $\mathbf{x}_j^c$.



(a) Row and column cluster centers



(b) An element of $\mathbf{X}$



(c) Membership probabilities of entities to cluster centers

Figure 4.4: Soft Two-Mode Clustering (TMC) problem for $K = L = 2$

Movements of $\mathbf{x}_i^r$ and $\mathbf{x}_j^c$ are provided by membership probabilities $p_k(\mathbf{x}_i^r)$ and $q_l(\mathbf{x}_j^c)$ as shown in Figure 4.4c. Membership probabilities depend on distances between entities and cluster centers as it will be explained in Section 4.2.1. The distance between a row entity $\mathbf{x}_i^r$ and row cluster center $\mathbf{v}_k^r$ is denoted by $d_k(\mathbf{x}_i^r)$. Likewise, the distance between a column entity $\mathbf{x}_j^c$ and column cluster center $\mathbf{v}_l^c$ is indicated by $\bar{d}_l(\mathbf{x}_j^c)$.

In Section 4.2.1, we explain how we adopt PDC principles for two-mode data sets. In Sections 4.2.2-4.2.6, the derivations of probabilities and center update procedure are introduced. We discuss the underlying optimization model and optimality conditions in Section 4.2.7. Later, the steps of the algorithm are provided in Section 4.2.8.

4.2.1 Principles

In the soft Two-Mode Clustering problem (TMC) both row and column entities are required to be clustered into K, and L many clusters, respectively. Thus, two principles for clustering row and column entities are proposed.

Principle 1. *For each row entity $\boldsymbol{x}_i^r \in \boldsymbol{X}$, and each row cluster k,*

$$p_k(\mathbf{x}_i^r)d_k(\mathbf{x}_i^r) = A(\mathbf{x}_i^r), \tag{4.3}$$

where $A(\boldsymbol{x}_i^r) =$ a constant, depending on $\boldsymbol{x}_i^r$.

Membership probability of row entity $\mathbf{x}_i^r$ to row cluster k is higher when the row entity is closer to the cluster center.

Note that problem properties imply

$$p_k(\mathbf{x}_i^r) = p_k(x_{i1}) = p_k(x_{i2}) = ... = p_k(x_{iM}), \quad i = 1, .., N, \tag{4.4}$$

which means that each element of a row entity, $x_{ij} \in \mathbf{x}_i^r$, has the same membership probability of $p_k(\mathbf{x}_i^r)$.

Principle 2. *For each column entity $\boldsymbol{x}_j^c \in \boldsymbol{X}$, and each column cluster l,*

$$q_l(\mathbf{x}_j^c)\bar{d}_l(\mathbf{x}_j^c) = B(\mathbf{x}_j^c), \tag{4.5}$$

where $B(\boldsymbol{x}_j^c) =$ a constant, depending on $\boldsymbol{x}_j^c$.

Membership probability of column entity $\mathbf{x}_j^c$ to column cluster l is higher when the column entity is closer to the cluster center.

Similarly, problem properties imply

$$q_l(\mathbf{x}_j^c) = q_l(x_{1j}) = q_l(x_{2j}) = ... = q_l(x_{Nj}), \quad j = 1, .., M, \tag{4.6}$$

which means that each element of a column entity, $x_{ij} \in \mathbf{x}_j^c$, has the same membership probability of $q_l(\mathbf{x}_j^c)$.

Lemma 1. Assuming $\mathbf{x}_i^r$ and $\mathbf{x}_j^c$ follow Principle 1 in (4.3) and Principle 2 in (4.5) respectively, the following equality can be obtained

$$p_k(\mathbf{x}_i^r)q_l(\mathbf{x}_j^c)d_k(\mathbf{x}_i^r)\bar{d}_l(\mathbf{x}_j^c) = A(\mathbf{x}_i^r)B(\mathbf{x}_j^c), \tag{4.7}$$

(4.4), and (4.6) with equality (4.7) imply

$$p_k(x_{ij})q_l(x_{ij})d_k(\mathbf{x}_i^r)\bar{d}_l(\mathbf{x}_j^c) = A(\mathbf{x}_i^r)B(\mathbf{x}_j^c). \tag{4.8}$$

By the equation (4.8), we derive a joint principle for each element $x_{ij} \in \mathbf{X}$ as follows.

Principle 3. *Following the Lemma 1, for each $\boldsymbol{x}_i^r, \boldsymbol{x}_j^c \in \boldsymbol{X}$, and each $\boldsymbol{V}_{kl}$ partition, for $x_{ij} \in \boldsymbol{x}_i^r$ and $x_{ij} \in \boldsymbol{x}_j^c$,*

$$R_{kl}(x_{ij})D_{kl}(\mathbf{x}_i^r, \mathbf{x}_j^c) = Z(x_{ij}), \tag{4.9}$$

where $R_{kl}(x_{ij}) = p_k(x_{ij})q_l(x_{ij})$ is the probability of an element x_{ij} is assigned to partition $\boldsymbol{V}_{kl}$, $D_{kl}(\boldsymbol{x}_i^r, \boldsymbol{x}_j^c) = d_k(\boldsymbol{x}_i^r)\bar{d}_l(\boldsymbol{x}_j^c)$ is the joint distance function of $\boldsymbol{x}_i^r$ and $\boldsymbol{x}_j^c$, and $Z(x_{ij}) = A(\boldsymbol{x}_i^r)B(\boldsymbol{x}_j^c)$ is a constant depending on x_{ij}.

Remark: The fact that $\sum_{k=1}^{K} p_k(x_{ij}) = 1$, and $\sum_{l=1}^{L} q_l(x_{ij}) = 1$ imply

$$\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(x_{ij})q_l(x_{ij}) = \sum_{k=1}^{K}\sum_{l=1}^{L} R_{kl}(x_{ij}) = 1.$$

An Explicit Example

Assume $N = M = 4$ and $K = L = 2$, and we follow Principle 3 for an element x_{34} to the submatrix $\mathbf{V}_{12}$.

Consider the membership of $\mathbf{x}_3^r$ to row cluster 1, and membership of $\mathbf{x}_4^c$ to column cluster 2, then using (4.3) and (4.5), we write

$$p_1(\mathbf{x}_3^r)d_1(\mathbf{x}_3^r) = A(\mathbf{x}_3^r) \quad \text{and} \quad q_2(\mathbf{x}_4^c)\bar{d}_2(\mathbf{x}_4^c) = B(\mathbf{x}_4^c). \tag{4.10}$$

We know from (4.4) and (4.6) that

$$p_1(\mathbf{x}_3^r) = p_1(x_{34}) \quad \text{and} \quad q_2(\mathbf{x}_4^c) = q_2(x_{34}). \tag{4.11}$$

Using (4.10) and (4.11)

$$p_1(x_{34})d_1(\mathbf{x}_3^r) = A(\mathbf{x}_3^r) \quad \text{and} \quad q_2(x_{34})\bar{d}_2(\mathbf{x}_4^c) = B(\mathbf{x}_4^c).$$

Therefore,

$$p_1(x_{34})q_2(x_{34})d_1(\mathbf{x}_3^r)\bar{d}_2(\mathbf{x}_4^c) = A(\mathbf{x}_3^r)B(\mathbf{x}_4^c).$$

4.2.2 Probabilities

Using the principles in Section 4.2.1, we find the membership probabilities of row and column entities. Moreover, joint membership probabilities are obtained by the problem properties explained in Section 4.2.1.

Row Membership Probabilities

From Principle 1, and the fact that probabilities $p_k(\mathbf{.})$ add to one, we get

Theorem 4. *Let the row cluster centers $\{\boldsymbol{v}_1^r, \boldsymbol{v}_2^r, ..., \boldsymbol{v}_K^r\}$ be given, let $\boldsymbol{x}_i^r$ be a row entity of data set $\boldsymbol{X}$, and let $\{d_k(\boldsymbol{x}_i^r) : k = 1, ..., K\}$ be its distance from the given cluster centers. Then the membership probabilities of $\boldsymbol{x}_i^r$ are*

$$p_k(\boldsymbol{x}_i^r) = \frac{\dfrac{1}{d_k(\boldsymbol{x}_i^r)}}{\displaystyle\sum_{s=1}^{K} \frac{1}{d_s(\boldsymbol{x}_i^r)}}. \tag{4.12}$$

Proof. Using (4.3) we write for s,k,

$$p_s(\mathbf{x}_i^r) = \left(\frac{p_k(\mathbf{x}_i^r)d_k(\mathbf{x}_i^r)}{d_s(\mathbf{x}_i^r)}\right).$$

Since $\sum_{s=1}^{K} p_s(\mathbf{x}_i^r) = 1$,

$$p_k(\mathbf{x}_i^r) d_k(\mathbf{x}_i^r) \sum_{s=1}^{K} \left(\frac{1}{d_s(\mathbf{x}_i^r)} \right) = 1.$$

$$p_k(\mathbf{x}_i^r) = \frac{1}{d_k(\mathbf{x}_i^r) \sum_{s=1}^{K} \left(\frac{1}{d_s(\mathbf{x}_i^r)} \right)} = \frac{\frac{1}{d_k(\mathbf{x}_i^r)}}{\sum_{s=1}^{K} \frac{1}{d_s(\mathbf{x}_i^r)}}.$$

□

In particular, for $K = 2$,

$$p_1(\mathbf{x}_i^r) = \frac{\frac{1}{d_1(\mathbf{x}_i^r)}}{\frac{1}{d_1(\mathbf{x}_i^r)} + \frac{1}{d_2(\mathbf{x}_i^r)}}, \quad p_2(\mathbf{x}_i^r) = \frac{\frac{1}{d_2(\mathbf{x}_i^r)}}{\frac{1}{d_1(\mathbf{x}_i^r)} + \frac{1}{d_2(\mathbf{x}_i^r)}}, \tag{4.13}$$

and using (4.4) we may write

$$p_1(x_{ij}) = \frac{\frac{1}{d_1(\mathbf{x}_i^r)}}{\frac{1}{d_1(\mathbf{x}_i^r)} + \frac{1}{d_2(\mathbf{x}_i^r)}}, \quad p_2(x_{ij}) = \frac{\frac{1}{d_2(\mathbf{x}_i^r)}}{\frac{1}{d_1(\mathbf{x}_i^r)} + \frac{1}{d_2(\mathbf{x}_i^r)}}.$$

Column Membership Probabilities

From Principle 2, and the fact that probabilities $q_l(\centerdot)$ add to one, we get

Theorem 5. *Let the column cluster centers $\{\boldsymbol{v}_1^c, \boldsymbol{v}_2^c, ..., \boldsymbol{v}_L^c\}$ be given, let $\boldsymbol{x}_j^c$ be a column entity of data set $\boldsymbol{X}$, and let $\{\bar{d}_l(\boldsymbol{x}_j^c) : l = 1, ..., L\}$ be its distance from the given cluster centers. Then the membership probabilities of $\boldsymbol{x}_j^c$ are*

$$q_l(\boldsymbol{x}_j^c) = \frac{\frac{1}{\bar{d}_l(\boldsymbol{x}_j^c)}}{\sum_{t=1}^{L} \bar{d}_t(\boldsymbol{x}_j^c)}. \tag{4.14}$$

Proof. Using (4.5) we write for t,l,

$$q_t(\mathbf{x}_j^c) = \left(\frac{q_l(\mathbf{x}_j^c) \bar{d}_l(\mathbf{x}_j^c)}{\bar{d}_t(\mathbf{x}_j^c)} \right).$$

Since $\sum_{t=1}^{L} q_t(\mathbf{x}_j^c) = 1$,

$$q_l(\mathbf{x}_j^c)\bar{d}_l(\mathbf{x}_j^c)\sum_{t=1}^{L}\left(\frac{1}{\bar{d}_t(\mathbf{x}_j^c)}\right) = 1.$$

$$q_l(\mathbf{x}_j^c) = \frac{1}{\bar{d}_l(\mathbf{x}_j^c)\sum_{t=1}^{L}\left(\frac{1}{\bar{d}_t(\mathbf{x}_j^c)}\right)} = \frac{\frac{1}{\bar{d}_l(\mathbf{x}_j^c)}}{\sum_{t=1}^{L}\frac{1}{\bar{d}_t(\mathbf{x}_j^c)}}.$$

□

In particular, for $L = 2$,

$$q_1(\mathbf{x}_j^c) = \frac{\frac{1}{\bar{d}_1(\mathbf{x}_j^c)}}{\frac{1}{\bar{d}_1(\mathbf{x}_j^c)} + \frac{1}{\bar{d}_2(\mathbf{x}_j^c)}}, \quad q_2(\mathbf{x}_j^c) = \frac{\frac{1}{\bar{d}_2(\mathbf{x}_j^c)}}{\frac{1}{\bar{d}_1(\mathbf{x}_j^c)} + \frac{1}{\bar{d}_2(\mathbf{x}_j^c)}}, \tag{4.15}$$

and using (4.6) we may write

$$q_1(x_{ij}) = \frac{\frac{1}{\bar{d}_1(\mathbf{x}_j^c)}}{\frac{1}{\bar{d}_1(\mathbf{x}_j^c)} + \frac{1}{\bar{d}_2(\mathbf{x}_j^c)}}, \quad q_2(x_{ij}) = \frac{\frac{1}{\bar{d}_2(\mathbf{x}_j^c)}}{\frac{1}{\bar{d}_1(\mathbf{x}_j^c)} + \frac{1}{\bar{d}_2(\mathbf{x}_j^c)}}.$$

The Joint Membership Probabilities

From Principle 3, and the fact that probabilities $R_{kl}(\mathbf{.})$ add to one, we get

Theorem 6. *Let the row cluster centers $\{\boldsymbol{v}_1^r, \boldsymbol{v}_2^r, ..., \boldsymbol{v}_K^r\}$ and column cluster centers $\{\boldsymbol{v}_1^c, \boldsymbol{v}_2^c, ..., \boldsymbol{v}_L^c\}$ be given, x_{ij} be an element of data set $\boldsymbol{X}$, and $\{D_{kl}(\boldsymbol{x}_i^r, \boldsymbol{x}_j^c) = d_k(\boldsymbol{x}_i^r)\bar{d}_l(\boldsymbol{x}_j^c) \mid k = 1, ..., K,\ l = 1, ..., L\}$ be its joint distance from the given cluster centers. Then the membership probabilities of x_{ij} are*

$$R_{kl}(x_{ij}) = \frac{\frac{1}{D_{kl}(\boldsymbol{x}_i^r, \boldsymbol{x}_j^c)}}{\sum_{s=1}^{K}\sum_{t=1}^{L}\frac{1}{D_{st}(\boldsymbol{x}_i^r, \boldsymbol{x}_j^c)}}. \tag{4.16}$$

Proof. Using (4.9) we write for s,t,k,l,

$$R_{st}(x_{ij}) = \left(\frac{R_{kl}(x_{ij})D_{kl}(\mathbf{x}_i^r, \mathbf{x}_j^c)}{D_{st}(\mathbf{x}_i^r, \mathbf{x}_j^c)}\right).$$

Since $\sum_{s=1}^{K}\sum_{t=1}^{L} R_{st}(x_{ij}) = 1$,

$$R_{kl}(x_{ij}) D_{kl}(\mathbf{x}_i^r, \mathbf{x}_j^c) \sum_{s=1}^{K}\sum_{t=1}^{L} \left(\frac{1}{D_{st}(\mathbf{x}_i^r, \mathbf{x}_j^c)} \right) = 1.$$

$$R_{kl}(x_{ij}) = \frac{1}{D_{kl}(\mathbf{x}_i^r, \mathbf{x}_j^c) \sum_{s=1}^{K}\sum_{t=1}^{L} \left(\frac{1}{D_{st}(\mathbf{x}_i^r, \mathbf{x}_j^c)} \right)} = \frac{\frac{1}{D_{kl}(\mathbf{x}_i^r, \mathbf{x}_j^c)}}{\sum_{s=1}^{K}\sum_{t=1}^{L} \frac{1}{D_{st}(\mathbf{x}_i^r, \mathbf{x}_j^c)}}.$$

□

Remark: From (4.12)-(4.16),

$$R_{kl}(x_{ij}) = \frac{\frac{1}{D_{kl}(\mathbf{x}_i^r, \mathbf{x}_j^c)}}{\sum_{s=1}^{K}\sum_{t=1}^{L} \frac{1}{D_{st}(\mathbf{x}_i^r, \mathbf{x}_j^c)}} = \frac{\frac{1}{d_k(\mathbf{x}_i^r)\bar{d}_l(\mathbf{x}_j^c)}}{\sum_{s=1}^{K}\sum_{t=1}^{L} \frac{1}{d_s(\mathbf{x}_i^r)\bar{d}_t(\mathbf{x}_j^c)}}$$
$$= \frac{\frac{1}{d_k(\mathbf{x}_i^r)\bar{d}_l(\mathbf{x}_j^c)}}{\sum_{s=1}^{K} \frac{1}{d_s(\mathbf{x}_i^r)} \sum_{t=1}^{L} \frac{1}{\bar{d}_t(\mathbf{x}_j^c)}} = p_k(x_{ij}) q_l(x_{ij}).$$

In particular, for $K = L = 2$,

$$R_{11}(x_{ij}) = \frac{\frac{1}{D_{11}(\mathbf{x}_i^r, \mathbf{x}_j^c)}}{\frac{1}{D_{11}(\mathbf{x}_i^r, \mathbf{x}_j^c)} + \frac{1}{D_{12}(\mathbf{x}_i^r, \mathbf{x}_j^c)} + \frac{1}{D_{21}(\mathbf{x}_i^r, \mathbf{x}_j^c)} + \frac{1}{D_{22}(\mathbf{x}_i^r, \mathbf{x}_j^c)}},$$

$$R_{12}(x_{ij}) = \frac{\frac{1}{D_{12}(\mathbf{x}_i^r, \mathbf{x}_j^c)}}{\frac{1}{D_{11}(\mathbf{x}_i^r, \mathbf{x}_j^c)} + \frac{1}{D_{12}(\mathbf{x}_i^r, \mathbf{x}_j^c)} + \frac{1}{D_{21}(\mathbf{x}_i^r, \mathbf{x}_j^c)} + \frac{1}{D_{22}(\mathbf{x}_i^r, \mathbf{x}_j^c)}},$$

$$R_{21}(x_{ij}) = \frac{\frac{1}{D_{21}(\mathbf{x}_i^r, \mathbf{x}_j^c)}}{\frac{1}{D_{11}(\mathbf{x}_i^r, \mathbf{x}_j^c)} + \frac{1}{D_{12}(\mathbf{x}_i^r, \mathbf{x}_j^c)} + \frac{1}{D_{21}(\mathbf{x}_i^r, \mathbf{x}_j^c)} + \frac{1}{D_{22}(\mathbf{x}_i^r, \mathbf{x}_j^c)}},$$

$$R_{22}(x_{ij}) = \frac{\frac{1}{D_{22}(\mathbf{x}_i^r, \mathbf{x}_j^c)}}{\frac{1}{D_{11}(\mathbf{x}_i^r, \mathbf{x}_j^c)} + \frac{1}{D_{12}(\mathbf{x}_i^r, \mathbf{x}_j^c)} + \frac{1}{D_{21}(\mathbf{x}_i^r, \mathbf{x}_j^c)} + \frac{1}{D_{22}(\mathbf{x}_i^r, \mathbf{x}_j^c)}}.$$

4.2.3 Extremal Principles

For simplicity consider the case of two row clusters and two column clusters, $K = L = 2$, (the results are easily extended to any row and column clusters).

Let $\mathbf{x}_i^r$ be a given row entity with distances $d_1(\mathbf{x}_i^r)$, $d_2(\mathbf{x}_i^r)$ to the row clusters. For any given column clusters, the probabilities in (4.13) are the optimal solutions $p_1(\mathbf{x}_i^r)$, $p_2(\mathbf{x}_i^r)$ of the following optimization problem

$$\begin{aligned} \underset{p_1(\mathbf{x}_i^r),\, p_2(\mathbf{x}_i^r)}{\text{minimize}} \quad & \sum_{j=1}^{M} d_1(\mathbf{x}_i^r) p_1(\mathbf{x}_i^r)^2 (\bar{d}_1(\mathbf{x}_j^c) q_1(\mathbf{x}_j^c)^2 + \bar{d}_2(\mathbf{x}_j^c) q_2(\mathbf{x}_j^c)^2) \\ & + d_2(\mathbf{x}_i^r) p_2(\mathbf{x}_i^r)^2 (\bar{d}_1(\mathbf{x}_j^c) q_1(\mathbf{x}_j^c)^2 + \bar{d}_2(\mathbf{x}_j^c) q_2(\mathbf{x}_j^c)^2) \\ \text{subject to} \quad & p_1(\mathbf{x}_i^r) + p_2(\mathbf{x}_i^r) = 1, \\ & p_1(\mathbf{x}_i^r), p_2(\mathbf{x}_i^r)) \geq 0, \end{aligned}$$

which can also be written as

$$\begin{aligned} \underset{p_1(\mathbf{x}_i^r),\, p_2(\mathbf{x}_i^r)}{\text{minimize}} \quad & [d_1(\mathbf{x}_i^r) p_1(\mathbf{x}_i^r)^2 + d_2(\mathbf{x}_i^r) p_2(\mathbf{x}_i^r)^2] \left(\sum_{j=1}^{M} \sum_{l=1}^{2} \bar{d}_l(\mathbf{x}_j^c) q_l(\mathbf{x}_j^c)^2 \right) \\ \text{subject to} \quad & p_1(\mathbf{x}_i^r) + p_2(\mathbf{x}_i^r) = 1, \\ & p_1(\mathbf{x}_i^r), p_2(\mathbf{x}_i^r) \geq 0. \end{aligned} \tag{4.17}$$

Here the term under summation is a constant term and can be ignored because probabilities and distances are all known.

The Lagrangian of the problem 4.17 is

$$\mathcal{L}(p_1(\mathbf{x}_i^r), p_2(\mathbf{x}_i^r), \lambda) = d_1(\mathbf{x}_i^r) p_1(\mathbf{x}_i^r)^2 + d_2(\mathbf{x}_i^r) p_2(\mathbf{x}_i^r)^2 - \lambda(p_1(\mathbf{x}_i^r) + p_2(\mathbf{x}_i^r) - 1),$$

and setting the partial derivatives with respect to $p_1(\mathbf{x}_i^r)$, $p_2(\mathbf{x}_i^r)$ equal to zero gives the Principle 1,

$$p_1(\mathbf{x}_i^r) d_1(\mathbf{x}_i^r) = p_2(\mathbf{x}_i^r) d_2(\mathbf{x}_i^r).$$

Similarly, now consider the same case $K = L = 2$ for a column entity $\mathbf{x}_j^c$ with any given row clusters, where probabilities and distances are known.

