THEORETICAL ADVANCES IN CLUSTERING WITH APPLICATIONS TO MATRIX FACTORIZATION

LIU ZHAOQIANG

NATIONAL UNIVERSITY OF SINGAPORE

2017

THEORETICAL ADVANCES IN CLUSTERING WITH APPLICATIONS TO MATRIX FACTORIZATION

LIU ZHAOQIANG

(B.Sc., Tsinghua University)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF MATHEMATICS
NATIONAL UNIVERSITY OF SINGAPORE
2017

Supervisors:
Professor Bao Weizhu
Assistant Professor Vincent Tan

Examiners:
Professor Toh Kim Chuan
Associate Professor Ji Hui
Associate Professor Nicolas Gillis, Université de Mons

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Liu Zhaoqiang

Liu Zhaoqiang

December 1, 2017

Acknowledgements

It is my great honor to thank those who made this thesis possible here.

First, I would like to sincerely thank my supervisor, Prof. Weizhu Bao. Prof. Bao is very rigorous and meticulous in research and helped me a lot in my research about computational mathematics. I think that I will always learn from his attitude towards research and try to be a rigorous researcher in the future. Furthermore, Prof. Bao was very kind and helpful when I was more interested in machine learning and data mining. He provided me great opportunities to enter machine learning field and recommended me to my co-supervisor Dr. Vincent Tan.

Second, I would like to express to sincere gratitude to my co-supervisor Dr. Tan. I am so lucky to have Dr. Tan as my co-supervisor to guide my research in machine learning. Dr. Tan provides me much valuable information about many cutting-edge research topics and helps me to find interesting directions to explore. I can always learn certain inspiring ideas from Dr. Tan during my regular meetings with him. Dr. Tan taught me how to write a nice paper very carefully and he helped me a lot to get my first paper published. Moreover, I have been deeply impressed by Dr. Tan's passion for research and by how productive he is. I admire his passion for research and I am trying to be so dedicated to research like he does.

Third, I heartfeltly thank my friends Jingrui Cheng at University of Wisconsin-Madison and Sheng Meng at NUS for their great help for my preparation for PhD qualifying examination. In addition, I would like to thank my friends, especially Quan Zhao, Xu Song, Bo Chen, for all the encouragement, emotional support, comradeship and entertainment they offered.

I would like to acknowledge my Bachelor's thesis supervisor Prof. Huaiyu Jian, who encouraged me to do a PhD.

Last but not least, I would like to thank my parents and my sister for their encouragement and unconditional support. To them I owe all that I am and all that I have ever accomplished.

Contents

Summary

Clustering is a fundamental task in machine learning that consists in grouping a set of objects such that the objects in the same group (called a cluster) are more similar than those in other groups. Clustering is a ubiquitous problem in various applications, such as analyzing the information contained in gene expression data, performing market research according to firms' financial characteristics or analyzing stock price behavior.

The main purpose of this thesis is to theoretically analyze the applications of clustering in various unsupervised learning problems, including the learning of mixture models and nonnegative matrix factorization (NMF).

The thesis mainly consists of two parts. The first part considers the informativeness of the k-means algorithm, which is perhaps the most popular clustering algorithm, for learning mixture models. The learning of mixture models can be viewed as a clustering problem. Indeed, given data samples independently generated from a mixture of distributions, we often would like to find the *correct target clustering* of the samples according to which component distribution they were generated from. For a clustering problem, practitioners often choose to use the simple k-means algorithm. k-means attempts to find an *optimal clustering* which minimizes the sum-of-squared distance between each point and its cluster center. In

Chapter 2 of this thesis, we provide sufficient conditions for the closeness of any optimal clustering and the correct target clustering assuming that the data samples are generated from a mixture of log-concave distributions. Moreover, we show that under similar or even weaker conditions on the mixture model, any optimal clustering for the samples with reduced dimensionality is also close to the correct target clustering. These results provide intuition for the informativeness of k-means (with and without dimensionality reduction) as an algorithm for learning mixture models. We verify the correctness of our theorems using numerical experiments and demonstrate using datasets with reduced dimensionality significant speed ups for the time required to perform clustering.

In the second part, we propose a geometric assumption on nonnegative data matrices such that under this assumption, we are able to provide upper bounds (both deterministic and probabilistic) on the relative error of nonnegative matrix factorization. The algorithm we propose first uses the geometric assumption to obtain an exact clustering of the columns of the data matrix; subsequently, it employs several rank-one NMFs to obtain the final decomposition. When applied to data matrices generated from our statistical model, we observe that our proposed algorithm produces factor matrices with comparable relative errors vis-à-vis classical NMF algorithms but with much faster speeds. On face image and hyperspectral imaging datasets, we demonstrate that our algorithm provides an excellent initialization for applying other NMF algorithms at a low computational cost. Finally, we show on face and text datasets that the combinations of our algorithm and several classical NMF algorithms outperform other algorithms in terms of clustering performance.

List of Tables

List of Figures

List of Symbols and Abbreviations

$\|\mathbf{X}\|_{\mathrm{F}}$	Frobenius norm of matrix $\mathbf{X}$
$\|\mathbf{X}\|_2$	2-norm of matrix $\mathbf{X}$ (also known as spectral norm)
$\mathrm{tr}(\mathbf{X})$	the trace of matrix $\mathbf{X}$
$\langle \mathbf{A}, \mathbf{B} \rangle$	matrix inner product, i.e., $\langle \mathbf{A}, \mathbf{B} \rangle := \mathrm{tr}(\mathbf{A}^T\mathbf{B})$
$[M]$	$[M] := \{1, 2, \cdots, M\}$ for any positive integer M
$\mathbf{V}(i, :)$	the i-th row of $\mathbf{V}$
$\mathbf{V}(:, j)$	the j-th column of $\mathbf{V}$
$\mathbf{X}(1\colon I, :\,)$	the first I rows of $\mathbf{X}$
$\mathbf{X}(:\,, 1\colon J)$	the first J columns of $\mathbf{X}$
$\mathbf{V}(:, \mathscr{K})$	the columns of $\mathbf{V}$ indexed by $\mathscr{K}$
$\mathbf{X}(i, j)$	the element in the (i, j)-th position of $\mathbf{X}$
$\mathbf{V}$	data matrix in $\mathbb{R}^{F \times N}$, where F is the dimensionality and N is the number of samples
$\mathbf{I}$	the identity matrix
$[\mathbf{V}_1, \mathbf{V}_2]$	the horizontal concatenation of two compatible matrices
$[\mathbf{V}_1; \mathbf{V}_2]$	the vertical concatenation of two compatible matrices
$\mathrm{diag}(\mathbf{w})$	the diagonal matrix with diagonal entries given by $\mathbf{w}$
$\mathbb{R}_+$	the set of nonnegative numbers

$\mathbb{R}_{++}$	the set of positive numbers
$\mathbb{N}_{+}$	the set of positive integers
$\xrightarrow{\mathrm{p}}$	convergence in probability
2D	two dimensions
3D	three dimensions
GMM	Gaussian mixture model
i.i.d.	independent and identically distributed
w.r.t.	with respect to
NMF	nonnegative matrix factorization
PCA	principal component analysis
SVD	singular value decomposition

Chapter 1

Introduction

1.1 Clustering

Generally speaking, clustering is the task of maximizing the similarity of objects within a cluster and minimizing the similarity of objects between different clusters. Clustering is a ubiquitous problem in various applications, such as analyzing the information contained in gene expression data [1], performing market research according to firms' financial characteristics or analyzing stock price behavior [2]. Objective-based clustering is a commonly-used technique for clustering. This is the procedure of minimizing a certain objective function to partition data samples into a fixed or appropriately-selected number of subsets known as *clusters*. The k-means algorithm [3] is perhaps the most popular objective-based clustering approach. Suppose we have a data matrix of N samples $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N] \in \mathbb{R}^{F \times N}$, a K*-clustering* (or simply a *clustering* or a *partition*) is defined as a set of pairwise disjoint index sets $\mathscr{C} := \{\mathscr{C}_1, \mathscr{C}_2, \dots, \mathscr{C}_K\}$ whose union is $\{1, 2, \dots, N\}$. The corresponding *sum-of-squares distortion measure* with respect to $\mathbf{V}$ and $\mathscr{C}$ is defined as

$$\mathcal{D}(\mathbf{V}, \mathscr{C}) := \sum_{k=1}^{K} \sum_{n \in \mathscr{C}_k} \|\mathbf{v}_n - \mathbf{c}_k\|_2^2, \qquad (1.1.1)$$

where $\mathbf{c}_k := \frac{1}{|\mathscr{C}_k|}\sum_{n\in\mathscr{C}_k}\mathbf{v}_n$ is the cluster center or centroid of the k-th cluster. The goal of the k-means algorithm is to find an *optimal clustering* $\mathscr{C}^{\mathrm{opt}}$ that satisfies

$$\mathcal{D}(\mathbf{V}, \mathscr{C}^{\mathrm{opt}}) = \min_{\mathscr{C}} \mathcal{D}(\mathbf{V}, \mathscr{C}), \tag{1.1.2}$$

where the minimization is taken over all K-clusterings. Optimizing this objective function is NP-hard [4]. Despite the wide usage of k-means and the theoretical analysis of k-means [5–7], there are few theoretical investigations with respect to *optimal clusterings*. Moreover, in real applications, such as clustering face images by identities, there are certain unknown correct target clusterings. While using k-means as a clustering algorithm, we make a *key implicit assumption* that any optimal clustering is close to the correct target clustering [8]. If there is an optimal clustering that is far away from the correct target clustering, then using k-means is meaningless because even if we obtain an optimal or approximately-optimal clustering, it may not be close to the desired correct target clustering.

Besides the k-means algorithm, there are many other objective-based clustering algorithms, such as k-medians [9] or min-sum clustering [10]. In addition, there are many other types of clustering algorithms besides objective-based clustering, for example, hierarchical clustering [11, 12]. Furthermore, it is well-known that the k-means algorithm is sensitive to initialization [5]. Various initialization techniques have been proposed, and the most popular one among them may be the k-means++ algorithm [5]. It uses a randomized greedy search algorithm to select initial cluster centroids from the samples. We use a deterministic greedy search algorithm in the clustering step of our algorithm for nonnegative matrix factorization and we derive corresponding theoretical guarantees for our algorithm.

1.2 The Learning of Mixture Models

Suppose there are K unknown distributions $F_1, F_2, \ldots, F_K$ and a probability vector $\mathbf{w} := [w_1, w_2, \ldots, w_K]$. The corresponding K-component mixture model is a generative model that assumes data samples are independently sampled such that the probability that each sample is generated from the k-th component is w_k, the *mixing weight for the k-th component*. Suppose that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$ are samples independently generated from a K-component mixture model, the *correct target clustering* $\mathscr{C} := \{\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_K\}$ satisfies the condition that $n \in \mathscr{C}_k$ if and only if $\mathbf{v}_n$ is generated from the k-th component. One of the most important goals of the learning of a mixture model is to find the correct target clustering of the samples (and thereby inferring the parameters of the model). The learning of mixture models, especially Gaussian mixture models (GMMs) [13–15], is of fundamental importance in machine learning and applied statistics.

1.3 Dimensionality Reduction

Due to the inherent inefficiencies in processing high-dimensional data, *dimensionality reduction* has received considerable attention. Applying dimensionality reduction techniques before clustering high-dimensional datasets can lead to significantly faster running times and reduced memory sizes. In addition, algorithms for learning GMMs usually include a dimensionality reduction step. For example, Dasgupta [16] shows that general ellipsoidal Gaussians become "more spherical" and thereby more amenable to (successful) analysis after a random projection onto a low-dimensional subspace. Vempala and Wang [17] show that reducing dimensionality by spectral decomposition leads to the amplification of the separation among Gaussian components. For performing the k-means algorithm to learn mixture models, we also consider reducing the dimensionality of data samples first before clustering. For the dimensionality reduction task, we start with considering principal component analysis (PCA) [18], which is perhaps the most popular dimensionality reduction technique. Because we need to perform singular value decomposition (SVD) in PCA, and SVD is time-consuming when the number of samples and the dimensionality of samples are large, we also consider performing randomized SVD [19,20] in PCA and performing random projection similarly to [16] for dimensionality reduction for the sake of achieving faster running time.

1.4 Nonnegative Matrix Factorization (NMF)

The nonnegative matrix factorization (NMF) problem can be formulated as follows: Given a nonnegative data matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ and a positive integer K, we seek nonnegative factor matrices $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times N}$, such that the distance (measured in some norm) between $\mathbf{V}$ and $\mathbf{WH}$ is minimized. Due to its non-subtractive, parts-based property which enhances interpretability, NMF has been widely used in machine learning [21] and signal processing [22] among others. In addition, there are many fundamental algorithms to approximately solve the NMF problem, including the multiplicative update algorithms [23], the alternating (nonnegative) least-squares-type algorithms [24–27], and the hierarchical alternating least square algorithms [28] (also called the rank-one residual iteration [29]). However, it is proved in [30] that NMF problem is NP-hard and all the basic algorithms simply either ensure that the sequence of objective functions is non-increasing or that the algorithm converges to the set of stationary points [29, 31, 32]. To the best of our knowledge, none of these algorithms is suitable for analyzing a bound on the approximation error of NMF.

In an effort to find computationally tractable algorithms for NMF and to provide theoretical guarantees on the errors of these algorithms, researchers have revisited the so-called *separability assumption* proposed by Donoho and Stodden [33]. An exact nonnegative factorization $\mathbf{V} = \mathbf{WH}$ is *separable* if for any $k \in \{1, 2, \ldots, K\}$, there is an $n(k) \in \{1, 2, \ldots, F\}$ such that $\mathbf{W}(n(k), j) = 0$ for all $j \neq k$ and $\mathbf{W}(n(k), k) > 0$. That is, an exact nonnegative factorization is separable if all the features can be represented as nonnegative linear combinations of K features. It is proved in [34] that under the separability condition, there is an algorithm that runs in time polynomial in F, N and K and outputs a separable nonnegative factorization $\mathbf{V} = \mathbf{W}^*\mathbf{H}^*$ with the number of columns of $\mathbf{W}^*$ being at most K. Furthermore, to handle noisy data, a perturbation analysis of their algorithm is presented. The authors assumed that $\mathbf{V}$ is normalized such that every row of it has unit ℓ_1 norm and $\mathbf{V}$ has a separable nonnegative factorization $\mathbf{V} = \mathbf{W}\mathbf{H}$. In addition, each row of $\mathbf{V}$ is perturbed by adding a vector of small ℓ_1 norm to obtain a new data matrix $\mathbf{V}'$. With additional assumptions on the noise and $\mathbf{H}$, their algorithm leads to an approximate nonnegative matrix factorization $\mathbf{W}'\mathbf{H}'$ of $\mathbf{V}'$ with a provable error bound for the ℓ_1 norm of each row of $\mathbf{V}' - \mathbf{W}'\mathbf{H}'$. To develop more efficient algorithms and to extend the basic formulation to more general noise models, a collection of elegant papers dealing with NMF under various separability conditions has emerged [35–41].

1.5 Purposes and Scope of This Thesis

There are three main contributions in Chapter 2.

1. We prove that if the data points are independently generated from a K-component spherical GMM with an appropriate separability assumption and the so-called non-degeneracy condition [42,43] (see Definition 1 to follow), then any optimal clustering of the data points is close to the correct target clustering with high probability provided the number of samples is commensurately large. We extend this result to mixtures of log-concave distributions.

2. We prove that under the same generation process, if the data points are projected onto a low-dimensional space using the first $K-1$ principal components of the empirical covariance matrix, then, under similar conditions, any optimal clustering for the data points with reduced dimensionality is close to the correct target clustering with high probability. Again, this result is extended to mixtures of log-concave distributions.

3. Lastly, we show that under appropriate conditions, any *approximately-optimal clustering* of the data points is close to the correct target clustering. This

enables us to use the theoretical framework provided herein to analyze various efficient dimensionality reduction techniques such as random projection and randomized singular value decomposition (SVD). It also allows us to combine our theoretical analyses with efficient variants of k-means that return approximately-optimal clusterings.

The main contributions in Chapter 3 are:

- Theoretical Contributions: We introduce a geometric assumption on the data matrix $\mathbf{V}$ that allows us to correctly group columns of $\mathbf{V}$ into disjoint subsets. This naturally suggests an algorithm that first clusters the columns and subsequently uses a rank-one approximate NMF algorithm [44] to obtain the final decomposition. We analyze the error performance and provide a deterministic upper bound on the relative error. We also consider a random data generation model and provide a probabilistic relative error bound. Our geometric assumption can be considered as a special case of the separability (or, more precisely, the near-separability) assumption [33]. However, there are certain key differences: First, because our assumption is based on a notion of clusterability [45], our proof technique is different from those in the literature that leverage the separability condition. Second, unlike most works that assume separability [35–39], we exploit the ℓ_2 norm of vectors instead of the ℓ_1 norm of vectors/matrices. Third, $\mathbf{V}$ does not need to be assumed to be normalized. As pointed out in [37], normalization, especially in the ℓ_1-norm for the rows of data matrices may deteriorate the clustering performance for text datasets significantly. Fourth, we provide an upper bound for relative error instead of the absolute error. Our work is the first to provide theoretical analyses for the relative error for near-separable-type NMF problems. Finally, we assume all the samples can be approximately represented by certain special samples (e.g., centroids) instead of using a small set of salient features to represent all the features. Mathematically, these two approximations may appear to be equivalent. However, our assumption and analysis techniques enable us to provide an efficient algorithm and tight probabilistic relative error bounds for the NMF approximation (cf. Theorem 8).

- Experimental Evaluations: Empirically, we show that this algorithm performs well in practice. When applied to data matrices generated from our statistical model, our algorithm yields comparable relative errors vis-à-vis several classical NMF algorithms including the multiplicative algorithm, the alternating nonnegative least square algorithm with block pivoting, and the hierarchical alternating least square algorithm. However, our algorithm is *significantly faster* as it simply involves calculating rank-one SVDs. It is also well-known that NMF is sensitive to initializations. The authors in [46, 47] use spherical k-means and an SVD-based technique to initialize NMF. We verify on several image and hyperspectral datasets that our algorithm, when combined with several classical NMF algorithms, achieves the best convergence rates and/or the smallest final relative errors. We also provide intuition for why our algorithm serves as an effective initializer for other NMF algorithms. Finally, combinations of our algorithm and several NMF algorithms achieve the best clustering performance for several face and text datasets. These experimental results substantiate that our algorithm can be used as a good initializer for standard NMF techniques.

This thesis is organized as follows.

In Chapter 2, we focus on using the k-means clustering and dimensionality reduction for learning mixtures models. More specifically, we mention some related works on learning Gaussian mixture models, k-means clustering, and dimensionality reduction in Section 2.1. Our main theoretical results concerning upper bounds on the misclassification error distance (ME distance) for spherical GMMs are presented in Section 2.2. In addition, numerical results examining the correctness of our bounds for spherical GMMs are also presented in this section. These results are extended to mixtures of log-concave distributions in Section 2.3. Other extensions (using other dimensionality-reduction techniques and the combination of our results with efficient clustering algorithms) are discussed in Section 2.4.

In Chapter 3, we propose a new initialization method for NMF which recovers a pair of factor matrices with provable relative error bounds under our geometric assumption. We discuss related works in Section 3.1. After that, we present our geometric assumption as well as certain useful lemmas in Section 3.2. In Section 3.3, we present our main theorem for deterministic data and we provide an algorithm named `cr1-nmf` which is able to recover a pair of factor matrices with relative error bounds. In addition, we verify in numerical experiments on several real datasets that this algorithm can be used as a good initialization method for NMF. We present our theorems for data matrices generated from a probabilistic model in Section 3.4. Automatically determining the latent dimensionality for NMF is an important practical problem and we consider this problem in Section 3.5. Finally, we verify the correctness of our theorems by synthetic experiments in Section 3.6.1 and we present the effectiveness and efficacy of our algorithm by real-data experiments in Section 3.6.2.

In Chapter 4, conclusions are drawn and we discuss some possible future works.

Chapter 2

k-Means Clustering and Dimension Reduction for Learning Mixture Models

We mentioned that the learning of mixture models is of fundamental importance in machine learning and applied statistics. In addition, as we will mention in Section 2.1.1 to follow, there are plenty of algorithms designed to learn mixture models. The learning of mixture models can be considered as a clustering problem that attempts to find the correct target clustering of samples according to which component distribution they were generating from. For a given clustering problem, it is natural to first try the popular k-means algorithm. However, a natural question beckons. Is it even appropriate to use the k-means algorithm to learn mixture models? What are some sufficient conditions such that the answer to the above question is "Yes"? In this chapter, we answer these questions by comparing the correct target clustering with any optimal clustering of the objective function of the k-means clustering. Furthermore, motivated by the reduction in complexity (running times and memory sizes) in applying clustering algorithms on reduced-dimensionality datasets, we also provide theoretical guarantees for the case in which the dataset first undergoes a dimensionality reduction step.

2.1 Background

In this section, we discuss some relevant existing works.

2.1.1 Learning Gaussian Mixture Models (GMMs)

Suppose there are K unknown distributions $F_1, F_2, \ldots, F_K$ and a probability vector $\mathbf{w} := [w_1, w_2, \ldots, w_K]$. The corresponding K-component mixture model is a generative model that assumes data samples are independently sampled such that the probability that each sample is generated from the k-th component is w_k, the *mixing weight for the k-th component.* For a K-component mixture model and for any $k \in [K]$, we *always* use $\mathbf{u}_k$ to denote the component mean vector, and use $\mathbf{\Sigma}_k$ to denote the component covariance matrix. When $\mathbf{\Sigma}_k = \sigma_k^2 \mathbf{I}$ for all $k \in [K]$, where $\mathbf{I}$ is the identity matrix, we say the mixture model is *spherical* and σ_k^2 is the variance of the k-th component. Suppose that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$ are samples independently generated from a K-component mixture model, the *correct target clustering* $\mathscr{C} := \{\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_K\}$ satisfies the condition that $n \in \mathscr{C}_k$ if and only if $\mathbf{v}_n$ is generated from the k-th component. One of the most important goals of the learning of a mixture model is to find the correct target clustering of the samples (and thereby inferring the parameters of the model).

The learning of mixture models, especially Gaussian mixture models (GMMs) [14,15], is of fundamental importance in machine learning. The EM algorithm [48,49] is widely used to estimate the parameters of a GMM. However, EM is a local-search heuristic approach for maximum likelihood estimation in the presence of incomplete data and in general, it cannot guarantee the parameters' convergence to global optima [50]. Recently, Hsu and Kakade [42] and Anandkumar et al. [43] provide approaches based on spectral decomposition to obtain consistent parameter estimates for spherical GMMs from first-, second- and third-order observable moments. To estimate parameters, they need to assume the so-called non-degeneracy condition for *spherical* GMMs with parameters $\{(w_k, \mathbf{u}_k, \sigma_k^2)\}_{k \in [K]}$.

Definition 1. *(Non-degeneracy condition) We say that a mixture model satisfies the* non-degeneracy condition *if its component mean vectors $\mathbf{u}_1, \ldots, \mathbf{u}_K$ span a K-dimensional subspace and the probability vector $\mathbf{w}$ has strictly positive entries.*

On the other hand, under certain separability assumptions on the Gaussian components, Dasgupta [16], Dasgupta and Schulman [51], Arora and Kannan [52], Vempala and Wang [17], and Kalai et al. [53] provide provably correct algorithms that guarantee most samples are correctly classified or that parameters are recovered with a certain accuracy with high probability. In particular, equipped with the following *separability* assumption

$$\|\mathbf{u}_i - \mathbf{u}_j\|_2 > C \max\{\sigma_i, \sigma_j\} \sqrt[4]{K \log \frac{F}{w_{\min}}}, \quad \forall, i, j \in [K], i \neq j, \tag{2.1.1}$$

for a spherical GMM, where $C > 0$ is a sufficiently large constant[1] and $w_{\min} := \min_{k \in [K]} w_k$, Vempala and Wang [17] present a simple spectral algorithm with running time polynomial in both F and K that correctly clusters random samples according to which spherical Gaussian they were generated from.

Remark 1. *We present more detailed discussion about previous work. In particular, we summarize known results for learning GMMs under pairwise separation assumptions. For simplicity, logarithmic factors in separation assumptions are ignored. Denote σ_k^2 as the maximum variance of the k-th component of the mixture along any direction. Dasgupta [16] provides an algorithm which is based on random projection to learn GMMs with the conditions that the mixing weights of all distributions are about the same, and the distance between any two different component mean vectors $\|\mathbf{u}_i - \mathbf{u}_j\|_2$ is at least $c\sqrt{F} \max\{\sigma_i, \sigma_j\}$, where c is a positive constant. Dasgupta and Schulman [51] further provide an EM based algorithm with the condition that $\|\mathbf{u}_i - \mathbf{u}_j\|_2$ is at least $cF^{\frac{1}{4}} \max\{\sigma_i, \sigma_j\}$. Arora and Kannan [52] also consider a learning algorithm with similar separation assumptions. Vempala and Wang [17]*

[1] Throughout, we use the generic notations C and $C_i, i \in \mathbb{N}$ to denote (large) positive constants. They depend on the parameters of the mixture model and may change from usage to usage.

are the first to consider using spectral algorithms (based on singular value decomposition) for dimensionality reduction to enable the learning of spherical GMMs. Their spectral algorithm requires a much weaker condition, i.e., $\|\mathbf{u}_i - \mathbf{u}_j\|_2$ *being at least* $cK^{\frac{1}{4}} \max\{\sigma_i, \sigma_j\}$*, for spherical GMMs. Achlioptas and McSherry [54] extend Vempala and Wang's results [17] to mixtures of arbitrary Gaussians with* $\|\mathbf{u}_i - \mathbf{u}_j\|_2$ *being at least* $c\left(K + \frac{1}{\sqrt{\min\{w_i, w_j\}}}\right) \max\{\sigma_i, \sigma_j\}$*. Kannan et al. [55] also present an algorithm for learning mixtures of arbitrary Gaussians with the corresponding separation between* $\mathbf{u}_i$ *and* $\mathbf{u}_j$ *being at least* $c(\sigma_i + \sigma_j)\frac{K^{\frac{3}{2}}}{w_{\min}^2}$*.*

Despite the large number of algorithms designed to find the (approximately) correct target clustering of a GMM, many practitioners use k-means because of its simplicity and successful applications in various fields. Kumar and Kannan [56] show that the k-means algorithm with a proper initialization can correctly cluster nearly all the data points generated from a GMM that satisfies a certain *proximity* assumption. Our theoretical results provide an explanation on *why* the k-means algorithm that attempts to find an optimal clustering is a good choice for learning mixture models. We compare and contrast our work to that of [56] in Remark 5.

2.1.2 A Lower Bound on Distortion and the ME Distance

Let $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N] \in \mathbb{R}^{F \times N}$ be a dataset and $\mathscr{C} := \{\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_K\}$ be a K-clustering. Let $\mathbf{H} \in \mathbb{R}^{K \times N}$ with elements $\mathbf{H}(k,n)$ be the clustering membership matrix satisfying $\mathbf{H}(k,n) = 1$ if $n \in \mathscr{C}_k$ and $\mathbf{H}(k,n) = 0$ if $n \notin \mathscr{C}_k$ for $(k,n) \in [K] \times [N]$. Let $n_k = |\mathscr{C}_k|$ and $\bar{\mathbf{H}} := \mathrm{diag}(\frac{1}{\sqrt{n_1}}, \frac{1}{\sqrt{n_2}}, \ldots, \frac{1}{\sqrt{n_K}})\mathbf{H}$ be the normalized version of $\mathbf{H}$. We have $\bar{\mathbf{H}}\bar{\mathbf{H}}^T = \mathbf{I}$ and the corresponding distortion can be written as [57]

$$\mathcal{D}(\mathbf{V}, \mathscr{C}) = \|\mathbf{V} - \mathbf{V}\bar{\mathbf{H}}^T\bar{\mathbf{H}}\|_\mathrm{F}^2 = \|\mathbf{V}\|_\mathrm{F}^2 - \mathrm{tr}(\bar{\mathbf{H}}\mathbf{V}^T\mathbf{V}\bar{\mathbf{H}}^T). \tag{2.1.2}$$

The *centering* of any data matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$ is a shift with respect to the mean vector $\bar{\mathbf{v}} = \frac{1}{N}\sum_{n=1}^N \mathbf{v}_n$ and the resultant data matrix $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N]$, with $\mathbf{z}_n = \mathbf{v}_n - \bar{\mathbf{v}}$ for $n \in [N]$, is said to be the *centralized matrix* of $\mathbf{V}$. Let $\mathbf{Z}$ be the centralized data matrix of $\mathbf{V}$ and define $\mathbf{S} := \mathbf{Z}^T\mathbf{Z}$. Note that $\mathcal{D}(\mathbf{V}, \mathscr{C}) = \mathcal{D}(\mathbf{Z}, \mathscr{C})$ for any clustering $\mathscr{C}$. Ding and He [57] make use of this property to provide a lower bound $\mathcal{D}^*(\mathbf{V})$ for distortion over all K-clusterings. That is, for any K-clustering $\mathscr{C}$,

$$\mathcal{D}(\mathbf{V}, \mathscr{C}) \geq \mathcal{D}^*(\mathbf{V}) := \mathrm{tr}(\mathbf{S}) - \sum_{k=1}^{K-1} \lambda_k(\mathbf{S}), \tag{2.1.3}$$

where $\lambda_1(\mathbf{S}) \geq \lambda_2(\mathbf{S}) \geq \ldots \geq 0$ are the eigenvalues of $\mathbf{S}$ sorted in non-increasing order.

For any two K-clusterings, the so-called *misclassification error (ME) distance* provides a quantitative comparison of their structures.

Definition 2. *(ME distance) The misclassification error distance of any two K-clusterings $\mathscr{C}^{(1)} := \{\mathscr{C}_1^{(1)}, \mathscr{C}_2^{(1)}, \ldots, \mathscr{C}_K^{(1)}\}$ and $\mathscr{C}^{(2)} := \{\mathscr{C}_1^{(2)}, \mathscr{C}_2^{(2)}, \ldots, \mathscr{C}_K^{(2)}\}$ is*

$$\mathrm{d}_{\mathrm{ME}}(\mathscr{C}^{(1)}, \mathscr{C}^{(2)}) := 1 - \frac{1}{N} \max_{\pi \in \mathcal{P}_K} \sum_{k=1}^{K} \left| \mathscr{C}_k^{(1)} \cap \mathscr{C}_{\pi(k)}^{(2)} \right|, \tag{2.1.4}$$

where $\mathcal{P}_K$ is the set of all permutations of $[K]$. It is known from [58] that the ME distance is indeed a distance.

For any $\delta, \delta' \in [0, K-1]$, define the functions

$$\tau(\delta, \delta') := 2\sqrt{\delta\delta'\Big(1 - \frac{\delta}{K-1}\Big)\Big(1 - \frac{\delta'}{K-1}\Big)}, \quad \text{and} \tag{2.1.5}$$

$$\tau(\delta) := \tau(\delta, \delta) = 2\delta\Big(1 - \frac{\delta}{K-1}\Big). \tag{2.1.6}$$

Combining Lemma 2 and Theorem 3 in [59], we have the following lemma.

Lemma 1. *Let $\mathscr{C} := \{\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_K\}$ and $\mathscr{C}' := \{\mathscr{C}_1', \mathscr{C}_2', \ldots, \mathscr{C}_K'\}$ be two K-clusterings of a dataset $\mathbf{V} \in \mathbb{R}^{F \times N}$. Let $p_{\max} := \max_k \frac{1}{N}|\mathscr{C}_k|$ and $p_{\min} := \min_k \frac{1}{N}|\mathscr{C}_k|$. Let $\mathbf{Z}$ be the centralized matrix of $\mathbf{V}$ and $\mathbf{S} = \mathbf{Z}^T\mathbf{Z}$. Define*

$$\delta := \frac{\mathcal{D}(\mathbf{V}, \mathscr{C}) - \mathcal{D}^*(\mathbf{V})}{\lambda_{K-1}(\mathbf{S}) - \lambda_K(\mathbf{S})}, \quad \textit{and} \quad \delta' := \frac{\mathcal{D}(\mathbf{V}, \mathscr{C}') - \mathcal{D}^*(\mathbf{V})}{\lambda_{K-1}(\mathbf{S}) - \lambda_K(\mathbf{S})}. \tag{2.1.7}$$

Then if $\delta, \delta' \leq \frac{1}{2}(K-1)$ and $\tau(\delta, \delta') \leq p_{\min}$, we have

$$\mathrm{d}_{\mathrm{ME}}(\mathscr{C}, \mathscr{C}') \leq \tau(\delta, \delta') p_{\max}. \tag{2.1.8}$$

This lemma says that any two "good" K-clusterings (in the sense that their distortions are sufficiently close to the lower bound of distortion $\mathcal{D}^*(\mathbf{V})$) are close to each other. In addition, we have the following useful corollary.

Corollary 1. *Let $\mathscr{C} := \{\mathscr{C}_1, \mathscr{C}_2, \dots, \mathscr{C}_K\}$ be a K-clustering of a dataset $\mathbf{V} \in \mathbb{R}^{F \times N}$ and define $p_{\max}$, $p_{\min}$, $\mathbf{Z}$, $\mathbf{S}$, and δ as in Lemma 1. Then if $\delta \le \frac{1}{2}(K-1)$ and $\tau(\delta) \le p_{\min}$, we have*

$$\mathrm{d}_{\mathrm{ME}}(\mathscr{C}, \mathscr{C}^{\mathrm{opt}}) \le p_{\max}\tau(\delta), \tag{2.1.9}$$

where $\mathscr{C}^{\mathrm{opt}}$ represents a K-clustering that minimizes the distortion for $\mathbf{V}$.

This corollary essentially says that if the distortion of a clustering is sufficiently close to the lower bound of distortion, then this clustering is close to any optimal clustering with respect to the ME distance.

2.1.3 Dimension Reduction by Principal Component Analysis (PCA)

Due to the inherent inefficiencies in processing high-dimensional data, *dimensionality reduction* has received considerable attention. Applying dimensionality reduction techniques before clustering high-dimensional datasets can lead to significantly faster running times and reduced memory sizes. In addition, algorithms for learning GMMs usually include a dimensionality reduction step. For example, Dasgupta [16] shows that general ellipsoidal Gaussians become "more spherical" and thereby more amenable to (successful) analysis after a random projection onto a low-dimensional subspace. Vempala and Wang [17] show that reducing dimensionality by spectral decomposition leads to the amplification of the separation among Gaussian components.

Principal component analysis (PCA) [18] is a popular strategy to compute the directions of maximal variances in vector-valued data and is widely used for dimensionality reduction. We write the singular value decomposition (SVD) of a *symmetric* matrix $\mathbf{A} \in \mathbb{R}^{F \times F}$ as $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ with $\mathbf{U} \in \mathbb{R}^{F \times F}$ being an orthogonal matrix and $\mathbf{D} \in \mathbb{R}^{F \times F}$ being a diagonal matrix. In addition, when $R := \mathrm{rank}(\mathbf{A}) < F$, the so-called compact SVD of $\mathbf{A}$ is written as $\mathbf{A} = \mathbf{U}_R\mathbf{D}_R\mathbf{U}_R^T$, where $\mathbf{U}_R := \mathbf{U}(:, 1\colon R)$ and $\mathbf{D}_R := \mathbf{\Sigma}(1\colon R, 1\colon R)$. For any dataset $\mathbf{V} \in \mathbb{R}^{F \times N}$ and any positive integer $k \leq F$, the so-called k-PCA for the dataset usually consists of two steps: (i) Obtain the centralized dataset $\mathbf{Z}$; (ii) Calculate the SVD of $\bar{\mathbf{\Sigma}}_N := \frac{1}{N}\mathbf{Z}\mathbf{Z}^T$, i.e., obtain $\bar{\mathbf{\Sigma}}_N = \mathbf{P}\mathbf{D}\mathbf{P}^T$, and project the dataset onto a k-dimensional space to obtain $\tilde{\mathbf{V}} := \mathbf{P}_k^T\mathbf{V}$, where $\mathbf{P}_k := \mathbf{P}(:, 1\colon k)$. For brevity, we say that $\tilde{\mathbf{V}}$ is the *post-k-PCA dataset* of $\mathbf{V}$ (or simply the *post-PCA* dataset). If only the projection step is performed (and not the centralizing step), we term the corresponding approach *k-PCA with no centering* or simply *k-SVD*, and we say that the corresponding $\tilde{\mathbf{V}}$ is the *post-k-SVD dataset* (or simply the *post-SVD* dataset) of $\mathbf{V}$.

