FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

SCALABLE CLUSTERING:

LARGE SCALE UNSUPERVISED LEARNING OF GAUSSIAN MIXTURE MODELS WITH OUTLIERS

By

YIJIA ZHOU

A Dissertation submitted to the
Department of Mathematics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2023

Yijia Zhou defended this dissertation on November 7, 2023.
The members of the supervisory committee were:

Adrian Barbu
Professor Directing Dissertation

Kyle A. Gallivan
Co-Professor Directing Dissertation

Gordon Erlebacher
University Representative

Giray Ökten
Committee Member

Mark Sussman
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

Dedicated to myself, who has endured countless hardships and setbacks while still maintaining kindness, tenacity, and courage.

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my supervisor, Dr. Adrian Barbu, for his patient guidance throughout my doctoral studies. His extensive expertise, perceptive insights, and constructive feedback have been instrumental in my successful completion of this dissertation.

I am equally indebted to my co-advisor, Dr. Kyle A. Gallivan, for his valuable suggestions, constructive criticisms, and profound academic insights. His contributions have significantly enriched my research, thereby enhancing the overall quality of this dissertation.

I extend my thanks to the entire committee, whose expertise and assistance have been invaluable during the dissertation process.

Lastly, I want to acknowledge the unwavering companionship and unconditional love from my friends. I genuinely believe they have always stood by me, offering their support no matter the circumstances.

Thank you all for your support and encouragement. I am very grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Herein is the comprehensive catalog of symbols employed in this dissertation.

Chapter 2	
Symbols	Meaning
N	the total number of objects
k	the total number of clusters
I	the total number of iterations
ε	the distance threshold
MinPts	the minimum number of data points required to form a dense region
σ	the influence of a point in its neighborhood
ξ	the importance of a density attractor
ε_1	distance parameter for spatial attributes such as latitude and longitude
ε_2	distance parameter for non-spatial attributes
$\Delta\varepsilon$	threshold value
p	the dimension of the observations
d	the number of principal components

Chapter 3	
Symbols	Meaning
N	the total number of observations
S	a subset of all the observations
n	the cardinality of S, $n = \|S\|$
p	the dimension of the observations $\mathbf{x}_i \in \mathbb{R}^p$
$l(\mathbf{x})$	the true label (cluster assignment) of observation $\mathbf{x}$
m	the number of positive clusters
T	the number of iterations (number of desired clusters) in Algorithm 1
S_k	the elements of S with label k, $S = \{\mathbf{x} \in S, l(\mathbf{x}) = k\}$
w_k	the weight of positive cluster k, $k = \overline{1, m}$
w_{-1}	the weight of negatives (observations $\mathbf{x}$ with $l(\mathbf{x}) = -1$)
$\boldsymbol{\mu}_k, \sigma_k$	true mean and standard deviation of positive cluster k
σ_{max}	the maximum standard deviation among all positive clusters
H	the set of all the negatives
F	a constant in the loss function (3.3). In this dissertation, $F = 2.5$
ρ	the bandwidth parameter in the loss function (3.3)
$R_\rho = \rho\sqrt{pF}$	radius of the support of the loss function (3.3)
$L(\mathbf{x}; \rho)$	loss function (3.3)
$\ell(\mathbf{x}; \rho)$	per observation loss function (3.4)

Chapter 4	
Symbols	Meaning
S'	a subset of all the observations
S''	a subset of all the observations
n_1	the cardinality of S', $n_1 = \|S'\|$
n_2	the cardinality of S'', $n_2 = \|S''\|$
m_1	the number of first-level positive clusters
m_2	the number of second-level positive clusters per each first-level cluster
m	the total number of second-level positive clusters, $m = m_1 \times m_2$
w_i	the weight of first-level positive cluster i, $i = \overline{1, m_1}$
w_{ij}	the weight of second-level cluster j within the first-level positive cluster i, $j = \overline{1, m_2}$
w_{-1}	the weight of first-level negatives
$w_{i,-1}$	the weight of second-level negatives for first-level positive cluster i
S_i	the elements of S with label $(i, \sim)$, $S = \{\mathbf{x} \in S, l(\mathbf{x}) = (i, \sim)\}$
S_{ij}	the elements of S with label (i, j), $S = \{\mathbf{x} \in S, l(\mathbf{x}) = (i, j)\}$
H_1	the set of all first-level negatives
H_{2i}	the set of second-level negatives from first positive cluster i, $i = \overline{1, m_1}$
$\sigma_{max,2} = \max_{i,j>0} \sigma_{ij}$	the maximum standard deviation among all the second-level positive clusters
$\sigma_{min,1} = \min_{i>0} \sigma_i$	the minimum standard deviation among all the first-level positive clusters
$\sigma_{max,1} = \max_{i>0} \sigma_i$	the maximum standard deviation among all the first-level positive clusters
ρ_1	the bandwidth parameter in the loss function (3.3)
ρ_2	the bandwidth parameter in the loss function (3.3)
$\boldsymbol{\mu}_i, \sigma_i$	true mean and standard deviation of first-level positive cluster $i, i = \overline{1, m_1}$
$\boldsymbol{\mu}_{ij}, \sigma_{ij}$	true mean and standard deviation of second-level positive cluster $i = \overline{1, m_1}, j = \overline{1, m_2}$

LIST OF ABBREVIATIONS

Herein is the compendium of abbreviations employed in this dissertation.

Abbreviation	Meaning
GMM	Gaussian Mixture Model
HGMM	Hierarchical Gaussian Mixture Model
CNN	Convolutional Neural Network

ABSTRACT

Clustering is a widely used technique with a long and rich history in a variety of areas. However, most existing algorithms do not scale well to large datasets, or are missing theoretical guarantees of convergence. In this dissertation, we develop a provably clustering algorithm namely Scalable Clustering by Robust Loss Minimization (SCRLM) that performs well on Gaussian Mixture Models with outliers. We derive theoretical guarantees that SCRLM obtains high accuracy with high probability under certain assumptions. Moreover, it can also be used as an initialization strategy for k-means clustering. Experiments on real-world large-scale datasets demonstrate the effectiveness of SCRLM when clustering a large number of clusters, and a k-means algorithm initialized by SCRLM outperforms most classic clustering methods in both speed and accuracy, while scaling well to large datasets such as ImageNet. We further extend SCRLM to Hierarchical SCRLM (HSCRLM) to handle hierarchical structures while maintaining robustness and theoretical guarantees. These advancements contribute to addressing modern clustering challenges.

CHAPTER 1

INTRODUCTION

1.1 Motivation, Problem and Thesis Statement

Clustering, or cluster analysis (Kaufman and Rousseeuw, 2009) is commonly defined as the grouping of similar objects into classes called clusters or defined more specifically as an unsupervised learning approach to classification of patterns into groups (clusters) based upon similarity, where a pattern is a representation of features or attributes of an object.

Over the past few decades, clustering has been widely applied to many fields, including information retrieval (Jardine and van Rijsbergen, 1971), image segmentation (Coleman and Andrews, 1979), pattern recognition (Diday et al., 1981), data mining (Mirkin, 2005) and disease diagnosis (Alashwal et al., 2019). In terms of traditional clustering techniques, they are broadly categorized into six main categories (Han et al., 2022): partitioning-based, hierarchical-based, density-based, distribution-based, grid-based and graph-based methods.

However, as the dataset grows in scale, traditional clustering algorithms become impractical due to their prohibitive computational cost and substantial memory requirements. To address these challenges, scalable clustering techniques have been developed.

Scalable clustering methods can be classified into three main categories (Mahdi et al., 2021): sampling-based (Zhao et al., 2019), projection-based (Thrun, 2018) and parallel-based (Brecheisen et al., 2006). Though these techniques provide more efficient and effective ways to cluster large-scale data, each type has its weakness. Sampling-based methods may not accurately capture the underlying distribution of the data if the sample is not representative enough. In addition, choosing an appropriate sample size can be challenging, and it may not be clear how many samples are needed to produce reliable results. Projection-based methods will result in information loss and inaccurate clustering results if the features of the data are not well captured. Parallel-based methods require specialized hardware which is challenging to implement in practice.

A crucial aspect of any clustering algorithm is the characterization of its performance. This requires efficient implementation and extensive well-designed empirical evaluation of model and real world problems. Ideally, the clustering algorithm would also have theoretical performance

guarantees in terms of guaranteed quality of the clustering and associated complexity bounds. The combination of the guarantees, bounds and empirical evidence rigorously characterizes the effectiveness and efficiency of the algorithm. Additionally, the combination provides guidance for the use of the algorithm on appropriate application problems.

This dissertation asserts the following thesis statement: It is possible to develop a clustering algorithm that

1. has theoretical performance guarantees for applications that satisfy reasonable assumptions on its cluster/noise structure;
2. has computational complexity that scales well as the dimension of the space in which the data resides, the number of clusters and the number of data grow to very large levels.

To support this assertion, new clustering algorithms called Scalable Clustering with Robust Loss Minimization (SCRLM) and Hierarchical Scalable Clustering with Robust Loss Minimization (HSCRLM) are proposed, analyzed and empirically evaluated in this dissertation.

1.2 Problem Model

The main application motivation is the problem of object recognition from images, where images often contain multiple regions or objects of interest, along with background noise. Inspired by the structure of such images and to facilitate the development of theoretical performance guarantees for SCRLM, a Gaussian Mixture Model (GMM) with outliers is introduced, where each Gaussian component represents one of the objects of interest (positives), while the outliers (negatives) correspond to the background images that cannot be clustered together.

A Hierarchical Gaussian Mixture Model (HGMM) is an extension of GMM that introduces a hierarchical structure to the model. In a GMM, the data is modeled as a mixture of several Gaussian distributions. In a HGMM, these Gaussian distributions are organized in a hierarchical manner, allowing for more complex and flexible representations of the data. Within each object of interest in a HGMM with outliers, there are different levels of sub-clustering, represented as first-level positives, second-level positives, up to the i-th level positives. Each level of sub-clustering can be represented by its own set of Gaussian mixture components. The outliers (negatives) at each level of sub-clustering can be identified as those instances that do not belong to any specific sub-clusters. The flexibility of the i-level HGMM model makes it well-suited for object recognition tasks involving images with complex object structures and multiple levels of classification. For simplicity, a two-level HGMM with outliers is used in this dissertation.

1.3 Related Work on Gaussian Mixture Models

The study of Gaussian Mixture Models can be traced back to Pearson (1894). The idea of using Gaussian mixtures in unsupervised learning was popularized by Duda et al. (1973). The Expectation Maximization (EM) algorithm (Dempster et al., 1977) was one of the first clustering algorithms for GMM. Xu and Jordan (1996) analyzed the convergence of EM for well-separated Gaussian mixtures. Dasgupta and Schulman (2007) proposed a two-round variant of the EM algorithm and showed that, with high probability, it can recover the parameters of the Gaussians to near-optimal precision. In recent years, approaches have been proposed to improve convergence guarantees and applied to different kinds of GMMs. Dwivedi et al. (2018) provided theoretical guarantees in two classes of misspecified mixture models and Segol and Nadler (2021) improved sample size requirements for accurate estimation by EM and gradient EM. Researchers also investigated the theoretical performance of spectral clustering in GMM. Vempala and Wang (2004) investigated the theoretical performance of spectral clustering in the isotropic GMM and proved that with high probability, exact recovery of the underlying cluster structure was achieved under a strong separation condition. Löffler et al. (2021) showed that spectral clustering is minimax optimal in GMMs with isotropic covariance, when the number of clusters is fixed and the signal-to-noise ratio is large enough.

An example of the proposed GMM with outliers in a 2D observation space is illustrated in Figure 1.1. There are five clusters (fishes, tigers, birds, cats and dogs) that are represented by colored dots and the remaining black squares that do not cluster together represent the outliers (e.g. textures).

1.4 Related Work on Hierarchical Gaussian Mixture Models

One of the earliest papers on Hierarchical Gaussian Mixture Models was by Liu et al. (2002), who proposed an efficient and effective method for speaker verification using HGMM. Experiments on benchmark datasets showed that HGMM outperformed GMM by achieving an 18% relative reduction in Equal Error Rate (EER). Since then, HGMMs have been applied to a wide range of tasks in machine learning and object/scene recognition. Eckart et al. (2018) presented a novel algorithm that achieves state-of-the-art speed and accuracy in adaptive 3D registration through its use of a HGMM representation which leads to a significant increase in speed, surpassing traditional GMM-based methods. In the area of environment perception and modeling for mobile autonomous

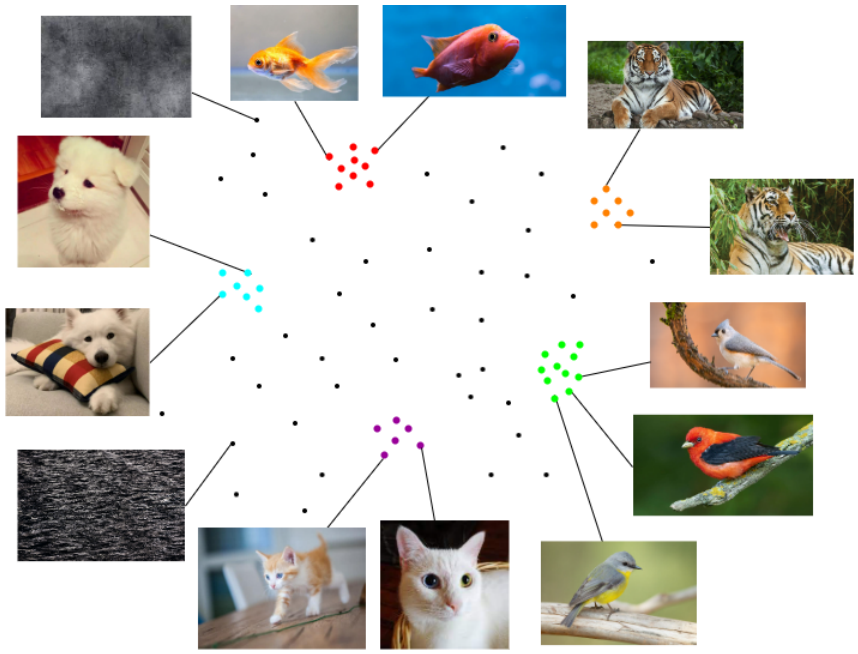


Figure 1.1: Gaussian Mixture Model with outliers in a 2D observation space.

systems, the use of HGMM (Srivastava and Michael, 2018) captured the structural dependencies between spatially distinct locations in the environment, which improved the accuracy of the environment representation which was important for mobile robots that need to navigate and interact with their surroundings in a safe and effective manner. In image segmentation, Shi et al. (2020) demonstrated the effectiveness of HGMM in improving accuracy for high-resolution remote sensing image segmentation compared to traditional GMM.

An example of the two-level HGMM with outliers in a 2D observation space used in this dissertation is illustrated in Figure 1.2. The first-level observations are depicted as large black symbols, with black squares representing first-level outliers (negatives) and black circles representing first-level positives. The three large black circles represent three different types of animals: cats, dogs, and horses, respectively. Meanwhile, the large black squares are textures that are not of interest. The second-level positives and negatives are indicated by small dots and small squares, respectively. For instance, in the cluster of dogs, there are three red dots that represent three different types of dogs: Samoyed, Bichon Frise, and Pomeranian. The second-level outliers (negatives) in the dog category are mutts, which will not be clustered together. Similarly, in the case of cats, there are three different types: Ragdoll, Shorthair, and Turkish, while the second-level outliers (negatives) are moggies.

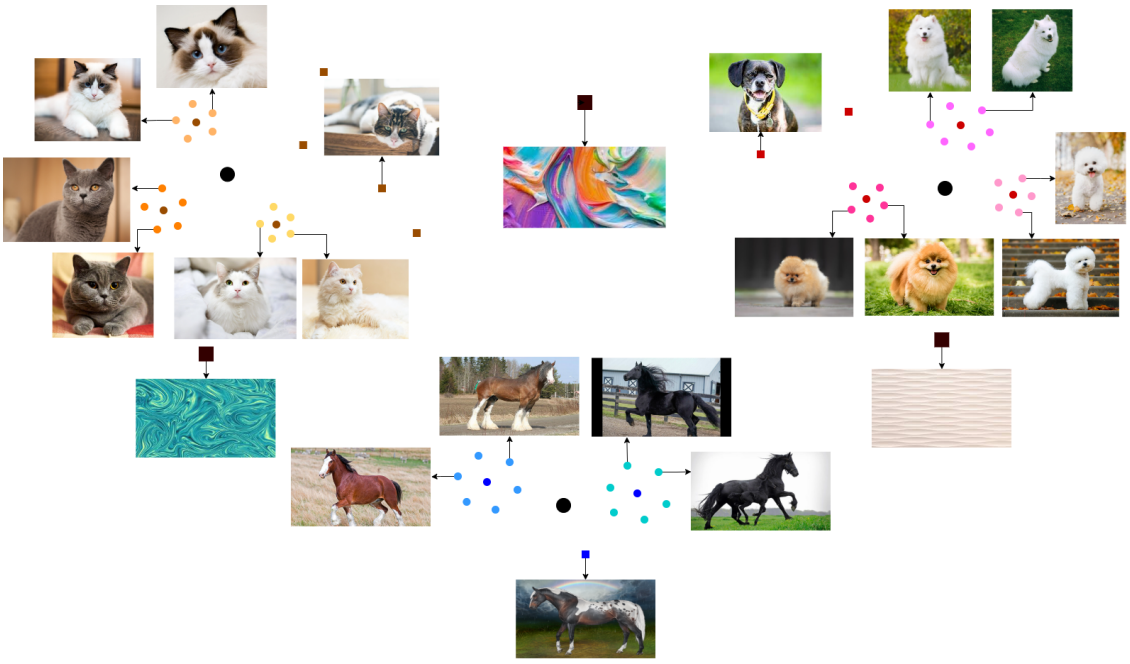


Figure 1.2: A two-level Hierarchical Gaussian Mixture Model with outliers.

1.5 Contributions of the Dissertation

The contributions of this dissertation are:

1. GMM and HGMM with outliers are introduced as two frameworks for characterizing application datasets, including those from image classification problems in computer vision, for the purpose of performance and complexity analysis.

2. Novel clustering algorithms, SCRLM and HSCRLM, are developed for GMM and HGMM with outliers respectively.

3. Theoretical guarantees are drived that showing SCRLM and HSCRLM are able to correctly cluster all the positives and detect all the negatives with high probability under certain assumptions for data from GMM and HGMM with outliers respectively.

4. The performance predictions are validated with experiments using simulated data and real data and it is shown that SCRLM outperforms other well-known algorithms when appropriate assumptions are met.

5. Experiments on real data demonstrate that SCRLM is very effective when the number of clusters and the data dimension are large, and it can be used as an initialization method for k-means clustering, outperforming k-means++ (Arthur and Vassilvitskii, 2007) in accuracy and computation time.

1.6 Overview of the Dissertation Structure

The structure of this dissertation is as follows. Chapter 2 offers an in-depth literature review that surveys a wide spectrum of traditional and scalable clustering techniques, providing a comprehensive foundation for the subsequent research. Chapter 3 develops a novel algorithm, SCRLM, to address the clustering problem in the proposed GMM with outliers, and the theoretical guarantees are derived. In Chapter 4, HSCRLM is introduced as an extension of SCRLM to tackle the hierarchical clustering problem, along with the theoretical guarantees within the framework of the proposed two-level HGMM with outliers. Chapter 5 presents experimental results, demonstrating the scalability, efficiency, and accuracy of both the proposed algorithms, SCRLM and HSCRLM, using synthetic and real data. Finally, Chapter 6 summarizes the key findings and draws conclusions from the research, with a discussion of future research work.

CHAPTER 2

LITERATURE REVIEW

2.1 Literature Review of Traditional Clustering Techniques

Traditional clustering methods are categorized into six groups: partitioning-based methods, hierarchical-based methods, density-based methods, model-based methods, grid-based methods and spectral graph-based methods. In this chapter, the state-of-the-art for each of these categories is discussed and their key advantages and disadvantages are identified.

2.1.1 Partitioning-based Clustering

A partitioning-based algorithm is a method used to arrange a set of N different objects into k partitions or clusters. The goal is to group the objects in such a way that each object belongs to one and only one cluster, and the objects within each cluster share some common characteristics or have a certain degree of similarity. An important characteristic of partitioning algorithms is that the number of clusters k must be specified ahead of time.

k-means. The k-means algorithm is one of the most popular partitioning-based clustering algorithms. It was initially proposed by Stuart Lloyd in 1957 for pulse-code modulation, but the detailed algorithm was published in (Lloyd, 1982). The algorithm starts by randomly selecting k initial cluster centroids. Each data point is assigned to the cluster whose centroid is closest to it. The closeness is typically measured using Euclidean distance, but other distance metrics can also be used. Convergence occurs when the assignment of data points to clusters and the centroids no longer change significantly. k-means is known for its simplicity and efficiency, making it a popular choice for a wide range of clustering tasks. However, it is worth noting that the algorithm may not always find the global optimum, and the quality of the clustering result depends on the initial centroid selection. It is also sensitive to outliers, and the clusters it forms are typically spherical and equally sized, which may not be suitable for all types of data.

k-means++. Since the initialization step of the k-means algorithm is indeed crucial for the quality of the resulting cluster assignments, k-means++ proposed by Arthur and Vassilvitskii (2007) provides a more principled way to initialize the centroids. The first cluster center is chosen uniformly

at random from the dataset, the remaining $k-1$ centers are selected by a weighted probability distribution. Each data point is chosen as the next center with a probability proportional to the squared distance to the closest already chosen center. In other words, points that are farther away from existing centers are more likely to be selected as the next center. It also provides theoretical guarantees that k-means++ is competitive with the optimal clustering. Specifically, it is proved to be $O(\log k)$ competitive, meaning that the cost of the clustering obtained with k-means++ is at most a constant factor times the cost of the optimal clustering, where this constant factor grows logarithmically with the number of clusters. Empirical investigations conducted by Fränti and Sieranoja (2019) have revealed that the effectiveness of k-means hinges entirely on the quality of its initialization where k-means++ outperforms other initialization methods, particularly in well-separated clusters. Due to its speed, simplicity, and good empirical performance, k-means++ stands as the predominant choice for k-means clustering. However, it can be computationally expensive when the number of clusters is large, and it is sensitive to outliers since the outliers may be far away from the main cluster that affect the selection of the initial centroids.

k-means$\|$ (k-means Parallel). A scalable variant of k-means++ called k-means$\|$ (Bahmani et al., 2012) is introduced by incorporating a parallel sampling process. It samples $O(k)$ points in each step and repeating for approximately $O(\log N)$ times. In the end, there are $O(k \log N)$ points selected and they are re-clustered into k initial centers. This approach excels not only as a parallel algorithm when compared to k-means++ but also exhibits slight improvements in performance under uniprocessor conditions. However, it's essential to note that the theoretical guarantees for k-means$\|$ are notably less robust than those of k-means++.

k-means$\|_{ER}$ (k-means Parallel Exponential Race). Following the initial sampling phase in the k-means$|$ algorithm, a refined variant known as k-means$|_{ER}$, as introduced by Makarychev et al. (2020), incorporates an iterative refinement step. This step involves the reassignment of data points to their closest cluster centers, updating the center locations, and removing empty clusters. Notably, this refinement process operates on a reduced subset of data points compared to the original dataset, resulting in computational efficiency without compromising the quality of clustering. The theoretical guarantee is improved by showing that the expected cost of the solution by k-means++ is bounded by a factor of at most $5(\ln k + 2)$ times the cost of the optimal solution and demonstrate that k-means$\|$ outperforms k-means++ due to its center pruning strategy.

Partitioning around Medoids. Compared to k-means, PAM (Kaufman, 1990) uses medoids as cluster centers to make it less sensitive to noise and outliers. Instead of minimizing a sum of squared Euclidean distances, it minimizes a sum of general pairwise dissimilarities. PAM is a heuristic algorithm and does not guarantee finding the globally optimal solution. Therefore, to obtain better clustering results, it is often necessary to run the PAM algorithm multiple times and select the best of the clustering results produced.

CLARANS. Clustering Large Application Based upon Randomized Search (CLARANS) (Ng and Han, 2002) is a clustering algorithm that relies on randomized search techniques. CLARANS employs a strategy of random movement to search for optimal clustering solutions within the dataset. It achieves this by iteratively generating candidate solutions and evaluating their quality to find the optimal clustering configuration. The quality of its results cannot be guaranteed and there is no theoretical guarantee on the number of steps required to reach a local optimum.

The advantages of partitioning-based clustering algorithms are their simplicity, ease of implementation, and efficiency. However, they are sensitive to the initial selection of cluster centers, and the results may be influenced by the initial values. Additionally, partitioning-based clustering algorithms are typically sensitive to outliers and noise, so data preprocessing and outlier handling are necessary when using these algorithms.

Table 2.1: Comparison of k-means, k-means++, k-means$\|$, k-means$\|_{ER}$, PAM, CLARANS.

Clustering Algorithm	Complexity	Sensitive to Outliers	Suitable for Non-Convex Clusters	Theoretical Guarantees	Scalability
k-means	$O(NkI)$	Yes	No	No	Good
k-means++	$O(NkI)$	Yes	No	Yes	Good
k-means$\|$	N/A	No	Yes	Yes	Excellent
k-means$\|_{ER}$	N/A	No	Yes	Yes	Excellent
PAM	$O(k(N-k)^2I)$	No	Yes	No	Poor
CLARANS	$O(N^2)$	No	Yes	No	Good

2.1.2 Hierarchical-based Clustering

Hierarchical-based clustering is a method of clustering that organizes a dataset into a hierarchical structure, creating a tree-like structure of clusters. They are either agglomerative or divisive:

1. Agglomerative methods are bottom-up approaches to clustering clustering that starts with each data point as a separate cluster and then repeatedly merges the closest clusters based

on distance or similarity until all data points are merged into a single cluster or the specified number of clusters is reached. Different distance measures and merging strategies such as minimum distance (single linkage), maximum distance (complete linkage), and average distance (average linkage) can be used during the merging process.

CURE (Clustering Using REpresentatives). CURE (Guha et al., 1998) is a data clustering algorithm for large databases that is more robust to outliers and captures clusters of different shapes and sizes by working with a representative sample and using incremental updates.

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies). BIRCH (Zhang et al., 1996) excels in scalability, making it particularly well-suited for clustering large-scale datasets. Its memory efficiency, achieved through the Clustering Feature (CF) approach, minimizes storage and memory requirements. Furthermore, BIRCH introduces a hierarchical structure in the form of the Clustering Feature Tree, enabling users to explore data at varying levels of granularity, which adds flexibility to data analysis. However, it is sensitive to the choice of parameters, such as the branching factor and may not perform well with datasets containing outliers.

CHAMELEON. CHAMELEON (Karypis et al., 1999) is a hierarchical clustering algorithm. It employs a two-step procedure: first, it creates a hierarchical tree structure, and then it utilizes this structure to automatically generate clusters at multiple levels of granularity. This hierarchical structure provides flexibility for users to explore data at various levels of detail, thereby accommodating a wide range of data analysis needs.

2. Divisive methods are top-down approaches to clustering that starts with all the data points as a single cluster and recursively dividing each cluster into smaller clusters until an acceptable clustering assignment is obtained. However, divisive methods are less commonly used primarily due to its complexity, the need for prior knowledge or assumptions about the data, and limited adaptability to complex data structures.

   DIANA (DIvisive ANAlysis Clustering). Introduced by Kaufman and Rousseeuw (2009), DIANA is a divisive hierarchical clustering method that divides clusters by identifying the cluster with the largest diameter and splitting it into smaller clusters. It has applications in pattern recognition and data mining.

2.1.3 Model-based Clustering

The basic idea of model-based clustering is to select a parametric model that represents the data distribution, such as GMM described earlier. The goal is to estimate the parameters of the GMM, including the means, variances, and mixture weights.

EM (Expectation Maximization). EM (Dempster et al., 1977) follows a straightforward two-step process, the Expectation (E-step) and Maximization (M-step). In the E-step, the algorithm computes the expected values of the data. This step involves calculating posterior probabilities, for each data point regarding its association with specific clusters within a probabilistic model. The Maximization (M-step), on the other hand, focuses on updating the model parameters to maximize the likelihood of the observed data. It takes into account the responsibilities obtained in the E-step and aims to derive new parameter estimates that best explain the observed data, given the expected values of the hidden variables.

Scaling EM. A scalable version of the Expectation-Maximization (EM) algorithm is introduced in the study by Bradley et al. (1998). This approach acknowledges that data can be managed in various ways: it can be compressed, stored in main memory, or discarded based on cluster membership certainty. If observations are confidently assigned to clusters, they are removed. If they belong to tightly-knit sub-clusters but cannot be discarded, they are compressed. Otherwise, they are kept in memory. Experimental results indicate that this Scalable EM approach outperforms the full in-memory EM algorithm when dealing with databases that fit within memory constraints.