If we let $\mathbf{x}_j^c$ be a column entity with distances $\bar{d}_1(\mathbf{x}_j^c)$, $\bar{d}_2(\mathbf{x}_j^c)$ to the column cluster centers, the probabilities in (4.15) are the optimal solutions $q_1(\mathbf{x}_j^c)$, $q_2(\mathbf{x}_j^c)$ of the fol-

lowing optimization problem

$$\begin{aligned} \underset{q_1(\mathbf{x}_j^c),\, q_2(\mathbf{x}_j^c)}{\text{minimize}} \quad & [\bar{d}_1(\mathbf{x}_j^c) q_1(\mathbf{x}_j^c)^2 + \bar{d}_2(\mathbf{x}_j^c) q_2(\mathbf{x}_j^c)^2] \left(\sum_{i=1}^{N} \sum_{k=1}^{2} d_k(\mathbf{x}_i^r) p_k(\mathbf{x}_i^r)^2 \right) \\ \text{subject to} \quad & q_1(\mathbf{x}_j^c) + q_2(\mathbf{x}_j^c) = 1, \\ & q_1(\mathbf{x}_j^c), q_2(\mathbf{x}_j^c) \geq 0. \end{aligned} \tag{4.18}$$

Again, the term under summation is constant as row membership probabilities and distances are known.

Thus, the Lagrangian of the problem 4.18 is

$$\mathcal{L}(q_1(\mathbf{x}_j^c), q_2(\mathbf{x}_j^c), \lambda) = \bar{d}_1(\mathbf{x}_j^c) q_1(\mathbf{x}_j^c)^2 + \bar{d}_2(\mathbf{x}_j^c) q_2(\mathbf{x}_j^c)^2 - \lambda(q_1(\mathbf{x}_j^c) + q_2(\mathbf{x}_j^c) - 1),$$

and setting the partial derivatives with respect to $q_1(\mathbf{x}_j^c)$, $q_2(\mathbf{x}_j^c)$ equal to zero gives the Principle 2,

$$q_1(\mathbf{x}_j^c) \bar{d}_1(\mathbf{x}_j^c) = q_2(\mathbf{x}_j^c) \bar{d}_2(\mathbf{x}_j^c).$$

Then the **optimization problem of row entities** of whole data set $\mathbf{X} = \{\mathbf{x}_1^r, \mathbf{x}_2^r, \ldots, \mathbf{x}_N^r\}$ for any given column clusters is

$$\begin{aligned} \underset{p_1(\mathbf{x}_i^r),\, p_2(\mathbf{x}_i^r)}{\text{minimize}} \quad & \sum_{i=1}^{N} [d_1(\mathbf{x}_i^r) p_1(\mathbf{x}_i^r)^2 + d_2(\mathbf{x}_i^r) p_2(\mathbf{x}_i^r)^2] \left(\sum_{j=1}^{M} \sum_{l=1}^{2} \bar{d}_l(\mathbf{x}_j^c) q_l(\mathbf{x}_j^c)^2 \right) \\ \text{subject to} \quad & p_1(\mathbf{x}_i^r) + p_2(\mathbf{x}_i^r) = 1, \quad i = 1, \ldots, N, \\ & p_1(\mathbf{x}_i^r), p_2(\mathbf{x}_i^r) \geq 0, \quad i = 1, \ldots, N. \end{aligned} \tag{4.19}$$

Following 4.18, the **optimization problem of column entities** of whole data set $\mathbf{X} = \{\mathbf{x}_1^c, \mathbf{x}_2^c, \ldots, \mathbf{x}_M^c\}$ for any given row clusters is

$$\begin{aligned} \underset{q_1(\mathbf{x}_j^c),\, q_2(\mathbf{x}_j^c)}{\text{minimize}} \quad & \sum_{j=1}^{M} [\bar{d}_1(\mathbf{x}_j^c) q_1(\mathbf{x}_j^c)^2 + \bar{d}_2(\mathbf{x}_j^c) q_2(\mathbf{x}_j^c)^2] \left(\sum_{i=1}^{N} \sum_{k=1}^{2} d_k(\mathbf{x}_i^r) p_k(\mathbf{x}_i^r)^2 \right) \\ \text{subject to} \quad & q_1(\mathbf{x}_j^c) + q_2(\mathbf{x}_j^c) = 1 \quad j = 1, \ldots, M, \\ & q_1(\mathbf{x}_j^c), q_2(\mathbf{x}_j^c) \geq 0, \quad j = 1, \ldots, M. \end{aligned} \tag{4.20}$$

4.2.4 Optimization Problem

Following the problems (4.19) and (4.20), the optimization model of two-mode clustering problem for K row and L column clusters is

$$\begin{aligned} \underset{p_k(\mathbf{x}_i^r),\, q_l(\mathbf{x}_j^c),\, \mathbf{v}_k^r,\, \mathbf{v}_l^c}{\text{minimize}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2 d_k(\mathbf{x}_i^r)\bar{d}_l(\mathbf{x}_j^c) \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i^r) = 1, \quad i = 1,\ldots,N, \\ & \sum_{l=1}^{L} q_l(\mathbf{x}_j^c) = 1, \quad j = 1,\ldots,M, \\ & p_k(\mathbf{x}_i^r), q_l(\mathbf{x}_j^c) \geq 0, \quad \forall i,j,k,l. \end{aligned} \tag{4.21}$$

4.2.5 Membership Problem

In the case the centers are given, the objective in (4.21) becomes quadratic function of $p_k(\mathbf{x}_i^r)$ and $q_l(\mathbf{x}_j^c)$. This model can be solved for $p_k(\mathbf{x}_i^r)$ by fixing $q_l(\mathbf{x}_j^c)$, which will be named as **row membership problem**, and for $q_l(\mathbf{x}_j^c)$ by fixing $p_k(\mathbf{x}_i^r)$, which is named as **column membership problem**.

Row Membership Problem

Let centers $\mathbf{v}_k^r$, $k = 1,\ldots,K$, and $\mathbf{v}_l^c$, $l = 1,\ldots,L$ are known and column probabilities $q_l(\mathbf{x}_j^c)$ are fixed, then the optimization problem in (4.21) is

$$\begin{aligned} \underset{p_k(\mathbf{x}_i^r)}{\text{minimize}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2 d_k(\mathbf{x}_i^r)\bar{d}_l(\mathbf{x}_j^c) \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i^r) = 1, \quad i = 1,...,N, \\ & p_k(\mathbf{x}_i^r) \geq 0, \quad i = 1,...,N, \quad k = 1,...,K, \end{aligned} \tag{4.22}$$

which is a convex optimization problem of $p_k(\mathbf{x}_i^r)$.

For simplicity we set $K = L = 2$. The problem in (4.22) can be decomposed for

each $\mathbf{x}_i^r$, and becomes

$$\begin{aligned}
\underset{p_1(\mathbf{x}_i^r),\, p_2(\mathbf{x}_i^r)}{\text{minimize}} \quad & \sum_{j=1}^{M} p_1(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 d_1(\mathbf{x}_i^r)\bar{d}_1(\mathbf{x}_j^c) + p_1(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 d_1(\mathbf{x}_i^r)\bar{d}_2(\mathbf{x}_j^c) \\
& + p_2(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 d_2(\mathbf{x}_i^r)\bar{d}_1(\mathbf{x}_j^c) + p_2(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 d_2(\mathbf{x}_i^r)\bar{d}_2(\mathbf{x}_j^c) \\
\text{subject to} \quad & p_1(\mathbf{x}_i^r) + p_2(\mathbf{x}_i^r) = 1, \quad i = 1, ..., N, \\
& p_1(\mathbf{x}_i^r), p_2(\mathbf{x}_i^r) \geq 0, \quad i = 1, ..., N.
\end{aligned} \tag{4.23}$$

Lagrangian of the problem in (4.23) is

$$\mathcal{L}(p_1(\mathbf{x}_i^r), p_2(\mathbf{x}_i^r), \lambda_i) = \Bigg(\sum_{j=1}^{M} p_1(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 d_1(\mathbf{x}_i^r)\bar{d}_1(\mathbf{x}_j^c) + p_1(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 d_1(\mathbf{x}_i^r)\bar{d}_2(\mathbf{x}_j^c) +$$

$$p_2(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 d_2(\mathbf{x}_i^r)\bar{d}_1(\mathbf{x}_j^c) + p_2(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 d_2(\mathbf{x}_i^r)\bar{d}_2(\mathbf{x}_j^c) \Bigg) - \lambda_i (p_1(\mathbf{x}_i^r) + p_2(\mathbf{x}_i^r) - 1)$$

$$\frac{\partial \mathcal{L}}{\partial p_1(\mathbf{x}_i^r)} = 2p_1(\mathbf{x}_i^r) d_1(\mathbf{x}_i^r)[q_1(\mathbf{x}_j^c)^2 \bar{d}_1(\mathbf{x}_j^c) + q_2(\mathbf{x}_j^c)^2 \bar{d}_2(\mathbf{x}_j^c)] - \lambda_i = 0 \tag{4.24}$$

$$\frac{\partial \mathcal{L}}{\partial p_2(\mathbf{x}_i^r)} = 2p_2(\mathbf{x}_i^r) d_2(\mathbf{x}_i^r)[q_1(\mathbf{x}_j^c)^2 \bar{d}_1(\mathbf{x}_j^c) + q_2(\mathbf{x}_j^c)^2 \bar{d}_2(\mathbf{x}_j^c)] - \lambda_i = 0 \tag{4.25}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = p_1(\mathbf{x}_i^r) + p_2(\mathbf{x}_i^r) - 1 = 0$$

From (4.24) and (4.25), we get

$$p_1(\mathbf{x}_i^r) d_1(\mathbf{x}_i^r) = p_2(\mathbf{x}_i^r) d_2(\mathbf{x}_i^r), \quad i = 1, ..., N,$$

which is Principle 1 in (4.3).

Column Membership Problem

Let centers $\mathbf{v}_k^r$, $k = 1, \ldots, K$, and $\mathbf{v}_l^c$, $l = 1, \ldots, L$ are known and row probabilities $p_k(\mathbf{x}_i^r)$ are fixed, then the optimization problem in (4.21) is

$$\begin{aligned}
\underset{q_l(\mathbf{x}_j^c)}{\text{minimize}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2 d_k(\mathbf{x}_i^r)\bar{d}_l(\mathbf{x}_j^c) \\
\text{subject to} \quad & \sum_{l=1}^{L} q_l(\mathbf{x}_j^c) = 1, \quad j = 1, ..., M, \\
& q_l(\mathbf{x}_j^c) \geq 0, \quad j = 1, ..., M, \quad l = 1, ..., L,
\end{aligned} \tag{4.26}$$

which is a convex optimization problem of $q_l(\mathbf{x}_j^c)$.

For simplicity we set $K = L = 2$. The problem in (4.26) can be decomposed for each $\mathbf{x}_j^c$, and becomes

$$\begin{aligned}
\underset{q_1(\mathbf{x}_j^c),\, q_2(\mathbf{x}_j^c)}{\text{minimize}} \quad & \sum_{i=1}^{N} p_1(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 d_1(\mathbf{x}_i^r)\bar{d}_1(\mathbf{x}_j^c) + p_1(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 d_1(\mathbf{x}_i^r)\bar{d}_2(\mathbf{x}_j^c) \\
& + p_2(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 d_2(\mathbf{x}_i^r)\bar{d}_1(\mathbf{x}_j^c) + p_2(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 d_2(\mathbf{x}_i^r)\bar{d}_2(\mathbf{x}_j^c) \\
\text{subject to} \quad & q_1(\mathbf{x}_j^c) + q_2(\mathbf{x}_j^c) = 1, \quad j = 1, ..., M, \\
& q_1(\mathbf{x}_j^c), q_2(\mathbf{x}_j^c) \geq 0, \quad j = 1, ..., M.
\end{aligned} \tag{4.27}$$

Lagrangian of the problem in (4.27) is

$$\mathcal{L}(q_1(\mathbf{x}_j^c), q_2(\mathbf{x}_j^c), \mu_i) = \Bigg(\sum_{i=1}^{N} p_1(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 d_1(\mathbf{x}_i^r)\bar{d}_1(\mathbf{x}_j^c) + p_1(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 d_1(\mathbf{x}_i^r)\bar{d}_2(\mathbf{x}_j^c) +$$

$$p_2(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 d_2(\mathbf{x}_i^r)\bar{d}_1(\mathbf{x}_j^c) + p_2(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 d_2(\mathbf{x}_i^r)\bar{d}_2(\mathbf{x}_j^c) \Bigg) - \mu_i (q_1(\mathbf{x}_j^c) + q_2(\mathbf{x}_j^c) - 1)$$

$$\frac{\partial \mathcal{L}}{\partial q_1(\mathbf{x}_j^c)} = 2q_1(\mathbf{x}_j^c)\bar{d}_1(\mathbf{x}_j^c)[p_1(\mathbf{x}_i^r)^2 d_1(\mathbf{x}_i^r) + p_2(\mathbf{x}_i^r)^2 d_2(\mathbf{x}_i^r)] - \mu_i = 0 \tag{4.28}$$

$$\frac{\partial \mathcal{L}}{\partial q_2(\mathbf{x}_j^c)} = 2q_2(\mathbf{x}_j^c)\bar{d}_2(\mathbf{x}_j^c)[p_1(\mathbf{x}_i^r)^2 d_1(\mathbf{x}_i^r) + p_2(\mathbf{x}_i^r)^2 d_2(\mathbf{x}_i^r)] - \mu_i = 0 \tag{4.29}$$

$$\frac{\partial \mathcal{L}}{\partial \mu_i} = q_1(\mathbf{x}_j^c) + q_2(\mathbf{x}_j^c) - 1 = 0$$

From (4.28) and (4.29), we get

$$q_1(\mathbf{x}_j^c)\bar{d}_1(\mathbf{x}_j^c) = q_2(\mathbf{x}_j^c)\bar{d}_2(\mathbf{x}_j^c), \quad j = 1, ..., M,$$

which is Principle 2 in (4.5).

4.2.6 Center Problem

Given the row and column membership probabilities ($p_k(\mathbf{x}_i^r)$ and $q_l(\mathbf{x}_j^c)$), row and column cluster centers ($\mathbf{v}_k^r$ and $\mathbf{v}_l^c$) become decision variables of the optimization problem in (4.21) as follows

$$\underset{\mathbf{v}_k^r,\, \mathbf{v}_l^c}{\text{minimize}} \quad \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2 d(\mathbf{x}_i^r, \mathbf{v}_k^r)\bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c). \tag{4.30}$$

When we fix the column centers $\mathbf{v}_l^c$, $l = 1, \ldots, L$ the problem in (4.30) can be solved for row centers $\mathbf{v}_k^r$, $k = 1, \ldots, K$. Similarly, column centers can be found by fixing the row centers as explained in the following parts.

Row Centers

Given the membership probabilities and fixed column centers ($\mathbf{v}_l^c$) we write the objective in (4.30) as a function of row centers $\mathbf{v}_k^r$ and get

$$f(\mathbf{v}_k^r) = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2 d(\mathbf{x}_i^r, \mathbf{v}_k^r)\bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c). \tag{4.31}$$

Theorem 7. *Let the distance function $d(\boldsymbol{x}_i^r, \boldsymbol{v}_k^r)$ be Euclidean,*

$$d(\boldsymbol{x}_i^r, \boldsymbol{v}_k^r) = \|\boldsymbol{x}_i^r - \boldsymbol{v}_k^r\|, \quad i = 1, \ldots, N, \quad k = 1, \ldots, K,$$

so that

$$f(\boldsymbol{v}_k^r) = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(\boldsymbol{x}_i^r)^2 q_l(\boldsymbol{x}_j^c)^2 \ \|\boldsymbol{x}_i^r - \boldsymbol{v}_k^r\| \ \bar{d}(\boldsymbol{x}_j^c, \boldsymbol{v}_l^c), \tag{4.32}$$

and let the probabilities be given for all $\boldsymbol{x}_i^r$ and $\boldsymbol{x}_j^c$. Also let column cluster centers $\boldsymbol{v}_l^c$ be known.

Then, the minimizers of $\boldsymbol{v}_k^r$ are given by

$$\boldsymbol{v}_k^r = \frac{\sum_{i=1}^{N} \frac{p_k(\boldsymbol{x}_i^r)^2}{d(\boldsymbol{x}_i^r, \boldsymbol{v}_k^r)} \boldsymbol{x}_i^r}{\sum_{i=1}^{N} \frac{p_k(\boldsymbol{x}_i^r)^2}{d(\boldsymbol{x}_i^r, \boldsymbol{v}_k^r)}}, \quad k = 1, \ldots, K. \tag{4.33}$$

Proof. The gradient of $d(\mathbf{x}_i^r, \mathbf{v}_k^r) = \|\mathbf{x}_i^r - \mathbf{v}_k^r\|$ with respect to $\mathbf{v}_k^r$ is

$$\nabla_{\mathbf{v}_k^r} \|\mathbf{x}_i^r - \mathbf{v}_k^r\| = -\frac{\mathbf{x}_i^r - \mathbf{v}_k^r}{\|\mathbf{x}_i^r - \mathbf{v}_k^r\|} = -\frac{\mathbf{x}_i^r - \mathbf{v}_k^r}{d(\mathbf{x}_i^r, \mathbf{v}_k^r)}.$$

Therefore, the gradient of (4.32) with respect to $\mathbf{v}_k^r$ is

$$\begin{aligned}\nabla_{\mathbf{v}_k^r} f(\mathbf{v}_k^r) &= -\left[\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{l=1}^{L} p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2 \frac{\mathbf{x}_i^r - \mathbf{v}_k^r}{d(\mathbf{x}_i^r, \mathbf{v}_k^r)} \bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c)\right] \\ &= -\sum_{i=1}^{N} \frac{\mathbf{x}_i^r - \mathbf{v}_k^r}{d(\mathbf{x}_i^r, \mathbf{v}_k^r)} p_k(\mathbf{x}_i^r)^2 \sum_{j=1}^{M}\sum_{l=1}^{L} q_l(\mathbf{x}_j^c)^2 \bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c), \quad k = 1, \ldots, K. \end{aligned} \tag{4.34}$$

Let us call the term $\sum_{j=1}^{M}\sum_{l=1}^{L} q_l(\mathbf{x}_j^c)^2\bar{d}(\mathbf{x}_j^c,\mathbf{v}_l^c)$ as w, then the equation (4.34) becomes

$$= -w\sum_{i=1}^{N}\frac{\mathbf{x}_i^r-\mathbf{v}_k^r}{d(\mathbf{x}_i^r,\mathbf{v}_k^r)}p_k(\mathbf{x}_i^r)^2.$$

By setting the gradient equal to zero, we get

$$\sum_{i=1}^{N}\frac{p_k(\mathbf{x}_i^r)^2\mathbf{x}_i^r}{d(\mathbf{x}_i^r,\mathbf{v}_k^r)} = \left[\sum_{i=1}^{N}\frac{p_k(\mathbf{x}_i^r)^2}{d(\mathbf{x}_i^r,\mathbf{v}_k^r)}\right]\mathbf{v}_k^r,$$

then

$$\mathbf{v}_k^r = \frac{\sum_{i=1}^{N}\frac{p_k(\mathbf{x}_i^r)^2}{d(\mathbf{x}_i^r,\mathbf{v}_k^r)}\mathbf{x}_i^r}{\sum_{i=1}^{N}\frac{p_k(\mathbf{x}_i^r)^2}{d(\mathbf{x}_i^r,\mathbf{v}_k^r)}}, \quad k=1,\ldots,K,$$

proving (4.33). □

Column Centers

Given the membership probabilities and fixed row centers ($\mathbf{v}_k^r$) we write the objective in (4.30) as a function of column centers $\mathbf{v}_l^c$ and get

$$f(\mathbf{v}_l^c) = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2 d(\mathbf{x}_i^r,\mathbf{v}_k^r)\bar{d}(\mathbf{x}_j^c,\mathbf{v}_l^c). \tag{4.35}$$

Theorem 8. *Let the distance function $\bar{d}(\boldsymbol{x}_j^c,\boldsymbol{v}_l^c)$ be Euclidean,*

$$\bar{d}(\boldsymbol{x}_j^c,\boldsymbol{v}_l^c) = \|\boldsymbol{x}_j^c-\boldsymbol{v}_l^c\|, \quad l=1,\ldots,L, \quad j=1,\ldots,M,$$

so that

$$f(\boldsymbol{v}_l^c) = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(\boldsymbol{x}_i^r)^2 q_l(\boldsymbol{x}_j^c)^2 d(\boldsymbol{x}_i^r,\boldsymbol{v}_k^r)\ \|\boldsymbol{x}_j^c-\boldsymbol{v}_l^c\|, \tag{4.36}$$

and let the probabilities be given for all $\boldsymbol{x}_i^r$ and $\boldsymbol{x}_j^c$. Also let row cluster centers $\boldsymbol{v}_k^r$ be known.

Then, the minimizers of $\boldsymbol{v}_l^c$ are given by

$$\boldsymbol{v}_l^c = \frac{\sum_{j=1}^{M}\frac{q_l(\boldsymbol{x}_j^c)^2}{\bar{d}(\boldsymbol{x}_j^c,\boldsymbol{v}_l^c)}\boldsymbol{x}_j^c}{\sum_{j=1}^{M}\frac{q_l(\boldsymbol{x}_j^c)^2}{\bar{d}(\boldsymbol{x}_j^c,\boldsymbol{v}_l^c)}}, \quad l=1,\ldots,L. \tag{4.37}$$

Proof. The gradient of $\bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c) = \| \mathbf{x}_j^c - \mathbf{v}_l^c \|$ with respect to $\mathbf{v}_l^c$ is

$$\nabla_{\mathbf{v}_l^c} \| \mathbf{x}_j^c - \mathbf{v}_l^c \| = -\frac{\mathbf{x}_j^c - \mathbf{v}_l^c}{\| \mathbf{x}_j^c - \mathbf{v}_l^c \|} = -\frac{\mathbf{x}_j^c - \mathbf{v}_l^c}{\bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c)}.$$

Therefore, the gradient of (4.36) with respect to $\mathbf{v}_l^c$ is

$$\begin{aligned} \nabla_{\mathbf{v}_l^c} f(\mathbf{v}_l^c) &= -\left[\sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2 d(\mathbf{x}_i^r, \mathbf{v}_k^r) \frac{\mathbf{x}_j^c - \mathbf{v}_l^c}{\bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c)} \right] \\ &= -\sum_{j=1}^{M} \frac{\mathbf{x}_j^c - \mathbf{v}_l^c}{\bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c)} q_l(\mathbf{x}_j^c)^2 \sum_{i=1}^{N} \sum_{k=1}^{K} p_k(\mathbf{x}_i^r)^2 d(\mathbf{x}_i^r, \mathbf{v}_k^r), \quad l = 1, \ldots, L \end{aligned} \tag{4.38}$$

Let us call the term $\sum_{i=1}^{N} \sum_{k=1}^{K} p_k(\mathbf{x}_i^r)^2 d(\mathbf{x}_i^r, \mathbf{v}_k^r)$ as w, then the equation (4.38) becomes

$$= -w \sum_{j=1}^{M} \frac{\mathbf{x}_j^c - \mathbf{v}_l^c}{\bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c)} q_l(\mathbf{x}_j^c)^2.$$

By setting the gradient equal to zero, we get

$$\sum_{j=1}^{M} \frac{q_l(\mathbf{x}_j^c)^2 \mathbf{x}_j^c}{\bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c)} = \left[\sum_{j=1}^{M} \frac{q_l(\mathbf{x}_j^c)^2}{\bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c)} \right] \mathbf{v}_l^c,$$

then

$$\mathbf{v}_l^c = \frac{\sum_{j=1}^{M} \frac{q_l(\mathbf{x}_j^c)^2}{\bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c)} \mathbf{x}_j^c}{\sum_{j=1}^{M} \frac{q_l(\mathbf{x}_j^c)^2}{\bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c)}}, \quad l = 1, \ldots, L,$$

proving (4.37). □

4.2.7 Discussion on the Optimality Conditions

Figure 4.5 illustrates position of an element x_{ij}, row and column cluster centers $\mathbf{v}_k^r$, $\mathbf{v}_l^c$ under optimal solution for $K = L = 2$. Distances of the x_{ij} from row and column cluster centers are denoted by $d_k(\mathbf{x}_i^r)$ and $\bar{d}_l(\mathbf{x}_j^c)$, respectively. Observe that distance of an element x_{ij} to a row cluster center $\mathbf{v}_k^r$ is equal to distance of row instance $\mathbf{x}_i^r$ that contains the element x_{ij} from row cluster center $\mathbf{v}_k^r$. Similarly, distance of an element x_{ij} to a column cluster center $\mathbf{v}_l^c$ is equal to distance of column instance $\mathbf{x}_j^c$ that contains the element x_{ij} from column cluster center $\mathbf{v}_l^c$.