When performing dimensionality reduction for clustering, it is important to compare any optimal clustering for the dataset with reduced dimensionality to any optimal clustering for the original dataset. More specifically, any optimal clustering for the dataset with reduced dimensionality should be close to any optimal clustering for the original dataset. However, existing works [60, 61] that combine k-means clustering and dimensionality reduction can only guarantee that the distortion of any optimal clustering for the dataset with reduced dimensionality, $\tilde{\mathscr{C}}^{\mathrm{opt}}$, can be bounded by a factor $\gamma > 1$ times the distortion of any optimal clustering for the original dataset, $\mathscr{C}^{\mathrm{opt}}$. That is,

$$\mathcal{D}(\mathbf{V}, \tilde{\mathscr{C}}^{\mathrm{opt}}) \leq \gamma\mathcal{D}(\mathbf{V}, \mathscr{C}^{\mathrm{opt}}). \tag{2.1.10}$$

As mentioned in [60], directly comparing the *structures* of $\tilde{\mathscr{C}}^{\mathrm{opt}}$ and $\mathscr{C}^{\mathrm{opt}}$ (instead of their distortions) is more interesting. In this chapter, we also prove that, if the samples are generated from a spherical GMM (or a mixture of log-concave distributions) which satisfies a separability assumption and the non-degeneracy condition, when the number of samples is sufficiently large, the ME distance between any optimal clustering for the original dataset and any optimal clustering for the post-PCA dataset can be bounded appropriately.

In addition, we can show that any optimal clustering of the dimensionality-reduced dataset is close to the correct target clustering by leveraging (2.1.10). This simple strategy seems to be adequate for data-independent dimensionality reduction techniques such as random projection. However, for data-dependent dimensionality reduction techniques such as PCA, we believe that it worth applying distinct proof techniques similar to those developed herein to obtain stronger theoretical results because of the generative models we assume. See Section 2.4.1 for a detailed discussion.

2.2 Error Bounds for Spherical GMMs

In this section, we assume the datasets are generated from spherical GMMs. Even though we can and will make statements for more general log-concave distributions (see Section 2.3), this assumption allows us to illustrate our results and mathematical ideas as cleanly as possible. We first present our main theorem for the upper bound of ME distance between any optimal clustering and the correct target clustering for the original dataset in Section 2.2.1. Then, in Section 2.2.2, we present our main theorem for the upper bound of ME distance between any optimal clustering and the correct target clustering for the dimensionality-reduced dataset. Finally, numerical results examining the correctness of our bounds are presented in Section 2.2.3.

2.2.1 The Theorem for Original Data

First, we show that with the combination of a new separability assumption and the non-degeneracy condition (cf. Definition 1) for a spherical GMM, any optimal clustering for a dataset generated from the spherical GMM is close to the correct target clustering with high probability when the number of samples is sufficiently large.

We adopt the following set of notations. Let $\mathbf{V}_1 \in \mathbb{R}^{F\times N_1}$ and $\mathbf{V}_2 \in \mathbb{R}^{F\times N_2}$, we denote by $[\mathbf{V}_1, \mathbf{V}_2]$ the horizontal concatenation of the two matrices. Let $\mathbf{\Sigma}_N :=$

$\frac{1}{N}\mathbf{V}\mathbf{V}^T$ and $\bar{\mathbf{\Sigma}}_N := \frac{1}{N}\mathbf{Z}\mathbf{Z}^T$, where $\mathbf{Z}$ is the centralized matrix of $\mathbf{V}$. Fix a mixture model with parameters $\{(w_k, \mathbf{u}_k, \mathbf{\Sigma}_k)\}_{k\in[K]}$ where w_k, $\mathbf{u}_k$ and $\mathbf{\Sigma}_k$ denote the mixing weight, the mean vector, and the covariance matrix of the k-th component. Let

$$\mathbf{\Sigma} := \sum_{k=1}^{K} w_k \left(\mathbf{u}_k\mathbf{u}_k^T + \mathbf{\Sigma}_k\right), \quad \text{and} \quad \mathbf{\Sigma}_0 := \sum_{k=1}^{K} w_k\mathbf{u}_k\mathbf{u}_k^T. \tag{2.2.1}$$

Denote $\bar{\mathbf{u}} := \sum_{k=1}^{K} w_k\mathbf{u}_k$ and write

$$\bar{\mathbf{\Sigma}} := \sum_{k=1}^{K} w_k \left((\mathbf{u}_k - \bar{\mathbf{u}})(\mathbf{u}_k - \bar{\mathbf{u}})^T + \mathbf{\Sigma}_k\right), \tag{2.2.2}$$

$$\text{and} \quad \bar{\mathbf{\Sigma}}_0 := \sum_{k=1}^{K} w_k(\mathbf{u}_k - \bar{\mathbf{u}})(\mathbf{u}_k - \bar{\mathbf{u}})^T, \tag{2.2.3}$$

and $\lambda_{\min} := \lambda_{K-1}(\bar{\mathbf{\Sigma}}_0)$. For a K-component spherical mixture model with covariance matrices $\sigma_k^2\mathbf{I}$ for $k \in [K]$, we write $\bar{\sigma}^2 := \sum_{k=1}^{K} w_k\sigma_k^2$.

For $p \in [0, \frac{1}{2}(K-1)]$, we define the function

$$\zeta(p) := \frac{p}{1 + \sqrt{1 - \frac{2p}{K-1}}}. \tag{2.2.4}$$

We have $\frac{1}{2}p \le \zeta(p) \le p$. Our first theorem reads:

Theorem 1. *Suppose all the columns of data matrix $\mathbf{V} \in \mathbb{R}^{F\times N}$ are independently generated from a K-component spherical GMM and $N > F > K$. Assume the spherical GMM satisfies the non-degeneracy condition. Let $w_{\min} := \min_k w_k$ and $w_{\max} := \max_k w_k$. Further assume*

$$\delta_0 := \frac{(K-1)\bar{\sigma}^2}{\lambda_{\min}} < \zeta(w_{\min}). \tag{2.2.5}$$

Let $\mathscr{C} := \{\mathscr{C}_1, \mathscr{C}_2, \dots, \mathscr{C}_K\}$ be the correct target K-clustering corresponding to the spherical GMM. Assume that $\epsilon > 0$ that satisfies

$$\epsilon \le \min\left\{\frac{w_{\min}}{2}, \lambda_{\min}, (K-1)\bar{\sigma}^2\right\} \quad \text{and} \quad \frac{(K-1)\bar{\sigma}^2 + \epsilon}{\lambda_{\min} - \epsilon} \le \zeta(w_{\min} - \epsilon). \tag{2.2.6}$$

Then for any $t \ge 1$, if $N \ge CF^5K^2t^2/\epsilon^2$, where $C > 0$ depends on $\{(w_k, \mathbf{u}_k, \sigma_k^2)\}_{k\in[K]}$, we have, with probability at least $1 - 36KF^2\exp(-t^2F)$,

$$\mathrm{d}_{\mathrm{ME}}(\mathscr{C}, \mathscr{C}^{\mathrm{opt}}) \le \tau\left(\frac{(K-1)\bar{\sigma}^2 + \epsilon}{\lambda_{\min} - \epsilon}\right)(w_{\max} + \epsilon), \tag{2.2.7}$$

where $\mathscr{C}^{\mathrm{opt}}$ is an optimal K-clustering for $\mathbf{V}$ and $\tau(\cdot)$ is defined in (2.2.12).

Remark 2. The condition (2.2.5) can be considered as a separability assumption. In particular, when $K = 2$, we have $\lambda_{\min} = w_1 w_2 \|\mathbf{u}_1 - \mathbf{u}_2\|_2^2$ and (2.2.5) becomes

$$\|\mathbf{u}_1 - \mathbf{u}_2\|_2 > \frac{\bar{\sigma}}{\sqrt{w_1 w_2 \zeta(w_{\min})}}, \tag{2.2.8}$$

which is similar to (2.1.1), the separability assumption of [17].

Remark 3. The separability condition in (2.2.5) is different from some other *pairwise separability* assumptions in the literature [16, 17, 51–53]. Our condition is a *global separability condition*. The intuitive reasons for this are twofold. First, we study the optimal solutions to the sum-of-squares distortion measure in (1.1.1). This is a global measure, involving all clusters, and so we believe a global separability condition is natural. Second, we leverage several technical lemmas in the literature, such as Lemma 1. These lemmas also involve global parameters such as $\lambda_{K-1}(\mathbf{S})$ and $\lambda_K(\mathbf{S})$, thus a global separability condition of the form of (2.2.5) seems unavoidable.

Remark 4. The non-degeneracy condition is used to ensure that $\lambda_{\min} > 0$. When $K = 2$, to ensure that $\lambda_{\min} > 0$, we only need to assume that the two component mean vectors are distinct. In particular, we do not require $\mathbf{u}_1$ and $\mathbf{u}_2$ to be linearly independent.

The proof of Theorem 1 is based on Corollary 1 and various concentration bounds. We need to make use of the following probabilistic lemmas. First, we present the following lemma from [62] that provides an upper bound for perturbation of eigenvalues when the matrix is perturbed.

Lemma 2. *If* $\mathbf{A}$ *and* $\mathbf{A} + \mathbf{E}$ *are in* $\mathbb{R}^{M \times M}$*, then*

$$|\lambda_m(\mathbf{A} + \mathbf{E}) - \lambda_m(\mathbf{A})| \le \|\mathbf{E}\|_2 \tag{2.2.9}$$

for any $m \in [M]$ *with* $\lambda_m(\mathbf{A})$ *being the* m*-th largest eigenvalue of* $\mathbf{A}$*.*

Because we will make use of the second-order moments of a mixture model, we present a simple lemma summarizing key facts.

Lemma 3. *Let $\mathbf{x}$ be a random sample from a K-component mixture model with parameters $\{(w_k, \mathbf{u}_k, \mathbf{\Sigma}_k\}_{k\in[K]}$. Then,*

$$\mathbb{E}\left(\mathbf{x}\mathbf{x}^T\right) = \sum_{k=1}^{K} w_k \left(\mathbf{u}_k\mathbf{u}_k^T + \mathbf{\Sigma}_k\right) = \mathbf{\Sigma}, \quad \textit{and} \tag{2.2.10}$$

$$\mathbb{E}\left((\mathbf{x}-\bar{\mathbf{u}})(\mathbf{x}-\bar{\mathbf{u}})^T\right) = \sum_{k=1}^{K} w_k \left((\mathbf{u}_k-\bar{\mathbf{u}})(\mathbf{u}_k-\bar{\mathbf{u}})^T + \mathbf{\Sigma}_k\right) = \bar{\mathbf{\Sigma}}. \tag{2.2.11}$$

To apply Corollary 1, we need to ensure that $\lambda_{K-1}(\mathbf{S}) - \lambda_K(\mathbf{S})$ which appears in the denominators of the expressions in (2.1.7) is positive. Note that if we assume all the columns of data matrix $\mathbf{V}$ are independently generated from a K-component spherical GMM, we have $\frac{1}{N}\left(\lambda_{K-1}(\mathbf{S}) - \lambda_K(\mathbf{S})\right) \xrightarrow{\text{p}} \lambda_{K-1}(\bar{\mathbf{\Sigma}}_0)$, where $\xrightarrow{\text{p}}$ represents convergence in probability as $N \to \infty$. Under the non-degeneracy condition, $\lambda_K(\mathbf{\Sigma}_0) > 0$. In addition, by the observation that $\bar{\mathbf{\Sigma}}_0 = \mathbf{\Sigma}_0 - \bar{\mathbf{u}}\bar{\mathbf{u}}^T$ and the following lemma in [62], we have $\lambda_{\min} = \lambda_{K-1}(\bar{\mathbf{\Sigma}}_0) \geq \lambda_K(\mathbf{\Sigma}_0) > 0$.

Lemma 4. *Suppose $\mathbf{B} = \mathbf{A} + \tau\mathbf{v}\mathbf{v}^T$ where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric, $\mathbf{v}$ has unit 2-norm (i.e., $\|\mathbf{v}\|_2 = 1$) and $\tau \in \mathbb{R}$. Then,*

$$\lambda_i(\mathbf{B}) \in \begin{cases} [\lambda_i(\mathbf{A}), \lambda_{i-1}(\mathbf{A})] & \textit{if} \quad \tau \geq 0,\, 2 \leq i \leq n \\ [\lambda_{i+1}(\mathbf{A}), \lambda_i(\mathbf{A})] & \textit{if} \quad \tau \leq 0,\, i \in [n-1] \end{cases}. \tag{2.2.12}$$

Furthermore, in order to obtain our probabilistic estimates, we need to make use of following concentration bounds for sub-Gaussian and sub-Exponential random variables. The definitions and relevant lemmas are extracted from [63].

Lemma 5. *(Hoeffding-type inequality) A* sub-Gaussian random variable *X is one such that $(\mathbb{E}|X|^p)^{1/p} \leq C\sqrt{p}$ for some $C > 0$ and for all $p \geq 1$. Let $X_1, \ldots, X_N$ be independent zero-mean sub-Gaussian random variables, then for every $\mathbf{a} = [a_1, a_2, \ldots, a_N]^T \in \mathbb{R}^N$ and every $t \geq 0$, it holds that*

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right) \leq \exp\left(1 - \frac{ct^2}{\|\mathbf{a}\|_2^2}\right), \tag{2.2.13}$$

where $c > 0$ is a constant.

Typical examples of sub-Gaussian random variables are Gaussian, Bernoulli and all bounded random variables. A random vector $\mathbf{x} \in \mathbb{R}^F$ is called sub-Gaussian if $\mathbf{x}^T\mathbf{z}$ is a sub-Gaussian random variable for any deterministic vector $\mathbf{z} \in \mathbb{R}^F$.

Lemma 6. *(Bernstein-type inequality) A* sub-Exponential random variable X *is one such that* $(\mathbb{E}|X|^p)^{1/p} \leq Cp$ *for some* $C > 0$ *and for all* $p \geq 1$*. Let* $X_1, \ldots, X_N$ *be independent zero-mean sub-Exponential random variables. It holds that*

$$\mathbb{P}\left(\left|\sum_{i=1}^N X_i\right| \geq \epsilon N\right) \leq 2\exp\left(-c \cdot \min\left(\frac{\epsilon^2}{M^2}, \frac{\epsilon}{M}\right)N\right), \tag{2.2.14}$$

where $c > 0$ *is an absolute constant and* $M > 0$ *is the maximum of the sub-Exponential norms*[2] *of* $\{X_i\}_{i=1}^N$*, i.e.,* $M = \max_{i\in[N]} \|X_i\|_{\Psi_1}$*.*

The set of sub-Exponential random variables includes those that have tails heavier than Gaussian. It is easy to see that a sub-Gaussian random variable is also sub-Exponential. The following lemma, which can be found in [63], is straightforward.

Lemma 7. *A random variable* X *is sub-Gaussian if and only if* X^2 *is sub-Exponential.*

Using this lemma, we see that Lemma 6 also provides a concentration bound for the sum of the squares of sub-Gaussian random variables. Finally, we can estimate empirical covariance matrices by the following lemma. Note that in this lemma, we do not need to assume that the sub-Gaussian distribution G in $\mathbb{R}^F$ has zero mean.

Lemma 8. *(Covariance estimation of sub-Gaussian distributions) Consider a sub-Gaussian distribution* G *in* $\mathbb{R}^F$ *with covariance matrix* $\mathbf{\Sigma}$*. Define the empirical covariance matrix* $\mathbf{\Sigma}_N := \frac{1}{N}\sum_{n=1}^N \mathbf{v}_n\mathbf{v}_n^T$ *where each* $\mathbf{v}_n$ *is an independent sample of* G*. Let* $\epsilon \in (0,1)$ *and* $t \geq 1$*. If* $N \geq C(t/\epsilon)^2F$ *for some constant* $C > 0$*, then with probability at least* $1 - 2\exp(-t^2F)$*,*

$$\|\mathbf{\Sigma}_N - \mathbf{\Sigma}\|_2 \leq \epsilon. \tag{2.2.15}$$

[2] The *sub-Exponential norm* of a sub-Exponential random variable X is defined as $\|X\|_{\Psi_1} := \sup_{p\geq1} p^{-1}(\mathbb{E}|X|^p)^{1/p}$.

The general idea of the proof of Theorem 1 is to first estimate the terms defining δ in (2.1.7) probabilistically. Next, we apply Corollary 1 to show that any optimal clustering for the original data matrix is close to the correct target clustering corresponding to the spherical GMM. Because the complete proof which contains various probabilistic estimates is somewhat lengthy, we provide the proof sketch below. Detailed calculations are deferred to the end of this section.

Proof Sketch of Theorem 1: We estimate every term in (2.1.7) for the correct target clustering $\mathscr{C}$. By Lemmas 5 and 6, we have for any $\epsilon \in (0,1)$,

$$\mathbb{P}\left(\left|\frac{1}{N}\mathcal{D}(\mathbf{V},\mathscr{C}) - F\bar{\sigma}^2\right| \geq \frac{\epsilon}{2}\right) \leq 2K((e+2)F+2)\mathrm{exp}\left(-C_1\frac{N\epsilon^2}{F^2K^2}\right), \tag{2.2.16}$$

where $C_1 > 0$ depends on $\{(w_k, \mathbf{u}_k, \sigma_k^2)\}_{k\in[K]}$. See the complete proof for the detailed calculation of this and other inequalities in this proof sketch. In particular, for the justification of (2.2.16), see the steps leading to (2.2.34). These simply involve the triangle inequality, the union bound, and careful probabilistic estimates. In addition, by Lemma 8, for any $t \geq 1$, if $N \geq C_2F^3K^2t^2/\epsilon^2$ (where $C_2 > 0$ also depends on $\{(w_k, \mathbf{u}_k, \sigma_k^2)\}_{k\in[K]}$),

$$\mathbb{P}\left(\|\bar{\mathbf{\Sigma}}_N - \bar{\mathbf{\Sigma}}\|_2 \geq \frac{\epsilon}{2}\right) \leq (9FKe+2K)\exp\left(-t^2F\right). \tag{2.2.17}$$

Therefore, by the matrix perturbation inequalities in Lemma 16, when $N \geq C_2F^3K^2t^2/\epsilon^2$, we have

$$\mathbb{P}\left(\left|\frac{1}{N}\lambda_{K-1}(\mathbf{S}) - \left(\lambda_{\min}+\bar{\sigma}^2\right)\right| \geq \frac{\epsilon}{2}\right) \leq (9FKe+2K)\exp\left(-t^2F\right), \tag{2.2.18}$$

$$\mathbb{P}\left(\left|\frac{1}{N}\lambda_{K}(\mathbf{S}) - \bar{\sigma}^2\right| \geq \frac{\epsilon}{2}\right) \leq (9FKe+2K)\exp\left(-t^2F\right). \tag{2.2.19}$$

Furthermore, if $N \geq C_2F^5K^2t^2/\epsilon^2$, we have

$$\mathbb{P}\left(\left|\frac{1}{N}\mathcal{D}^*(\mathbf{V}) - (F-K+1)\bar{\sigma}^2\right| \geq \frac{\epsilon}{2}\right) \\ \leq (F-K+1)(9FKe+2K)\exp(-t^2F). \tag{2.2.20}$$

Combining these results, appealing to Corollary 1, the union bound, and the property that both $\tau(\cdot)$ and $\zeta(\cdot)$ are continuous and monotonically increasing functions on $[0, \frac{1}{2}(K-1)]$, we obtain (2.2.7) as desired. □

Note that both $\tau(\cdot)$ and $\zeta(\cdot)$ are continuous and monotonically increasing on $[0, \frac{1}{2}(K-1)]$. When the mixing weights are skewed (leading to a small $w_{\min}$), we require a strong separability assumption in (2.2.6). This is consistent with the common knowledge [64] that imbalanced clusters are more difficult to disambiguate for k-means. When δ_0 is small (i.e., the data is well-separated) and N is large (so ϵ and t can be chosen to be sufficiently small and large respectively), we have with probability close to 1 that the upper bound on the ME distance given by (2.2.7) is close to 0.

When the ME distance between any optimal clustering for k-means $\mathscr{C}^{\mathrm{opt}}$ and the correct target clustering $\mathscr{C}$ of the samples generated from a spherical GMM is small (and thus the implicit assumption of k-means is satisfied), we can readily perform k-means to find $\mathscr{C}^{\mathrm{opt}}$ to infer $\mathscr{C}$. The tightness of the upper bound in (2.2.7) is assessed numerically in Section 2.2.3.

Remark 5. The result by [56] (discussed in Section 2.1.1) may, at a first glance, appear to be similar to Theorem 1 in the sense that both results show that under appropriate conditions, k-means is a good choice for learning certain mixture models. However, there is a salient difference. The analysis of [56] is based on a variant of k-means algorithm, while we only analyze the *objective function* of k-means (in (1.1.1)) which determines all optimal clusterings. Since there are multiple ways to approximately minimize the ubiquitous but intractable sum-of-squares distortion measure in (1.1.1), our analysis is partly *algorithm-independent* and thus fundamental in the theory of clustering. Our analysis and theoretical results, in fact, *underpin why* the separability assumptions of various forms appear to be necessary to make theoretical guarantees for using k-means to learn mixture models.

Finally, we present the complete proof of Theorem 1.

Complete Proof of Theorem 1: To apply Corollary 1, we first estimate $\frac{1}{N}\mathcal{D}(\mathbf{V}, \mathscr{C})$.

We have

$$\frac{1}{N}\mathcal{D}(\mathbf{V},\mathscr{C}) = \frac{1}{N}\sum_{k=1}^{K}\sum_{n\in\mathscr{C}_k}\|\mathbf{v}_n-\mathbf{c}_k\|_2^2 = \frac{1}{N}\sum_{k=1}^{K}\left(\sum_{n\in\mathscr{C}_k}\|\mathbf{v}_n\|_2^2 - n_k\|\mathbf{c}_k\|_2^2\right) \tag{2.2.21}$$

$$= \sum_{k=1}^{K}\frac{n_k}{N}\left(\frac{\sum_{n\in\mathscr{C}_k}\|\mathbf{v}_n\|_2^2}{n_k} - \|\mathbf{c}_k\|_2^2\right), \tag{2.2.22}$$

where $n_k := |\mathscr{C}_k|$. By Lemma 5 and by the property that if X has a sub-Gaussian distribution,[3] $\|X-\mathbb{E}X\|_{\Psi_2}\le 2\|X\|_{\Psi_2}$ (similarly, if X has a sub-Exponential distribution, $\|X-\mathbb{E}X\|_{\Psi_1}\le 2\|X\|_{\Psi_1}$) [63], we have for $k\in[K]$,

$$\mathbb{P}\left(\left|\frac{n_k}{N}-w_k\right|\ge\frac{w_k}{2}\right)\le e\exp(-C_0N), \tag{2.2.23}$$

where $C_0>0$ is a constant depending on w_k, $k\in[K]$. Then with probability at least $1-Ke\exp(-C_0N)$, we have $\frac{n_k}{N}\ge\frac{w_k}{2}$ for all $k\in[K]$. For brevity, we only consider this case and replace n_k with N in the following inequalities. In addition, for any $0<\epsilon\le\frac{w_{\min}}{2}<1$,

$$\mathbb{P}\left(\left|\frac{n_k}{N}-w_k\right|\ge\epsilon\right)\le e\exp(-\tau_kN\epsilon^2), \tag{2.2.24}$$

where $\tau_k>0$ is a constant depending on w_k.

By Lemma 7, the square of each entry of a random vector generated from a spherical Gaussian distribution has a sub-Exponential distribution. In addition, for random vector $\mathbf{v}$ generated from the k-th component of the spherical GMM, we have $\mathbb{E}[\mathbf{v}(f)^2]=\mathbf{u}_k(f)^2+\sigma_k^2$ for any $f\in[F]$. By Lemma 6, for any $k\in[K]$,

$$\mathbb{P}\left(\left|\frac{1}{n_k}\sum_{n\in\mathscr{C}_k}\|\mathbf{v}_n\|_2^2-\left(\|\mathbf{u}_k\|_2^2+F\sigma_k^2\right)\right|\ge\epsilon\right)$$

$$\le\sum_{f=1}^{F}\mathbb{P}\left(\left|\frac{1}{n_k}\sum_{n\in\mathscr{C}_k}\mathbf{v}_n(f)^2-\left(\mathbf{u}_k(f)^2+\sigma_k^2\right)\right|\ge\frac{\epsilon}{F}\right) \tag{2.2.25}$$

$$\le 2F\exp\left(-\xi_k\frac{N\epsilon^2}{F^2}\right), \tag{2.2.26}$$

[3] The *sub-Gaussian norm* of a sub-Gaussian random variable X is defined as $\|X\|_{\Psi_2} := \sup_{p\ge1}p^{-1/2}\left(\mathbb{E}|X|^p\right)^{1/p}$.

where $\xi_k > 0$ is a constant depending on $w_k, \sigma_k^2, \mathbf{u}_k$. Similarly, by Lemma 5, there exists $\zeta_k > 0$ depending on w_k, σ_k^2 and $\mathbf{u}_k$, such that

$$\mathbb{P}\left(\left|\|\mathbf{c}_k\|_2^2 - \|\mathbf{u}_k\|_2^2\right| \geq \epsilon\right) \leq \sum_{f=1}^{F} \mathbb{P}\left(\left|\mathbf{c}_k(f)^2 - \mathbf{u}_k(f)^2\right| \geq \frac{\epsilon}{F}\right) \tag{2.2.27}$$

$$\leq 2eF \exp\left(-\zeta_k \frac{N\epsilon^2}{F^2}\right). \tag{2.2.28}$$

The final bound in (2.2.28) holds because for any f, if $\mathbf{u}_k(f) = 0$, we have

$$\mathbb{P}\left(\left|\mathbf{c}_k(f)^2 - \mathbf{u}_k(f)^2\right| \geq \frac{\epsilon}{F}\right) = \mathbb{P}\left(|\mathbf{c}_k(f)| \geq \sqrt{\frac{\epsilon}{F}}\right). \tag{2.2.29}$$

On the other hand, if $\mathbf{u}_k(f) \neq 0$,

$$\begin{aligned}&\mathbb{P}\left(\left|\mathbf{c}_k(f)^2 - \mathbf{u}_k(f)^2\right| \geq \frac{\epsilon}{F}\right)\\ &\leq \mathbb{P}\left(|\mathbf{c}_k(f) - \mathbf{u}_k(f)| \geq \frac{\epsilon}{3|\mathbf{u}_k(f)|F}\right) + \mathbb{P}\left(|\mathbf{c}_k(f) + \mathbf{u}_k(f)| \geq 3|\mathbf{u}_k(f)|\right).\end{aligned} \tag{2.2.30}$$

Now let $d_k := \frac{1}{n_k}\sum_{n \in \mathscr{C}_k} \|\mathbf{v}_n\|_2^2 - \|\mathbf{c}_k\|_2^2$ and $L_k := \|\mathbf{u}_k\|_2^2 + F\sigma_k^2$. Then, there exists a constant $C_1 > 0$ depending on $\{(w_k, \sigma_k^2, \mathbf{u}_k)\}_{k \in [K]}$ such that

$$\mathbb{P}\left(\left|\frac{1}{N}\mathcal{D}(\mathbf{V}, \mathscr{C}) - F\bar{\sigma}^2\right| \geq \frac{\epsilon}{2}\right) \leq \sum_{k=1}^{K} \mathbb{P}\left(\left|\frac{n_k}{N}d_k - w_k F\sigma_k^2\right| \geq \frac{\epsilon}{2K}\right) \tag{2.2.31}$$

$$\leq \sum_{k=1}^{K} \mathbb{P}\left(\left|\frac{n_k}{N} - w_k\right| F\sigma_k^2 \geq \frac{\epsilon}{4K}\right) + \sum_{k=1}^{K} \mathbb{P}\left(\left|d_k - F\sigma_k^2\right| \frac{n_k}{N} \geq \frac{\epsilon}{4K}\right) \tag{2.2.32}$$

$$\begin{aligned}&\leq \sum_{k=1}^{K} \mathbb{P}\left(\left|\frac{n_k}{N} - w_k\right| \geq \frac{\epsilon}{4KF\sigma_k^2}\right) + \sum_{k=1}^{K} \mathbb{P}\left(\frac{n_k}{N} \geq 2w_k\right)\\ &\quad + \sum_{k=1}^{K} \mathbb{P}\left(\left|\frac{1}{n_k}\sum_{n \in \mathscr{C}_k} \|\mathbf{v}_n\|_2^2 - L_k\right| \geq \frac{\epsilon}{16 w_k K}\right)\\ &\quad + \sum_{k=1}^{K} \mathbb{P}\left(\left|\|\mathbf{c}_k\|_2^2 - \|\mathbf{u}_k\|_2^2\right| \geq \frac{\epsilon}{16 w_k K}\right)\end{aligned} \tag{2.2.33}$$

$$\leq 2K\left((e+2)F + e\right)\exp\left(-C_1 \frac{N\epsilon^2}{F^2K^2}\right), \tag{2.2.34}$$

where (2.2.31) is a consequence of (2.2.22). This proves (2.2.16).

Now we estimate the positive eigenvalues of $\mathbf{S} := \mathbf{Z}^T\mathbf{Z}$, where $\mathbf{Z}$ is the centralized data matrix of $\mathbf{V}$. Equivalently, we may consider the eigenvalues of $\bar{\mathbf{\Sigma}}_N := \frac{1}{N}\mathbf{Z}\mathbf{Z}^T =$

$\frac{1}{N}\sum_{n=1}^{N}(\mathbf{v}_n - \bar{\mathbf{v}})(\mathbf{v}_n - \bar{\mathbf{v}})^T$, where $\bar{\mathbf{v}} = \frac{1}{N}\sum_n \mathbf{v}_n$. The expectation of centralized covariance matrix for the spherical GMM is $\sum_{k=1}^{K} w_k(\mathbf{u}_k - \bar{\mathbf{u}})(\mathbf{u}_k - \bar{\mathbf{u}})^T + \bar{\sigma}^2\mathbf{I} = \bar{\mathbf{\Sigma}}$. For any $f \in [F]$,

$$\mathbb{P}\left(|\bar{\mathbf{v}}(f) - \bar{\mathbf{u}}(f)| \geq \epsilon\right) \leq \sum_{k=1}^{K} \mathbb{P}\left(\left|\frac{n_k}{N} \cdot \frac{\sum_{n\in\mathscr{C}_k} \mathbf{v}_n(f)}{n_k} - w_k\mathbf{u}_k(f)\right| \geq \frac{\epsilon}{K}\right) \tag{2.2.35}$$

$$\leq 3Ke\exp\left(-C_2\frac{N\epsilon^2}{K^2}\right), \tag{2.2.36}$$

and

$$\mathbb{P}\left(\|\bar{\mathbf{v}}\bar{\mathbf{v}}^T - \bar{\mathbf{u}}\bar{\mathbf{u}}^T\|_2 \geq \frac{\epsilon}{2}\right)$$

$$\leq \mathbb{P}\left(\|(\bar{\mathbf{v}} - \bar{\mathbf{u}})(\bar{\mathbf{v}} - \bar{\mathbf{u}})^T\|_2 \geq \frac{\epsilon}{6}\right) + \mathbb{P}\left(\|(\bar{\mathbf{v}} - \bar{\mathbf{u}})\bar{\mathbf{u}}^T\|_2 \geq \frac{\epsilon}{6}\right)$$

$$+ \mathbb{P}\left(\|\bar{\mathbf{u}}(\bar{\mathbf{v}} - \bar{\mathbf{u}})^T\|_2 \geq \frac{\epsilon}{6}\right) \tag{2.2.37}$$

$$\leq \mathbb{P}\left(\|\bar{\mathbf{v}} - \bar{\mathbf{u}}\|_2^2 \geq \frac{\epsilon}{6}\right) + 2\mathbb{P}\left(\|\bar{\mathbf{v}} - \bar{\mathbf{u}}\|_2 \geq \frac{\epsilon}{6\|\bar{\mathbf{u}}\|_2}\right). \tag{2.2.38}$$

Hence, similarly to (2.2.28), and by (2.2.36), we obtain

$$\mathbb{P}\left(\|\bar{\mathbf{v}}\bar{\mathbf{v}}^T - \bar{\mathbf{u}}\bar{\mathbf{u}}^T\|_2 \geq \frac{\epsilon}{2}\right) \leq 9FKe\exp\left(-C_3\frac{N\epsilon^2}{F^2K^2}\right), \tag{2.2.39}$$

where $C_3 > 0$ is a constant depending on $\{(w_k, \mathbf{u}_k, \sigma_k^2)\}_{k\in[K]}$. Note that

$$\mathbf{\Sigma}_N - \mathbf{\Sigma} = \sum_{k=1}^{K} \frac{n_k}{N} \sum_{n\in\mathscr{C}_k} \frac{\mathbf{v}_n\mathbf{v}_n^T}{n_k} - \sum_{k=1}^{K} w_k(\mathbf{u}_k\mathbf{u}_k^T + \sigma_k^2\mathbf{I}). \tag{2.2.40}$$

Then similar to (2.2.33) and by Lemma 8, we have that for any $t \geq 1$, if $N \geq C_4F^3K^2t^2/\epsilon^2$, where $C_4 > 0$ is a constant depending on $\{(w_k, \mathbf{u}_k, \sigma_k^2)\}_{k\in[K]}$,

$$\mathbb{P}\left(\|\bar{\mathbf{\Sigma}}_N - \bar{\mathbf{\Sigma}}\|_2 \geq \frac{\epsilon}{2}\right) = \mathbb{P}\left(\left\|(\mathbf{\Sigma}_N - \mathbf{\Sigma}) - \left(\bar{\mathbf{v}}\bar{\mathbf{v}}^T - \bar{\mathbf{u}}\bar{\mathbf{u}}^T\right)\right\|_2 \geq \frac{\epsilon}{2}\right) \tag{2.2.41}$$

$$\leq \mathbb{P}\left(\|\mathbf{\Sigma}_N - \mathbf{\Sigma}\|_2 \geq \frac{\epsilon}{4}\right) + \mathbb{P}\left(\|\bar{\mathbf{v}}\bar{\mathbf{v}}^T - \bar{\mathbf{u}}\bar{\mathbf{u}}^T\|_2 \geq \frac{\epsilon}{4}\right) \tag{2.2.42}$$

$$\leq (2K + 9FKe)\exp\left(-t^2F\right). \tag{2.2.43}$$

This proves (2.2.17).

Now, if $N \geq C_4 F^5 K^2 t^2/\epsilon^2$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{N}\mathcal{D}^*(\mathbf{V}) - (F-K+1)\bar{\sigma}^2\right| \geq \frac{\epsilon}{2}\right) \\
&= \mathbb{P}\left(\left|\frac{1}{N}\sum_{k=K}^{F} \lambda_k(\mathbf{S}) - (F-K+1)\bar{\sigma}^2\right| \geq \frac{\epsilon}{2}\right) && (2.2.44)\\
&= \mathbb{P}\left(\left|\sum_{k=K}^{F} \lambda_k(\bar{\mathbf{\Sigma}}_N) - (F-K+1)\bar{\sigma}^2\right| \geq \frac{\epsilon}{2}\right) && (2.2.45)\\
&\leq \sum_{k=K}^{F} \mathbb{P}\left(|\lambda_k(\bar{\mathbf{\Sigma}}_N) - \lambda_k(\bar{\mathbf{\Sigma}})| \geq \frac{\epsilon}{2(F-K+1)}\right) && (2.2.46)\\
&\leq (F-K+1)\mathbb{P}\left(\|\bar{\mathbf{\Sigma}}_N - \bar{\mathbf{\Sigma}}\|_2 \geq \frac{\epsilon}{2(F-K+1)}\right) && (2.2.47)\\
&\leq (F-K+1)(9FKe+2K)\exp(-t^2F). && (2.2.48)
\end{aligned}$$

This proves (2.2.20).