Model-based clustering offers several advantages. It excels in handling clusters of diverse shapes and sizes, not confined by predefined distance or density metrics. It provides a robust probabilistic framework, enabling uncertainty assessment and facilitating model selection. Moreover, it adeptly manages missing data and extends its applicability to a wide spectrum of data types, including continuous, categorical, and mixed data. However, model-based clustering also has some limitations. It necessitates assumptions about data distributions, which might not consistently align with real-world datasets. It can be sensitive to model choice and the determination of an optimal cluster count, both of which present challenges. Furthermore, the computational complexity of model-based clustering may surpass that of alternative clustering methods.

2.1.4 Density-based Clustering

The basic idea of density-based clustering algorithms is to discover clustering structures based on the density of data points. Unlike partitioning-based methods, they do not require specifying the number of clusters in advance. Instead, they automatically identify high-density regions of data points to form clusters.

DBSCAN. The most representative density-based clustering algorithm is density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996) which requires two param-

eters, the cluster radius and the minimum number of points in each cluster, to form a dense region. Therefore, it is robust to outliers and effective in discovering clusters of arbitrary shapes. However, it is difficult to choose the appropriate parameter values if there is a large difference between the densities of the clusters. Things become worse when the points are in a high dimensional space.

OPTICS (Ordering Points To Identify the Clustering Structure). OPTICS (Ankerst et al., 1999) is an extension to DBSCAN. It stands out by providing more flexibility in input parameters. The core idea behind OPTICS is the concept of reachability distance. This distance metric combines both distance and density aspects, allowing it to adapt to different cluster densities and shapes effectively. One of the most valuable features is the OPTICS plot, offering a graphical representation of data density and clustering hierarchy. It aids in understanding the underlying data patterns. However, OPTICS remains sensitive to parameter settings, particularly the neighborhood radius, demanding careful tuning for optimal results.

DENCLUE (DENsity-based CLUstEring). DENCLUE (Hinneburg et al., 1998) works by summing the influence functions of individual data points. Clusters are identified through the determination of density-attractors, which are instrumental in characterizing clusters of arbitrary shapes using a simple equation rooted in the overall density function. Within the DENCLUE framework, two pivotal parameters, σ and ξ, play a crucial role. σ governs a point's influence within its neighborhood, and ξ dictates the significance of density attractors. This allows for a reduction in the number of density attractors, enhancing the algorithm's efficiency. DENCLUE demonstrates its proficiency in clustering high-dimensional feature vectors.

Many variants have emerged in the past decades to overcome the disadvantages of DBSCAN such as IDBSCAN (Borah and Bhattacharyya, 2004), ST-DBSCAN (Birant and Kut, 2007), DMDBSCAN (Elbatta and Ashour, 2013), and VDBSCAN (Liu et al., 2007). Table 2.2 contains a comparison of these algorithms in terms of four aspects: time complexity, required input parameters, support of varied density and insensitive to order of inputs.

2.1.5 Grid-based Clustering

Grid-based approaches divide the data space into a grid of cells. Each cell in this grid either contains data points or remains empty, forming a natural organization of the data. Grid-based clustering methods are particularly suited for handling large datasets efficiently. They offer advantages such as reduced computational complexity and the ability to discover clusters with varying

Table 2.2: Comparison of different density-based spatial clustering algorithms (Ahmed and Razak, 2014).

Name	Complexity	Input Parameters	Insensitivity	Varied Density
DBSCAN	$O(N \log N)$	ε, MinPts	No	No
DENCLUE	$O(\log \lvert D \rvert)$	σ, ξ	No	No
OPTICS	$O(N \log N)$	ε, MinPts	No	No
IDBSCAN	$O(N \log N)$	ε, MinPts	No	No
ST-DBSCAN	$O(N \log N)$	ε_1,ε_2, MinPts, $\Delta\varepsilon$	No	No
DMDBSCAN	$O(N \log N)$	N/A	Yes	Yes
VDBSCAN	$O(N \log N)$	N/A	Yes	Yes

shapes and densities. Some representative grid-based clustering algorithms are STING (WANG, 1997), WaveCluster (Sheikholeslami et al., 1998) and CLIQUE (Agrawal et al., 1998).

STING. STING operates by dividing the spatial domain into a grid structure and employs statistical tests to detect regions of the grid with significantly different data point densities, indicating potential spatial clusters. This grid-based, statistical approach makes STING adaptable to various types of spatial data and scalable to handle large datasets efficiently.

WaveCluster. WaveCluster leverages the multi-resolution capabilities of wavelet transforms to proficiently uncover clusters of arbitrary shapes at varying levels of precision. This approach involves iterative applications of wavelet transforms, yielding clusters across a spectrum of scales, from fine-grained to coarser representations. Notably, the wavelet transform filters play a pivotal role in automatically eliminating outliers from the data. An additional strength of WaveCluster lies in its robustness to variations in the order of the input observations, ensuring consistent and reliable results regardless of input sequence. However, it is not suitable for high-dimensional datasets (Andritsos et al., 2002).

CLIQUE. CLIQUE is known for its efficiency in identifying dense clusters within large spatial databases. What sets CLIQUE apart is its focus on identifying high-density clusters with the smallest possible dimensions, making it particularly well-suited for high-dimensional data. The algorithm employs a step-by-step strategy that incrementally increases data dimensions, starting from lower dimensions and gradually expanding to higher dimensions in order to uncover high-density regions. This capability enables CLIQUE to identify clusters of various shapes and sizes, irrespective of the data's dimensionality.

2.1.6 Spectral Graph-based Clustering

In contrast to many clustering methods that exclusively rely on distance measures, spectral clustering capitalizes on the attributes of a graph representation derived from the pairwise relationships within the data. Through harnessing the eigenvalues and eigenvectors of either the similarity matrix or the Laplacian matrix linked with the data, spectral clustering facilitates a transformation of the data into a reduced-dimensional space, thereby enhancing the distinctiveness and separability of clusters.

The origins of spectral clustering trace back to the early work of Donath and Hoffman (1973), who initially propose constructing graph partitions based on eigenvectors of the adjacency matrix. Subsequently, it gain prominence through influential studies such as (Shi and Malik, 2000), (Meilă and Shi, 2001), (Ng et al., 2002), and (Von Luxburg, 2007). This approach is characterized by its straightforward implementation and efficient solvability through standard linear algebra software, though computational demands escalate unless the graph is sparse.

From a theoretical standpoint, Kannan et al. (2004) utilizes the theorem from (Sinclair and Jerrum, 1989) to establish a worst-case guarantee for the algorithm in (Shi and Malik, 2000). Furthermore, Liu and Han (2018) introduces a self-tuning clustering method capable of automatically computing the scale and number of groups, addressing previous challenges associated with multi-scale data.

2.2 Literature Review of Scalable Clustering Techniques

Improving the computational efficiency and reducing the running time are key issues when dealing with large scale data. This section discusses some scalable clustering techniques. From the data perspective, we can either choose to reduce the sample size, which can be referred as sampling-based clustering or to reduce the data dimension, which is called projection-based clustering. In addition to these two methods, parallel clustering techniques, where the main idea is to divide the original data into small pieces and process them on different machines simultaneously are also used.

2.2.1 Sampling-based Clustering

Sampling-based clustering refers to a class of clustering algorithms that utilize random sampling techniques to handle large datasets efficiently. These algorithms aim to approximate the clustering structure of the entire dataset by working with a representative subset of the data.

The basic idea behind sampling-based clustering is to select a subset of the data points as representatives or prototypes and perform clustering on this smaller subset. The clustering results obtained from the subset are then used to infer the clustering structure of the entire dataset.

Sampling-based clustering algorithms have several advantages. They can handle large datasets that do not fit into memory by working with a smaller subset. They also reduce the computational complexity of clustering algorithms, making them more scalable. Additionally, sampling can help mitigate the effect of outliers and noise in the dataset.

However, sampling-based clustering also has limitations. The quality of clustering results depends on the representativeness and size of the sampled subset. If the sample is not representative or too small, important clusters may be missed. Moreover, the sampling process introduces randomness, which can lead to variations in the clustering results across different runs.

Overall, sampling-based clustering provides a trade-off between computational efficiency and clustering accuracy. It is particularly useful for large datasets where traditional clustering algorithms may be computationally expensive or memory-intensive.

Algorithms in this category are BIRCH (Zhang et al., 1996), CURE (Guha et al., 1998) and CLARANS (Ng and Han, 2002). The complexities of BIRCH, CURE, and CLARANS have been described in the previous sections.

2.2.2 Projection-based Clustering

Dimension reduction is a technique that aims to transform a high-dimensional dataset with a potentially large number of variables or features into a lower-dimensional representation, while preserving or maximizing relevant information. A large number of dimension reduction techniques have been discussed in (Pratihar, 2009), which are mainly classified into two categories: linear and non-linear methods. Nonlinear dimensionality methods are discussed in depth in (Lee and Verleysen, 2007).

In terms of linear dimension reduction methods, simple linear functions are used to transform high dimensional data into lower dimensions. Typical examples of linear projections include principal component analysis (PCA) (Pearson, 1901), projection pursuit mapping (Friedman and Tukey, 1974), factor analysis (Rummel, 1988), and independent component analysis (ICA)(Comon, 1991) (Hyvärinen and Oja, 2000).

Typical examples of nonlinear dimension reduction techniques are multidimensional scaling (Cox and Cox, 2008), curvilinear component analysis (CCA) (Demartines and Hérault, 1997),

the t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008), neighborhood retrieval visualizer (NeRV) (Venna et al., 2010), and polar swarm (Pswarm) (Thrun and Ultsch, 2021).

A comparative study of dimensionality reduction techniques is presented in (Ayesha et al., 2020). PCA and its variants (i.e., robust PCA, sparse PCA) are still widely used due to their simplicity and efficiency. The presence of outliers can adversely affect the performance of PCA. Table 2.3 summarizes the complexities of different methods to compute PCA. From the table, we observe that only stochastic SVD and probabilistic PCA have an efficient time complexity of $O(Npd)$. Moreover, probabilistic PCA has a lower communication complexity compared to stochastic SVD. Therefore, Elgamal et al. (2015) presents a scalable approach of PCA namely sPCA based on the probabilistic PCA (PPCA) algorithm (Tipping and Bishop, 1999). It employs several optimizations to support large datasets on distributed clusters. ICA works well with Gaussian data, but its variants have the capability to deal with Gaussian and non-Gaussian data.

Table 2.3: Comparison of different methods for computing PCA of an $N \times p$ matrix to produce d principal components (Elgamal et al., 2015).

Method to Compute PCA	Time Complexity	Communication Complexity
Eigen decomp. of covariance matrix	$O(Np \times \min(N, p))$	$O\left(p^2\right)$
SVD-Bidiag	$O\left(Np^2 + p^3\right)$	$O\left(\max\left((N+p)d, p^2\right)\right)$
Stochastic SVD (SSVD)	$O(Npd)$	$O\left(\max\left(Nd, d^2\right)\right)$
Probabilistic PCA (PPCA)	$O(Npd)$	$O(pd)$

2.2.3 Parallel-based Clustering

Parallel-based clustering process large datasets efficiently by harnessing the power of parallel computing architectures. This approach seeks to leverage the inherent parallelism in clustering algorithms to expedite their execution, making it particularly well-suited for big data applications.

One of the fundamental paradigms in parallel clustering is the MapReduce framework (Dean and Ghemawat, 2008), which has been widely adopted in parallelizing clustering algorithms. Many classical clustering algorithms, including k-means and DBSCAN, have been adapted to run efficiently in parallel MapReduce environments such as Multiplex k-means (Li et al., 2014) and MR-DBSCAN (He et al., 2014). These adaptations enable the scalability of these algorithms to massive datasets distributed across clusters of computers.

In addition to MapReduce, other parallel computing paradigms, such as Spark (Zaharia et al., 2010) and Peer-to-peer networks (Milojicic et al., 2002) have gained prominence in parallel clustering. Apache Spark (Han et al., 2016), in particular, has become popular due to its in-memory processing capabilities and ease of use for distributed data processing, making it a natural choice for parallelizing clustering algorithms.

Recent advancements in parallel-based clustering have extended beyond parallelizing existing algorithms. Novel clustering algorithms designed explicitly for parallel architectures have emerged, often taking advantage of the parallelism inherent in dense linear algebra operations, which are prevalent in many clustering techniques. Furthermore, hybrid approaches that combine the strengths of parallel and distributed computing have shown promise. These approaches distribute data across multiple clusters or nodes and then apply parallel algorithms locally, effectively achieving both horizontal and vertical scalability which are presented in Figure 2.1.

Parallel-based clustering has opened new avenues for addressing computational challenges associated with massive datasets. However, most parallel clustering algorithms are only able to deal with one single type of data, they fail to handle the real-time, heterogeneous, multi-view and multi-model big data.

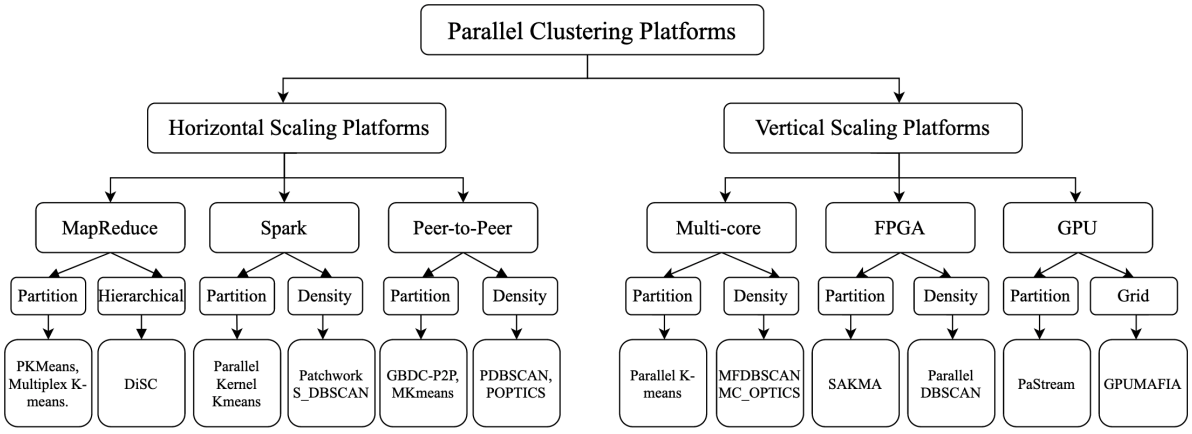


Figure 2.1: Parallel-based clustering algorithms (Dafir et al., 2021).

CHAPTER 3

SCALABLE CLUSTERING BY ROBUST LOSS MINIMIZATION IN GAUSSIAN MIXTURE MODEL WITH OUTLIERS

3.1 Problem Formulation

Given a set $X = \{\mathbf{x}_i \in \mathbb{R}^p, i = 1, \ldots, N\}$ of N points from a Gaussian Mixture Model with outliers containing m Gaussians, the goal is to group these points into m compact subsets. In this dissertation, only large and high-dimensional data and a reasonably large number of clusters, specifically, $N \approx 10^6, p \approx 10^3, m \approx 10^3$, are considered.

A Gaussian Mixture Model with outliers with parameters $\Theta = (w_{-1}, w_1, \mu_1, \Sigma_1, ..., w_m, \mu_m, \Sigma_m)$ is a weighted sum of m component Gaussian densities and outliers as given by the equation,

$$p(\mathbf{x} \mid \Theta) = \sum_{i=1}^{m} w_i \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_i, \Sigma_i) + w_{-1} O(\mathbf{x}) \tag{3.1}$$

where $\mathbf{x} \in \mathbb{R}^p$ is a p-dimensional data point, $w_i, i \in \{-1, 1, 2, \ldots, m\}$, are the mixture weights, $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_i, \Sigma_i), i = 1, \ldots, m$, are the component Gaussian densities and $O(\mathbf{x})$ is the distribution of the outliers.

It is assumed that each component density is a p-variate isotropic Gaussian function of the form,

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_i, \Sigma_i) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_i, \sigma_i^2\right) = \frac{1}{(2\pi)^{p/2} \left|\sigma_i^2 I_p\right|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T (\sigma_i^2 I_p)^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right\} \tag{3.2}$$

with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\Sigma_i = \sigma_i^2 I_p$.

Let $l(\mathbf{x}) \in \{-1, 1, 2, \ldots, m\}$ be the label of observation $\mathbf{x}$, i.e. the mixture component from which it was generated. The samples $\mathbf{x}_i$ with $l(\mathbf{x}_i) > 0$ are called positives and the outliers (with $l(\mathbf{x}_i) = -1$) are also called negatives.

Inspired by real data examples, where the observations are standardized feature vectors generated by a convolutional neural network (CNN) from real images of certain objects or background, in this dissertation, it is assumed that the m centers $\boldsymbol{\mu}_i, i = 1, \ldots, m$ and all outliers are generated from $O(\mathbf{x}) = \mathcal{N}(\mathbf{0}, I_p)$ and the label i positives are generated with frequency w_i from $\mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2)$, where

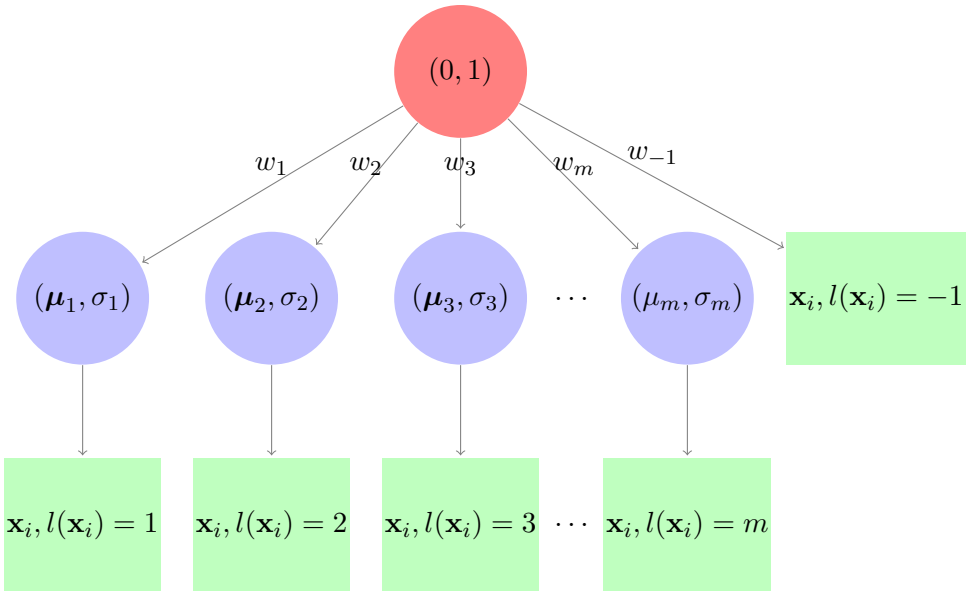


Figure 3.1: Structure of the Gaussian Mixture Model with outliers used in this dissertation.

$\sigma_i < 1, \forall i$. The structure of the Gaussian Mixture Model with outliers used in this dissertation is shown in Figure 3.1.

The problem of interest is to cluster a set of unlabeled observations generated from such a Gaussian Mixture Model with outliers and recover the labels $l(\mathbf{x}_i)$ as well as $\boldsymbol{\mu}_i$.

3.2 Robust Loss Function

A loss minimization approach is taken, using the following loss function,

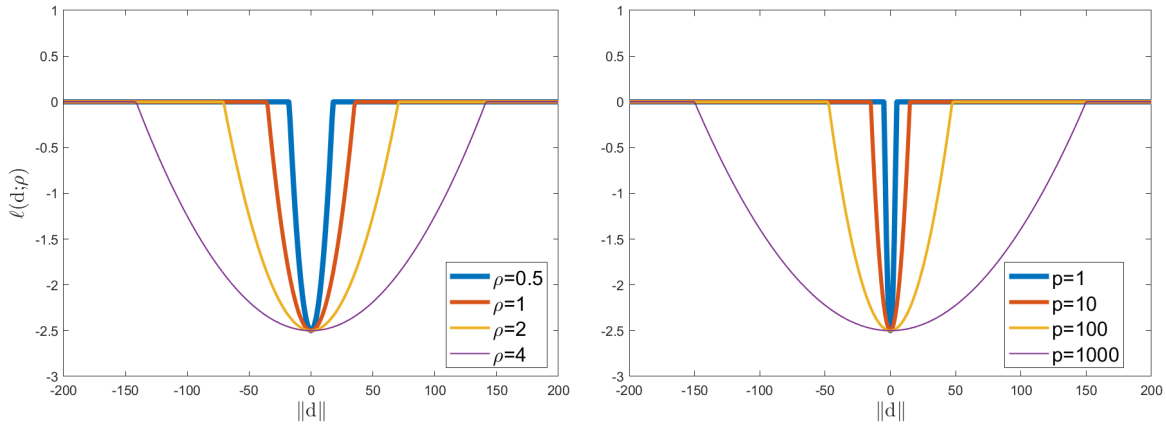
$$L(\mathbf{x}; \rho) = \sum_{i=1}^{N} \ell\left(\mathbf{x}_i - \mathbf{x}; \rho\right) = \sum_{i=1}^{N} \min\left(\frac{\|\mathbf{x}_i - \mathbf{x}\|^2}{p\rho^2} - F, 0\right) \tag{3.3}$$

where the per-observation loss, illustrated in Figure 3.2 is,

$$\ell(\mathbf{d}; \rho) = \min\left(\frac{\|\mathbf{d}\|^2}{p\rho^2} - F, 0\right). \tag{3.4}$$

The loss function $\ell(\mathbf{d}; \rho)$ is zero outside a ball of radius $R_\rho = \rho\sqrt{pF}$. The constant F is set in this dissertation to $F = 2.5$ (See Remark 1). The idea of the algorithm is to find the cluster centers $\boldsymbol{\mu}_i, i = 1, 2, ..., m$ as local minima of the loss function (3.3). For computational reasons, the centers $\boldsymbol{\mu}$ are sought among a subsample $S \subset \{1, \ldots, N\}$ of the observations $\mathbf{x}_i, i = 1, \ldots, N$,

$$\boldsymbol{\mu} = \mathbf{x}_k, \text{ where } k = \underset{i \in S}{\operatorname{argmin}}\, L\left(\mathbf{x}_i; \rho\right). \tag{3.5}$$



a) $p = 500$ b) $\rho = 3$

Figure 3.2: The robust loss function $\ell(\mathbf{d}; \rho)$ for different values of p and ρ.

After one cluster center $\boldsymbol{\mu} = \mathbf{x}_k$ has been found, all samples from S within the radius $\rho\sqrt{pF}$ from the $\boldsymbol{\mu}$ are considered as belonging to this cluster and are removed. The process is repeated until $\min_{j \in S} L(\mathbf{x}_j; \rho) = -F$. The process is illustrated in Figure 3.3. Suppose there are two clusters and some outliers, as shown in Figure 3.3 a). The first cluster center is found as the sample with minimum loss (3.3) among all subsamples. Then, all subsamples within the given radius $R = \rho\sqrt{pF}$ to this center are identified as belongings to this cluster (Figure 3.3 b)) and are removed. Then the process is repeated to find another cluster center (Figure 3.3 c)) and all subsamples within the same radius are removed. Now all the subsamples have loss values $-F$, which means there are no clusters left, so the remaining subsamples are negatives (Figure 3.3 d)).

Once all the cluster centers are found, the label of each observation is assigned to its nearest center based on the norm distances. If the nearest distance is greater than $\rho\sqrt{pF}$, it is classified as a negative. The procedure is summarized in Algorithm 1.

In clustering, accuracy is a measure that quantifies the similarity between the cluster assignments ($\hat{\mathbf{l}}$) and the ground truth labels ($\mathbf{l}$). It is calculated by considering all possible permutations of cluster assignments and finding the permutation that maximizes the sum of correctly assigned data points. The formula is defined as follows:

$$\text{Accuracy}(\mathbf{l}, \hat{\mathbf{l}}) = \frac{1}{N} \max_{\pi \in P} \sum_{i=1}^{N} I(\pi(\hat{\mathbf{l}}_i) = \mathbf{l}_i)$$

where P is the set of all possible permutations of cluster assignments, π represents a permutation of cluster assignments from the set P. I is an indicator function that evaluates to 1 if the cluster

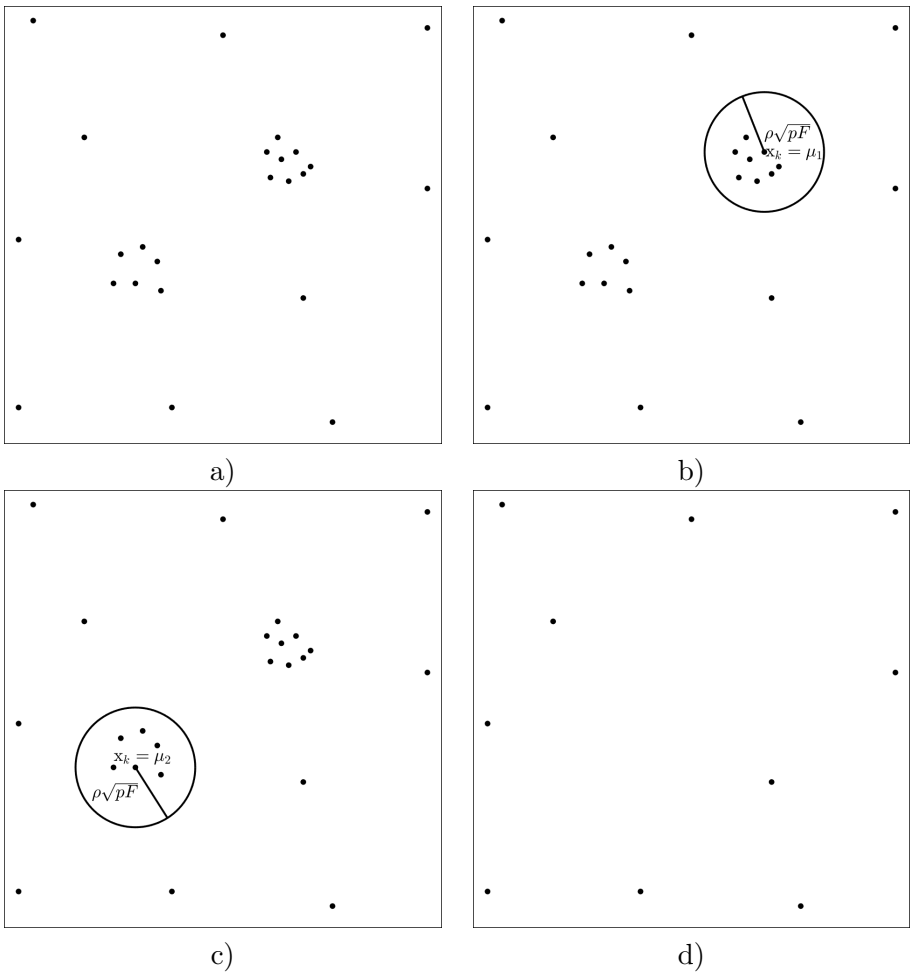


Figure 3.3: Diagram illustrating Algorithm 1.

assignment after applying permutation matches the true cluster label and 0 otherwise. The accuracy can be computed in polynomial time using the Hungarian algorithm.

3.3 Theoretical Guarantees

In this section, it is proven that SCRLM obtains high accuracy with high probability under Assumption 1 in the GMM with outliers. Corollary 1 to Corollary 6 are stated in Section 3.3.1 followed by basic propositions in Section 3.3.2. The main accuracy guarantees are stated as Theorem 1 and Corollary 7 in Section 3.3.3.

Let $\sigma_{max} = \max_{i \geq 0} \sigma_i$ be the maximum over the true standard deviations of the positive clusters. The following is the main assumption that is needed for the theoretical guarantees.

Assumption 1. $\sigma_{max} \leq \rho < \sqrt{0.6}$, *where* ρ *is the bandwidth parameter for the loss function* (3.3).

Algorithm 1 Scalable Clustering by Robust Loss Minimization (SCRLM)

Input: $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^p$, the number of subsamples n, the bandwidth parameter ρ, the desired number of clusters T.

Output: the number of clusters m, the centers μ_t for $t = 1, \ldots, m$, cluster labels $l_1, \ldots, l_N \in \{-1, 1, 2, \ldots, m\}$.

Randomly select a set of n observations $S = \{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_n}\} \subset X$ without replacement from X.

Compute $L(\mathbf{x}_j; \rho) = \sum_{i=1}^{N} \ell\left(\mathbf{x}_i - \mathbf{x}_{\boldsymbol{j}}; \rho\right)\ \forall j \in S$, where $\ell(\mathbf{d}; \rho) = \min\left(\frac{\|\mathbf{d}\|^2}{p\rho^2} - F, 0\right)$.

for $t = 1$ to T **do**

  Find $k = \underset{j \in S}{\operatorname{argmin}}\, L\left(\mathbf{x}_{\boldsymbol{j}}; \rho\right)$

  if $L\left(\mathbf{x}_{\boldsymbol{k}}; \rho\right) < -F$ **then**

    Obtain one positive cluster center as $\boldsymbol{\mu}_t = \mathbf{x}_k$

    Update $S \leftarrow S - \left\{\mathbf{x} \in S, \|\mathbf{x} - \boldsymbol{\mu}_t\| < \rho\sqrt{pF}\right\}$

  else

    break

  end if

end for

for $i = 1$ to N **do**

  Compute $k = \underset{j}{\operatorname{argmin}} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|$

  if $\|\mathbf{x}_i - \boldsymbol{\mu}_k\| < \rho\sqrt{pF}$ **then**

    $l_i = k$

  else

    $l_i = -1$

  end if

end for

3.3.1 Preliminaries

In this section, the basic separation and concentration results for pairs of training examples are presented. These results use the following Lemma from Wainwright (2019).