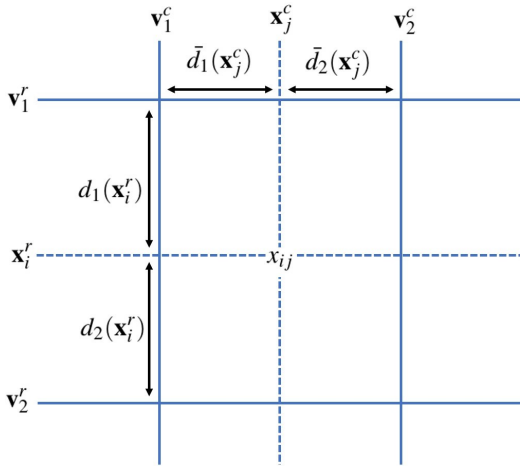


Figure 4.5: Positions of an element x_{ij}, and row and column cluster centers $\mathbf{v}_k^r$, $\mathbf{v}_l^c$ in the optimal solution, where $K = L = 2$.

Following the Principle 3 in (4.9), in the optimal solution, the membership probability of x_{ij} to partition $\mathbf{V}_{kl}$ is higher when the corresponding $D_{kl}(x_{ij})$ is smaller. For $K = L = 2$, the following equalities should hold for the optimality,

$$\begin{aligned} R_{11}(x_{ij})D_{11}(\mathbf{x}_i^r, \mathbf{x}_j^c) &= R_{12}(x_{ij})D_{12}(\mathbf{x}_i^r, \mathbf{x}_j^c) \\ &= R_{21}(x_{ij})D_{21}(\mathbf{x}_i^r, \mathbf{x}_j^c) = R_{22}(x_{ij})D_{22}(\mathbf{x}_i^r, \mathbf{x}_j^c). \end{aligned} \tag{4.39}$$

We may rewrite the equation (4.39) in terms of $d_k(\mathbf{x}_i^r)$ and $\bar{d}_l(\mathbf{x}_j^c)$ as

$$\begin{aligned} R_{11}(x_{ij})d_1(\mathbf{x}_i^r)\bar{d}_1(\mathbf{x}_j^c) &= R_{12}(x_{ij})d_1(\mathbf{x}_i^r)\bar{d}_2(\mathbf{x}_j^c) \\ &= R_{21}(x_{ij})d_2(\mathbf{x}_i^r)\bar{d}_1(\mathbf{x}_j^c) = R_{22}(x_{ij})d_2(\mathbf{x}_i^r)\bar{d}_2(\mathbf{x}_j^c). \end{aligned} \tag{4.40}$$

(4.40) implies the following

$$R_{11}(x_{ij})d_1(\mathbf{x}_i^r) = R_{21}(x_{ij})d_2(\mathbf{x}_i^r), \tag{4.41a}$$

$$R_{11}(x_{ij})\bar{d}_1(\mathbf{x}_j^c) = R_{12}(x_{ij})\bar{d}_2(\mathbf{x}_j^c), \tag{4.41b}$$

$$R_{12}(x_{ij})d_1(\mathbf{x}_i^r) = R_{22}(x_{ij})d_2(\mathbf{x}_i^r), \tag{4.41c}$$

$$R_{21}(x_{ij})\bar{d}_1(\mathbf{x}_j^c) = R_{22}(x_{ij})\bar{d}_2(\mathbf{x}_j^c). \tag{4.41d}$$

The equalities in (4.41a)-(4.41d) have an analogy with a fundamental concept of physics. The *torque* or *moment* of a force is the propensity of the force to rotate the body that is applied to. It is calculated by multiplying the force and its perpendicular distance to a turning point. To illustrate, examine the Figure 4.6a, if we consider $R_{11}(x_{ij})$ and $R_{21}(x_{ij})$ as forces applied to blue branches with lengths of $d_1(\mathbf{x}_i^r)$ and $d_2(\mathbf{x}_i^r)$, then the torque applied by force $R_{11}(x_{ij})$ is equal to $R_{11}(x_{ij})d_1(\mathbf{x}_i^r)$ and its sign is positive due to right hand side rule, and the torque applied by force $R_{21}(x_{ij})$ is equal to $R_{21}(x_{ij})d_2(\mathbf{x}_i^r)$, which has negative sign. Thus, the net torque is equal to $R_{11}(x_{ij})d_1(\mathbf{x}_i^r) - R_{21}(x_{ij})d_2(\mathbf{x}_i^r)$.

In the case that the net torque is 0, the position of x_{ij} is stabilized, which means that there is no better place for x_{ij} to move. In terms of clustering, it means that we need to find forces ($R_{kl}(x_{ij})$s) that has net torque of 0 for all possible turning points for the optimality. When we consider all possible turning points for x_{ij} as in Figure 4.6a-4.6d, making the net torque 0 leads us to the equalities in (4.41a)-(4.41d).

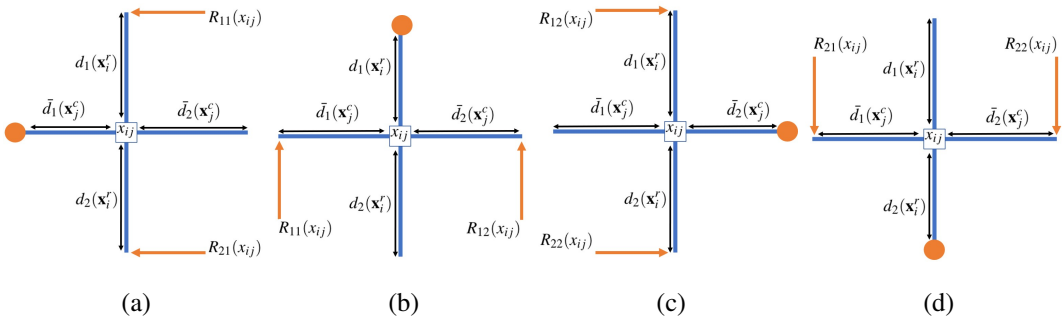


Figure 4.6: Physical representation of optimality conditions, given the turning points (orange dots) located on the (a) left, (b) top, (c) right, and (d) bottom.

4.2.8 Algorithm

We propose an iterative algorithm for soft two-mode clustering by implementing the ideas presented in the previous sections. The steps of the algorithm are as follows:

Algorithm 1: Proposed Algorithm 1

Step 1. Given the data set $\mathbf{X}$, and $\epsilon > 0$ initialize the cluster centers $\{\mathbf{v}_k^r, \mid k = 1, ..., K\}$ and $\{\mathbf{v}_l^c, \mid l = 1, ..., L\}$

Step 2. Compute the distance of $d_k(\mathbf{x}_i^r)$ and $\bar{d}_l(\mathbf{x}_j^c)$ for all $\mathbf{x}_i^r, \mathbf{x}_j^c \in \mathbf{X}$, respectively by

$$d_k(\mathbf{x}_i^r) = \| \mathbf{x}_i^r - \mathbf{v}_k^r \|, \quad \text{and} \quad \bar{d}_l(\mathbf{x}_j^c) = \| \mathbf{x}_j^c - \mathbf{v}_l^c \|$$

Step 3. Compute the membership probability of $p_k(\mathbf{x}_i^r)$ and $q_l(\mathbf{x}_j^c)$ for all $\mathbf{x}_i^r, \mathbf{x}_j^c \in \mathbf{X}$ respectively by

$$p_k(\mathbf{x}_i^r) = \frac{\dfrac{1}{d_k(\mathbf{x}_i^r)}}{\sum_{s=1}^{K} \dfrac{1}{d_s(\mathbf{x}_i^r)}}, \quad \text{and} \quad q_l(\mathbf{x}_j^c) = \frac{\dfrac{1}{\bar{d}_l(\mathbf{x}_j^c)}}{\sum_{t=1}^{L} \dfrac{1}{\bar{d}_t(\mathbf{x}_j^c)}} \quad \text{as in (4.12) and (4.14)}$$

Step 4. Update the row cluster centers $\mathbf{v}_k^r$, $k = 1, \ldots, K$ and column cluster centers $\mathbf{v}_l^c$, $l = 1, \ldots, L$ simultaneously by

$$\mathbf{v}_k^{r+} = \frac{\sum_{i=1}^{N} \dfrac{p_k(\mathbf{x}_i^r)^2}{d(\mathbf{x}_i^r, \mathbf{v}_k^r)} \mathbf{x}_i^r}{\sum_{i=1}^{N} \dfrac{p_k(\mathbf{x}_i^r)^2}{d(\mathbf{x}_i^r, \mathbf{v}_k^r)}}, \quad \text{and} \quad \mathbf{v}_l^{c+} = \frac{\sum_{j=1}^{M} \dfrac{q_l(\mathbf{x}_j^c)^2}{\bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c)} \mathbf{x}_j^c}{\sum_{j=1}^{M} \dfrac{q_l(\mathbf{x}_j^c)^2}{\bar{d}(\mathbf{x}_j^c, \mathbf{v}_l^c)}} \quad \text{as in (4.33) and (4.37)}$$

Step 5. If $\sum_{k=1}^{K} \| \mathbf{v}_k^{r+} - \mathbf{v}_k^r \| + \sum_{l=1}^{L} \| \mathbf{v}_l^{c+} - \mathbf{v}_l^c \| < \epsilon$, stop. Else, continue with Step 2.

Note that in **Step 1**, row and column entities are randomly assigned into K and L clusters, respectively as an initilization. A row cluster center $\mathbf{v}_k^r$ is calculated by row average of its assigned row entities. Similarly, a column cluster center $\mathbf{v}_l^c$ is computed as the column mean of its column entities.

4.3 Alternative Approach for Two-Mode Probabilistic Distance Clustering

In Algorithm 1 (See Section 4.2), the distance between a row entity $\mathbf{x}_i^r$ and a row cluster center $\mathbf{v}_k^r$, and the distance between a column entity $\mathbf{x}_j^c$ and a column cluster center $\mathbf{v}_l^c$ are considered in the objective of (4.21) as $d_k(\mathbf{x}_i^r)\bar{d}_l(\mathbf{x}_j^c)$.

In this section, we change the representation of cluster centers and consider each cluster center as a function of the submatrix elements, i.e., mean of the values of elements in the partition. We define submatrix centers v_{kl} for each partition $\mathbf{V}_{kl}$ and consider minimizing total distance of the elements x_{ij} to their probabilistically assigned partition centers v_{kl}, (see Section 4.3.1). Since the distance function measures the difference between values of v_{kl} and x_{ij}, the Euclidean distance function reduces to the rectilinear norm. Note that we consider a soft assignment scheme such that each element x_{ij} contributes to the center update procedure with a weight (membership). Underlying principles are discussed in Section 4.3.1-4.3.5. Later, a gradient descent algorithm is proposed in Section 4.3.6.

4.3.1 Optimization Problem

Considering the problem in (4.21), instead of $d_k(\centerdot)\bar{d}_l(\centerdot)$, now we use the distance function as a function of the center v_{kl} and the element x_{ij} as $|x_{ij} - v_{kl}|$, satisfying that $x_{ij} \in \mathbf{x}_i^r$ and $\mathbf{x}_j^c$. Thus, the problem in (4.21) becomes

$$\begin{aligned}
\underset{p_k(\mathbf{x}_i^r),\, q_l(\mathbf{x}_j^c),\, v_{kl}}{\text{minimize}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2 |x_{ij} - v_{kl}| \\
\text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i^r) = 1, \quad i = 1, \ldots, N, \\
& \sum_{l=1}^{L} q_l(\mathbf{x}_j^c) = 1, \quad j = 1, \ldots, M, \\
& p_k(\mathbf{x}_i^r), q_l(\mathbf{x}_j^c) \geq 0, \quad \forall i, j, k, l.
\end{aligned} \tag{4.42}$$

where $p_k(\mathbf{x}_i^r)$ are row cluster membership probabilities, and $q_l(\mathbf{x}_j^c)$ are column cluster membership probabilities. Moreover, v_{kl} are submatrix centers.

One can observe that the problem in (4.42) can be expressed in terms of x_{ij} by fol-

lowing the problem properties (4.4) and (4.6) as follows

$$\begin{aligned} \underset{p_k(x_{ij}),\, q_l(x_{ij}),\, v_{kl}}{\text{minimize}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(x_{ij})^2 q_l(x_{ij})^2 |x_{ij} - v_{kl}| \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(x_{ij}) = 1, \quad i = 1, \ldots, N, \\ & \sum_{l=1}^{L} q_l(x_{ij}) = 1, \quad j = 1, \ldots, M, \\ & p_k(x_{ij}), q_l(x_{ij}) \geq 0, \quad \forall i, j, k, l. \end{aligned} \tag{4.43}$$

In (4.43), if the membership probabilities are given, the only decision variables will be cluster centers v_{kl}, $k = 1, \ldots, K$ and $l = 1, \ldots, L$. Thus, the objective function of the problem becomes the function of submatrix centers, and this problem is called **center problem**. Similarly, if the centers are given, then the objective becomes a function of probabilities, and the problem is called **membership problem** in the subsequent sections.

4.3.2 Membership Problem

As mentioned above, in the case the centers are given, the objective of (4.42) becomes a quadratic function of $p_k(\mathbf{x}_i^r)$ and $q_l(\mathbf{x}_j^c)$. This model can be solved for $p_k(\mathbf{x}_i^r)$ by fixing $q_l(\mathbf{x}_j^c)$, which will be named as **row membership problem**, and for $q_l(\mathbf{x}_j^c)$ by fixing $p_k(\mathbf{x}_i^r)$, which is named as **column membership problem** in the following sections.

Row Membership Problem

Let centers v_{kl}, $k = 1, \ldots, K$ and, $l = 1, \ldots, L$ are known and column probabilities $q_l(\mathbf{x}_j^c)$ are fixed, then the optimization problem in (4.42) is

$$\begin{aligned} \underset{p_k(\mathbf{x}_i^r)}{\text{minimize}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2 |x_{ij} - v_{kl}| \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i^r) = 1, \quad i = 1, ..., N, \\ & p_k(\mathbf{x}_i^r) \geq 0, \quad i = 1, ..., N, \quad k = 1, ..., K, \end{aligned} \tag{4.44}$$

which is a convex optimization problem of $p_k(\mathbf{x}_i^r)$.

Again, for simplicity we set $K = L = 2$. Then Problem in (4.44) can be decomposed for each $\mathbf{x}_i^r$, and it becomes

$$\begin{aligned} \underset{p_1(\mathbf{x}_i^r), p_2(\mathbf{x}_i^r)}{\text{minimize}} \quad & \sum_{j=1}^{M} p_1(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 |x_{ij} - v_{11}| + p_1(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 |x_{ij} - v_{12}| \\ & + p_2(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 |x_{ij} - v_{21}| + p_2(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 |x_{ij} - v_{22}| \\ \text{subject to} \quad & p_1(\mathbf{x}_i^r) + p_2(\mathbf{x}_i^r) = 1, \quad i = 1, \ldots, N, \\ & p_1(\mathbf{x}_i^r), p_2(\mathbf{x}_i^r) \geq 0, \quad i = 1, \ldots, N. \end{aligned} \tag{4.45}$$

Lagrangian of the problem (4.45) is

$$\begin{aligned} \mathcal{L}(p_1(\mathbf{x}_i^r), p_2(\mathbf{x}_i^r), \lambda_i) = \Bigg(& \sum_{j=1}^{M} p_1(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 |x_{ij} - v_{11}| + p_1(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 |x_{ij} - v_{12}| \\ & + p_2(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 |x_{ij} - v_{21}| + p_2(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 |x_{ij} - v_{22}| \Bigg) - \lambda_i (p_1(\mathbf{x}_i^r) + p_2(\mathbf{x}_i^r) - 1) \end{aligned}$$

$$\frac{\partial \mathcal{L}}{\partial p_1(\mathbf{x}_i^r)} = 2p_1(\mathbf{x}_i^r)\left[\sum_{j=1}^{M} q_1(\mathbf{x}_j^c)^2 |x_{ij} - v_{11}| + q_2(\mathbf{x}_j^c)^2 |x_{ij} - v_{12}|\right] - \lambda_i = 0 \tag{4.46}$$

$$\frac{\partial \mathcal{L}}{\partial p_2(\mathbf{x}_i^r)} = 2p_2(\mathbf{x}_i^r)\left[\sum_{j=1}^{M} q_1(\mathbf{x}_j^c)^2 |x_{ij} - v_{21}| + q_2(\mathbf{x}_j^c)^2 |x_{ij} - v_{22}|\right] - \lambda_i = 0 \tag{4.47}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = p_1(\mathbf{x}_i^r) + p_2(\mathbf{x}_i^r) - 1 = 0$$

(4.46) and (4.47) yield

$$p_1(\mathbf{x}_i^r)\left[\sum_{j=1}^{M} q_1(\mathbf{x}_j^c)^2|x_{ij}-v_{11}| + q_2(\mathbf{x}_j^c)^2|x_{ij}-v_{12}|\right]$$
$$= p_2(\mathbf{x}_i^r)\left[\sum_{j=1}^{M} q_1(\mathbf{x}_j^c)^2|x_{ij}-v_{21}| + q_2(\mathbf{x}_j^c)^2|x_{ij}-v_{22}|\right], \quad i=1,\ldots,N.$$

When we denote the term $\left[\sum_{j=1}^{M} q_1(\mathbf{x}_j^c)^2|x_{ij}-v_{11}| + q_2(\mathbf{x}_j^c)^2|x_{ij}-v_{12}|\right]$ as $\delta_1(\mathbf{x}_i^r)$, and $\left[\sum_{j=1}^{M} q_1(\mathbf{x}_j^c)^2|x_{ij}-v_{21}| + q_2(\mathbf{x}_j^c)^2|x_{ij}-v_{22}|\right]$ as $\delta_2(\mathbf{x}_i^r)$, we get

$$p_1(\mathbf{x}_i^r)\delta_1(\mathbf{x}_i^r) = p_2(\mathbf{x}_i^r)\delta_2(\mathbf{x}_i^r), \quad i=1,\ldots,N, \tag{4.48}$$

which resembles the Principle 1 in (4.3).

Column Membership Problem

Similarly, if centers v_{kl}, $k=1,\ldots,K$ and $l=1,\ldots,L$ are known and row probabilities $p_k(\mathbf{x}_i^r)$ are fixed, then the optimization problem in (4.42) is

$$\begin{aligned} \underset{q_l(\mathbf{x}_j^c)}{\text{minimize}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2|x_{ij}-v_{kl}| \\ \text{subject to} \quad & \sum_{l=1}^{L} q_l(\mathbf{x}_j^c) = 1, \quad j=1,\ldots,M, \\ & q_l(\mathbf{x}_j^c) \geq 0, \quad j=1,\ldots,M, \quad l=1,\ldots,L, \end{aligned} \tag{4.49}$$

which is now a convex optimization problem of $q_l(\mathbf{x}_j^c)$.

Again, we set $K=L=2$, for simplicity. Then Problem in (4.49) can be decomposed for each $\mathbf{x}_j^c$, and it becomes

$$\begin{aligned} \underset{q_1(\mathbf{x}_j^c),\, q_2(\mathbf{x}_j^c)}{\text{minimize}} \quad & \sum_{i=1}^{N} p_1(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2|x_{ij}-v_{11}| + p_1(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2|x_{ij}-v_{12}| \\ & + p_2(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2|x_{ij}-v_{21}| + p_2(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2|x_{ij}-v_{22}| \\ \text{subject to} \quad & q_1(\mathbf{x}_j^c) + q_2(\mathbf{x}_j^c) = 1, \quad j=1,\ldots,M, \\ & q_1(\mathbf{x}_j^c), q_2(\mathbf{x}_j^c) \geq 0, \quad j=1,\ldots,M. \end{aligned} \tag{4.50}$$

Lagrangian of problem (4.50) is

$$\mathcal{L}(q_1(\mathbf{x}_j^c), q_2(\mathbf{x}_j^c), \mu_j) = \Bigg(\sum_{i=1}^{N} p_1(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 |x_{ij} - v_{11}| + p_1(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 |x_{ij} - v_{12}|$$
$$+ p_2(\mathbf{x}_i^r)^2 q_1(\mathbf{x}_j^c)^2 |x_{ij} - v_{21}| + p_2(\mathbf{x}_i^r)^2 q_2(\mathbf{x}_j^c)^2 |x_{ij} - v_{22}| \Bigg) - \mu_j (q_1(\mathbf{x}_j^c) + q_2(\mathbf{x}_j^c) - 1)$$

$$\frac{\partial \mathcal{L}}{\partial q_1(\mathbf{x}_j^c)} = 2q_1(\mathbf{x}_j^c) \Bigg[\sum_{i=1}^{N} p_1(\mathbf{x}_i^r)^2 |x_{ij} - v_{11}| + p_2(\mathbf{x}_i^r)^2 |x_{ij} - v_{21}| \Bigg] - \mu_j = 0 \quad (4.51)$$

$$\frac{\partial \mathcal{L}}{\partial q_2(\mathbf{x}_j^c)} = 2q_2(\mathbf{x}_j^c) \Bigg[\sum_{i=1}^{N} p_1(\mathbf{x}_i^r)^2 |x_{ij} - v_{12}| + p_2(\mathbf{x}_i^r)^2 |x_{ij} - v_{22}| \Bigg] - \mu_j = 0 \quad (4.52)$$

$$\frac{\partial \mathcal{L}}{\partial \mu_j} = q_1(\mathbf{x}_j^c) + q_2(\mathbf{x}_j^c) - 1 = 0$$

(4.51) and (4.52) yield

$$q_1(\mathbf{x}_j^c) \Bigg[\sum_{i=1}^{N} p_1(\mathbf{x}_i^r)^2 |x_{ij} - v_{11}| + p_2(\mathbf{x}_i^r)^2 |x_{ij} - v_{21}| \Bigg]$$
$$= q_2(\mathbf{x}_j^c) \Bigg[\sum_{i=1}^{N} p_1(\mathbf{x}_i^r)^2 |x_{ij} - v_{12}| + p_2(\mathbf{x}_i^r)^2 |x_{ij} - v_{22}| \Bigg], \quad j = 1, \ldots, M.$$

When we denote the terms $\Bigg[\sum_{i=1}^{N} p_1(\mathbf{x}_i^r)^2 |x_{ij} - v_{11}| + p_2(\mathbf{x}_i^r)^2 |x_{ij} - v_{21}| \Bigg]$ as $\bar{\delta}_1(\mathbf{x}_j^c)$, and $\Bigg[\sum_{i=1}^{N} p_1(\mathbf{x}_i^r)^2 |x_{ij} - v_{12}| + p_2(\mathbf{x}_i^r)^2 |x_{ij} - v_{22}| \Bigg]$ as $\bar{\delta}_2(\mathbf{x}_j^c)$, we get

$$q_1(\mathbf{x}_j^c) \bar{\delta}_1(\mathbf{x}_j^c) = q_2(\mathbf{x}_j^c) \bar{\delta}_2(\mathbf{x}_j^c), \quad j = 1, \ldots, M, \quad (4.53)$$

which resembles the Principle 2 in (4.5).