In addition, if $N \geq C_4 F^3 K^2 t^2/\epsilon^2$, similarly,

$$\mathbb{P}\left(\left|\frac{1}{N}\lambda_{K-1}(\mathbf{S}) - \left(\lambda_{K-1}(\bar{\mathbf{\Sigma}}_0) + \bar{\sigma}^2\right)\right| \geq \frac{\epsilon}{2}\right) \leq (9FKe+2K)\exp\left(-t^2F\right), \tag{2.2.49}$$

$$\mathbb{P}\left(\left|\frac{1}{N}\lambda_K(\mathbf{S}) - \bar{\sigma}^2\right| \geq \frac{\epsilon}{2}\right) \leq (9FKe+2K)\exp\left(-t^2F\right). \tag{2.2.50}$$

This proves (2.2.18) and (2.2.19).

Finally, let $p_{\min} = \frac{1}{N}\min_k |\mathscr{C}_k|$ and $p_{\max} = \frac{1}{N}\max_k |\mathscr{C}_k|$. Then if $\epsilon > 0$ satisfies (2.2.6), when $N \geq CF^5K^2t^2/\epsilon^2$ with $C > 0$ being a constant depending on $\{(w_k, \sigma_k^2, \mathbf{u}_k)\}_{k\in[K]}$, with probability at least $1 - 36F^2K\exp\left(-t^2F\right)$,

$$\delta := \frac{\mathcal{D}(\mathbf{V},\mathscr{C}) - \mathcal{D}^*(\mathbf{V})}{\lambda_{K-1}(\mathbf{S}) - \lambda_K(\mathbf{S})} \leq \frac{(K-1)\bar{\sigma}^2 + \epsilon}{\lambda_{\min} - \epsilon} \tag{2.2.51}$$

Thus by (2.2.6),

$$\delta \leq \zeta(w_{\min} - \epsilon) \leq \zeta(p_{\min}). \tag{2.2.52}$$

This is equivalent to

$$\tau(\delta) \leq p_{\min}. \tag{2.2.53}$$

Therefore, by Corollary 1, if $N \geq CF^5K^2t^2/\epsilon^2$ (where $C > 0$ is an appropriate constant depending on $\{(w_k, \mathbf{u}_k, \sigma_k^2)\}_{k\in[K]}$),

$$\begin{aligned} \mathrm{d}_{\mathrm{ME}}(\mathscr{C}, \mathscr{C}^{\mathrm{opt}}) &\leq \tau(\delta)p_{\max} \leq \tau(\delta)(w_{\max} + \epsilon) && (2.2.54) \\ &\leq \tau\left(\frac{(K-1)\bar{\sigma}^2 + \epsilon}{\lambda_{\min} - \epsilon}\right)(w_{\max} + \epsilon) && (2.2.55) \end{aligned}$$

with probability at least $1 - 36KF^2 \exp(-t^2F)$. □

2.2.2 The Theorem for Post-PCA Data

Next, we show that under similar assumptions for the generating process and with a *weaker* separability assumption for the spherical GMM, any optimal clustering for the post-PCA dataset is also close to the correct target clustering with high probability when N is large enough.

Theorem 2. *Let the dataset* $\mathbf{V} \in \mathbb{R}^{F\times N}$ *be generated under the same conditions given in Theorem 1 with the separability assumption (2.2.5) being modified to*

$$\delta_1 := \frac{(K-1)\bar{\sigma}^2}{\lambda_{\min} + \bar{\sigma}^2} < \zeta(w_{\min}). \qquad (2.2.56)$$

Let $\tilde{\mathbf{V}} \in \mathbb{R}^{(K-1)\times N}$ *be the post-*$(K-1)$*-PCA dataset of* $\mathbf{V}$*. Then, for any* $\epsilon > 0$ *that satisfies*

$$\epsilon \leq \min\left\{\frac{w_{\min}}{2}, \lambda_{\min} + \bar{\sigma}^2, (K-1)\bar{\sigma}^2\right\}, \quad \text{and} \quad \frac{(K-1)\bar{\sigma}^2 + \epsilon}{\lambda_{\min} + \bar{\sigma}^2 - \epsilon} \leq \zeta(w_{\min} - \epsilon), \qquad (2.2.57)$$

and for any $t \geq 1$*, when* $N \geq CF^3K^5t^2/\epsilon^2$*, where* $C > 0$ *depends on* $\{(w_k, \mathbf{u}_k, \sigma_k^2)\}_{k\in[K]}$*, we have, with probability at least* $1 - 165KF\exp(-t^2K)$*,*

$$\mathrm{d}_{\mathrm{ME}}(\mathscr{C}, \tilde{\mathscr{C}}^{\mathrm{opt}}) \leq \tau\left(\frac{(K-1)\bar{\sigma}^2 + \epsilon}{\lambda_{\min} + \bar{\sigma}^2 - \epsilon}\right)(w_{\max} + \epsilon), \qquad (2.2.58)$$

where $\mathscr{C}$ *is the correct target clustering and* $\tilde{\mathscr{C}}^{\mathrm{opt}}$ *is an optimal* K*-clustering for* $\tilde{\mathbf{V}}$.

Remark 6. Vempala and Wang [17] analyze the SVD of $\mathbf{\Sigma}_N$ (corresponding to PCA with no centering) instead of $\bar{\mathbf{\Sigma}}_N$ (corresponding to PCA). The proof of Corollary

3 in [17] is based on the key observation that the subspace spanned by the first K singular vectors of $\mathbf{\Sigma}_N$ lies close to the subspace spanned by the K component mean vectors of the spherical GMM with high probability when N is large. By performing K-SVD (cf. Section 2.1.3) on $\mathbf{\Sigma}_N$, we have the following corollary.

Corollary 2. *Let the dataset $\mathbf{V} \in \mathbb{R}^{F\times N}$ be generated under the same conditions given in Theorem 1. Let $\tilde{\mathbf{V}}$ be the post-K-SVD dataset of $\mathbf{V}$, then for any positive ϵ satisfying (2.2.6) and for any $t \geq 1$, if $N \geq CF^3K^5t^2/\epsilon^2$, then with probability at least $1-167KF\exp\left(-t^2K\right)$, the same upper bound in (2.2.7) holds for $\mathrm{d}_{\mathrm{ME}}(\mathscr{C}, \tilde{\mathscr{C}}^{\mathrm{opt}})$, where $\tilde{\mathscr{C}}^{\mathrm{opt}}$ is an optimal K-clustering for $\tilde{\mathbf{V}}$.*

Combining the results of Theorems 1 and 2, by the triangle inequality for ME distance, we obtain the following corollary concerning an upper bound for $\mathrm{d}_{\mathrm{ME}}(\mathscr{C}^{\mathrm{opt}}, \tilde{\mathscr{C}}^{\mathrm{opt}})$, the ME distance between any optimal clustering of the original dataset and any optimal clustering of the post-PCA dataset.

Corollary 3. *Let the dataset $\mathbf{V} \in \mathbb{R}^{F\times N}$ be generated under the same conditions given in Theorem 1. Let $\tilde{\mathbf{V}}$ be the post-$(K-1)$-PCA dataset of $\mathbf{V}$, then for any positive ϵ satisfying (2.2.6) and for any $t \geq 1$, if $N \geq CF^5K^5t^2/\epsilon^2$, then with probability at least $1-201KF^2\exp\left(-t^2K\right)$, $\mathrm{d}_{\mathrm{ME}}(\mathscr{C}^{\mathrm{opt}}, \tilde{\mathscr{C}}^{\mathrm{opt}})$ is upper bounded by the sum of the right-hand-sides of (2.2.7) and (2.2.58).*

The proof of Theorem 2 hinges mainly on the fact that the subspace spanned by the first $K-1$ singular vectors of $\bar{\mathbf{\Sigma}}_N$ is "close" to the subspace spanned by the first $K-1$ singular vectors of $\bar{\mathbf{\Sigma}}_0$. See Lemma 9 to follow for a precise statement. Note that the assumption in (2.2.56) is weaker than (2.2.5) and the upper bound given by (2.2.58) is smaller than that in (2.2.7) (if all the parameters are the same). In addition, when $K = 2$, by applying PCA to the original dataset as described in Theorem 2, we obtain a 1-dimensional dataset, which is easier to cluster optimally compared to the 2-dimensional dataset obtained by performing PCA with no centering as described in Remark 6. These comparisons also provide a theoretical basis for the fact that centering can result in a stark difference in PCA.

Now, we prove Theorem 2. Following the notations in Section 2.1.3, we write $\bar{\mathbf{\Sigma}}_N = \mathbf{P}\mathbf{D}\mathbf{P}^T$, $\mathbf{P}_{K-1} = \mathbf{P}(:\,,1\colon K-1)$, and $\mathbf{P}_{-(K-1)} = \mathbf{P}(:\,,K\colon F)$. We also denote $\tilde{\mathbf{V}} = \mathbf{P}_{K-1}^T\mathbf{V}$ as the post-$(K-1)$-PCA dataset of $\mathbf{V}$. Instead of using $\mathbf{P}_{K-1}$ which is correlated to the samples, we consider the SVD of $\bar{\mathbf{\Sigma}}_0$ and project the original data matrix onto $\mathbb{R}^{K-1}$ using the first $K-1$ singular vectors of $\bar{\mathbf{\Sigma}}_0$. We can similarly estimate the terms in (2.1.7) for the corresponding $(K-1)$-dimensional spherical GMM. Furthermore, we estimate the difference between the results obtained from projecting the original data matrix onto $\mathbb{R}^{K-1}$ using the first $K-1$ singular vectors of $\bar{\mathbf{\Sigma}}_0$ and the results obtained from projecting the original data matrix onto $\mathbb{R}^{K-1}$ using the columns of $\mathbf{P}_{K-1}$.

Proof of Theorem 2: By the non-degeneracy condition and Lemma 4, we have $\mathrm{rank}(\bar{\mathbf{\Sigma}}_0) = K-1$. Let the compact SVD of $\bar{\mathbf{\Sigma}}_0$ be

$$\bar{\mathbf{\Sigma}}_0 = \mathbf{Q}_{K-1}\mathbf{E}_{K-1}\mathbf{Q}_{K-1}^T, \tag{2.2.59}$$

where $\mathbf{Q}_{K-1} \in \mathbb{R}^{F\times(K-1)}$ has orthonormal columns and $\mathbf{E}_{K-1} \in \mathbb{R}^{(K-1)\times(K-1)}$ is a diagonal matrix. Since $\mathbf{Q}_{K-1}^T\mathbf{Q}_{K-1} = \mathbf{I}$, by the property of Gaussians, we know if $\mathbf{x}$ is a random vector with a spherical Gaussian distribution $\mathcal{N}(\mathbf{u},\sigma^2\mathbf{I})$ in $\mathbb{R}^F$, then $\mathbf{Q}_{K-1}^T\mathbf{x}$ is a random vector with a spherical Gaussian distribution $\mathcal{N}(\mathbf{Q}_{K-1}^T\mathbf{u},\sigma^2\mathbf{I})$ in $\mathbb{R}^{K-1}$. Let $\hat{\mathbf{V}} := \mathbf{Q}_{K-1}^T\mathbf{V}$, $\hat{\mathbf{Z}}$ be the centralized matrix of $\hat{\mathbf{V}}$ and $\hat{\mathbf{S}} := \hat{\mathbf{Z}}^T\hat{\mathbf{Z}}$. Denote $\hat{\bar{\mathbf{u}}} = \mathbf{Q}_{K-1}^T\bar{\mathbf{u}}$ and $\hat{\mathbf{u}}_k = \mathbf{Q}_{K-1}^T\mathbf{u}_k$ for all $k \in [K]$. Let $\hat{\bar{\mathbf{\Sigma}}}_0 := \sum_{k=1}^K w_k(\hat{\mathbf{u}}_k - \hat{\bar{\mathbf{u}}})(\hat{\mathbf{u}}_k - \hat{\bar{\mathbf{u}}})^T$. Define $\mathbf{X} := [\sqrt{w_1}(\mathbf{u}_1 - \bar{\mathbf{u}}),\dots,\sqrt{w_K}(\mathbf{u}_K - \bar{\mathbf{u}})] \in \mathbb{R}^{F\times K}$ and let $\hat{\mathbf{X}} := \mathbf{Q}_{K-1}^T\mathbf{X}$. Select $\mathbf{Q}_{-(K-1)} \in \mathbb{R}^{F\times(F-K+1)}$ such that $[\mathbf{Q}_{K-1},\mathbf{Q}_{-(K-1)}]$ is an orthogonal matrix. We have

$$\mathbf{X}^T\mathbf{X} - \hat{\mathbf{X}}^T\hat{\mathbf{X}} = \mathbf{X}^T\mathbf{X} - \mathbf{X}^T(\mathbf{Q}_{K-1}\mathbf{Q}_{K-1}^T)\mathbf{X} = \mathbf{X}^T\mathbf{Q}_{-(K-1)}\mathbf{Q}_{-(K-1)}^T\mathbf{X} = 0. \tag{2.2.60}$$

Thus we have

$$\lambda_{\min} = \lambda_{K-1}(\mathbf{X}\mathbf{X}^T) = \lambda_{K-1}(\mathbf{X}^T\mathbf{X}) = \lambda_{K-1}(\hat{\mathbf{X}}^T\hat{\mathbf{X}}) = \lambda_{K-1}(\hat{\bar{\mathbf{\Sigma}}}_0). \tag{2.2.61}$$

Then similar to that in Theorem 1, for any $\epsilon \in (0,1)$,

$$\mathbb{P}\left(\left|\frac{1}{N}\mathcal{D}(\hat{\mathbf{V}},\mathscr{C}) - (K-1)\bar{\sigma}^2\right| \geq \frac{\epsilon}{2}\right) \leq 2K((e+2)K+e)\exp\left(-C_3\frac{N\epsilon^2}{K^4}\right). \tag{2.2.62}$$

In addition, for any $t \geq 1$, if $N \geq C_4K^5t^2/\epsilon^2$,

$$\mathbb{P}\left(\left|\frac{1}{N}\lambda_{K-1}(\hat{\mathbf{S}}) - (\lambda_{\min}+\bar{\sigma}^2)\right| \geq \frac{\epsilon}{2}\right) \leq 9(eK^2+2K)e\exp(-t^2K), \tag{2.2.63}$$

where $C_3, C_4 > 0$ depend on $\{(w_k, \mathbf{u}_k, \sigma_k^2)\}_{k\in[K]}$. Note that $\frac{1}{N}\mathcal{D}^*(\hat{\mathbf{V}}) = \frac{1}{N}\lambda_K(\hat{\mathbf{S}}) = 0$. Thus, we only need to estimate $\frac{1}{N}\left|\mathcal{D}(\hat{\mathbf{V}},\mathscr{C}) - \mathcal{D}(\tilde{\mathbf{V}},\mathscr{C})\right|$ and $\frac{1}{N}\left|\lambda_{K-1}(\hat{\mathbf{S}}) - \lambda_{K-1}(\tilde{\mathbf{S}})\right|$, where $\tilde{\mathbf{S}} := \tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{Z}}$ is the centralized matrix of $\tilde{\mathbf{V}}$. By (2.1.2) and writing $\mathbf{R} := \mathbf{Q}_{K-1}\mathbf{Q}_{K-1}^T - \mathbf{P}_{K-1}\mathbf{P}_{K-1}^T$, we have

$$\frac{1}{N}\left|\mathcal{D}(\hat{\mathbf{V}},\mathscr{C}) - \mathcal{D}(\tilde{\mathbf{V}},\mathscr{C})\right| = \left|\left\langle\frac{1}{N}\left(\mathbf{V}\mathbf{V}^T - \mathbf{V}\bar{\mathbf{H}}^T\bar{\mathbf{H}}\mathbf{V}^T\right), \mathbf{R}\right\rangle\right| \tag{2.2.64}$$

$$\leq \left(1 + \|\bar{\mathbf{H}}^T\|_{\mathrm{F}}^2\right)\left(\frac{1}{N}\|\mathbf{V}\|_{\mathrm{F}}^2\right)\|\mathbf{R}\|_{\mathrm{F}} \tag{2.2.65}$$

$$= (1+K)\left(\frac{1}{N}\|\mathbf{V}\|_{\mathrm{F}}^2\right)\|\mathbf{R}\|_{\mathrm{F}}. \tag{2.2.66}$$

Note that

$$\mathbb{E}\left[\frac{1}{N}\|\mathbf{V}\|_{\mathrm{F}}^2\right] = \sum_{k=1}^K w_k\left(\|\mathbf{u}_k\|_2^2 + F\sigma_k^2\right). \tag{2.2.67}$$

In addition, by Lemma 16 and routine calculations,

$$\frac{1}{N}\left|\lambda_{K-1}(\hat{\mathbf{S}}) - \lambda_{K-1}(\tilde{\mathbf{S}})\right| \leq \left\|\frac{1}{N}(\hat{\mathbf{S}}-\tilde{\mathbf{S}})\right\|_2 \leq \left\|\frac{1}{N}(\hat{\mathbf{S}}-\tilde{\mathbf{S}})\right\|_{\mathrm{F}} = \left\|\frac{1}{N}\mathbf{Z}^T\mathbf{R}\mathbf{Z}\right\|_{\mathrm{F}} \tag{2.2.68}$$

$$\leq \frac{1}{N}\|\mathbf{R}\|_{\mathrm{F}}\|\mathbf{Z}\|_{\mathrm{F}}^2 \leq \|\mathbf{R}\|_{\mathrm{F}}\left(\frac{1}{N}\|\mathbf{V}\|_{\mathrm{F}}^2 + \|\bar{\mathbf{v}}\|_2^2\right). \tag{2.2.69}$$

Thus in (2.2.66) and (2.2.69), we need to bound $\|\mathbf{R}\|_{\mathrm{F}}$. According to (2.2.17), $\|\bar{\mathbf{\Sigma}}_N - \bar{\mathbf{\Sigma}}\|_2$ can be bounded probabilistically. By Lemma 9 to follow, the upper bound of $\|\mathbf{R}\|_{\mathrm{F}}$ can be deduced by the upper bound of $\|\bar{\mathbf{\Sigma}}_N - \bar{\mathbf{\Sigma}}\|_2$. By leveraging additional concentration bounds for sub-Gaussian and sub-Exponential distributions given in Lemmas 5 and 6, we deduce that if $N \geq C_5 F^3 K^5 t^2/\epsilon^2$,

$$\mathbb{P}\left(\frac{1}{N}\left|\mathcal{D}(\hat{\mathbf{V}},\mathscr{C}) - \mathcal{D}(\tilde{\mathbf{V}},\mathscr{C})\right| \geq \frac{\epsilon}{2}\right) \leq 48KF\exp(-t^2F), \tag{2.2.70}$$

$$\mathbb{P}\left(\frac{1}{N}\left|\lambda_{K-1}(\hat{\mathbf{S}}) - \lambda_{K-1}(\tilde{\mathbf{S}})\right| \geq \frac{\epsilon}{2}\right) \leq 48KF\exp(-t^2F), \tag{2.2.71}$$

where $C_5 > 0$ depends on $\{(w_k, \mathbf{u}_k, \sigma_k^2)\}_{k\in[K]}$. The proofs of (2.2.70) and (2.2.71) are omitted for the sake of brevity. See inequalities (2.3.61) and (2.3.62) in Section 2.3.2 for calculations similar to those to obtain these bounds. Combining these bounds with (2.2.62) and (2.2.63) and by using Corollary 1, we obtain (2.2.58) as desired. □

The following is a lemma essential for establishing upper bounds of (2.2.66) and (2.2.69) in the proof of Theorem 2. Note that if we view the Grassmannian manifold as a metric measure space, the distance between subspaces $\mathcal{E}$ and $\mathcal{F}$ can be defined as [65]

$$\mathrm{d}_{\mathcal{S}}(\mathcal{E},\mathcal{F}) := \|\mathbf{P}_{\mathcal{E}} - \mathbf{P}_{\mathcal{F}}\|_{\mathrm{F}}, \tag{2.2.72}$$

where $\mathbf{P}_{\mathcal{E}}$ and $\mathbf{P}_{\mathcal{F}}$ are the orthogonal projections onto $\mathcal{E}$ and $\mathcal{F}$. Because $\mathbf{Q}_{K-1}\mathbf{Q}_{K-1}^T$ and $\mathbf{P}_{K-1}\mathbf{P}_{K-1}^T$ are the orthogonal projection matrices for projections onto the subspaces spanned by the columns of $\mathbf{Q}_{K-1}$ and $\mathbf{P}_{K-1}$ respectively, $\|\mathbf{R}\|_{\mathrm{F}}$ is a measure of the distance between these two subspaces.

Lemma 9. *For $\epsilon > 0$, if $\|\bar{\mathbf{\Sigma}}_N - \bar{\mathbf{\Sigma}}\|_2 \leq \epsilon$, then*

$$\|\mathbf{R}\|_{\mathrm{F}} \leq \frac{4\sqrt{K}\epsilon}{\lambda_{\min}}. \tag{2.2.73}$$

Proof. By Lemma 16, $|\lambda_F(\bar{\mathbf{\Sigma}}_N) - \bar{\sigma}^2| = |\lambda_F(\bar{\mathbf{\Sigma}}_N) - \lambda_F(\bar{\mathbf{\Sigma}})| \leq \epsilon$. Then

$$\|\bar{\mathbf{\Sigma}}_N - \lambda_F(\bar{\mathbf{\Sigma}}_N)\mathbf{I} - \bar{\mathbf{\Sigma}}_0\|_2 \leq 2\epsilon. \tag{2.2.74}$$

Because $\bar{\mathbf{\Sigma}}_N - \lambda_F(\bar{\mathbf{\Sigma}}_N)\mathbf{I}$ is also positive semidefinite, the SVD is

$$\bar{\mathbf{\Sigma}}_N - \lambda_F(\bar{\mathbf{\Sigma}}_N)\mathbf{I} = \mathbf{P}\bar{\mathbf{D}}\mathbf{P}^T, \tag{2.2.75}$$

where $\bar{\mathbf{D}} := \mathbf{D} - \lambda_F(\bar{\mathbf{\Sigma}}_N)\mathbf{I}$. Let $\bar{\mathbf{D}}_{K-1} = \bar{\mathbf{D}}(1\colon K-1, 1\colon K-1)$. Note that $\operatorname{rank}(\bar{\mathbf{\Sigma}}_0) = K-1$. By Lemma 16, $\lambda_K(\mathbf{P}\bar{\mathbf{D}}\mathbf{P}^T) \le 2\epsilon$ and thus we have

$$\|\mathbf{P}\bar{\mathbf{D}}\mathbf{P}^T - \mathbf{P}_{K-1}\bar{\mathbf{D}}_{K-1}\mathbf{P}_{K-1}^T\|_2 = \lambda_K(\mathbf{P}\bar{\mathbf{D}}\mathbf{P}^T) \le 2\epsilon. \tag{2.2.76}$$

Therefore, there exists a matrix $\mathbf{E}_0$ with $\|\mathbf{E}_0\|_2 \le 4\epsilon$, such that

$$\bar{\mathbf{\Sigma}}_0 = \mathbf{P}_{K-1}\bar{\mathbf{D}}_{K-1}\mathbf{P}_{K-1}^T + \mathbf{E}_0. \tag{2.2.77}$$

That is,

$$\mathbf{Q}_{K-1}\mathbf{E}_{K-1}\mathbf{Q}_{K-1}^T = \mathbf{P}_{K-1}\bar{\mathbf{D}}_{K-1}\mathbf{P}_{K-1}^T + \mathbf{E}_0. \tag{2.2.78}$$

Recall that $\mathbf{P}_{-(K-1)} := \mathbf{P}(:, K : F)$. We obtain

$$\mathbf{Q}_{K-1}\mathbf{Q}_{K-1}^T = \mathbf{E}_0\mathbf{Q}_{K-1}\mathbf{E}_{K-1}^{-1}\mathbf{Q}_{K-1}^T + \mathbf{P}_{K-1}\bar{\mathbf{D}}_{K-1}\mathbf{P}_{K-1}^T\mathbf{Q}_{K-1}\mathbf{E}_{K-1}^{-1}\mathbf{Q}_{K-1}^T \tag{2.2.79}$$

$$= \mathbf{P}_{K-1}\mathbf{P}_{K-1}^T\mathbf{P}_{K-1}\bar{\mathbf{D}}_{K-1}\mathbf{P}_{K-1}^T\bar{\mathbf{\Sigma}}_0^+ + \mathbf{E}_0\bar{\mathbf{\Sigma}}_0^+ \tag{2.2.80}$$

$$= \mathbf{P}_{K-1}\mathbf{P}_{K-1}^T + \mathbf{P}_{-(K-1)}\mathbf{P}_{-(K-1)}^T\mathbf{E}_0\bar{\mathbf{\Sigma}}_0^+, \tag{2.2.81}$$

where $\bar{\mathbf{\Sigma}}_0^+ := \mathbf{Q}_{K-1}\mathbf{E}_{K-1}^{-1}\mathbf{Q}_{K-1}^T$ is the Moore-Penrose generalized inverse (or pseudoinverse) of $\bar{\mathbf{\Sigma}}_0$ and its largest eigenvalue is $\lambda_{\min}^{-1}$. Because $\mathbf{P}_{-(K-1)}\mathbf{P}_{-(K-1)}^T$ projects vectors in $\mathbb{R}^F$ onto the linear space spanned by the orthonormal columns of $\mathbf{P}_{-(K-1)}$, we have

$$\|\mathbf{R}\|_{\mathrm{F}} = \|\mathbf{Q}_{K-1}\mathbf{Q}_{K-1}^T - \mathbf{P}_{K-1}\mathbf{P}_{K-1}^T\|_{\mathrm{F}} \tag{2.2.82}$$

$$\le \|\mathbf{E}_0\bar{\mathbf{\Sigma}}_0^+\|_{\mathrm{F}} \le \|\mathbf{E}_0\|_2\sqrt{K}\|\bar{\mathbf{\Sigma}}_0^+\|_2 \le \frac{4\sqrt{K}\epsilon}{\lambda_{\min}}, \tag{2.2.83}$$

where we use the inequality that $\|\mathbf{M}\mathbf{N}\|_{\mathrm{F}} \le \|\mathbf{M}\|_2\|\mathbf{N}\|_{\mathrm{F}}$ for any two compatible matrices $\mathbf{M}$ and $\mathbf{N}$. The inequality $\|\bar{\mathbf{\Sigma}}_0^+\|_{\mathrm{F}} \le \sqrt{K}\|\bar{\mathbf{\Sigma}}_0^+\|_2$ arises because $\bar{\mathbf{\Sigma}}_0^+$ only contains $K-1$ positive eigenvalues. □

Remark 7. By Lemmas 8 and 9, we see that the subspace spanned by the first $K-1$ singular vectors of $\bar{\mathbf{\Sigma}}_0$ lies close to the subspace spanned by the first $K-1$

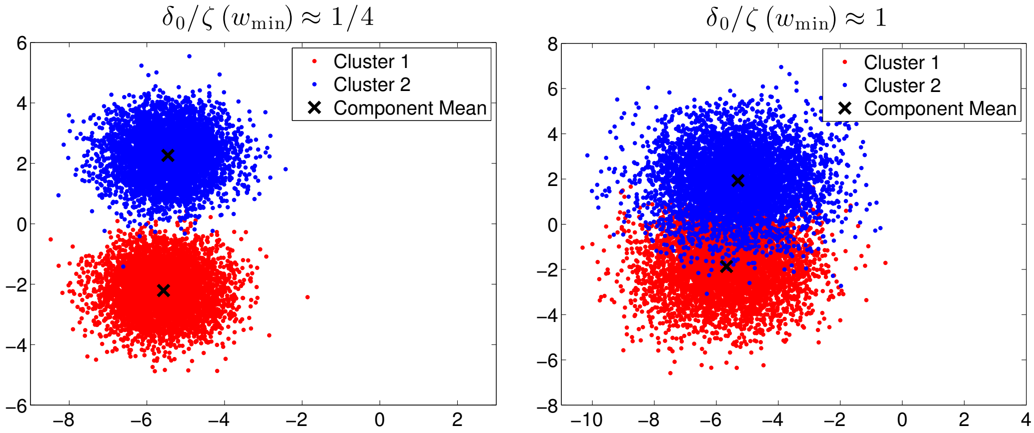

Figure 2.1: Visualization of post-2-SVD datasets.

singular vectors of $\bar{\mathbf{\Sigma}}_N$ when the number of samples is sufficiently large. Note that $\sum_{k=1}^K w_k(\mathbf{u}_k - \bar{\mathbf{u}})(\mathbf{u}_k - \bar{\mathbf{u}})^T = \bar{\mathbf{\Sigma}}_0 = \mathbf{Q}_{K-1}\mathbf{E}_{K-1}\mathbf{Q}_{K-1}^T$. We also obtain that the subspace spanned by $\mathbf{u}_k - \bar{\mathbf{u}}, k \in [K]$ is close to the subspace spanned by the first $K-1$ singular vectors of $\bar{\mathbf{\Sigma}}_N$. Note that $\mathbf{\Sigma}_0 = \sum_{k=1}^K w_k \mathbf{u}_k \mathbf{u}_k^T$. A similar result can be obtained for K-SVD to corroborate an observation by [17] (cf. Remark 6).

2.2.3 Numerical Results

To verify the correctness of the upper bounds given in Theorems 1 and 2 and the efficacy of clustering post-PCA samples, we perform numerical experiments on synthetic datasets. We sample data points from two types of 2-component spherical GMMs. The dimensionality of the data points is $F = 100$, and the number of samples N ranges from 1000 to 10000. Component mean vectors are randomly and uniformly picked from the hypercube $[0,1]^F$. Equal mixing weights are assigned to the components. After fixing the mixing weights and the component mean vectors, $\lambda_{\min}$ is fixed. For all $k \in [K]$, we set the variances to be

$$\sigma_k^2 = \frac{\lambda_{\min}\zeta(w_{\min} - \varepsilon)}{4(K-1)}, \quad \text{corresponding to} \quad \frac{\delta_0}{\zeta(w_{\min})} \approx \frac{1}{4}, \quad \text{or} \tag{2.2.84}$$

$$\sigma_k^2 = \frac{\lambda_{\min}\zeta(w_{\min} - \varepsilon)}{K-1}, \quad \text{corresponding to} \quad \frac{\delta_0}{\zeta(w_{\min})} \approx 1, \tag{2.2.85}$$

where $\varepsilon = 10^{-6}$. In all figures, left and right plots correspond to (2.2.84) and (2.2.85) respectively.

We observe from Figure 2.1 that for (2.2.84), the clusters are well-separated, while for (2.2.85), the clusters are moderately well-separated. For both cases, the separability assumption (2.2.5) is satisfied. Similar to that in [60], we use the command `kmeans(V', K, 'Replicates', 10, 'MaxIter', 1000)` in Matlab to obtain an approximately-optimal clustering of $\mathbf{V}$. Here `V'` represents the transpose of $\mathbf{V}$. This command means that we run k-means for 10 times with distinct initializations and pick the best outcome. For each run, the maximal number of iterations is set to be 1000. Define $\mathrm{d}_{\mathrm{org}} := \mathrm{d}_{\mathrm{ME}}(\mathscr{C}, \mathscr{C}^{\mathrm{opt}})$ and define the (expected) upper bound for $\mathrm{d}_{\mathrm{ME}}(\mathscr{C}, \mathscr{C}^{\mathrm{opt}})$ as $\bar{\mathrm{d}}_{\mathrm{org}} := \tau(\delta_0) w_{\max}$ (provided by Theorem 1). Similarly, we define $\mathrm{d}_{\mathrm{pca}} := \mathrm{d}_{\mathrm{ME}}(\mathscr{C}, \tilde{\mathscr{C}}^{\mathrm{opt}})$ and the (expected) upper bound for $\mathrm{d}_{\mathrm{ME}}(\mathscr{C}, \tilde{\mathscr{C}}^{\mathrm{opt}})$ is defined as $\bar{\mathrm{d}}_{\mathrm{pca}} := \tau(\delta_1) w_{\max}$ (given by Theorem 2). We use a superscript "emp" to represent the corresponding empirical value. For example, $\delta_0^{\mathrm{emp}} := \frac{\mathcal{D}(\mathbf{V},\mathscr{C}) - \mathcal{D}^*(\mathbf{V})}{\lambda_{K-1}(\mathbf{S}) - \lambda_K(\mathbf{S})}$ is an approximation of δ_0 (calculated from the samples), and $\bar{\mathrm{d}}_{\mathrm{org}}^{\mathrm{emp}} := \tau(\delta_0^{\mathrm{emp}}) p_{\max}$ is an approximation of $\bar{\mathrm{d}}_{\mathrm{org}}$, where $p_{\max} := \max_k \frac{1}{N} |\mathscr{C}_k|$.

Our numerical results are reported in Figure 2.2. We observe that the empirical values of upper bounds are close to the corresponding expected upper bounds. This observation verifies the correctness of the probabilistic estimates. For the well-separated case in (2.2.84), we observe that the upper bounds for ME distance are small compared to the moderately well-separated case in (2.2.85). For the former, the true distances $\mathrm{d}_{\mathrm{org}}$ and $\mathrm{d}_{\mathrm{pca}}$ are both close to 0, even when the number of samples is 1000 (a small number in this scenario). The k-means algorithm can easily find an approximately-optimal clustering, which is also the approximately-correct clustering. For the moderately well-separated case, we observe that the upper bounds given in Theorem 1 and Theorem 2 are informative. In particular, they are only approximately 2.5 times the corresponding true distances.

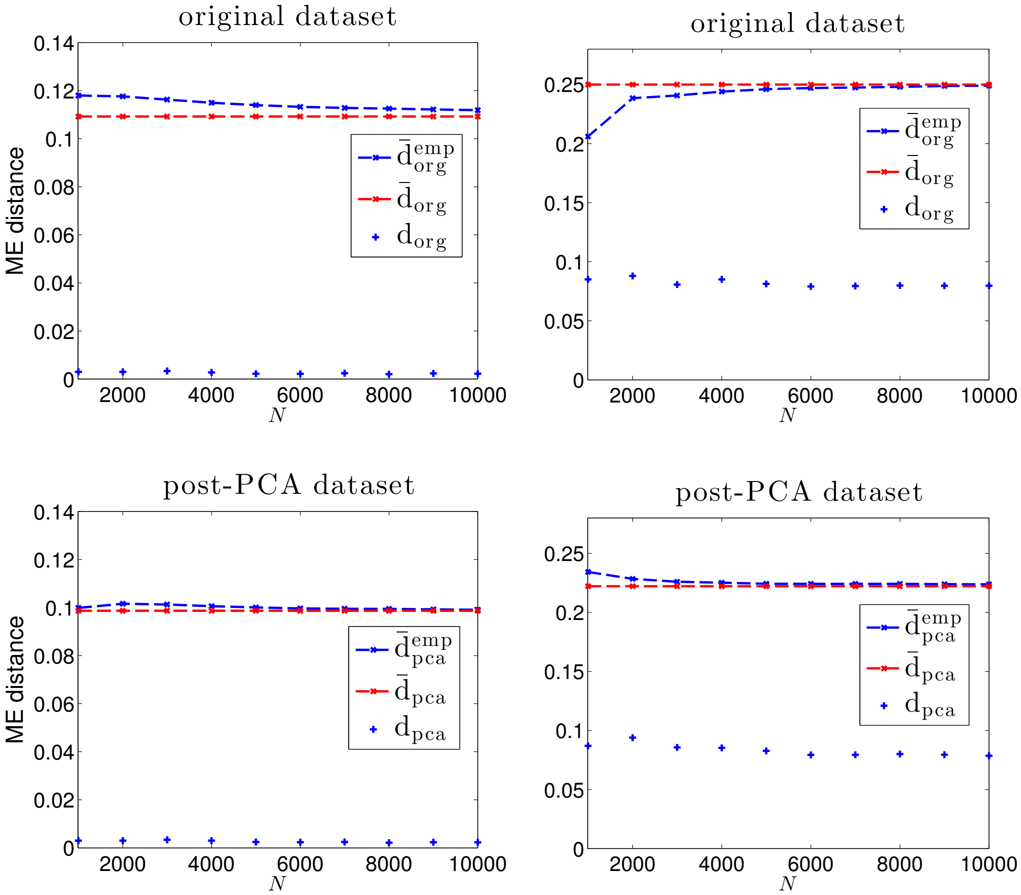

Figure 2.2: True distances and their corresponding upper bounds.