Lemma 1. *((Wainwright, 2019), Example 2.5) If $Z_1, \ldots, Z_n$ are i.i.d Gaussian random variables $Z_i \sim \mathcal{N}(0, 1)$, then for any $\epsilon \in (0, 1)$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i^2 - 1\right| \geq \epsilon\right) \leq 2\exp\left\{-n\epsilon^2/8\right\}.$$

Lemma 2. *If $Z_1, \ldots, Z_n$ are i.i.d Bernoulli random variables $Z_i \sim B(w)$, then for any $\epsilon \in (0, 1)$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - w\right| \geq \epsilon\right) \leq 2\exp\left\{-2n\epsilon^2\right\}.$$

Corollary 1. *If* $\mathbf{x} = (X_1, \ldots, X_p)$ *is a multivariate Gaussian random variable* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I_p)$*, then* $\mathbb{E}\left(\|\mathbf{x}\|^2\right) = p$ *and for any* $\epsilon \in (0, 1)$*,*

$$\mathbb{P}\left(\left|\frac{1}{p}\|\mathbf{x}\|^2 - 1\right| \geq \epsilon\right) \leq 2\exp\{-p\epsilon^2/8\}.$$

Proof. Follows from Lemma 1 above taking $Z_i = X_i, i = 1, \ldots, p$. □

Corollary 2. *If* $\mathbf{x} = (X_1, \ldots, X_p), \mathbf{y} = (Y_1, \ldots, Y_p)$ *are independent multivariate Gaussian random variables* $\mathbf{x}, \mathbf{y} \sim \mathcal{N}(\mathbf{0}, I_p)$*, then* $\mathbb{E}\left(\|\mathbf{x} - \mathbf{y}\|^2\right) = 2p$ *and for any* $\epsilon \in (0, 1)$*,*

$$\mathbb{P}\left(\left|\frac{1}{2p}\|\mathbf{x} - \mathbf{y}\|^2 - 1\right| \geq \epsilon\right) \leq 2\exp\{-p\epsilon^2/8\}.$$

Proof. Follows from Lemma 1 above taking $Z_i = (X_i - Y_i)/\sqrt{2}, i = 1, \ldots, p$. □

Using these results, it follows that with high probability the negatives are well-separated from each other.

Corollary 3 (Separation between negatives)**.** *For two negatives* $\mathbf{x}_i$ *and* $\mathbf{x}_k$*, with probability at least* $1 - 2\exp\{-p/128\}$*, the separation satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > 1.5p, \quad \mathbf{x}_i, \mathbf{x}_k \in H.$$

Proof. Since $\mathbf{x}_k \sim \mathcal{N}(\mathbf{0}, I_p)$ and $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, I_p)$, then $\mathbf{x}_i - \mathbf{x}_k \sim \mathcal{N}(\mathbf{0}, 2I_p)$, thus $\mathbb{E}\left(\|\mathbf{x}_i - \mathbf{x}_k\|^2\right) = 2p$. According to Corollary 2, it follows that

$$\mathbb{P}\left(\left|\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2p} - 1\right| \geq \epsilon\right) \leq 2\exp\left\{-p\epsilon^2/8\right\},$$

then

$$\mathbb{P}\left(\|\mathbf{x}_i - \mathbf{x}_k\|^2 \leq 2p(1-\epsilon)\right) \leq 2\exp\left\{-p\epsilon^2/8\right\}.$$

Then with high probability at least $1 - 2\exp\{-p\epsilon^2/8\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > 2p(1-\epsilon).$$

Now take $\epsilon = 1/4$ so that with high probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > 1.5p.$$

□

It then follows that the positives from the same cluster are within a certain radius from each other with high probability.

Corollary 4 (Concentration of positives in the same cluster)**.** *For any positive cluster S_j with mean $\boldsymbol{\mu}_j$ and covariance matrix $\sigma_j^2 I_p$, with probability at least $1-2\exp\{-p/128\}$, the concentration is bounded as*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 < 2.5p\sigma_j^2, \quad \mathbf{x}_i, \mathbf{x}_k \in S_j.$$

Proof. Since $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma_j^2 I_p)$ and $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma_j^2 I_p)$, then $\mathbf{x}_i - \mathbf{x}_k \sim \mathcal{N}(\mathbf{0}, 2\sigma_j^2 I_p)$, thus

$$\mathbb{E}\left(\|\mathbf{x}_i - \mathbf{x}_k\|^2\right) = 2p\sigma_j^2.$$

According to Corollary 1, it follows that

$$\mathbb{P}\left(\left|\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2p\sigma_j^2} - 1\right| \geq \epsilon\right) \leq 2\exp\left\{-p\epsilon^2/8\right\},$$

then

$$\mathbb{P}(\|\mathbf{x}_i - \mathbf{x}_k\|^2 \geq 2p\sigma_j^2(1+\epsilon)) \leq 2\exp\{-p\epsilon^2/8\}.$$

Take $\epsilon = 1/4$, yields

$$\mathbb{P}(\|\mathbf{x}_i - \mathbf{x}_k\|^2 \geq 2.5p\sigma_j^2) \leq 2\exp\{-p/128\}.$$

Therefore, with probability at least $1-2\exp\{-p/128\}$, the concentration is bounded as

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 < 2.5p\sigma_j^2.$$

□

Then, it is proven that the positives are well-separated from the negatives with high probability.

Corollary 5 (Separation between positives and negatives)**.** *For negative $\mathbf{x}_i$ and positive $\mathbf{x}_k$ from cluster S_j with mean $\boldsymbol{\mu}_j$ and covariance matrix $\sigma_j^2 I_p$, with probability at least $1-2\exp\{-p/128\}$, the separation satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > p(1.5 + 0.75\sigma_j^2), \quad \mathbf{x}_i \in H, \mathbf{x}_k \in S_j.$$

Proof. Since $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma_j^2 I_p)$ and $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, I_p)$, then $\mathbf{x}_i - \mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma_j^2 I_p + I_p)$, thus $\mathbf{x}_i - \mathbf{x}_k = \boldsymbol{\mu}_j + \boldsymbol{\epsilon_1}\sqrt{\sigma_j^2+1}$ with $\boldsymbol{\epsilon_1} \sim \mathcal{N}(\mathbf{0}, I_p)$. Since $\boldsymbol{\mu}_j \sim \mathcal{N}(\mathbf{0}, I_p)$, then $\mathbf{x}_i - \mathbf{x}_k$ is a Gaussian with $\mathbb{E}(\mathbf{x}_i - \mathbf{x}_k) = \mathbf{0}$ and

$$\mathbb{E}\left(\|\mathbf{x}_i - \mathbf{x}_k\|^2\right) = \mathbb{E}\left[\left(\boldsymbol{\mu}_j + \boldsymbol{\epsilon_1}\sqrt{\sigma_j^2+1}\right)^T \left(\boldsymbol{\mu}_j + \boldsymbol{\epsilon_1}\sqrt{\sigma_j^2+1}\right)\right] = \mathbb{E}(\|\boldsymbol{\mu}_j\|^2) + (\sigma_j^2+1)\mathbb{E}\left({\boldsymbol{\epsilon_1}}^T\boldsymbol{\epsilon_1}\right),$$

thus

$$\mathbb{E}\left(\|\mathbf{x}_i - \mathbf{x}_k\|^2\right) = p + (\sigma_j^2 + 1)\mathbb{E}\left(\|\boldsymbol{\epsilon_1}\|^2\right) = (2 + \sigma_j^2)p.$$

According to Corollary 1, it follows immediately that,

$$\mathbb{P}\left(\left|\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{(2+\sigma_j^2)p} - 1\right| \geq \epsilon\right) \leq 2\exp\left\{-p\epsilon^2/8\right\},$$

then

$$\mathbb{P}\left(\|\mathbf{x}_i - \mathbf{x}_k\|^2 \leq p(2+\sigma_j^2)(1-\epsilon)\right) \leq 2\exp\left\{-p\epsilon^2/8\right\}.$$

Then with probability at least $1 - 2\exp\{-p\epsilon^2/8\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > p(2+\sigma_j^2)(1-\epsilon).$$

Now take $\epsilon = 1/4$, so that with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > p(1.5 + 0.75\sigma_j^2).$$

□

Moreover, positives from different clusters are also well-separated from each other with high probability.

Corollary 6 (Separation between positives in different clusters)**.** *For positive* $\mathbf{x}_i$ *from cluster* S_i *with true mean* $\boldsymbol{\mu}_i$ *and covariance matrix* $\sigma_i^2 I_p$ *and positive* $\mathbf{x}_k$ *from another cluster* S_j *with true mean* $\boldsymbol{\mu}_j$ *and covariance matrix* $\sigma_j^2 I_p$*, with probability at least* $1 - 2\exp\{-p/128\}$*, the separation satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2), \quad \mathbf{x}_i \in S_i, \mathbf{x}_k \in S_j.$$

Proof. Since $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma_j^2 I_p)$ and $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 I_p)$, then $\mathbf{x}_i - \mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i, \sigma_j^2 I_p + \sigma_i^2 I_p)$, thus $\mathbf{x}_i - \mathbf{x}_k = \boldsymbol{\mu}_j - \boldsymbol{\mu}_i + \boldsymbol{\epsilon_1}\sqrt{\sigma_j^2 + \sigma_i^2}$ with $\boldsymbol{\epsilon_1} \sim \mathcal{N}(\mathbf{0}, I_p)$. Since $\boldsymbol{\mu}_j \sim \mathcal{N}(\mathbf{0}, I_p)$ and $\boldsymbol{\mu}_i \sim \mathcal{N}(\mathbf{0}, I_p)$, then $\boldsymbol{\mu}_j - \boldsymbol{\mu}_i \sim \mathcal{N}(\mathbf{0}, 2I_p)$, then $\mathbf{x}_i - \mathbf{x}_k$ is a Gaussian with $\mathbb{E}(\mathbf{x}_i - \mathbf{x}_k) = \mathbf{0}$ and

$$\mathbb{E}\left(\|\mathbf{x}_i - \mathbf{x}_k\|^2\right) = \mathbb{E}\left[\left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i + \boldsymbol{\epsilon_1}\sqrt{\sigma_j^2 + \sigma_i^2}\right)^T \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i + \boldsymbol{\epsilon_1}\sqrt{\sigma_j^2 + \sigma_i^2}\right)\right],$$

thus

$$\mathbb{E}\left(\|\mathbf{x}_i - \mathbf{x}_k\|^2\right) = \mathbb{E}(\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\|^2) + (\sigma_j^2 + \sigma_i^2)\mathbb{E}\left(\boldsymbol{\epsilon_1}^T\boldsymbol{\epsilon_1}\right) = 2p + (\sigma_j^2 + \sigma_i^2)\mathbb{E}\left(\|\boldsymbol{\epsilon_1}\|^2\right) = (2 + \sigma_i^2 + \sigma_j^2)p.$$

According to Corollary 1, it follows that

$$\mathbb{P}\left(\left|\frac{\|\mathbf{x}_i-\mathbf{x}_k\|^2}{(2+\sigma_i^2+\sigma_j^2)p}-1\right|\geq\epsilon\right)\leq 2\exp\left\{-p\epsilon^2/8\right\},$$

then

$$\mathbb{P}\left(\|\mathbf{x}_i-\mathbf{x}_k\|^2\leq p(2+\sigma_i^2+\sigma_j^2)(1-\epsilon)\right)\leq 2\exp\left\{-p\epsilon^2/8\right\}.$$

Then with probability at least $1-2\exp\{-p\epsilon^2/8\}$, the separation satisfies

$$\|\mathbf{x}_i-\mathbf{x}_k\|^2>p(2+\sigma_i^2+\sigma_j^2)(1-\epsilon).$$

Now take $\epsilon=1/4$ so that with probability at least $1-2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i-\mathbf{x}_k\|^2>p(1.5+0.75\sigma_i^2+0.75\sigma_j^2).$$

□

Remark 1. *Observed from Corollary 4 where we take $\epsilon=1/4$, so that we obtain $F=2.5$.*

The previous corollaries are used to prove that with high probability, all positives from each cluster are within $2.5p\rho^2$ of each other, and $2.5p\rho^2$ away from the other clusters and from the negatives.

Proposition 1. *Given a set X of N samples from a GMM with outliers, and $\sigma_{max}\leq\rho<\sqrt{0.6}$, then with probability at least $1-6N^2\exp\{-p/128\}$, the distance between positives within a cluster satisfies*

$$\|\mathbf{x}_i-\mathbf{x}_j\|^2<2.5p\rho^2,\quad\forall\mathbf{x}_i,\mathbf{x}_j\ s.t.\ l(\mathbf{x}_i)=l(\mathbf{x}_j)>0,$$

and the distance between positives from a cluster and other samples not in that cluster satisfies

$$\|\mathbf{x}_i-\mathbf{x}_j\|^2>2.5p\rho^2\quad\forall\mathbf{x}_i,\mathbf{x}_j\ s.t.\ l(\mathbf{x}_j)\neq l(\mathbf{x}_i)>0.$$

Proof. From Corollary 4, with probability at least $1-2\exp\{-p/128\}$, the distance between two positives in the same cluster is bounded as

$$\|\mathbf{x}_i-\mathbf{x}_j\|^2<2.5p\sigma_{l(\mathbf{x}_i)}^2,\quad l(\mathbf{x}_i)=l(\mathbf{x}_j)>0.$$

Using the union bound, with probability at least $1-2N^2\exp\{-p/128\}$, the distances between all positives in the same cluster are bounded as

$$\|\mathbf{x}_i-\mathbf{x}_j\|^2<2.5p\sigma_{l(\mathbf{x}_i)}^2\leq 2.5p\rho^2,\quad\forall\mathbf{x}_i,\mathbf{x}_j\text{ s.t. }l(\mathbf{x}_i)=l(\mathbf{x}_j)>0.$$

From Corollary 5, with probability at least $1-2\exp\{-p/128\}$, the distance between a positive and a negative satisfies

$$\|\mathbf{x}_i-\mathbf{x}_j\|^2 > p(1.5+0.75\sigma^2_{l(\mathbf{x}_i)}), \quad l(\mathbf{x}_i)>0, l(\mathbf{x}_j)=-1.$$

Using the union bound, with probability at least $1-2N^2\exp\{-p/128\}$, the distance between any positive and negative satisfies

$$\|\mathbf{x}_i-\mathbf{x}_j\|^2 > p(1.5+0.75\sigma^2_{l(\mathbf{x}_i)}), \quad \forall \mathbf{x}_i,\mathbf{x}_j, \text{ s.t. } l(\mathbf{x}_i)>0, l(\mathbf{x}_j)=-1.$$

Given $\sigma_{max} \leq \rho < \sqrt{\frac{1.5+0.75\sigma_k^2}{2.5}}$, since $0<\sigma_k<1$, therefore, with $\sigma_{max} \leq \rho < \sqrt{\frac{1.5}{2.5}} = \sqrt{0.6}$, then with probability at least $1-2N^2\exp\{-p/128\}$, the distance between any positive and negative satisfies

$$\|\mathbf{x}_i-\mathbf{x}_j\|^2 > 2.5p\rho^2, \quad \forall \mathbf{x}_i,\mathbf{x}_j, \text{ s.t. } l(\mathbf{x}_i)>0, l(\mathbf{x}_j)=-1.$$

From Corollary 6, with probability at least $1-2\exp\{-p/128\}$, the distance between two positives from different clusters satisfies

$$\|\mathbf{x}_i-\mathbf{x}_j\|^2 > p(1.5+0.75(\sigma^2_{l(\mathbf{x}_i)}+\sigma^2_{l(\mathbf{x}_j)})), \quad 0<l(\mathbf{x}_i)\neq l(\mathbf{x}_j)>0.$$

Using the union bound, with probability at least $1-2N^2\exp\{-p/128\}$, the distance between any two positives from different clusters satisfies

$$\|\mathbf{x}_i-\mathbf{x}_j\|^2 > p(1.5+0.75(\sigma^2_{l(\mathbf{x}_i)}+\sigma^2_{l(\mathbf{x}_j)})), \quad \forall \mathbf{x}_i,\mathbf{x}_j \text{ s.t. } 0<l(\mathbf{x}_i)\neq l(\mathbf{x}_j)>0.$$

Given $\sigma_{max} \leq \rho < \sqrt{\frac{1.5+0.75(\sigma_i^2+\sigma_j^2)}{2.5}}$, since $0<\sigma_i,\sigma_j<1$, therefore, with $\sigma_{max} \leq \rho < \sqrt{0.6}$, then with probability at least $1-2N^2\exp\{-p/128\}$, the distance between any two positives from different clusters satisfies

$$\|\mathbf{x}_i-\mathbf{x}_j\|^2 > 2.5p\rho^2, \quad \forall \mathbf{x}_i,\mathbf{x}_j \text{ s.t. } 0<l(\mathbf{x}_i)\neq l(\mathbf{x}_j)>0.$$

Therefore, with probability at least $1-4N^2\exp\{-p/128\}$, the distance between any positive and any sample not from that cluster satisfies

$$\|\mathbf{x}_i-\mathbf{x}_j\|^2 > 2.5p\rho^2, \quad \forall \mathbf{x}_i,\mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j)\neq l(\mathbf{x}_i)>0.$$

Therefore, with probability at least $1-6N^2\exp\{-p/128\}$, the following bounds on positives within a cluster and between clusters are satisfied

$$\|\mathbf{x}_i-\mathbf{x}_j\|^2 > 2.5p\rho^2, \quad \forall \mathbf{x}_i,\mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j)\neq l(\mathbf{x}_i)>0,$$

and

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) = l(\mathbf{x}_i) > 0.$$

□

Finally, a bound for the probability that a sample S has at least k elements from each positive cluster is proven.

Lemma 3. *If the clusters have weights $w_1, ..., w_m$, with $\sum_{k=1}^m w_k \leq 1$, then the probability that a sample S of size $|S| = n$ contains at least one observation from each cluster is at least*

$$\mathbb{P}(|\{\mathbf{x} \in S, l(\mathbf{x}) = k\}| \geq 1, \forall k = \overline{1, m}) \geq 1 - \sum_{k=1}^{m} (1 - w_k)^n.$$

Proof. The probability that S contains no elements from cluster k is

$$\mathbb{P}(l(\mathbf{x}) \neq k, \forall \mathbf{x} \in S) = (1 - w_k)^n.$$

Then using the union bound, the probability that there is a k such that S does not contain any elements from cluster k is

$$\mathbb{P}(\exists k, l(\mathbf{x}) \neq k, \forall \mathbf{x} \in S) \leq \sum_{k=1}^{m} (1 - w_k)^n,$$

which implies the result. □

Lemma 4. *With probability at least $1 - 2\exp\left\{-2(wn - k + 1)^2/n\right\}$, a subset S of size $|S| = n$ samples contains at least k observation from positive cluster with weight w.*

Proof. According to Lemma 2,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - w\right| \geq \epsilon\right) \leq 2\exp\left\{-2n\epsilon^2\right\}.$$

Take $\epsilon = w - (k-1)/n$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Z_i - nw\right| \geq nw - k + 1\right) \leq 2\exp\left\{-2n\epsilon^2\right\}.$$

$$\mathbb{P}\left(\sum_{i=1}^{n} Z_i - nw \geq nw - k + 1 \text{ or } \sum_{i=1}^{n} Z_i - nw \leq -nw + k - 1\right) \leq 2\exp\left\{-2n\epsilon^2\right\}.$$

$$\mathbb{P}\left(\sum_{i=1}^{n} Z_i \leq k - 1\right) \leq 2\exp\left\{-2(wn - k + 1)^2/n\right\}.$$

Therefore, with probability at least $1 - 2\exp\left\{-2(wn - k + 1)^2/n\right\}$, $\sum_{i=1}^n Z_i \geq k$. □

3.3.2 Loss Bounds

In this section, the concentration and separation results are used to obtain bounds on the loss function values. First, it is proven that, with high probability, the loss value of a negative is $-F$.

Proposition 2. *Given a set X of N samples from a GMM with outliers, randomly select a set S of $|S| = n$ subsamples from it, with $\sigma_{max} \leq \rho < \sqrt{0.6}$, then for a negative sample $\mathbf{x}_j \in S$ with $l(\mathbf{x}_j) = -1$, with probability at least $1 - 4N\exp\{-p/128\}$, the loss satisfies $L(\mathbf{x}_j; \rho) = -F$.*

Proof. From Corollary 3, for $\mathbf{x}_i, l(\mathbf{x}_i) = -1$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p, \quad l(\mathbf{x}_i) = -1, i \neq j.$$

Using the union bound, with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p, \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = -1, i \neq j.$$

Given $\sigma_{max} \leq \rho < \sqrt{0.6}$, then with probability at least $1-2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p > 2.5p\rho^2, \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = -1, i \neq j.$$

From Corollary 5, for $\mathbf{x}_i, l(\mathbf{x}_i) > 0$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5 + 0.75\sigma^2_{l(\mathbf{x}_i)}), \quad l(\mathbf{x}_i) > 0.$$

Using the union bound, with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5 + 0.75\sigma^2_{l(\mathbf{x}_i)}), \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) > 0.$$

Given $\sigma_{max} \leq \rho < \sqrt{0.6}$, then with probability at least $1-2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5 + 0.75\sigma^2_{l(\mathbf{x}_i)}) > 2.5p\rho^2 \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) > 0.$$

Therefore, with probability at least $1 - 4N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 2.5p\rho^2, \quad \forall i \neq j.$$

Therefore, with probability at least $1 - 4N\exp\{-p/128\}$, it follows that

$$\ell(\|\mathbf{x}_j - \mathbf{x}_i\|; \rho) = \min\left(\frac{\|\mathbf{x}_j - \mathbf{x}_i\|^2}{p\rho^2} - 2.5, 0\right) = 0, \forall i \neq j.$$

Therefore, with probability at least $1 - 4N\exp\{-p/128\}$, the loss satisfies

$$L(\mathbf{x}_j;\rho) = \sum_{i=1}^{N} \ell\left(\|\mathbf{x}_j - \mathbf{x}_i\|;\rho\right) = -F,$$

since $\ell\left(\|\mathbf{x}_j - \mathbf{x}_j\|;\rho\right) = \ell(0;\rho) = -F$. □

Next, it is proven that, with high probability, the loss value of a positive is less than $-F$.

Proposition 3. *Given a set X of N samples from a GMM with outliers, randomly select a set S of $|S| = n$ subsamples from it, then with $\sigma_{max} \leq \rho < \sqrt{0.6}$, then for a positive sample $\mathbf{x}_j \in S_k, l(\mathbf{x}_j) = k > 0$, then with probability at least $1 - 2\exp\{-p/128\} - \exp\{-(N-1)w_k\}$, the loss is bounded as $L(\mathbf{x}_j;\rho) < -F$.*

Proof. The probability that a sample of size $N-1$ contains no elements from cluster S_k is

$$(1 - w_k)^{N-1} \leq \exp\{-(N-1)w_k\}.$$

Therefore, with probability at least $1 - \exp\{-(N-1)w_k\}$, there is at least one more sample $\mathbf{x}_a \in X, a \neq j$ besides $\mathbf{x}_j$ in cluster S_k.
From Corollary 4, with probability at least $1 - 2\exp\{-p/128\}$, the distance between $\mathbf{x}_a$ and $\mathbf{x}_j$ is bounded as

$$\|\mathbf{x}_j - \mathbf{x}_a\|^2 < 2.5p\sigma_k^2.$$

Given $\sigma_{max} \leq \rho < \sqrt{0.6}$, with probability at least $1 - 2\exp\{-p/128\}$, the distance between $\mathbf{x}_a$ and $\mathbf{x}_j$ is bounded as

$$\|\mathbf{x}_j - \mathbf{x}_a\|^2 < 2.5p\sigma_k^2 \leq 2.5p\sigma_{max}^2 \leq 2.5p\rho^2.$$

Therefore, with probability at least $1 - 2\exp\{-p/128\} - \exp\{-(N-1)w_k\}$, the following equality holds

$$\ell(\|\mathbf{x}_j - \mathbf{x}_a\|;\rho) = \min\left(\frac{\|\mathbf{x}_j - \mathbf{x}_a\|^2}{p\rho^2} - 2.5, 0\right) < 0.$$

Therefore, with probability at least $1 - 2\exp\{-p/128\} - \exp\{-(N-1)w_k\}$, the loss is bounded above as

$$L(\mathbf{x}_j;\rho) = \sum_{i=1}^{N} \ell\left(\|\mathbf{x}_j - \mathbf{x}_i\|;\rho\right) \leq \ell\left(\|\mathbf{x}_j - \mathbf{x}_a\|;\rho\right) + \ell\left(\|\mathbf{x}_j - \mathbf{x}_j\|;\rho\right) < -F.$$

□

Proposition 4. *Given a set X of N samples from a GMM with outliers, with $w_i \geq a/m, i = \overline{1,m}$ for some $a > 0$ and $\sigma_{max} \leq \rho < \sqrt{0.6}$, randomly select a set S of $|S| = n$ subsamples from it, then with probability at least $1 - 2m\exp\{-p/128\} - m\exp\{-na/m\} - m\exp\{-a(N-1)/m\}$ for each $k = \overline{1,m}$ there exists $\mathbf{x}_j \in S_k = \{\mathbf{x} \in S, l(\mathbf{x}) = k\}$ such that $L(\mathbf{x}_j; \rho) < -F$.*

Proof. According to Lemma 3, the probability that a sample S of size n contains at least one observation from each cluster is

$$1 - \sum_{i=1}^{m}(1-w_i)^n \geq 1 - m(1-a/m)^n \geq 1 - m\exp\{-na/m\},$$

and without loss of generality let $\mathbf{x}_j$ be the observation from cluster $S_k, k = \overline{1,m}$. Applying Proposition 3 repeatedly to these m samples and using the union bound, with probability at least

$$1 - 2m\exp\{-p/128\} - \sum_{i=1}^{m}\exp\{-(N-1)w_i\},$$

the loss is bounded as $L(\mathbf{x}_j; \rho) < -F$.

Since $\forall w_i \geq a/m$, therefore $\sum_{i=1}^{m}\exp\{-(N-1)w_i\} \leq m\exp\{-a(N-1)/m\}$. Therefore, with probability at least

$$1 - m\exp\{-na/m\} - 2m\exp\{-p/128\} - m\exp\{-a(N-1)/m\},$$

for each $k = \overline{1,m}$ there exists $\mathbf{x}_j \in S_k$ such that $L(\mathbf{x}_j; \rho) < -F$. □

3.3.3 Accuracy Guarantees

The following theorem guarantees that Algorithm 1 (SCRLM) can detect all outliers and cluster all positives correctly with high probability.

Theorem 1. *Given a set X of N samples from a GMM with outliers, with $w_i \geq a/m, i = \overline{1,m}$ for some $a > 0$ and $\sigma_{max} \leq \rho < \sqrt{0.6}$, then the SCRLM Algorithm 1 that selects n subsamples has 100% accuracy with probability at least $1 - 10N^2\exp\{-p/128\} - m\exp\{-na/m\} - 2m\exp\{-p/128\} - m\exp\{-a(N-1)/m\}$.*

Proof. From Proposition 2, with probability at least $1 - 4N\exp\{-p/128\}$, for a negative sample $\mathbf{x}_j$, $L(\mathbf{x}_j; \rho) = -F$, then for all the negatives, with probability at least $1 - 4N^2\exp\{-p/128\}$, the loss satisfies $L(\mathbf{x}_j; \rho) = -F$.

From Proposition 4, with probability at least

$$1 - m\exp\{-na/m\} - 2m\exp\{-p/128\} - m\exp\{-a(N-1)/m\},$$

for each $k = \overline{1,m}$ there is $\mathbf{x}_j \in S_k, L(\mathbf{x}_j;\rho) < -F$.

Combining Proposition 2 and Proposition 4, with probability at least

$$1 - 4N^2 \exp\{-p/128\} - m\exp\{-na/m\} - 2m\exp\{-p/128\} - m\exp\{-a(N-1)/m\},$$

only positives will be selected at step 8 of SCRLM.

From Proposition 1, with probability at least $1 - 6N^2\exp\{-p/128\}$, all positives are correctly identified in Steps 9 and 17 and removed from negatives.

So with probability at least

$$1 - 10N^2 \exp\{-p/128\} - m\exp\{-na/m\} - 2m\exp\{-p/128\} - m\exp\{-a(N-1)/m\},$$

SCRLM will have 100% accuracy. □

Based on Theorem 1, Corollary 7 establishes theoretical bounds for parameters p and n.