4.3.3 Underlying Principles

Consider the term $\delta_k(\mathbf{x}_i^r)$ in (4.48), which is

$$\delta_k(\mathbf{x}_i^r) = \sum_{j=1}^{M} \sum_{l=1}^{L} q_l(\mathbf{x}_j^c)^2 |x_{ij} - v_{kl}|, \quad i = 1, \ldots, N. \quad (4.54)$$

From (4.54), one can observe that $\delta_k(\mathbf{x}_i^r)$ depends on column membership probabilities $q_l(\mathbf{x}_j^c)$.

Recall the problem property (4.6) in Section 4.2.1. It states that each element in a column entity $x_{ij} \in \mathbf{x}_j^c$ should have the same column membership probability of $q_l(\mathbf{x}_j^c)$. Thus, we can rewrite (4.54) in terms of x_{ij} as follows

$$\delta_k(\mathbf{x}_i^r) = \sum_{j=1}^{M} \sum_{l=1}^{L} q_l(x_{ij})^2 |x_{ij} - v_{kl}|, \quad x_{ij} \in \mathbf{x}_j^c, \quad i = 1, \ldots, N. \tag{4.55}$$

When a row entity $\mathbf{x}_i^r$ is assigned to a row cluster k with a probability of $p_k(\mathbf{x}_i^r)$, its elements $x_{ij} \in \mathbf{x}_i^r$ are probabilistically assigned to submatrix centers $\{v_{kl} : l = 1, ..., L\}$ with the respective column membership probabilities of $q_l(x_{ij})$. Thus, row assignments are dependent to column assignments as in (4.54) or (4.55).

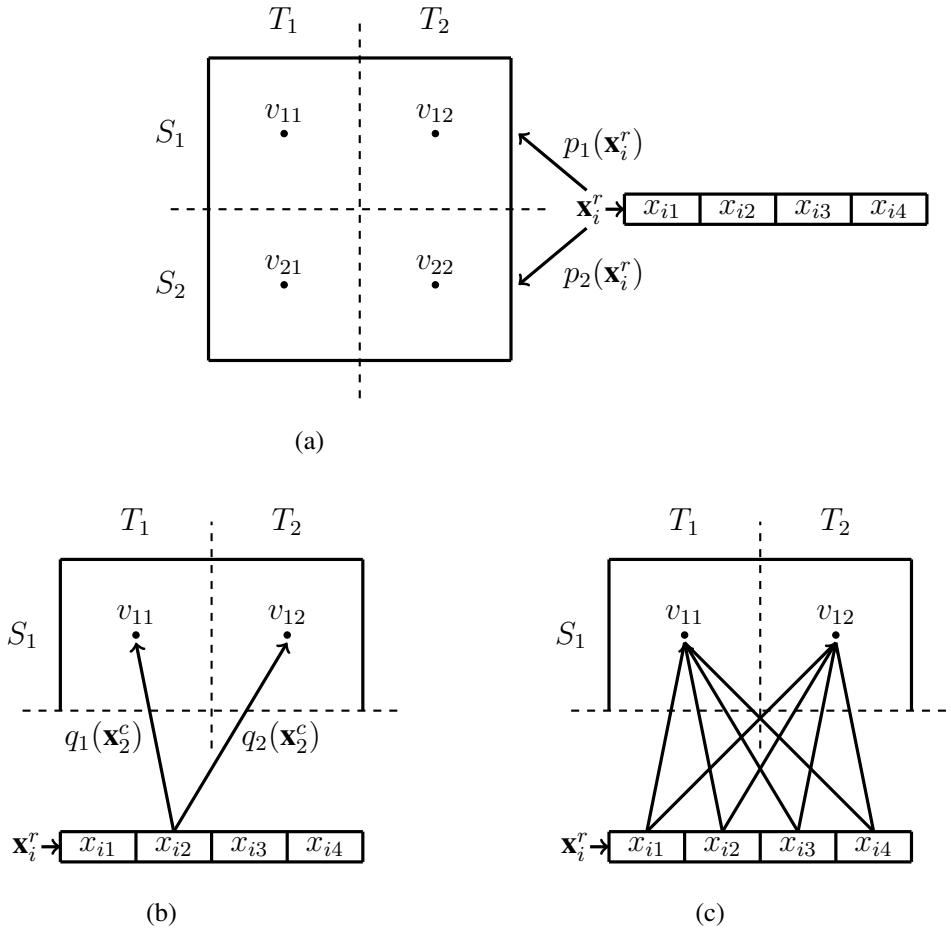


Figure 4.7: (a) Representation of partitions when $K = L = 2$, (b) considering assignment of $\mathbf{x}_i^r$ to row cluster $k = 1$, the soft assignments of row elements x_{ij} to corresponding submatrix centers , (c) soft assignments of $\forall x_{ij} \in \mathbf{x}_i^r$

To illustrate, consider Figure 4.7. Let $K = L = 2$, and any row entity $\mathbf{x}_i^r$ can be assigned to row clusters with its membership probabilities $p_1(\mathbf{x}_i^r)$ and $p_2(\mathbf{x}_i^r)$ (see Figure 4.7a). We evaluate the membership of a row entity $\mathbf{x}_i^r$ to the row cluster 1 and consider the submatrix centers v_{11} and v_{12} as they belong to row cluster 1. Let row entity $\mathbf{x}_i^r$ has four elements $\{x_{i1}, x_{i2}, x_{i3}, x_{i4}\}$, which belong to column entities $\mathbf{x}_1^c$, $\mathbf{x}_2^c$, $\mathbf{x}_3^c$ and $\mathbf{x}_4^c$, respectively. Thus, if $\mathbf{x}_i^r$ is assigned to row cluster 1 with its membership probability of $p_1(\mathbf{x}_i^r)$, then; for example, the element x_{i2} gets assigned to v_{11} and v_{12} with its column membership probabilities $q_1(\mathbf{x}_2^c)$ and $q_2(\mathbf{x}_2^c)$ (See Figure 4.7b). This will be considered for each element of $\mathbf{x}_i^r$ as in Figure 4.7c.

Following the illustration in Figure 4.7, for each $\mathbf{x}_i^r$, $i \in \{1, \ldots, N\}$, given a row cluster k and $L = 2$, (4.54) will be

$$\delta_k(\mathbf{x}_i^r) = \sum_{j=1}^{M} q_1(\mathbf{x}_j^c)^2 |x_{ij} - v_{k1}| + q_2(\mathbf{x}_j^c)^2 |x_{ij} - v_{k2}|,$$

which is equivalent to

$$\delta_k(\mathbf{x}_i^r) = \sum_{j=1}^{M} q_1(\mathbf{x}_j^c)(q_1(\mathbf{x}_j^c)|x_{ij} - v_{k1}|) + q_2(\mathbf{x}_j^c)(q_2(\mathbf{x}_j^c)|x_{ij} - v_{k2}|). \tag{4.56}$$

The terms $(q_1(\mathbf{x}_j^c)|x_{ij} - v_{k1}|)$ and $(q_2(\mathbf{x}_j^c)|x_{ij} - v_{k2}|)$ represent the weighted distance of element x_{ij} to the column clusters 1 and 2, respectively. Also, from the fact that probabilities add up to one, (4.56) is the sum of convex combination of weighted distances to column clusters. Therefore, we construct Principle 4 as follows

Principle 4. *For each row entity $\boldsymbol{x}_i^r \in \boldsymbol{X}$, and each row cluster k,*

$$p_k(\mathbf{x}_i^r)\delta_k(\mathbf{x}_i^r) = E(\mathbf{x}_i^r), \tag{4.57}$$

where $E(\boldsymbol{x}_i^r) =$ a constant, depending on $\boldsymbol{x}_i^r$.

Membership probability of a row entity $\mathbf{x}_i^r$ to row cluster k is higher when the convex combination of row entity elements' weighted distances to column clusters $\delta_k(\mathbf{x}_i^r)$ is smaller.

A similar principle can be derived for column entities as follows.

Principle 5. *For each column entity $\boldsymbol{x}_j^c \in \boldsymbol{X}$, and each column cluster l,*

$$q_l(\mathbf{x}_j^c)\bar{\delta}_l(\mathbf{x}_j^c) = F(\mathbf{x}_j^c), \tag{4.58}$$

where $F(\boldsymbol{x}_j^c) =$ a constant, depending on $\boldsymbol{x}_j^c$.

Membership probability of a column entity $\mathbf{x}_j^c$ to column cluster l is higher when the convex combination of column entity elements' weighted distances to row clusters $\bar{\delta}_l(\mathbf{x}_j^c)$ is smaller.

4.3.4 Probabilities

From Principle 4 in (4.57), and the fact that probabilities $p_k(\boldsymbol{.})$ add to one, we get

Theorem 9. *Let the submatrix centers $\{v_{11}, v_{12}, ..., v_{kl}\}$ be given, let $\boldsymbol{x}_i^r$ be a row entity of data set $\boldsymbol{X}$, and $\{\delta_k(\boldsymbol{x}_i^r) : k = 1, ..., K\}$ be its weighted expected distance from the given cluster centers. Then the membership probabilities of $\boldsymbol{x}_i^r$ are*

$$p_k(\boldsymbol{x}_i^r) = \frac{\dfrac{1}{\delta_k(\boldsymbol{x}_i^r)}}{\displaystyle\sum_{s=1}^{K} \frac{1}{\delta_s(\boldsymbol{x}_i^r)}}. \tag{4.59}$$

Proof. Using (4.57) we write for s,k,

$$p_s(\mathbf{x}_i^r) = \left(\frac{p_k(\mathbf{x}_i^r)\delta_k(\mathbf{x}_i^r)}{\delta_s(\mathbf{x}_i^r)}\right).$$

Since $\displaystyle\sum_{s=1}^{K} p_s(\mathbf{x}_i^r) = 1$,

$$p_k(\mathbf{x}_i^r)\delta_k(\mathbf{x}_i^r) \sum_{s=1}^{K} \left(\frac{1}{\delta_s(\mathbf{x}_i^r)}\right) = 1.$$

$$p_k(\mathbf{x}_i^r) = \frac{1}{\delta_k(\mathbf{x}_i^r) \displaystyle\sum_{s=1}^{K} \left(\frac{1}{\delta_s(\mathbf{x}_i^r)}\right)} = \frac{\dfrac{1}{\delta_k(\mathbf{x}_i^r)}}{\displaystyle\sum_{s=1}^{K} \frac{1}{\delta_s(\mathbf{x}_i^r)}}.$$

□

Similarly, column membership probabilities are

$$q_l(\mathbf{x}_j^c) = \frac{\dfrac{1}{\bar{\delta}_l(\mathbf{x}_j^c)}}{\sum\limits_{t=1}^{L} \dfrac{1}{\bar{\delta}_t(\mathbf{x}_j^c)}}. \tag{4.60}$$

In particular, for $K = L = 2$,

$$p_1(\mathbf{x}_i^r) = \frac{\dfrac{1}{\bar{\delta}_1(\mathbf{x}_i^r)}}{\dfrac{1}{\bar{\delta}_1(\mathbf{x}_i^r)} + \dfrac{1}{\bar{\delta}_2(\mathbf{x}_i^r)}}, \quad p_2(\mathbf{x}_i^r) = \frac{\dfrac{1}{\bar{\delta}_2(\mathbf{x}_i^r)}}{\dfrac{1}{\bar{\delta}_1(\mathbf{x}_i^r)} + \dfrac{1}{\bar{\delta}_2(\mathbf{x}_i^r)}},$$
$$q_1(\mathbf{x}_j^c) = \frac{\dfrac{1}{\bar{\bar{\delta}}_1(\mathbf{x}_j^c)}}{\dfrac{1}{\bar{\bar{\delta}}_1(\mathbf{x}_j^c)} + \dfrac{1}{\bar{\bar{\delta}}_2(\mathbf{x}_j^c)}}, \quad q_2(\mathbf{x}_j^c) = \frac{\dfrac{1}{\bar{\bar{\delta}}_2(\mathbf{x}_j^c)}}{\dfrac{1}{\bar{\bar{\delta}}_1(\mathbf{x}_j^c)} + \dfrac{1}{\bar{\bar{\delta}}_2(\mathbf{x}_j^c)}}.$$

4.3.5 Center Problem

Given the membership probabilities, the objective of optimization problem in (4.42) is

$$f(v_{kl}) = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2 |x_{ij} - v_{kl}|. \tag{4.61}$$

Theorem 10. *For (4.61), the minimizers of v_{kl} are given by*

$$v_{kl} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} \dfrac{p_k(\boldsymbol{x}_i^r)^2 q_l(\boldsymbol{x}_j^c)^2}{|x_{ij} - v_{kl}|} x_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{M} \dfrac{p_k(\boldsymbol{x}_i^r)^2 q_l(\boldsymbol{x}_j^c)^2}{|x_{ij} - v_{kl}|}}, \quad k = 1, \ldots, K, \quad l = 1, \ldots, L. \tag{4.62}$$

Proof. The gradient of $|x_{ij} - v_{kl}|$ with respect to v_{kl} is

$$\nabla_{v_{kl}} |x_{ij} - v_{kl}| = -\frac{x_{ij} - v_{kl}}{|x_{ij} - v_{kl}|}.$$

Thus, the gradient (4.61) with respect to v_{kl} is

$$\nabla_{v_{kl}} f(v_{kl}) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2 \frac{x_{ij} - v_{kl}}{|x_{ij} - v_{kl}|}, \quad \forall k, l. \tag{4.63}$$

By setting the gradient equal to zero, we get

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2}{|x_{ij}-v_{kl}|}x_{ij} = \left(\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2}{|x_{ij}-v_{kl}|}\right)v_{kl},$$

then

$$v_{kl} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2}{|x_{ij}-v_{kl}|}x_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2}{|x_{ij}-v_{kl}|}},$$

proving (4.62). □

4.3.6 Algorithm

We propose an iterative algorithm by following the ideas presented in the previous sections. The steps of the algorithm are given as follows:

Algorithm 2: Proposed Algorithm 2

Step 1. Given the data set $\mathbf{X}$, and $\epsilon > 0$ initialize the membership probabilities $p_k(\mathbf{x}_i^r)$ and $q_l(\mathbf{x}_j^c)$ for all $\mathbf{x}_i^r, \mathbf{x}_j^c \in \mathbf{X}$.

Step 2. Update the the cluster centers v_{kl}, $k = 1, \ldots, K$, and $l = 1, \ldots, L$ by

$$v_{kl}^{+} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2}{|x_{ij}-v_{kl}|}x_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{p_k(\mathbf{x}_i^r)^2 q_l(\mathbf{x}_j^c)^2}{|x_{ij}-v_{kl}|}} \quad \text{as in (4.62)}$$

Step 3. Compute the membership probabilities of $p_k(\mathbf{x}_i^r)$ and $q_l(\mathbf{x}_j^c)$ for all $\mathbf{x}_i^r, \mathbf{x}_j^c \in \mathbf{X}$ respectively by

$$p_k(\mathbf{x}_i^r) = \frac{\frac{1}{\delta_k(\mathbf{x}_i^r)}}{\sum_{s=1}^{K}\frac{1}{\delta_s(\mathbf{x}_i^r)}}, \quad \text{and} \quad q_l(\mathbf{x}_j^c) = \frac{\frac{1}{\bar{\delta}_l(\mathbf{x}_j^c)}}{\sum_{t=1}^{L}\frac{1}{\bar{\delta}_t(\mathbf{x}_j^c)}} \quad \text{as in (4.59) and (4.60)}$$

where

$\delta_k(\mathbf{x}_i^r) = \sum_{j=1}^{M}\sum_{l=1}^{L} q_l(\mathbf{x}_j^c)^2 |x_{ij} - v_{kl}^{+}|$

$\bar{\delta}_l(\mathbf{x}_j^c) = \sum_{i=1}^{N}\sum_{k=1}^{K} p_k(\mathbf{x}_i^r)^2 |x_{ij} - v_{kl}^{+}|$

Step 4. If $\sum_{k=1}^{K}\sum_{l=1}^{L} \| v_{kl}^{+} - v_{kl} \| < \epsilon$, stop. Otherwise, continue with Step 2.

Note that in **Step 1**, membership probabilities are determined randomly for each entity and cluster centers are calculated as the average of elements assigned to a submatrix.

CHAPTER 5

EXPERIMENTAL STUDY

5.1 Experiment Settings

As performances of clustering techniques are significantly affected by the features of the data sets, we evaluate the performance of our proposed algorithms by generating data sets with different properties.

We use a data generation method having a similar approach in [35]. To create a data set $\mathbf{X}_{N\times M}$, we need row and column memberships, $\mathbf{P}_{N\times K}$ and $\mathbf{Q}_{M\times L}$, and submatrix centers, $\mathbf{V}_{K\times L}$. Lastly, error perturbations to each element of the data set, $\mathbf{E}_{N\times M}$, are added. Thus, the data set is obtained by

$$\mathbf{X} = \mathbf{PVQ}' + \mathbf{E}. \tag{5.1}$$

The data generation algorithm is given below.

Algorithm 3: Main Data Generation Algorithm

Input: N, M, K, L, σ

Step 1. Generate membership $\{p_{ik} \in \mathbf{P} \mid p_{ik} = \{0, 1\}\}$ for each row entity i to a row cluster $\{k \mid k = 1, .., K\}$ with discrete uniform distribution.

Step 2. Generate membership $\{q_{jl} \in \mathbf{Q} \mid q_{jl} = \{0, 1\}\}$ for each column entity j to a column cluster $\{l \mid l = 1, .., L\}$ with discrete uniform distribution.

Step 3. Generate a random value from interval $\{[t - 0.2,\ t + 0.2] \mid t = 1, ..., K \times L\}$ for each submatrix center $v_{kl} \in \mathbf{V}$. Then, shuffle $\mathbf{V}$.

Step 4. Generate a perturbation value e_{ij} for each element x_{ij}, where $e_{ij} \in \mathbf{E}$ is normally distributed with a mean of 0 and standard deviation σ.

Step 5. Compute $\mathbf{X} = \mathbf{PVQ}' + \mathbf{E}$.

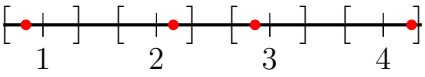
Output: **X**, **P**, **Q**

The input parameters N, M, K, and L affect the problem's difficulty in terms of complexity (see (4.2)). We set the following levels to these parameters for the experimental study:

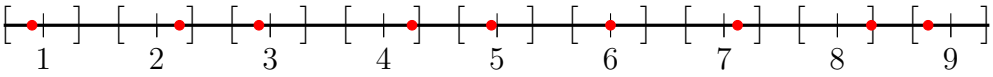
- Small size data set: $N = M = 20$, medium size data set: $N = M = 50$, large size data set: $N = M = \{80, 150\}$.
- $K = L = 2$, $K = L = 3$, $K = L = 4$, and $K = L = 5$.

The selection of submatrix centers affects the difficulty of the problem in terms of the disjointness of partitions. The distance between submatrix centers is the first factor that affects the separability of the submatrices. As it can be observed from Figure 5.1, we ensure a certain minimum distance between submatrix centers by selecting them from the interval $[t - 0.2,\ t + 0.2]$ at Step 3. The orientation of submatrices also affects the separability of the data set. Consider the data sets in Figure 5.2. Although these data sets are formed from the same submatrices, they differ due to orientation. The data set in Figure 5.2a has clear boundaries between its submatrices. On the other hand, data sets in Figure 5.2b and Figure 5.2c have blurry boundaries between their some submatrices, making them harder to analyze than the one in5.2a. To simulate various orientations, we shuffle the submatrix centers in Step 3.

The size of the error perturbations also affects the clarity of submatrices, and it is controlled by the term standard deviation σ in Step 4. To see the impact of parameter σ, consider the data sets in Figure 5.3. For the lower level of σ, the submatrix distinction is more visible as in Figure 5.3a. When σ increases, the visibility of submatrices diminishes. The submatrix structure is almost lost for a high level of σ, as in Figure 5.3c. Thus, as σ increases, the data set becomes harder to analyze.

