Interpretation and Validation for Unsupervised Learning

Hanyu Zhang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Marina Meila-Predoviciu, Chair

Yen-Chi Chen

Zaid Harchaoui

Program Authorized to Offer Degree:
Department of Statistics

University of Washington

Abstract

Interpretation and Validation for Unsupervised Learning

Hanyu Zhang

Chair of the Supervisory Committee:
Marina Meila-Predoviciu
Department of Statistics

This thesis studies two major problems in unsupervised learning: manifold learning and clustering. The motivation of this research is to establish mathematically rigorous methods that enable practitioners to have better understanding of what the algorithm is doing, even if there is no ground truth label for unsupervised learning problems. Specifically, we propose two criterion for a useful unsupervised learning paradigm: interpretability and stability.

In the first part (chapter2-chapter 3), we propose a framework that allows domain experts to include a set of dictionary functions that can help provide manifold embedding coordinates with physical meaning. We first discuss mathematical foundation of this frameowrk. Based on this framework, we develop two algorithms. TSLasso obtains a manifold embedding function $\hat{\phi}$ that directly consists of functions from this dictionary as a valid *parametrization* of the data manifold. ManifoldLasso works with existing manifold embedding coordinates and outputs a subset of functions that parametrize the existing manifold embedding coordinates.

In the second part of the thesis (chapter 4-chapter 6), we introduce the stability of clustering to quantitatively validate a clustering result so that it is possible for practitioners to avoid these unwanted phenomena. Our target is to establish a generic notion $(\gamma, \epsilon)-$stability and show how this can be applied to real statistical tasks.

In chapter 5, we quantify population stability with respect to K-means clustering as a quantity for an arbitrary population P. With very mild assumptions on P, we show

this quantity of P relates to that of a finite sample drawn from P: if any optimal K-means clusterings of P is not stable, then with high probability any global optimizers of K-means on $i.i.d.$ sample of P is not stable; on the other hand, if population P allows one stable clustering with low K-means loss, then global optimizers of K-means clustering on i.i.d. sample is with high probability stable. We develop an algorithm to compute an upper bound of stability metric with respect to K-means clustering. As a byproduct, it provides an upper bound on the discrepancy between the global optimal K-means clustering assignment with the computed ones. We also provide emprical validation of this method.

In chapter 6, we focus on model-based clustering through fitting mixtures of spherical Gaussians (sGMM). Fitting sGMM is essentially a parameter estimation problem, and clustering assignments are based on the estimation. This thesis discusses mainly the parametric stability of sGMM: We show that if any two sGMMs are close, then their parameters are pairwise close. This result is proved with different assumptions on the model class of sGMMs. We can also see from numeric example that with the assumptions on the separation of different components in a Gaussian mixture, we obtain a precise upper bound on the parameter distances.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

LIST OF AlgorithmS

GLOSSARY

$[\boldsymbol{K}]$	The set $\{1, 2, \cdots, K\}$ when K is a natural number
$v, x, \cdots$	Vectors
$\mathbf{A}, \mathbf{X}, \cdots$	Matrices
$\mathcal{N}_i$	Set of indices of neighbors of point ξ_i
$\boldsymbol{\Xi}_i$	Local data set by considering difference $(\xi_j - \xi_i)_{j \in \mathcal{N}_i}$
$\mathbb{R}^d$	d dimensional Euclidean space
$\mathbb{S}^{d-1}$	Unit circle with dimension $d-1$
$\mathbb{N}$	Set of natural numbers
C^k	Has continuous derivatives up to order k. Specifically k could be ∞
$C^k(\mathcal{X})$	Space of C^k function on a space $\mathcal{X}$
$\mathcal{M}$	A manifold
$\{\xi_i\}_{i=1}^n$	A set of n data points, equivalent to the set $\{\xi_1, \xi_2, \cdots, \xi_n\}$
$\mathcal{D}$	Data set, usually consisting of data points $\{\xi_i\}_{i=1}^n$
$\|\|v\|\|$	L_2 norm of a vector v in Euclidean space
$\|\|\mathbf{A}\|\|_2$	Spectral norm of a squared matrix $\mathbf{A}$
$\|\|\mathbf{A}\|\|_F$	Frobenius norm of an arbitrary matrix $\mathbf{A}$
$\mathbf{A} \succeq 0$	$\mathbf{A}$ is positive semidefinite
Tr	Trace of a square matrix