Corollary 7. *Given a set X of N samples from a GMM with outliers, with $w_i \geq a/m, i = \overline{1,m}$ for some $a > 0$ and $\sigma_{max} \leq \rho < \sqrt{0.6}$, then for any δ, if*

$$p > 128(2\log N + \log\frac{40}{\delta}),$$

$$n > \frac{m}{a}(\log m + \log\frac{4}{\delta}),$$

$$p > 128(\log m + \log\frac{8}{\delta}),$$

$$N > \frac{m}{a}(\log m + \log\frac{4}{\delta}) + 1,$$

the SCRLM Algorithm 1 that selects n subsamples will have 100% accuracy with probability at least $1 - \delta$.

Proof. The condition

$$p > 128(2\log N + \log\frac{40}{\delta})$$

is equivalent to

$$10N^2\exp\{-p/128\} < \frac{\delta}{4}.$$

The condition

$$n > \frac{m}{a}(\log 4m - \log\delta)$$

is equivalent to

$$m\exp(-na/m) < \frac{\delta}{4}.$$

The condition

$$p > 128(\log 8m - \log \delta)$$

is equivalent to

$$2m \exp\{-p/128\} < \frac{\delta}{4}.$$

Finally, the condition

$$N > \frac{m}{a}(\log 4m - \log \delta) + 1.$$

is equivalent to:

$$m \exp\{-a(N-1)/m\} < \frac{\delta}{4}.$$

These conditions together imply that

$$1 - 10N^2 \exp\{-p/128\} - m \exp(-na/m) - 2m \exp\{-p/128\} - m \exp\{-a(N-1)/m\} > 1 - \delta.$$

According to Theorem 1, SCRLM has 100% accuracy with probability at least $1 - \delta$. □

3.4 Computational Complexity

Computing $L(\mathbf{x}_j; \rho), j \in S$ in Step 4 of Algorithm 1 is $O(nNp)$. Each iteration of steps 6-12 is $O(np)$, so steps 5-13 take $O(nmp)$. Similarly, steps 14-21 take $O(Nmp)$. Therefore, the computation complexity of Algorithm 1 is $O(nNp + nmp + Nmp) = O(nNp + Nmp)$. From Corollary 7 one could see that the subsample size n should be chosen on the order of $O(m \log m)$. Therefore, the computational complexity of Algorithm 1 is $O(mpN \log m)$, it is linear in the dimension p and the number of observations N and log-linear in the number of clusters m.

CHAPTER 4

HIERARCHICAL SCALABLE CLUSTERING BY ROBUST LOSS MINIMIZATION IN HIERARCHICAL GAUSSIAN MIXTURE MODEL WITH OUTLIERS

4.1 Problem Formulation

A Hierarchical Gaussian Mixture Model (HGMM) with outliers is a probabilistic model that assumes data is generated from a multi-level hierarchical structure of Gaussian mixture components. In the model, each level of the hierarchy represents a different granularity of the data distribution. At every level, the data is described by a mixture of Gaussian components and outlier components. At each level, the Gaussian components capture the characteristics of the data, while the outlier components account for noise. For simplicity, this dissertation focuses on a two-level HGMM with outliers.

Given a set $X = \mathbf{x}_i \in \mathbb{R}^p, i = 1, \ldots, N$ of N points sampled from a two-level HGMM with outliers, the model has m_1 Gaussian mixture components at the first level of the hierarchy, and each component at the first level is further decomposed into m_2 mixture components at the second level of the hierarchy. The goal is to group these points into $m = m_1 \times m_2$ compact subsets.

The probability density function (pdf) of a two-level HGMM with outliers can be expressed as:

$$p(\mathbf{x} \mid \Theta) = \sum_{i=1}^{m_1} w_i \left(\sum_{j=1}^{m_2} w_{ij} \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{ij}, \Sigma_{ij}) + w_{i,-1} O_1(\mathbf{x}) \right) + w_{-1} O(\mathbf{x}) \tag{4.1}$$

where $\mathbf{x} \in \mathbb{R}^p$ is a p-dimensional data point, m_1 is the number of Gaussian components at the first level, and m_2 is the number of Gaussian subcomponents within each first-level Gaussian component. The weights w_i and w_{ij} represent the mixing proportions of the Gaussian components at the first and second levels, respectively. Outliers are introduced at the first level with the components denoted by $O(\mathbf{x})$ and its mixing proportion represented by w_{-1}. The second-level outlier components are denoted by $O_1(\mathbf{x})$, with their mixing proportions represented by $w_{i,-1}$. The Gaussian distributions

at the second level are $\mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_{ij}, \Sigma_{ij}\right), i = 1, \ldots, m_1, j = 1, \ldots, m_2$ where

$$\mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_{ij}, \Sigma_{ij}\right) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_{ij}, \sigma_{ij}^2\right) = \frac{1}{(2\pi)^{p/2}\left|\sigma_{ij}^2 I_p\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_{ij}\right)^T \left(\sigma_{ij}^2 I_p\right)^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_{ij}\right)\right\} \tag{4.2}$$

with mean vector $\boldsymbol{\mu}_{ij}$ and covariance matrix $\Sigma_{ij} = \sigma_{ij}^2 I_p$.

Let $l(\mathbf{x}) \in \{-1 \cup (i,j) \mid i \in \{1, 2, \ldots, m_1\}, j \in \{-1, 1, 2, \ldots, m_2\}\}$ be the label of observation $\mathbf{x}$. The samples $\mathbf{x}_i$ with $l(\mathbf{x}_i) = -1$ are called first-level outliers (negatives), the samples $\mathbf{x}_i$ with $l(\mathbf{x}_i) = (i, -1)$ are called second-level outliers (negatives) belonging to the first-level positive cluster i and the samples $\mathbf{x}_i$ with $l(\mathbf{x}_i) = (i, j)$ where $j > 0$ are called second-level positives.

The two-level HGMM with outliers (Figure 4.1) used in this dissertation is generated as follows:

1. The first-level cluster centers $\boldsymbol{\mu}_i, i = 1, \ldots, m_1$ and first-level negatives are generated from $O(\mathbf{x}) = \mathcal{N}(\mathbf{0}, I_p)$ with weights $w_i, i = -1, 1, \ldots, m_1$.

2. The second-level cluster centers $\boldsymbol{\mu}_{ij}, i = 1, \ldots, m_1, j = 1, \ldots, m_2$ and second-level negatives are generated from $O_1(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i)$ with weights $w_{ij}, i = 1, \ldots, m_1, j = -1, 1, \ldots, m_2$ around each first-level center $\boldsymbol{\mu}_i$.

3. Then, second-level positives are generated from $\mathcal{N}(\boldsymbol{\mu}_{ij}, \sigma_{ij})$, $i = 1, \ldots, m_1, j = -1, 1, \ldots, m_2$.

The problem of interest is to cluster a set of unlabeled observations generated from such a two-level hierarchical Gaussian Mixture Model with outliers and recover the labels $l(\mathbf{x}_i)$ and $\boldsymbol{\mu}_i$.

4.2 Hierarchical Scalable Clustering by Robust Loss Minimization

Algorithm 2 Hierarchical Scalable Clustering by Robust Loss Minimization (HSCRLM)

1: **Input:** $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^p$, the number of subsamples n_1 and n_2, the bandwidth parameter ρ_1 and ρ_2, the desired number of cluster T.
2: **Output:** the number of first-level clusters m_1, the number of second-level clusters m_2 per first-level, cluster labels $l_1, \ldots, l_N \in \{-1, (1,-1), (1,1), (1,2), \ldots, (m_1, m_2)\}$.
3: Apply SCRLM(X, n_1, N, ρ_1, T) to set X to obtain $l_i, i = 1, \ldots, N$
4: **for** $i = 1$ to m_1 **do**
5: Let $I_i = \{j \in \overline{1, N}, l_j = i\}$ and denote $X_i = X_{I_i}$.
6: Apply SCRLM$(X_i, n_2, |X_i|, \rho_2, T)$ to obtain $q_j, j \in I_i$.
7: $l_j = (i, q_j), j \in I$
8: **end for**

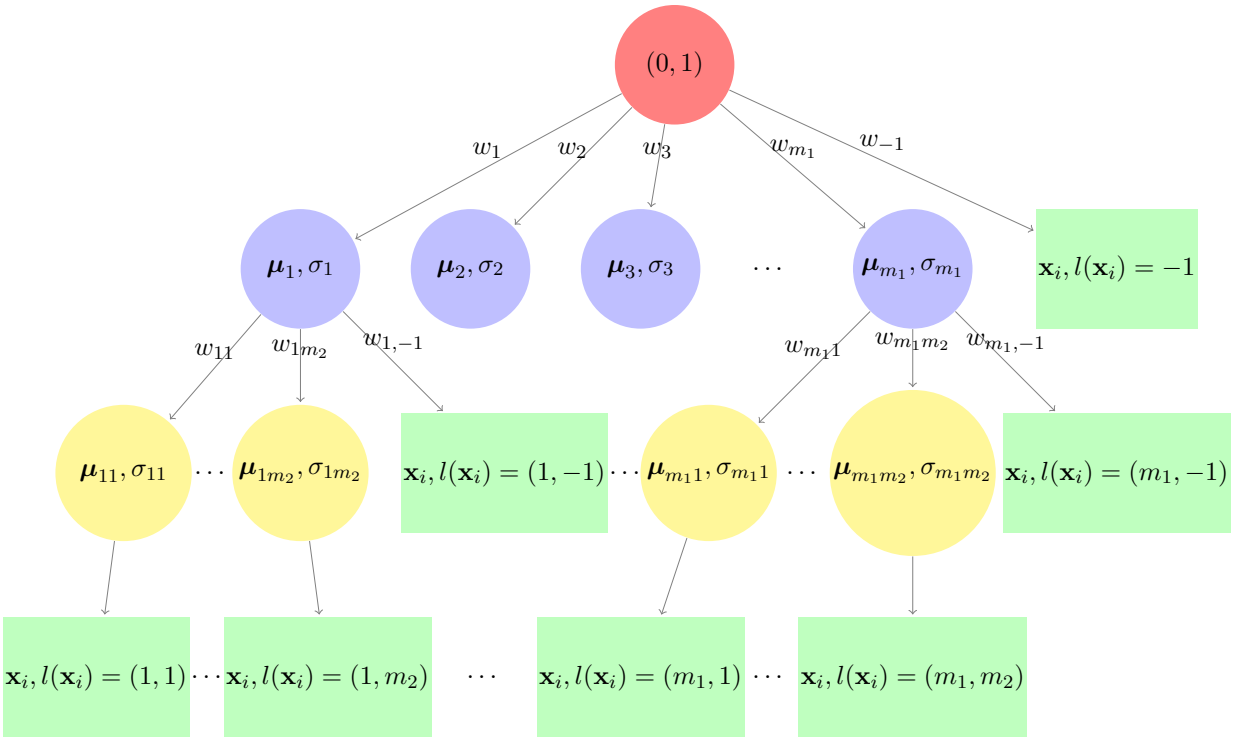


Figure 4.1: Structure of the two-level HGMM with outliers used in this dissertation.

4.3 Hierarchical Classification

Once the first-level and second-level cluster centers have been obtained through the preceding Algorithm 2, the subsequent Algorithm 3 presents a powerful approach for the efficient hierarchical classification of data points. This methodology offers a substantial reduction in computational complexity compared to exhaustive searching, especially when dealing with a large number of cluster centers. In contrast to the computationally intensive process of computing distances between the input vector and all cluster centers, Algorithm 3 adopts a judicious strategy. It begins by discerningly selecting the top k closest first-level clusters, chosen based on their proximity to the input vector, thereby effectively pruning less relevant clusters from consideration. This initial step alone significantly diminishes the computational burden. Subsequently, within these selected first-level clusters, a refinement procedure is applied to pinpoint the nearest second-level cluster center. This two-tiered selection process further enhances computational efficiency, as it reduces the number of distance calculations required.

Algorithm 3 Hierarchical Classification

Input: $\mathbf{x}$, the number of first-level clusters m_1, the number of second-level clusters per first-level cluster m_2, the first-level cluster centers $\boldsymbol{\mu}_i$ for $i = 1, \ldots, m_1$, the second-level cluster centers $\boldsymbol{\mu}_{ij}$ for $i = 1, \ldots, m_1, j = 1, \ldots, m_2$.
Output: label of $\mathbf{x}, l(\mathbf{x}) \in \{-1, (1,-1), (1,1), (1,2), \ldots, (m_1, m_2)\}$.
Compute $k = \underset{i}{\operatorname{argmin}} \|\mathbf{x} - \boldsymbol{\mu}_i\|$
if $\|\mathbf{x} - \boldsymbol{\mu}_k\| > \rho_1\sqrt{pF}$ **then**
    $l(\mathbf{x}) = -1$
else
    Find indices J of top k closest super class centers: $J = \{i_1, i_2, ..., i_k\}$ corresponding to the first k smallest distances, let $U = J \times \{1, ..., m_2\}$, compute $(i, j) = \underset{(i,j)\in U}{\operatorname{argmin}} \|\mathbf{x} - \boldsymbol{\mu}_{ij}\|$
    if $||\mathbf{x} - \boldsymbol{\mu}_{ij}|| > \rho_2\sqrt{pF}$ **then**
        $l(\mathbf{x}) = (i, -1)$
    else
        $l(\mathbf{x}) = (i, j)$
    end if
end if

4.4 Theoretical Guarantees

In this section, two theorems are proven:

1. SCRLM obtains high accuracy with high probability under Assumption 2 in the two-level HGMM with outliers.
2. HSCRLM obtains high accuracy with high probability under Assumption 3 and 4 in the two-level HGMM with outliers.

4.4.1 Assumptions

Let $\sigma_{max,2} = \max_{i,j>0} \sigma_{ij}$ represent the maximum standard deviation among all second-level positive clusters, and $\sigma_{min,1} = \min_{i>0} \sigma_i$ denote the minimum standard deviation among all first-level positive clusters. Based on these values, the following are the main assumptions required for the theoretical guarantees to hold.

Assumption 2. *$\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, where ρ is the bandwidth parameter for the loss function* (3.3).

Assumption 3. *$\sqrt{\sigma_{max,1}^2 + \sigma_{max,2}^2} \leq \rho_1 < \sqrt{0.6}$, where ρ_1 is the bandwidth parameter for the loss function* (3.3).

Assumption 4. *$\sigma_{max,2} \leq \rho_2 < \sqrt{0.6}\sigma_{min,1}$, where ρ_2 is the bandwidth parameter for the loss function* (3.3).

4.4.2 Preliminaries

Corollary 8 (Separation between first-level negatives)**.** *For two first-level negatives $\mathbf{x}_i$ and $\mathbf{x}_k$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > 1.5p, \quad \mathbf{x}_i, \mathbf{x}_k \in H_1.$$

Proof. Same as Corollary 3 □

Corollary 9 (Separation between second-level negatives from different first-level clusters)**.** *For two second-level negatives $\mathbf{x}_i$ and $\mathbf{x}_k$ from different first-level clusters, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > p(1.5 + 0.75(\sigma_i^2 + \sigma_j^2)), \quad \mathbf{x}_i \in H_{2i}, \mathbf{x}_k \in H_{2j}.$$

Proof. Same as Corollary 6. □

Corollary 10 (Separation between second-level negatives from the same first-level cluster)**.** *For two second-level negatives $\mathbf{x}_i$ and $\mathbf{x}_k$ from the same first-level cluster, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > 1.5p\sigma_i^2, \quad \mathbf{x}_i, \mathbf{x}_k \in H_{2i}.$$

Proof. Since $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 I_p)$ and $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 I_p)$, then $\mathbf{x}_i - \mathbf{x}_k \sim \mathcal{N}(0, 2\sigma_i^2 I_p)$. According to Corollary 1, it follows that

$$\mathbb{P}\left(\left|\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2p\sigma_i^2} - 1\right| \geq \epsilon\right) \leq 2\exp\left\{-p\epsilon^2/8\right\},$$

then

$$\mathbb{P}(\|\mathbf{x}_i - \mathbf{x}_k\|^2 \leq 2p\sigma_i^2(1-\epsilon)) \leq 2\exp\{-p\epsilon^2/8\}.$$

Take $\epsilon = 1/4$, yields

$$\mathbb{P}(\|\mathbf{x}_i - \mathbf{x}_k\|^2 \leq 1.5p\sigma_i^2) \leq 2\exp\{-p/128\}.$$

Therefore, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > 1.5p\sigma_i^2.$$

□

Corollary 11 (Separation between first-level negatives and second-level negatives)**.** *For first-level negative* $\mathbf{x}_i$ *second-level negative* $\mathbf{x}_k$*, with probability at least* $1 - 2\exp\{-p/128\}$*, the separations satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > p(1.5 + 0.75\sigma_i^2), \quad \mathbf{x}_i \in H_1, \mathbf{x}_k \in H_{2i}.$$

Proof. Same as Corollary 5. □

Corollary 12 (Concentration of second-level positives in the same cluster)**.** *For any second-level positive cluster* S_{ij} *with mean* $\boldsymbol{\mu}_{ij}$ *and covariance matrix* $\sigma_{ij}^2 I_p$*, with probability at least* $1 - 2\exp\{-p/128\}$*, the concentration satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 < 2.5p\sigma_{ij}^2, \quad \mathbf{x}_i, \mathbf{x}_k \in S_{ij}.$$

Proof. Since $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_{ij}, \sigma_{ij}^2 I_p)$ and $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_{ij}, \sigma_{ij}^2 I_p)$, then $\mathbf{x}_i - \mathbf{x}_k \sim \mathcal{N}(0, 2\sigma_{ij}^2 I_p)$, thus $\mathbb{E}\left(\|\mathbf{x}_i - \mathbf{x}_k\|^2\right) = 2p\sigma_{ij}^2$. According to Corollary 1, it follows that

$$\mathbb{P}\left(\left|\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2p\sigma_{ij}^2} - 1\right| \geq \epsilon\right) \leq 2\exp\left\{-p\epsilon^2/8\right\},$$

then

$$\mathbb{P}(\|\mathbf{x}_i - \mathbf{x}_k\|^2 \geq 2p\sigma_{ij}^2(1+\epsilon)) \leq 2\exp\{-p\epsilon^2/8\}.$$

Take $\epsilon = 1/4$, yields

$$\mathbb{P}(\|\mathbf{x}_i - \mathbf{x}_k\|^2 \geq 2.5p\sigma_{ij}^2) \leq 2\exp\{-p/128\}.$$

Therefore, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 < 2.5p\sigma_{ij}^2.$$

□

Corollary 13 (Separation between second-level positives from the same first-level cluster)**.** *For second-level positive* $\mathbf{x}_{ij}$ *from cluster* S_{ij} *with true mean* $\{\boldsymbol{\mu}_{ij}$ *and covariance matrix* $\sigma_{ij}^2 I_p$ *and second-level positive* $\mathbf{x}_{ik}$ *from another cluster* S_{ik} *with true mean* $\boldsymbol{\mu}_{ik}$ *and covariance matrix* $\sigma_{ik}^2 I_p$*, where cluster* S_{ij} *and* S_{ij} *are both from the same first-level cluster* S_i *with true mean* $\boldsymbol{\mu}_i$ *and covariance matrix* $\sigma_i^2 I_p$*, with probability at least* $1 - 2\exp\{-p/128\}$*, the separation satisfies*

$$\|\mathbf{x}_{ij} - \mathbf{x}_{ik}\|^2 > p(1.5\sigma_i^2 + 0.75(\sigma_{ij}^2 + \sigma_{ik}^2)), \quad \mathbf{x}_{ij} \in S_{ij}, \mathbf{x}_{ik} \in S_{ik}$$

Proof. Since $\mathbf{x}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}_{ij}, \sigma_{ij}^2 I_p)$, $\mathbf{x}_{ik} \sim \mathcal{N}(\boldsymbol{\mu}_{ik}, \sigma_{ik}^2 I_p)$, then $\mathbf{x}_{ij} - \mathbf{x}_{ik} \sim \mathcal{N}(\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_{ik}, \sigma_{ij}^2 I_p + \sigma_{ik}^2 I_p)$, thus $\mathbf{x}_{ij} - \mathbf{x}_{ik} = \boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_{ik} + \boldsymbol{\epsilon_1}\sqrt{\sigma_{ij}^2 + \sigma_{ik}^2}$ with $\boldsymbol{\epsilon_1} \sim \mathcal{N}(\mathbf{0}, I_p)$. Since $\boldsymbol{\mu}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 I_p)$, $\boldsymbol{\mu}_{ik} \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 I_p)$, then $\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_{ik} \sim \mathcal{N}\left(\mathbf{0}, 2\sigma_i^2 I_p\right)$, then $\mathbf{x}_{ij} - \mathbf{x}_{ik}$ is a Gaussian with $\mathbb{E}\left(\mathbf{x}_{ij} - \mathbf{x}_{ik}\right) = \mathbf{0}$ and

$$\mathbb{E}\left(\|\mathbf{x}_{ij} - \mathbf{x}_{ik}\|^2\right) = \mathbb{E}\left[\left(\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_{ik} + \boldsymbol{\epsilon_1}\sqrt{\sigma_{ij}^2 + \sigma_{ik}^2}\right)^T \left(\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_{ik} + \boldsymbol{\epsilon_1}\sqrt{\sigma_{ij}^2 + \sigma_{ik}^2}\right)\right],$$

thus

$$\begin{aligned}
\mathbb{E}\left(\|\mathbf{x}_{ij} - \mathbf{x}_{ik}\|^2\right) &= \mathbb{E}(\|\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_{ik}\|^2) + (\sigma_{ij}^2 + \sigma_{ik}^2)\mathbb{E}\left(\boldsymbol{\epsilon_1}^T \boldsymbol{\epsilon_1}\right) \\
&= 2\sigma_i^2 p + (\sigma_{ij}^2 + \sigma_{ik}^2)\mathbb{E}\left(\|\boldsymbol{\epsilon_1}\|^2\right) \\
&= (2\sigma_i^2 + \sigma_{ij}^2 + \sigma_{ik}^2)p.
\end{aligned}$$

According to Corollary 1, the probabilities are bounded

$$\mathbb{P}\left(\left|\frac{\|\mathbf{x}_{ij} - \mathbf{x}_{ik}\|^2}{2\sigma_i^2 p + \sigma_{ij}^2 p + \sigma_{ik}^2 p} - 1\right| \geq \epsilon\right) \leq 2\exp\{-p\epsilon^2/8\},$$

then

$$\mathbb{P}\left(\|\mathbf{x}_{ij} - \mathbf{x}_{ik}\|^2 \leq p(2\sigma_i^2 + \sigma_{ij}^2 + \sigma_{ik}^2)(1-\epsilon)\right) \leq 2\exp\left\{-p\epsilon^2/8\right\}.$$

Then with high probability at least $1 - 2\exp\{-p\epsilon^2/8\}$, the separation satisfies

$$\|\mathbf{x}_{ij} - \mathbf{x}_{ik}\|^2 > p(2\sigma_i^2 + \sigma_{ij}^2 + \sigma_{ik}^2)(1-\epsilon).$$

Now take $\epsilon = 1/4$ so that with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_{ij} - \mathbf{x}_{ik}\|^2 > p(1.5\sigma_i^2 + 0.75\sigma_{ij}^2 + 0.75\sigma_{ik}^2).$$

□

Corollary 14 (Separation between second-level positives from different first-level clusters)**.** *For second-level positive $\mathbf{x}_{ik}$ from cluster S_{ik} with true mean $\boldsymbol{\mu}_{ik}$ and covariance matrix $\sigma_{ik}^2 I_p$ and second-level positive $\mathbf{x}_{jk}$ from another cluster S_{jk} with true mean $\boldsymbol{\mu}_{jk}$ and covariance matrix $\sigma_{jk}^2 I_p$, where cluster S_{ik} is from the first-level cluster S_i with true mean $\boldsymbol{\mu}_i$ and covariance matrix $\sigma_i^2 I_p$, and S_{jk} from the first-level cluster S_j with true mean $\boldsymbol{\mu}_j$ and covariance matrix $\sigma_j^2 I_p$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies*

$$\|\mathbf{x}_{ik} - \mathbf{x}_{jk}\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2 + 0.75\sigma_{ij}^2 + 0.75\sigma_{ik}^2), \quad \mathbf{x}_{ik} \in S_{ik}, \mathbf{x}_{jk} \in S_{jk}$$

Proof. Since $\mathbf{x}_{ik} \sim \mathcal{N}(\boldsymbol{\mu}_{ik}, \sigma_{ik}^2 I_p)$, $\mathbf{x}_{jk} \sim \mathcal{N}(\boldsymbol{\mu}_{jk}, \sigma_{jk}^2 I_p)$. then $\mathbf{x}_{ik} - \mathbf{x}_{jk} \sim \mathcal{N}(\boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{jk}, \sigma_{ik}^2 I_p + \sigma_{jk}^2 I_p)$, thus $\mathbf{x}_{ik} - \mathbf{x}_{jk} = \boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{jk} + \boldsymbol{\epsilon_1}\sqrt{\sigma_{ik}^2 + \sigma_{jk}^2}$ with $\boldsymbol{\epsilon_1} \sim \mathcal{N}(\mathbf{0}, I_p)$. Since $\boldsymbol{\mu}_{ik} \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 I_p)$, $\boldsymbol{\mu}_{jk} \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma_j^2 I_p)$, then $\boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{jk} \sim \mathcal{N}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j, \sigma_i^2 I_p + \sigma_j^2 I_p\right)$, thus $\boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{jk} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j + \boldsymbol{\epsilon_2}\sqrt{\sigma_i^2 + \sigma_j^2}$ with $\boldsymbol{\epsilon_2} \sim \mathcal{N}(\mathbf{0}, I_p)$. So $\mathbf{x}_{ik} - \mathbf{x}_{jk} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j + \boldsymbol{\epsilon_1}\sqrt{\sigma_{ik}^2 + \sigma_{jk}^2} + \boldsymbol{\epsilon_2}\sqrt{\sigma_i^2 + \sigma_j^2}$. Since $\mu_i \sim \mathcal{N}(\mathbf{0}, I_p)$ and $\mu_j \sim \mathcal{N}(\mathbf{0}, I_p)$, then $\mathbf{x}_{ik} - \mathbf{x}_{jk}$ is a Gaussian with $\mathbb{E}(\mathbf{x}_{ij} - \mathbf{x}_{ik}) = \mathbf{0}$ and $\mathbb{E}\left(\|\mathbf{x}_{ik} - \mathbf{x}_{jk}\|^2\right) = \mathbb{E}\left[\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j + \boldsymbol{\epsilon_1}\sqrt{\sigma_{ij}^2 + \sigma_{ik}^2} + \boldsymbol{\epsilon_2}\sqrt{\sigma_i^2 + \sigma_j^2}\right)^T \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j + \boldsymbol{\epsilon_1}\sqrt{\sigma_{ij}^2 + \sigma_{ik}^2} + \boldsymbol{\epsilon_2}\sqrt{\sigma_i^2 + \sigma_j^2}\right)\right]$, thus

$$\mathbb{E}\left(\|\mathbf{x}_{ik} - \mathbf{x}_{jk}\|^2\right) = (2 + \sigma_i^2 + \sigma_j^2 + \sigma_{ij}^2 + \sigma_{ik}^2)p.$$

According to Corollary 1, the probabilities are bounded

$$\mathbb{P}\left(\left|\frac{\|\mathbf{x}_{ik} - \mathbf{x}_{jk}\|^2}{2p + \sigma_i^2 p + \sigma_j^2 p + \sigma_{ik}^2 p + \sigma_{jk}^2 p} - 1\right| \geq \epsilon\right) \leq 2\exp\{-p\epsilon^2/8\},$$

then

$$\mathbb{P}\left(\|\mathbf{x}_{ik} - \mathbf{x}_{jk}\|^2 \leq (1-\epsilon)(\sigma_{ik}^2 p + \sigma_{jk}^2 p + \sigma_i^2 + +\sigma_j^2 + 2p)\right) \leq 2\exp\{-p\epsilon^2/8\}$$

Now take $\epsilon = 1/4$ so that with probability at least $1 - 2\exp\{-p/128\}$, the bound becomes

$$\|\mathbf{x}_{ik} - \mathbf{x}_{jk}\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2 + 0.75\sigma_{ij}^2 + 0.75\sigma_{ik}^2)$$