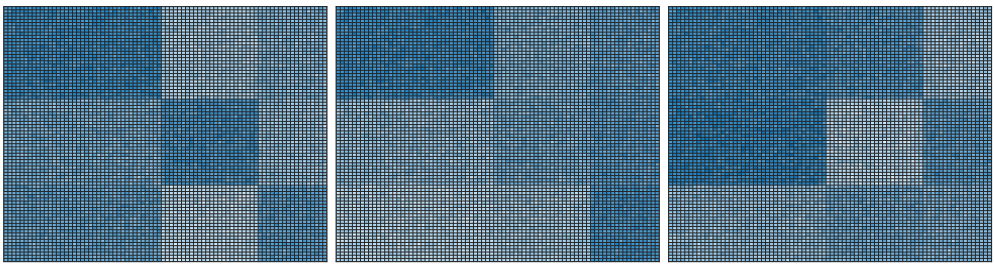


(a) An example realization of submatrix centers for $K = L = 2$



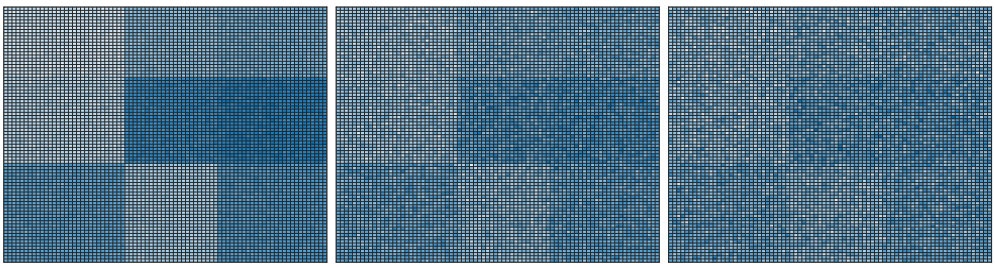
(b) An example realization of submatrix centers for $K = L = 3$

Figure 5.1: Submatrix center locations, represented by red dots, determined by Step 3 for different number of partitions



(a) (b) (c)

Figure 5.2: Orientation of submatrices for a data set with $N = M = 80$ and $K = L = 3$

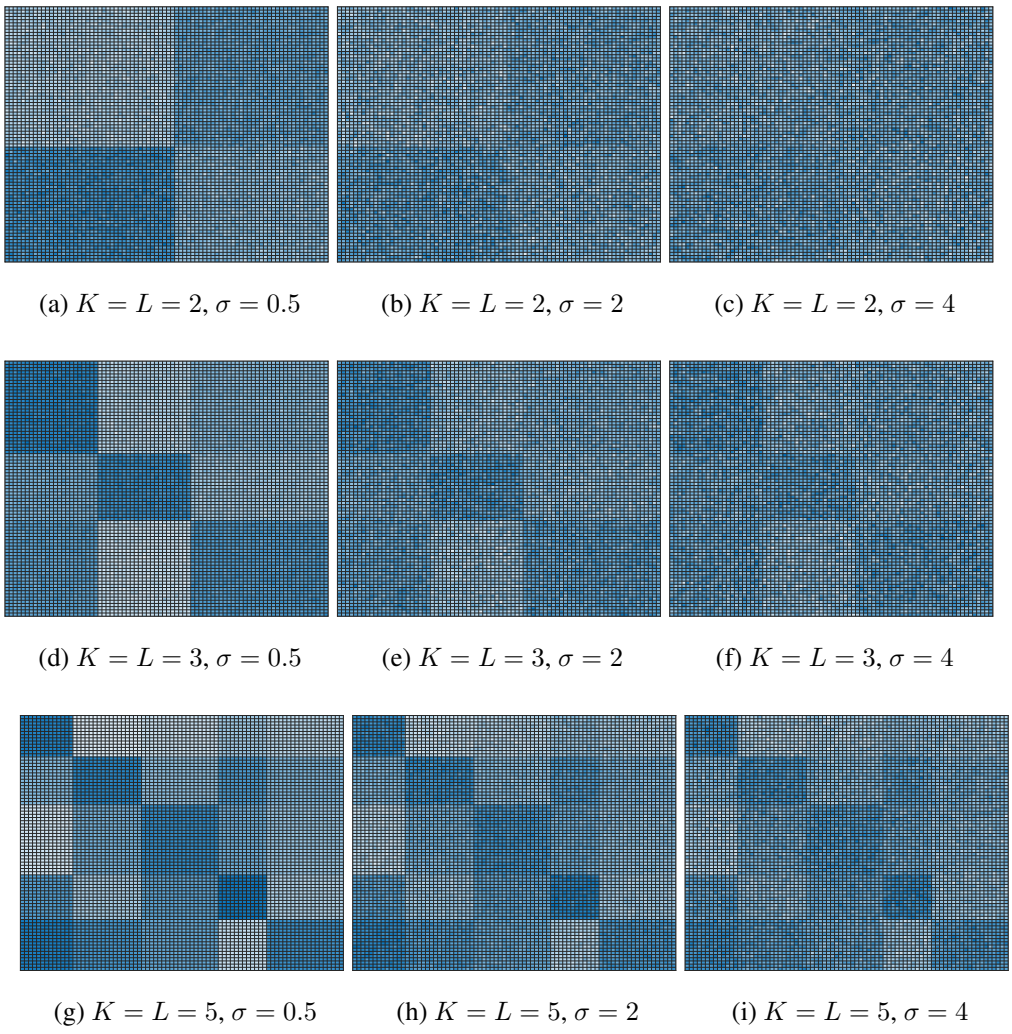


(a) (b) (c)

Figure 5.3: Different error perturbation levels for a data set with $N = M = 80$ and $K = L = 3$, (a) $\sigma = 0.5$, (b) $\sigma = 2$, (c) $\sigma = 4$

Since center locations are selected from incompatible ranges for different partitioning levels (see Figure 5.1), the effect of a σ level on the data separability alters with the number of partitions. Consider the Figure 5.4, for $K = L = 2$, differentiation of submatrices significantly decreases when $\sigma = 2$. On the other hand, for $K = L = 5$, even with a larger level of $\sigma = 4$, the boundaries of submatrices are still clear (see Figure 5.4i). To overcome this issue, we determine four levels of σ as *low*, *moderate*, *high moderate*, and *high* from functions of K and L:

- Low error level: $\sigma = 0.02KL$, moderate error level: $\sigma = 0.1KL$, high moderate error level: $\sigma = 0.2KL$, high error level: $\sigma = 0.4KL$.



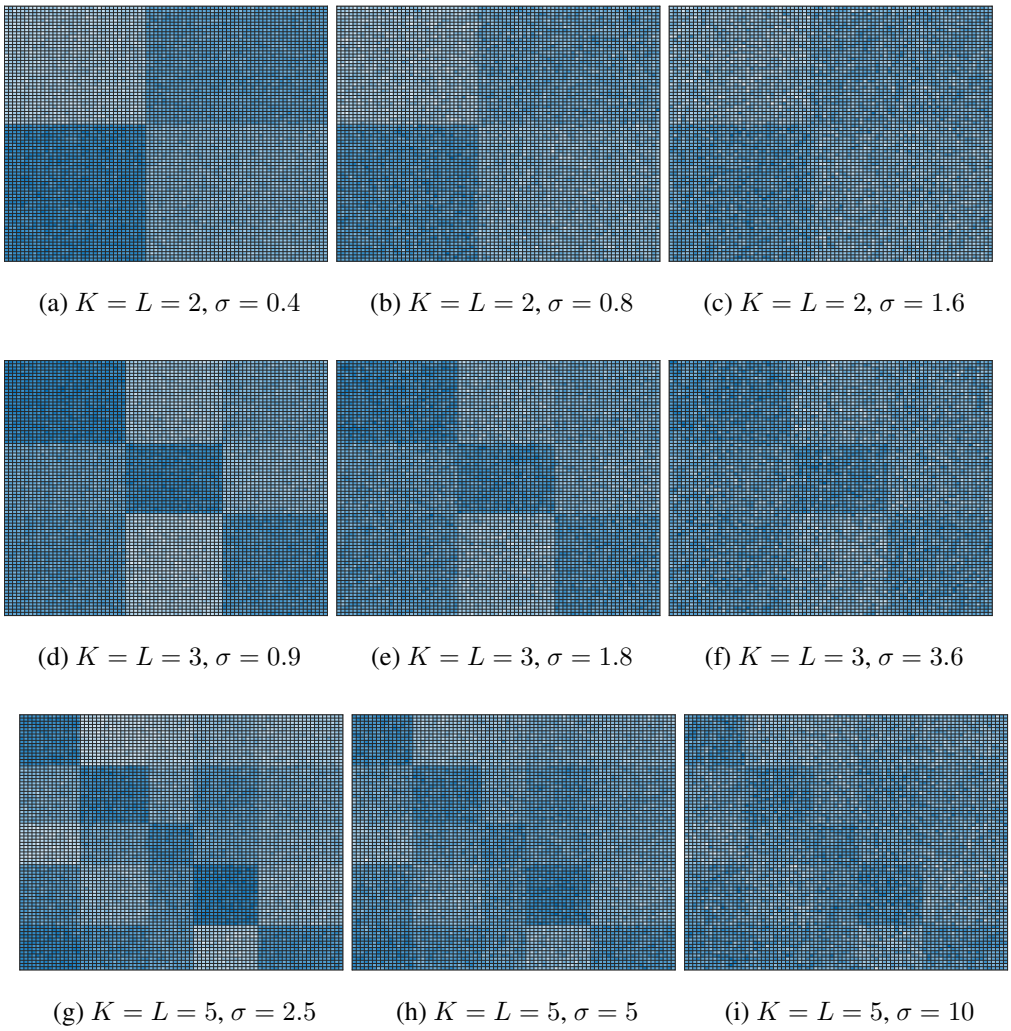
(a) $K = L = 2, \sigma = 0.5$ (b) $K = L = 2, \sigma = 2$ (c) $K = L = 2, \sigma = 4$

(d) $K = L = 3, \sigma = 0.5$ (e) $K = L = 3, \sigma = 2$ (f) $K = L = 3, \sigma = 4$

(g) $K = L = 5, \sigma = 0.5$ (h) $K = L = 5, \sigma = 2$ (i) $K = L = 5, \sigma = 4$

Figure 5.4: The same σ levels for different K, L levels

Data sets with different σ levels obtained from functions of K, L can be seen in Figure 5.5.



(a) $K = L = 2, \sigma = 0.4$ (b) $K = L = 2, \sigma = 0.8$ (c) $K = L = 2, \sigma = 1.6$

(d) $K = L = 3, \sigma = 0.9$ (e) $K = L = 3, \sigma = 1.8$ (f) $K = L = 3, \sigma = 3.6$

(g) $K = L = 5, \sigma = 2.5$ (h) $K = L = 5, \sigma = 5$ (i) $K = L = 5, \sigma = 10$

Figure 5.5: Data sets with various σ levels depending on a function of $K = L = \{2, 3, 5\}$

As a result, we have 4 levels of number of entities (N, M), 4 levels of number of partitions (K, L), and 4 levels of error perturbation, making a total of 64 combinations. We aim to test the proposed algorithms with various submatrix center generations. Therefore, for each parameter combination, we generate 30 data sets. As a result, we create a total of 1920 data sets for computational experiments.

5.2 Performance Measures

Since we have two proposed algorithms, each with different objectives, we need external measures to evaluate and compare their performances. Rand index (RI) is widely used in the literature to assess the performance of clustering approaches [41]. RI is calculated with the following ratio:

$$RI(Y, Z) = \frac{|A \cap C| + |B \cap D|}{|A \cap C| + |A \cap D| + |B \cap C| + |B \cap D|}, \tag{5.2}$$

where

A**:** Set of entity pairs belonging to same cluster in reference partitioning **Y**,
B**:** Set of entity pairs belonging to different cluster in reference partitioning **Y**,
C**:** Set of entity pairs belonging to same cluster in a partitioning **Z**,
D**:** Set of entity pairs belonging to different cluster in a partitioning **Z**.

In the original definition of RI, the reference partition **Y** and a partition to be evaluated **Z** are both hard assignments, where each data point is strictly assigned to a cluster. In our study, we have strict reference partitions and soft partitions to be compared. Therefore, we use Campello's fuzzy rand index (FRI) given in [42]. FRI is computed with revising the set definitions in (5.2) as follows:

$A(i_1, i_2)$**:** For reference partitioning **Y**, the membership probability of an entity pair (i_1, i_2) to the same cluster k:
$max\{min\{\mathbf{Y}(i_1, k_1), \mathbf{Y}(i_2, k_1)\}, min\{\mathbf{Y}(i_1, k_2), \mathbf{Y}(i_2, k_2)\}, ...\}$

$B(i_1, i_2)$**:** For reference partitioning **Y**, the membership probability of an entity pair (i_1, i_2) to the different clusters:
$max\{min\{\mathbf{Y}(i_1, k_1), \mathbf{Y}(i_2, k_2)\}, min\{\mathbf{Y}(i_1, k_2), \mathbf{Y}(i_2, k_1)\}, ...\}$, where $k_1 \neq k_2$,

$C(i_1, i_2)$**:** For a partitioning **Z**, the membership probability of an entity pair (i_1, i_2) to the same cluster k:
$max\{min\{\mathbf{Z}(i_1, k_1), \mathbf{Z}(i_2, k_1)\}, min\{\mathbf{Z}(i_1, k_2), \mathbf{Z}(i_2, k_2)\}, ...\}$,

$D(i_1, i_2)$**:** For a partitioning **Z**, the membership probability of an entity pair (i_1, i_2) to the different clusters:

$max\{min\{\mathbf{Z}(i_1,k_1),\mathbf{Z}(i_2,k_2)\},min\{\mathbf{Z}(i_1,k_2),\mathbf{Z}(i_2,k_1)\},...\}$, where $k_1 \neq k_2$.

Then,

$$|A \cap C| = \sum_{i_1=1}^{i_2-1}\sum_{i_2=2}^{N} min\{A(i_1,i_2),C(i_1,i_2)\} \tag{5.3}$$

$$|A \cap D| = \sum_{i_1=1}^{i_2-1}\sum_{i_2=2}^{N} min\{A(i_1,i_2),D(i_1,i_2)\} \tag{5.4}$$

$$|B \cap C| = \sum_{i_1=1}^{i_2-1}\sum_{i_2=2}^{N} min\{B(i_1,i_2),C(i_1,i_2)\} \tag{5.5}$$

$$|B \cap D| = \sum_{i_1=1}^{i_2-1}\sum_{i_2=2}^{N} min\{B(i_1,i_2),D(i_1,i_2)\} \tag{5.6}$$

Fuzzy Rand Index (FRI) is measured by substituting (5.3)-(5.6) into (5.2).

Since we deal with two-mode partitioning, FRI is computed by joint probabilities $\mathbf{R}_{kl}(x_{ij})$, which estimates the probability of an element x_{ij} is assigned to submatrix $\mathbf{V}_{kl}$, (see Principle 3 in (4.9)). Reference partitioning $\mathbf{Y}$ is computed with original row and column clusters, $\mathbf{P}$ and $\mathbf{Q}$, as

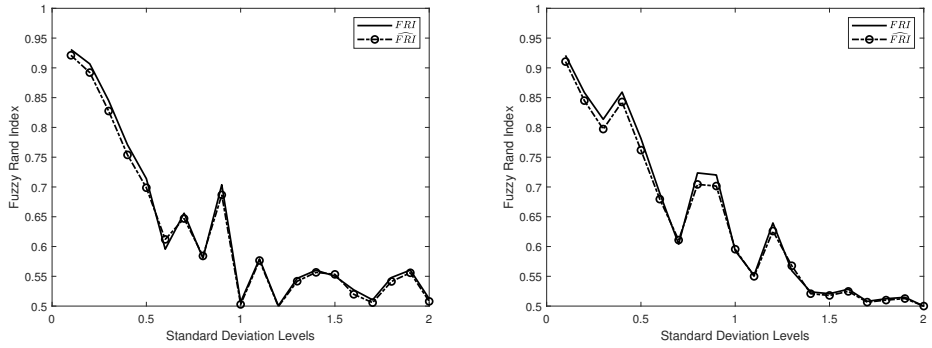
$$\mathbf{Y}_{kl}(x_{ij}) = p_{ik}q_{jl}, \quad \forall i,j,k,l. \tag{5.7}$$

Since FRI is computationally time consuming, we approximate $FRI(\mathbf{Y},\mathbf{R})$ as

$$\widehat{FRI}(\mathbf{Y},\mathbf{R}) = f(FRI(\mathbf{P},\mathbf{P}_r))f(FRI(\mathbf{Q},\mathbf{Q}_r)), \tag{5.8}$$

where $\mathbf{P}_r$ and $\mathbf{Q}_r$ are resulting row and column clusters obtained from proposed algorithms, and $f(\centerdot)$ is a scaling function.

As it can be observed from Figure 5.6, $\widehat{FRI}$ provides a good approximation of FRI. For lower levels of σ, both indexes are high. As σ increases, indexe values decrease since the data becomes harder to partition. In terms of CPU time, the average computing time of FRI is 0.1 and 53.1 seconds for 10×10 and 20×20 data sets, respectively. On the other hand, $\widehat{FRI}$ takes 0.001 and 0.002, respectively.



(a) $N = M = 10$, $K = L = 2$ (b) $N = M = 20$, $K = L = 2$

Figure 5.6: Behavior of FRI and $\widehat{FRI}$ for several σ levels

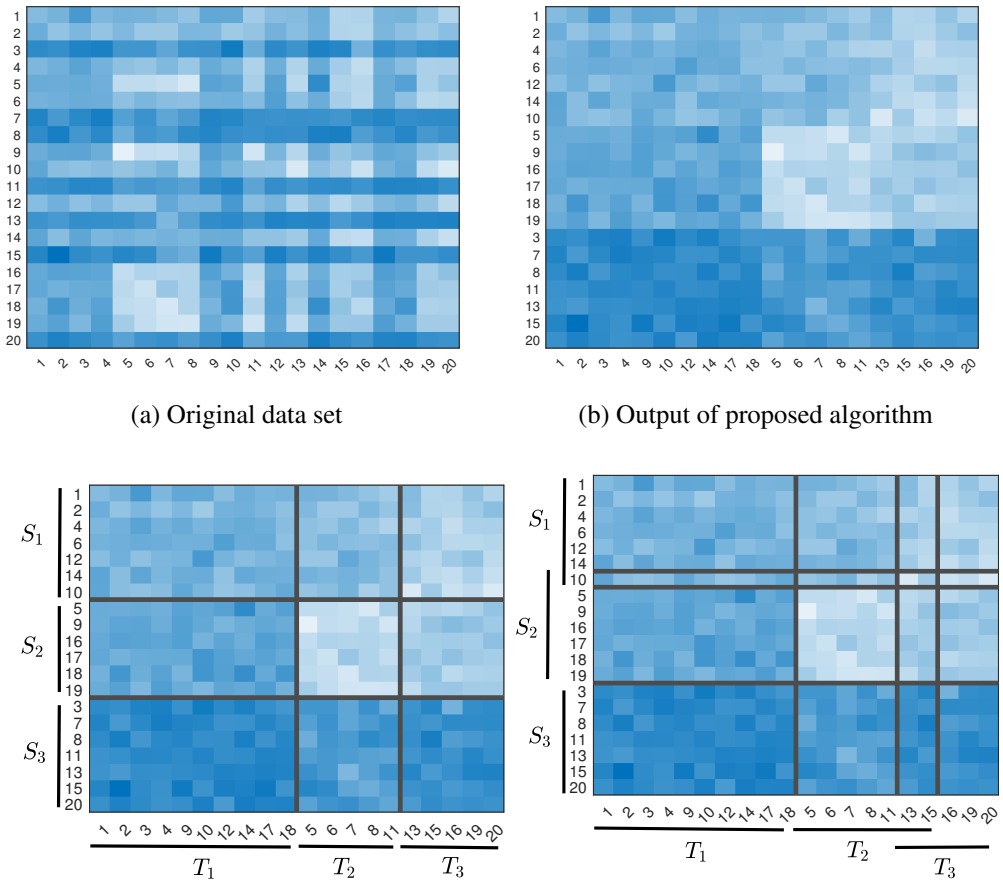
Besides $\widehat{FRI}$, we also compute an approximated Rand Index $\widehat{RI}$ with obtaining hard assignments of row and column entities based on their soft assignments. The approximation function is derived with a similar approach given in (5.8). We apply two hardening methods to resulting clusters, $\mathbf{P}_r$ and $\mathbf{Q}_r$

Strict Hardening: Entities are strictly assigned to one cluster. Hard assignments are obtained by assigning an entity to a cluster according to the highest membership probability. The resulting Rand Index is denoted by $\widehat{RI}^s$.

Overlapping Hardening: Overlapping clusters are allowed. Thus, hard assignments are obtained by allowing each entity to be assigned to multiple clusters at which its membership probability exceeds specific cut points, $\tau^r = 1/K$ and $\tau^c = 1/L$ for rows and columns, respectively. We denote the resulting Rand Index by $\widehat{RI}^o$.

Figure 5.7 represents hard assignments of entities by *strict* and *overlapping* hardening for resulting clusters $\mathbf{P}_r$ and $\mathbf{Q}_r$ of a given data set.

Lastly, CPU time for each algorithm is also considered as a performance measure. Proposed algorithms are coded in MATLAB R2020b, and they are executed on an Intel(R) Core(TM) i7-10510U 2.30 GHz processor with 16 GB RAM.



(a) Original data set

(b) Output of proposed algorithm

(c) Strict hardening

(d) Overlapping hardening

Figure 5.7: Resulting row and column partitions with strict and overlapping hardening

5.3 Experiment Results

In this section, performances of proposed algorithms are compared based on measures introduced in Section 5.2. The effect of initialization on the performance of the algorithms is avoided by adopting a multi-start approach. Each algorithm is executed 100 times, and the best solution in terms of the objective function is reported as the final result. Thus, the CPU time is calculated as a sum of 100 starts for each algorithm in seconds. As stated before, 30 data sets are generated for each parameter combination. Average performances of Proposed Algorithm 1 (**PA1**) and Proposed Algorithm 2 (**PA2**) are reported in Table 5.1 and Table 5.2. The first column shows the size of the data set, and the second column is for the number of row and column clusters.

The third column represents the different levels of error perturbation. Each algorithm is evaluated with $\widehat{FRI}$, $\widehat{RI}^s$, $\widehat{RI}^o$, and CPU in the last two columns.

From Table 5.1 and 5.2, it can be observed that **PA1** outperforms **PA2** on the average for most of the parameter combinations. We conduct one-sided paired t-test to compare the performance of the algorithms statistically considering $\widehat{FRI}$ values. For each parameter combination, we compare the mean $\widehat{FRI}$ values of **PA1** and **PA2**, denoted as $\overline{FRI_1}$ and $\overline{FRI_2}$, respectively. Assuming that mean difference is normally distributed, the first hypothesis is as follows:

$$H_0 : \overline{FRI_1} - \overline{FRI_2} \leq 0$$
$$H_1 : \overline{FRI_1} - \overline{FRI_2} > 0$$

Resulting p-values are given in Appendix A Table A.1 at fourth and ninth columns. For 52 out of 64 parameter combinations, the null hypothesis is rejected as p-values are less than the significance level of 0.05. That means **PA1** outperforms **PA2** statistically for those parameter combinations. We also construct a hypothesis for the reverse direction as:

$$H_0 : \overline{FRI_2} - \overline{FRI_1} \leq 0$$
$$H_1 : \overline{FRI_2} - \overline{FRI_1} > 0$$

Resulting p-values are given in Appendix A Table A.1 at fifth and tenth columns. For 8 out of 64 parameter combinations, **PA2** outperforms **PA1** statistically. All these combinations have *high* σ level. There are 16 parameter combination with *high* level of σ. In 5 of them, **PA1** is statistically better than **PA2** based on the first hypothesis test. On the contrary, in 8 of them **PA2** statistically outperforms **PA1**. For the other 3 combinations none of the algorithm outperforms each other. Remind that high σ level refers to the datasets where intrinsic structure of submatrices is hardly identified.

Additionally, we conduct one-sided paired t-test on $\widehat{RI}^s$ values. Results are given in Appendix A Table A.2. The results indicate that based on $\widehat{RI}^s$, **PA1** outperforms **PA2** statistically in 58 out of 64 parameter combinations. For the remaining 6 parameter combinations, it cannot be concluded that an algorithm beats the other one.