$\mathbf{A}^\top$	Transpose of matrix $\mathbf{A}$
$\mathbf{A}^\dagger$	Pseudoinverse of matrix $\mathbf{A}$
$\mathcal{T}_\xi\mathcal{M}$	Tangent space of $\mathcal{M}$ at point ξ
$\mathfrak{d}$	Derivation of function on a manifold
d	Differential forms; differentials
$\langle\cdot,\cdot\rangle$	Inner product
$\frac{\partial f}{\partial x}$	Partial derivatives, also served as basis of tangent spaces
∇	Gradient operator in Euclidean space
Δ	Laplacian in Euclidean space
$\mathbf{1}_{Conditions}$	Indicator function that takes value one when condition holds, otherwise being zero
$\overline{\mathbf{1}}$	Vector with all entries being one
$\mathrm{grad}_{\mathcal{M}} f$	Gradient operator on manifold $\mathcal{M}$
$\Delta_{\mathcal{M}} f$	Laplacian operator on manifold $\mathcal{M}$
$\mathfrak{M}$	A model class
D	Difference on function values
$\mathbf{I}_d$	$d \times d$ Identity matrix
det	Determinant of a squared matrix
$\mathfrak{S}_K$	The set of permutations on K elements
$\mathbb{E}$	Expectation
Cov	Covariance
$P(\cdot)$	Probability of certain event

$\mathcal{N}_d$	Multivariate Gaussian distribution in d dimension
$O(a_n)$	Big O notation: $b_n = O(a_n)$ means there exists a constant $C > 0$ s.t. $b_n \leq Ca_n$ holds as $n \to \infty$
$o(a_n)$	Small O notation: $b_n = o(a_n)$ means $\lim_{n\to\infty} b_n/a_n = 0$
$\Omega(a_n)$	Big Ω notation, $b_n = \Omega(a_n)$ means there exists a constant $C > 0$ s.t. $b_n \geq Ca_n$ holds as $n \to \infty$
$\asymp$	$b_n \asymp a_n$ means that $b_n = O(a_n)$ and $b_n = \Omega(a_n)$ both hold as $n \to \infty$

ACKNOWLEDGMENTS

I am grateful to my advisor, Professor Marina Meila, for her unwavering support throughout my graduate studies. Her inspiration and encouragement have allowed me to freely explore a range of interesting problems. I would also like to express my appreciation to my committee members, Yen-Chi Chen, Zaid Harchaoui, and Kevin Jamieson, as well as other faculty members in the Statistics Department at UW, including Fang Han, Ema Perkovic, for their invaluable insights and guidance. I am also thankful to my collaborators, such as Sida Peng from Microsoft and Wei Sun from Fred Hutch Cancer Research Center, for their contributions to various research projects.

I would like to extend my gratitude to my groupmates Samson Koelle, Yu-Chia Chen, Zhenman Yuan, and James Buenfil, as well as my friends in the department, including Zhaoqi Li, Wenyu Chen, Jerry Wei, and Yikun Zhang.

Lastly, I would like to thank my wife, Hanying Leng, whom I met and married during my PhD program, and my family in China for their support and encouragement, particularly during the unprecedented challenges posed by the pandemic.

DEDICATION

To my dear wife, Hanying

Chapter 1

INTRODUCTION

1.1 Overview

Understanding data is one of the core missions of data science. Given a dataset, or a set of observations, it is often interesting to provide a compressed description or statistical model that fits the data. Such tasks are often examples of *unsupervised learning*. In contrast to *supervised learning*, another powerful machine learning technique, unsupervised learning tasks come with no labeled data and no ground truths. Furthermore, often these tasks are viewed as a complicated version of descriptive statistics and can be the cornerstone for further data analysis procedures. As we will see shortly, the results of unsupervised learning algorithms are often mysterious. They are often hard to interpret, and these outputs may change drastically even if the algorithm is re-run on the same data set. In this dissertation, we aim to provide a set of algorithms and mathematical results that help users to understand and trust results returned from two sets of typical unsupervised learning tasks: manifold learning and clustering.