From Figure 2.3, we observe that performing k-means for the original (high-dimensional) datasets is significantly slower than performing k-means for the corresponding post-PCA datasets (the reported running times for post-PCA datasets are the sums of the running times for performing PCA and for performing k-means on the post-PCA datasets) when the number of samples is large. This difference is more pronounced for the moderately well-separated case. For this case and $N = 10000$, we have an *order of magnitude speed up*. The running time for larger N can be less than the running time for smaller N because the number of iterations for k-means are possibly different. All the results are averaged over 10 runs. All experiments we run on an Intel Core i7 CPU at 2.50GHz and 16GB of memory, and the Matlab

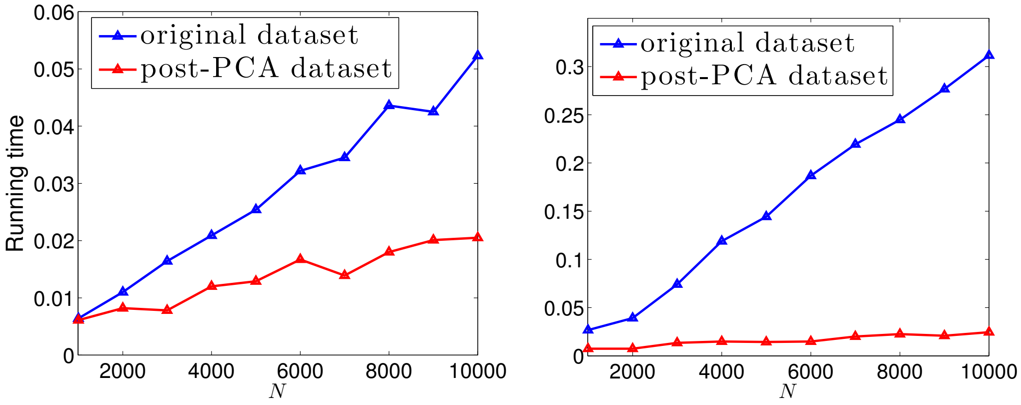

Figure 2.3: Comparisons of running times in seconds. Notice the tremendous speed up of clustering on the post-PCA dataset compared to the original one.

version is 8.3.0.532 (R2014a).

2.3 Extension to Mixtures of Log-Concave Distributions

In this section, we extend the results in Section 2.2 to mixtures of *log-concave distributions*. This is motivated partly by Arora and Kannan [52] who mentioned the algorithm they design for analyzing the learning of GMMs may be extended to log-concave distributions (besides Gaussians). Also, Kannan et al. [55] generalize the work of Vempala and Wang [17] from spherical GMMs to mixtures of log-concave distributions. Furthermore, Brubaker [66] considers the robust learning of mixtures of log-concave distributions. We also consider the learning of mixtures of log-concave distributions although the structure of the theoretical analysis is mostly similar to that for spherical GMMs. In Section 2.3.1, before presenting our theorem and proof for the original dataset generated from a mixture of log-concave distributions, we provide some necessary preliminary definitions and results for log-concave distributions and random variables. To prove our main results concerning such distributions for dimensionality-reduced datasets, we need to employ a slightly different proof strategy vis-à-vis the one for spherical GMMs. We discuss these differences and present the theorem and proof for dimensionality-reduced datasets in Section 2.3.2 .

2.3.1 The Theorem for Original Data

A function $f : \mathbb{R}^F \to \mathbb{R}_+$ is *log-concave* if its logarithm is concave. That is, for any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^F$ and any $\alpha \in [0, 1]$,

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \geq f(\mathbf{x})^\alpha f(\mathbf{y})^{1-\alpha}. \tag{2.3.1}$$

A distribution is *log-concave* if its probability density (or mass) function is log-concave. We say a random variable is *log-concave* if its distribution is log-concave. There are many distributions that are log-concave, including Gaussian distributions, exponential distributions, and Laplace distributions. In particular, distributions that belong to exponential families are log-concave. Log-concave distributions have several desirable properties. For example, the sum of two independent log-concave random variables (i.e., the convolution of two log-concave distributions) is also log-concave. In addition, the linear projection of a log-concave distribution onto a lower-dimensional space remains log-concave. To start off, we need to estimate the deviation of an empirical covariance matrix from the true covariance matrix. We leverage the following lemma due to [55].

Lemma 10. *Let $\epsilon, \eta \in (0, 1)$, and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N$ be zero-mean i.i.d. random vectors from a log-concave distribution in $\mathbb{R}^F$. Then there is an absolute constant $C > 0$ such that if $N > C\frac{F}{\epsilon^2}\log^5\left(\frac{F}{\epsilon\eta}\right)$, with probability at least $1 - \eta$,*

$$(1-\epsilon)\mathbb{E}\big[(\mathbf{v}^T\mathbf{y})^2\big] \leq \frac{1}{N}\sum_{n=1}^{N}(\mathbf{v}^T\mathbf{y}_n)^2 \leq (1+\epsilon)\mathbb{E}\big[(\mathbf{v}^T\mathbf{y})^2\big], \quad \forall\, \mathbf{v} \in \mathbb{R}^F. \tag{2.3.2}$$

Note that Lemma 10 provides an estimate for the empirical covariance matrix. This is because that for any symmetric matrix $\mathbf{M}$, $\|\mathbf{M}\|_2 = \sup_{\mathbf{v}\in\mathbb{R}^F, \|\mathbf{v}\|_2=1} |\langle \mathbf{M}\mathbf{v}, \mathbf{v}\rangle|$ and (2.3.2) is equivalent to

$$\|\mathbf{\Sigma}_N - \mathbf{\Sigma}\|_2 \leq \epsilon\|\mathbf{\Sigma}\|_2. \tag{2.3.3}$$

Using Lemma 10, we also have the following corollary which provides a useful concentration bound for the sum of the squares of log-concave random variables.

Corollary 4. *Let $\epsilon \in (0,1)$, and $y_1, y_2, \ldots, y_N$ be i.i.d. samples from a log-concave distribution with expectation μ and variance σ^2. Then if ϵ is sufficiently small, there is an absolute constant $C > 0$ such that for $N > C\frac{1}{\epsilon^2}\log^5\left(\frac{1}{\epsilon\eta}\right)$, with probability at least $1-\eta$,*

$$\left|\frac{1}{N}\sum_{n=1}^{N} y_n^2 - (\mu^2+\sigma^2)\right| \le \epsilon. \tag{2.3.4}$$

Proof. When $\mu = 0$, we can apply Lemma 10 with $F = 1$ directly. When $\mu \neq 0$, we have

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N} y_n^2 - (\mu^2+\sigma^2)\right| > \epsilon\right)\\ &\le \mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N}(y_n-\mu)^2 - \sigma^2\right| > \frac{\epsilon}{2}\right) + \mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N} y_n - \mu\right| > \frac{\epsilon}{4|\mu|}\right).\end{aligned} \tag{2.3.5}$$

By Lemma 10, there is a $C > 0$ such that if $N \ge C\frac{1}{\epsilon^2}\log^5\left(\frac{1}{\epsilon\eta}\right)$, the first term in (2.3.5) is less than or equal $\frac{\eta}{2}$. By Lemmas 5 and 6, the second term is less than or equal $2\exp(-c\epsilon^2 N)$, where $c > 0$. When $N \ge \frac{1}{c\epsilon^2}\log(\frac{4}{\eta})$, we have $\frac{\eta}{2} \ge 2\exp(-c\epsilon^2 N)$. If ϵ is sufficiently small, we have $\frac{1}{c\epsilon^2}\log(\frac{4}{\eta}) \le C\frac{1}{\epsilon^2}\log^5\left(\frac{1}{\epsilon\eta}\right)$. Thus we obtain (2.3.4) as desired. □

We will make use of the following lemma [62] concerning the eigenvalues of a matrix which is the sum of two matrices.

Lemma 11. *If $\mathbf{A}$ and $\mathbf{A}+\mathbf{E}$ are both n-by-n symmetric matrices, then*

$$\lambda_k(\mathbf{A}) + \lambda_n(\mathbf{E}) \le \lambda_k(\mathbf{A}+\mathbf{E}) \le \lambda_k(\mathbf{A}) + \lambda_1(\mathbf{E}), \quad \forall\, k \in [n]. \tag{2.3.6}$$

For any $k \in [K]$, let $\sigma_{k,\max}^2$ and $\sigma_{k,\min}^2$ be the maximal and minimal eigenvalues of $\mathbf{\Sigma}_k$ respectively, and define $\bar{\sigma}_{\max}^2 := \sum_{k=1}^K w_k\sigma_{k,\max}^2$ and $\bar{\sigma}_{\min}^2 := \sum_{k=1}^K w_k\sigma_{k,\min}^2$. Other notations are the same as that in previous theorems. Then in a similar manner as Theorem 1 for spherical GMMs, we have the following theorem for mixtures of log-concave distributions.

Theorem 3. *Suppose that all the columns of data matrix* $\mathbf{V} \in \mathbb{R}^{F\times N}$ *are independently generated from a mixture of* K *log-concave distributions and* $N > F > K$*. Assume the mixture model satisfies the non-degeneracy condition. We further assume that*

$$0 < \delta_2 := \frac{F\bar{\sigma}_{\max}^2 - (F-K+1)\bar{\sigma}_{\min}^2}{\lambda_{\min} + \bar{\sigma}_{\min}^2 - \bar{\sigma}_{\max}^2} < \zeta(w_{\min}). \tag{2.3.7}$$

Then for any sufficiently small $\epsilon \in (0,1)$ *and any* $t \geq 1$*, if* $N \geq C\frac{K^2F^4}{\epsilon^2}\log^5\left(\frac{K^2F^3}{\epsilon\eta}\right)$*, where* $C > 0$ *depends on the parameters of the mixture model, we have, with probability at least* $1-\eta$*,*

$$\mathrm{d}_{\mathrm{ME}}(\mathscr{C}, \mathscr{C}^{\mathrm{opt}}) \leq \tau\left(\frac{F\bar{\sigma}_{\max}^2 - (F-K+1)\bar{\sigma}_{\min}^2 + \epsilon}{\lambda_{\min} + \bar{\sigma}_{\min}^2 - \bar{\sigma}_{\max}^2 - \epsilon}\right)(w_{\max} + \epsilon), \tag{2.3.8}$$

where $\mathscr{C}^{\mathrm{opt}}$ *is an optimal* K*-clustering for* $\mathbf{V}$*.*

The proof, which is presented below, is mostly similar to that for Theorem 1, except that we employ different concentration inequalities and slightly different bounding strategies (e.g., Lemma 11 is required).

Proof of Theorem 3. We have that the two inequalities concerning w_k in (2.2.23) and (2.2.24) still hold. In addition, for any $k \in [K]$, if $N \geq C_1\frac{F^2}{\epsilon^2}\log^5\left(\frac{F}{\epsilon\eta}\right)$ with $C_1 > 0$ being sufficiently large,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n_k}\sum_{n\in\mathcal{I}_k}\|\mathbf{v}_n\|_2^2 - \left(\|\mathbf{u}_k\|_2^2 + \mathrm{tr}(\mathbf{\Sigma}_k)\right)\right| \geq \epsilon\right) \\
&\leq \sum_{f=1}^{F}\mathbb{P}\left(\left|\frac{1}{n_k}\sum_{n\in\mathcal{I}_k}v_n(f)^2 - \left(u_k(f)^2 + \mathbf{\Sigma}_k(f,f)\right)\right| \geq \frac{\epsilon}{F}\right) \leq F\eta.
\end{aligned} \tag{2.3.9}$$

By Lemma 6, similarly, we have for a sufficiently small $c > 0$,

$$\mathbb{P}\left(\left|\|\mathbf{c}_k\|_2^2 - \|\mathbf{u}_k\|_2^2\right| \geq \epsilon\right) \leq 4F\exp\left(-c\frac{N\epsilon^2}{F^2}\right). \tag{2.3.10}$$

Therefore, by $\sum_{k=1}^{K} w_k\mathrm{tr}(\mathbf{\Sigma}_k) \leq F\bar{\sigma}_{\max}^2$, similar to that for spherical GMMs, if $N \geq C_1\frac{F^2K^2}{\epsilon^2}\log^5\left(\frac{F^2K^2}{\epsilon\eta}\right)$,

$$\mathbb{P}\left(\frac{1}{N}\mathcal{D}(\mathbf{V}, \mathscr{I}) - F\bar{\sigma}_{\max}^2 \geq \frac{\epsilon}{2}\right) \leq \eta. \tag{2.3.11}$$

Similarly, we have

$$\mathbb{P}\left(|\bar{\mathbf{v}}(f) - \bar{\mathbf{u}}(f)| \geq \epsilon\right) \leq (4+e)K \exp\left(-c\frac{N\epsilon^2}{K^2}\right). \tag{2.3.12}$$

We also have that

$$\begin{align}
&\mathbb{P}\left(\|\bar{\mathbf{\Sigma}}_N - \bar{\mathbf{\Sigma}}\|_2 \geq \frac{\epsilon}{2}\right) \nonumber\\
&\leq \mathbb{P}\left(\|\mathbf{\Sigma}_N - \mathbf{\Sigma}\|_2 \geq \frac{\epsilon}{4}\right) + \mathbb{P}\left(\|\bar{\mathbf{v}}\bar{\mathbf{v}}^T - \bar{\mathbf{u}}\bar{\mathbf{u}}^T\|_2 \geq \frac{\epsilon}{4}\right) \tag{2.3.13}\\
&\leq \sum_{k=1}^{K} \mathbb{P}\left(\left\|\frac{n_k}{N}\frac{\sum_{n\in\mathscr{C}_k}\mathbf{v}_n\mathbf{v}_n^T}{n_k} - w_k(\mathbf{u}_k\mathbf{u}_k^T + \mathbf{\Sigma}_k)\right\|_2 \geq \frac{\epsilon}{4K}\right) \nonumber\\
&\quad + \mathbb{P}\left(\|\bar{\mathbf{v}}\bar{\mathbf{v}}^T - \bar{\mathbf{u}}\bar{\mathbf{u}}^T\|_2 \geq \frac{\epsilon}{4}\right) \tag{2.3.14}\\
&\leq \sum_{k=1}^{K} \mathbb{P}\left(\left\|\left(\frac{n_k}{N} - w_k\right)(\mathbf{u}_k\mathbf{u}_k^T + \mathbf{\Sigma}_k)\right\|_2 \geq \frac{\epsilon}{8K}\right) + \sum_{k=1}^{K}\mathbb{P}\left(\frac{n_k}{N} \geq 2w_k\right) \nonumber\\
&\quad + \sum_{k=1}^{K}\mathbb{P}\left(\left\|\frac{\sum_{n\in\mathscr{C}_k}\mathbf{v}_n\mathbf{v}_n^T}{n_k} - (\mathbf{u}_k\mathbf{u}_k^T + \mathbf{\Sigma}_k)\right\|_2 \geq \frac{\epsilon}{16Kw_k}\right) + \mathbb{P}\left(\|\bar{\mathbf{v}}\bar{\mathbf{v}}^T - \bar{\mathbf{u}}\bar{\mathbf{u}}^T\|_2 \geq \frac{\epsilon}{4}\right). \tag{2.3.15}
\end{align}$$

Note that the following bound for $\bar{\mathbf{\Sigma}}_N - \bar{\mathbf{\Sigma}}$ is slightly different from (2.2.43) because Lemma 10 requires that the log-concave distribution has *zero mean*. Furthermore, recall that $\mathbf{c}_k := \frac{1}{|\mathscr{C}_k|}\sum_{n\in\mathscr{C}_k}\mathbf{v}_n$, we have

$$\begin{align}
&\mathbb{P}\left(\left\|\frac{1}{n_k}\sum_{n\in\mathscr{C}_k}\mathbf{v}_n\mathbf{v}_n^T - (\mathbf{u}_k\mathbf{u}_k^T + \mathbf{\Sigma}_k)\right\|_2 \geq \frac{\epsilon}{16Kw_k}\right) \nonumber\\
&\leq \mathbb{P}\left(\left\|\frac{1}{n_k}\sum_{n\in\mathscr{C}_k}(\mathbf{v}_n - \mathbf{u}_k)(\mathbf{v}_n - \mathbf{u}_k)^T - \mathbf{\Sigma}_k\right\|_2 \geq \frac{\epsilon}{32Kw_k}\right) \nonumber\\
&\qquad + \mathbb{P}\left(\|\mathbf{c}_k\mathbf{u}_k^T + \mathbf{u}_k\mathbf{c}_k^T - 2\mathbf{u}_k\mathbf{u}_k^T\|_2 \geq \frac{\epsilon}{32Kw_k}\right) \tag{2.3.16}\\
&\leq \mathbb{P}\left(\left\|\frac{1}{n_k}\sum_{n\in\mathscr{C}_k}(\mathbf{v}_n - \mathbf{u}_k)(\mathbf{v}_n - \mathbf{u}_k)^T - \mathbf{\Sigma}_k\right\|_2 \geq \frac{\epsilon}{32Kw_k}\right) \nonumber\\
&\qquad + 2\mathbb{P}\left(\|\mathbf{c}_k - \mathbf{u}_k\|_2 \geq \frac{\epsilon}{64Kw_k\|\mathbf{u}_k\|_2}\right). \tag{2.3.17}
\end{align}$$

Therefore, we have that if $N \geq C_1\frac{F^2K^2}{\epsilon^2}\log^5\left(\frac{FK^2}{\epsilon\eta}\right)$, with probability at least $1-\eta$,

$$\|\bar{\mathbf{\Sigma}}_N - \bar{\mathbf{\Sigma}}\|_2 < \frac{\epsilon}{2}. \tag{2.3.18}$$

Now, note that by Lemma 11, for $k \leq F < N$,

$$\lambda_k\left(\bar{\mathbf{\Sigma}}\right) = \lambda_k\left(\bar{\mathbf{\Sigma}}_0 + \sum_{k=1}^{K} w_k \mathbf{\Sigma}_k\right) \geq \lambda_k(\bar{\mathbf{\Sigma}}_0) + \lambda_F\left(\sum_{k=1}^{K} w_k \mathbf{\Sigma}_k\right) \tag{2.3.19}$$

$$\geq \lambda_k(\bar{\mathbf{\Sigma}}_0) + \sum_{k=1}^{K} w_k \lambda_F(\mathbf{\Sigma}_k) = \lambda_k(\bar{\mathbf{\Sigma}}_0) + \bar{\sigma}^2_{\min}. \tag{2.3.20}$$

Therefore, if $N \geq C_1 \frac{F^4K^2}{\epsilon^2} \log^5\left(\frac{F^2K^2}{\epsilon\eta}\right)$, we have

$$\mathbb{P}\left(\frac{1}{N}\mathcal{D}^*(\mathbf{V}) - (F-K+1)\bar{\sigma}^2_{\min} \leq -\frac{\epsilon}{2}\right)$$

$$= \mathbb{P}\left(\frac{1}{N}\sum_{k=K}^{F} \lambda_k(\mathbf{S}) - (F-K+1)\bar{\sigma}^2_{\min} \leq -\frac{\epsilon}{2}\right) \tag{2.3.21}$$

$$\leq \sum_{k=K}^{F} \mathbb{P}\left(|\lambda_k(\bar{\mathbf{\Sigma}}_N) - \lambda_k(\bar{\mathbf{\Sigma}})| \geq \frac{\epsilon}{2(F-K+1)}\right) \tag{2.3.22}$$

$$\leq 2(F-K+1)\eta. \tag{2.3.23}$$

Or more concisely, if $N \geq C_1 \frac{F^4K^2}{\epsilon^2} \log^5\left(\frac{F^3K^2}{\epsilon\eta}\right)$,

$$\mathbb{P}\left(\frac{1}{N}\mathcal{D}^*(\mathbf{V}) - (F-K+1)\bar{\sigma}^2_{\min} \leq -\frac{\epsilon}{2}\right) \leq \eta. \tag{2.3.24}$$

Similarly, by the inequalities $\lambda_{K-1}(\bar{\mathbf{\Sigma}}) \geq \lambda_{K-1}(\bar{\mathbf{\Sigma}}_0) + \bar{\sigma}^2_{\min}$ and $\lambda_K(\bar{\mathbf{\Sigma}}) \leq \bar{\sigma}^2_{\max}$, if $N \geq C_1 \frac{F^2K^2}{\epsilon^2} \log^5\left(\frac{FK^2}{\epsilon\eta}\right)$,

$$\mathbb{P}\left(\frac{1}{N}\lambda_{K-1}(\mathbf{S}) - \left(\lambda_{\min} + \bar{\sigma}^2_{\min}\right) \leq -\frac{\epsilon}{2}\right) \leq \eta, \tag{2.3.25}$$

$$\mathbb{P}\left(\frac{1}{N}\lambda_K(\mathbf{S}) - \bar{\sigma}^2_{\max} \geq \frac{\epsilon}{2}\right) \leq \eta. \tag{2.3.26}$$

Combining these results with Corollary 1, we similarly obtain the conclusion we desire. □

2.3.2 The Theorem for Post-PCA Data

For non-spherical GMMs or more general mixtures of log-concave distributions, we cannot expect a result similar to that mentioned in Remarks 6 and 7. That is, in general, the subspace spanned by the top K singular vectors of $\mathbf{\Sigma}_N$ does not converge to the subspace spanned by the K component mean vectors, where "convergence" is in the sense that the distance defined in (2.2.72) vanishes as $N \to \infty$. An illustration of this fact is presented in [55, Figure 1]. However, we can still provide an upper bound for the distance of this two subspaces for more general mixture models. It is proved in [55] that the subspace spanned by the top K singular vectors of $\mathbf{\Sigma}_N$ is close (in terms of sample variances) to the means of samples generated from each component of the mixture (see Lemma 12 to follow). Define the maximum variance of a set of sample points generated from the k-th component of the mixture along any direction in a subspace $\mathcal{S}$ as

$$\sigma_{k,\mathcal{S}}^2(\mathbf{V}) := \max_{\mathbf{z}\in\mathcal{S},\|\mathbf{z}\|_2=1} \frac{1}{n_k} \sum_{n\in\mathscr{C}_k} \left|\mathbf{z}^T(\mathbf{v}_n - \mathbf{c}_k)\right|^2, \tag{2.3.27}$$

where $\mathbf{c}_k := \frac{1}{|\mathscr{C}_k|}\sum_{n\in\mathscr{C}_k} \mathbf{v}_n$ is the centroid of the points in $\mathscr{C}_k$. Recall that from Section 2.1.2, we denote $\bar{\mathbf{v}} := \frac{1}{N}\sum_{n=1}^N \mathbf{v}_n$ and $\mathbf{Z}$ as the centralized matrix of $\mathbf{V}$. Let $\bar{\mathbf{c}}_k = \mathbf{c}_k - \bar{\mathbf{v}}$. We have $\sigma_{k,\mathcal{S}}^2(\mathbf{Z}) = \sigma_{k,\mathcal{S}}^2(\mathbf{V})$ for any subspace $\mathcal{S}$ and the the following lemma which is similar to Theorem 1 in [55] holds. Note that this lemma holds not only for mixtures of log-concave distributions, but also for *any* mixture model.

Lemma 12. *Let $\mathscr{C}$ be the correct target clustering corresponding to the mixture. Let $\mathcal{W}$ be the subspace spanned by the top $K-1$ left singular vectors of $\mathbf{Z}$. Then*

$$\sum_{k=1}^K n_k \mathrm{d}(\bar{\mathbf{c}}_k, \mathcal{W})^2 \le (K-1)\sum_{k=1}^K n_k \sigma_{k,\mathcal{W}}^2(\mathbf{V}), \tag{2.3.28}$$

where $\mathrm{d}(\bar{\mathbf{c}}_k, \mathcal{W})$ denotes the orthogonal distance of $\bar{\mathbf{c}}_k$ from subspace $\mathcal{W}$ for any $k \in [K]$.

In addition, recall the notations $\mathbf{P}_{K-1}, \mathbf{P}_{-(K-1)}$, and $\mathbf{Q}_{K-1}$ (cf. (2.2.59)) from Section 2.2.2. Let $\bar{\mathbf{u}}_k := \mathbf{u}_k - \bar{\mathbf{u}}$ for $k \in [K]$. Since $\sum_{k=1}^K w_k \bar{\mathbf{u}}_k = 0$, $\mathrm{rank}(\bar{\mathbf{\Sigma}}_0) \le K-1$. It is easy to see that $\mathrm{d}(\mathbf{x}, \mathcal{W})^2 = \|\mathbf{x} - \mathbf{P}_{K-1}\mathbf{P}_{K-1}^T\mathbf{x}\|_2^2$ for any vector $\mathbf{x}$ and by denoting $\sigma_{k,\max}^2$ as the maximal eigenvalue of $\mathbf{\Sigma}_k$ (the covariance matrix of the k-th component), we obtain the following corollary of Lemma 12.

Corollary 5. *If we further assume that the mixture is a mixture of log-concave distributions, then for any sufficiently small $\epsilon \in (0,1)$ and all $\eta > 0$, if $N \ge$*

$C\frac{F^2K^4}{\epsilon^2}\log^5\left(\frac{FK^3}{\epsilon\eta}\right)$, *with probability at least* $1-\eta$,

$$\sum_{k=1}^{K} w_k \mathrm{d}(\bar{\mathbf{u}}_k, \mathcal{W})^2 \le \left((K-1)\sum_{k=1}^{K} w_k \sigma_{k,\max}^2\right) + \epsilon. \tag{2.3.29}$$

Proof. Consider,

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{k=1}^{K} w_k \mathrm{d}(\bar{\mathbf{c}}_k, \mathcal{W})^2 - \sum_{k=1}^{K} w_k \mathrm{d}(\bar{\mathbf{u}}_k, \mathcal{W})^2\right| \ge \frac{\epsilon}{2}\right) \\
&\le \sum_{k=1}^{K} \mathbb{P}\left(\left|\|\mathbf{P}_{-(K-1)}\mathbf{P}_{-(K-1)}^T\bar{\mathbf{c}}_k\|_2^2 - \|\mathbf{P}_{-(K-1)}\mathbf{P}_{-(K-1)}^T\bar{\mathbf{u}}_k\|_2^2\right| \ge \frac{\epsilon}{2Kw_k}\right) && (2.3.30)\\
&\le \sum_{k=1}^{K} \mathbb{P}\left(\left|\|\bar{\mathbf{c}}_k\|_2^2 - \|\bar{\mathbf{u}}_k\|_2^2\right| \ge \frac{\epsilon}{2Kw_k}\right) && (2.3.31)\\
&\le 8eK^2F\exp\left(-C_1\frac{N\epsilon^2}{F^2K^4}\right), && (2.3.32)
\end{aligned}$$

where (2.3.31) is because that $\mathbf{P}_{-(K-1)}\mathbf{P}_{-(K-1)}^T$ is a projection matrix and $C_1 > 0$ is a constant depending on parameters of the mixture model. On the other hand,

$$\sigma_{k,\mathcal{W}}^2(\mathbf{V}) \le \lambda_1\left(\frac{1}{n_k}\sum_{n\in\mathscr{C}_k}(\mathbf{v}_n - \mathbf{c}_k)(\mathbf{v}_n - \mathbf{c}_k)^T\right), \tag{2.3.33}$$

where $\lambda_1(\cdot)$ denotes the largest eigenvalue of a matrix. Consequently, for some fixed $\eta' \in (0,1)$, if $N \ge C_2\frac{F^2K^4}{\epsilon^2}\log^5\left(\frac{FK^2}{\epsilon\eta'}\right)$ (where $C_2 > 0$ is a constant depending on parameters of the mixture model), we have

$$\begin{aligned}
&\mathbb{P}\left((K-1)\sum_{k=1}^{K} w_k\sigma_{k,\mathcal{W}}^2(\mathbf{V}) - (K-1)\sum_{k=1}^{K} w_k\sigma_{k,\max}^2 \geq \frac{\epsilon}{2}\right)\\
&\leq \sum_{k=1}^{K}\mathbb{P}\left(\sigma_{k,\mathcal{W}}^2(\mathbf{V}) - \sigma_{k,\max}^2 \geq \frac{\epsilon}{2K^2 w_k}\right) &(2.3.34)\\
&\leq \sum_{k=1}^{K}\mathbb{P}\left(\lambda_1\left(\frac{1}{n_k}\sum_{n\in\mathscr{C}_k}(\mathbf{v}_n-\mathbf{c}_k)(\mathbf{v}_n-\mathbf{c}_k)^T\right) - \sigma_{k,\max}^2 \geq \frac{\epsilon}{2K^2 w_k}\right) &(2.3.35)\\
&\leq \sum_{k=1}^{K}\mathbb{P}\left(\left\|\frac{1}{n_k}\sum_{n\in\mathscr{C}_k}(\mathbf{v}_n-\mathbf{c}_k)(\mathbf{v}_n-\mathbf{c}_k)^T - \mathbf{\Sigma}_k\right\|_2 \geq \frac{\epsilon}{2K^2 w_k}\right) &(2.3.36)\\
&\leq \sum_{k=1}^{K}\mathbb{P}\left(\left\|\frac{1}{n_k}\sum_{n\in\mathscr{C}_k}(\mathbf{v}_n-\mathbf{u}_k)(\mathbf{v}_n-\mathbf{u}_k)^T - \mathbf{\Sigma}_k\right\|_2 \geq \frac{\epsilon}{4K^2 w_k}\right)\\
&\qquad + \sum_{k=1}^{K}\mathbb{P}\left(\|\mathbf{c}_k-\mathbf{u}_k\|_2^2 \geq \frac{\epsilon}{4K^2 w_k}\right) &(2.3.37)\\
&\leq 2K\eta', &(2.3.38)
\end{aligned}$$

where (2.3.36) is due to Lemma 16 and (2.3.37) is due to the fact that

$$\frac{1}{n_k}\sum_{n\in\mathscr{C}_k}(\mathbf{v}_n-\mathbf{c}_k)(\mathbf{v}_n-\mathbf{c}_k)^T = \left[\frac{1}{n_k}\sum_{n\in\mathscr{C}_k}(\mathbf{v}_n-\mathbf{u}_k)(\mathbf{v}_n-\mathbf{u}_k)^T\right] - (\mathbf{u}_k-\mathbf{c}_k)(\mathbf{u}_k-\mathbf{c}_k)^T. \tag{2.3.39}$$

Then with probability at least $1-(8e+2)K\eta'$,

$$\sum_{k=1}^{K} w_k \mathrm{d}(\mathbf{u}_k-\bar{\mathbf{u}},\mathcal{W})^2 \leq \left((K-1)\sum_{k=1}^{K} w_k\sigma_{k,\max}^2\right) + \epsilon. \tag{2.3.40}$$

Now, we define $\eta := (8e+2)K\eta'$. Then, we obtain (2.3.29) with the sample complexity result as in the statement of Corollary 5. □

Because Corollary 5 only provides a guarantee for the closeness of centralized component mean vectors to the SVD subspace of centralized samples, we need to further extend it to show that the subspace spanned by all the centralized component mean vectors is close to the SVD subspace of centralized samples. This is described in the following lemma.

Lemma 13. *Let* $\mathbf{V} \in \mathbb{R}^{F \times N}$ *be generated from a mixture of* K *log-concave distributions that satisfies the non-degeneracy condition (cf. Definition 1). Using the notations defined above (see, in particular, the definition of* $\lambda_{\min}$ *in Section 2.2.1),*

$$\left\|\mathbf{P}_{K-1}\mathbf{P}_{K-1}^T - \mathbf{Q}_{K-1}\mathbf{Q}_{K-1}^T\right\|_{\mathrm{F}}^2 \le \frac{2\sum_{k=1}^K w_k \mathrm{d}(\bar{\mathbf{u}}_k, \mathcal{W})^2}{\lambda_{\min}}. \tag{2.3.41}$$

Proof. We have that

$$\sum_{k=1}^K w_k \mathrm{d}(\bar{\mathbf{u}}_k, \mathcal{W})^2 = \sum_{k=1}^K w_k \|\bar{\mathbf{u}}_k - \mathbf{P}_{K-1}\mathbf{P}_{K-1}^T\bar{\mathbf{u}}_k\|_2^2 \tag{2.3.42}$$

$$= \sum_{k=1}^K w_k \bar{\mathbf{u}}_k^T \mathbf{P}_{-(K-1)}\mathbf{P}_{-(K-1)}^T \bar{\mathbf{u}}_k = \mathrm{tr}(\mathbf{A}^T\mathbf{A}), \tag{2.3.43}$$

where $\mathbf{A} := \mathbf{P}_{-(K-1)}^T\mathbf{U}\mathbf{D}$, $\mathbf{U} := [\bar{\mathbf{u}}_1, \ldots, \bar{\mathbf{u}}_K]$, and $\mathbf{D} := \mathrm{diag}(\sqrt{w_1}, \ldots, \sqrt{w_K})$. Recall that we write the SVD of $\bar{\mathbf{\Sigma}}_0 := \sum_{k=1}^K w_k \bar{\mathbf{u}}_k \bar{\mathbf{u}}_k^T$ as $\bar{\mathbf{\Sigma}}_0 = \mathbf{Q}_{K-1}\mathbf{E}_{K-1}\mathbf{Q}_{K-1}^T$. We have

$$\mathrm{tr}(\mathbf{A}^T\mathbf{A}) = \mathrm{tr}(\mathbf{U}\mathbf{D}^2\mathbf{U}^T\mathbf{P}_{-(K-1)}\mathbf{P}_{-(K-1)}^T) \tag{2.3.44}$$

$$= \mathrm{tr}(\mathbf{E}_{K-1}\mathbf{Q}_{K-1}^T\mathbf{P}_{-(K-1)}\mathbf{P}_{-(K-1)}^T\mathbf{Q}_{K-1}) \tag{2.3.45}$$

$$\ge \lambda_{\min}\mathrm{tr}(\mathbf{Q}_{K-1}^T\mathbf{P}_{-(K-1)}\mathbf{P}_{-(K-1)}^T\mathbf{Q}_{K-1}), \tag{2.3.46}$$

where the inequality is because the diagonal entries of $\mathbf{Q}_{K-1}^T\mathbf{P}_{-(K-1)}\mathbf{P}_{-(K-1)}^T\mathbf{Q}_{K-1}$ are nonnegative.

Let $\beta := \sum_{k=1}^K w_k \mathrm{d}(\bar{\mathbf{u}}_k, \mathcal{W})^2$. By combining the above inequalities, we have

$$\|\mathbf{P}_{-(K-1)}^T\mathbf{Q}_{K-1}\|_{\mathrm{F}}^2 \le \frac{\beta}{\lambda_{\min}}. \tag{2.3.47}$$

Since

$$\|\mathbf{P}_{K-1}^T\mathbf{Q}_{K-1}\|_{\mathrm{F}}^2 + \|\mathbf{P}_{-(K-1)}^T\mathbf{Q}_{K-1}\|_{\mathrm{F}}^2 = \|\mathbf{P}^T\mathbf{Q}_{K-1}\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{Q}_{K-1}^T\mathbf{P}\mathbf{P}^T\mathbf{Q}_{K-1}) \tag{2.3.48}$$

$$= K - 1 = \|\mathbf{P}_{K-1}^T\mathbf{Q}\|_{\mathrm{F}}^2 = \|\mathbf{P}_{K-1}^T\mathbf{Q}_{K-1}\|_{\mathrm{F}}^2 + \|\mathbf{P}_{K-1}^T\mathbf{Q}_{-(K-1)}\|_{\mathrm{F}}^2, \tag{2.3.49}$$

where $\mathbf{Q}_{-(K-1)} \in \mathbb{R}^{F \times (F-K+1)}$ is selected such that $[\mathbf{Q}_{K-1}, \mathbf{Q}_{-(K-1)}]$ is orthogonal, we also have that $\|\mathbf{P}_{K-1}^T\mathbf{Q}_{-(K-1)}\|_{\mathrm{F}}^2 \le \frac{\beta}{\lambda_{\min}}$. Now we have

$$\begin{aligned}
&\|\mathbf{Q}_{K-1}\mathbf{Q}_{K-1}^T - \mathbf{P}_{K-1}\mathbf{P}_{K-1}^T\|_{\mathrm{F}}^2 \\
&= \mathrm{tr}(\mathbf{Q}_{K-1}^T\mathbf{P}_{-(K-1)}\mathbf{P}_{-(K-1)}^T\mathbf{Q}_{K-1}) + \mathrm{tr}(\mathbf{P}_{K-1}^T\mathbf{Q}_{-(K-1)}\mathbf{Q}_{-(K-1)}^T\mathbf{P}_{K-1}) \quad (2.3.50)\\
&= \|\mathbf{P}_{-(K-1)}^T\mathbf{Q}_{K-1}\|_{\mathrm{F}}^2 + \|\mathbf{P}_{K-1}^T\mathbf{Q}_{-(K-1)}\|_{\mathrm{F}}^2 \le \frac{2\beta}{\lambda_{\min}}, \quad (2.3.51)
\end{aligned}$$

concluding the proof of Lemma 13. □

Combining the results of Corollary 5 and Lemma 13, we obtain an upper bound of the distance between the two subspaces, and we can prove a result concerning optimal clusterings of the dimensionality-reduced dataset (via PCA) generated from a mixture of log-concave distributions. This parallels the procedures for the proof strategy of Theorem 2.