Table 5.1: Performances of algorithms for small and medium size data sets

$N = M$	$K = L$	σ	**PA1**				**PA2**			
			$\widehat{FRI}$	$\widehat{RI}^s$	$\widehat{RI}^o$	CPU	$\widehat{FRI}$	$\widehat{RI}^s$	$\widehat{RI}^o$	CPU
20	2	low	0.93	0.99	0.99	0.37	0.89	0.98	0.98	0.33
		moderate	0.78	1.00	1.00	0.32	0.64	0.93	0.93	0.41
		high moderate	0.63	0.98	0.98	1.94	0.54	0.89	0.89	0.40
		high	0.53	0.83	0.83	4.23	0.50	0.72	0.72	0.28
	3	low	0.93	1.00	1.00	1.25	0.62	0.85	0.74	0.65
		moderate	0.73	0.97	0.94	2.62	0.53	0.77	0.66	0.79
		high moderate	0.58	0.87	0.76	4.46	0.51	0.70	0.63	0.67
		high	0.50	0.68	0.63	4.66	0.50	0.62	0.60	0.52
	4	low	0.93	1.00	0.99	1.58	0.53	0.76	0.65	1.21
		moderate	0.71	0.95	0.87	1.99	0.51	0.71	0.64	1.32
		high moderate	0.56	0.82	0.69	5.07	0.50	0.68	0.63	1.26
		high	0.50	0.66	0.62	1.51	0.50	0.62	0.60	0.85
	5	low	0.93	0.99	0.99	2.64	0.51	0.70	0.64	2.11
		moderate	0.67	0.90	0.79	2.39	0.51	0.67	0.63	2.23
		high moderate	0.55	0.80	0.66	4.32	0.50	0.64	0.62	1.88
		high	0.50	0.65	0.63	1.44	0.50	0.59	0.60	1.59
50	2	low	0.94	1.00	1.00	0.30	0.79	0.99	0.99	1.64
		moderate	0.77	1.00	1.00	0.41	0.66	0.97	0.97	2.04
		high moderate	0.64	1.00	1.00	1.81	0.54	0.90	0.90	1.86
		high	0.51	0.89	0.88	7.24	0.50	0.76	0.76	1.18
	3	low	0.94	1.00	1.00	0.62	0.58	0.84	0.73	3.19
		moderate	0.73	1.00	0.95	2.04	0.52	0.79	0.67	4.77
		high moderate	0.59	0.94	0.78	10.07	0.51	0.72	0.64	3.77
		high	0.50	0.73	0.66	2.84	0.50	0.68	0.63	2.51
	4	low	0.93	1.00	1.00	0.89	0.52	0.73	0.64	7.31
		moderate	0.69	0.96	0.86	4.45	0.51	0.71	0.64	9.04
		high moderate	0.54	0.81	0.66	11.61	0.50	0.67	0.63	4.99
		high	0.50	0.69	0.63	1.82	0.50	0.63	0.61	4.39
	5	low	0.93	1.00	1.00	1.48	0.51	0.70	0.63	13.10
		moderate	0.68	0.96	0.82	9.86	0.50	0.66	0.62	12.54
		high moderate	0.52	0.74	0.65	30.64	0.50	0.63	0.62	7.92
		high	0.50	0.66	0.64	1.87	0.50	0.60	0.61	6.69

Table 5.2: Performances of algorithms for large size data sets

$N = M$	$K = L$	σ	PA1				PA2			
			$\widehat{FRI}$	$\widehat{RI}^s$	$\widehat{RI}^o$	CPU	$\widehat{FRI}$	$\widehat{RI}^s$	$\widehat{RI}^o$	CPU
80	2	low	0.94	1.00	1.00	0.51	0.78	1.00	1.00	8.24
		moderate	0.78	1.00	1.00	0.73	0.64	0.95	0.95	10.62
		high moderate	0.62	1.00	1.00	4.94	0.56	0.94	0.94	10.88
		high	0.50	0.79	0.79	8.49	0.50	0.84	0.84	6.03
	3	low	0.93	1.00	1.00	1.33	0.58	0.83	0.72	17.18
		moderate	0.74	1.00	0.97	3.71	0.51	0.78	0.67	21.07
		high moderate	0.57	0.92	0.76	25.97	0.51	0.76	0.66	22.74
		high	0.50	0.74	0.66	2.97	0.50	0.70	0.64	13.23
	4	low	0.93	1.00	1.00	2.44	0.52	0.77	0.64	61.83
		moderate	0.73	1.00	0.94	10.00	0.51	0.72	0.63	60.63
		high moderate	0.53	0.78	0.65	48.38	0.50	0.67	0.63	45.16
		high	0.50	0.69	0.63	4.48	0.50	0.63	0.62	30.94
	5	low	0.93	1.00	0.99	3.38	0.51	0.69	0.63	91.04
		moderate	0.69	0.97	0.86	19.65	0.50	0.67	0.62	84.22
		high moderate	0.52	0.73	0.64	116.45	0.50	0.63	0.63	55.60
		high	0.50	0.66	0.63	4.15	0.50	0.61	0.62	43.09
150	2	low	0.94	1.00	1.00	4.64	0.78	1.00	1.00	63.42
		moderate	0.78	1.00	1.00	3.07	0.64	0.96	0.96	88.54
		high moderate	0.64	1.00	1.00	11.01	0.55	0.93	0.93	77.25
		high	0.50	0.78	0.77	49.26	0.50	0.83	0.83	48.52
	3	low	0.93	1.00	1.00	5.83	0.60	0.87	0.72	121.93
		moderate	0.74	1.00	0.96	17.47	0.52	0.79	0.67	191.69
		high moderate	0.58	0.92	0.78	41.00	0.51	0.75	0.66	135.61
		high	0.50	0.74	0.66	10.42	0.50	0.72	0.66	85.91
	4	low	0.93	1.00	1.00	9.14	0.51	0.75	0.63	308.65
		moderate	0.72	0.99	0.93	40.77	0.51	0.72	0.64	314.32
		high moderate	0.53	0.78	0.66	84.51	0.50	0.69	0.63	241.40
		high	0.50	0.71	0.65	12.93	0.50	0.66	0.62	136.73
	5	low	0.93	1.00	1.00	12.03	0.51	0.71	0.62	497.55
		moderate	0.65	0.95	0.80	119.54	0.50	0.66	0.63	376.35
		high moderate	0.51	0.70	0.64	177.31	0.50	0.64	0.63	235.01
		high	0.50	0.66	0.64	13.08	0.50	0.62	0.62	207.20

The performances of the algorithms under different parameter levels are summarized in Table 5.3. The number of entities does not alter the performance of the algorithms. However, as the number of entities increases, CPU times increase as well. As the number of clusters gets larger, the performance measures ($\widehat{FRI}$, $\widehat{RI}^s$, $\widehat{RI}^o$) slightly decreases for both algorithms. On the other hand, CPU times increase. The size of the error directly affects the difficulty of data sets. Performances of the algorithms decrease significantly for the larger size of error perturbations.

Interestingly, CPU times for the high error perturbed data sets are lower than the error level of moderate or high moderate. A high level of σ distorts the submatrix structures in the data sets, and so the submatrices get almost undefined (recall Figure 5.5). Thus, algorithms tend to assign equal probability to each entity for these data sets, and then the algorithms converge fast.

Table 5.3: Average performances for each level of each design feature

Design Feature	Level	**PA1**				**PA2**			
		$\widehat{FRI}$	$\widehat{RI}^s$	$\widehat{RI}^o$	CPU	$\widehat{FRI}$	$\widehat{RI}^s$	$\widehat{RI}^o$	CPU
Number of entities ($N = M$)	20	0.68	0.88	0.84	2.55	0.55	0.74	0.70	1.03
	50	0.68	0.90	0.84	5.50	0.54	0.75	0.71	5.43
	80	0.68	0.89	0.85	16.10	0.54	0.76	0.71	36.41
	150	0.68	0.89	0.84	38.25	0.54	0.77	0.72	195.63
Number of clusters ($K = L$)	2	0.71	0.95	0.95	6.20	0.63	0.91	0.91	20.10
	3	0.69	0.91	0.84	8.58	0.53	0.76	0.67	39.14
	4	0.67	0.86	0.80	15.10	0.51	0.70	0.63	76.88
	5	0.66	0.84	0.77	32.51	0.50	0.65	0.62	102.38
Size of error (σ)	low	0.93	1.00	1.00	3.03	0.61	0.82	0.75	74.96
	moderate	0.72	0.98	0.92	14.94	0.55	0.78	0.72	73.79
	high moderate	0.57	0.86	0.77	36.22	0.51	0.74	0.70	52.90
	high	0.50	0.72	0.68	8.21	0.50	0.68	0.66	36.85

Since **PA1** mostly outperforms **PA2**, we conduct our additional experiments for **PA1** and name it *"Two-Mode Probabilistic Distance Clustering"* (**TMPDC**) in the rest of the analysis.

5.4 Computational Results on Noisy Data Sets

We aim to evaluate the performance of **TMPDC** on noise-added data sets. Thus, we generate data sets $\mathbf{X}$ with a low level of error by our main generation algorithm explained in Section 5.1. Then, some noisy row and column entities are included to these data sets such that the new entities may belong to more than one cluster with some proportion. The steps of this data generation algorithm are given below.

Algorithm 4: Noise Included Data Generation Algorithm

Input: $\mathbf{X}$, $\mathbf{P}$, $\mathbf{Q}$, N, M, α

Step 1. Compute the row cluster centers $\{\mathbf{v}_k^r \mid k = 1, \ldots, K\}$ by averaging the row entities assigned to cluster k, $\{\mathbf{x}_i^r \mid p_{ik} = 1\}$.

Step 2. Generate αN many row entities by computing convex combinations of row cluster centers as
$\mathbf{x}_i^r = \sum_{k=1}^{K} \lambda_k \mathbf{v}_k^r, \quad i = 1, \ldots, \alpha N$, where $\sum_{k=1}^{K} \lambda_k = 1$ and each λ_k is selected randomly.

Step 3. Update $\mathbf{X}$ by adding new row entities to the original data set, and update $\mathbf{P}$ by adding λ_k values for each new row entity. Then set $N = N + \alpha N$.

Step 4. Compute the column cluster centers $\{\mathbf{v}_l^c \mid l = 1, \ldots, L\}$ by averaging the column entities assigned to cluster l, $\{\mathbf{x}_j^c \mid q_{jl} = 1\}$.

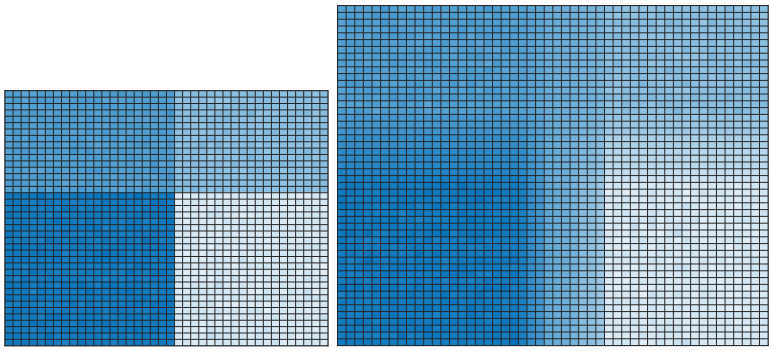
Step 5. Generate αM many column entities by computing convex combinations of column cluster centers as
$\mathbf{x}_j^c = \sum_{l=1}^{L} \lambda_l \mathbf{v}_l^c, \quad j = 1, \ldots, \alpha M$, where $\sum_{l=1}^{L} \lambda_l = 1$ and each λ_l is selected randomly.

Step 6. Update $\mathbf{X}$ by adding new column entities to the original data set, and update $\mathbf{Q}$ by adding λ_l values for each new column entity. Then set $M = M + \alpha M$.

Step 7. Shuffle row and columns of the data set $\mathbf{X}$.

Output: Updated $\mathbf{X}$, $\mathbf{P}$, $\mathbf{Q}$, N, M

Figure 5.8 represents our noise added data set generation technique. Assume that we want to create a data set as $N = M = 50$ with moderate level of noise ($\alpha = 0.2$) and $K = L = 2$. A 40×40 data set $\mathbf{X}$ is generated with first using main data generation algorithm (Algorithm 3) as in Figure 5.8a. Then 10 row and column entities are generated as explained in Steps 2-6 of Algorithm 4.



(a) $N = M = 40$ (b) $N = M = 50$

Figure 5.8: Representation of noise added data generation method, (a) the initial data set, (b) noise added data

The parameter α controls the complexity of the data set in terms of noise. We select three levels for α, which are

- Low noise $\alpha = 0.1$, moderate noise $\alpha = 0.2$, high noise $\alpha = 0.4$.

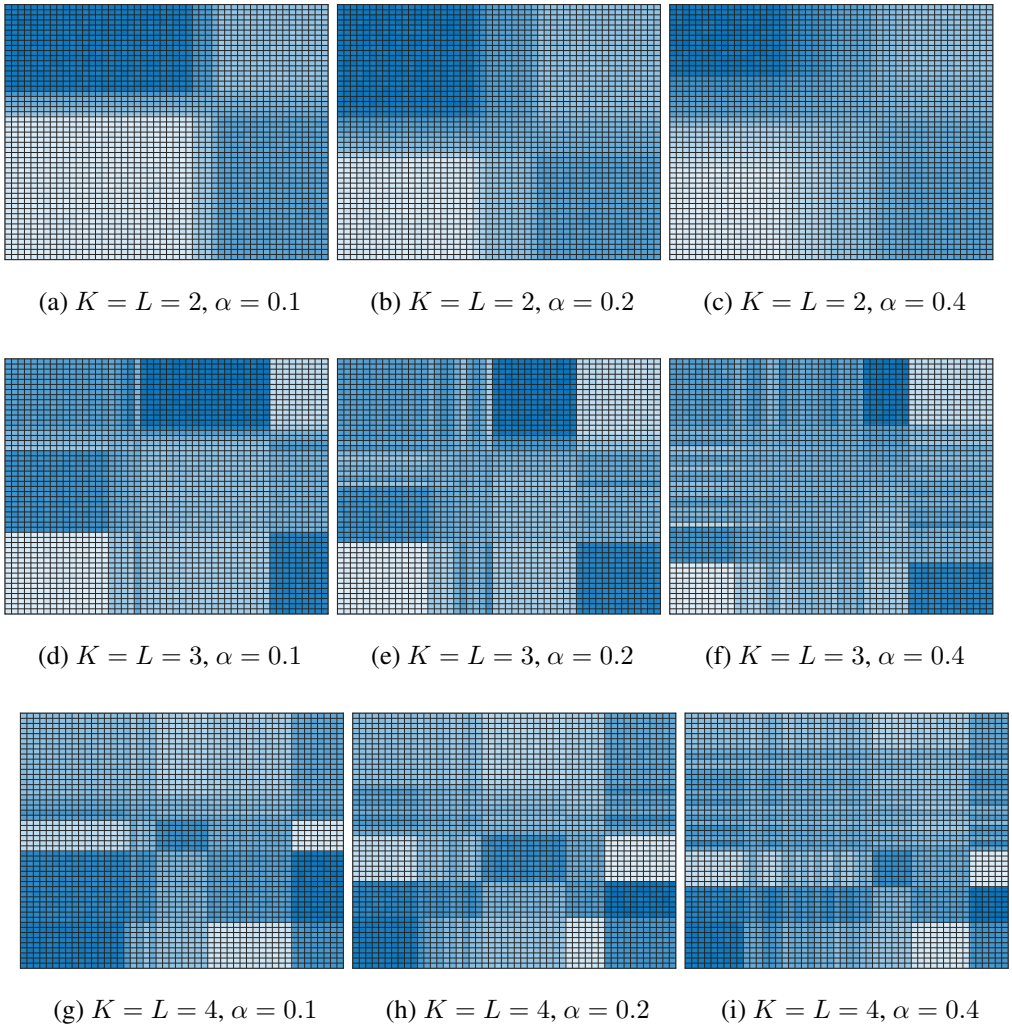
The effect of different α levels on the submatrices can be observed in Figure 5.9. As α increases, distortion of the submatrix structures becomes more severe.

The number of entities and the number of clusters control the difficulty in terms of computational time, and the following levels are used for experiments

- Small size data set: $N = M = 20$, medium size data set: $N = M = 50$, large size data set: $N = M = \{80, 150\}$.
- $K = L = 2$, $K = L = 3$, $K = L = 4$, and $K = L = 5$.

For computational experiments, we have 4 levels of number of entities (N,M), 4

levels of number of clusters (K,L), and 3 levels of noise that makes total of 48 combinations. For each parameter combination, 30 data sets are generated. As a result, 1440 data sets are created for this experimental study.



(a) $K = L = 2, \alpha = 0.1$ (b) $K = L = 2, \alpha = 0.2$ (c) $K = L = 2, \alpha = 0.4$

(d) $K = L = 3, \alpha = 0.1$ (e) $K = L = 3, \alpha = 0.2$ (f) $K = L = 3, \alpha = 0.4$

(g) $K = L = 4, \alpha = 0.1$ (h) $K = L = 4, \alpha = 0.2$ (i) $K = L = 4, \alpha = 0.4$

Figure 5.9: Representation of low, moderate, and high α levels for different K, L levels

5.4.1 Results

Average performances out of 30 trials for each parameter combination are reported in Table 5.4-Table 5.5. Note that in this setting, membership probabilities **P** and **Q** have some probabilistic assignments (λ_k and λ_l values of noise entities are in-

cluded as probabilities). Thus, while computing $\widehat{RI}^s$ index, we use strict hardening for reference clusters (**P** and **Q**). Similarly, to estimate $\widehat{RI}^o$, overlapping hardening is employed for reference clusters.

Table 5.4: Performance of **TMPDC** on the noise added small and medium size data sets

$N = M$	$K = L$	α	$\widehat{FRI}$	$\widehat{RI}^s$	$\widehat{RI}^o$	CPU
20	2	0.1	0.84	1.00	1.00	0.24
		0.2	0.77	0.98	0.98	0.34
		0.4	0.65	0.91	0.91	0.65
	3	0.1	0.82	0.94	0.87	0.55
		0.2	0.73	0.89	0.79	0.68
		0.4	0.59	0.74	0.66	1.03
	4	0.1	0.79	0.89	0.81	0.79
		0.2	0.71	0.85	0.73	0.85
		0.4	0.56	0.68	0.58	0.73
	5	0.1	0.79	0.88	0.80	0.88
		0.2	0.68	0.79	0.69	0.63
		0.4	0.55	0.65	0.56	0.50
50	2	0.1	0.84	1.00	1.00	0.43
		0.2	0.77	1.00	1.00	0.67
		0.4	0.67	0.94	0.94	1.68
	3	0.1	0.83	0.95	0.88	0.96
		0.2	0.75	0.89	0.78	2.29
		0.4	0.60	0.76	0.65	4.53
	4	0.1	0.82	0.93	0.83	2.71
		0.2	0.71	0.86	0.73	4.10
		0.4	0.56	0.77	0.59	4.03
	5	0.1	0.79	0.93	0.81	2.38
		0.2	0.68	0.87	0.70	2.99
		0.4	0.56	0.80	0.58	3.05

Table 5.5: Performance of **TMPDC** on the noise added large size data sets

$N = M$	$K = L$	α	$\widehat{FRI}$	$\widehat{RI}^s$	$\widehat{RI}^o$	CPU
80	2	0.1	0.84	1.00	1.00	0.77
		0.2	0.77	1.00	1.00	1.27
		0.4	0.66	0.97	0.97	3.46
	3	0.1	0.83	0.94	0.86	2.03
		0.2	0.75	0.89	0.78	3.73
		0.4	0.59	0.74	0.65	8.28
	4	0.1	0.83	0.93	0.84	6.63
		0.2	0.72	0.86	0.73	17.54
		0.4	0.56	0.78	0.59	18.11
	5	0.1	0.81	0.93	0.81	12.76
		0.2	0.66	0.87	0.67	14.54
		0.4	0.53	0.79	0.56	19.69
150	2	0.1	0.84	1.00	1.00	3.67
		0.2	0.77	1.00	1.00	5.82
		0.4	0.67	1.00	1.00	15.04
	3	0.1	0.83	0.95	0.87	6.97
		0.2	0.75	0.90	0.78	12.74
		0.4	0.60	0.75	0.66	51.68
	4	0.1	0.84	0.94	0.84	21.86
		0.2	0.72	0.87	0.73	53.97
		0.4	0.55	0.77	0.59	76.21
	5	0.1	0.83	0.93	0.83	68.28
		0.2	0.66	0.86	0.67	137.36
		0.4	0.53	0.78	0.56	113.54

The overall performance of **TMPDC** for each parameter combination is summarized in Table 5.6. As the size of the data set increases, the performance measures do not alter, but CPU times increase. Increasing the number of partitions results in decreased $\widehat{FRI}$ and $\widehat{RI}$ values and increased CPU times. Compared to computational exper-

iments (see Table 5.3), the number of partitions affects the algorithm performance slightly more under noise. Consider the case of $K = 2$, and a noise entity is added to row entities, then it will be obtained by a convex combination of two cluster centers. On the other hand, if it is added to a data set with $K = 5$, the noise entity will be obtained from a convex combination of five cluster centers. That is why the effect of the number of partitions on the performance is more significant. As the level of noise increases, submatrix structures of data sets get disperse. Thus, the performance of the algorithm decreases for a larger level of α. In this case, data becomes harder to analyze; therefore, convergence takes more time. Thus, CPU times are higher for greater levels of α.

Table 5.6: Average performance of **TMPDC** for each level of each design feature

Design Feature	Level	$\widehat{FRI}$	$\widehat{RI}^s$	$\widehat{RI}^o$	CPU
Number of entities ($N = M$)	20	0.71	0.85	0.78	0.66
	50	0.72	0.89	0.79	2.48
	80	0.71	0.89	0.79	9.07
	150	0.72	0.90	0.79	47.26
Number of clusters ($K = L$)	2	0.76	0.98	0.98	2.84
	3	0.72	0.86	0.77	7.96
	4	0.70	0.84	0.72	17.29
	5	0.67	0.84	0.69	31.38
Level of noise (α)	0.1	0.82	0.95	0.88	8.24
	0.2	0.72	0.90	0.80	16.22
	0.4	0.59	0.80	0.69	20.14

5.5 Computational Results on Binary Data Sets

Up to this section, we discuss data sets with continuous elements. Data sets with binary elements are input for various problems in the literature, such as social network analysis or part-machine grouping technology. Therefore, in this section, we generate binary data sets and evaluate the performance of **TMPDC**.

The steps of the binary data set generation algorithm are given below.

Algorithm 5: Binary Data Set Generation Algorithm

Input: N, M, K, L, β

Step 1. Generate membership $\{p_{ik} \in \mathbf{P} \mid p_{ik} = \{0, 1\}\}$ for each row entity i to a row cluster $\{k \mid k = 1, .., K\}$ with discrete uniform distribution.