Manifold learning Manifold learning is one of the handy tools for revealing the intrinsic structure of high-dimensional data. Nowadays, data (images, texts, RNA sequence data, etc.) often come with hundreds, thousands, or even millions of covariates, which on the one hand, provides more information, but on the other hand, hinders understanding data in various ways. Recent developments [Bellman, 1966] in high dimensional probability and statistics have illustrated challenges of analyzing high-dimensional data, the so-called *cure of dimensionality*. Widely accepted manifold assumption proposes that high dimensional data concentrate near a lower dimensional manifold. Hence a popular topic is to recover the structure of this low dimensional manifold. Twenty years after the first manifold learning algorithm, Isomap [Bernstein et al., 2000], was invented, various algorithms appeared and

can provide a low-dimensional representation of the data while retaining critical geometric features. However, it is hard to include domain knowledge and interpret the result. In the first part (chapter 2 to chapter 3) of this dissertation, we will establish a mathematical framework and algorithmic methodology to provide a solution to this interpretability issue.

Clustering Clustering analysis aims to find subgroups of data so that the observations within each group are more similar than the observations not from this group. Though many different clustering algorithms exist, recent result in [Jin et al., 2016b] demonstrate that many clustering algorithms can provide different results even given the same data set since most clustering algorithms, as optimization problems, are combinatorial optimization problems and are often NP-hard. In the second part of the thesis (chapter 4-chapter 6), we focus on providing a clustering stability guarantee under different clustering paradigms so that to determine whether the output of a clustering algorithm is trustworthy.

Contents of this dissertation have been extracted from, with modifications, the following manuscripts: Chapter 3 is taken and reorganized from *Manifold Coordinates with Physical Meaning* and *Dictionary-based Manifold Learning*, both co-authored with Samson Koelle, Marina Meila and Yu-chia Chen. Chapter 5 is taken from *Distribution free optimality intervals for clustering*, co-authored with Marina meila. Chapter 6 is taken from *Parameter Stability of Spherical Gaussian Mixture Models*, co-authored with Marina Meila. The author wants to point out that the experiments in chapter 3 is implemented by Samson Koelle and in chapter 5 is implemented by Marina Meila. The author include these contents for the completeness of the thesis. However, the major contribution of the author is in the theory part of the thesis.

1.2 Part I: Manifold embedding with physical meaning

In this dissertation we focus on the manifold learning paradigm called manifold embedding algorithms, which learn a mapping $\hat{\phi}$ that sends high dimensional data points to low dimensional representations. The key assumption of manifold learning is data are sampled from a distribution that is supported close to a lower dimensional manifold embedded in the high dimensional space. Since the proposal of Isomap in 2000 [Bernstein et al., 2000], various

manifold embedding algorithms have appeared [Belkin and Niyogi, 2002, Zhang and Zha, 2004]. In the limiting scenario that distribution is supported on a manifold $\mathcal{M}$ and sample size tends to infinity, these methods provide consistent results in the sense that learned mapping $\hat{\phi}$ will converge to a smooth embedding of $\mathcal{M}$ to the low dimension space. In chapter 2, we provide a more comprehensive review of the mathematical background and progress of manifold learning.

These manifold embedding algorithms are handy for scientists to study complicated systems. Take the molecular dynamic simulation(MDS) study as an example. In an MDS study, raw configuration data are locations of atoms in a molecule system and can easily reach hundreds of dimensions. Manifold embedding as a dimension reduction technique is very useful for scientists to find only a few variables to describe the state of the studied molecular system. These variables, called collective variables, can govern the movement of a molecule.