Now, we demonstrate that under similar assumptions on the generating process of the samples (compared to those in Theorem 3), any optimal clustering for the post-PCA (cf. Section 2.1.3) dataset is also close to the correct target clustering with high probability.

Theorem 4. *Define $\bar{L} := \sum_{k=1}^K w_k\left(\|\mathbf{u}_k\|_2^2 + \mathrm{tr}(\mathbf{\Sigma}_k)\right)$. Let the dataset $\mathbf{V} \in \mathbb{R}^{F\times N}$ be generated under the same conditions given in Theorem 3 with the separability assumption in (2.3.7) being modified to*

$$0 < \delta_3 := \frac{(K-1)\bar{\sigma}_{\max}^2 + a}{\lambda_{\min} + \bar{\sigma}_{\min}^2 - b} < \zeta(w_{\min}), \tag{2.3.52}$$

where

$$a := (1+K)\bar{L}\sqrt{\frac{2(K-1)\bar{\sigma}_{\max}^2}{\lambda_{\min}}}, \quad \text{and} \quad b := (\bar{L} - \|\bar{\mathbf{u}}\|_2^2)\sqrt{\frac{2(K-1)\bar{\sigma}_{\max}^2}{\lambda_{\min}}}. \tag{2.3.53}$$

Let $\tilde{\mathbf{V}} \in \mathbb{R}^{F\times(K-1)}$ be the post-$(K-1)$-PCA dataset of $\mathbf{V}$. Then for any sufficiently small $\epsilon \in (0,1)$, if $N \ge C\frac{F^2K^6}{\epsilon^2}\log^5\left(\frac{F^2K^4}{\epsilon\eta}\right)$, where $C > 0$ depends on the parameters of the mixture model, we have, with probability at least $1-\eta$,

$$\mathrm{d}_{\mathrm{ME}}(\mathscr{C}, \tilde{\mathscr{C}}^{\mathrm{opt}}) \le \tau\left(\frac{(K-1)\bar{\sigma}_{\max}^2 + a + \epsilon}{\lambda_{\min} + \bar{\sigma}_{\min}^2 - b - \epsilon}\right)(w_{\max} + \epsilon), \tag{2.3.54}$$

where $\mathscr{C}$ is the correct target clustering and $\tilde{\mathscr{C}}^{\mathrm{opt}}$ is an optimal K-clustering for $\tilde{\mathbf{V}}$.

Proof. We use the same notations as those in the proof of Theorem 2 and in the statement of Theorem 4. Since $\mathbf{Q}_{K-1} \in \mathbb{R}^{F\times(K-1)}$ has orthonormal columns,

$$\operatorname{tr}(\mathbf{Q}_{K-1}^T \mathbf{\Sigma}_k \mathbf{Q}_{K-1}) \le \sum_{j=1}^{K-1} \lambda_j(\mathbf{\Sigma}_k) \le (K-1)\sigma_{k,\max}^2, \quad \forall\, k \in [K]. \tag{2.3.55}$$

Therefore, similar to that for the case for the original dataset (cf. the inequality in (2.3.11)), if $N \ge C_1 \frac{K^4}{\epsilon^2} \log^5\left(\frac{K^4}{\epsilon\eta}\right)$ with $C_1 > 0$ being sufficiently large,

$$\mathbb{P}\left(\frac{1}{N}\mathcal{D}(\hat{\mathbf{V}}, \mathcal{I}) - (K-1)\bar{\sigma}_{\max}^2 \ge \frac{\epsilon}{2}\right) \le \eta. \tag{2.3.56}$$

In addition, we have $\lambda_{K-1}(\mathbf{Q}_{K-1}^T \mathbf{\Sigma}_k \mathbf{Q}_{K-1}) \ge \lambda_F(\mathbf{\Sigma}_k) = \sigma_{k,\min}^2$.[4] Similarly, we have $\lambda_1(\mathbf{Q}_{K-1}^T \mathbf{\Sigma}_k \mathbf{Q}_{K-1}) \le \lambda_1(\mathbf{\Sigma}_k) = \sigma_{k,\max}^2$. Thus if $N \ge C_1 \frac{K^4}{\epsilon^2} \log^5\left(\frac{K^3}{\epsilon\eta}\right)$,

$$\mathbb{P}\left(\frac{1}{N}\lambda_{K-1}(\hat{\mathbf{S}}) - (\lambda_{\min} + \bar{\sigma}_{\min}^2) \le -\frac{\epsilon}{2}\right) \le \eta. \tag{2.3.57}$$

Recall that we write $\mathbf{R} := \mathbf{Q}_{K-1}\mathbf{Q}_{K-1}^T - \mathbf{P}_{K-1}\mathbf{P}_{K-1}^T$. Let $r := \sqrt{\frac{2(K-1)\bar{\sigma}_{\max}^2}{\lambda_{\min}}}$. Combining Corollary 5 and Lemma 13, we have that if $N \ge C\frac{F^2K^4}{\epsilon^2}\log^5\left(\frac{FK^3}{\epsilon\eta}\right)$, with probability at least $1-\eta$,

$$\|\mathbf{R}\|_{\mathrm{F}} \le r + C_2\epsilon, \tag{2.3.58}$$

where $C_2 > 0$ is sufficiently large. In addition, using the inequalities

$$\frac{1}{N}\left|\mathcal{D}(\hat{\mathbf{V}}, \mathscr{I}) - \mathcal{D}(\tilde{\mathbf{V}}, \mathscr{I})\right| \le \frac{1+K}{N}\|\mathbf{V}\|_{\mathrm{F}}^2\|\mathbf{R}\|_{\mathrm{F}} \tag{2.3.59}$$

$$\frac{1}{N}\left\|\hat{\mathbf{S}} - \tilde{\mathbf{S}}\right\|_2 \le \|\mathbf{R}\|_{\mathrm{F}}\|\mathbf{Z}\|_{\mathrm{F}}^2, \tag{2.3.60}$$

[4] Indeed, assume, to the contrary, that $\lambda_{K-1}(\mathbf{Q}_{K-1}^T\mathbf{\Sigma}_k\mathbf{Q}_{K-1}) < \sigma_{k,\min}^2$. Then there is a $\lambda < \sigma_{k,\min}^2$ and a corresponding unit vector $\mathbf{x} \in \mathbb{R}^{K-1}$, such that $\mathbf{Q}_{K-1}^T\mathbf{\Sigma}_k\mathbf{Q}_{K-1}\mathbf{x} = \lambda\mathbf{x}$. Thus, $\sigma_{k,\min}^2\|\mathbf{x}\|_2^2 = \sigma_{k,\min}^2\|\mathbf{Q}_{K-1}\mathbf{x}\|_2^2 \le \mathbf{x}^T\mathbf{Q}_{K-1}^T\mathbf{\Sigma}_k\mathbf{Q}_{K-1}\mathbf{x} = \lambda\|\mathbf{x}\|_2^2 < \sigma_{k,\min}^2\|\mathbf{x}\|_2^2$, which is a contradiction.

we deduce that

$$\mathbb{P}\left(\frac{1}{N}\left|\mathcal{D}(\hat{\mathbf{V}},\mathscr{I})-\mathcal{D}(\tilde{\mathbf{V}},\mathscr{I})\right|-a\geq\frac{\epsilon}{2}\right)$$
$$\leq\mathbb{P}\left(\frac{1}{N}\|\mathbf{V}\|_{\mathrm{F}}^2\|\mathbf{R}\|_{\mathrm{F}}-\bar{L}r\geq\frac{\epsilon}{2(1+K)}\right)\tag{2.3.61}$$
$$\leq\mathbb{P}\left(\left(\frac{1}{N}\|\mathbf{V}\|_{\mathrm{F}}^2-\bar{L}\right)r\geq\frac{\epsilon}{4(1+K)}\right)+\mathbb{P}\left(\frac{1}{N}\|\mathbf{V}\|_{\mathrm{F}}^2\geq\bar{L}+1\right)$$
$$+\mathbb{P}\left(\|\mathbf{R}\|_{\mathrm{F}}-r\geq\frac{\epsilon}{4(1+K)(\bar{L}+1)}\right).\tag{2.3.62}$$

Therefore, by (2.3.9) and (2.3.58), we obtain that if $N\geq C_1\frac{F^2K^6}{\epsilon^2}\log^5\left(\frac{F^2K^4}{\epsilon\eta}\right)$,

$$\mathbb{P}\left(\frac{1}{N}\left|\mathcal{D}(\hat{\mathbf{V}},\mathscr{I})-\mathcal{D}(\tilde{\mathbf{V}},\mathscr{I})\right|-a\geq\frac{\epsilon}{2}\right)\leq\eta.\tag{2.3.63}$$

Similarly, when $N\geq C_1\frac{F^2K^6}{\epsilon^2}\log^5\left(\frac{F^2K^4}{\epsilon\eta}\right)$,

$$\mathbb{P}\left(\frac{1}{N}\left\|\hat{\mathbf{S}}-\tilde{\mathbf{S}}\right\|_2-b\geq\frac{\epsilon}{2}\right)\leq\eta.\tag{2.3.64}$$

Combining these results with Corollary 1, we obtain the desired conclusion. □

Note that by using the fact that $\mathbf{w}=(w_1,w_2,\dots,w_K)$ is a probability vector and the non-degeneracy condition, we have $\sum_{k=1}^K w_k\|\mathbf{u}_k\|_2^2>\|\bar{\mathbf{u}}\|_2^2$ and thus $b>0$. The proof of Theorem 4 is similar to that for Theorem 2, except that the estimate of $\|\mathbf{P}_{K-1}\mathbf{P}_{K-1}^T-\mathbf{Q}_{K-1}\mathbf{Q}_{K-1}^T\|_{\mathrm{F}}$ is obtained differently. The separability assumption (2.3.7) for mixtures of log-concave distributions reduces to (2.2.5) for spherical GMMs because in this case, we have $\bar{\sigma}_{\max}^2=\bar{\sigma}_{\min}^2=\bar{\sigma}^2$. If the mixture model is non-spherical, the separability assumption (2.3.7) is generally stricter than (2.2.5). This is especially the case when $\bar{\sigma}_{\max}^2\gg\bar{\sigma}_{\min}^2$. This implies non-spherical mixture models are generally more difficult to disambiguate and learn. For dimensionality-reduced datasets (using PCA), the separability assumption in (2.3.52) is stricter than the separability assumption in (2.2.56), even for spherical GMMs, because of the presence of the additional positive terms a and b in (2.3.53). In addition, unlike that for spherical GMMs, the separability assumption in (2.3.52) for dimensionality-reduced datasets may also be stricter than the separability assumption in (2.3.7) also because of the same additional positive terms a and b.

2.4 Discussion and Other Perspectives

In this section, we discuss several interesting and practically-relevant extensions of the preceding results. In Section 2.4.1, we show that Theorems 2 and 4 may be readily extended to other dimensionality-reduction techniques besides PCA/SVD by leveraging results such as (2.1.10). In Section 2.4.2, we show that we can apply efficient clustering algorithms to obtain an approximately-optimal clustering which is also close to the correct target clustering.

2.4.1 Other Dimensionality-Reduction Techniques

Our results can be used to prove similar upper bounds for the ME distance between any *approximately-optimal* clustering and the correct target clustering. The following corollary follows easily from Lemma 1.

Corollary 6. *Consider a K-clustering $\mathscr{C}$ with corresponding δ (cf. Lemma 1) and a K-clustering $\mathscr{C}'$ that satisfies*

$$\mathcal{D}(\mathbf{V}, \mathscr{C}') \le \gamma \mathcal{D}(\mathbf{V}, \mathscr{C}^{\mathrm{opt}}), \tag{2.4.1}$$

for some $\gamma \ge 1$. Then if

$$\delta_\gamma := \frac{\gamma \mathcal{D}(\mathbf{V}, \mathscr{C}) - \mathcal{D}^*(\mathbf{V})}{\lambda_{K-1}(\mathbf{S}) - \lambda_K(\mathbf{S})}, \tag{2.4.2}$$

satisfies

$$\delta_\gamma \le \frac{K-1}{2}, \quad \text{and} \quad \tau(\delta_\gamma) \le p_{\min}, \tag{2.4.3}$$

we have

$$\mathrm{d}_{\mathrm{ME}}(\mathscr{C}', \mathscr{C}) \le p_{\max} \tau(\delta). \tag{2.4.4}$$

Proof. We have $\delta \le \delta_\gamma \le \frac{1}{2}(K-1)$ and $\tau(\delta, \delta_\gamma) \le \tau(\delta_\gamma) \le p_{\min}$. Lemma 1 thus yields $\mathrm{d}_{\mathrm{ME}}(\mathscr{C}, \mathscr{C}') \le p_{\max}\tau(\delta)$. □

According to the above corollary, we deduce that if we make a stronger separability assumption as in (2.4.3), we can bound the ME distance between any approximately-optimal clustering and the correct target clustering. Therefore, by leveraging (2.1.10), our theoretical results for dimensionality reduction by PCA (in Theorems 2 and 4) can be extended to other dimensionality-reduction techniques such as random projection [67–70] and randomized SVD [19, 20, 70]. We describe these dimensionality-reduction techniques in the following and provide known results for γ satisfying (2.1.10).

- A *random projection* from F dimensions to $D < F$ dimensions is represented by a $D \times F$ matrix, which can be generated as follows [71]: (i) Set each entry of the matrix to be an i.i.d. $\mathcal{N}(0, 1)$ random variable; (ii) Orthonormalize the rows of the matrix. Theoretical guarantees for dimensionality-reduction via random projection are usually established by appealing to the well-known Johnson-Lindenstrauss lemma [72] which says that pairwise distances and inner products are approximately preserved under the random projection.

- Because computing an exact SVD is generally expensive, *randomized SVD* has gained tremendous interest for solving large-scale problems. For a data matrix $\mathbf{V} \in \mathbb{R}^{F \times N}$, to reduce the dimensionality of the columns from F to $K < F$, one performs a randomized SVD using an $F \times K$ matrix $\mathbf{Z}_K$. More specifically, we can adopt the following procedure [60]: (i) Generate a $D \times F$ $(D > K)$ matrix $\mathbf{L}$ whose entries are i.i.d. $\mathcal{N}(0, 1)$ random variables; (ii) Let $\mathbf{A} = \mathbf{L}\mathbf{V} \in \mathbb{R}^{D \times N}$ and orthonormalize the rows of $\mathbf{A}$ to construct a matrix $\mathbf{B}$; (iii) Let $\mathbf{Z}_K \in \mathbb{R}^{F \times K}$ be the matrix of top K left singular vectors of $\mathbf{V}\mathbf{B}^T \in \mathbb{R}^{F \times D}$. Such $\mathbf{Z}_K$ is expected to satisfy $\|\mathbf{V} - \mathbf{Z}_K\mathbf{Z}_K^T\mathbf{V}\|_\mathrm{F} \approx \|\mathbf{V} - \mathbf{P}_K\mathbf{P}_K^T\mathbf{V}\|_\mathrm{F}$, where $\mathbf{P}_K$ is the matrix of top K left singular vectors of $\mathbf{V}$. The key advantage of randomized SVD over exact SVD is that when $D \ll \min\{F, N\}$, the computation of the randomized SVD is significantly faster than the computing of an exact SVD.

- We may also employ feature selection techniques such as those described in [73]

Table 2.1: Summary of Feature Extraction Techniques

Technique	Reference	Dimensions	γ
PCA/SVD	[75] [61]	K $\lceil K/\epsilon \rceil$	2 $1+\epsilon$
random projection	[61]	$O(K/\epsilon^2)$	$1+\epsilon$
randomized SVD	[61]	$\lceil K/\epsilon \rceil$	$1+\epsilon$

and [74].

A subset of the results of [61] is presented in Table 2.1. This table shows the γ such that (2.1.10) is satisfied for various reduced dimensions and dimensionality reduction techniques. Even though the results in Table 2.1 appear promising, we observe from numerical experiments that for dimensionality reduction by PCA/SVD, if the data matrix is generated from a spherical GMM, even if it is moderately-well separated (cf. Section 2.2.3 to follow), $\mathcal{D}(\mathbf{V}, \tilde{\mathscr{C}}^{\mathrm{opt}}) \approx \mathcal{D}(\mathbf{V}, \mathscr{C}^{\mathrm{opt}})$. That is, in this case, $\gamma \approx 1$ even though the reduced dimensionality is $K-1$ or K . Furthermore, we show in Theorem 2 that for dimensionality reduction by PCA, we require a weaker separability assumption (compared to that in Theorem 1). However, from Table 2.1, for SVD, if the reduced dimensionality is K, then $\gamma = 2$ and we will require a stronger separability assumption according to Corollary 6. Therefore, the results for PCA/SVD in Table 2.1 are generally pessimistic. This is reasonable because PCA/SVD is data-dependent and we have assumed specific generative mixture models for our data matrices.

2.4.2 Other Efficient Clustering Algorithms

Although k-means is a popular heuristic algorithm that attempts to minimize the sum-of-squares distortion, in general, minimizing this objective is NP-hard and k-means only converges to a locally optimal solution. In addition, k-means is sensitive to initialization [5]. Fortunately, there are variants of k-means with judiciously chosen initializations that possess theoretical guarantees, e.g., k-means++ [5]. In addition, efficient variants of k-means [45,76] have been proposed to find approximately-optimal clusterings under appropriate conditions. Our theoretical results can be easily combined with these efficient algorithms to produce approximately-optimal clusterings which are also close to the correct target clustering. We demonstrate this by using a result in [76]. Namely, if we denote the optimal distortion with $k \in \mathbb{N}$ clusters as OPT_k, Theorem 4.13 in [76] states that:

Lemma 14. *If $\frac{\mathrm{OPT}_K}{\mathrm{OPT}_{K-1}} \le \epsilon^2$ for a small enough $\epsilon > 0$, the randomized algorithm presented before Theorem 4.13 in [76] returns a solution of cost at most $\left(\frac{1-\epsilon^2}{1-37\epsilon^2}\right)\mathrm{OPT}_K$ with probability $1 - O(\sqrt{\epsilon})$ in time $O(FNK + K^3F)$.*

We demonstrate that this lemma and the proposed algorithm can be combined with our theoretical results to produce further interesting results. For simplicity, we assume the data matrix $\mathbf{V} \in \mathbb{R}^{F\times N}$ is generated from a K-component spherical GMM. Then by previous calculations, the lower bound of the distortion for $K-1$ clusters is $\mathcal{D}^*_{K-1} := \sum_{k=K-1}^{F} \lambda_k(\mathbf{S})$ (cf. Section 2.1.2). As $N \to \infty$, $\mathcal{D}^*_{K-1}$ converges to $\lambda_{\min} + (F-K+2)\bar{\sigma}^2$ in probability. In addition, the distortion for the correct target clustering (with K clusters) converges to $F\bar{\sigma}^2$ in probability. Therefore, if the number of samples is large enough, with high probability,

$$\frac{\mathrm{OPT}_K}{\mathrm{OPT}_{K-1}} \le \frac{F\bar{\sigma}^2}{\lambda_{\min} + (F-K+2)\bar{\sigma}^2}. \tag{2.4.5}$$

Thus if $\bar{\sigma}^2$ is sufficiently small or $\lambda_{\min}$ is sufficiently large, by Lemma 14, we can use the algorithm suggested therein to obtain an approximately-optimal clustering for the original dataset. In addition, by Theorem 1 and Corollary 6, under an appropriate separability assumption, the approximately-optimal clustering is close to the correct target clustering that we ultimately seek.

We have provided one example of an efficient algorithm to obtain an approximately-optimal clustering. Interested readers may refer to the paper by

Ackerman and Ben-David [45] which discusses other computationally efficient algorithms with guarantees.

Chapter 3

A New Initialization Method for NMF

As we mentioned in Chapter 1, classical algorithms for NMF [23, 29, 31, 32] typically have no theoretical guarantees beyond guaranteeing that the sequence of the values of the objective function is non-increasing and the iterates converge to a stationary point. Further, there are no analyses concerning error bounds. Recently, near-separable NMF has become popular and researchers are able to derive error bound analysis for near-separable NMF. However, usually, to leverage on convex analysis technique, a strong assumption that the data matrix $\mathbf{V}$ is normalized such that each column (or row) of it has unit ℓ_1 norm is made. As pointed out in [37], normalization, especially in the ℓ_1-norm for the rows of data matrices may degrade the clustering performance for text datasets significantly. In this chapter, we propose a geometric assumption which can be considered as a special case of the near-separability assumption. We design an algorithm named `cr1-nmf`. It first uses the geometric assumption to obtain an exact clustering of the columns of the data matrix; subsequently, it employs several rank-one NMFs to obtain the final decomposition. In particular, we are able to derive error bound analysis without the normalization assumption. Numerical experiment results reveal that our algorithm can be used as a good initializer for classical NMF algorithms.

3.1 Background

In this section, we describe some works that are related to ours.

3.1.1 Near-Separable NMF

Arora et al. [34] provide an algorithm that runs in time polynomial in F, N and K to find the correct factor matrices under the separability condition. Furthermore, the authors consider the near-separable case and prove an approximation error bound when the original data matrix $\mathbf{V}$ is slightly perturbed from being separable. The algorithm and the theorem for near-separable case is also presented in [36]. The main ideas behind the theorem are as follows: first, $\mathbf{V}$ must be normalized such that every row of it has unit ℓ_1 norm; this assumption simplifies the conical hull for exact NMF to a convex hull. Second, the rows of $\mathbf{H}$ need to be robustly simplicial, i.e., every row of $\mathbf{H}$ should not be contained in the convex hull of all other rows and the largest perturbation of the rows of $\mathbf{V}$ should be bounded by a function of the smallest distance from a row of $\mathbf{H}$ to the convex hull of all other rows. Later we will show in Section 3.2 that our geometric assumption stated in inequality (3.2.2) is similar to this key idea in [34]. Although we impose a clustering-type generating assumption for data matrix, we do not need the normalization assumption in [34], which is stated in [37] that may lead to bad clustering performance for text datasets. In addition, because we do not impose this normalization assumption, instead of providing an upper bound on the approximation error, we provide the upper bound for relative error, which is arguably more natural.

3.1.2 Initialization Techniques for NMF

Similar to k-means, NMF can easily be trapped at bad local optima and is sensitive to initialization. We find that our algorithm is particularly amenable to provide good initial factor matrices for subsequently applying standard NMF algorithms. Thus, here we mention some works on initialization for NMF. Spherical k-means (`spkm`) is a simple clustering method and it is shown to be one of the most efficient algorithms for document clustering [77]. The authors in [46] consider using `spkm` for initializing the left factor matrix $\mathbf{W}$ and observe a better convergence rate compared to random initialization. Other clustering-based initialization approaches for NMF including divergence-based k-means [78] and fuzzy clustering [79]. It is also natural to consider using singular value decomposition (SVD) to initialize NMF. In fact, if there is no nonnegativity constraint, we can obtain the best rank-K approximation of a given matrix directly using SVD, and there are strong relations between NMF and SVD. For example, we can obtain the best rank-one NMF from the best rank-one SVD (see Lemma 17), and if the best rank-two approximation matrix of a nonnegative data matrix is also nonnegative, then we can also obtain best rank-two NMF [44]. Moreover, for a general positive integer K, it is shown in [47] that nonnegative double singular value decomposition (`nndsvd`), a deterministic SVD-based approach, can be used to enhance the initialization of NMF, leading to a faster reduction of the approximation error of many NMF algorithms. The CUR decomposition-based initialization method [80] is another factorization-based initialization approach for NMF. We compare our algorithm to widely-used algorithms for initializing NMF in Section 3.6.2.

3.2 Our Problem Formulation

In this section, we first present our geometric assumption and prove that the exact clustering can be obtained for the normalized data points under the geometric assumption. Next, we introduce several useful lemmas in preparation for the proofs of the main theorems in subsequent sections.

Our Geometric Assumption for the data matrix $\mathbf{V}$ is presented in the following. We assume the columns of $\mathbf{V}$ lie in K circular cones which satisfy a geometric assumption presented in (3.2.2) to follow. We define *circular cones* as follows:

Definition 3. *Given* $\mathbf{u} \in \mathbb{R}_+^F$ *with unit* ℓ_2 *norm and an angle* $\alpha \in (0, \pi/2)$, *the*

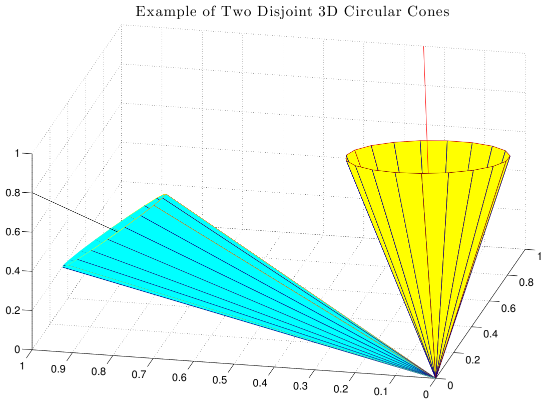

Figure 3.1: Illustration of the geometric assumption in (3.2.2). Here $\alpha_1 = \alpha_2 = 0.2$ and $\beta_{12} = 0.9 > 3\alpha_1 + \alpha_2$.

circular cone with respect to (w.r.t.) $\mathbf{u}$ and α *is defined as*

$$\mathcal{C}(\mathbf{u}, \alpha) := \left\{\mathbf{x} \in \mathbb{R}^F \setminus \{0\} : \frac{\mathbf{x}^T \mathbf{u}}{\|\mathbf{x}\|_2} \geq \cos \alpha\right\}. \tag{3.2.1}$$

In other words, $\mathcal{C}(\mathbf{u}, \alpha)$ contains all $\mathbf{x} \in \mathbb{R}^F \setminus \{0\}$ for which the angle between $\mathbf{u}$ and $\mathbf{x}$ is not larger than α. We say that α and $\mathbf{u}$ are respectively the *size angle* and *basis vector* of the circular cone. In addition, the corresponding truncated circular cone in nonnegative orthant is $\mathcal{C}(\mathbf{u}, \alpha) \cap \mathcal{P}$, where $\mathcal{P} := \mathbb{R}_+^F$.

We assume that there are K truncated circular cones $C_1 \cap \mathcal{P}, \ldots, C_K \cap \mathcal{P}$ with corresponding basis vectors and size angles, i.e., $C_k := \mathcal{C}(\mathbf{u}_k, \alpha_k)$ for $k \in [K]$. Let $\beta_{ij} := \arccos\left(\mathbf{u}_i^T \mathbf{u}_j\right)$. We make the geometric assumption that the columns of our data matrix $\mathbf{V}$ lie in K truncated circular cones which satisfy

$$\min_{i,j \in [K], i \neq j} \beta_{ij} > \max_{i,j \in [K], i \neq j} \alpha_i + 3\alpha_j. \tag{3.2.2}$$

If we sort $\alpha_1, \ldots, \alpha_K$ as $\hat{\alpha}_1, \ldots, \hat{\alpha}_K$ such that $\hat{\alpha}_1 \geq \hat{\alpha}_2 \geq \ldots \geq \hat{\alpha}_K$, (3.2.2) is equivalent to

$$\min_{i,j \in [K], i \neq j} \beta_{ij} > 3\hat{\alpha}_1 + \hat{\alpha}_2. \tag{3.2.3}$$

The size angle α_k is a measure of perturbation in k-th circular cone and $\beta_{ij}, i \neq j$ is a measure of distance between the i-th basis vector and the j-th basis vector. Thus, (3.2.2) is similar to the second idea in [34] (cf. Section 3.1.1), namely, that the largest perturbation of the rows of $\mathbf{V}$ is bounded by a function of the smallest distance from a row of $\mathbf{H}$ to the convex hull of all other rows. This assumption is realistic for datasets whose samples can be clustered into distinct types; for example, image datasets in which images either contain a distinct foreground (e.g., a face) embedded on a background, or they only comprise a background. See Figure 3.1 for an illustration of the geometric assumption in (3.2.2) and refer to Figure 1 in [36] for an illustration of the separability condition.

Now we discuss the relation between our geometric assumption and the separability and near-separability [34, 36] conditions that have appeared in the literature (and discussed in Section 1.4). Consider a data matrix $\mathbf{V}$ generated under the extreme case of our geometric assumption that all the size angles of the K circular cones are zero. Then every column of $\mathbf{V}$ is a nonnegative multiple of a basis vector of a circular cone. This means that all the columns of $\mathbf{V}$ can be represented as nonnegative linear combinations of K columns, i.e., the K basis vectors $\mathbf{u}_1, \ldots, \mathbf{u}_K$. This can be considered as a special case of separability assumption. When the size angles are not all zero, our geometric assumption can be considered as a special case of the near-separability assumption.

In Lemma 15, we show that Algorithm 1, which has time complexity $O(KFN)$, correctly clusters the columns of $\mathbf{V}$ under the geometric assumption.

Lemma 15. *Under the geometric assumption on* $\mathbf{V}$*, if Algorithm 1 is applied to* $\mathbf{V}$*, then the columns of* $\mathbf{V}$ *are partitioned into* K *subsets, such that the data points in the same subset are generated from the same truncated circular cone.*

Proof. We normalize $\mathbf{V}$ to obtain $\mathbf{V}'$, such that all the columns of $\mathbf{V}'$ have unit ℓ_2 norm. From the definition, we know if a data point is in a truncated circular cone, then the normalized data point is also in the truncated circular cone. Then for any two columns $\mathbf{x}$, $\mathbf{y}$ of $\mathbf{V}'$ that are in the same truncated circular cone $C_k \cap \mathbb{R}_+^F$, $k \in [K]$, the largest possible angle between them is $\min\{2\alpha_k, \pi/2\}$, and thus the distance $\|\mathbf{x}-\mathbf{y}\|_2$ between these two data points is not larger than $\sqrt{2\left(1-\cos\left(2\alpha_k\right)\right)}$. On the other hand, for any two columns $\mathbf{x}$, $\mathbf{y}$ of $\mathbf{V}'$ that are in two truncated circular cones

Algorithm 1 Greedy clustering method with geometric assumption in (3.2.2)

```
Input: Data matrix $\mathbf{V} \in \mathbb{R}_+^{F\times N}$, $K \in \mathbb{N}$
Output: A set of non-empty, pairwise disjoint index sets $\mathscr{I}_1, \mathscr{I}_2, \dots, \mathscr{I}_K \subseteq [N]$ such that their union is $[N]$
1) Normalize $\mathbf{V}$ to obtain $\mathbf{V}'$, such that all the columns of $\mathbf{V}'$ have unit $\ell_2$ norm.
2) Arbitrarily pick a point $\mathbf{z}_1 \in \mathbf{V}'$ (i.e., $\mathbf{z}_1$ is a column in $\mathbf{V}'$) as the first centroid.
3) for $k = 1$ to $K-1$ do
        $\mathbf{z}_{k+1} := \arg\min_{\mathbf{z}\in\mathbf{V}'}\{\max\{\mathbf{z}_i^T\mathbf{z}, i \in [k]\}\}$    (3.2.4)
    and set $\mathbf{z}_{k+1}$ be the $(k+1)$-st centroid.
4) $\mathscr{I}_k := \{n \in [N] : k = \arg\max_{j\in[K]} \mathbf{z}_j^T\mathbf{V}'(:,n)\}$ for all $k \in [K]$.
```

$C_i \cap \mathbb{R}_+^F, C_j \cap \mathbb{R}_+^F, i \neq j$, the smallest possible angle between them is $\beta_{ij} - \alpha_i - \alpha_j$, thus the smallest possible distance between them is $\sqrt{2\left(1-\cos\left(\beta_{ij}-\alpha_i-\alpha_j\right)\right)}$. Then under the geometric assumption (3.2.2), the distance between any two unit data points in distinct truncated circular cones is larger than the distance between any two unit data points in the same truncated circular cone. Hence, Algorithm 1 returns the correct clusters. □

Now we present the following two useful lemmas. Lemma 16 provides an upper bound for perturbations of singular values. Lemma 17 shows that we can directly obtain the best rank-one nonnegative matrix factorization from the best rank-one SVD.

Lemma 16 (Perturbation of singular values [62])**.** *If* $\mathbf{A}$ *and* $\mathbf{A}+\mathbf{E}$ *are in* $\mathbb{R}^{F\times N}$*, then*

$$\sum_{p=1}^{P} \left(\sigma_p(\mathbf{A}+\mathbf{E}) - \sigma_p(\mathbf{A})\right)^2 \leq \|\mathbf{E}\|_{\mathrm{F}}^2, \tag{3.2.5}$$

where $P = \min\{F, N\}$ *and* $\sigma_p(\mathbf{A})$ *is the* p*-th largest singular value of* $\mathbf{A}$*. In addition, we have*

$$|\sigma_p(\mathbf{A}+\mathbf{E}) - \sigma_p(\mathbf{A})| \leq \sigma_1(\mathbf{E}) = \|\mathbf{E}\|_2 \tag{3.2.6}$$

for any $p \in [P]$.

Lemma 17 (Rank-One Approximate NMF [44])**.** *Let* $\sigma \mathbf{u}\mathbf{v}^T$ *be the rank-one truncated singular value decomposition of a matrix* $\mathbf{V} \in \mathbb{R}_+^{F \times N}$. *Then* $\mathbf{u}' := \sigma|\mathbf{u}|$, $\mathbf{v}' := |\mathbf{v}|$ *solves*

$$\min_{\mathbf{x} \in \mathbb{R}_+^F, \mathbf{y} \in \mathbb{R}_+^N} \|\mathbf{V} - \mathbf{x}\mathbf{y}^T\|_{\mathrm{F}}. \tag{3.2.7}$$

3.3 Relative Error Bounds for Non-Probabilistic Data

In this section, we first present a deterministic theorem concerning an upper bound for the relative error of NMF. Subsequently, we provide several extensions of this theorem.