Step 2. Generate membership $\{q_{jl} \in \mathbf{Q} \mid q_{jl} = \{0, 1\}\}$ for each column entity j to a column cluster $\{l \mid l = 1, .., L\}$ with discrete uniform distribution.

Step 3. Compute $\mathbf{X} = \mathbf{PQ}'$.

Step 4. Count the number of ones in $\mathbf{X}$, denoted by C.

Step 5. Change βC many ones of $\mathbf{X}$ to zero, and βC many zeros to one, where $\beta \in [0, 1]$.

Output: **X**, **P**, **Q**

The parameter β controls the percentage of ones (zeros) converted to zeros (ones). When $\beta = 0$, it means that the zero and one blocks (submatrices) are clearly defined in the data set. When β increases, the data set gets more challenging to analyze. We set three levels for β, which are

- Low $\beta = 0.1$, moderate $\beta = 0.2$, high $\beta = 0.3$.

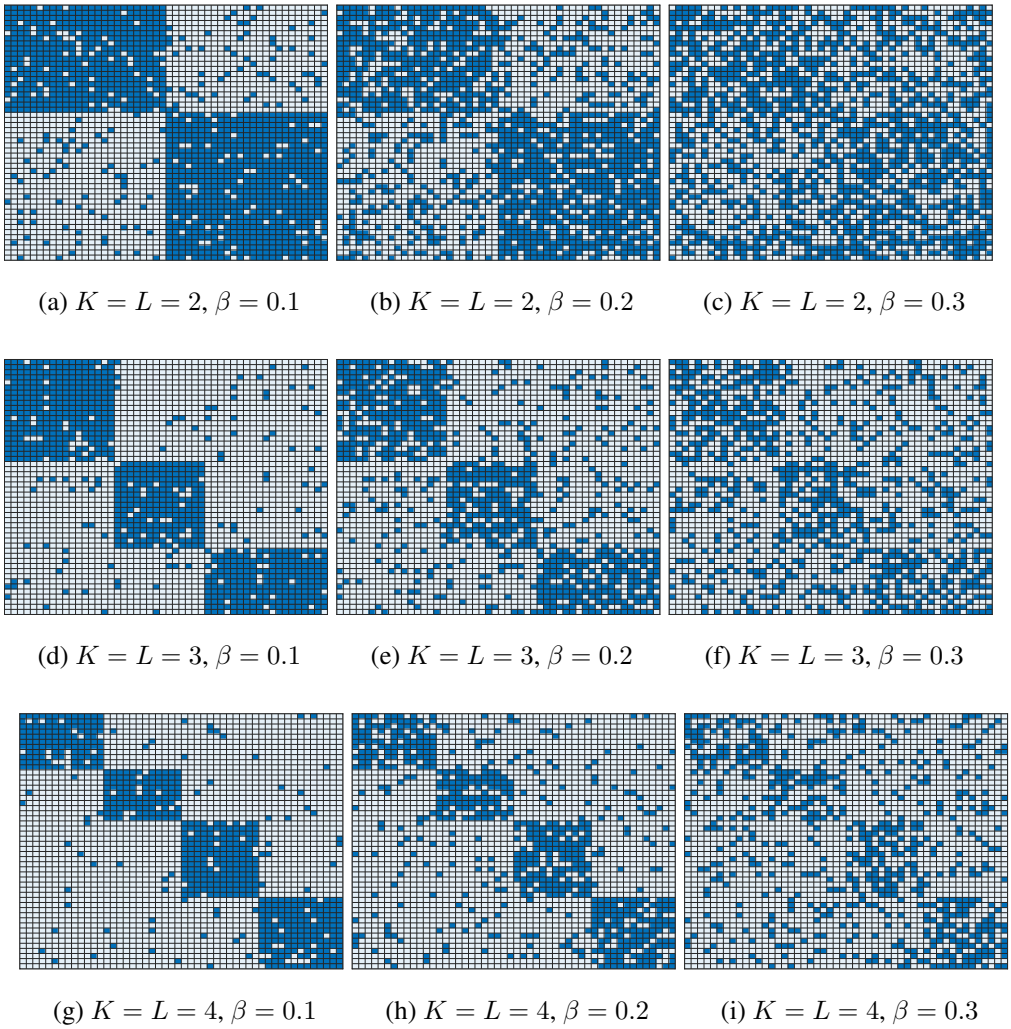
Figure 5.10 shows the effect of different β levels on the submatrices. As β increases, submatrix structures get uncertain.

Due to the nature of binary data sets, as number of partitions increases the sizes of the submatrices consisting of ones decreases. Therefore, we use larger data sets for these experiments. The following levels are set for binary data experiments:

- $N = M = 50$, $N = M = 80$, $N = M = 150$, and $N = M = 300$.

- $K = L = 2$, $K = L = 3$, $K = L = 4$, and $K = L = 5$.

In the total of 48 different parameter combinations are tried for data set generations (4 levels for the number of entities (N, M), 4 levels for the number of clusters (K, L), and 3 levels of β). For each combination, 30 different data sets are generated.



(a) $K = L = 2$, $\beta = 0.1$ (b) $K = L = 2$, $\beta = 0.2$ (c) $K = L = 2$, $\beta = 0.3$

(d) $K = L = 3$, $\beta = 0.1$ (e) $K = L = 3$, $\beta = 0.2$ (f) $K = L = 3$, $\beta = 0.3$

(g) $K = L = 4$, $\beta = 0.1$ (h) $K = L = 4$, $\beta = 0.2$ (i) $K = L = 4$, $\beta = 0.3$

Figure 5.10: Representation of β levels for different K, L levels

5.5.1 Results

The performance of **TMPDC** on the binary data sets is given in Table 5.7- Table 5.8. The overall performance of **TMPDC** is summarized in Table 5.9. It can be observed that as the size of the data set increases, the CPU time increases as well. In

contrast to previous experiments, the number of clusters affects the performance of the algorithm. This is due to the nature of binary data sets. As the number of row and column clusters increases, the sparsity in data becomes severe, which directly affects the performance. For $K = L = 2$, even with the largest level of β, **TMPDC** performs well in terms of $\widehat{RI}^s$. On the contrary, the results are not promising for $K = L = 5$. As expected, for the higher levels of β the performance of **TMPDC** deteriorates.

Table 5.7: Performance of **TMPDC** on the binary data sets for $N = M = \{50, 80\}$

$N = M$	$K = L$	β	$\widehat{FRI}$	$\widehat{RI}^s$	$\widehat{RI}^o$	CPU
50	2	0.1	0.70	1.00	1.00	0.75
		0.2	0.56	1.00	1.00	3.42
		0.3	0.50	0.99	0.99	0.51
	3	0.1	0.70	1.00	0.98	1.45
		0.2	0.51	0.79	0.69	9.54
		0.3	0.50	0.79	0.69	0.90
	4	0.1	0.67	0.97	0.92	5.06
		0.2	0.50	0.67	0.63	5.26
		0.3	0.50	0.71	0.63	0.99
	5	0.1	0.59	0.88	0.73	6.52
		0.2	0.50	0.62	0.62	2.38
		0.3	0.50	0.66	0.61	1.12
80	2	0.1	0.70	1.00	1.00	0.80
		0.2	0.55	1.00	1.00	6.56
		0.3	0.50	1.00	1.00	0.92
	3	0.1	0.70	1.00	1.00	2.69
		0.2	0.50	0.79	0.68	21.65
		0.3	0.50	0.78	0.69	1.50
	4	0.1	0.68	0.99	0.95	18.93
		0.2	0.50	0.67	0.62	6.08
		0.3	0.50	0.70	0.61	2.34
	5	0.1	0.54	0.79	0.68	13.69
		0.2	0.50	0.64	0.61	4.12
		0.3	0.50	0.68	0.60	2.47

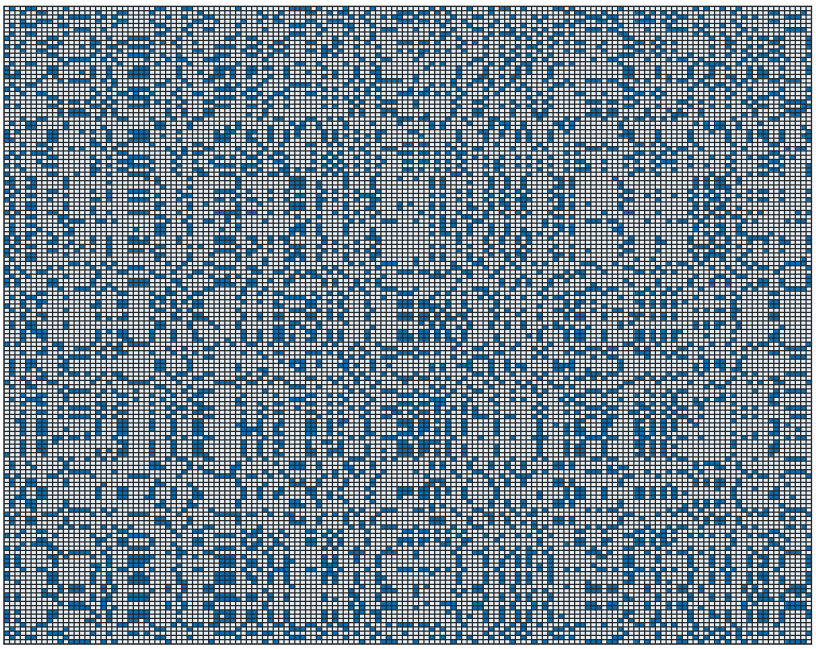
Table 5.8: Performance of **TMPDC** on the binary data sets for $N = M = \{150, 300\}$

$N = M$	$K = L$	β	$\widehat{FRI}$	$\widehat{RI}^s$	$\widehat{RI}^o$	CPU
150	2	0.1	0.70	1.00	1.00	3.85
		0.2	0.55	1.00	1.00	34.51
		0.3	0.50	1.00	1.00	4.24
	3	0.1	0.70	1.00	1.00	11.59
		0.2	0.50	0.77	0.68	33.10
		0.3	0.50	0.80	0.68	6.25
	4	0.1	0.62	0.91	0.84	73.47
		0.2	0.50	0.70	0.62	13.72
		0.3	0.50	0.74	0.60	6.76
	5	0.1	0.51	0.65	0.61	34.69
		0.2	0.50	0.66	0.59	11.93
		0.3	0.50	0.71	0.59	7.88
300	2	0.1	0.70	1.00	1.00	14.50
		0.2	0.56	1.00	1.00	110.38
		0.3	0.50	1.00	1.00	14.99
	3	0.1	0.70	1.00	1.00	37.43
		0.2	0.50	0.77	0.68	69.13
		0.3	0.50	0.80	0.68	21.86
	4	0.1	0.55	0.75	0.71	220.34
		0.2	0.50	0.68	0.61	40.82
		0.3	0.50	0.76	0.60	22.62
	5	0.1	0.50	0.62	0.59	74.47
		0.2	0.50	0.67	0.57	38.26
		0.3	0.50	0.75	0.56	25.24

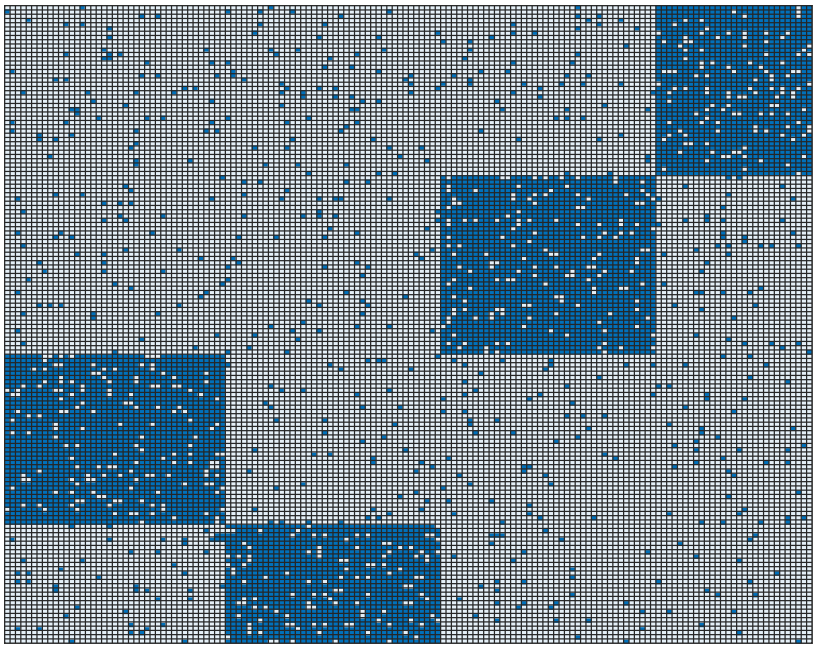
Table 5.9: Average performance of **TMPDC** for binary data sets

Design Feature	Level	$\widehat{FRI}$	$\widehat{RI}^s$	$\widehat{RI}^o$	CPU
Number of entities ($N = M$)	50	0.56	0.84	0.79	3.16
	80	0.56	0.84	0.79	6.81
	150	0.55	0.83	0.77	20.17
	300	0.54	0.82	0.75	57.50
Number of clusters ($K = L$)	2	0.59	1.00	1.00	16.28
	3	0.57	0.86	0.79	18.09
	4	0.54	0.77	0.70	34.70
	5	0.51	0.69	0.61	18.56
Level of β	0.1	0.64	0.91	0.88	32.51
	0.2	0.51	0.78	0.73	25.68
	0.3	0.50	0.81	0.72	7.54

Figure 5.11 presents a large size data set studied in this section and its solution given by **TMPDC**.



(a) Original data



(b) **TMPDC** solution

Figure 5.11: **TMPDC** solution for an $N = M = 150$ data set for $K = L = 4$

5.6 How Can a Decision-Maker Benefit From Two-Mode Clustering with Soft Assignments?

In this section, we explain how to use the results of **TMPDC**. In this section, we explain how to use the results of TMPDC. We examined two examples: a synthetic data set with continuous elements and a binary data set that stores part-machine incidences.

Patient-Symptom Data Set

Assume that a decision-maker has 14 patients who suffer from two distinct diseases. However, we do not know which patients have which of these diseases. Patients are examined for 15 symptoms that are associated with severity values between 0 and 7. Given the patient i and symptom j the severity value of x_{ij} takes 0 if the patient does not show this symptom. When the x_{ij} is 7, that means the patient i exhibits symptom j intensely. The resulting data set is given in Figure 5.12, where darker colors refer to higher severity.

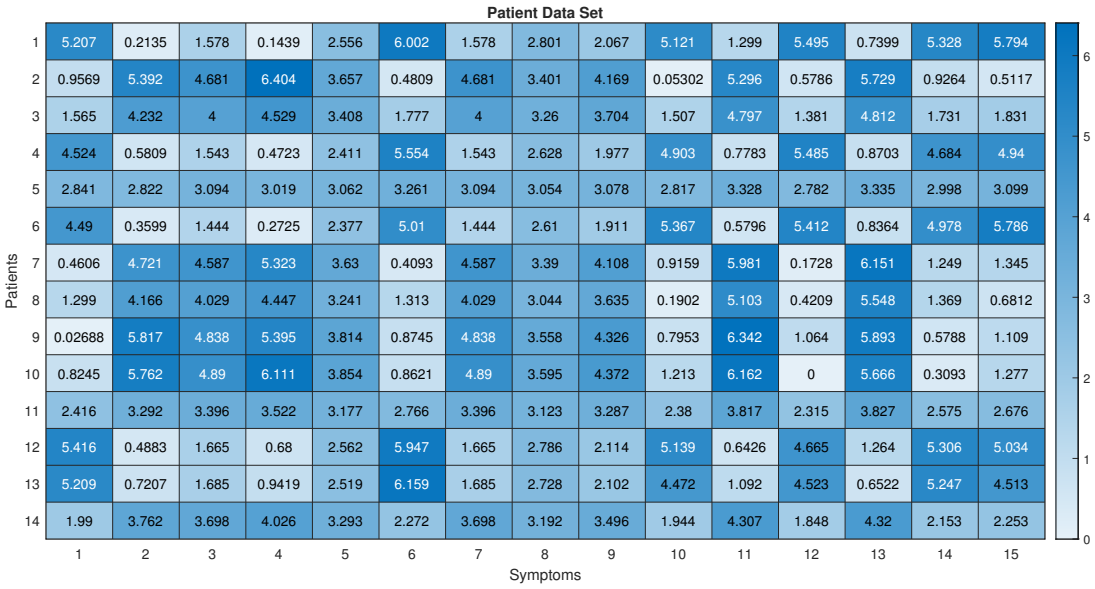


Figure 5.12: Severity values of symptoms for each patient.

In such an environment, the decision-maker wants to answer the following questions:

- Which groups of patients have the same disease?

- Which two groups of symptoms are intensely observed in which group of patients? Can we identify the symptoms that can best distinguish these two diseases?

- Are there any patients that can have both diseases?

- Are there any symptoms that are not explanatory for these two diseases?

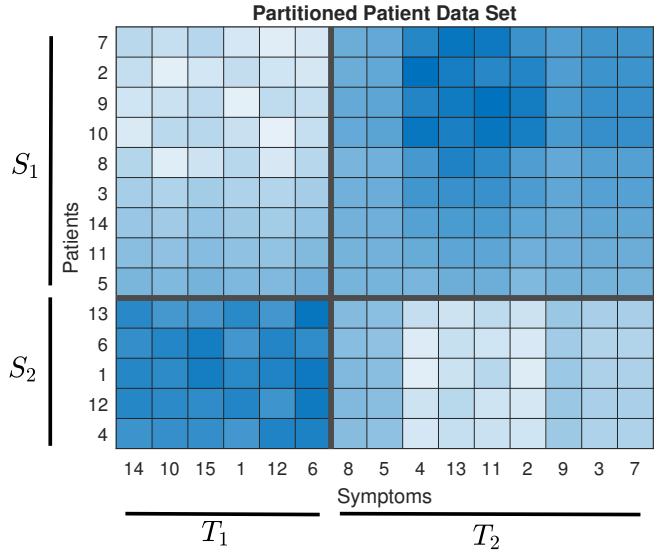
To summarize, the decision-maker aims to discover two groups of patients and two groups of symptoms simultaneously such that one of the symptom groups is fiercely perceived in one of the patient groups and weakly in the other one. Therefore, this is a two-mode clustering problem with $N = 14$, $M = 15$, $K = L = 2$.

We use **TMPDC** as a tool to solve the problem. The solution of **TMPDC** is hardened to get the partitions. The hard assignments are obtained by concerning probability thresholds denoted as τ^r and τ^c for row and column membership probabilities. In other words, we make the hard assignments by following rules:

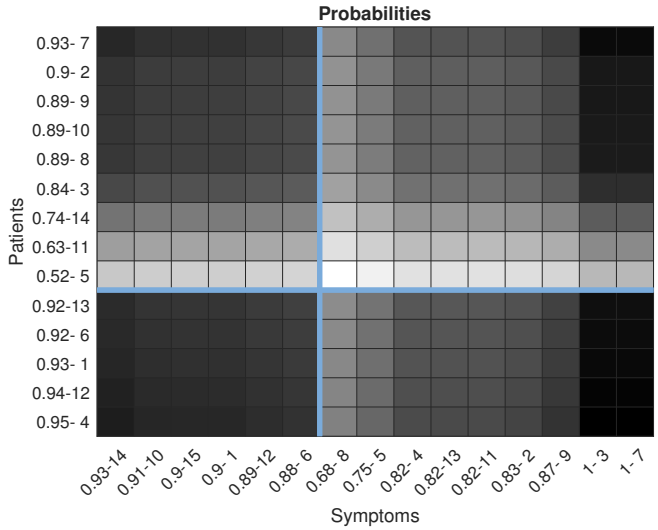
- The patient i is a member of patient group k, if its membership probability $p_k(\mathbf{x}_i^r)$ is greater than the threshold τ^r.

- The symptom j is a member of symptom group l, if its membership probability $q_l(\mathbf{x}_j^c)$ is greater than the threshold τ^c.

If the decision-maker desires to see the assignments of all patients and symptoms, then $\tau^r = \tau^c = 0.5$. From a clustering perspective, partitions are to be obtained such that they cover the whole set of modes. According to these threshold values, the results are given in Figure 5.13. In panel (a), the indexes of rows and columns refer to the patient and symptom numbers, respectively. Panel (b) is to show soft assignment probabilities. The membership probabilities are provided next to the entity number in the rows and columns for each patient and symptom.



(a)



(b)

Figure 5.13: (a) **TMPDC** solution of the patient data set for $\tau^r = \tau^c = 0.5$, where partitions are represented by black lines, (b) membership probabilities of patients, and symptoms

The resulting partitions in Figure 5.13a indicate:

- The first group of patients are $S_1 = \{7, 2, 9, 10, 8, 3, 14, 11, 5\}$ and suffer from disease 1, whereas the second group of patients $S_2 = \{13, 6, 1, 12, 4\}$ suffer

from the disease 2.

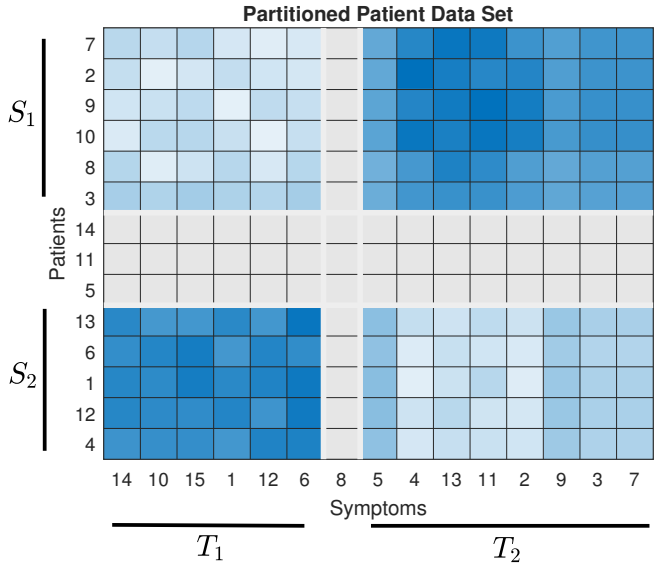
- The first and second group of symptoms are $T_1 = \{14, 10, 15, 1, 12, 6\}$ and $T_2 = \{8, 5, 4, 13, 11, 2, 9, 3, 7\}$, respectively.
- Assigned elements of S_k and T_l form submatrices $\mathbf{V}_{kl}$. For instance, $\mathbf{V}_{11}$ contains elements x_{ij} such that $i \in S_1$ and $j \in T_1$.
- The submatrix centers v_{kl} measures the strengths of the relationship between patient and symptom groups and is calculated as the mean of submatrices. Thus, for the first disease, symptom group T_1 has a low value of severity $v_{11} = 1.434$, while the symptom group T_2 has a high value of severity as $v_{12} = 4.229$. Similarly, patients of second disease show symptoms T_1 intensely $v_{21} = 5.190$, whereas they show symptoms T_2 weakly $v_{22} = 1.458$.