□

Corollary 15 (Separation between first-level negatives and second-level positives)**.** *For second-level negative $\mathbf{x}_{ij}$ and first-level negative $\mathbf{x}_k$, with probability at least $1 - 2\exp\{-p/128\}$, the separations satisfies*

$$\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_{ij}^2)$$

Proof. Since $\mathbf{x}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}_{ij}, \sigma_{ij}^2 I_p)$ and $\mathbf{x}_k \sim \mathcal{N}(\mathbf{0}, I_p)$, then $\mathbf{x}_{ij} - \mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_{ij}, \sigma_{ij}^2 I_p + I_p)$, thus $\mathbb{E}\left(\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2\right) = \mathbb{E}(\|\boldsymbol{\mu}_{ij}\|^2) + (\sigma_{ij}^2 + 1)p = \mathbb{E}(\|\boldsymbol{\mu}_i\|^2) + \sigma_i^2 p + (\sigma_{ij}^2 + 1)p = 2p + \sigma_i^2 p + \sigma_{ij}^2 p$. According to Corollary 1,

$$\mathbb{P}\left(\left|\frac{\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2}{2p + \sigma_i^2 p + \sigma_{ij}^2 p} - 1\right| \geq \epsilon\right) \leq 2\exp\left\{-p\epsilon^2/8\right\}$$

so

$$\mathbb{P}\left(\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2 \leq (2p + \sigma_i^2 p + \sigma_{ij}^2 p)(1-\epsilon)\right) \leq 2\exp\left\{-p\epsilon^2/8\right\}$$

Then with probability at least $1 - 2\exp\{-p\epsilon^2/8\}$, the separation satisfies

$$\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2 > (2p + \sigma_i^2 p + \sigma_{ij}^2 p)(1-\epsilon)$$

Now take $\epsilon = 1/4$ so that with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_{ij}^2)$$

□

Corollary 16 (Separation between second-level negatives and second-level positives from the same first-level cluster)**.** *For second-level negative* $\mathbf{x}_{ij}$ *and second-level negative* $\mathbf{x}_k$*, with probability at least* $1 - 2\exp\{-p/128\}$*, the separations satisfies*

$$\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2 > p(1.5\sigma_i^2 + 0.75\sigma_{ij}^2)$$

Proof. Since $\mathbf{x}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}_{ij}, \sigma_{ij}^2 I_p)$ and $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 I_p)$, then $\mathbf{x}_{ij} - \mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_i, \sigma_{ij}^2 I_p + \sigma_i^2 I_p)$, thus $\mathbb{E}\left(\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2\right) = \mathbb{E}(\|\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_i\|^2) + (\sigma_{ij}^2 + \sigma_i^2)p$.Since $\boldsymbol{\mu}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 I_p)$, then $\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_i \sim \mathcal{N}(0, \sigma_i^2 I_p)$, thus $\mathbb{E}(\|\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_i\|^2) = \sigma_i^2 p$, then $\mathbb{E}\left(\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2\right) = \sigma_i^2 p + (\sigma_{ij}^2 + \sigma_i^2)p = 2\sigma_i^2 p + \sigma_{ij}^2 p$. According to Corollary 1,

$$\mathbb{P}\left(\left|\frac{\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2}{(2\sigma_i^2 + \sigma_{ij}^2)p} - 1\right| \geq \epsilon\right) \leq 2\exp\left\{-p\epsilon^2/8\right\}$$

so

$$\mathbb{P}\left(\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2 \leq (2\sigma_i^2 p + \sigma_{ij}^2 p)(1 - \epsilon)\right) \leq 2\exp\left\{-p\epsilon^2/8\right\}$$

Then with probability at least $1 - 2\exp\{-p\epsilon^2/8\}$, the separation satisfies

$$\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2 > (2\sigma_i^2 p + \sigma_{ij}^2 p)(1 - \epsilon)$$

Now take $\epsilon = 1/4$ so that with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > p(1.5\sigma_i^2 + 0.75\sigma_{ij}^2)$$

□

Corollary 17 (Separation between second-level negatives and second-level positives from different first-level clusters)**.** *For second-level positive* $\mathbf{x}_{ij}$ *and second-level negative* $\mathbf{x}_k$*, with probability at least* $1 - 2\exp\{-p/128\}$*, the separations satisfies*

$$\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2 + 0.75\sigma_{ij}^2)$$

Proof. Since $\mathbf{x}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}_{ij}, \sigma_{ij}^2 I_p)$ and $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma_j^2 I_p)$, then $\mathbf{x}_{ij} - \mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_j, \sigma_{ij}^2 I_p + \sigma_j^2 I_p)$, thus $\mathbb{E}\left(\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2\right) = \mathbb{E}(\|\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_j\|^2) + (\sigma_{ij}^2 + \sigma_j^2)p = \mathbb{E}(\|\boldsymbol{\mu}_i\|^2) + \sigma_i^2 p + p + (\sigma_{ij}^2 + \sigma_j^2)p = (2 + \sigma_i^2 + \sigma_j^2 + \sigma_{ij}^2)p$. According to Corollary 1,

$$\mathbb{P}\left(\left|\frac{\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2}{2p + \sigma_i^2 p + + \sigma_i^2 p + \sigma_{ij}^2 p} - 1\right| \geq \epsilon\right) \leq 2\exp\left\{-p\epsilon^2/8\right\}$$

so

$$\mathbb{P}\left(\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2 \leq (2p + \sigma_i^2 p + + \sigma_i^2 p + \sigma_{ij}^2 p)(1 - \epsilon)\right) \leq 2\exp\left\{-p\epsilon^2/8\right\}$$

Then with probability at least $1 - 2\exp\{-p\epsilon^2/8\}$, the separation satisfies

$$\|\mathbf{x}_{ij} - \mathbf{x}_k\|^2 > (2p + \sigma_i^2 p + + \sigma_j^2 p + \sigma_{ij}^2 p)(1 - \epsilon)$$

Now take $\epsilon = 1/4$ so that with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2 + 0.75\sigma_{ij}^2)$$

□

Corollary 18 (Concentration of second-level negatives)**.** *For second-level negatives* $\mathbf{x}_i$ *and* $\mathbf{x}_k$ *from the same first-level cluster, with probability at least* $1 - 2\exp\{-p/128\}$*, the concentration satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 < 2.5p\sigma_j^2, \quad \mathbf{x}_i, \mathbf{x}_k \in H_{2i}.$$

Proof. Same as Corollary 4 □

Corollary 19 (Concentration of second-level positives)**.** *For second-level positive* $\mathbf{x}_{ij}$ *from cluster* S_{ij} *with true mean* $\boldsymbol{\mu}_{ij}$ *and covariance matrix* $\sigma_{ij}^2 I_p$ *and second-level positive* $\mathbf{x}_{ik}$ *from another cluster* S_{ik} *with true mean* $\boldsymbol{\mu}_{ik}$ *and covariance matrix* $\sigma_{ik}^2 I_p$*, where cluster* S_{ij} *and* S_{ij} *are both from the same first-level cluster* S_i *with true mean* $\boldsymbol{\mu}_i$ *and covariance matrix* $\sigma_i^2 I_p$*, with probability at least* $1 - 2\exp\{-p/128\}$*, the concentration satisfies*

$$\|\mathbf{x}_{ij} - \mathbf{x}_{ik}\|^2 < p(2.5\sigma_i^2 + 1.25(\sigma_{ij}^2 + \sigma_{ik}^2)), \quad \mathbf{x}_{ij} \in S_{ij}, \mathbf{x}_{ik} \in S_{ik}$$

Proof. Since $\mathbf{x}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}_{ij}, \sigma_{ij}^2 I_p)$, $\mathbf{x}_{ik} \sim \mathcal{N}(\boldsymbol{\mu}_{ik}, \sigma_{ik}^2 I_p)$, then

$$\mathbb{E}\left(\|\mathbf{x}_{ij} - \mathbf{x}_{ik}\|^2\right) = \mathbb{E}(\|\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_{ik}\|^2) + (\sigma_{ij}^2 + \sigma_{ik}^2)p = 2\sigma_i^2 p + (\sigma_{ij}^2 + \sigma_{ik}^2)p = (2\sigma_i^2 + \sigma_{ij}^2 + \sigma_{ik}^2)p.$$

According to Corollary 1, the probabilities are bounded

$$\mathbb{P}\left(\left|\frac{\|\mathbf{x}_{ij}-\mathbf{x}_{ik}\|^2}{2\sigma_i^2+\sigma_{ij}^2p+\sigma_{ik}^2p}-1\right|\geq\epsilon\right)\leq 2\exp\{-p\epsilon^2/8\}$$

$$\mathbb{P}\left(\frac{\|\mathbf{x}_{ij}-\mathbf{x}_{ik}\|^2}{2\sigma_i^2+\sigma_{ij}^2p+\sigma_{ik}^2p}\geq 1+\epsilon\right)\leq 2\exp\{-p\epsilon^2/8\}$$

so

$$\mathbb{P}\left(\|\mathbf{x}_{ij}-\mathbf{x}_{ik}\|^2\geq(1+\epsilon)(2\sigma_i^2+\sigma_{ij}^2p+\sigma_{ik}^2p)\right)\leq 2\exp\{-p\epsilon^2/8\}$$

Therefore, with probability at least $1-4\exp\{-p\epsilon^2/8\}$, the concentration is bounded below as

$$\|\mathbf{x}_{ij}-\mathbf{x}_{ik}\|^2<p(2\sigma_i^2+\sigma_{ij}^2+\sigma_{ik}^2)(1+\epsilon)$$

Now take $\epsilon=1/4$ so that with probability at least $1-2\exp\{-p/128\}$, the bound becomes

$$\|\mathbf{x}_{ij}-\mathbf{x}_{ik}\|^2<p(2.5\sigma_i^2+1.25(\sigma_{ij}^2+\sigma_{ik}^2))$$

□

Corollary 20 (Concentration of second-level negatives and second-level positives from the same first-level cluster)**.** *For second-level positive* $\mathbf{x}_i$ *and second-level negative* $\mathbf{x}_k$ *from the same first-level cluster, with probability at least* $1-2\exp\{-p/128\}$*, the concentration satisfies*

$$\|\mathbf{x}_i-\mathbf{x}_k\|^2<p(2.5\sigma_i^2+1.25\sigma_{ij}^2)$$

Proof. Since $\mathbf{x}_i\sim\mathcal{N}(\boldsymbol{\mu}_{ij},\sigma_{ij}^2I_p)$ and $\mathbf{x}_k\sim\mathcal{N}(\boldsymbol{\mu}_i,\sigma_i^2I_p)$, then $\mathbf{x}_i-\mathbf{x}_k\sim\mathcal{N}(\boldsymbol{\mu}_{ij}-\boldsymbol{\mu}_i,\sigma_{ij}^2I_p+\sigma_i^2I_p)$, thus $\mathbb{E}\left(\|\mathbf{x}_i-\mathbf{x}_k\|^2\right)=(\sigma_{ij}^2+2\sigma_i^2)p$. According to Corollary 1,

$$\mathbb{P}\left(\left|\frac{\|\mathbf{x}_i-\mathbf{x}_k\|^2}{(\sigma_{ij}^2+2\sigma_i^2)p}-1\right|\geq\epsilon\right)\leq 2\exp\left\{-p\epsilon^2/8\right\}$$

so

$$\mathbb{P}\left(\|\mathbf{x}_i-\mathbf{x}_k\|^2\geq(\sigma_{ij}^2p+2\sigma_i^2p)(1+\epsilon)\right)\leq 2\exp\left\{-p\epsilon^2/8\right\}$$

Then with probability at least $1-2\exp\{-p\epsilon^2/8\}$, the concentration satisfies

$$\|\mathbf{x}_i-\mathbf{x}_k\|^2<(\sigma_{ij}^2p+2\sigma_i^2p)(1+\epsilon)$$

Now take $\epsilon=1/4$ so that with probability at least $1-2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i-\mathbf{x}_k\|^2<p(2.5\sigma_i^2+1.25\sigma_{ij}^2)$$

□

Proposition 5. *Given a set X of N samples from a two-level HGMM with outliers and $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then with probability at least $1 - 12N^2 \exp\{-p/128\}$, the distance between second-level positives within a cluster satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \ s.t. \ l(\mathbf{x}_i) = l(\mathbf{x}_j) = (a, b), a > 0, b > 0,$$

and the distance between second-level positives from a cluster and other samples not in that cluster satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2.5p\rho^2 \quad \forall \mathbf{x}_i, \mathbf{x}_j \ s.t. \ l(\mathbf{x}_j) \neq l(\mathbf{x}_i).$$

Proof. From Corollary 12, with probability at least $1 - 2\exp\{-p/128\}$, the distance between two positives in the same cluster is bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\sigma_{ij}^2, \quad l(\mathbf{x}_i) = l(\mathbf{x}_j) = (a, b), a > 0, b > 0.$$

Using the union bound, with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distances between all positives in the same cluster are bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\sigma_{ij}^2 \leq 2.5p\sigma_{max,2}^2 \leq 2.5p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = l(\mathbf{x}_j) = (a, b), a > 0, b > 0.$$

From Corollary 13, with probability at least $1 - 2\exp\{-p/128\}$, the distance between two positives from the same first cluster but different second-level clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5\sigma_i^2 + 0.75(\sigma_{ij}^2 + \sigma_{ik}^2)), \quad l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (a, d), b \neq d, a, b, d > 0.$$

Using the union bound, with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any two positives from the same first cluster but different second-level clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5\sigma_i^2 + 0.75(\sigma_{ij}^2 + \sigma_{ik}^2)), \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (a, d), b \neq d, a, b, d > 0.$$

Given $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any two positives from the same first cluster but different second-level clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 1.5p\sigma_{min,1}^2 > 2.5p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (a, d), b \neq d, a, b, d > 0.$$

From Corollary 14, with probability at least $1 - 2\exp\{-p/128\}$, the distance between two second-level positives from different first clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2 + 0.75\sigma_{ij}^2 + 0.75\sigma_{ik}^2),$$

where

$$l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (c, d), a \neq c, a, b, c, d > 0.$$

Using the union bound, with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any two second-level positives from different first clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2 + 0.75\sigma_{ij}^2 + 0.75\sigma_{ik}^2),$$

where

$$\forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (c, d), a, b, c, d > 0.$$

Given $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any two second-level positives from different first clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 1.5p > 2.5p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (c, d), a, b, c, d > 0.$$

From Corollary 15, with probability at least $1 - 2\exp\{-p/128\}$, the distance between a second-level positive and a first-level negative satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_{ij}^2), \quad l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = -1.$$

Using the union bound, with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any second-level positive and any first-level negative satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_{ij}^2), \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = -1, a, b > 0.$$

Given $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any second-level positive and any first-level negative satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 1.5p > 2.5p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = -1, a, b > 0.$$

From Corollary 16, with probability at least $1 - 2\exp\{-p/128\}$, the distance between a second-level positive and a second-level negative from the same first-level cluster satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5\sigma_i^2 + 0.75\sigma_{ij}^2), \quad l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (a, -1), a, b > 0.$$

Using the union bound, with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any second-level positive and any second-level negative from the same first-level cluster satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5\sigma_i^2 + 0.75\sigma_{ij}^2), \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (a, -1), a, b > 0.$$

Given $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any second-level positive and any second-level negative from the same first-level cluster satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 1.5p\sigma_{min,1}^2 > 2.5p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (a, -1), a, b > 0.$$

From Corollary 17, with probability at least $1 - 2\exp\{-p/128\}$, the distance between a second-level positive and a second-level negative from different first-level clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2 + 0.75\sigma_{ij}^2), \quad l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (c, -1), a, b, c > 0, a \neq c.$$

Using the union bound, with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any second-level positive and second-level negative from different first-level clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2 + 0.75\sigma_{ij}^2),$$

where

$$\forall \mathbf{x}_i, \mathbf{x}_j, \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (c, -1), a, b, c > 0, a \neq c.$$

Given $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any second-level positive and second-level negative from different first-level clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 1.5p > 2.5p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j, \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (c, -1), a, b, c > 0, a \neq c.$$

Therefore, with probability at least $1 - 10N^2 \exp\{-p/128\}$, the distance between any positive and any sample not from that cluster satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2.5p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) \neq l(\mathbf{x}_i) = (a, b).$$

Therefore, with probability at least $1 - 12N^2 \exp\{-p/128\}$, the following bounds on second-level positives within a cluster and between clusters are satisfied

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2.5p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) \neq l(\mathbf{x}_i) = (a, b), a > 0, b > 0,$$

and

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) = l(\mathbf{x}_i) = (a, b), a > 0, b > 0.$$

□

4.4.3 Loss Bounds

Proposition 6. *Given a set X of N samples from a two-level HGMM with outliers and $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then for a first-level negative sample $\mathbf{x}_j, l(\mathbf{x}_j) = -1$, with probability at least $1 - 6N\exp\{-p/128\}$, the loss satisfies $L(\mathbf{x}_j; \rho) = -F$.*

Proof. From Corollary 8, for $\mathbf{x}_i, l(\mathbf{x}_i) = -1$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p, \quad l(\mathbf{x}_i) = -1, i \neq j.$$

Using the union bound, with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p > 2.5p\rho^2, \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = -1, i \neq j.$$

From Corollary 11, for $\mathbf{x}_i, l(\mathbf{x}_i) = (a, -1), a > 0$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2)$$

Using the union bound, with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5 + 0.75\sigma_i^2) > 2.5p\rho^2, \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (a, -1), m > 0.$$

From Corollary 15, for $\mathbf{x}_i, l(\mathbf{x}_i) = (a, b), a, b > 0$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_{ij}^2)$$

Using the union bound, with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_{ij}^2) > 1.5p > 2.5p\rho^2, \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (a, b), a, b > 0.$$

Therefore, with probability at least $1 - 6N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 2.5p\rho^2, \quad \forall i \neq j.$$

Therefore, with probability at least $1 - 6N\exp\{-p/128\}$, it follows that

$$\ell(\|\mathbf{x}_j - \mathbf{x}_i\|; \rho) = \min\left(\frac{\|\mathbf{x}_j - \mathbf{x}_i\|^2}{p\rho^2} - 2.5, 0\right) = 0, \forall i \neq j.$$

Therefore, with probability at least $1 - 6N\exp\{-p/128\}$, the loss satisfies

$$L(\mathbf{x}_j; \rho) = \sum_{i=1}^{N} \ell\left(\|\mathbf{x}_j - \mathbf{x}_i\|; \rho\right) = -F,$$

since $\ell\left(\|\mathbf{x}_j - \mathbf{x}_j\|; \rho\right) = \ell(0; \rho) = -F$. □

Proposition 7. *Given a set X of N samples from a two-level HGMM with outliers and $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then for a second-level negative sample $\mathbf{x}_j, l(\mathbf{x}_j) = (a, -1), a > 0$, with probability at least $1 - 8N\exp\{-p/128\}$, the loss satisfies $L(\mathbf{x}_j; \rho) = -F$.*

Proof. From Corollary 9, for $\mathbf{x}_i, l(\mathbf{x}_i) = (c, -1), c > 0, a \neq c$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5 + 0.75(\sigma_i^2 + \sigma_j^2)), \quad l(\mathbf{x}_i) = (c, -1), a \neq c.$$

Using the union bound, with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5 + 0.75(\sigma_i^2 + \sigma_j^2)), \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (c, -1), a \neq c.$$

Given $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p > 2.5p\rho^2, \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (c, -1), a \neq c.$$

From Corollary 10, for $\mathbf{x}_i, l(\mathbf{x}_i) = (a, -1)$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p\sigma_i^2, \quad l(\mathbf{x}_i) = (a, -1), i \neq j.$$

Using the union bound, with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p\sigma_i^2, \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (a, -1), i \neq j.$$

Given $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p\sigma_{min,1}^2 > 2.5p\rho^2 \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (a, -1), i \neq j.$$

From Corollary 16, for $\mathbf{x}_i, l(\mathbf{x}_i) = (a, b), b > 0$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5\sigma_i^2 + 0.75\sigma_{ij}^2), \quad l(\mathbf{x}_i) = (a, b), b > 0.$$

Using the union bound, with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5\sigma_i^2 + 0.75\sigma_{ij}^2), \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (a, b), b > 0.$$

Given $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p > 2.5p\rho^2 \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (a, b), a, b > 0.$$

From Corollary 17, for $\mathbf{x}_i, l(\mathbf{x}_i) = (c,d), c,d > 0$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2 + 0.75\sigma_{ij}^2) \quad l(\mathbf{x}_i) = (c,d), c,d > 0.$$

Using the union bound, with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2 + 0.75\sigma_{ij}^2), \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (c,d), c,d > 0.$$

Given $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p > 2.5p\rho^2 \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (c,d), c,d > 0.$$

Therefore, with probability at least $1 - 8N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 2.5p\rho^2, \quad \forall i \neq j.$$

Therefore, with probability at least $1 - 8N\exp\{-p/128\}$, it follows that

$$\ell(\|\mathbf{x}_j - \mathbf{x}_i\|; \rho) = \min\left(\frac{\|\mathbf{x}_j - \mathbf{x}_i\|^2}{p\rho^2} - 2.5, 0\right) = 0, \forall i \neq j.$$

Therefore, with probability at least $1 - 8N\exp\{-p/128\}$, the loss satisfies

$$L(\mathbf{x}_j; \rho) = \sum_{i=1}^{N} \ell\left(\|\mathbf{x}_j - \mathbf{x}_i\|; \rho\right) = -F,$$

since $\ell\left(\|\mathbf{x}_j - \mathbf{x}_j\|; \rho\right) = \ell(0; \rho) = -F$. □

Proposition 8. *Given a set X of N samples from a two-level HGMM with outliers and $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then for a second-level positive sample $\mathbf{x}_j, l(\mathbf{x}_j) = (a,b), a > 0, b > 0$, with probability at least $1 - 2\exp\{-p/128\} - \exp\{-(N-1)w_i w_{ij}\}$, the loss is bounded as $L(\mathbf{x}_j; \rho) < -F$.*

Proof. The probability that a sample of size $N-1$ contains no elements from cluster S_{ij} is

$$(1 - w_i w_{ij})^{N-1} \leq \exp\{-(N-1)w_i w_{ij}\}.$$

Therefore, with probability at least $1 - \exp\{-(N-1))w_i w_{ij}\}$, there is at least one more sample $\mathbf{x}_a, a \neq j$ besides $\mathbf{x}_j$ in cluster S_{ij}.

From Corollary 12, with probability at least $1 - 2\exp\{-p/128\}$, the distance between $\mathbf{x}_a$ and $\mathbf{x}_j$ is bounded as

$$\|\mathbf{x}_j - \mathbf{x}_a\|^2 < 2.5p\sigma_{ij}^2.$$

Given $\sigma_{max,2} \leq \rho$, with probability at least $1 - 2\exp\{-p/128\}$, the distance between $\mathbf{x}_a$ and $\mathbf{x}_j$ is bounded as

$$\|\mathbf{x}_j - \mathbf{x}_a\|^2 < 2.5p\sigma_{ij}^2 \leq 2.5p\sigma_{max,2}^2 \leq 2.5p\rho^2.$$

Therefore, with probability at least $1-2\exp\{-p/128\}-\exp\{-(N-1)w_iw_{ij}\}$, the following equality holds

$$\ell(\|\mathbf{x}_j - \mathbf{x}_a\|;\rho) = \min\left(\frac{\|\mathbf{x}_j - \mathbf{x}_a\|^2}{p\rho^2} - 2.5, 0\right) < 0.$$

Therefore, with probability at least $1 - 2\exp\{-p/128\} - \exp\{-(N-1)w_iw_{ij}\}$, the loss is bounded above as

$$L(\mathbf{x}_j;\rho) = \sum_{i=1}^{N} \ell\left(\|\mathbf{x}_j - \mathbf{x}_i\|;\rho\right) \leq \ell\left(\|\mathbf{x}_j - \mathbf{x}_a\|;\rho\right) + \ell\left(\|\mathbf{x}_j - \mathbf{x}_j\|;\rho\right) < -F.$$

□

Proposition 9. *Given a set X of N samples from a two-level HGMM with outliers, with $w_i \geq a_1/m_1, i = \overline{1,m_1}$ for some $a_1 > 0$, $w_{ij} \geq a_2/m_2, j = \overline{1,m_2}$ for some $a_2 > 0$ and $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, randomly select a set S of $|S| = n$ subsamples from it, then with probability at least $1 - m\exp\{-na_1a_2/m\} - 2m\exp\{-p/128\} - m\exp\{-a_1a_2(N-1)/m\}$ for each $i = \overline{1,m_1}, j = \overline{1,m_2}$ there exists $\mathbf{x}_k \in S_{ij} = \{\mathbf{x} \in S, l(\mathbf{x}) = (i,j)\}$ such that $L(\mathbf{x}_k;\rho) < -F$.*

Proof. According to Lemma 3, the probability that a sample S of size n contains at least one observation from each cluster is

$$1 - \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(1 - w_iw_{ij})^n \geq 1 - m_1m_2(1 - a_1a_2/m_1m_2)^n \geq 1 - m\exp\{-na_1a_2/m\},$$

and without loss of generality let $\mathbf{x}_k$ be the observation from cluster $S_{ij}, i = \overline{1,m_1}, j = \overline{1,m_2}$. Applying Proposition 8 repeatedly to m samples and using the union bound, with probability at least $1-2m\exp\{-p/128\}-\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\exp\{-(N-1)w_iw_{ij}\}$, the loss is bounded as $L(\mathbf{x}_j;\rho) < -F$. Since $\forall w_i \geq a_1/m_1, \forall w_{ij} \geq a_2/m_2$, therefore $\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\exp\{-(N-1)w_iw_{ij}\} \leq m\exp\{-a_1a_2(N-1)/m\}$. Therefore, with probability at least

$$1 - m\exp\{-na_1a_2/m\} - 2m\exp\{-p/128\} - m\exp\{-a_1a_2(N-1)/m\},$$

for each $i = \overline{1,m_1}, j = \overline{1,m_2}$ there exists $\mathbf{x}_j \in S_{ij}$ such that $L(\mathbf{x}_j;\rho) < -F$. □

4.4.4 Accuracy Guarantees

Theorem 2. *Given a set X of N samples from a two-level HGMM with outliers, with $w_i \geq a_1/m_1, i = \overline{1, m_1}$ for some $a_1 > 0$, $w_{ij} \geq a_2/m_2, j = \overline{1, m_2}$ for some $a_2 > 0$ and $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then SCRLM Algorithm 1 that selects n subsamples has 100% accuracy with probability at least $1-26N^2\exp\{-p/128\}-m\exp\{-na_1a_2/m\}-2m\exp\{-p/128\}-m\exp\{-a_1a_2(N-1)/m\}$.*

Proof. Combining Proposition 6 and Proposition 7, for a negative sample $\mathbf{x}_j$, with probability at least $1 - 14N\exp\{-p/128\}$, $L(\mathbf{x}_j;\rho) = -F$, then for all the negatives, with probability at least $1 - 14N^2\exp\{-p/128\}$, the loss satisfies $L(\mathbf{x}_j;\rho) = -F$.
From Proposition 9, with probability at least

$$1 - m\exp\{-na_1a_2/m\} - 2m\exp\{-p/128\} - m\exp\{-a_1a_2(N-1)/m\}$$

there is $\mathbf{x}_k \in S_{ij}, L(\mathbf{x}_j;\rho) < -F$.
Combining Proposition 6, Proposition 7 and Proposition 9, with probability at least

$$1 - 14N^2\exp\{-p/128\} - m\exp\{-na_1a_2/m\} - 2m\exp\{-p/128\} - m\exp\{-a_1a_2(N-1)/m\},$$

only positives will be selected at step 8 of SCRLM.
From Proposition 5, with probability at least $1 - 12N^2\exp\{-p/128\}$, all positives are correctly identified in Steps 9 and 17 and removed from negatives.
So with probability at least

$$1 - 26N^2\exp\{-p/128\} - m\exp\{-na_1a_2/m\} - 2m\exp\{-p/128\} - m\exp\{-a_1a_2(N-1)/m\},$$

SCRLM will have 100% accuracy. □

Based on Theorem 2, Corollary 21 establishes theoretical bounds for parameters p and n.

Corollary 21. *Given a set X of N samples from a two-level HGMM with outliers, with $w_i \geq a_1/m_1, w_{ij} \geq a_2/m_2, i = \overline{1, m_1}, j = \overline{1, m_2}$ for some $a_1, a_2 > 0$ and $\sigma_{max,2} \leq \rho < \sqrt{0.6}\sigma_{min,1}$, then for any δ, if*

$$p > 128(2\log N + \log\frac{104}{\delta}),$$
$$n > \frac{m}{a_1a_2}(\log m + \log\frac{4}{\delta}),$$

the SCRLM Algorithm 1 that selects n subsamples will have 100% accuracy with probability at least $1 - \delta$.