Theorem 5. *Suppose all the data points in data matrix* $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ *are drawn from* K *truncated circular cones* $C_1 \cap \mathbb{R}_+^F, \dots, C_K \cap \mathbb{R}_+^F$, *where* $C_k := \mathcal{C}(\mathbf{u}_k, \alpha_k)$ *for* $k \in [K]$. *Then there is a pair of factor matrices* $\mathbf{W}^* \in \mathbb{R}_+^{F \times K}$, $\mathbf{H}^* \in \mathbb{R}_+^{K \times N}$, *such that*

$$\frac{\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_{\mathrm{F}}}{\|\mathbf{V}\|_{\mathrm{F}}} \le \max_{k \in [K]} \{\sin \alpha_k\}. \tag{3.3.1}$$

Proof. Define $\mathscr{I}_k := \{n \in [N] : \mathbf{v}_n \in C_k \cap \mathbb{R}_+^F\}$ (if a data point $\mathbf{v}_n$ is contained in more than one truncated circular cones, we arbitrarily assign the data point to any truncated circular cone that it is contained in). Then $\mathscr{I}_1, \mathscr{I}_2, \dots, \mathscr{I}_K \subseteq [N]$ are disjoint index sets and their union is $[N]$. Any two data points $\mathbf{V}(:, j_1)$ and $\mathbf{V}(:, j_2)$ are in the same circular cones if j_1 and j_2 are in the same index set. Let $\mathbf{V}_k = \mathbf{V}(:, \mathscr{I}_k)$ and without loss of generality, suppose that $\mathbf{V}_k \in C_k$ for $k \in [K]$. For any $k \in [K]$ and any column $\mathbf{z}$ of $\mathbf{V}_k$, suppose the angle between $\mathbf{z}$ and $\mathbf{u}_k$ is β, we have $\beta \le \alpha_k$ and $\mathbf{z} = \|\mathbf{z}\|_2(\cos\beta)\mathbf{u}_k + \mathbf{y}$, with $\|\mathbf{y}\|_2 = \|\mathbf{z}\|_2(\sin\beta) \le \|\mathbf{z}\|_2(\sin\alpha_k)$. Thus $\mathbf{V}_k$ can be written as the sum of a rank-one matrix $\mathbf{A}_k$ and a perturbation matrix $\mathbf{E}_k$. By Lemma 17, we can find the best rank-one approximate NMF of $\mathbf{V}_k$ from the singular value decomposition of $\mathbf{V}_k$. Suppose $\mathbf{w}_k^* \in \mathbb{R}_+^F$ and $\mathbf{h}_k \in \mathbb{R}_+^{|\mathscr{I}_k|}$

solve the best rank-one approximate NMF. Let $\mathbf{S}_k := \mathbf{w}_k^* \mathbf{h}_k^T$ be the best rank-one approximation matrix of $\mathbf{V}_k$. Let $P_k = \min\{F, |\mathscr{I}_k|\}$, then by Lemma 16, we have

$$\|\mathbf{V}_k - \mathbf{S}_k\|_{\mathrm{F}}^2 = \sum_{p=2}^{P_k} \sigma_p^2 \left(\mathbf{V}_k\right) = \sum_{p=2}^{P_k} \sigma_p^2 \left(\mathbf{A}_k + \mathbf{E}_k\right) \leq \|\mathbf{E}_k\|_{\mathrm{F}}^2. \tag{3.3.2}$$

From the previous result, we know that

$$\frac{\|\mathbf{E}_k\|_{\mathrm{F}}^2}{\|\mathbf{V}_k\|_{\mathrm{F}}^2} = \frac{\sum_{\mathbf{z} \in \mathbf{V}_k} \|\mathbf{z}\|_2^2 \sin^2 \beta_{\mathbf{z}}}{\sum_{\mathbf{z} \in \mathbf{V}_k} \|\mathbf{z}\|_2^2} \leq \sin^2 \alpha_k, \tag{3.3.3}$$

where $\beta_{\mathbf{z}}$ denotes the angle between $\mathbf{z}$ and $\mathbf{u}_k$, $\beta_{\mathbf{z}} \leq \alpha_k$, and $\mathbf{z} \in \mathbf{V}_k$ runs over all columns of the matrix $\mathbf{V}_k$.

Define $\mathbf{h}_k^* \in \mathbb{R}_+^N$ as $\mathbf{h}_k^*(n) = \mathbf{h}_k(n)$, if $n \in \mathscr{I}_k$ and $\mathbf{h}_k^*(n) = 0$ if $n \notin \mathscr{I}_k$. Let $\mathbf{W}^* := \left[\mathbf{w}_1^*, \mathbf{w}_2^*, \ldots, \mathbf{w}_K^*\right]$ and $\mathbf{H}^* := \left[\, (\mathbf{h}_1^*)^T ; (\mathbf{h}_2^*)^T \ldots ; (\mathbf{h}_K^*)^T \,\right]$, then we have

$$\begin{aligned} \frac{\|\mathbf{V} - \mathbf{W}^* \mathbf{H}^*\|_{\mathrm{F}}^2}{\|\mathbf{V}\|_{\mathrm{F}}^2} &= \frac{\sum_{k=1}^K \|\mathbf{V}_k - \mathbf{w}_k^* \mathbf{h}_k^T\|_{\mathrm{F}}^2}{\|\mathbf{V}\|_{\mathrm{F}}^2} && (3.3.4) \\ &\leq \frac{\sum_{k=1}^K \|\mathbf{V}_k\|_{\mathrm{F}}^2 \sin^2 \alpha_k}{\sum_{k=1}^K \|\mathbf{V}_k\|_{\mathrm{F}}^2}. && (3.3.5) \end{aligned}$$

Thus we have (3.3.1) as desired. □

In practice, to obtain the tightest possible upper bound for (3.3.1), we need to solve the following optimization problem:

$$\min \max_{k \in [K]} \alpha(\mathbf{V}_k), \tag{3.3.6}$$

where $\alpha(\mathbf{V}_k)$ represents the smallest possible size angle corresponding to $\mathbf{V}_k$ (defined in (3.3.11)) and the minimization is taken over all possible clusterings of the columns of $\mathbf{V}$. We consider finding an optimal size angle and a corresponding basis vector for any data matrix, which we hereby write as $\mathbf{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_M] \in \mathbb{R}_+^{F \times M}$ where $M \in \mathbb{N}_+$. This is solved by the following optimization problem:

$$\begin{aligned} \text{minimize}_{\alpha, \mathbf{u}} \quad & \alpha \\ \text{subject to} \quad & \mathbf{x}_m^T \mathbf{u} \geq \cos \alpha, \quad m \in [M], \\ & \mathbf{u} \geq 0, \quad \|\mathbf{u}\|_2 = 1, \quad \alpha \geq 0, \end{aligned} \tag{3.3.7}$$

where $\mathbf{u} \geq 0$ denotes element-wise nonnegativity and the decision variables are $(\alpha, \mathbf{u})$. Alternatively, consider

$$\begin{aligned} \text{maximize}_{\alpha,\mathbf{u}} \quad & \cos\alpha \\ \text{subject to} \quad & \mathbf{x}_m^T\mathbf{u} \geq \cos\alpha, \quad m \in [M], \\ & \mathbf{u} \geq 0, \quad \|\mathbf{u}\|_2 = 1. \end{aligned} \tag{3.3.8}$$

Similar to the primal optimization problem for linearly separable support vector machines [81], we can obtain the optimal $\mathbf{u}$ and α for (3.3.8) by solving

$$\begin{aligned} \text{minimize}_{\mathbf{u}} \quad & \frac{1}{2}\|\mathbf{u}\|_2^2 \\ \text{subject to} \quad & \mathbf{x}_m^T\mathbf{u} \geq 1, \quad m \in [M], \quad \mathbf{u} \geq 0, \end{aligned} \tag{3.3.9}$$

where the decision variable here is only $\mathbf{u}$. Note that (3.3.9) is a quadratic programming problem and can be easily solved by standard convex optimization software. Suppose $\hat{\mathbf{u}}$ is the optimal solution of (3.3.9), then $\mathbf{u}^* := \hat{\mathbf{u}}/\|\hat{\mathbf{u}}\|_2$ and $\alpha^* := \arccos\left(1/\|\hat{\mathbf{u}}\|_2\right)$ is the optimal basis vector and size angle.

We now state and prove a relative error bound of the proposed approximate NMF algorithm detailed in Algorithm 2 under our geometric assumption. We see that if the size angles of all circular cones are small compared to the angle between the basis vectors of any two circular cones, then exact clustering is possible, and thus the relative error of the best approximate NMF of an arbitrary nonnegative matrix generated from these circular cones can be appropriately controlled by these size angles. Note that first factor of the SVD can be computed for example with the power method [62]. Recall that as mentioned in Section 3.2, Theorem 6 is similar to the corresponding theorem for the near-separable case in [34] in terms of the geometric condition imposed.

Theorem 6. *Under the geometric assumption given in Section 3.2 for generating* $\mathbf{V} \in \mathbb{R}_+^{F\times N}$*, Algorithm 2 outputs* $\mathbf{W}^* \in \mathbb{R}_+^{F\times K}$*,* $\mathbf{H}^* \in \mathbb{R}_+^{K\times N}$*, such that*

$$\frac{\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_{\mathrm{F}}}{\|\mathbf{V}\|_{\mathrm{F}}} \leq \max_{k\in[K]}\{\sin\alpha_k\}. \tag{3.3.10}$$

Algorithm 2 Clustering and Rank One NMF (`cr1-nmf`)

```
Input: Data matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$, $K \in \mathbb{N}$
Output: Factor matrices $\mathbf{W}^* \in \mathbb{R}_+^{F \times K}$, $\mathbf{H}^* \in \mathbb{R}_+^{K \times N}$
1) Use Algorithm 1 to find a set of non-empty, pairwise disjoint index sets $\mathscr{I}_1, \mathscr{I}_2, \ldots, \mathscr{I}_K \subseteq [N]$.
2) for $k = 1$ to $K$ do
       $\mathbf{V}_k := \mathbf{V}(:, \mathscr{I}_k);$                                              (3.3.11)
       $[\mathbf{U}_k, \mathbf{\Sigma}_k, \mathbf{X}_k] := \mathrm{svd}(\mathbf{V}_k);$                          (3.3.12)
       $\mathbf{w}_k^* := |\mathbf{U}_k(:,1)|, \quad \mathbf{h}_k := \mathbf{\Sigma}_k(1,1)|\mathbf{X}_k(:,1)|;$      (3.3.13)
       $\mathbf{h}_k^* := \mathrm{zeros}(1,N), \mathbf{h}_k^*(\mathscr{I}_k) = \mathbf{h}_k.$                (3.3.14)
3) $\mathbf{W}^* := [\mathbf{w}_1^*, \ldots, \mathbf{w}_K^*]$, $\mathbf{H}^* := [(\mathbf{h}_1^*)^T; \ldots; (\mathbf{h}_K^*)^T]$.
```

Proof. From Lemma 15, under the geometric assumption in Section 3.2, we can obtain a set of non-empty, pairwise disjoint index sets $\mathscr{I}_1, \mathscr{I}_2, \ldots, \mathscr{I}_K \subseteq [N]$ such that their union is $[N]$ and two data points $\mathbf{V}(:, j_1)$ and $\mathbf{V}(:, j_2)$ are in the same circular cones if and only if j_1 and j_2 are in the same index set. Then from Theorem 5, we can obtain $\mathbf{W}^*$ and $\mathbf{H}^*$ with the same upper bound on the relative error. □

In addition, for any fixed $K < \min\{F, N\}$, we can use K truncated circular cones with same size angle to cover nonnegative unit sphere, or equivalently, cover the nonnegative orthant $\mathcal{P}$. Then by applying Theorem 5, we can obtain a general upper bound for relative error.

Theorem 7. *Given a data matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ and $K \in \mathbb{N}$ with $F \leq N$, $K < F$ (if $F > N$, we consider the transpose of $\mathbf{V}$). If we define the minimum of the relative error of the order-K NMF of $\mathbf{V}$ to be*

$$J(\mathbf{V}, K) := \frac{\min_{\mathbf{W} \in \mathbb{R}_+^{F \times K}, \mathbf{H} \in \mathbb{R}_+^{K \times N}} \|\mathbf{V} - \mathbf{WH}\|_F}{\|\mathbf{V}\|_F}, \tag{3.3.15}$$

we have the bound for $J(\mathbf{V}, K)$

$$J(\mathbf{V}, K) \leq \frac{\sqrt{F - 3 + 4\sin^2 \dfrac{\pi}{4K} / \sin^2 \dfrac{\pi}{2K}}}{\sqrt{F - 2 + 4\sin^2 \dfrac{\pi}{4K} / \sin^2 \dfrac{\pi}{2K}}}. \tag{3.3.16}$$

When $K = 1$, we can take the circular cone $\mathcal{C}(\mathbf{u}, \alpha)$ with $\mathbf{u} = \mathbf{e}/\sqrt{F}$ and $\alpha = \arccos 1/\sqrt{F}$ to cover the nonnegative orthant, where $\mathbf{e}$ is the vector with all 1's. $\sin\alpha = \sqrt{F-1}/\sqrt{F}$, and this coincides with (3.3.16) for $K = 1$. From the best rank-one approximation error in Frobenius norm of SVD, it is easy to see the equality can be achieved by taking identity matrix, i.e., $\mathbf{V} = \mathbf{I}_F$.

3.4 Relative Error Bounds for Probabilistic Data

We now provide a tighter relative error bound by assuming a probabilistic model. For simplicity, we assume a straightforward and easy-to-implement statistical model for the sampling procedure. We first present the proof of the tighter relative error bound corresponding to the probabilistic model in Theorem 8 to follow, then we show that the upper bound for relative error is tight if we assume all the circular cones are contained in nonnegative orthant in Theorem 9.

We assume the following generating process for each column $\mathbf{v}$ of $\mathbf{V}$ in Theorem 8 to follow.

1. sample $k \in [K]$ with equal probability $1/K$;
2. sample the squared length l from the exponential distribution[1] $\mathrm{Exp}(\lambda_k)$ with parameter (inverse of the expectation) λ_k;
3. uniformly sample a unit vector $\mathbf{z} \in C_k$ w.r.t. the angle between $\mathbf{z}$ and $\mathbf{u}_k$;[2]
4. if $\mathbf{z} \notin \mathbb{R}_+^F$, set all negative entries of $\mathbf{z}$ to zero, and rescale $\mathbf{z}$ to be a unit vector;

[1] $\mathrm{Exp}(\lambda)$ is the function $x \mapsto \lambda \exp(-\lambda x) 1\{x \geq 0\}$.

[2] This means we first uniformly sample an angle $\beta \in [0, \alpha_k]$ and subsequently uniformly sample a vector $\mathbf{z}$ from the set $\{\mathbf{x} \in \mathbb{R}^F : \|\mathbf{x}\|_2 = 1, \mathbf{x}^T \mathbf{u}_k = \cos\beta\}$

5. let $\mathbf{v} = \sqrt{l}\mathbf{z}$;

Theorem 8. *Suppose the K truncated circular cones $C_k \cap \mathbb{R}_+^F$ with $C_k := \mathcal{C}(\mathbf{u}_k, \alpha_k) \in \mathbb{R}^F$ for $k \in [K]$ satisfy the geometric assumption given by (3.2.2). Let $\boldsymbol{\lambda} := (\lambda_1; \lambda_2; \dots; \lambda_K) \in \mathbb{R}_{++}^K$. We generate the columns of a data matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ from the above generative process. Let $f(\alpha) := \frac{1}{2} - \frac{\sin 2\alpha}{4\alpha}$, then for a small $\epsilon > 0$, with probability at least $1 - 8\exp(-\xi N \epsilon^2)$, one has*

$$\frac{\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_{\mathrm{F}}}{\|\mathbf{V}\|_{\mathrm{F}}} \leq \sqrt{\frac{\sum_{k=1}^K f(\alpha_k)/\lambda_k}{\sum_{k=1}^K 1/\lambda_k}} + \epsilon, \tag{3.4.1}$$

where the constant $\xi > 0$ depends only on λ_k and $f(\alpha_k)$ for all $k \in [K]$.

Remark 8. *The assumption in Step 1 in the generating process that the data points are generated from K circular cones with equal probability can be easily generalized to unequal probabilities. The assumption in Step 2 that the square of the length of a data point is sampled from an exponential distribution can be easily extended any non-negative sub-Exponential distribution (cf. Lemma 6), or equivalently, the length of a data point is sampled from a nonnegative sub-Gaussian distribution (cf. Lemma 5 in Section 2.2.1).*

The relative error bound produced by Theorem 8 is better than that of Theorem 6, i.e., the former is more conservative. This can be seen from (3.4.3) to follow, or from the inequality $\alpha \leq \tan \alpha$ for $\alpha \in [0, \pi/2)$. We also observe this in the experiments in Section 3.6.1.

Theorem 8 is proved by combining the large deviation bound in Lemma 6 with the deterministic bound on the relative error in Theorem 6.

Proof of Theorem 8. From (3.3.2) and (3.3.3) in the proof of Theorem 6, to obtain an upper bound for the square of the relative error, we consider the following random variable

$$D_N := \frac{\sum_{n=1}^N L_n^2 \sin^2 B_n}{\sum_{n=1}^N L_n^2}, \tag{3.4.2}$$

where L_n is the random variable corresponding to the length of the n-th point, and B_n is the random variable corresponding to the angle between the n-th point and $\mathbf{u}_k$ for some $k \in [K]$ such that the point is in $C_k \cap \mathbb{R}_+^F$. We first consider estimating the above random variable with the assumption that all the data points are generated from a single truncated circular cone $C \cap \mathbb{R}_+^F$ with $C := \mathcal{C}(\mathbf{u}, \alpha)$ (i.e., assume $K = 1$), and the square of lengths are generated according to the exponential distribution $\mathrm{Exp}(\lambda)$. Because we assume each angle β_n for $n \in [N]$ is sampled from a uniform distribution on $[0, \alpha]$, the expectation of $\sin^2 B_n$ is

$$\mathbb{E}\left[\sin^2 B_n\right] = \int_0^\alpha \frac{1}{\alpha} \sin^2 \beta \,\mathrm{d}\beta = \frac{1}{2} - \frac{\sin 2\alpha}{4\alpha} = f(\alpha). \tag{3.4.3}$$

Here we only need to consider vectors $\mathbf{z} \in \mathbb{R}_+^F$ whose angles with $\mathbf{u}$ are not larger than α. Otherwise, we have $\mathbb{E}[\sin^2 B_n] \le f(\alpha)$. Our probabilistic upper bound also holds in this case.

Since the length and the angle are independent, we have

$$\mathbb{E}\left[D_N\right] = \mathbb{E}\left[\mathbb{E}\left[D_N | L_1, \ldots, L_N\right]\right] = f(\alpha), \tag{3.4.4}$$

and we also have

$$\mathbb{E}\left[L_n^2 \sin^2 B_n\right] = \mathbb{E}\left[L_n^2\right] \mathbb{E}\left[\sin^2 B_n\right] = \frac{f(\alpha)}{\lambda}. \tag{3.4.5}$$

Define $X_n := L_n^2$ for all $n \in [N]$. Let

$$H_N := \frac{\sum_{n=1}^N X_n}{N}, \quad \text{and} \quad G_N := \frac{\sum_{n=1}^N X_n \sin^2 B_n}{N}. \tag{3.4.6}$$

We have for all $n \in [N]$,

$$\mathbb{E}[X_n^p] = \lambda^{-p}\Gamma(p+1) \le \lambda^{-p} p^p, \qquad \forall\, p \ge 1, \tag{3.4.7}$$

where $\Gamma(\cdot)$ is the gamma function. Thus $\|X_n\|_{\Psi_1} \le \lambda^{-1}$, and X_n is sub-Exponential. By the triangle inequality, we have $\|X_n - \mathbb{E}X_n\|_{\Psi_1} \le \|X_n\|_{\Psi_1} + \|\mathbb{E}X_n\|_{\Psi_1} \le 2\|X_n\|_{\Psi_1}$. Hence, by Lemma 6, for all $\epsilon > 0$, we have (2.2.14) where M can be taken as $M = 2/\lambda$. Because

$$\left(\mathbb{E}\left[\left(X_n \sin^2 B_n\right)^p\right]\right)^{1/p} \le \lambda^{-1} p \sin^2 \alpha \le \lambda^{-1} p, \tag{3.4.8}$$

we have a similar large deviation result for G_N.

On the other hand, for all $\epsilon > 0$

$$\mathbb{P}\left(|D_N - f(\alpha)| \geq \epsilon\right) = \mathbb{P}\left(\left|\frac{G_N}{H_N} - f(\alpha)\right| \geq \epsilon\right) \tag{3.4.9}$$

$$\leq \mathbb{P}\left(|\lambda G_N - f(\alpha)| \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(\left|\frac{G_N}{H_N} - \lambda G_N\right| \geq \frac{\epsilon}{2}\right). \tag{3.4.10}$$

For the second term, by fixing small constants $\delta_1, \delta_2 > 0$, we have

$$\mathbb{P}\left(\left|\frac{G_N}{H_N} - \lambda G_N\right| \geq \frac{\epsilon}{2}\right) = \mathbb{P}\left(\frac{|1-\lambda H_N| G_N}{H_N} \geq \frac{\epsilon}{2}\right) \tag{3.4.11}$$

$$\leq \mathbb{P}\left(\frac{|1-\lambda H_N| G_N}{H_N} \geq \frac{\epsilon}{2}, H_N \geq \frac{1}{\lambda} - \delta_1, G_N \leq \frac{f(\alpha)}{\lambda} + \delta_2\right)$$

$$+ \mathbb{P}\left(H_N < \frac{1}{\lambda} - \delta_1\right) + \mathbb{P}\left(G_N > \frac{f(\alpha)}{\lambda} + \delta_2\right). \tag{3.4.12}$$

Combining the large deviation bounds for H_N and G_N in (2.2.14) with the above inequalities, if we set $\delta_1 = \delta_2 = \epsilon$ and take ϵ sufficiently small,

$$\mathbb{P}\left(|D_N - f(\alpha)| \geq \epsilon\right) \leq 8 \exp\left(-\xi N \epsilon^2\right), \tag{3.4.13}$$

where ξ is a positive constant depending on λ and $f(\alpha)$.

Now we turn to the general case in which $K \in \mathbb{N}$. We have

$$\mathbb{E}\left[X_n\right] = \frac{\sum_{k=1}^K 1/\lambda_k}{K}, \quad \text{and} \tag{3.4.14}$$

$$\mathbb{E}\left[X_n \sin^2 B_n\right] = \frac{\sum_{k=1}^K f(\alpha_k)/\lambda_k}{K}, \tag{3.4.15}$$

and for all $p \geq 1$,

$$\left(\mathbb{E}[X_n^p]\right)^{1/p} = \left(\frac{\sum_{k=1}^K \lambda_k^{-p} \Gamma(p+1)}{K}\right)^{1/p} \leq \frac{p}{\min_k \lambda_k}. \tag{3.4.16}$$

Similar to (3.4.13), we have

$$\mathbb{P}\left(\left|D_N - \frac{\sum_{k=1}^K f(\alpha_k/\lambda_k)}{\sum_{k=1}^K 1/\lambda_k}\right| \geq \epsilon\right) \leq 8 \exp\left(-\xi N \epsilon^2\right), \tag{3.4.17}$$

and thus, if we let $\Delta := \sqrt{\frac{\sum_{k=1}^K f(\alpha_k)/\lambda_k}{\sum_{k=1}^K 1/\lambda_k}}$, we have

$$\mathbb{P}\left(\left|\sqrt{D_N} - \Delta\right| \leq \epsilon\right) \geq \mathbb{P}\left(\left|D_N - \Delta^2\right| \leq \Delta\epsilon\right) \tag{3.4.18}$$

$$\geq 1 - 8\exp\left(-\xi N \Delta^2 \epsilon^2\right). \tag{3.4.19}$$

This completes the proof of (3.4.1). □

Furthermore, if the K circular cones $\mathcal{C}_1, \dots, \mathcal{C}_K$ are contained in the nonnegative orthant $\mathbb{R}_+^F$, we do not need to project the data points not in $\mathbb{R}_+^F$ onto $\mathbb{R}_+^F$. Then we can prove that the upper bound in Theorem 8 is asymptotically tight, i.e.,

$$\frac{\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_{\mathrm{F}}}{\|\mathbf{V}\|_{\mathrm{F}}} \xrightarrow{\mathrm{P}} \sqrt{\frac{\sum_{k=1}^K f(\alpha_k)/\lambda_k}{\sum_{k=1}^K 1/\lambda_k}}, \text{ as } N \to \infty. \tag{3.4.20}$$

Theorem 9. *Suppose the data points of $\mathbf{V} \in \mathbb{R}_+^{F\times N}$ are generated as given in Theorem 8 with all the circular cones being contained in the nonnegtive orthant, then Algorithm 2 produces $\mathbf{W}^* \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H}^* \in \mathbb{R}_+^{K\times N}$ with the property that for any $\epsilon \in (0,1)$ and $t \geq 1$, if $N \geq c(t/\epsilon)^2 F$, then with probability at least $1 - 6K\exp(-t^2 F)$ one has*

$$\left| \frac{\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_{\mathrm{F}}}{\|\mathbf{V}\|_{\mathrm{F}}} - \sqrt{\frac{\sum_{k=1}^K f(\alpha_k)/\lambda_k}{\sum_{k=1}^K 1/\lambda_k}} \right| \leq c\epsilon \tag{3.4.21}$$

where c is a constant depending on K and α_k, λ_k for $k \in [K]$.

To prove Theorem 9, we first provide a few definitions and lemmas. Consider the following condition that ensures that the circular cone $C(\mathbf{u}, \alpha)$ is entirely contained in the non-negative orthant $\mathbb{R}_+^F$.

Lemma 18. *If $\mathbf{u} = (u(1), u(2), \dots, u(F))$ is a positive unit vector and $\alpha > 0$ satisfies*

$$\alpha \leq \arccos \sqrt{1 - u_{\min}^2}, \tag{3.4.22}$$

where $u_{\min} := \min_f u(f)$, then $\mathcal{C}(\mathbf{u}, \alpha) \subseteq \mathbb{R}_+^F$.

Proof of Lemma 18. Because any nonnegative vector $\mathbf{x}$ is spanned by basis vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_F$, given a positive unit vector $\mathbf{u}$, to find the largest size angle, we only need to consider the angle between $\mathbf{u}$ and $\mathbf{e}_f$, $f \in [F]$. Take any $f \in [F]$, if the angle β between $\mathbf{u}$ and $\mathbf{e}_f$ is not larger than $\pi/4$, we can obtain the unit vector symmetric to $\mathbf{e}_f$ w.r.t. $\mathbf{u}$ in the plane spanned by $\mathbf{u}$ and $\mathbf{e}_f$ is also nonnegative. In fact, the vector is $2(\cos\beta)\mathbf{u} - \mathbf{e}_f$. Because $u(f) = \cos\beta$ and $\beta \leq \pi/4$, we have $2\cos^2\beta \geq 1$

and the vector is nonnegative. If $\beta > \pi/4$, i.e., $u(f) < 1/\sqrt{2}$, we can take the extreme nonnegative unit vector $\mathbf{z}$ in the span of $\mathbf{u}$ and $\mathbf{e}_f$, i.e.,

$$\mathbf{z} = \frac{\mathbf{u} - u(f)\mathbf{e}_f}{\|\mathbf{u} - u(f)\mathbf{e}_f\|_2}, \tag{3.4.23}$$

and it is easy to see $\mathbf{u}^T\mathbf{z} = \sqrt{1-u(f)^2}$. Hence the angle between $\mathbf{z}$ and $\mathbf{u}$ is $\pi/2 - \beta < \pi/4$. Therefore, the largest size angle $\alpha_{\mathbf{e}_f}$ w.r.t. $\mathbf{e}_f$ is

$$\alpha_{\mathbf{e}_f} := \begin{cases} \arccos u(f), & \text{if } u(f) \geq 1/\sqrt{2} \\ \arccos \sqrt{1-u(f)^2}, & \text{if } u(f) < 1/\sqrt{2} \end{cases} \tag{3.4.24}$$

or equivalently, $\alpha_{\mathbf{e}_f} = \min\{\arccos u(f), \arccos\sqrt{1-u(f)^2}\}$. Thus, the largest size angle corresponding to $\mathbf{u}$ is

$$\min_f \left\{ \min\{\arccos u(f), \arccos\sqrt{1-u(f)^2}\}\right\} \tag{3.4.25}$$

Let $u_{\max} := \max_f u(f)$ and $u_{\min} := \min_f u(f)$. Then the largest size angle corresponding to $\mathbf{u}$ is

$$\min\left\{\arccos u_{\max}, \arccos\sqrt{1-u_{\min}^2}\right\}. \tag{3.4.26}$$

Because $u_{\max}^2 + u_{\min}^2 \leq 1$ for $F > 1$, the expression in (3.4.26) equals $\arccos\sqrt{1-u_{\min}^2}$ and this completes the proof. □

Lemma 19. *Define $f(\beta) := \frac{1}{2} - \frac{\sin(2\beta)}{4\beta}$ and $g(\beta) := \frac{1}{2} + \frac{\sin(2\beta)}{4\beta}$ for $\beta \in \left(0, \frac{\pi}{2}\right]$. Let $\mathbf{e}_f$, $f \in [F]$ be the unit vector with only the f-th entry being 1, and C be the circular cone with basis vector $\mathbf{u} = \mathbf{e}_f$, size angle being α, and the inverse expectation parameter for the exponential distribution being λ. Then if the columns of the data matrix $\mathbf{V} \in \mathbb{R}^{F\times N}$ are generated as in Theorem 8 from C ($K = 1$) and with no projection to the nonnegative orthant (Step 4 in the generating process), we have*

$$\mathbb{E}\left(\frac{\mathbf{V}\mathbf{V}^T}{N}\right) = \frac{\mathbf{D}_f}{\lambda} \tag{3.4.27}$$

where $\mathbf{D}_f$ is a diagonal matrix with the f-th diagonal entry being $g(\alpha)$ and other diagonal entries being $f(\alpha)/(F-1)$.

Proof of Lemma 19. Each column $\mathbf{v}_n$, $n \in [N]$ can be generated as follows: First, uniformly sample a $\beta_n \in [0, \alpha]$ and sample a positive scalar l_n from the exponential distribution $\mathrm{Exp}(\lambda)$, then we can write $\mathbf{v}_n = \sqrt{l_n}\,[\cos\beta_n \mathbf{e}_f + \sin\beta_n \mathbf{y}_n]$, where $\mathbf{y}_n$ can be generated from sampling $y_n(1), \dots, y_n(f-1), y_n(f+1), \dots, y_n(F)$ from the standard normal distribution $\mathcal{N}(0,1)$, and setting $y_n(j) = y_n(j)/\sqrt{\sum_{i\neq f} y_n(i)^2}$, $j \neq f$, $y_n(f) = 0$. Then

$$
\begin{aligned}
&\mathbb{E}\left[v_n(f_1)v_n(f_2)\right] \\
&= \mathbb{E}\big[l_n((\cos^2\beta)e_f(f_1)e_f(f_2) + (\sin^2\beta)y_n(f_1)y_n(f_2))\big] && (3.4.28)\\
&= \begin{cases} 0, & f_1 \neq f_2, \\ g(\alpha)/\lambda, & f_1 = f_2 = f, \quad , \\ f(\alpha)/\left((F-1)\lambda\right), & f_1 = f_2 \neq f. \end{cases} && (3.4.29)
\end{aligned}
$$

where $e_f(f_1) = 1\{f = f_1\}$ is the f_1-th entry of the vector $\mathbf{e}_f$. Thus $\mathbb{E}\left(\mathbf{V}\mathbf{V}^T/N\right) = \mathbb{E}\left(\mathbf{v}_n\mathbf{v}_n^T\right) = \mathbf{D}_f/\lambda$. □

Proof of Theorem 9. Similar to Theorem 6, we have

$$
\begin{aligned}
\frac{\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_{\mathrm{F}}^2}{\|\mathbf{V}\|_{\mathrm{F}}^2} &= \frac{\sum_{k=1}^K \|\mathbf{V}_k - \mathbf{w}_k^*\mathbf{h}_k^T\|_{\mathrm{F}}^2}{\sum_{k=1}^K \|\mathbf{V}_k\|_F^2} && (3.4.30)\\
&= \frac{\sum_{k=1}^K \left(\|\mathbf{V}_k\|_F^2 - \sigma_1^2\left(\mathbf{V}_k\right)\right)}{\sum_{k=1}^K \|\mathbf{V}_k\|_F^2} && (3.4.31)\\
&= 1 - \frac{\sum_{k=1}^K \sigma_1^2\left(\mathbf{V}_k\right)}{\sum_{k=1}^K \|\mathbf{V}_k\|_{\mathrm{F}}^2}. && (3.4.32)
\end{aligned}
$$

Take any $k \in [K]$ and consider $\sigma_1^2\left(\mathbf{V}_k\right)$. Define the index $f_k := \mathrm{argmin}_{f\in[F]}\mathbf{u}_k$ and the orthogonal matrix $\mathbf{P}_k$ as in (3.6.1). The columns of $\mathbf{V}_k$ can be considered as Householder transformations of the data points generated from the circular cone $C_{f_k}^0 := \mathcal{C}\left(\mathbf{e}_{f_k}, \alpha_k\right)$ (the circular cone with basis vector $\mathbf{e}_{f_k}$ and size angle α_k), i.e., $\mathbf{V}_k = \mathbf{P}_k\mathbf{X}_k$, where $\mathbf{X}_k$ contains the corresponding data points in $C_{f_k}^0$. In addition, denoting N_k as the number of data points in $\mathbf{V}_k$, we have

$$
\frac{\sigma_1^2\left(\mathbf{V}_k\right)}{N_k} = \frac{\sigma_1^2\left(\mathbf{V}_k^T\right)}{N_k} = \lambda_{\max}\left(\frac{\mathbf{V}_k\mathbf{V}_k^T}{N_k}\right) \qquad (3.4.33)
$$

where $\lambda_{\max}\left(\mathbf{V}_k\mathbf{V}_k^T/N_k\right)$ represents the largest eigenvalue of $\mathbf{V}_k\mathbf{V}_k^T/N_k$. Take any $\mathbf{v} \in \mathbf{V}_k$. Note that $\mathbf{v}$ can be written as $\mathbf{v} = \mathbf{P}_k\mathbf{x}$ with $\mathbf{x}$ being generated from $C_{f_k}^0$. Now, for all unit vectors $\mathbf{z} \in \mathbb{R}^F$, we have

$$\begin{aligned} \|\mathbf{v}\|_{\Psi_2} &= \|\mathbf{P}_k\mathbf{x}\|_{\Psi_2} = \|\mathbf{x}\|_{\Psi_2} && (3.4.34)\\ &= \sup_{\|\mathbf{z}\|_2=1}\sup_{p\geq 1} p^{-1/2}\left(\mathbb{E}\left(|\mathbf{x}^T\mathbf{z}|^p\right)\right)^{1/p} && (3.4.35)\\ &\leq \sup_{p\geq 1} p^{-1/2}\mathbb{E}\left(\|\mathbf{x}\|_2^p\right)^{1/p} && (3.4.36)\\ &= \|\|\mathbf{x}\|_2\|_{\Psi_2} \leq \sqrt{\|\|\mathbf{x}\|_2^2\|_{\Psi_1}} \leq 1/\sqrt{\lambda_k}. && (3.4.37) \end{aligned}$$

That is, all columns are sampled from a sub-Gaussian distribution. By Lemma 19,

$$\mathbb{E}\left(\mathbf{v}\mathbf{v}^T\right) = \mathbb{E}\left(\mathbf{P}_k\mathbf{x}\mathbf{x}^T\mathbf{P}_k^T\right) = \mathbf{P}_k\mathbf{D}_{f_k}\mathbf{P}_k^T/\lambda_k. \tag{3.4.38}$$

By Lemma 8, we have for $\epsilon \in (0,1), t \geq 1$ and if $N_k \geq \xi_k(t/\epsilon)^2F$ (ξ_k is a positive constant depending on λ_k), with probability at least $1-2\exp(-t^2F)$,

$$\begin{aligned} &\left|\lambda_{\max}\left(\mathbf{V}_k\mathbf{V}_k^T/N_k\right) - \lambda_{\max}\left(\mathbb{E}\left(\mathbf{v}\mathbf{v}^T\right)\right)\right| \\ &\quad \leq \|\mathbf{V}_k\mathbf{V}_k^T/N_k - \mathbb{E}\left(\mathbf{v}\mathbf{v}^T\right)\|_2 \leq \epsilon, && (3.4.39) \end{aligned}$$

where the first inequality follows from Lemma 16. Because $\lambda_{\max}\left(\mathbb{E}\left(\mathbf{v}\mathbf{v}^T\right)\right) = g(\alpha_k)/\lambda_k$, we can obtain that with probability at least $1-4K\exp(-t^2F)$,

$$\begin{aligned} &\left|\sum_{k=1}^K \frac{\sigma_1^2\left(\mathbf{V}_k\right)}{N} - \sum_{k=1}^K \frac{g(\alpha_k)}{K\lambda_k}\right| \\ &= \left|\sum_{k=1}^K \lambda_{\max}\left(\frac{\mathbf{V}_k\mathbf{V}_k^T}{N_k}\right)\frac{N_k}{N} - \sum_{k=1}^K \frac{g(\alpha_k)}{K\lambda_k}\right| && (3.4.40)\\ &\leq 2K\epsilon, && (3.4.41) \end{aligned}$$

where the final inequality follows from the triangle inequality and (3.4.39). From the proof of Theorem 8, we know that with probability at least $1-2\exp(-c_1N\epsilon^2)$,

$$\left|\frac{\|\mathbf{V}\|_{\mathrm{F}}^2}{N} - \frac{\sum_{k=1}^K 1/\lambda_k}{K}\right| \leq \epsilon. \tag{3.4.42}$$

Taking N to be sufficiently large such that $t^2F \le c_1 N\epsilon^2$, we have with probability at least $1 - 6K\exp(-t^2F)$,

$$\frac{\sum_{k=1}^K g(\alpha_k)/\lambda_k}{\sum_{k=1}^K 1/\lambda_k} - c_2\epsilon \le \frac{\sum_{k=1}^K \sigma_1^2(\mathbf{V}_k)}{\sum_{k=1}^K \|\mathbf{V}_k\|_{\mathrm{F}}^2} \tag{3.4.43}$$

$$\le \frac{\sum_{k=1}^K g(\alpha_k)/\lambda_k}{\sum_{k=1}^K 1/\lambda_k} + c_3\epsilon. \tag{3.4.44}$$

Note that $g(\alpha_k) + f(\alpha_k) = 1$. As a result, we have

$$\frac{\sum_{k=1}^K f(\alpha_k)/\lambda_k}{\sum_{k=1}^K 1/\lambda_k} - c_3\epsilon \le \frac{\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_{\mathrm{F}}^2}{\|\mathbf{V}\|_{\mathrm{F}}^2} \tag{3.4.45}$$

$$\le \frac{\sum_{k=1}^K f(\alpha_k)/\lambda_k}{\sum_{k=1}^K 1/\lambda_k} + c_2\epsilon. \tag{3.4.46}$$

Thus, with probability at least $1 - 6K\exp(-t^2F)$, we have

$$\left| \frac{\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_{\mathrm{F}}}{\|\mathbf{V}\|_{\mathrm{F}}} - \sqrt{\frac{\sum_{k=1}^K f(\alpha_k)/\lambda_k}{\sum_{k=1}^K 1/\lambda_k}} \right| \le c_4\epsilon, \tag{3.4.47}$$

where c_4 depends on K and $\{(\alpha_k, \lambda_k) : k \in [K]\}$. □

3.5 Automatically Determining the Latent Dimension

Automatically determining the latent dimensionality K is an important problem in NMF. Unfortunately, the usual and popular approach for determining the latent dimensionality of nonnegative data matrices based on Bayesian automatic relevance determination by Tan and Févotte [82] does not work well for data matrices generated under the geometric assumption given in Section 3.2. This is because in [82], $\mathbf{W}$ and $\mathbf{H}$ are assumed to be generated from the *same* distribution. Under the geometric assumption, $\mathbf{V}$ has well clustered columns and the corresponding coefficient matrix $\mathbf{H}$ can be approximated by a clustering membership indicator matrix with columns that are 1-sparse (i.e., only contains one non-zero entry). Thus, $\mathbf{W}$ and $\mathbf{H}$ have very different statistics. While there are many approaches [83–85] to learn the number of clusters in clustering problems, most methods lack strong theoretical guarantees.