Additional information can be provided based on membership probabilities. In Figure 5.13b, the membership degrees of patients and symptoms to their groups are provided. The elements are the joint membership probabilities $R_{kl}(x_{ij}) = p_k(\mathbf{x}_i^r)q_l(\mathbf{x}_j^c)$. Therefore, the following information can be derived:

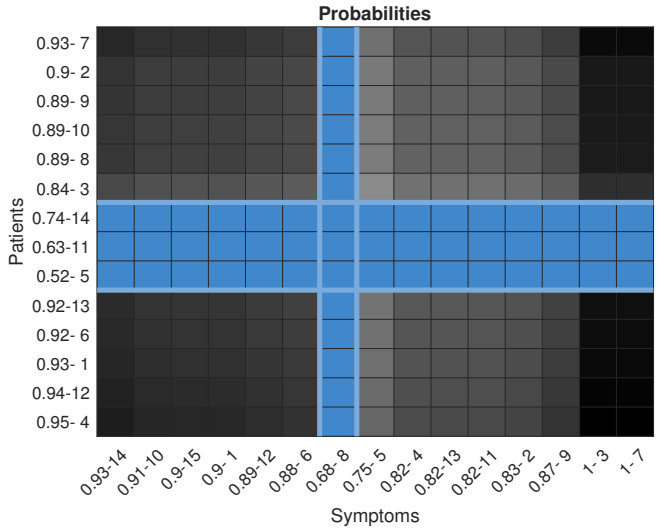
- Patient 7 is a member of S_1 with the highest probability. The membership probability of patient 5 is too low that we can conclude she can have both diseases.
- Membership probabilities of all patients in S_2 are significantly high. Thus, it can be referenced that S_2 is a more reliable group relative to S_1.
- Symptom 14 is the most crucial symptom of classifying disease 2, and symptoms 3 and 7 are critical for diagnosing disease 1.
- Symptoms 5 and 8 can be observed in both diseases relative to other symptoms. Thus, they may not be explanatory enough.

The decision-maker may be more risk-averse and want to have a more convenient solution, which can be achieved by changing threshold values. For instance, she can regard the assignments of patients having probability higher than 0.8 and symptoms having probability higher than 0.7. The resulting assignments for $\tau^r = 0.8$ and $\tau^c =$

0.7 are given in Figure 5.14. In panel (a), the light gray regions refer to the undecided patients or symptoms. The corresponding probabilities are shown by blue regions in panel (b).



(a)



(b)

Figure 5.14: (a) **TMPDC** solution of the patient data set for $\tau^r = 0.8$ and $\tau^c = 0.7$, (b) membership probabilities of patients and symptoms

According to new partitions in 5.14a:

- Patients 14, 11 and 5 are excluded from the analysis. Thus, the first group of patients $S_1 = \{7, 2, 9, 10, 8, 3\}$ and the second one is $S_2 = \{13, 6, 1, 12, 4\}$.
- Symptom 8 is not explanatory for new threshold. Therefore, the first symptom set is $T_1 = \{14, 10, 15, 1, 12, 6\}$ and the second one is $T_2 = \{5, 4, 13, 11, 2, 9, 3, 7\}$.
- The submatrix centers v_{kl} is updated as $v_{11} = 0.891$, $v_{12} = 4.786$, $v_{21} = 5.190$ and $v_{22} = 1.301$. This update is important to evaluate new patients. For instance, with the first partition (Figure 5.13a), we expect a patient who suffers from disease 1 will have symptom severity of around 1.434 over symptom group T_1. With new assignments (Figure 5.14a), we expect them to have a severity value of approximately 0.891 for the same symptom group.

Note that even more aggressive thresholds can be considered, i.e., $\tau^r = \tau^c = 0.9$. However, this will decrease the sample size of patients and symptoms. Therefore, it is the expert's decision between analysis level and sample size.

Part-Machine Grouping Technology

Group technology (GT) is a method used to obtain a cellular manufacturing layout, which is an alternative for a job shop layout. It aims to enhance production efficiency by grouping machines and parts according to process requirements [43]. The grouped parts that have similar processing needs are called *part families*. The machines that can process specific part families are named *machine cells*. In a cellular layout, machines of a cell are located closer to each other to decrease the cost and traffic intensity of production logistic activities of a manufacturing environment.

The input of the GT problem is a part-machine incidence matrix that represents the machine requirements of parts. If a part i is processed by a machine j, the element x_{ij} of the incidence matrix takes a value of one; otherwise, it takes zero.

As an example, we study a part-machine incidence matrix of Chan and Milner [44] given in Figure 5.15. In this example, there are 10 parts as first mode entities and

15 machines as second mode entities ($N = 10$, $M = 15$). The number of machine cells and part families is three ($K = L = 3$). In Figure 5.15, the blue cells represent ones, and the light blue cells show the zeros. This instance was solved by an exact algorithm in [32] for the hard partitioning case. Thus, we have an exact solution to compare our results.

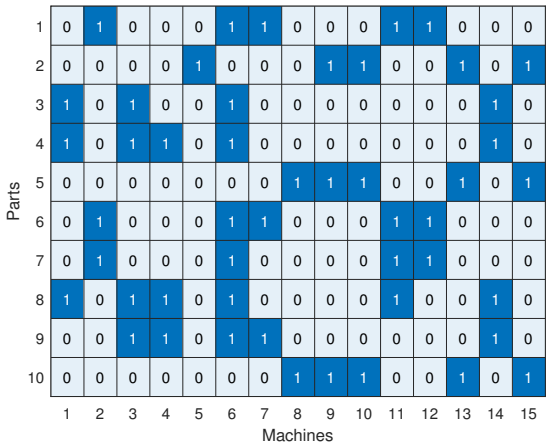


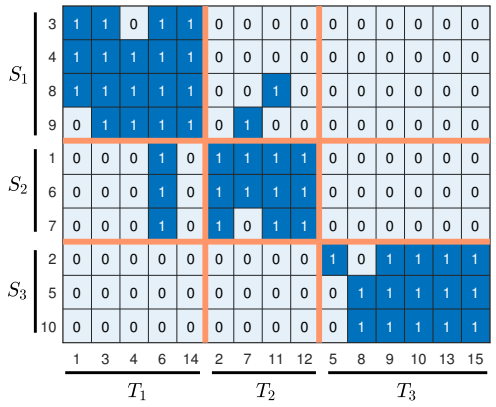
Figure 5.15: Part-machine incidence matrix

We employ **TMPDC** to find embedded submatrices of the part-machine incidence matrix. Figure 5.16b shows resulting assignment probabilities $\mathbf{P}_r$ and $\mathbf{Q}_r$. In this figure, the first values on axes are the highest assignment probabilities of entities to their respective clusters. The second value shows the entity number. For example, entity 2 is assigned to the first row cluster with a membership probability of 0.53. Note that we sorted the assigned entities according to membership probabilities in descending order for each cluster. The elements of the matrix in Figure 5.16b are joint membership probabilities of part-machine pairs to their assigned submatrices ($R_{kl}(x_{ij}) = p_k(\mathbf{x}_i^r)q_l(\mathbf{x}_j^c)$). For instance, part 5 and machine 3 pair is assigned to submatrix $\mathbf{V}_{11}$ with a probability of 1. Thus, darker cells imply higher joint membership probabilities.

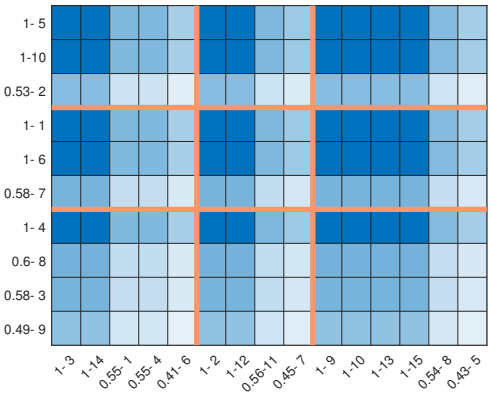
According to memberships, part families and machine cells are obtained with strict hardening, as in Figure 5.16c. The resulting submatrices are precisely compatible with the exact hard partitioning solution in Figure 5.16a. According to the solution, the part families S_1, S_2, and S_3 are mainly processed in machine cells T_3, T_2, and T_1, respectively. Therefore, it is reasonable to locate machines in a specific cell closer to

each other.

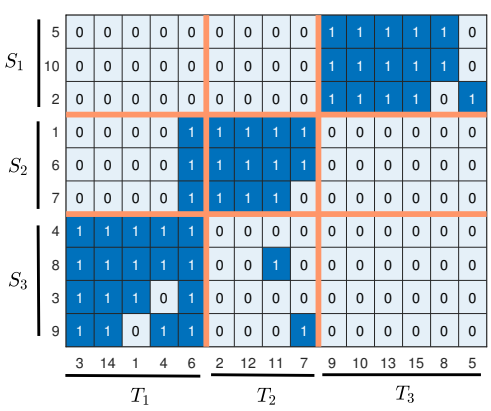
The resulting membership probabilities can also be interpreted with overlapping hardening as in Figure 5.16d. In this case, an entity may belong to more than one cluster. We use cut points $\tau^r = 1/K$ for row entities and $\tau^c = 1/L$ for column entities. The resulting assignments indicate that machine 6 belongs to both machine cells T_1 and T_2. Thus, it will be a logical action to locate cells T_1 and T_2 closer. Moreover, machine 6 should be located in T_1 at a position that is closest to cell T_2. It can be concluded that unlike the well-known methods for GT, our soft assignment approach can determine not only the machine cells but also the relative locations of cells.



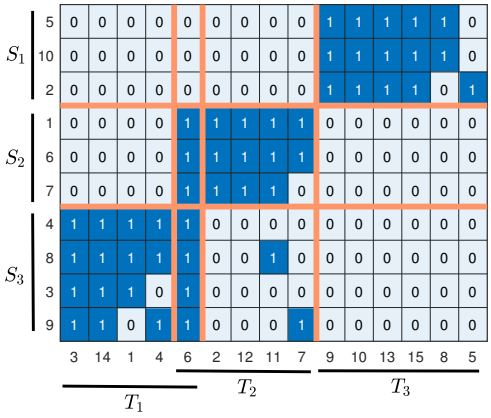
(a) Exact hard partitioning solution in [32]



(b) **TMPDC** assignment probabilities



(c) Solution under strict hardening



(d) Solution under overlapping hardening

Figure 5.16: Results for GT example

CHAPTER 6

CONCLUSION

In this thesis, we study soft clustering techniques. Our focus is Probabilistic Distance Clustering (PDC), a soft clustering algorithm for one-mode data sets. The study is constructed with two phases.

In the first phase, PDC principles are examined for one-mode problems by showing that these principles explain soft approaches in different literature concepts. From marketing, we studied the Huff Model, which finds given the facility locations, visiting probabilities of customers. In clustering perspective, it aims to find soft assignments to given cluster centers. We show that the probability definition of the Huff Model follows PDC principles. Then, we discuss two famous soft clustering approaches K-Harmonic Means Clustering (KHM) and Fuzzy c-Means Algorithm (FCM). We prove that PDC is a generalization of KHM and FCM. Moreover, we criticize the probability definition of KHM. Later, we discuss Gravity p-Median Problem (GPM) from location literature and prove that the objective of GPM is not compatible with its probability representation. We revise the objective function of GPM. This is an essential contribution to literature. By revised GPM, the assumption that hub locations will be on nodes of the network can be relaxed. Thus, we pave the way for the development of problem-specific algorithms.

In the second phase, we focus on Two-Mode Clustering (TMC) problems with soft assignments. To the best of our knowledge, the only soft assignment approach developed for TMC is Two-Mode Fuzzy C-Means. However, the algorithm's performance with soft memberships has not been discussed in the literature. It is used as an interstep to a well-known hard assignment algorithm, the Two-Mode KL-Means Algorithm. We derive the theoretical properties of TMC with soft assignments and

define two-mode PDC principles. By following those principles, we construct optimization models for the problem and propose two novel algorithms. Since this is a pioneer study for soft TMC, evaluating the performance of proposed approaches is a new research question. A full factorial design is adopted for the experimental study. Since the performance of clustering techniques is affected by the properties of data sets, we use generated data sets for our experiments. We modify Rosmalen's data generation method [35] and develop three two-mode data generation algorithms. We believe that these algorithms will be a benchmark in TMC literature for both hard and soft assignment approaches. We include various properties to our synthetic data sets by setting levels to data generation parameters. Moreover, external performance measures are proposed to test soft TMC algorithms. An extensive experimental study is conducted with a total of 4800 generated data sets. Lastly, we show that a soft TMC approach yields extra valuable information to decision-makers relative to hard TMC methods by employing our proposed algorithm, **TMPDC**, on two two-mode clustering problems.

To sum up, this study answers the following research questions

- Why PDC principles are so explanatory?
- How can a soft two-mode clustering problem be formulated and solved?
- Why it is needed to a soft assignment approach for two-mode clustering problems? How can a decision-maker benefit from a soft TMC solution?

As a future research direction, it will be a worthwhile effort to develop new solution methods for the revised Gravity p-Median problem. Moreover, for soft TMC, generalized versions of **TMPDC** can be obtained by using exponents of membership probabilities and distance as in one-mode PDC principles. This effort will yield additional parameters to the problem. The effect of new parameters on solution quality can be examined.

.

REFERENCES

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[2] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[3] C. K. Reddy and B. Vinzamuri, "A survey of partitional and hierarchical clustering algorithms," in *Data clustering*, pp. 87–110, Chapman and Hall/CRC, 2018.

[4] V. Mehta, S. Bawa, and J. Singh, "Analytical review of clustering techniques and proximity measures," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5995–6023, 2020.

[5] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[6] L. RDUSSEEUN and P. KAUFMAN, "Clustering by means of medoids," 1987.

[7] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, pp. 191–203, Jan. 1984.

[8] B. Zhang, M. Hsu, and U. Dayal, "K-harmonic means -a spatial clustering algorithm with boosting," in *Temporal, Spatial, and Spatio-Temporal Data Mining*, pp. 31–45, Springer Berlin Heidelberg, 2001.

[9] A. Ben-Israel and C. Iyigun, "Probabilistic d-clustering," *Journal of Classification*, vol. 25, pp. 5–26, June 2008.

[10] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the emalgorithm," *Journal of the Royal statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[11] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, "Epidemic spreading in real networks: An eigenvalue viewpoint," in *22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings.*, pp. 25–34, IEEE, 2003.

[12] S. L. Hakimi, "Optimum distribution of switching centers in a communication network and some related graph theoretic problems," *Operations research*, vol. 13, no. 3, pp. 462–475, 1965.

[13] T. Drezner and Z. Drezner, "The gravity p-median model," *European Journal of Operational Research*, vol. 179, pp. 1239–1251, June 2007.

[14] D. L. Huff, "Defining and estimating a trading area," *Journal of Marketing*, vol. 28, pp. 34–38, July 1964.

[15] M. D. Beule, D. V. den Poel, and N. V. de Weghe, "An extended huff-model for robustly benchmarking and predicting retail network performance," *Applied Geography*, vol. 46, pp. 80–89, Jan. 2014.

[16] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, June 1965.

[17] B. Zhang, "Generalized k-harmonic means – boosting in unsupervised learning," tech. rep., 2000.

[18] M. Teboulle, "A unified continuous optimization framework for center-based clustering methods," *Journal of Machine Learning Research*, vol. 8, no. 3, pp. 65–102, 2007.

[19] C. Iyigun and A. Ben-Israel, "PROBABILISTIC DISTANCE CLUSTERING ADJUSTED FOR CLUSTER SIZE," *Probability in the Engineering and Informational Sciences*, vol. 22, pp. 603–621, Sept. 2008.

[20] J. D. Carroll and P. Arabie, "Multidimensional scaling," *Annual Review of Psychology*, vol. 31, pp. 607–649, Jan. 1980.

[21] T. Kong, K. Seong, K. Song, and K. Lee, "Two-mode modularity clustering of parts and activities for cell formation problems," *Computers & Operations Research*, vol. 100, pp. 77–88, Dec. 2018.

[22] W. S. Desarbo, "Gennclus: New models for general nonhierarchical clustering analysis," *Psychometrika*, vol. 47, pp. 449–475, Dec. 1982.

[23] B. Wang, Y. Miao, H. Zhao, J. Jin, and Y. Chen, "A biclustering-based method for market segmentation using customer pain points," *Engineering Applications of Artificial Intelligence*, vol. 47, pp. 101–109, Jan. 2016.

[24] S. P. Borgatti and M. G. Everett, "Network analysis of 2-mode data," *Social Networks*, vol. 19, pp. 243–269, Aug. 1997.

[25] S. P. Borgatti, "Social network analysis, two-mode concepts in," in *Encyclopedia of Complexity and Systems Science*, pp. 8279–8291, Springer New York, 2009.

[26] R. Rathipriya, K. Thangavel, and J. Bagyamani, "Binary particle swarm optimization based biclustering of web usage data," *International Journal of Computer Applications*, vol. 25, pp. 43–49, July 2011.

[27] S. Balbi, "Beyond the curse of multidimensionality: high dimensional clustering in text mining," *Statistica Applicata-Italian Journal of Applied Statistics*, vol. 22, no. 1, pp. 53–63, 2010.

[28] V. Raponi, F. Martella, and A. Maruotti, "A biclustering approach to university performances: an italian case study," *Journal of Applied Statistics*, vol. 43, no. 1, pp. 31–45, 2016.

[29] R. Henriques and S. C. Madeira, "Bicpam: Pattern-based biclustering for biomedical data analysis," *Algorithms for Molecular Biology*, vol. 9, no. 1, pp. 1–30, 2014.

[30] M. A. Gara, S. Rosenberg, and L. Goldberg, "Dsm-iii—r as a taxonomy: A cluster analysis of diagnoses and symptoms.," *Journal of Nervous and Mental Disease*, 1992.

[31] I. V. Mechelen, H.-H. Bock, and P. D. Boeck, "Two-mode clustering methods: astructuredoverview," *Statistical Methods in Medical Research*, vol. 13, pp. 363–394, Oct. 2004.

[32] M. J. Brusco and P. Doreian, "An exact algorithm for the two-mode KL-means partitioning problem," *Journal of Classification*, vol. 32, pp. 481–515, Sept. 2015.

[33] R. C. Tryon, "Cluster analysis. edwards brothers," *Ann Arbor, Michigan*, p. 122, 1939.

[34] V. Maurizio, "Double k-means clustering for simultaneous classification of objects and variables," in *Advances in classification and data analysis*, pp. 43–52, Springer, 2001.

[35] J. Van Rosmalen, P. J. Groenen, J. Trejos, and W. Castillo, "Optimization strategies for two-mode partitioning," *Journal of Classification*, vol. 26, no. 2, pp. 155–181, 2009.

[36] W. J. Heiser and P. J. Groenen, "Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima," *Psychometrika*, vol. 62, no. 1, pp. 63–83, 1997.

[37] J. Trejos and W. Castillo, "Simulated annealing optimization for two-mode partitioning," in *Classification and Information Processing at the Turn of the Millennium*, pp. 135–142, Springer, 2000.

[38] W. Castillo and J. Trejos, "Two-mode partitioning: review of methods and application of tabu search," *Classification, clustering, and data analysis*, pp. 43–51, 2002.

[39] J. Hansohm, "Two-mode clustering with genetic algorithms," in *Classification, automation, and new media*, pp. 87–93, Springer, 2002.

[40] M. Brusco and P. Doreian, "A real-coded genetic algorithm for two-mode kl-means partitioning with application to homogeneity blockmodeling," *Social Networks*, vol. 41, pp. 26–35, 2015.

[41] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, pp. 846–850, Dec. 1971.

[42] R. Campello, "A fuzzy extension of the rand index and other related indexes for

clustering and classification assessment," *Pattern Recognition Letters*, vol. 28, pp. 833–841, May 2007.

[43] J. L. Burbidge, *The introduction of group technology*. John Wiley & Sons, 1975.

[44] H. M. Chan and D. Milner, "Direct clustering algorithm for group formation in cellular manufacture," *Journal of Manufacturing systems*, vol. 1, no. 1, pp. 65–75, 1982.

APPENDIX A

STATISTICAL COMPARISON OF ALGORITHMS

Table A.1: One-sided pair t-test for $\widehat{FRI}$

$N = M$	$K = L$	σ	p-value		$N = M$	$K = L$	σ	p-value	
			$\overline{FRI_1}$-$\overline{FRI_2}$	$\overline{FRI_2}$-$\overline{FRI_1}$				$\overline{FRI_1}$-$\overline{FRI_2}$	$\overline{FRI_2}$-$\overline{FRI_1}$
20	2	low	*0.06*	*0.94*	80	2	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	0.00	*1.00*			high	0.01	*0.99*
	3	low	0.00	*1.00*		3	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	*0.12*	*0.88*			high	*1.00*	0.00
	4	low	0.00	*1.00*		4	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	0.04	*0.96*			high	*1.00*	0.00
	5	low	0.00	*1.00*		5	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	*0.13*	*0.87*			high	*1.00*	0.00
50	2	low	0.00	*1.00*	150	2	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	0.00	*1.00*			high	0.02	*0.98*
	3	low	0.00	*1.00*		3	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	*0.18*	*0.82*			high	*1.00*	0.00
	4	low	0.00	*1.00*		4	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	*1.00*	0.00			high	*1.00*	0.00
	5	low	0.00	*1.00*		5	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	*1.00*	0.00			high	*1.00*	0.00

Table A.2: One-sided pair t-test for $\widehat{RI}^s$

$N=M$	$K=L$	σ	p-value		$N=M$	$K=L$	σ	p-value	
			$\overline{RI_1^s}$-$\overline{RI_2^s}$	$\overline{RI_2^s}$-$\overline{RI_1^s}$				$\overline{RI_1^s}$-$\overline{RI_2^s}$	$\overline{RI_2^s}$-$\overline{RI_1^s}$
20	2	low	*0.22*	*0.78*	80	2	low	*0.10*	*0.90*
		moderate	0.00	*1.00*			moderate	0.01	*0.99*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	0.00	*1.00*			high	*0.89*	*0.11*
	3	low	0.00	*1.00*		3	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	0.00	*1.00*			high	0.00	*1.00*
	4	low	0.00	*1.00*		4	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	0.00	*1.00*			high	0.00	*1.00*
	5	low	0.00	*1.00*		5	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	0.00	*1.00*			high	0.00	*1.00*
50	2	low	*0.05*	*0.95*	150	2	low	*0.05*	*0.95*
		moderate	0.02	*0.98*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	0.00	*1.00*			high	*0.88*	*0.12*
	3	low	0.00	*1.00*		3	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	0.00	*1.00*			high	*0.06*	*0.94*
	4	low	0.00	*1.00*		4	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	0.00	*1.00*			high	0.00	*1.00*
	5	low	0.00	*1.00*		5	low	0.00	*1.00*
		moderate	0.00	*1.00*			moderate	0.00	*1.00*
		high moderate	0.00	*1.00*			high moderate	0.00	*1.00*
		high	0.00	*1.00*			high	0.00	*1.00*