Proof. The condition

$$p > 128(2\log N + \log\frac{104}{\delta})$$

is equivalent to

$$26N^2\exp\{-p/128\} < \frac{\delta}{4}.$$

The condition

$$n > \frac{m}{a_1 a_2}(\log 4m - \log\delta)$$

is equivalent to

$$m\exp(-na_1a_2/m) < \frac{\delta}{4}.$$

The condition

$$p > 128(\log 8m - \log\delta)$$

is equivalent to

$$2m\exp\{-p/128\} < \frac{\delta}{4}.$$

Finally, the condition

$$N > \frac{m}{a_1 a_2}(\log 4m - \log\delta) + 1.$$

is equivalent to:

$$m\exp\{-a_1a_2(N-1)/m\} < \frac{\delta}{4}.$$

Since $n < N$ and $m < N$, then

$$p > 128(2\log N + \log\frac{104}{\delta}) > 128(\log 8m - \log\delta)$$

$$N > n > \frac{m}{a_1 a_2}(\log 4m - \log\delta).$$

These conditions together imply that

$$1 - 26N^2\exp\{-p/128\} - m\exp(-na_1a_2/m) - 2m\exp\{-p/128\} - m\exp\{-a_1a_2(N-1)/m\} > 1-\delta.$$

According to Theorem 2, SCRLM has 100% accuracy with probability at least $1-\delta$. □

Proposition 10. *Given a set X of N samples from a two-level HGMM with outliers and $\sqrt{\sigma^2_{max,1} + \sigma^2_{max,2}} \le \rho_1 < \sqrt{0.6}$, then with probability at least $1 - 18N^2\exp\{-p/128\}$, the distance between first-level positives within a cluster satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \ s.t.\ l(\mathbf{x}_i) = l(\mathbf{x}_j) = (a, \sim), a > 0,$$

and the distance between first-level positives from a cluster and other samples not in that cluster satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2.5p\rho_1^2 \quad \forall \mathbf{x}_i, \mathbf{x}_j \ s.t. \ l(\mathbf{x}_j) \neq l(\mathbf{x}_i).$$

Proof. From Corollary 12, with probability at least $1 - 2\exp\{-p/128\}$, the distance between two second-level positives in the same cluster is bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\sigma_{ij}^2, \quad l(\mathbf{x}_i) = l(\mathbf{x}_j) = (a, b), a, b > 0.$$

Using the union bound, with probability at least $1 - 2N^2\exp\{-p/128\}$, the distances between all second-level positives in the same cluster are bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\sigma_{ij}^2 < 2.5p\sigma_{max,2}^2 \leq 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = l(\mathbf{x}_j) = (a, b), a, b > 0.$$

From Corollary 18, with probability at least $1 - 2\exp\{-p/128\}$, the distance between two second-level negatives in the same first-level cluster is bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\sigma_j^2, \quad l(\mathbf{x}_i) = l(\mathbf{x}_j) = (a, -1), a > 0.$$

Using the union bound, with probability at least $1 - 2N^2\exp\{-p/128\}$, the distances between all second-level negatives in the same first-level cluster are bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\sigma_j^2 < 2.5p\sigma_{max,1}^2 \leq 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = l(\mathbf{x}_j) = (a, -1), a > 0.$$

From Corollary 19, with probability at least $1 - 2\exp\{-p/128\}$, the distance between two positives from the same first-level cluster but different second-level clusters is bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < p(2.5\sigma_i^2 + 1.25(\sigma_{ij}^2 + \sigma_{ik}^2)), \quad l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (a, d), b \neq d, a, b, d > 0.$$

Using the union bound, with probability at least $1 - 2N^2\exp\{-p/128\}$, the distances between two positives from the same first-level cluster but different second-level clusters are bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p(\sigma_{max,1}^2 + \sigma_{max,2}^2) \leq 2.5p\rho_1^2,$$

where

$$\forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (a, d), b \neq d, a, b, d > 0.$$

From Corollary 20, with probability at least $1 - 2\exp\{-p/128\}$, the distance between two positives from the same first cluster but different second-level clusters is bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p(\sigma_i^2 + 0.5\sigma_{ij}^2), \quad l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (a, -1), a, b > 0.$$

Using the union bound, with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any second-level positives and second-level negatives from the same first cluster are bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p(\sigma^2_{max,1} + \sigma^2_{max,2}) \leq 2.5p\rho_1^2, \quad l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (a, -1), a, b > 0.$$

Therefore, with probability at least $1 - 8N \exp\{-p/128\}$, the distance between two first-level positive is bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) = (a, \sim), l(\mathbf{x}_i) = (a, \sim). \tag{4.3}$$

Using the union bound, with probability at least $1 - 8N^2 \exp\{-p/128\}$, the distance between any first-level positives are bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) = (a, \sim), l(\mathbf{x}_i) = (a, \sim).$$

From Corollary 9, with probability at least $1-2\exp\{-p/128\}$, the distance between two second-level negatives from different first clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2), \quad l(\mathbf{x}_i) = (a, -1), l(\mathbf{x}_j) = (c, -1), a \neq c, a, c > 0.$$

Using the union bound, with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any two second-level negatives from different first clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 1.5p > 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, -1), l(\mathbf{x}_j) = (c, -1), a, c > 0.$$

From Corollary 11, with probability at least $1 - 2\exp\{-p/128\}$, the distance between second-level negative and first-level negative satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2), \quad l(\mathbf{x}_i) = (a, -1), l(\mathbf{x}_j) = -1, a > 0.$$

Using the union bound, with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distances between any second-level negatives and first-level negatives satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 1.5p > 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, -1), l(\mathbf{x}_j) = -1, a > 0.$$

From Corollary 14, with probability at least $1 - 2\exp\{-p/128\}$, the distance between two second-level positives from different first clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2 + 0.75\sigma_{ij}^2 + 0.75\sigma_{ik}^2),$$

where

$$l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (c, d), a \neq c, a, b, c, d > 0.$$

Using the union bound, with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any two second-level positives from different first clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 1.5p > 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (c, d), a, b, c, d > 0.$$

From Corollary 15, with probability at least $1 - 2\exp\{-p/128\}$, the distance between a second-level positive and a first-level negative satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_{ij}^2), \quad l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = -1.$$

Using the union bound, with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any second-level positive and any first-level negative satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 1.5p > 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = -1, a, b > 0.$$

From Corollary 17, with probability at least $1 - 2\exp\{-p/128\}$, the distance between a second-level positive and a second-level negative from different first-level clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_j^2 + 0.75\sigma_{ij}^2), \quad l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (c, -1), a, b, c > 0, a \neq c.$$

Using the union bound, with probability at least $1 - 2N^2 \exp\{-p/128\}$, the distance between any second-level positive and second-level negative from different first-level clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 1.5p > 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j, \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (c, -1), a, b, c > 0, a \neq c.$$

Therefore, with probability at least $1 - 10N^2 \exp\{-p/128\}$, the distance between any first-level positive and any sample not from that cluster satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) \neq l(\mathbf{x}_i) = (a, \sim).$$

Therefore, with probability at least $1 - 18N^2 \exp\{-p/128\}$, the following bounds on first-level positives within a cluster and between clusters are satisfied

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2.5p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) \neq l(\mathbf{x}_i) = (a, \sim), a > 0,$$

and

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) = l(\mathbf{x}_i) = (a, \sim), a > 0.$$

□

Proposition 11. *Given a set X of N samples from a two-level HGMM with outliers and $\sqrt{\sigma_{max,1}^2+\sigma_{max,2}^2} \le \rho_1 < \sqrt{0.6}$, then for a first-level negative sample $\mathbf{x}_j, l(\mathbf{x}_j) = -1$, with probability at least $1 - 6N\exp\{-p/128\}$, the loss satisfies $L(\mathbf{x}_j;\rho) = -F$.*

Proof. From Corollary 8, for $\mathbf{x}_i, l(\mathbf{x}_i) = -1$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p, \quad l(\mathbf{x}_i) = -1, i \neq j.$$

Using the union bound, with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p > 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = -1, i \neq j.$$

From Corollary 11, for $\mathbf{x}_i, l(\mathbf{x}_i) = (a,-1), a > 0$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2)$$

Using the union bound, with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5 + 0.75\sigma_i^2) > 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (a,-1), m > 0.$$

From Corollary 15, for $\mathbf{x}_i, l(\mathbf{x}_i) = (a,b), a, b > 0$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_{ij}^2)$$

Using the union bound, with probability at least $1 - 2N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5 + 0.75\sigma_i^2 + 0.75\sigma_{ij}^2) > 1.5p > 2.5p\rho_1^2, \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (a,b), a, b > 0.$$

Therefore, with probability at least $1 - 6N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 2.5p\rho_1^2, \quad \forall i \neq j.$$

Therefore, with probability at least $1 - 6N\exp\{-p/128\}$, it follows that

$$\ell(\|\mathbf{x}_j - \mathbf{x}_i\|;\rho_1) = \min\left(\frac{\|\mathbf{x}_j - \mathbf{x}_i\|^2}{p\rho_1^2} - 2.5, 0\right) = 0, \forall i \neq j.$$

Therefore, with probability at least $1 - 6N\exp\{-p/128\}$, the loss satisfies

$$L(\mathbf{x}_j;\rho_1) = \sum_{i=1}^{N} \ell\left(\|\mathbf{x}_j - \mathbf{x}_i\|;\rho_1\right) = -F,$$

since $\ell\left(\|\mathbf{x}_j - \mathbf{x}_j\|;\rho_1\right) = \ell(0;\rho_1) = -F$. □

Proposition 12. *Given a set X of N samples from a two-level HGMM with outliers and $\sqrt{\sigma_{max,1}^2+\sigma_{max,2}^2} \le \rho_1 < \sqrt{0.6}$, then for a first-level positive sample $\mathbf{x}_j$, $l(\mathbf{x}_j) = (k, \sim), k > 0$, with probability at least $1 - 8\exp\{-p/128\} - \exp\{-(N-1)w_k\}$, the loss is bounded as $L(\mathbf{x}_j; \rho_1) < -F$.*

Proof. The probability that a sample of size $N-1$ contains no elements from the first-level cluster S_k is

$$(1-w_k)^{N-1} \le \exp\{-(N-1)w_k\}.$$

Therefore, with probability at least $1 - \exp\{-(N-1))w_k\}$, there is at least one more sample $\mathbf{x}_a, a \ne j$ besides $\mathbf{x}_j$ in the first-level cluster S_k.

From (4.3), with probability at least $1-8\exp\{-p/128\}$, the distance between $\mathbf{x}_a$ and $\mathbf{x}_j$ is bounded as

$$\|\mathbf{x}_j - \mathbf{x}_a\|^2 < 2.5p\rho_1^2.$$

Therefore, with probability at least $1-8\exp\{-p/128\} - \exp\{-(N-1))w_k\}$, the following equality holds

$$\ell(\|\mathbf{x}_j - \mathbf{x}_a\|; \rho_1) = \min\left(\frac{\|\mathbf{x}_j - \mathbf{x}_a\|^2}{p\rho_1^2} - 2.5, 0\right) < 0.$$

Therefore, with probability at least $1 - 8\exp\{-p/128\} - \exp\{-(N-1))w_k\}$, the loss is bounded above as

$$L(\mathbf{x}_j; \rho_1) = \sum_{i=1}^{N} \ell(\|\mathbf{x}_j - \mathbf{x}_i\|; \rho_1) \le \ell(\|\mathbf{x}_j - \mathbf{x}_a\|; \rho_1) + \ell(\|\mathbf{x}_j - \mathbf{x}_j\|; \rho_1) < -F.$$

□

Proposition 13. *Given a set X of N samples from a two-level HGMM with outliers, with $w_k \ge a_1/m_1, k = \overline{1, m_1}$ for some $a_1 > 0$, and $\sqrt{\sigma_{max,1}^2+\sigma_{max,2}^2} \le \rho_1 < \sqrt{0.6}$, randomly select a set S' of $|S'| = n_1$ subsamples from it, then with probability at least $1 - m_1\exp\{-n_1a_1/m_1\} - 8m_1\exp\{-p/128\} - m_1\exp\{-a_1(N-1)/m_1\}$, for each $k = \overline{1, m_1}$, there exists $\mathbf{x}_j \in \{\mathbf{x} \in S', l(\mathbf{x}) = (k, \sim)\}$ such that $L(\mathbf{x}_j; \rho_1) < -F$.*

Proof. According to Lemma 3, the probability that a sample S' of size n_1 contains at least one observation from each cluster is

$$1 - \sum_{k=1}^{m_1}(1-w_k)^{n_1} \ge 1 - m_1(1-a_1/m_1)^{n_1} \ge 1 - m_1\exp\{-n_1a_1/m_1\},$$

and without loss of generality let $\mathbf{x}_j$ be the observation from cluster $S_i, i = \overline{1, m_1}$. Applying Proposition 12 repeatedly to m_1 samples and using the union bound, with probability at least

$1 - 8m_1 \exp\{-p/128\} - \sum_{i=1}^{m_1} \exp\{-(N-1)w_i\}$, the loss is bounded as $L(\mathbf{x}_j, \rho_1) < -F$.
Since $\forall w_k \geq a_1/m_1$, therefore $\sum_{k=1}^{m_1} \exp\{-(N-1)w_k\} \leq m_1 \exp\{-a_1(N-1)/m_1\}$. Therefore, with probability at least $1 - m_1 \exp\{-n_1 a_1/m_1\} - 8m_1 \exp\{-p/128\} - m_1 \exp\{-a_1(N-1)/m_1\}$, for each $k = \overline{1, m_1}$ there exists $\mathbf{x}_j \in S_i$ such that $L(\mathbf{x}_j; \rho_1) < -F$. □

Theorem 3. *Given a set X of N samples from a two-level HGMM with outliers, with $w_i \geq a_1/m_1, i = \overline{1, m_1}$ for some $a_1 > 0$, $\sqrt{\sigma_{max,1}^2 + \sigma_{max,2}^2} \leq \rho_1 < \sqrt{0.6}$, then the HSCRLM Algorithm 2 Step 3 has 100% accuracy with probability at least $1 - 24N^2 \exp\{-p/128\} - m_1 \exp\{-n_1 a_1/m_1\} - 8m_1 \exp\{-p/128\} - m_1 \exp\{-a_1(N-1)/m_1\}$.*

Proof. Combining Proposition 11, for a first-level negative sample $\mathbf{x}_j$, with probability at least $1 - 6N \exp\{-p/128\}$, $L(\mathbf{x}_j; \rho) = -F$, then for all the first-level negatives, with probability at least $1 - 6N^2 \exp\{-p/128\}$, the loss satisfies $L(\mathbf{x}_j; \rho_1) = -F$.
From Proposition 13, with probability at least

$$1 - m_1 \exp\{-n_1 a_1/m_1\} - 8m_1 \exp\{-p/128\} - m_1 \exp\{-a_1(N-1)/m_1\},$$

there is $\mathbf{x}_j \in S_i, L(\mathbf{x}_j; \rho_1) < -F$.
Combining Proposition 11 and Proposition 13, with probability at least

$$1 - 6N^2 \exp\{-p/128\} - m_1 \exp\{-n_1 a_1/m_1\} - 8m_1 \exp\{-p/128\} - m_1 \exp\{-a_1(N-1)/m_1\},$$

only first-level positives will be selected.
From Proposition 10, with probability at least $1 - 18N^2 \exp\{-p/128\}$, all first-level positives are correctly identified and removed from first-level negatives.
So with probability at least

$$1 - 24N^2 \exp\{-p/128\} - m_1 \exp\{-n_1 a_1/m_1\} - 8m_1 \exp\{-p/128\} - m_1 \exp\{-a_1(N-1)/m_1\},$$

HSCRLM Algorithm 2 Step 3 will have 100% accuracy. □

Proposition 14. *Given a set X of N samples from a two-level HGMM with outliers and $\sigma_{max,2} \leq \rho_2 < \sqrt{0.6}\sigma_{min,1}$, then with probability at least $1 - 6N^2 \exp\{-p/128\}$, the distance between second-level positives within a cluster satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\rho_2^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \ s.t. \ l(\mathbf{x}_i) = l(\mathbf{x}_j) = (a, b), a > 0, b > 0$$

and the distance between second-level positives from a cluster and other samples not in that cluster satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2.5p\rho_2^2 \quad \forall \mathbf{x}_i, \mathbf{x}_j \ s.t. \ l(\mathbf{x}_j) \neq l(\mathbf{x}_i).$$

Proof. From Corollary 12, with probability at least $1 - 2\exp\{-p/128\}$, the distance between two second-level positives in the same second-level cluster is bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\sigma_{ij}^2, \quad l(\mathbf{x}_i) = l(\mathbf{x}_j) = (a, b), a, b > 0.$$

Using the union bound, with probability at least $1 - 2N^2\exp\{-p/128\}$, the distances between all second-level positives in the same second-level cluster are bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\sigma_{ij}^2 < 2.5p\sigma_{max,2}^2 \leq 2.5p\rho_2^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = l(\mathbf{x}_j) = (a, b), a, b > 0.$$

From Corollary 16, with probability at least $1-2\exp\{-p/128\}$, the distance between a second-level positive and a second-level negative from same first-level clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5\sigma_i^2 + 0.75\sigma_{ij}^2), \quad l(\mathbf{x}_i) = (a, b, l(\mathbf{x}_j) = (a, -1), a > 0.$$

Using the union bound, with probability at least $1 - 2N^2\exp\{-p/128\}$, the distance between any second-level positives and second-level negatives from the same first clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 1.5p\sigma_{min,1}^2 > 2.5p\rho_2^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (a, -1), a > 0.$$

From Corollary 13, with probability at least $1 - 2\exp\{-p/128\}$, the distance between two second-level positives from the same first-level cluster satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(1.5\sigma_i^2 + 0.75(\sigma_{ij}^2 + \sigma_{ik}^2)), \quad l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (a, d), a, b, d > 0, b \neq d.$$

Using the union bound, with probability at least $1 - 2N^2\exp\{-p/128\}$, the distance between any second-level positives from the same first-level cluster satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 1.5p\sigma_{min,1}^2 > 2.5p\rho_2^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j, \text{ s.t. } l(\mathbf{x}_i) = (a, b), l(\mathbf{x}_j) = (a, d), a, b, d > 0, b \neq d.$$

Therefore, with probability at least $1 - 4N^2\exp\{-p/128\}$, the distance between any second-level positive and any sample not from that cluster satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2.5p\rho_2^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) \neq l(\mathbf{x}_i) = (a, b).$$

Therefore, with probability at least $1 - 6N^2 \exp\{-p/128\}$, the following bounds on second-level positives within a cluster and between clusters are satisfied

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2.5p\rho_2^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) \neq l(\mathbf{x}_i) = (a, b), a, b > 0,$$

and

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2.5p\rho_2^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) = l(\mathbf{x}_i) = (a, b), a, b > 0.$$

□

Proposition 15. *Given a set X of N samples from a two-level HGMM with outliers and $\sigma_{max,2} \leq \rho_2 < \sqrt{0.6}\sigma_{min,1}$, then for a second-level negative sample $\mathbf{x}_j, l(\mathbf{x}_j) = (a, -1), a > 0$, with probability at least $1 - 4|X_i| \exp\{-p/128\}$, the loss satisfies $L(\mathbf{x}_j; \rho_2) = -F$ for $i \in \{1, ..., m_1\}$.*

Proof. From Corollary 10, for $\mathbf{x}_i, l(\mathbf{x}_i) = (a, -1)$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p\sigma_i^2, \quad l(\mathbf{x}_i) = (a, -1), i \neq j.$$

Using the union bound, with probability at least $1 - 2N \exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p\sigma_i^2, \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (a, -1), i \neq j.$$

Given $\sigma_{max,2} \leq \rho_2 < \sqrt{0.6}\sigma_{min,1}$, then with probability at least $1 - 2N \exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p\sigma_{min,1}^2 > 2.5p\rho_2^2 \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (a, -1), i \neq j.$$

From Corollary 16, for $\mathbf{x}_i, l(\mathbf{x}_i) = (a, b), b > 0$, with probability at least $1 - 2\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5\sigma_i^2 + 0.75\sigma_{ij}^2), \quad l(\mathbf{x}_i) = (a, b), b > 0.$$

Using the union bound, with probability at least $1 - 2N \exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(1.5\sigma_i^2 + 0.75\sigma_{ij}^2), \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (a, b), b > 0.$$

Given $\sigma_{max,2} \leq \rho_2 < \sqrt{0.6}\sigma_{min,1}$, then with probability at least $1 - 2N \exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 1.5p\sigma_{min,1}^2 > 2.5p\rho_2^2 \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = (a, b), a, b > 0.$$

Therefore, with probability at least $1 - 4N\exp\{-p/128\}$, the separation satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 2.5p\rho_2^2, \quad \forall i \neq j.$$

Therefore, with probability at least $1 - 4N\exp\{-p/128\}$, it follows that

$$\ell(\|\mathbf{x}_j - \mathbf{x}_i\|; \rho_2) = \min\left(\frac{\|\mathbf{x}_j - \mathbf{x}_i\|^2}{p\rho_2^2} - 2.5, 0\right) = 0, \forall i \neq j.$$

Therefore, with probability at least $1 - 4N\exp\{-p/128\}$, the loss satisfies

$$L(\mathbf{x}_j; \rho_2) = \sum_{i=1}^{N} \ell\left(\|\mathbf{x}_j - \mathbf{x}_i\|; \rho_2\right) = -F,$$

since $\ell\left(\|\mathbf{x}_j - \mathbf{x}_j\|; \rho_2\right) = \ell(0; \rho_2) = -F$. □

Proposition 16. *Given a set X of N samples from a two-level HGMM with outliers with $N > 2n_2m_1/a$, and $\sigma_{max,2} \leq \rho_2 < \sqrt{0.6}\sigma_{min,1}$, then for a second-level positive sample from first level cluster S_i and second-level cluster S_{ij}, then with probability at least $1-2\exp\{-p/128\}-\exp\{-(N_i-1)w_{ij}\}$, the loss is bounded as $L(\mathbf{x}_j; \rho_2) < -F$ for $i \in \{1, ..., m_1\}$.*

Proof. The probability that a sample of size $N_i - 1$ contains no elements from the second-level cluster S_{ij} is

$$(1 - w_{ij})^{N_i - 1} \leq \exp\{-(N_i - 1)w_{ij}\}.$$

Therefore, with probability at least $1 - \exp\{-(N_i - 1)w_{ij}\}$, there is at least one more sample $\mathbf{x}_a, a \neq j$ besides $\mathbf{x}_j$ in the second-level cluster S_{ij}.

From Corollary 12, with probability at least $1 - 2\exp\{-p/128\}$, the distance between $\mathbf{x}_a$ and $\mathbf{x}_j$ is bounded as

$$\|\mathbf{x}_j - \mathbf{x}_a\|^2 < 2.5p\rho_2^2.$$

Therefore, with probability at least $1 - 2\exp\{-p/128\} - \exp\{-(N_i - 1)w_{ij}\}$, the following equality holds

$$\ell(\|\mathbf{x}_j - \mathbf{x}_a\|; \rho_2) = \min\left(\frac{\|\mathbf{x}_j - \mathbf{x}_a\|^2}{p\rho_2^2} - 2.5, 0\right) < 0.$$

Therefore, with probability at least $1 - 2\exp\{-p/128\} - \exp\{-(N_i - 1)w_{ij}\}$, the loss is bounded above as

$$L(\mathbf{x}_j; \rho_2) = \sum_{i=1}^{N_i} \ell\left(\|\mathbf{x}_j - \mathbf{x}_i\|; \rho_2\right) \leq \ell\left(\|\mathbf{x}_j - \mathbf{x}_a\|; \rho_2\right) + \ell\left(\|\mathbf{x}_j - \mathbf{x}_j\|; \rho_2\right) < -F.$$

□

Proposition 17. *Given a set X of N samples from a two-level HGMM with outliers, with $w_{ij} \geq a_2/m_2, j = \overline{1, m_2}$ for some $a_2 > 0$, and $\sigma_{max,2} \leq \rho_2 < \sqrt{0.6}\sigma_{min,1}$, randomly select a set S_2 of $|S_2| = n_2$ subsamples from it, then with probability at least $1 - m_2 \exp\{-n_2 a_2/m_2\} - 2m_2 \exp\{-p/128\} - m_2 \exp\{-a_2(N_i - 1)/m_2\}$ for each $j = \overline{1, m_2}$ there exists $\mathbf{x}_j \in \{\mathbf{x} \in S_2, l(\mathbf{x}) = (i, j)\}$ such that $L(\mathbf{x}_j; \rho_2) < -F$.*

Proof. According to Lemma 4, the probability that a sample S'' of size n_2 contains at least one observation from each cluster is

$$1 - \sum_{j=1}^{m_2} (1 - w_{ij})^{n_2} \geq 1 - m_2(1 - a_2/m_2)^{n_2} \geq 1 - m_2 \exp\{-n_2 a_2/m_2\},$$

and without loss of generality let $\mathbf{x}_j$ be the observation from cluster $S_{ij}, i = \overline{1, m_1}, j = \overline{1, m_2}$. Applying Proposition 17 repeatedly to m_2 samples and using the union bound, with probability at least $1 - 2m_2 \exp\{-p/128\} - \sum_{j=1}^{m_2} \exp\{-(N_i - 1)w_{ij}\}$, the loss is bounded as $L(\mathbf{x}_j; \rho_2) < -F$. Since $\forall w_{ij} \geq a_2/m_2$, therefore $\sum_{j=1}^{m_2} \exp\{-(N_i - 1)w_{ij}\} \leq m_2 \exp\{-a_2(N_i - 1)/m_2\}$. Therefore, with probability at least $1 - m_2 \exp\{-n_2 a_2/m_2\} - 2m_2 \exp\{-p/128\} - m_2 \exp\{-a_2(N_i - 1)/m_2\}$, for each $j = \overline{1, m_2}$ there exists $\mathbf{x}_j \in S_{ij}$ such that $L(\mathbf{x}_j; \rho_2) < -F$. □

Theorem 4. *Given a set X of N samples from a two-level HGMM with outliers, with $w_{ij} \geq a_2/m_2, j = \overline{1, m_2}$ for some $a_2 > 0$ and $\sigma_{max,2} \leq \rho_2 < \sqrt{0.6}\sigma_{min,1}$, then the HSCRLM Algorithm 2 Step 6 has 100% accuracy with probability at least $1 - 10N_i^2 \exp\{-p/128\} - m_2 \exp\{-n_2 a_2/m_2\} - 2m_2 \exp\{-p/128\} - m_2 \exp\{-a_2(N_i - 1)/m_2\}$.*

Proof. Combining Proposition 15, for a second-level negative sample $\mathbf{x}_j$, with probability at least $1 - 4N_i \exp\{-p/128\}$, $L(\mathbf{x}_j; \rho) = -F$, then for all the second-level negatives, with probability at least $1 - 4N_i^2 \exp\{-p/128\}$, the loss satisfies $L(\mathbf{x}_j; \rho_1) = -F$.

From Proposition 17, with probability at least

$$1 - m_2 \exp\{-n_2 a_2/m_2\} - 2m_2 \exp\{-p/128\} - m_2 \exp\{-a_2(N_i - 1)/m_2\},$$

there is $\mathbf{x}_j \in S_i, L(\mathbf{x}_j; \rho_1) < -F$.

Combining Proposition 11 and Proposition 13, with probability at least

$$1 - 4N_i^2 \exp\{-p/128\} - m_2 \exp\{-n_2 a_2/m_2\} - 2m_2 \exp\{-p/128\} - m_2 \exp\{-a_2(N_i - 1)/m_2\},$$

only second-level positives will be selected.

From Proposition 14, with probability at least $1 - 6N_i^2 \exp\{-p/128\}$, all second-level positives are

correctly identified and removed from second-level negatives.