By assuming the generative procedure for $\mathbf{V}$ proposed in Theorem 8, we consider a simple approach for determining K based on the maximum of the ratios between adjacent singular values. We provide a theoretical result for the correctness of this approach. Our method consists in estimating the correct number of circular cones $\hat{K}$ as follows:

$$\hat{K} := \underset{k\in\{K_{\min},\ldots,K_{\max}\}}{\arg\max} \frac{\sigma_k(\mathbf{V})}{\sigma_{k+1}(\mathbf{V})}. \tag{3.5.1}$$

Here $K_{\min} > 1$ and $K_{\max} < \mathrm{rank}(\mathbf{V})$ are selected based on domain knowledge. The main ideas that underpin (3.5.1) are (i) the approximation error for the best rank-k approximation of a data matrix in the Frobenius norm and (ii) the so-called elbow method [86] for determining the number of clusters. More precisely, let $\mathbf{V}_k$ be the best rank-k approximation of $\mathbf{V}$. Then $\|\mathbf{V}-\mathbf{V}_k\|_{\mathrm{F}}^2 = \sum_{j=k+1}^{r} \sigma_j^2(\mathbf{V})$, where r is the rank of $\mathbf{V}$. If we increase k to $k+1$, the square of the best approximation error decreases by $\sigma_{k+1}^2(\mathbf{V})$. The elbow method chooses a number of clusters k so that the decrease in the objective function value from k clusters to $k+1$ clusters is small compared to the decrease in the objective function value from $k-1$ clusters to k clusters. Although this approach seems to be simplistic, interestingly, the following theorem tells that under appropriate assumptions, we can correctly find the number of circular cones with high probability.

Theorem 10. *Suppose that the data matrix $\mathbf{V} \in \mathbb{R}_+^{F\times N}$ is generated according to the generative process given in Theorem 8 where K is the true number of circular cones. Further assume that the size angles for K circular cones are all equal to α, the angles between distinct basis vectors of the circular cones are all equal to β, and the parameters (inverse expectations) for the exponential distributions are all equal to λ. In addition, we assume all the circular cones are contained in the nonnegative orthant $\mathbb{R}_+^F$ (cf. Theorem 9) and $K \in \{K_{\min},\ldots,K_{\max}\}$ with $K_{\min} > 1$ and $K_{\max} < \mathrm{rank}(\mathbf{V})$. Then, for any $t \geq 1$, and sufficiently small ϵ satisfying*

(3.5.17) *in the proof, if $N \geq c(t/\epsilon)^2 F$ (for a constant $c > 0$ depending only on λ, α and β), with probability at least $1 - 2\left(K_{\max} - K_{\min} + 1\right)\exp\left(-t^2 F\right)$,*

$$\frac{\sigma_K(\mathbf{V})}{\sigma_{K+1}(\mathbf{V})} = \max_{j \in \{K_{\min},\dots,K_{\max}\}} \frac{\sigma_j(\mathbf{V})}{\sigma_{j+1}(\mathbf{V})}. \tag{3.5.2}$$

Before proving Theorem 10, we first state and prove the following lemma.

Lemma 20. *Suppose data matrix $\mathbf{V}$ is generated as in Theorem 8 with all the circular cones being contained in $\mathbb{R}_+^F$, then the expectation of the covariance matrix $\mathbf{v}_1\mathbf{v}_1^T$ is*

$$\begin{aligned}\mathbb{E}\left[\mathbf{v}_1\mathbf{v}_1^T\right] &= \frac{\sum_{k=1}^K f(\alpha_k)/\lambda_k}{K(F-1)}\mathbf{I} \\ &+ \frac{1}{K}\sum_{k=1}^K \frac{g(\alpha_k) - f(\alpha_k)/(F-1)}{\lambda_k}\mathbf{u}_k\mathbf{u}_k^T,\end{aligned} \tag{3.5.3}$$

where $\mathbf{v}_1$ denotes the first column of $\mathbf{V}$.

Proof. From the proof in Lemma 19, we know if we always take $\mathbf{e}_1$ to be the original vector for the Householder transformation, the corresponding Householder matrix for the k-th circular cone $\mathcal{C}_k$ is given by (3.6.1) and we have

$$\mathbb{E}\left[\mathbf{v}_1\mathbf{v}_1^T\right] = \frac{1}{K}\sum_{k=1}^K \frac{\mathbf{P}_k\mathbf{D}_k\mathbf{P}_k^T}{\lambda_k}, \tag{3.5.4}$$

where $\mathbf{D}_k$ is a diagonal matrix with the first diagonal entry being $g(\alpha_k) := \frac{1}{2} + \frac{\sin(2\alpha_k)}{4\alpha_k}$ and other diagonal entries are

$$\frac{f(\alpha_k)}{F-1} = \frac{\frac{1}{2} - \frac{\sin(2\alpha_k)}{4\alpha_k}}{F-1}. \tag{3.5.5}$$

We simplify $\mathbf{P}_k\mathbf{D}_k\mathbf{P}_k^T$ using the property that all the $F-1$ diagonal entries of $\mathbf{D}_k$ are the same. Namely, we can write

$$\mathbf{P}_k = \mathbf{I} - 2\mathbf{z}_k\mathbf{z}_k^T = \mathbf{I} - \frac{(\mathbf{e}_1 - \mathbf{u}_k)(\mathbf{e}_1 - \mathbf{u}_k)^T}{1 - u_k(1)} \tag{3.5.6}$$

$$= \begin{bmatrix} u_k(1) & u_k(2) & \cdots & u_k(F) \\ u_k(2) & 1 - \frac{u_k(2)^2}{1-u_k(1)} & \cdots & -\frac{u_k(2)u_k(F)}{1-u_k(1)} \\ \vdots & \vdots & \ddots & \vdots \\ u_k(F) & -\frac{u_k(F)u_k(2)}{1-u_k(1)} & \cdots & 1 - \frac{u_k(F)^2}{1-u_k(1)} \end{bmatrix}. \tag{3.5.7}$$

Note that $\mathbf{P}_k = \left[\mathbf{p}_1^k, \mathbf{p}_2^k, \dots, \mathbf{p}_F^k\right]$ is symmetric and the first column of $\mathbf{P}_k$ is $\mathbf{u}_k$. Let $\mathbf{D}_k$ be the diagonal matrix with diagonal entries being $d_1, d_2, \dots, d_F$. Then we have

$$
\begin{aligned}
\mathbf{P}_k\mathbf{D}_k\mathbf{P}_k^T &= \sum_{j=1}^{K} d_j \mathbf{p}_j^k (\mathbf{p}_j^k)^T && (3.5.8)\\
&= d_1 \mathbf{u}_k\mathbf{u}_k^T + d_2 \sum_{j=2}^{K} \mathbf{p}_j^k (\mathbf{p}_j^k)^T && (3.5.9)\\
&= g\left(\alpha_k\right)\mathbf{u}_k\mathbf{u}_k^T + \frac{f(\alpha_k)}{F-1}\left(\mathbf{I} - \mathbf{u}_k\mathbf{u}_k^T\right) && (3.5.10)\\
&= \frac{f(\alpha_k)}{F-1}\mathbf{I} + \left(g(\alpha_k) - \frac{f(\alpha_k)}{F-1}\right)\mathbf{u}_k\mathbf{u}_k^T. && (3.5.11)
\end{aligned}
$$

Thus, we obtain (3.5.3) as desired. □

We are now ready to prove Theorem 10.

Proof of Theorem 10. Define

$$
\begin{aligned}
a &:= \frac{\sum_{k=1}^K f(\alpha)/\lambda}{K(F-1)} = \frac{f(\alpha)/\lambda}{F-1}, \quad \text{and} && (3.5.12)\\
b &:= \frac{g(\alpha) - f(\alpha)/(F-1)}{K\lambda}. && (3.5.13)
\end{aligned}
$$

By exploiting the assumption that all the α_k's and λ_k's are the same, we find that

$$
\mathbb{E}\left[\mathbf{v}_1\mathbf{v}_1^T\right] = a\mathbf{I} + b\sum_{k=1}^{K}\mathbf{u}_k\mathbf{u}_k^T. \qquad (3.5.14)
$$

Let $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_K]$. We only need to consider the eigenvalues of $\sum_{k=1}^K \mathbf{u}_k\mathbf{u}_k^T = \mathbf{U}\mathbf{U}^T$. The matrix $\mathbf{U}^T\mathbf{U}$ has same non-zero eigenvalues as that of $\mathbf{U}\mathbf{U}^T$. Furthermore,

$$
\begin{aligned}
\mathbf{U}^T\mathbf{U} &= \begin{bmatrix} 1 & \cos\beta & \cdots & \cos\beta \\ \cos\beta & 1 & \cdots & \cos\beta \\ \vdots & \vdots & \ddots & \vdots \\ \cos\beta & \cos\beta & \cdots & 1 \end{bmatrix} && (3.5.15)\\
&= (\cos\beta)\mathbf{e}\mathbf{e}^T + (1-\cos\beta)\mathbf{I} && (3.5.16)
\end{aligned}
$$

where $\mathbf{e} \in \mathbb{R}^K$ is the vector with all entries being 1. Therefore, the eigenvalues of $\mathbf{U}^T\mathbf{U}$ are $1+(K-1)\cos\beta, 1-\cos\beta, \dots, 1-\cos\beta$. Thus, the vector of eigenvalues of $\mathbb{E}\left[\mathbf{v}_1\mathbf{v}_1^T\right]$ is $[a+b(1+(K-1)\cos\beta), a+b(1-\cos\beta), \dots, a+b(1-\cos\beta), a, a, \dots, a]$.

By Lemmas 16 and 8, we deduce that for any $t \geq 1$ and a sufficiently small $\epsilon > 0$, such that

$$\frac{a+\epsilon}{a-\epsilon} < \frac{a+b(1-\cos\beta)-\epsilon}{a+\epsilon}, \tag{3.5.17}$$

then if $N \geq c(t/\epsilon)^2 F$ (where $c > 0$ depends only on λ, α, and β), then with probability at least $1 - 2\left(K_{\max} - K_{\min} + 1\right)\exp\left(-t^2 F\right)$, Eqn. (3.5.2) holds. □

In Section 3.6.1, we show numerically that the proposed method in (3.5.1) works well even when the geometric assumption is only *approximately satisfied* (see Section 3.6.1 for a formal definition) assuming that N is sufficiently large. This shows that the determination of the correct number of clusters is robust to noise.

Remark 9. *The conditions of Theorem 10 may appear to be rather restrictive. However, we make them only for the sake of convenience in presentation. We do not need to assume that the parameters of the exponential distribution are equal if, instead of $\sigma_j(\mathbf{V})$, we consider the singular values of a normalized version of $\mathbf{V}$. The assumptions that all the size angles are the same and the angles between distinct basis vectors are the same can also be relaxed. The theorem continues to hold even when the geometric assumption in (3.2.2) is not satisfied, i.e., $\beta \leq 4\alpha$. However, we empirically observe in Section 3.6.1 that if $\mathbf{V}$ satisfies the geometric assumption (even approximately), the results are superior compared to the scenario when the assumption is significantly violated.*

Remark 10. *We may replace the assumption that the circular cones are contained in the nonnegative orthant by removing Step 4 in the generating process (projection onto $\mathcal{P}$) in the generative procedure in Theorem 8. Because we are concerned with finding the number of clusters (or circular cones) rather than determining the true latent dimensionality of an NMF problem (cf. [82]), we can discard the nonnegativity* *constraint. The number of clusters serves as a proxy for the latent dimensionality of NMF.*

3.6 Numerical Results

3.6.1 Experiments on Synthetic Data

To verify the correctness of our bounds, to observe the computational efficiency of the proposed algorithm, and to check if the procedure for estimating K is effective, we first perform numerical simulations on synthetic datasets. All the experiments were executed on a Windows machine whose processor is an Intel(R) Core(TM) i5-3570, the CPU speed is 3.40 GHz, and the installed memory (RAM) is 8.00 GB. The Matlab version is 7.11.0.584 (R2010b). The Matlab codes for running the experiments can be found at `https://github.com/zhaoqiangliu/cr1-nmf`.

Comparison of Relative Errors and Running Times

To generate the columns of $\mathbf{V}$, given an integer $k \in [K]$ and an angle $\beta \in [0, \alpha_k]$, we uniformly sample a vector $\mathbf{z}$ from $\{\mathbf{x} : \mathbf{x}^T \mathbf{u}_k = \cos\beta\}$, i.e., $\mathbf{z}$ is a unit vector such that the angle between $\mathbf{z}$ and $\mathbf{u}_k$ is β. To achieve this, note that if $\mathbf{u}_k = \mathbf{e}_f, f \in [F]$ ($\mathbf{e}_f$ is the vector with only the f-th entry being 1), this uniform sampling can easily be achieved. For example, we can take $\mathbf{x} = (\cos\beta)\mathbf{e}_f + (\sin\beta)\mathbf{y}$, where $y(f) = 0$, $y(i) = s(i)/\sqrt{\sum_{j \neq f} s(j)^2}, i \neq f$, and $s(i) \sim \mathcal{N}(0,1), i \neq f$. We can then use a Householder transformation [87] to map the unit vector generated from the circular cone with basis vector $\mathbf{e}_f$ to the unit vector generated from the circular cone with basis vector $\mathbf{u}_k$. The corresponding Householder transformation matrix is (if $\mathbf{u}_k = \mathbf{e}_f$, $\mathbf{P}_k$ is set to be the identity matrix $\mathbf{I}$)

$$\mathbf{P}_k = \mathbf{I} - 2\mathbf{z}_k \mathbf{z}_k^T, \quad \text{where} \quad \mathbf{z}_k = \frac{\mathbf{e}_f - \mathbf{u}_k}{\|\mathbf{e}_f - \mathbf{u}_k\|_2}. \tag{3.6.1}$$

In this set of experiments, we set the size angles α to be the same for all the circular cones. The angle between any two basis vectors is set to be $4\alpha + \Delta\alpha$ where $\Delta\alpha :=$

0.01. The parameter for the exponential distribution $\boldsymbol{\lambda} := 1./(1 : K)$. We increase N from 10^2 to 10^4 logarithmically. We fix the parameters $F = 1600$, $K = 40$ and $\alpha = 0.2$ or 0.3. The results shown in Figure 3.2. In the left plot of Figure 3.2, we compare the relative errors of Algorithm 2 (`cr1-nmf`) with the derived relative error bounds. In the right plot, we compare the relative errors of our algorithm with the relative errors of three classical algorithms: (i) the multiplicative update algorithm [23] (`mult`); (ii) the alternating nonnegative least-squares algorithm with block-pivoting (`nnlsb`), which is reported to be one of the best alternating nonnegative least-squares-type algorithm for NMF in terms of both running time and approximation error [26]; (iii) and the hierarchical alternating least squares algorithm [28] (`hals`). In contrast to these three algorithms, our algorithm is not iterative. The iteration numbers for `mult` and `hals` are set to 100, while the iteration number for `nnlsb` is set to 20, which is sufficient (in our experiments) for approximate convergence. For statistical soundness of the results of the plots on the left, 50 data matrices $\mathbf{V} \in \mathbb{R}_{+}^{F\times 10000}$ are independently generated and for each data matrix $\mathbf{V}$, we run our algorithm for 20 runs. For the plots on the right, 10 data matrices $\mathbf{V}$ are independently generated and all the algorithms are run for 10 times for each $\mathbf{V}$. We also compare the running time for these algorithms when they first achieve the approximation error smaller than or equal the approximation error of Algorithm 2. The running times are shown in Table 3.1. Because the running times for $\alpha = 0.2$ and $\alpha = 0.3$ are similar, we only present the running times for the former.

From Figure 3.2, we observe that the relative errors obtained from Algorithm 2 are smaller than the theoretical relative error bounds. When $\alpha = 0.2$, the relative error of Algorithm 2 appears to converge to the probabilistic relative error bound as N becomes large, but when $\alpha = 0.3$, there is a gap between the relative error and the probabilistic relative error bound. From Theorems 8 and 9, we know that this difference is due to the projection of the cones to the nonnegative orthant. If there is no projection (this may violate the nonnegative constraint), the probabilistic relative error bound is tight as N tends to infinity. We conclude that when the size

Table 3.1: Running times in seconds of various algorithms ($\alpha = 0.2$)

N	`cr1-nmf`	`mult`	`nnlsb`	`hals`
10^2	**0.03**±0.03	1.56±0.76	5.82 ± 1.15	0.46±0.20
10^3	**0.26**±0.10	9.54±5.91	6.44 ± 2.70	3.01±1.85
10^4	**1.85**±0.22	85.92±54.51	27.84 ± 8.62	17.39±5.77

angle α is large, the projection step causes a larger gap between the relative error and the probabilistic relative error bound. We observe from Figure 3.2 that there are large oscillations for `mult`. Other algorithms achieve similar approximation errors. Table 3.1 shows that classical NMF algorithms require significantly more time (at least an order of magnitude for large N) to achieve the same relative error compared to our algorithm.

Automatically Determining K

We now verify the efficacy and the robustness of the proposed method in (3.5.1) for automatically determining the correct number of circular cones. We generated the data matrix $\hat{\mathbf{V}} := [\mathbf{V} + \delta\mathbf{E}]_+$, where each entry of $\mathbf{E}$ is sampled i.i.d. from the standard normal distribution, $\delta > 0$ corresponds to the noise magnitude, and $[\cdot]_+$ represents the projection to nonnegative orthant operator. We generated the nominal/noiseless data matrix $\mathbf{V}$ by setting $\alpha = 0.3$, the true number of circular cones $K = 40$, and other parameters similarly to the procedure in Section 3.6.1. The noise magnitude δ is set to be either 0.1 or 0.5; the former simulates a relatively clean setting in which the geometric assumption is approximately satisfied, while in the latter, $\hat{\mathbf{V}}$ is far from a matrix that satisfies the geometric assumption, i.e., a very noisy scenario. We generated 1000 perturbed data matrices $\hat{\mathbf{V}}$ independently. From Figure 3.3 in which the true $K = 40$, we observe that, as expected, the method in (3.5.1) works well if the noise level is small. Somewhat surprisingly, it also works well even when the noise level is relatively high (e.g., $\delta = 0.5$) if the number of data points N is also commensurately large (e.g., $N \geq 5 \times 10^3$).

3.6.2 Experiments on Real Datasets

Initialization Performance in Terms of the Relative Error

Because real datasets do not, in general, strictly satisfy the geometric assumption, our algorithm `cr1-nmf`, does not achieve as low a relative error compared to other NMF algorithms. However, similar to the popular spherical k-means (`spkm`; we use 10 iterations to produce its initial left factor matrix $\mathbf{W}$) algorithm [46], our algorithm may be used as *initialization method* for NMF. In this section, we compare `cr1-nmf` to other classical and popular initialization approaches for NMF. These include random initialization (`rand`), `spkm`, and the `nndsvd` initialization method [47] (`nndsvd`). We empirically show that our algorithm, when used as an initializer, achieves the best performance when combined with classical NMF algorithms. The specifications of the real datasets and the running times for the initialization methods are presented in Tables 3.2 and 3.3 respectively. We use face datasets (CK, faces94, Georgia Tech face datasets) because this type of datasets is popularly used for NMF. In fact, the invention of NMF is motivated by the finding that NMF can learn the parts of objects in face datasets [88]. We use hyperspectral imaging datasets (PaviaU dataset) and text datasets (tr11 and wap datasets, used for evaluating the clustering performance of various methods) because they are popularly used in the analysis of NMF with separability assumptions [34, 39]. These variants of NMF are closely related to our model, i.e., NMF with a geometric assumption. In addition, faces94 and Georgia Tech are balanced datasets, while CK, tr11 and wap datasets are unbalanced (for example, for CK dataset, the largest cluster contains 173 samples and the smallest cluster only contains 24 samples). We think that these differences in the datasets can help us to make a more comprehensive evaluation.

We use `mult`, `nnlsb`, and `hals` as the classical NMF algorithms that are combined

[3] http://www.consortium.ri.cmu.edu/ckagree/

[4] http://cswww.essex.ac.uk/mv/allfaces/faces94.html

[5] http://www.anefian.com/research/face_reco.htm

[6] http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

Table 3.2: Information for real datasets used

Dataset Name	F	N	K	Description
CK[3]	49×64	8795	97	face dataset
faces94[4]	200×180	3040	152	face dataset
Georgia Tech[5]	480×640	750	50	face dataset
PaviaU[6]	207400	103	9	hyperspectral

Table 3.3: Running times for initialization

Dataset Name	`cr1-nmf`	`spkm`	`nndsvd`
CK	$\mathbf{3.30} \pm 0.10$	6.68 ± 0.71	9.45 ± 0.12
faces94	$\mathbf{14.50} \pm 0.20$	32.23 ± 2.28	32.81 ± 0.29
Georgia Tech	$\mathbf{18.90} \pm 1.13$	24.77 ± 3.58	21.28 ± 0.35
PaviaU	$\mathbf{0.73} \pm 0.11$	2.47 ± 0.48	0.84 ± 0.12

with the initialization approaches. Note that for `nnlsb`, we only need to initialize the left factor matrix $\mathbf{W}$. This is because the initial $\mathbf{H}$ can be obtained from initial $\mathbf{W}$ using [26, Algorithm 2]. Also note that by the following lemma, the pair $(\mathbf{W}^*, \mathbf{H}^*)$ produced by Algorithm 2 is a fixed point for `mult`, so we use a small perturbation of $\mathbf{H}^*$ as an initialization for the right factor matrix.

Lemma 21. *The $(\mathbf{W}^*, \mathbf{H}^*)$ pair generated by Algorithm 2 remains unchanged in the iterations of standard multiplicative update algorithm [23] for NMF.*

Proof. There is at most one non-zero entry in each column of $\mathbf{H}^*$. When updating $\mathbf{H}^*$, the zero entries remain zero. For the non-zero entries of $\mathbf{H}^*$, we consider partitioning $\mathbf{V}$ into K submatrices corresponding to the K circular cones. Clearly,

$$\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_{\mathrm{F}}^2 = \sum_{k=1}^{K} \|\mathbf{V}_k - \mathbf{w}_k\mathbf{h}_k^T\|_{\mathrm{F}}^2, \tag{3.6.2}$$

where $\mathbf{V}_k \in \mathbb{R}^{F \times |\mathcal{I}_k|}$ and $\mathbf{h}_k \in \mathbb{R}_+^{|\mathcal{I}_k|}$. Because of the property of rank-one NMF (Lemma 17), for any k, when $\mathbf{w}_k$ is fixed, $\mathbf{h}_k \in \mathbb{R}_+^{|\mathcal{I}_k|}$ minimizes $\|\mathbf{V}_k - \mathbf{w}_k\mathbf{h}^T\|_{\mathrm{F}}^2$.

Table 3.4: Shift number for initialization approaches

	CK	faces94	Georgia Tech	PaviaU
`cr1-nmf+mult`	3	2	3	2
`spkm+mult`	6	5	4	7
`nndsvd+mult`	8	5	3	2
`cr1-nmf+hals`	2	2	2	1
`spkm+hals`	5	4	3	5
`nndsvd+hals`	7	4	2	1

Also, for the standard multiplicative update algorithm, the objective function is non-increasing for each update [23]. Thus $\mathbf{h}_k$ for each $k \in [K]$ (i.e., $\mathbf{H}^*$) will remain unchanged. A completely symmetric argument holds for $\mathbf{W}^*$. □

For `spkm`, similarly to [46, 47], we initialize the right factor matrix randomly. In addition, to ensure a fair comparison between these initialization approaches, we need to shift the iteration numbers appropriately, i.e., the initialization method that takes a longer time should start with a commensurately smaller iteration number when combined one of the three classical NMF algorithms. Table 3.4 reports the number of shifts. Note that unlike `mult` and `hals`, the running times for different iterations of `nnlsb` can be significantly different. We observe that for most datasets, when run for the same number of iterations, random initialization and `nndsvd` initialization not only result in larger relative errors, but they also take a much longer time than `spkm` and our initialization approach. Because initialization methods can also affect the running time of each iteration of `nnlsb` significantly, we do not report shifts for initialization approaches when combined with `nnlsb`. Table 3.5 reports running times that various algorithms first achieve a fixed relative error $\epsilon > 0$ for various initialization methods when combined with `nnlsb`. Our proposed algorithm is clearly superior.

We present supplementary plots for Figure 3.4 and Table 3.5. Here, we use running time instead of number of iterations for the horizontal axis. In particular, we display the performance of `nnlsb` for a fixed, reasonable, runtime with and without initialization. We only present the results for `nnlsb` because of two reasons. Firstly, when combined with initialization methods, compared to `mult` and `hals`, `nnlsb` is unusual because the running times for different iterations of `nnlsb` can differ significantly. Secondly, the plots for `mult` and `hals` are similar to that for `nnlsb`. From Figure 3.5, we further observe that our initialization method is superior to other initialization methods on the four datasets.

We observe from Figure 3.4 that our algorithm almost always outperforms all other initialization approaches in terms of convergence speed and/or the final relative error when combined with classical NMF algorithms for the selected real datasets (except that `nndsvd`+`hals` performs the best for PaviaU). In addition, we present the results from the Georgia Tech image dataset. For ease of illustration, we only display the results for 3 individuals (there are images for 50 individuals in total) for the various initialization methods combined with `mult`. Several images of these 3 individuals are presented in Figure 3.6. The basis images produced at the 20th iteration are presented in Figure 3.7 (more basis images obtained at other iteration numbers are presented in the supplementary material). We observe from the basis images in Figure 3.7 that our initialization method is clearly superior to `rand` and `nndsvd`. In the supplementary material, we additionally present an illustration of Table 3.5 as a figure where the horizontal and vertical axes are the running times (instead of number of iterations) and the relative errors respectively. These additional plots substantiate our conclusion that Algorithm 2 serves as a good initializer for various other NMF algorithms.

Intuition for the Advantages of `cr1-nmf` over `spkm` as an Initializer for NMF Algorithms

From Figure 3.4, we see that the difference between the results obtained from using `spkm` as initialization method and the corresponding results obtained from using our initialization approach appears to be rather insignificant. However, from Table 3.5, which reports the running time to *first* achieve specified relative errors $\epsilon > 0$ for the initialization methods combined with `nnlsb` (note that `nnlsb` only needs to use the initial left factor matrix, and thus we can compare the initial estimated basis vectors obtained by `spkm` and `cr1-nmf` directly), we see that our initialization approach is clearly faster than `spkm`.

In addition, consider the scenario where there are duplicate or near-duplicate samples. Concretely, assume the data matrix $\mathbf{V} := \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}_{+}^{2\times 3}$ and $K = 1$. Then the left factor matrix produced by rank-one NMF is $\mathbf{w} = [1; 0]$ and the normalized mean vector (centroid for `spkm`) is $\bar{\mathbf{u}} := [\frac{2}{\sqrt{5}}; \frac{1}{\sqrt{5}}]$. The approximation error w.r.t. $\mathbf{w}$ is $\|\mathbf{V} - \mathbf{w}\mathbf{w}^T\mathbf{V}\|_{\mathrm{F}} = 1$, while the approximation error w.r.t. $\bar{\mathbf{u}}$ is $\|\mathbf{V} - \bar{\mathbf{u}}\bar{\mathbf{u}}^T\mathbf{V}\|_{\mathrm{F}} \approx 1.0954$. Note that `spkm` is more constrained since it implicitly outputs a binary right factor matrix $\mathbf{H} \in \{0, 1\}^{K\times N}$ while rank-one NMF (cf. Lemma 17) does not impose this stringent requirement. Hence `cr1-nmf` generally leads to a smaller relative error compared to `spkm`.

Initialization Performance in Terms of Clustering

We now compare clustering performances using various initialization methods. To obtain a comprehensive evaluation, we use three widely-used evaluation metrics, namely, the normalized mutual information [89] (nmi), the Dice coefficient [90] (Dice) and the purity [91, 92]. The clustering results for the CK and tr11[7] datasets are shown in Tables 3.6 and 3.7 respectively. Clustering results for other datasets are shown in the supplementary material (for space considerations). We run the standard k`-means` and `spkm` clustering algorithms for at most 1000 iterations and terminate the algorithm if the cluster memberships do not change. All the classical NMF algorithms are terminated if the variation of the product of factor matrices is

[7] The tr11 dataset can be found at http://glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/datasets.tar.gz. It is a canonical example of a text dataset and contains 6429 terms and 414 documents. The number of clusters/topics is $K = 9$.

small over 10 iterations. Note that `nndsvd` is a deterministic initialization method, so its clustering results are the same across different runs. We observe from Tables 3.6 and 3.7 and those in the supplementary material that our initialization approach almost always outperforms all others (under all the three evaluation metrics).

In addition, we present the clustering performance of faces94 and wap[8] datasets. Information concerning the faces94 dataset is presented in Section 3.6.2. The other dataset under consideration, wap, is a text dataset with 8460 terms, 1560 documents, and the number of clusters is $K = 20$. From Tables 3.8 and 3.9, we observe that our initialization approach outperforms all others for all the three clustering evaluation metrics, except for the faces94 dataset where `spkm` with the `nnlsb` NMF algorithm.

[8] http://glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/datasets.tar.gz

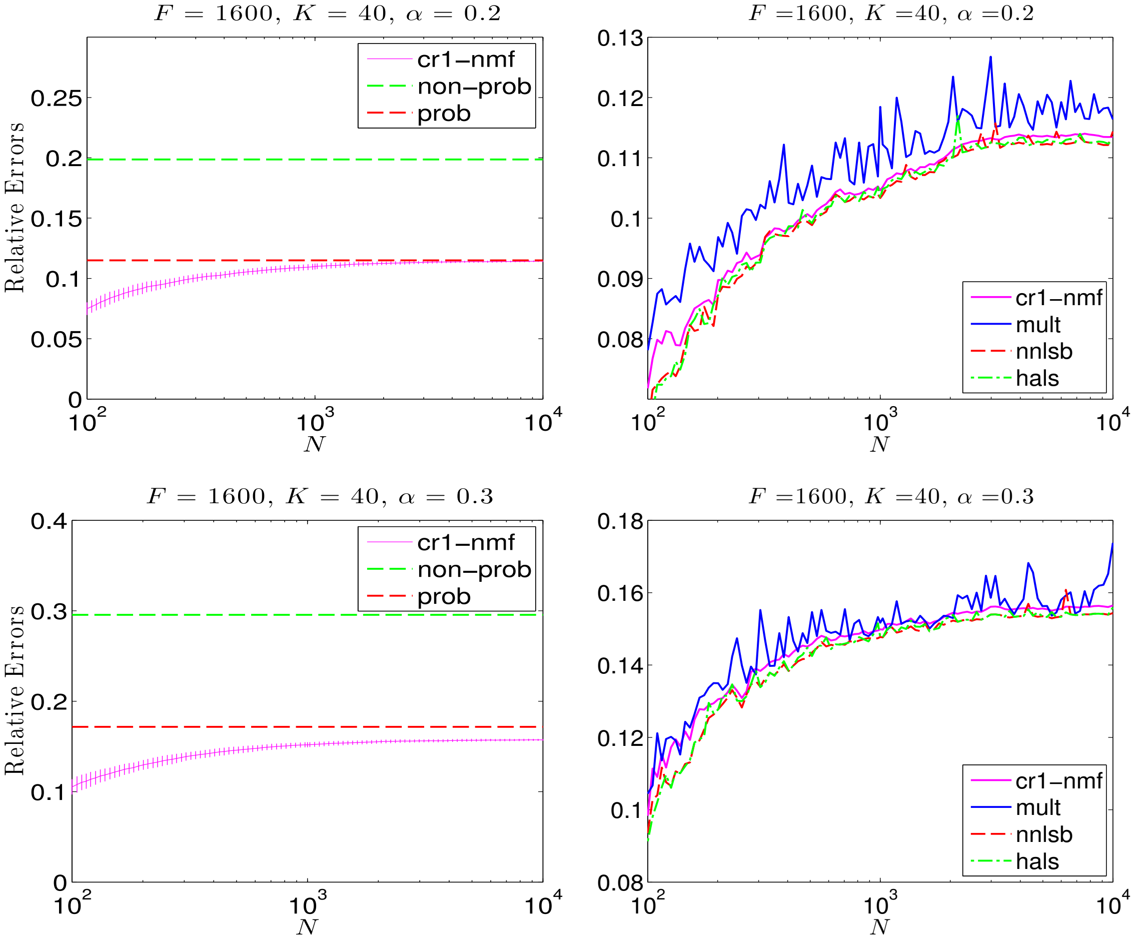

Figure 3.2: Errors and performances of various algorithms. On the left plot, we compare the empirical performance to the theoretical non-probabilistic and probabilistic bounds given by Theorems 6 and 8 respectively. On the right plot, we compare the empirical performance to other NMF algorithms.

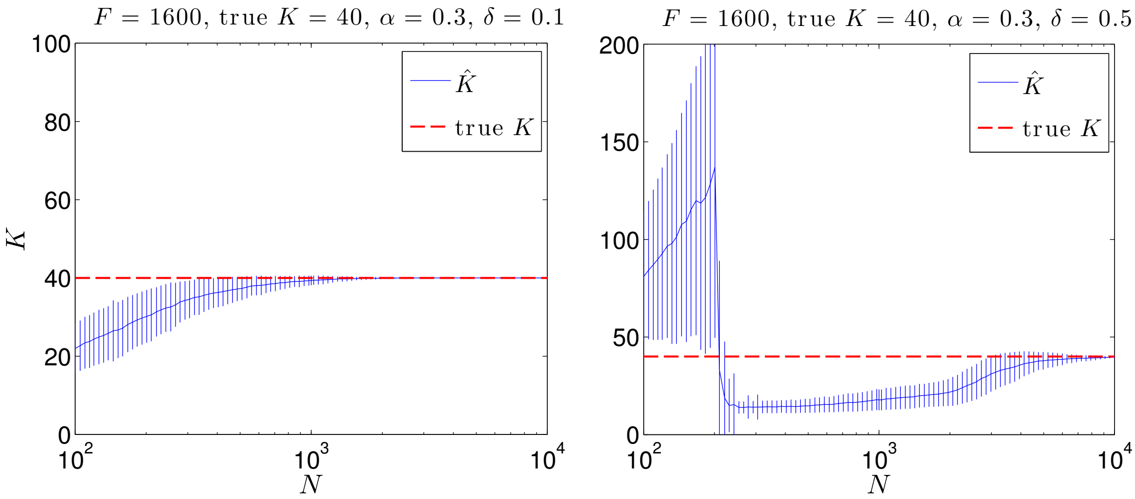

Figure 3.3: Estimated number of circular cones K with different noise levels. The error bars denote one standard deviation away from the mean.