There are two challenges in using manifold embedding paradigms for this task. First, the procedure of learning $\hat{\phi}$ does not include any domain knowledge of chemists. Therefore, the discovered variables are not physically meaningful, hindering the algorithm outputs' interpretation. Second, the learned $\hat{\phi}$ does not have a functional form. Therefore, it is hard to consistently map data not in the sample to their low dimensional representation.

In contrast, when scientists describe/model a system using knowledge from their domain, often the resulting model is in terms of domain-relevant features, which are continuous functions of other domain variables (e.g., equations of motion, rotation of angles, etc.). The key question in the first part of this thesis is: is it possible to use a subset of these physically meaningful functions to describe the data manifold? If such a subset of functions exists, we will call them a *parametrization* of the data manifold.

Contributions On a high level, a user can propose any reasonable domain-related smooth functions on the manifold data. In this dissertation, we introduce a framework in chapter 3 that allows domain experts to include a set of dictionary functions that can help provide manifold embedding coordinates with physical meaning. Based on this framework, we develop two algorithms. TSLasso obtains a manifold embedding function $\hat{\phi}$ that directly

consists of functions from this dictionary as a valid *parametrization* of the data manifold. ManifoldLasso works with existing manifold embedding coordinates and outputs a subset of functions that parametrize the existing manifold embedding coordinates.

1.3 Part II: Stability of clustering

Clustering is a broad topic that applies to various data types. In chapter 4, we review some of the remarkable results established for clusterings. These results generally show that if the data are indeed clusterable or come from a specific probabilistic model, then current algorithms can efficiently find or approximate the correct clustering.

However, there is no standard method to validate such assumptions. Sometimes the population P or the data set does not have a meaningful sub-group structure (under the K-means paradigm) with the given number of clusters, while K-means can still output some results. In this thesis, we introduce the stability of clustering to quantitatively validate a clustering result so that it is possible for practitioners to avoid these unwanted phenomena. Our target is to establish a stability guarantee theorem in the following form:

Theorem 1.1 (Generic theorem)**.** *Given a population P (could be empirical distribution on finite sample), a clustering $\mathcal{C}$ and a clustering paradigm $\mathfrak{M}$, $\mathcal{C}$ is (γ, ϵ)-stable if all clustering $\mathcal{C}'$ with $g(\mathcal{C}', P) \leq g(\mathcal{C}, P) + \gamma$ must have $d(\mathcal{C}, \mathcal{C}') \leq \epsilon$.*

We identify three important components in our framework.

- Clustering paradigm $\mathfrak{M}$: It could be a loss-based clustering, model-based clustering, etc.

- Clustering performance $g(\mathcal{C}, \mathcal{D})$: It can often be a loss function, a likelihood function, or other quantities that measure the performance of a clustering. $g(\mathcal{C}, \mathcal{D})$ should be different for different clustering paradigms. Without loss of generality, we regard a smaller value of $g(\mathcal{C}, \mathcal{D})$ means better clustering.

- Clustering distance $d(\mathcal{C}, \mathcal{C}')$: This is a closeness measure for two different clusterings $\mathcal{C}, \mathcal{C}'$. Similarly, this distance is adaptive to the choice of clustering paradigms.

Establishing a general guarantee like 1.1 for all clustering algorithms is challenging. Some previous work [Wan and Meilă, 2016] have established such results for graph clustering settings. In this thesis, we mainly consider clusterings of data in Euclidean space $\mathbb{R}^d$. The second part of this thesis instantiates theorem 1.1 in multiple cases.

Contributions We focus on K-means clustering in chapter 5. We quantify population stability with respect to K-means clustering as a quantity for an arbitrary population P. With very mild assumptions on P, we show this quantity of P relates to that of a finite sample drawn from P: if any optimal K-means clusterings of P is not stable, then with high probability any global optimizers of K-means on *i.i.d.* sample of P is not stable; on the other hand, if population P allows one stable clustering with low K-means loss, then global optimizers of K-means clustering on i.i.d. sample is with high probability stable. We develop an algorithm to compute an upper bound of stability metric with respect to K-means clustering. As a byproduct, it provides an upper bound on the discrepancy between the global optimal K-means clustering assignment with the computed ones.