So with probability at least

$$1 - 10N_i^2 \exp\{-p/128\} - m_2 \exp\{-n_2 a_2/m_2\} - 2m_2 \exp\{-p/128\} - m_2 \exp\{-a_2(N_i - 1)/m_2\},$$

HSCRLM Step 6 will have 100% accuracy. □

Theorem 5. *Given a set X of N samples from a two-level HGMM with outliers, with $w_i \geq a_1/m_1, w_{ij} \geq a_2/m_2, i = \overline{1, m_1}, j = \overline{1, m_2}$ for some $a_1, a_2 > 0$ and $\sqrt{\sigma_{max,1}^2 + \sigma_{max,2}^2} \leq \rho_1 < \sqrt{0.6}$, $\sigma_{max,2} \leq \rho_2 < \sqrt{0.6}\sigma_{min,1}$, then the HSCRLM Algorithm 2 Step 4-8 has 100% accuracy with probability at least $1-10N^2 \exp\{-p/128\}-m \exp\{-n_2 a_2/m_2\}-2m \exp\{-p/128\}-m \exp\{-a_2(N_i-1)/m_2\}$.*

Proof. From Lemma 4, with probability at least $1-2\exp\{-Nw_i^2/2\}$, there are at least $Nw_i/2+1$ observations in X_i. Therefore, with probability at least $1-2\exp\{-Na_1^2/2m_1^2\}-10N_i^2 \exp\{-p/128\}-m_2 \exp\{-n_2 a_2/m_2\}-2m_2 \exp\{-p/128\}-m_2 \exp\{-a_1 a_2 N/2m\}$, HSCRLM Algorithm 2 Step 6 can achieve 100% accuracy. HSCRLM Algorithm 2 Step 4-8 apply Step 6 m_1 times, Therefore, with probability $1-2m_1 \exp\{-Na_1^2/2m_1^2\}-10N^2 \exp\{-p/128\}-m \exp\{-n_2 a_2/m_2\}-2m \exp\{-p/128\}-m \exp\{-a_1 a_2 N/2m\}$, the HSCRLM Algorithm 2 Step 4-8 has 100% accuracy. □

Theorem 6. *Given a set X of N samples from a two-level HGMM with outliers, with $w_i \geq a_1/m_1, w_{ij} \geq a_2/m_2, i = \overline{1, m_1}, j = \overline{1, m_2}$ for some $a_1, a_2 > 0$ and $\sqrt{\sigma_{max,1}^2 + \sigma_{max,2}^2} \leq \rho_1 < \sqrt{0.6}$, $\sigma_{max,2} \leq \rho_2 < \sqrt{0.6}\sigma_{min,1}$, then the HSCRLM Algorithm 2 has 100% accuracy with probability at least $1 - (34N^2 + 8m_1 + 2m)\exp\{-p/128\} - m_1 \exp\{-n_1 a_1/m_1\} - m_1 \exp\{-a_1(N-1)/m_1\} - 2m_1 \exp\{-Na_1^2/2m_1^2\} - m \exp\{-n_2 a_2/m_2\} - m \exp\{-a_1 a_2 N/2m\}$.*

Proof. Combining Theorem 3 and Theorem 5, we obtain with probability at least

$$\begin{aligned} &1 - 24N^2 \exp\{-p/128\} - m_1 \exp\{-n_1 a_1/m_1\} - 8m_1 \exp\{-p/128\} \\ &- m_1 \exp\{-a_1(N-1)/m_1\} - 2m_1 \exp\{-Na_1^2/2m_1^2\} - 10N^2 \exp\{-p/128\} \\ &- m \exp\{-n_2 a_2/m_2\} - 2m \exp\{-p/128\} - m \exp\{-a_1 a_2 N/2m\}, \end{aligned}$$

the HSCRLM Algorithm 2 has 100%. Therefore, with probability at least

$$\begin{aligned} &1 - (34N^2 + 8m_1 + 2m)\exp\{-p/128\} - m_1 \exp\{-n_1 a_1/m_1\} - m_1 \exp\{-a_1(N-1)/m_1\} \\ &- 2m_1 \exp\{-Na_1^2/2m_1^2\} - m \exp\{-n_2 a_2/m_2\} - m \exp\{-a_1 a_2 N/2m\}, \end{aligned}$$

the HSCRLM Algorithm 2 has 100%. □

Corollary 22. *Given a set X of N samples from a two-level HGMM with outliers, with $w_i \geq a_1/m_1, w_{ij} \geq a_2/m_2, i = \overline{1, m_1}, j = \overline{1, m_2}$ for some $a_1, a_2 > 0$ and $\sqrt{\sigma_{max,1}^2 + \sigma_{max,2}^2} \leq \rho_1 < \sqrt{0.6}$, $\sigma_{max,2} \leq \rho_2 < \sqrt{0.6}\sigma_{min,1}$, then for any δ, if*

$$p > 128(\log(204N^2 + 48m_1 + 12m) - \log\delta)$$

$$n_1 > \frac{m_1}{a_1}(\log 6m_1 - \log\delta)$$

$$n_2 > \frac{m_2}{a_2}(\log 6m - \log\delta)$$

$$N > \frac{2m_1^2}{a_1^2}\log\frac{12m_1}{\delta}$$

$$N > \frac{2m}{a_1 a_2}\log\frac{6m}{\delta}$$

the HSCRLM Algorithm 2 will have 100% *accuracy with probability at least* $1 - \delta$.

Proof. The condition

$$p > 128(\log(204N^2 + 48m_1 + 12m) - \log\delta)$$

is equivalent to

$$(34N^2 + 8m_1 + 2m)\exp\{-p/128\} < \frac{\delta}{6}.$$

the condition

$$n_1 > \frac{m_1}{a_1}(\log 6m_1 - \log\delta)$$

is equivalent to:

$$m_1 \exp\{-n_1 a_1/m_1\} < \frac{\delta}{6}.$$

the condition

$$n_2 > \frac{m_2}{a_2}(\log 6m - \log\delta)$$

is equivalent to:

$$m\exp\{-n_2 a_2/m_2\} < \frac{\delta}{6}.$$

The condition

$$N > \frac{2m_1^2}{a_1^2}\log\frac{12m_1}{\delta}$$

is equivalent to

$$2m_1\exp\{-Na_1^2/2m_1^2\} < \frac{\delta}{6}.$$

The condition

$$N > \frac{2m}{a_1 a_2} \log \frac{6m}{\delta}$$

is equivalent to

$$m \exp\{-a_1 a_2 N/2m\} < \frac{\delta}{6}.$$

The condition

$$N > \frac{m_1}{a_1}(\log 6m_1 - \log \delta) + 1.$$

is equivalent to:

$$m_1 \exp\{-a_1(N-1)/m_1\} < \frac{\delta}{6}.$$

□

4.5 Computational Complexity

4.5.1 Computational Complexity of Algorithm 2

Step 3 takes $O(m_1 pN \log m_1)$ which is explained in Section 3.4. Steps 4-8 take $O(mpN \log m_2)$. The computational complexity of HSCRLM (Algorithm 2) is $O(m_1 Np \log m_1 + mNp \log m_2) = O(mNp \log m_2)$. If $m_1 = m_2 = \sqrt{m}$, then the computational complexity of Algorithm 2 is $O(mNp \log \sqrt{m})$.

4.5.2 Computational Complexity of Algorithm 3

Step 3 takes $O(m_1 p)$, Step 7 takes $O(km_2 p)$ The computational complexity of hierarchical classification (Algorithm 3) is $O(m_1 p + km_2 p) = O(m_1 p + kmp/m_1)$. If m_1 is set to be $\sqrt{km}$, then the total complexity is $O(\sqrt{kmp})$.

CHAPTER 5

EMPIRICAL EVALUATION

5.1 Overview

This chapter presents an empirical evaluation of the performance of SCRLM using synthetic data and real datasets from computer vision. First, the tightness of the parameter bounds given in the theoretical guarantees are evaluated using synthetic data. Then, the effectiveness of SCRLM in real applications is evaluated using five real image datasets.

To evaluate the clustering performance on synthetic and real datasets, two evaluation measures are defined for a true labeling vector $\mathbf{l} \in \{1, ..., m\}^N$ and an obtained labeling vector $\hat{\mathbf{l}} \in \{1, ..., T\}^N$.

1. Accuracy$(\mathbf{l}, \hat{\mathbf{l}}) = \frac{1}{N} \max_{\pi \in P} \sum_{i=1}^{N} I(\pi(\hat{\mathbf{l}}_i) = \mathbf{l}_i)$
2. Purity$(\mathbf{l}, \hat{\mathbf{l}}) = \frac{1}{N} \sum_{i=1}^{T} \max_j |\hat{\mathbf{l}}^{-1}(i) \cap \mathbf{l}^{-1}(j)|$

where T is the parameter specifying the maximum number of clusters allowed and $\mathbf{l}^{-1}(j) = \{i, \mathbf{l}_i = j\}$. Purity assesses the homogeneity of clusters by formalizing the process of assigning each cluster to the true class label that is most frequent within that cluster. It then quantifies this assignment by computing the ratio of correctly assigned data points to the total number of data points.

In order to assess the effectiveness of SCRLM, its performance is compared with the following clustering methods: k-means++ (Arthur and Vassilvitskii, 2007), Complete Linkage Clustering (CL) (Johnson, 1967), Spectral Clustering (SC) (Ng et al., 2002), Tensor Decomposition (TD) (Hsu and Kakade, 2013), Expectation Maximization (EM) (Dempster et al., 1977) and t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008).

For consistency in comparing the accuracy and running time, experiments use an implementation of SCRLM and the state-of-the-art algorithms in MATLAB. For k-means++, the built in function `kmeans` that implements k-means++ is used. The built-in function `linkage`, `spectralcluster` and `fitgmdist` are used for CL, SC and EM respectively. For TD, Theorem 2 from (Hsu and Kakade, 2013) has been implemented in MATLAB. For t-SNE+k-means++, the built-in `tsne` function is used to generate a matrix of two-dimensional embeddings followed by an application of k-means++ to obtain the final results. For SCRLM+k-means, k-means is applied using the initial centers obtained from SCRLM.

5.2 Simulation Experiments

This section shows experiments on synthetic data generated from a GMM with outliers described in Equation 3.1 and a HGMM with outliers described in Equation 4.1.

5.2.1 Comparison of Observed and Theoretical Accuracy of SCRLM

This section evaluates the tightness of the theoretical bounds for Algorithm 1. Synthetic data is generated with 20% outliers. The minimum and maximum weights for the positive clusters are taken to be $0.7/m$ and $0.9/m$ respectively. The standard deviations σ_i of positive clusters are linearly increasing with i from 1/16 to 1/4. The experiments use $\rho = 0.5$.

The regions for different parameter combinations where the theoretical bound guarantees of achieving 100% accuracy with at least 99% probability are compared with similar regions obtained experimentally. The theoretical regions are described below on a case-by-case basis. The experimental regions are obtained by running Algorithm 1 with different parameter combinations on a log log plot. For each parameter combination the algorithm is run 100 times and the number of times the algorithm has 100% accuracy is recorded. The area where at least 99 of the 100 runs had 100% accuracy is shown in light gray in Figure 5.1.

Figure 5.1 a) displays the results for the data dimension p vs. the sample size N, keeping the number of clusters m fixed to $m = 3$ and the subsample size $n = \lceil \frac{m}{a}(\log m + \log \frac{4}{\delta}) \rceil$. According to Corollary 7, the theoretical p in this case should be at least

$$p > \lceil 128(2 \log N + \log \frac{40}{\delta}) \rceil$$

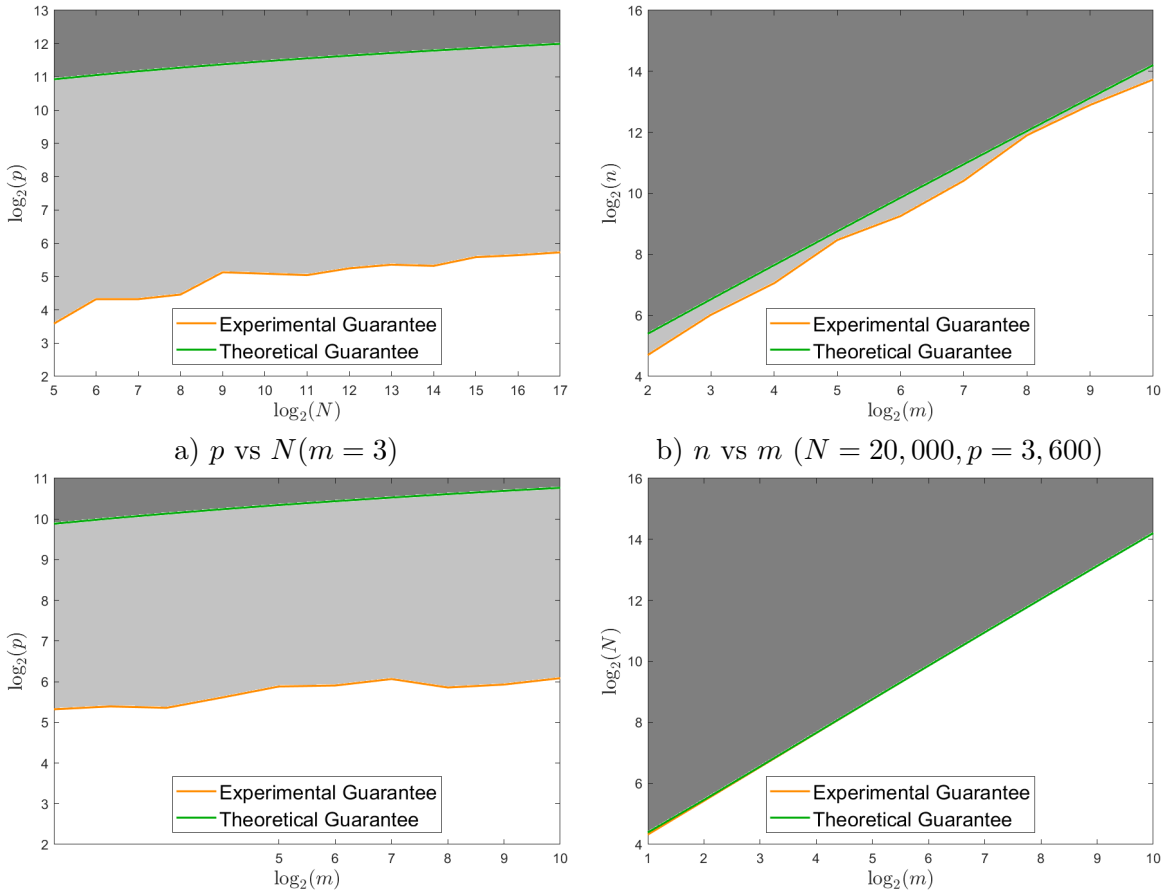
when $\delta = 0.01$. This area is shown in dark gray in Figure 5.1 a). From the plot, it can be seen that the theoretical bound on p is not very tight, since there is a large gap, by a factor of over 64 between the dark region (theoretical) and the light gray region (experimental).

Figure 5.1 b) displays the results for the subsample size n vs. the number of clusters m, when the sample size is $N = 20,000$ and $p = 3600$. According to Corollary 7, the theoretical n is at least

$$n > \lceil \frac{m}{a}(\log m + \log \frac{4}{\delta}) \rceil.$$

The plot indicates that the theoretical bound for n is tight, by a factor around 1.2.

Figure 5.1 c) displays the results for the data dimension p vs. the number of clusters m, when N is fixed to be $N = 20,000$ and $n = \lceil \frac{m}{a}(\log m + \log \frac{4}{\delta}) \rceil$. According to Corollary 7, the theoretical



a) p vs $N(m = 3)$

b) n vs m $(N = 20,000, p = 3,600)$

c) p vs m $(N = 20,000)$

d) N vs m $(p = 3,600)$

Figure 5.1: Comparison between the parameter combinations where the SCRLM algorithm is theoretically guaranteed to have 100% accuracy for 99% of the time with the experimental findings in GMM with outliers.

p should be at least

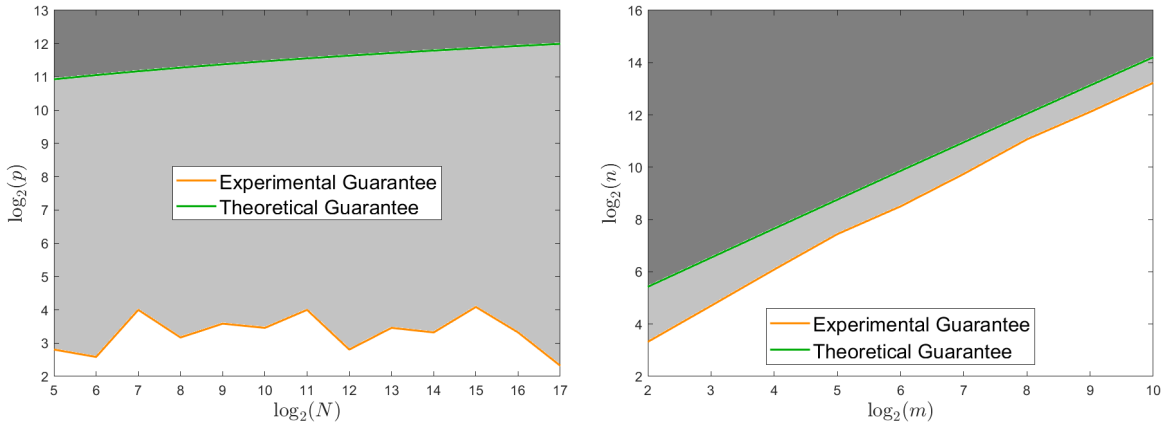
$$p > \lceil 128(\log m + \log \frac{8}{\delta}) \rceil$$

when $\delta = 0.01$. The plot indicates that the bound on p in not very tight, off by a factor over 32.

Figure 5.1 d) displays the results for the sample size N vs. the number of clusters m, when $p = 3600$ and $n = \lceil \frac{m}{a}(\log m + \log \frac{4}{\delta}) \rceil$. According to Corollary 7, the theoretical N is at least

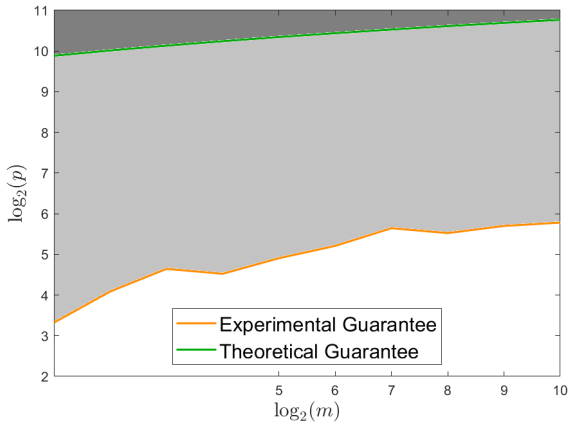
$$N > \lceil \frac{m}{a}(\log m + \log \frac{4}{\delta}) + 1 \rceil.$$

The theoretical bound almost overlaps the experimental bound in this case since the smallest N one can pick is n.

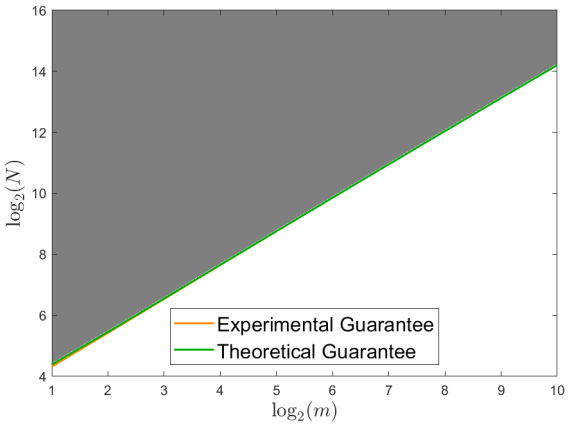


a) p vs $N(m = 3)$

b) n vs m $(N = 20,000, p = 3,600)$



c) p vs m $(N = 20,000)$



d) N vs m $(p = 3,600)$

Figure 5.2: Comparison between the parameter combinations where the SCRLM algorithm is theoretically guaranteed to have 100% accuracy for 99% of the time with the experimental findings in GMM without outliers.

The empirical results support the conclusions that the theoretical bound for p is conservative and accurate results are obtained with smaller values of p in practice, but the theoretical bounds for N and n are in good agreement with values needed in practice.

We also generate data in GMM without outliers and compare the theoretical bounds with the experimental bounds in Figure 5.2. The minimum and maximum weights for the positive clusters are taken to be $0.7/m$ and $1.3/m$ respectively. The standard deviations σ_i of positive clusters and ρ are taken to be the same as Section 5.2.1. From Figure 5.1 and Figure 5.2, it is observed that the experimental bounds in GMM with outliers is tighter than the bounds in GMM without outliers.

5.2.2 Comparison of Observed and Theoretical Accuracy of HSCRLM

This section evaluates the tightness of the theoretical bounds for Algorithm 2. Synthetic data is generated with 10% first-level outliers and 10% second-level outliers per first-level cluster. The minimum and maximum weights for the first and second-level positive clusters are taken to be $0.85/m$ and $0.95/m$ respectively. The standard deviations σ_i of first-level positive clusters are linearly increasing with i from 0.5 to 0.6, The standard deviations σ_{ij} of second-level positive clusters are linearly increasing with j from 0.05 to 0.3. The experiments use $\rho_1 = 0.7$ and $\rho_2 = 0.35$.

Figure 5.3 a) displays the results for the data dimension p vs. the sample size N, keeping the number of first-level and second-level clusters $m_1 = m_2 = 4$ and the subsample size $n_1 = \lceil \frac{m_1}{a_1}(\log m_1 + \log \frac{6}{\delta}) \rceil$ and $n_2 = \lceil \frac{m_2}{a_2}(\log m + \log \frac{6}{\delta}) \rceil$. According to Corollary 22, the theoretical p in this case should be at least

$$p > \lceil 128(\log (204N^2 + 48m_1 + 12m) - \log \delta) \rceil.$$

when $\delta = 0.01$. From the plot, it can be seen that the theoretical bound on p is not very tight, since there is a large gap, by a factor of over 16 between the dark region (theoretical) and the light gray region (experimental).

Figure 5.3 b) displays the results for the subsample size n_1 vs. the number of first-level clusters m_1, when the sample size is $N = 800,000$ and $p = 4800$. According to Corollary 22, the theoretical n_1 is at least

$$n_1 > \lceil \frac{m_1}{a_1}(\log m_1 + \log \frac{6}{\delta}) \rceil.$$

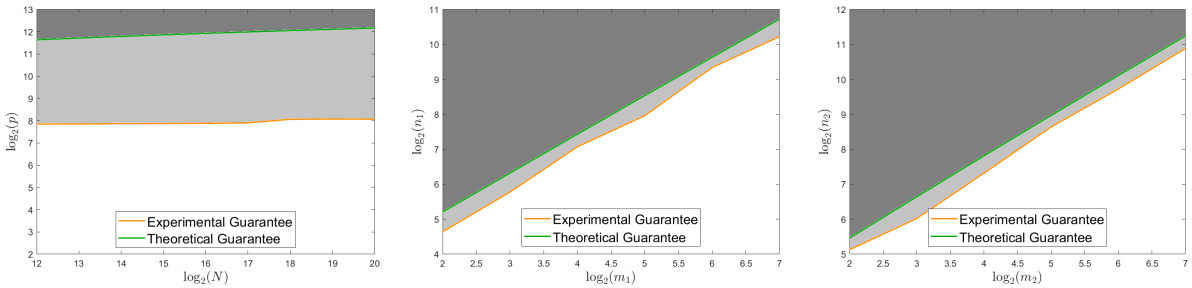
The plot indicates that the theoretical bound for n_1 is tight, by a factor around 1.39.

Figure 5.3 c) displays the results for the subsample size n_2 vs. the number of second-level clusters m_2, when the sample size is $N = 800,000$ and $p = 4800$. According to Corollary 22, the theoretical n_2 is at least

$$n_2 > \lceil \frac{m_2}{a_2}(\log m + \log \frac{6}{\delta}) \rceil.$$

The plot indicates that the theoretical bound for n_2 is tight, by a factor around 1.34.

The empirical results support the conclusions that the theoretical bound for p is conservative and accurate results are obtained with smaller values of p in practice, but the theoretical bounds for n_1 and n_2 are in good agreement with values needed in practice.



a) p vs N b) n_1 vs m_1 c) n_2 vs m_2

Figure 5.3: Comparison between the parameter combinations where the HSCRLM algorithm is theoretically guaranteed to have 100% accuracy for 99% of the time with the experimental findings.

5.2.3 Stability of SCRLM Relative to the Bandwidth Parameter

This following experiments evaluate the tightness of the theoretical bounds of ρ for Algorithm 1. The experiments use $\sigma_{max} = 0.25$.

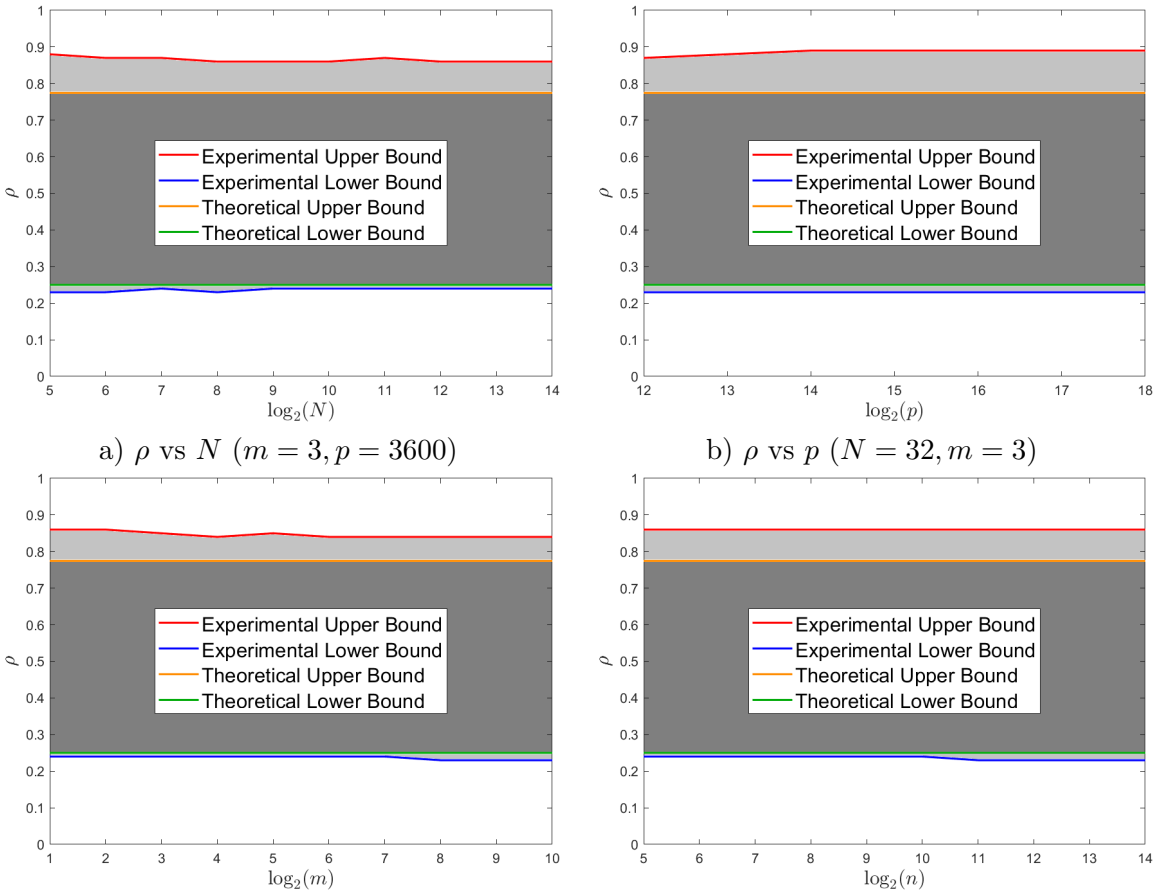
Figure 5.4 a) displays the results for the bandwidth parameter ρ vs. the sample size N, keeping the number of clusters m fixed to $m = 3$ and the subsample size $n = \lceil \frac{m}{a}(\log m + \log \frac{4}{\delta}) \rceil$. Figure 5.4 b) displays the results for the bandwidth parameter ρ vs. the data dimension p, when $N = 32$, $m = 3$ and $n = \lceil \frac{m}{a}(\log m + \log \frac{4}{\delta}) \rceil$. Figure 5.4 c) displays the results for the the bandwidth parameter ρ vs. the number of clusters m, when N is fixed to be $N = 20000$, p is fixed to be 3700 and $n = \lceil \frac{m}{a}(\log m + \log \frac{4}{\delta}) \rceil$. Figure 5.4 d) displays the results for the bandwidth parameter ρ vs. the number of subsamples n, when $N = 20000$, $p = 4200$ and $m = 3$.

According to Assumption 1, for all of the experiments, the theoretical upper bound of ρ is $\sqrt{0.6}$, and the theoretical lower bound of ρ is $\sigma_{max} = 0.25$. From Figure 5.4, it can be seen that the theoretical upper bound on ρ is not very tight with a difference of more than 0.1, but the theoretical lower bound on ρ is very tight with the difference less than 0.02.

The empirical results support the conclusions that the theoretical upper bound for ρ is not tight, that 100% accuracy can be achieved with $\rho > \sigma_{max}$ in practice, but the theoretical lower bounds for ρ are in good agreement with values needed in practice.

5.2.4 Comparison with other clustering methods

For these simulations, the data is generated with different number of clusters (m), different dimension (p) and different number of observations (N). The data is generated to contain 80% positives and 20% negatives (outliers). The number of desired clusters is specified as $m + 1$ for



a) ρ vs N $(m = 3, p = 3600)$ b) ρ vs p $(N = 32, m = 3)$

c) ρ vs m $(N = 20,000, p = 3600)$ d) ρ vs n $(N = 20,000, p = 3600, m = 3)$

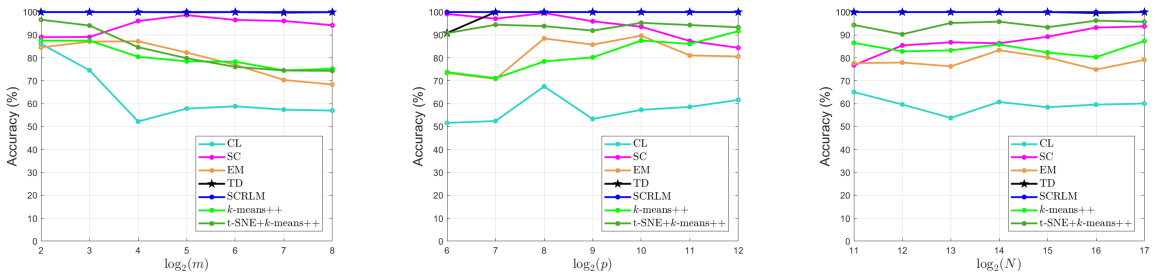
Figure 5.4: Evaluation of tightness of the bandwidth parameter ρ.

all methods other than SCRLM. For SCRLM, the number of desired clusters T was selected to be $T = N$ and thus the actual number of clusters was found automatically. From Figure 5.5, it can be seen that only SCRLM and TD are able to detect outliers, the other methods are very sensitive to outliers and SCRLM is much faster than TD which is shown in Figure 5.6.