Table 3.5: Running times when algorithm first achieve relative error ϵ for initialization methods combined with `nnlsb`

CK	$\epsilon = 0.105$	$\epsilon = 0.100$	$\epsilon = 0.095$
`rand`	727.53±23.93	1389.32±61.32	–
`cr1-nmf`	**40.27**±1.96	**71.77**±2.83	**129.62**±5.98
`spkm`	79.37±2.52	91.23±2.69	240.12±5.32
`nndsvd`	309.25±6.24	557.34±7.59	1309.51±21.97

faces94	$\epsilon = 0.140$	$\epsilon = 0.135$	$\epsilon = 0.131$
`rand`	2451.8±26.6	7385.8±49.6	–
`cr1-nmf`	**338.8**±11.1	**706.3**±13.3	**3585.2**±49.4
`spkm`	465.3±13.5	1231.1±28.5	5501.4±134.4
`nndsvd`	1531.5±6.4	3235.8±12.1	10588.6±35.9

Georgia Tech	$\epsilon = 0.185$	$\epsilon = 0.18$	$\epsilon = 0.175$
`rand`	3766.7±92.8	5003.7±126.8	7657.4±285.9
`cr1-nmf`	**147.3**±2.8	**308.2**±7.8	**1565.0**±59.5
`spkm`	253.2±20.1	537.4±43.4	2139.2±142.9
`nndsvd`	2027.0±7.0	2819.4±9.5	4676.4±15.3

PaviaU	$\epsilon = 0.0230$	$\epsilon = 0.0225$	$\epsilon = 0.0220$
`rand`	192.51±16.11	224.65±16.17	289.48±16.74
`cr1-nmf`	**13.30**±0.40	**16.93**±0.61	**30.06**±0.94
`spkm`	32.00±3.16	40.27±4.39	52.40±6.29
`nndsvd`	79.92±0.84	106.29±0.91	160.10±0.92

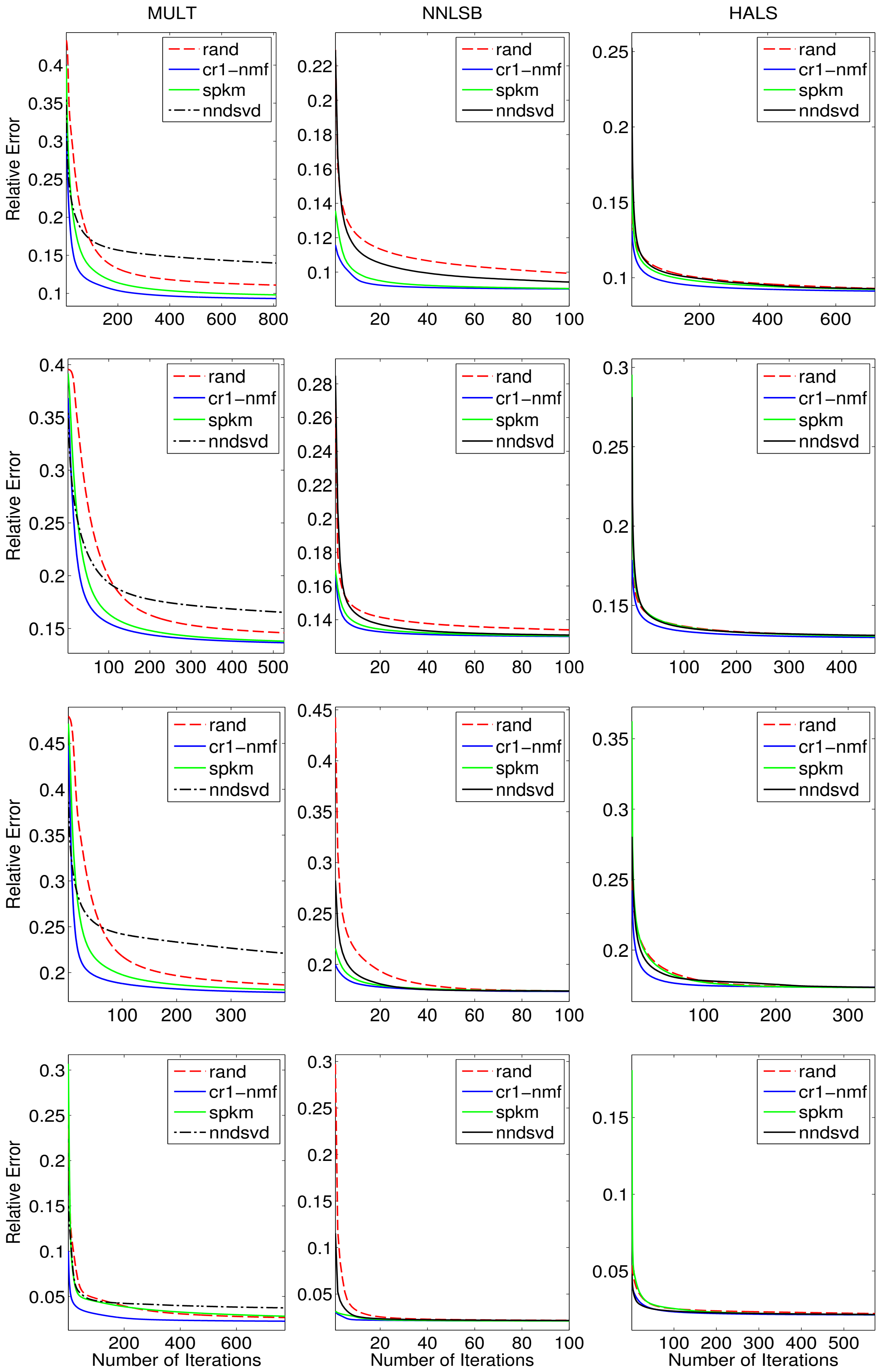

Figure 3.4: The first to fourth rows are the numerical results for CK, faces94, Georgia

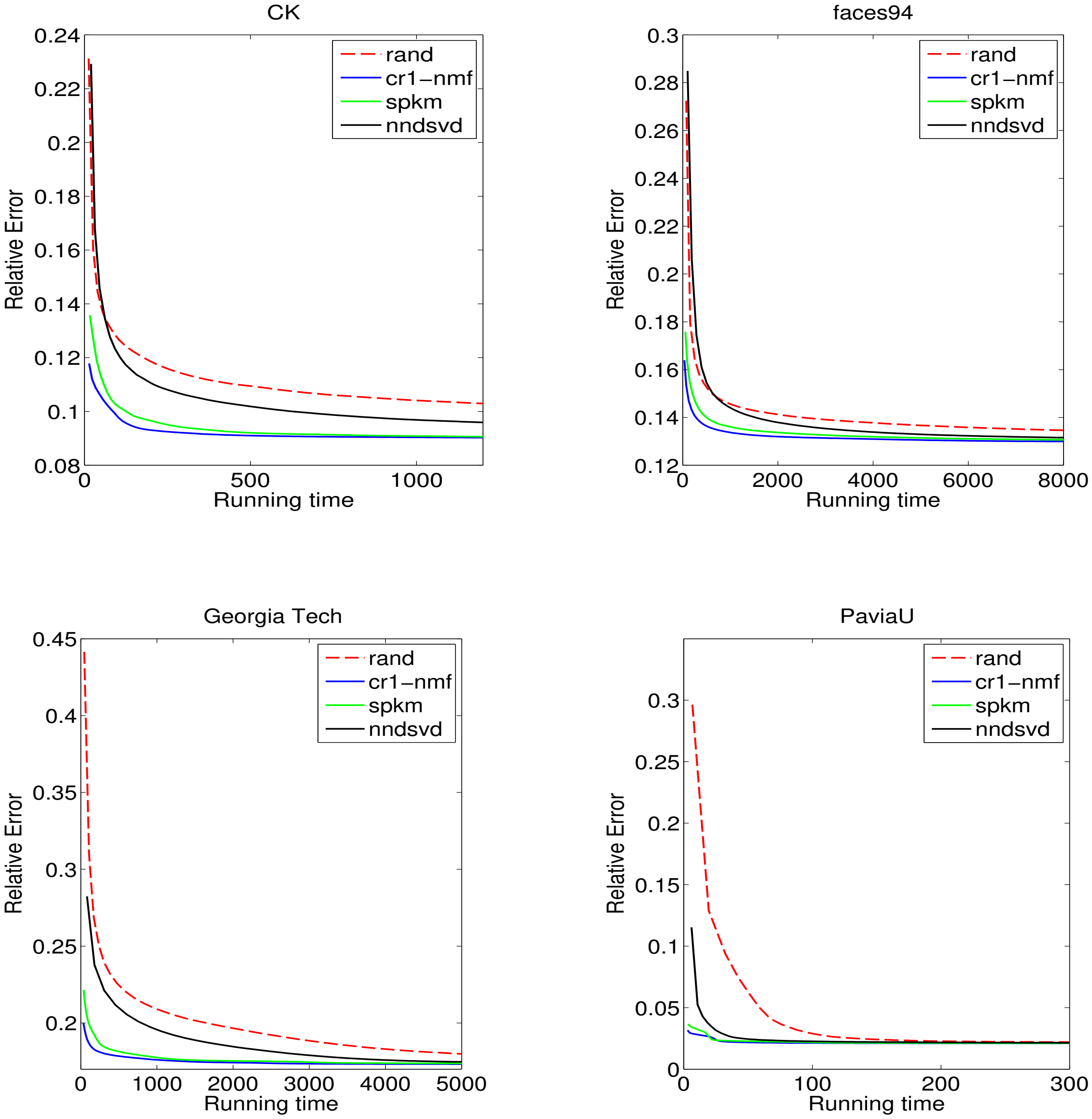

Figure 3.5: Numerical results for CK, faces94, Georgia Tech, and PaviaU datasets processed by `nnlsb` with different initialization methods.

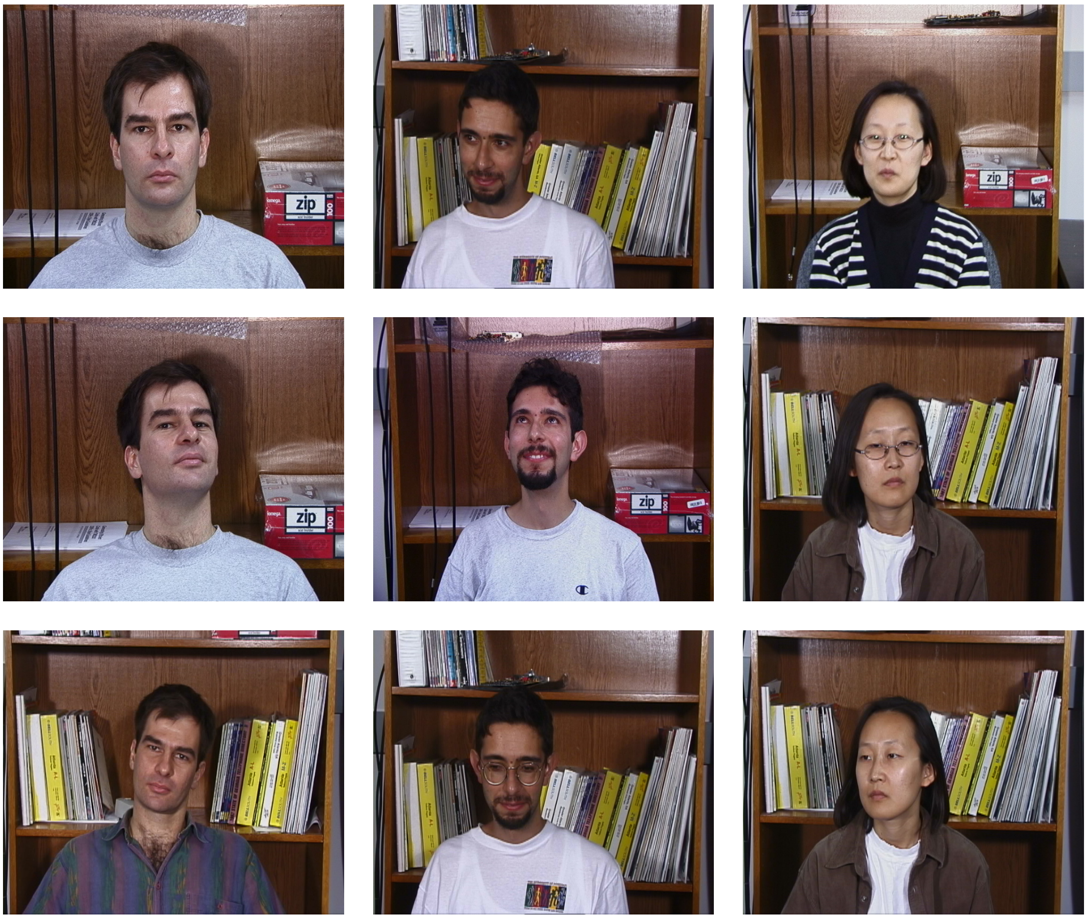

Figure 3.6: Images of 3 individuals in Georgia Tech dataset.

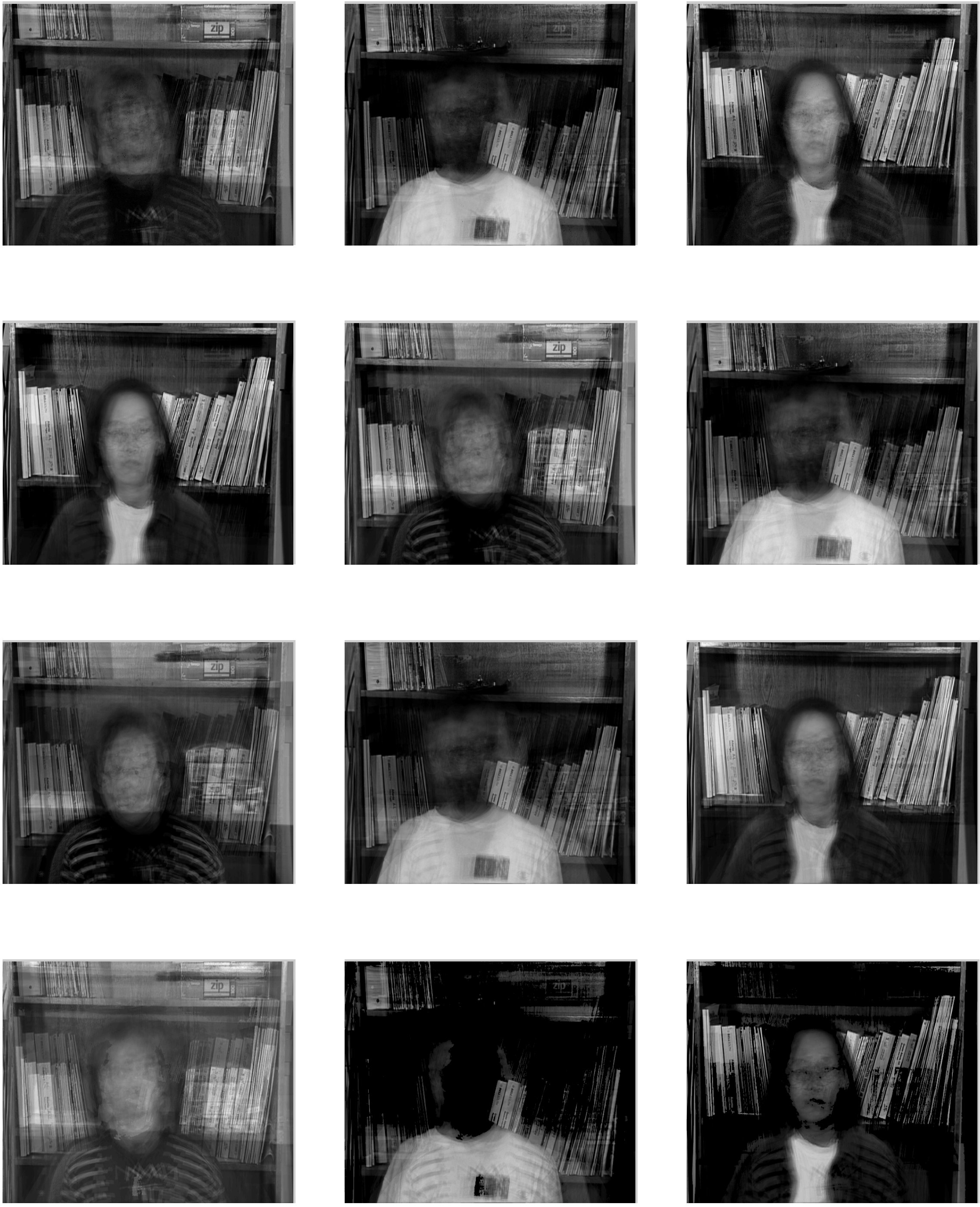

Figure 3.7: Basis images of 3 individuals in Georgia Tech dataset obtained at the 20th iteration. The first to fourth rows pertain to `rand`, `cr1-nmf`, `spkm`, and `nndsvd` initializations respectively.

Table 3.6: Clustering performances for initialization methods combined with classical NMF algorithms for the CK dataset

	nmi	Dice	purity
k-means	0.941±0.008	0.773±0.030	0.821±0.023
spkm	0.940±0.010	0.765±0.036	0.815±0.031
rand+mult	0.919±0.009	0.722±0.026	0.753±0.025
cr1-nmf+mult	**0.987**±0.002	**0.944**±0.006	**0.961**±0.006
spkm+mult	0.969±0.005	0.875±0.020	0.911±0.018
nndsvd+mult	0.870±0.000	0.614±0.000	0.619±0.000
rand+nnlsb	0.918±0.011	0.727±0.026	0.756±0.027
cr1-nmf+nnlsb	**0.986**±0.003	**0.940**±0.011	**0.959**±0.010
spkm+nnlsb	0.984±0.004	0.929±0.014	0.956±0.012
nndsvd+nnlsb	0.899±0.000	0.688±0.000	0.724±0.000
rand+hals	0.956±0.007	0.826±0.017	0.859±0.022
cr1-nmf+hals	**0.974**±0.006	**0.889**±0.015	**0.925**±0.016
spkm+hals	0.964±0.005	0.854±0.015	0.885±0.020
nndsvd+hals	0.942±0.000	0.786±0.000	0.830±0.000

Table 3.7: Clustering performances for initialization methods combined with classical NMF algorithms for the tr11 dataset

	nmi	Dice	purity
k-`means`	0.520±0.061	0.470±0.042	0.673±0.059
`spkm`	0.504±0.103	0.454±0.085	0.664±0.091
`rand`+`mult`	0.595±0.040	0.540±0.050	0.764±0.025
`cr1-nmf`+`mult`	**0.649**±0.049	**0.610**±0.052	**0.791**±0.023
`spkm`+`mult`	0.608±0.052	0.550±0.061	0.773±0.031
`nndsvd`+`mult`	0.580±0.000	0.515±0.000	0.761±0.000
`rand`+`nnlsb`	0.597±0.030	0.537±0.040	0.765±0.018
`cr1-nmf`+`nnlsb`	**0.655**±0.046	**0.615**±0.050	**0.794**±0.023
`spkm`+`nnlsb`	0.618±0.052	0.563±0.065	0.776±0.027
`nndsvd`+`nnlsb`	0.585±0.000	0.512±0.000	0.766±0.000
`rand`+`hals`	0.609±0.044	0.555±0.056	0.772±0.024
`cr1-nmf`+`hals`	**0.621**±0.052	**0.580**±0.062	**0.778**±0.026
`spkm`+`hals`	0.619±0.052	0.567±0.061	0.776±0.027
`nndsvd`+`hals`	0.583±0.000	0.511±0.000	0.768±0.000

Table 3.8: Clustering performance for initialization methods combined with classical NMF algorithms for the faces94 dataset

	nmi	Dice	purity
`k-means`	0.945±0.007	0.788±0.027	0.778±0.026
`spkm`	0.939±0.005	0.779±0.019	0.775±0.021
`rand` + `mult`	0.840±0.007	0.595±0.019	0.577±0.019
`cr1-nmf` + `mult`	**0.941**±0.003	**0.835**±0.009	**0.834**±0.009
`spkm` + `mult`	0.938±0.005	0.825±0.016	0.826±0.016
`nndsvd` + `mult`	0.589±0.000	0.217±0.000	0.172±0.000
`rand` + `nnlsb`	0.842±0.007	0.607±0.015	0.587±0.018
`cr1-nmf` + `nnlsb`	0.952±0.002	**0.873**±0.009	**0.873**±0.008
`spkm` + `nnlsb`	**0.953**±0.004	0.871±0.013	0.870±0.015
`nndsvd` + `nnlsb`	0.924±0.000	0.798±0.000	0.785±0.000
`rand` + `hals`	0.894±0.007	0.725±0.016	0.708±0.017
`cr1-nmf` + `hals`	**0.924**±0.004	**0.801**±0.010	**0.791**±0.008
`spkm` + `hals`	0.903±0.005	0.746±0.007	0.734±0.007
`nndsvd` + `hals`	0.871±0.000	0.700±0.000	0.678±0.000

Table 3.9: Clustering performance for initialization methods combined with classical NMF algorithms for the wap dataset

	nmi	Dice	purity
`k-means`	0.528±0.030	0.360±0.042	0.604±0.033
`spkm`	0.518±0.012	0.351±0.023	0.614±0.027
`rand` + `mult`	0.572±0.017	0.400±0.022	0.655±0.025
`cr1-nmf` + `mult`	**0.598**±0.020	**0.433**±0.030	**0.688**±0.022
`spkm` + `mult`	0.588±0.019	0.421±0.025	0.678±0.023
`nndsvd` + `mult`	0.589±0.000	0.429±0.000	0.679±0.000
`rand` + `nnlsb`	0.583±0.013	0.410±0.018	0.668±0.016
`cr1-nmf` + `nnlsb`	**0.600**±0.022	**0.432**±0.027	**0.688**±0.020
`spkm` + `nnlsb`	0.592±0.019	0.428±0.027	0.684±0.025
`nndsvd` + `nnlsb`	0.588±0.000	0.427±0.000	0.678±0.000
`rand` + `hals`	0.584±0.015	0.416±0.018	0.669±0.013
`cr1-nmf` + `hals`	**0.602**±0.015	**0.435**±0.022	**0.694**±0.019
`spkm` + `hals`	0.584±0.011	0.420±0.015	0.678±0.020
`nndsvd` + `hals`	0.594±0.000	0.421±0.000	0.681±0.000

Chapter 4

Conclusions and Future Works

In Chapter 2, we propose a fundamental understanding about when optimizing the objective function of the k-means algorithm returns a clustering that is close to the correct target clustering. To improve computational and memory issues, various dimensionality reduction techniques such as PCA are also considered.

In Chapter 3, we propose a new geometric assumption for the purpose of performing NMF. In contrast to the separability condition [33, 34, 36], under our geometric assumption, we are able to prove several novel deterministic and probabilistic results concerning the relative errors of learning the factor matrices. We are also able to provide a theoretically-grounded method of choosing the number of clusters (i.e., the number of circular cones) K. We show experimentally on synthetic datasets that satisfy the geometric assumption that our algorithm performs exceedingly well in terms of accuracy and speed. Our method also serves a fast and effective initializer for running NMF on real datasets. Finally, it outperforms other competing methods on various clustering tasks.

For using k-means and dimensionality reduction techniques to learn mixture models, several natural questions arise from the work in Chapter 2.

1. Instead of the separability assumptions made herein, we may consider modifying our analyses so that we eventually make less restrictive *pairwise* separability assumptions. This may enable us to make more direct comparisons between

our separability assumptions and similar assumptions in the literature, such as those in [17] and [55].

2. Brubaker [66] considers the *robust* learning of mixtures of log-concave distributions. Similarly, we may extend our work to the robust learning of noisy mixtures in which there may be outliers in the data.

3. Besides studying the sum-of-squares distortion measure for k-means in (1.1.1), it may be fruitful to analyze other objective functions such as those for k-medians [9] or min-sum clustering [10]. These may result in alternative separability assumptions and further insights on the fundamental limits of various clustering tasks.

4. We have provided *upper* bounds on the ME distance under certain sufficient (separability) conditions. It would be fruitful to also study *necessary* conditions on the separability of the mixture components to ensure that the ME distance is small. This will possibly result in new separability assumptions which will, in turn, aid in assessing the tightness of our bounds and how they may be improved.

For the work presented in Chapter 3, we plan to explore the following extensions.

1. First, we hope to prove theoretical guarantees for the scenario when $\mathbf{V}$ only satisfies an *approximate version* of the geometric assumption, i.e., we only have access to $\hat{\mathbf{V}} := [\mathbf{V} + \delta\mathbf{E}]_+$ (cf. Section 3.6.1) where $\delta \approx 0$.

2. Second, here we focused on upper bounds on the relative error. To assess the tightness of these bounds, we hope to prove *minimax lower* bounds on the relative error similarly to Jung et al. [93].

3. Third, as mentioned in Section 3.2, our geometric assumption in (3.2.2) can be considered as a special case of the near-separability assumption for NMF [33]. To the best of our knowledge, there is no theoretical guarantee for the relative error under the near-separability assumption.

4. For large-scale data, it is often desirable to perform NMF in an *online* fashion [94, 95], i.e., each data point $\mathbf{v}_n$ arrives in a sequential manner. We would like to develop online versions of the algorithm herein.

5. It would be fruitful to leverage the theoretical results for k-means++ [5] to provide guarantees for a probabilistic version of our initialization method. Note that our method is deterministic while k-means++ is probabilistic, so a probabilistic variant of Algorithm 2 may have to be developed for fair comparisons with k-means++.

6. We may also extend our Theorem 10 to near-separable data matrices, possibly with additional assumptions.

Bibliography

[1] K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

[2] J. A. Hartigan. *Clustering algorithms*, volume 209. Wiley New York, 1975.

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

[4] S. Dasgupta. *The hardness of k-means clustering.* Department of Computer Science and Engineering, University of California, San Diego, 2008.

[5] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proc. SODA*, pages 1027–1035. SIAM, 2007.

[6] L. Bottou and Y. Bengio. Convergence properties of the k-means algorithms. In *Proc. NIPS*, pages 585–592. Morgan Kaufmann Publishers Inc., 1995.

[7] J. Blömer, C. Lammersen, M. Schmidt, and C. Sohler. Theoretical analysis of the k-means algorithm–A survey. In *Algorithm Engineering*, pages 81–116. Springer, 2016.

[8] M. F. Balcan, A. Blum, and A. Gupta. Approximate clustering without the approximation. In *Proc. SODA*, pages 1068–1077. SIAM, 2009.

[9] P. S. Bradley, O. L. Mangasarian, and W. N. Street. Clustering via concave minimization. *Proc. NIPS*, pages 368–374, 1997.

[10] Y. Bartal, M. Charikar, and D. Raz. Approximating min-sum k-clustering in metric spaces. In *Proc. STOC*, pages 11–20. ACM, 2001.

[11] F. Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359, 1983.

[12] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *Proc. KDD workshop on text mining*, volume 400, pages 525–526. Boston, 2000.

[13] K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[14] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions.* Wiley Series in Probability and Statistics, 1985.

[15] C. M. Bishop. *Pattern recognition and machine learning.* Springer, 2006.

[16] S. Dasgupta. Learning mixtures of Gaussians. In *Proc. FOCS*, pages 634–644. IEEE, 1999.

[17] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *Proc. FOCS*, pages 113–122. IEEE, 2002.

[18] I. T. Jolliffe. *Principal component analysis.* Springer-Verlag, 1986.

[19] P. Drineas and V. Vinay. Clustering in large graphs and matrices. In *Proc. SODA*, pages 291–299. SIAM, 1999.

[20] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[21] A. Cichocki, R. Zdunek, A. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. John Wiley & Sons, 2009.

[22] I. Buciu. Non-negative matrix factorization, a new tool for feature extraction: theory and applications. *International Journal of Computers, Communications and Control*, 3:67–74, 2008.

[23] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Proc. NIPS*, pages 556–562, 2000.

[24] M. Chu, F. Diele, R. Plemmons, and S. Ragni. Optimality, computation, and interpretation of nonnegative matrix factorizations. *SIAM Journal on Matrix Analysis*, 2004.

[25] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.

[26] J. Kim and H. Park. Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In *Proc. ICDM*, pages 353–362, Dec 2008.

[27] J. Kim and H. Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, 33(6):3261–3281, 2011.

[28] A. Cichocki, R. Zdunek, and S. I. Amari. Hierarchical ALS algorithms for nonnegative matrix and 3d tensor factorization. In *Proc. ICA*, pages 169–176, Sep 2007.

[29] N.-D. Ho, P. Van Dooren, and V. D. Blondel. *Descent methods for Nonnegative Matrix Factorization*. Springer Netherlands, 2011.

[30] S. A. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20:1364–1377, 2009.

[31] C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, Oct. 2007.

[32] C.-J. Lin. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 18(6):1589–1596, Nov 2007.

[33] D. Donoho and V. Stodden. When does non-negative matrix factorization give correct decomposition into parts? In *Proc. NIPS*, pages 1141–1148. MIT Press, 2004.

[34] S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization–provably. In *Proc. STOC*, pages 145–162, May 2012.

[35] N. Gillis and S. A. Vavasis. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):698–714, 2014.

[36] V. Bittorf, B. Recht, C. Ré, and J. A. Tropp. Factoring nonnegative matrices with linear programs. In *Proc. NIPS*, pages 1214–1222, 2012.

[37] A. Kumar, V. Sindhwani, and P. Kambadur. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *Proc. ICML*, pages 231–239, Jun 2013.

[38] A. Benson, J. Lee, B. Rajwa, and D. Gleich. Scalable methods for nonnegative matrix factorizations of near-separable tall-and-skinny matrice. In *Proc. NIPS*, pages 945–953, 2014.

[39] N. Gillis and R. Luce. Robust near-separable nonnegative matrix factorization using linear optimization. *Journal of Machine Learning Research*, 15(1):1249–1280, 2014.

[40] N. Gillis. Robustness analysis of hottopixx, a linear programming model for factoring nonnegative matrices. *SIAM Journal on Matrix Analysis and Applications*, 34(3):1189–1212, 2013.

[41] N. Gillis. The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12(257), 2014.

[42] D. J. Hsu and S. M. Kakade. Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In *Proc. ITCS*, pages 11–20. ACM, 2013.

[43] A. Anandkumar, R. Ge, D. J. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[44] E. F. Gonzalez. *Efficient alternating gradient-type algorithms for the approximate non-negative matrix factorization problem.* PhD thesis, Rice University, Houston, Texas, 2009.

[45] M. Ackerman and S. Ben-David. Clusterability: A theoretical study. In *Proc. AISTATS*, volume 5, pages 1–8, 2009.

[46] S. Wild, J. Curry, and A. Dougherty. Improving non-negative matrix factorizations through structured initialization. *Pattern Recognition*, 37:2217–2232, 2004.

[47] C. Boutsidis and E. Gallopoulos. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41:1350–1362, 2008.

[48] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

[49] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[50] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, pages 95–103, 1983.

[51] S. Dasgupta and L. J. Schulman. A two-round variant of EM for Gaussian mixtures. In *Proc. UAI*, pages 152–159. Morgan Kaufmann Publishers Inc., 2000.

[52] S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proc. STOC*, pages 247–257. ACM, 2001.

[53] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *Proc. STOC*, pages 553–562. ACM, 2010.

[54] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

[55] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proc. COLT*, pages 444–457. Springer, 2005.

[56] A. Kumar and R. Kannan. Clustering with spectral norm and the k-means algorithm. In *Proc. FOCS*, pages 299–308. IEEE, 2010.

[57] C. Ding and X. He. k-means clustering via principal component analysis. In *Proc. ICML*, page 29. ACM, 2004.

[58] M. Meilă. Comparing clusterings: an axiomatic view. In *Proc. ICML*, pages 577–584. ACM, 2005.

[59] M. Meilă. The uniqueness of a good optimum for k-means. In *Proc. ICML*, pages 625–632. ACM, 2006.

[60] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas. Randomized dimensionality reduction for k-means clustering. *IEEE Transactions on Information Theory*, 61(2):1045–1062, 2015.

[61] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proc. STOC*, pages 163–172. ACM, 2015.

[62] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[63] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[64] H. Xiong, J. Wu, and J. Chen. k-means clustering versus validation measures: A data-distribution perspective. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):318–331, 2009.

[65] R. Vershynin. *High dimensional probability.* To be published by Cambridge University Press, 2016.

[66] S. C. Brubaker. Robust PCA and clustering in noisy mixtures. In *Proc. SODA*, pages 1078–1087. SIAM, 2009.

[67] S. Vempala. *The random projection method*, volume 65. American Mathematical Soc., 2005.

[68] E. Bingham and H. Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proc. SIGKDD*, pages 245–250. ACM, 2001.

[69] X. Z. Fern and C. E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proc. ICML*, volume 3, pages 186–193, 2003.

[70] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proc. FOCS*, pages 143–152. IEEE, 2006.

[71] S. Dasgupta. Experiments with random projection. In *Proc. UAI*, pages 143–151. Morgan Kaufmann Publishers Inc., 2000.

[72] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26(189-206):1, 1984.

[73] C. Boutsidis, P. Drineas, and M. W. Mahoney. Unsupervised feature selection for the k-means clustering problem. In *Proc. NIPS*, pages 153–161, 2009.

[74] C. Boutsidis and M. Magdon-Ismail. Deterministic feature selection for k-means clustering. *IEEE Transactions on Information Theory*, 59(9):6099–6110, 2013.

[75] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1):9–33, 2004.

[76] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k-means problem. In *Proc. FOCS*, pages 165–176. IEEE, 2006.

[77] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42:143–175, 2001.

[78] Y. Xue, C. S. Chen, Y. Chen, and W. S. Chen. Clustering-based initialization for non-negative matrix factorization. *Applied Mathematics and Computation*, 205:525–536, 2008.

[79] Z. Zheng, J. Yang, and Y. Zhu. Initialization enhancer for non-negative matrix factorization. *Engineering Applications of Artificial Intelligence*, 20:101–110, 2007.

[80] A. N. Langville, C. D. Meyer, and R. Albright. Initializations for the nonnegative matrix factorization. In *Proc. SIGKDD*, pages 23–26, Aug 2006.

[81] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proc. COLT*, pages 144–152, Jul 1992.

[82] V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1592–1605, 2013.

[83] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[84] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proc. ICML*, 2000.

[85] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63:411–423, 2001.

[86] R. L. Thorndike. Who belongs in the family. *Psychometrika*, pages 267–276, 1953.

[87] R. L. Burden and J. D. Faires. *Numerical Analysis.* Thomson/Brooks/Cole, 8 edition, 2005.

[88] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[89] A. Strehl and J. Ghosh. Cluster ensembles–a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, pages 583–617, 2002.

[90] G. Salton. *Automatic text processing : the transformation, analysis, and retrieval of information by computer.* Reading: Addison-Wesley, 1989.

[91] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to information retrieval.* Cambridge University Press, 2008.

[92] Y. Li and A. Ngom. The non-negative matrix factorization toolbox for biological data mining. *Source Code for Biology and Medicine*, 8, 2013.

[93] A. Jung, Y. C. Eldar, and N. Görtz. On the minimax risk of dictionary learning. *IEEE Transactions on Information Theory*, 62(3):1501–1515, 2016.

[94] R. Zhao and V. Y. F. Tan. Online nonnegative matrix factorization with outliers. *IEEE Transactions on Signal Processing*, 65(3):555–570, 2017.

[95] R. Zhao, V. Y. F. Tan, and H. Xu. Online nonnegative matrix factorization with general divergences. In *Proc. AISTATS*, 2017. arXiv:1608.00075.