5.2.5 Tuning of Bandwidth Parameter

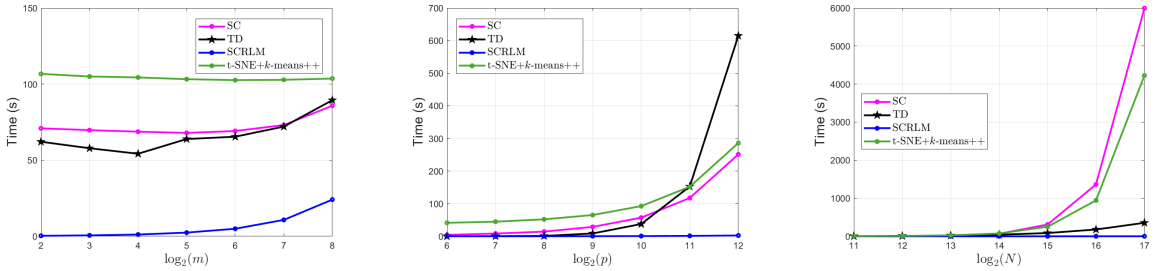
This section describes how to tune the hyper-parameter ρ based on data distribution. It is followed by the following process.

1. Subsampling the Data. First of all, we randomly select a representative subset, or subsample, from the complete dataset. This subsample should maintain the essential characteristics and patterns present in the entire data population. From Corollary 7, n should be chosen by $\lceil \frac{m}{a}(\log m + \log \frac{4}{\delta}) \rceil$.



a) Accuracy vs m b) Accuracy vs p c) Accuracy vs N

Figure 5.5: Accuracy of clustering algorithms on simulation data.



a) Time vs m b) Time vs p c) Time vs N

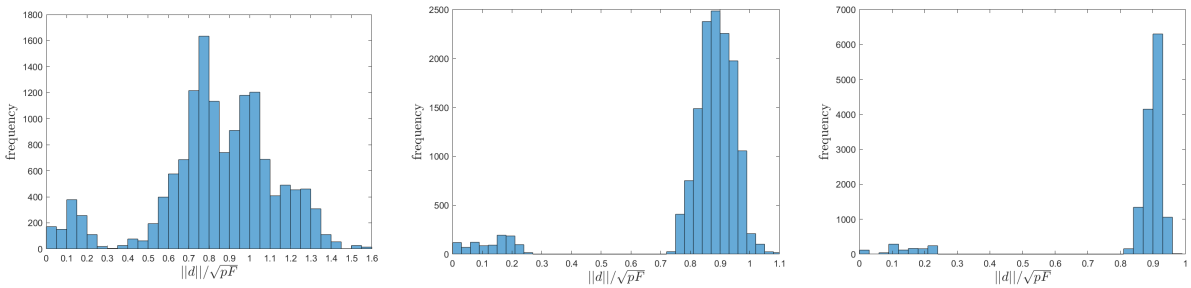
Figure 5.6: Time of clustering algorithms on simulation data.

2. Computation of Scaled Pairwise Distances. Secondly, we employ the Euclidean distance to calculate the pairwise distances between data points within the selected subsample. After computing the distances, divide each distance value by $\sqrt{pF}$. The resulting pairwise distance matrix quantifies the dissimilarity or similarity between each pair of data points.

3. Construction of Scaled Distance Histogram. Then we formulate a histogram that represents the distribution of the scaled pairwise distances within the subsample. The histogram's bin structure captures the range of scaled distance values, allowing for insights into the distribution characteristics.

4. Observation of Gap in Histogram. Finally, we examine the histogram of pairwise distances for patterns and gaps. A noticeable gap or separation between clusters of distances suggests the presence of distinct groups or clusters within the subsamples.

From Figure 5.7, when the dimension becomes bigger, we have a wider range of choice of ρ. For example, when $p = 100$, an appropriate ρ is chosen between $[0.3, 0.7]$, when $p = 1000$, an appropriate ρ can be chosen between $[0.25, 0.8]$.



a) $p = 10$ b) $p = 100$ c) $p = 1000$

Figure 5.7: Tuning of bandwidth parameter ρ ($N = 20000, m = 10, w_{-1} = 0.2$).

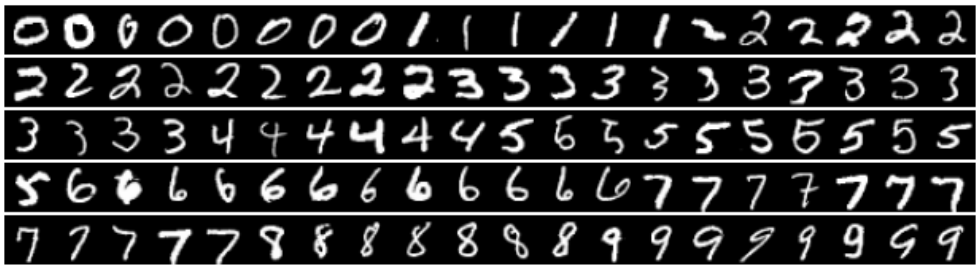


Figure 5.8: Variability of MNIST, cluster centers obtained by SCRLM ($T = 100$).

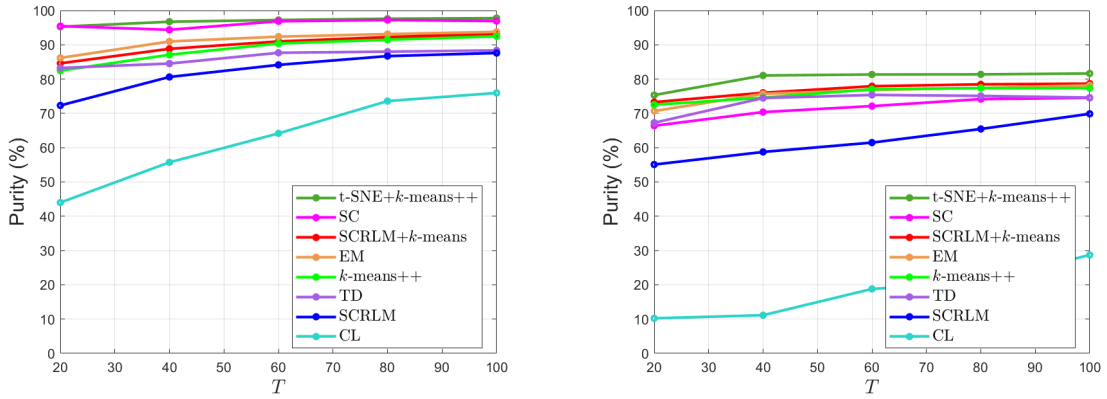
5.3 Real Data Experiments

To show that the SCRLM is an effective method, it was applied to four real datasets: MNIST (Deng, 2012), CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009) and ImageNet ILSVRC-2012 (Russakovsky et al., 2015).

MNIST (Deng, 2012) has 70,000 images of handwritten digits from 0 to 9 with 60,000 images used for training and 10,000 images used for testing. CIFAR-10 (Krizhevsky et al., 2009) consists of 60000 images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. CIFAR-100 (Krizhevsky et al., 2009) is just like the CIFAR-10, except it has 100 classes containing 600 images each. The ImageNet (Russakovsky et al., 2015) validation dataset has 50000 observations on 1000 classes with 50 observation per class and the ImageNet training dataset has almost 1.3 million observations on 1000 classes.

5.3.1 Data Preprocessing

Feature extraction for image data obtains a compact feature vector from the interesting parts of an image. The model SimCLR (Chen et al., 2020) was used to obtain a version of the MNIST



a) MNIST b) CIFAR-10

Figure 5.9: Purity vs number of clusters T of clustering algorithms on MNIST and CIFAR-10.

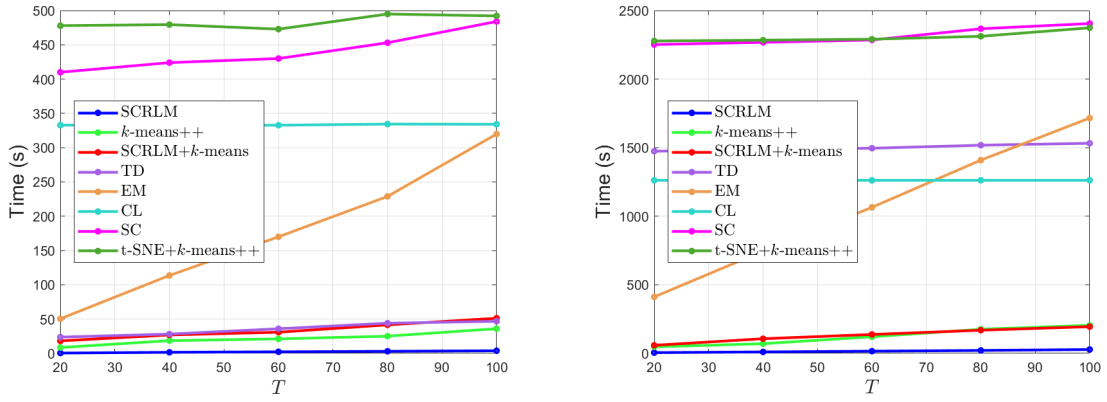
dataset as real vectors with dimension $p = 512$. The images from the CIFAR-10 and CIFAR-100 were resized to 144×144 pixels, then a pre-trained CNN, CLIP ResNet50 $\times$ 64 (Radford et al., 2021) with average pooling was used to obtain a $p = 4096$ dimensional feature vector for each image. The images from the ImageNet were resized to 224×224 pixels, then a $p = 640$ dimensional feature vector for each image was obtained using CLIP ResNet50 $\times$ 4 (Radford et al., 2021) and attention pooling.

5.3.2 Results

Figure 5.8 shows the cluster centers obtained by SCRLM when the number of desired clusters T is set to be 100 in MNIST. It can be seen that each cluster center is a good representation of that cluster. The variations of simple digits like 1 and 4 are relatively small, while complex digits like 2 and 3 have more variations. This shows MNIST is likely to have a hierarchical structure that can be used to cluster data when the number of clusters has a range of values.

Figure 5.9 and 5.10 support the conclusion that the SCRLM-based methods are superior to other methods for problems with a large number of clusters. From the plot, it can be seen that the purity of SCRLM and SCRLM+k-means increases as the number of clusters increases. However, the purity of TD does not have an obvious increase as the number of clusters increases, and the running time of EM increases significantly as the number of clusters increases. Therefore, SCRLM+k-means is the most efficient in producing a particular level of accuracy within a particular time.

The comparison of accuracy and time is shown in Figure 5.11 and summarized in Tables 5.1 and 5.2. In all the cases, SCRLM outperforms all other methods in terms of running time. EM performs



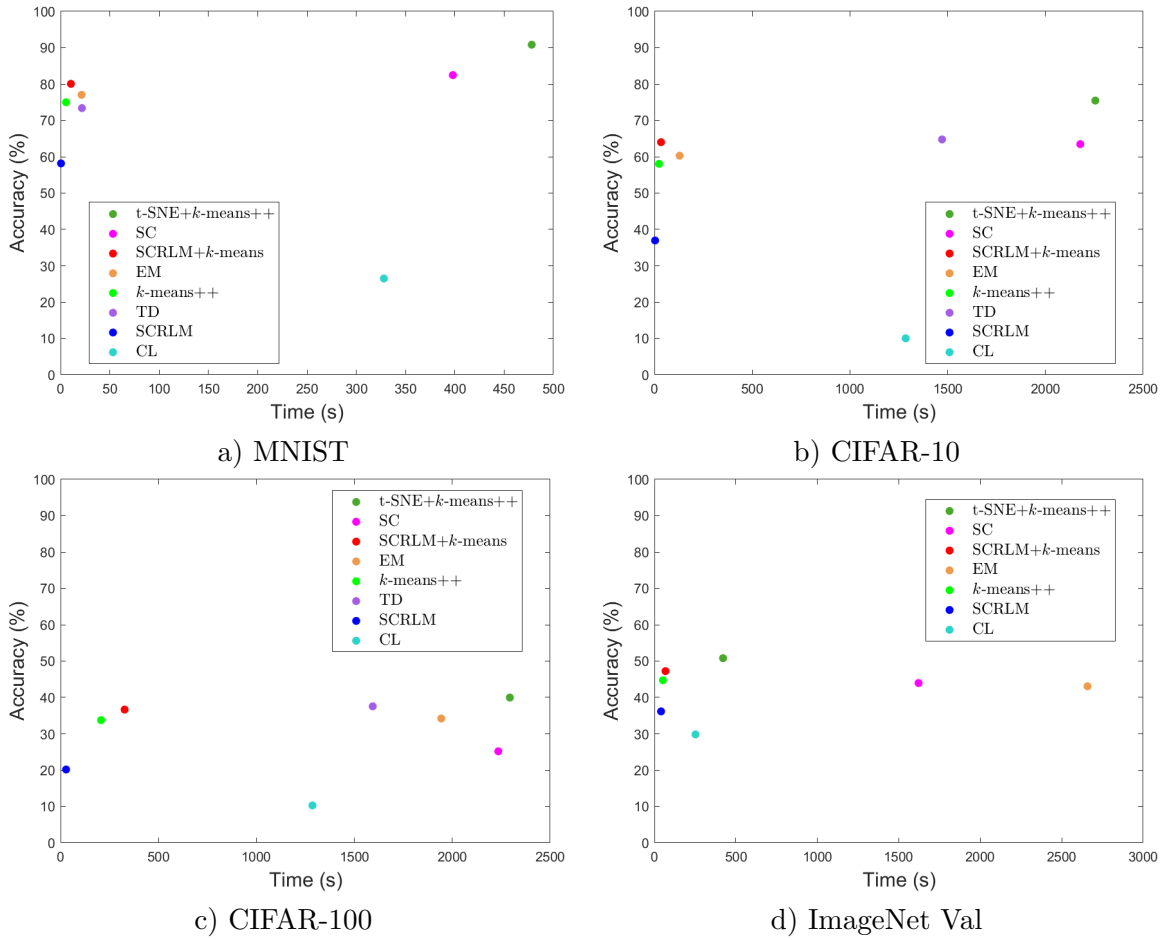
a) MNIST b) CIFAR-10

Figure 5.10: Computation time vs number of clusters T of clustering algorithms on MNIST and CIFAR-10.

well when the number of clusters is small but has prohibitive computation cost for CIFAR-100 and ImageNet validation datasets. t-SNE and TD achieve the best accuracy but only have acceptable running time when the dimension is small. Therefore, only SCRLM, SCRLM+k-means and k-means++ are compared for the ImageNet training dataset. From Tables 5.1 and 5.2 it can be seen that SCRLM+k-means achieves a higher accuracy on ImageNet than k-means++ in far less time, by a factor of 3.83. This demonstrates that SCRLM can be used as an initialization technique for k-means clustering that has a better performance than k-means++.

Table 5.1: Accuracy of clustering algorithms on five image datasets.

Accuracy(%)	MNIST	CIFAR-10	CIFAR-100	ImageNet val	ImageNet
CL	26.50	10.05	10.29	29.83	-
SC	82.46	63.47	25.17	43.96	-
EM	77.03	60.29	34.21	43.07	-
TD	73.38	64.76	37.55	-	-
t-SNE+k-means++	90.83	75.45	39.97	50.81	-
k-means++	74.99	58.06	33.75	44.73	47.71
SCRLM	58.17	36.96	20.17	36.16	34.01
SCRLM+k-means	80.06	64.00	36.66	47.24	48.61



a) MNIST b) CIFAR-10

c) CIFAR-100 d) ImageNet Val

Figure 5.11: Accuracy vs time of different clustering algorithms on four image datasets.

Table 5.2: Computation time of clustering algorithms on five image datasets.

Time(s)	MNIST	CIFAR-10	CIFAR-100	ImageNet val	ImageNet
CL	328	1,285	1,286	252	-
SC	398	2,178	2,235	1,621	-
EM	21.3	129	1,944	2,658	-
TD	21.7	1,471	1,594	-	-
t-SNE+k-means++	478	2,255	2,294	421	-
k-means++	5.61	23.8	207	52.1	10,005
SCRLM	0.46	3.25	28.2	40.6	1,269
SCRLM+k-means	10.5	33.6	327	67.8	2,610

CHAPTER 6

CONCLUSIONS

6.1 Summary of Completed Work

In this dissertation, two innovative clustering methods: SCRLM and its hierarchical counterpart, HSCRLM are introduced. SCRLM is designed to address the challenge of clustering large-scale GMM in the presence of outliers. The fundamental assumptions underpinning this algorithm include the assumption of isotropic Gaussians for the foreground (positives) clusters and a constraint on the range of values for the bandwidth parameter of the loss function. Notably, SCRLM stands out from many conventional clustering methods by offering robust theoretical guarantees. With high confidence, it excels in outlier detection and ensures accurate clustering, particularly when dealing with a large number of clusters and dimensions. Moreover, it can be used as an initialization strategy for k-means clustering and was observed to have better performance than other centroid initialization methods in extensive experiments.

HSCRLM, an extension of SCRLM, is designed to tackle the hierarchical domain of HGMM with outliers. It inherits the robustness and scalability of its predecessor while accommodating the hierarchical structure of complex data. Just like SCRLM, HSCRLM offers robust theoretical guarantees, ensuring accurate clustering and outlier detection within the hierarchical context.

In conclusion, this dissertation marks a significant milestone in the pursuit of robust clustering solutions for modern data analysis challenges. SCRLM and HSCRLM stand as robust, theoretically grounded algorithms, and SCRLM's versatility as a clustering method and k-means initializer holds great promise.

6.2 Future Research

While this dissertation has laid a strong foundation for the SCRLM and HSCRLM algorithms, there remain intriguing avenues for future research in the realm of robust clustering. One promising direction involves the parallelization of SCRLM to harness the computational power of modern parallel and distributed computing environments. Developing strategies for efficient parallel execution can significantly enhance SCRLM and HSCRLM's scalability, allowing it to handle even larger datasets in a time-efficient manner. Additionally, exploring the integration of SCRLM with emerging technologies, such as deep learning and online learning, could further extend its capabilities. Investigating the adaptability to diverse data types beyond Gaussian Mixtures and the development of automated parameter tuning methods are areas ripe for exploration. Furthermore, real-world applications across various domains, such as healthcare and finance, offer opportunities to validate SCRLM's robustness and utility in practical settings.

BIBLIOGRAPHY

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of data*, pages 94–105. 13

Ahmed, K. N. and Razak, T. A. (2014). A comparative study of different density based spatial clustering algorithms. *International Journal of Computer Applications*, 975:8887. viii, 13

Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., and Moustafa, A. A. (2019). The application of unsupervised clustering methods to Alzheimer's disease. *Frontiers in Computational Neuroscience*, 13:31. 1

Andritsos, P. et al. (2002). Data clustering techniques. *Rapport technique, University of Toronto. Department of Computer Science*. 13

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. *ACM SIGMOD Record*, 28(2):49–60. 12

Arthur, D. and Vassilvitskii, S. (2007). K-means++ the advantages of careful seeding. In *Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. 5, 7, 67

Ayesha, S., Hanif, M. K., and Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, 59:44–58. 16

Bahmani, B., Moseley, B., Vattani, A., Kumar, R., and Vassilvitskii, S. (2012). Scalable k-means++. *Proceedings of the Very Large Data Bases Endowment*, 5(7):622–633. 8

Birant, D. and Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, 60(1):208–221. 12

Borah, B. and Bhattacharyya, D. K. (2004). An improved sampling-based DBSCAN for large spatial databases. In *International Conference on Intelligent Sensing and Information Processing*, pages 92–96. IEEE. 12

Bradley, P. S., Fayyad, U., Reina, C., et al. (1998). Scaling EM (expectation-maximization) clustering to large databases. *Microsoft Research*, pages 0–25. 11

Brecheisen, S., Kriegel, H.-P., and Pfeifle, M. (2006). Parallel density-based clustering of complex objects. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 179–188. Springer. 1

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. 75

Coleman, G. B. and Andrews, H. C. (1979). Image segmentation by clustering. *Proceedings of the IEEE*, 67(5):773–785. 1

Comon, P. (1991). Independent component analysis. *Higher Order Statistics*, pages 29–38. 15

Cox, M. A. and Cox, T. F. (2008). Multidimensional scaling. In *Handbook of data visualization*, pages 315–347. Springer. 15

Dafir, Z., Lamari, Y., and Slaoui, S. C. (2021). A survey on parallel clustering algorithms for big data. *Artificial Intelligence Review*, 54(4):2411–2443. ix, 17

Dasgupta, S. and Schulman, L. J. (2007). A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research*, 8:203–226. 3

Dean, J. and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113. 16

Demartines, P. and Hérault, J. (1997). Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1):148–154. 15

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22. 3, 11, 67

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142. 75

Diday, E., Govaert, G., Lechevallier, Y., and Sidi, J. (1981). Clustering in pattern recognition. In *Digital Image Processing*, pages 19–58. Springer. 1

Donath, W. E. and Hoffman, A. J. (1973). Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425. 14

Duda, R. O., Hart, P. E., et al. (1973). *Pattern classification and scene analysis*, volume 3. Wiley New York. 3

Dwivedi, R., Khamaru, K., Wainwright, M. J., Jordan, M. I., et al. (2018). Theoretical guarantees for EM under misspecified Gaussian mixture models. *Advances in Neural Information Processing Systems*, 31. 3

Eckart, B., Kim, K., and Kautz, J. (2018). HGMR: Hierarchical Gaussian Mixtures for Adaptive 3D Registration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 705–721. 3

Elbatta, M. T. and Ashour, W. M. (2013). A dynamic method for discovering density varied clusters. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 6(1). 12

Elgamal, T., Yabandeh, M., Aboulnaga, A., Mustafa, W., and Hefeeda, M. (2015). sPCA: Scalable principal component analysis for big data on distributed platforms. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 79–91. viii, 16

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231. 11

Fränti, P. and Sieranoja, S. (2019). How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, 93:95–112. 8

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 100(9):881–890. 15

Guha, S., Rastogi, R., and Shim, K. (1998). Cure: An efficient clustering algorithm for large databases. *ACM SIGMOD Record*, 27(2):73–84. 10, 15

Han, D., Agrawal, A., Liao, W.-K., and Choudhary, A. (2016). A novel scalable DBSCAN algorithm with Spark. In *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 1393–1402. IEEE. 17

Han, J., Pei, J., and Tong, H. (2022). *Data mining: concepts and techniques*. Morgan kaufmann. 1

He, Y., Tan, H., Luo, W., Feng, S., and Fan, J. (2014). MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data. *Frontiers of Computer Science*, 8(1):83–99. 16

Hinneburg, A., Keim, D. A., et al. (1998). An efficient approach to clustering in large multimedia databases with noise. In *KDD*, volume 98, pages 58–65. 12

Hsu, D. and Kakade, S. M. (2013). Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, pages 11–20. 67

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430. 15

Jardine, N. and van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7(5):217–240. 1

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254. 67

Kannan, R., Vempala, S., and Vetta, A. (2004). On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3):497–515. 14

Karypis, G., Han, E.-H., and Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75. 10

Kaufman, L. (1990). Partitioning around medoids (program pam). *Finding groups in data*, 344:68–125. 9

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons. 1, 10

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.(2009). *Technical report, University of Toronto.* 75

Lee, J. A. and Verleysen, M. (2007). *Nonlinear dimensionality reduction*, volume 1. Springer. 15

Li, C., Zhang, Y., Jiao, M., and Yu, G. (2014). Mux-Kmeans: multiplex Kmeans for clustering large-scale data set. In *Proceedings of the 5th ACM workshop on Scientific cloud computing*, pages 25–32. 16

Liu, J. and Han, J. (2018). Spectral clustering. In *Data Clustering*, pages 177–200. Chapman and Hall/CRC. 14

Liu, M., Chang, E., and Dai, B.-q. (2002). Hierarchical gaussian mixture model for speaker verification. In *Proceedings of the 7th International Conference on Spoken Language Processing*. 3

Liu, P., Zhou, D., and Wu, N. (2007). VDBSCAN: varied density based spatial clustering of applications with noise. In *2007 International Conference on Service Systems and Service Management*, pages 1–4. IEEE. 12

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137. 7

Löffler, M., Zhang, A. Y., and Zhou, H. H. (2021). Optimality of spectral clustering in the gaussian mixture model. *The Annals of Statistics*, 49(5):2506–2530. 3

Mahdi, M. A., Hosny, K. M., and Elhenawy, I. (2021). Scalable clustering algorithms for big data: A review. *IEEE Access*, 9:80015–80027. 1

Makarychev, K., Reddy, A., and Shan, L. (2020). Improved guarantees for k-means++ and k-means++ parallel. *Advances in Neural Information Processing Systems*, 33:16142–16152. 8

Meilă, M. and Shi, J. (2001). A random walks view of spectral segmentation. In *International Workshop on Artificial Intelligence and Statistics*, pages 203–208. 14

Milojicic, D. S., Kalogeraki, V., Lukose, R., Nagaraja, K., Pruyne, J., Richard, B., Rollins, S., and Xu, Z. (2002). Peer-to-peer computing. 17

Mirkin, B. (2005). *Clustering for data mining: a data recovery approach*. Chapman and Hall/CRC. 1

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856. 14, 67

Ng, R. T. and Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, 14(5):1003–1016. 9, 15

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110. 3

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572. 15

Pratihar, D. K. (2009). Non-linear dimensionality reduction techniques. In *Encyclopedia of Data Warehousing and Mining, Second Edition*, pages 1416–1424. IGI Global. 15

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. 76

Rummel, R. J. (1988). *Applied factor analysis*. Northwestern University Press. 15

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252. 75

Segol, N. and Nadler, B. (2021). Improved convergence guarantees for learning Gaussian mixture models by EM and gradient EM. *Electronic Journal of Statistics*, 15(2):4510–4544. 3

Sheikholeslami, G., Chatterjee, S., and Zhang, A. (1998). WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases. In *Proceedings of the 24rd International Conference on Very Large Data Bases*, pages 428–439. 13

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905. 14

Shi, X., Li, Y., and Zhao, Q. (2020). Flexible hierarchical gaussian mixture model for high-resolution remote sensing image segmentation. *Remote Sensing*, 12(7):1219. 4

Sinclair, A. and Jerrum, M. (1989). Approximate counting, uniform generation and rapidly mixing markov chains. *Information and Computation*, 82(1):93–133. 14

Srivastava, S. and Michael, N. (2018). Efficient, multifidelity perceptual representations via hierarchical gaussian mixture models. *IEEE Transactions on Robotics*, 35(1):248–260. 4

Thrun, M. C. (2018). *Projection-based clustering through self-organization and swarm intelligence: combining cluster analysis with the visualization of high-dimensional data*. Springer. 1

Thrun, M. C. and Ultsch, A. (2021). Swarm intelligence for self-organized clustering. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*, pages 5125–5129. 16

Tipping, M. E. and Bishop, C. M. (1999). Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482. 16

Van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605. 16, 67

Vempala, S. and Wang, G. (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860. 3

Venna, J., Peltonen, J., Nybo, K., Aidos, H., and Kaski, S. (2010). Information Retrieval Perspective to Nonlinear Dimensionality Reduction for Data Visualization. *Journal of Machine Learning Research*, 11:451–490. 16

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416. 14

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press. 22

WANG, W. (1997). STING: A statistical information grid approach to spatial data mining. In *Proceedings of Very Large Data Bases Conference*, pages 186–195. 13

Xu, L. and Jordan, M. I. (1996). On convergence properties of the em algorithm for gaussian mixtures. *Neural Computation*, 8(1):129–151. 3

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., and Stoica, I. (2010). Spark: Cluster computing with working sets. In *2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 10)*. 17

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM SIGMOD Record*, 25(2):103–114. 10, 15

Zhao, X., Liang, J., and Dang, C. (2019). A stratified sampling based clustering algorithm for large-scale data. *Knowledge-Based Systems*, 163:416–428. 1

BIOGRAPHICAL SKETCH

Yijia Zhou was born in 1996 in Changzhou, Jiangsu province of China. From a young age, Yijia had a passion for mathematics that ultimately led her to pursue a career in this field. After completing her Bachelor's degree in Financial Mathematics at Southern University of Science and Technology in 2017, she enrolled in the Ph.D. program of the Department of Mathematics at Florida State University in the same year. There, she worked under the guidance of professors Adrian Barbu and Kyle A. Gallivan.

Yijia's research focuses on designing and analyzing algorithms for unsupervised learning. She also works on sampling techniques for subsampling large data to efficiently explore, summarize, and learn. Her interests and experiences range a broad spectrum of machine learning and data science and she is passionate about using her research to address challenging, real-world problems.