In chapter 6, we focus on model-based clustering through fitting mixtures of spherical Gaussians (sGMM). Fitting sGMM is essentially a parameter estimation problem, and clustering assignments are based on the estimation. This thesis discusses mainly the parametric stability of sGMM: We show that if any two sGMMs are close, then their parameters are pairwise close. This result is proved with different assumptions on the model class of sGMMs. As we shall see, with the assumptions on the separation of different components in a Gaussian mixture, we obtain a precise upper bound on the parameter distances.

Part I

INTERPRETABLE MANIFOLD LEARNING THROUGH SPARSE RECOVERY

Chapter 2

MATHEMATICAL FOUNDATIONS OF INTERPRETABLE MANIFOLD LEARNING

2.1 Differential geometry

2.1.1 Manifold, tangent space and embedding

Intuitively, the notion of a manifold is a generalization of curves and surfaces. Readers can consult [Lee, 2003, do Carmo, 1992] for the formal definitions and a detailed treatment of manifolds and differential geometry. In this thesis, smooth usually means C^∞.

We start with the mathematical definition of a smooth manifold.

Definition 2.1 (Smooth manifold)**.** *$\mathcal{M}$ is a manifold of dimension d (also called a d-manifold) if it is a topological space with the following property:*

- *$\mathcal{M}$ is Hausdorff: any two distinct points can be separated by disjoint open sets.*
- *$\mathcal{M}$ is second-countable: $\mathcal{M}$ has a countable base.*
- *For each $\xi \in \mathcal{M}$, there exist an open set $U \subset \mathcal{M}$ that contains ξ, an open subset $\hat{U} \subset \mathbb{R}^d$ and a function $\varphi : U \to \hat{U}$ such that φ is bijective and continuous from $U \to \hat{U}$ and φ^{-1} is also continuous.*

The pair (U, φ) is called a (coordinate) chart, and the component functions of φ are denoted by $(x^1(\xi), \cdots, x^d(\xi))$. φ is called local coordinate. Two charts $(U, \varphi), (V, \psi)$ on a d dimensional manifold $\mathcal{M}$ are (smoothly) compatible if whenever $U \cap V \neq \emptyset$, the map $\varphi \circ \psi^{-1}$ is a diffeomorphism. A smooth atlas is a collection of mutually smoothly compatible charts whose domains cover $\mathcal{M}$. Finally, a smooth manifold is a manifold with a smooth atlas.

This definition intuitively states that, locally around each of its points, a manifold behaves like an open subset in $\mathbb{R}^d$. The assumption of Hausdorff and second-countable are

technical requirements. These two conditions are easy to satisfy for most standard objects. Euclidean spaces, compact metric spaces, and separable metric spaces satisfy this assumption.

Definition 2.2 (Smooth functions)**.** *Suppose $\mathcal{M}, \mathcal{N}$ are two smooth manifolds. A function $f : \mathcal{M} \to \mathbb{R}$ is smooth (or C^k) if for any point $\xi \in \mathcal{M}$, there exist a coordinate chart (U, φ) containing ξ such that $f \circ \varphi^{-1}$ is smooth (or C^k).*

Similarly a smooth mapping $F : \mathcal{M} \to \mathcal{N}$ is smooth (or C^k) if for any point $\xi \in \mathcal{M}$, there exists a coordinate chart (U, φ) containing ξ and (V, ψ) containing $F(\xi)$ such that $\psi \circ F \circ \varphi^{-1}$ is smooth (or C^k).

We denote the space of all smooth (or C^k) functions on $\mathcal{M}$ by $C^\infty(\mathcal{M})$ (or $C^k(\mathcal{M})$).

Definition 2.3 (Differentials, tangent vectors and tangent spaces)**.** *A* derivation *at ξ is a linear operator $\mathfrak{d} : C^\infty(\mathcal{M}) \to \mathbb{R}$ such that for any smooth functions $f, g \in C^\infty(\mathcal{M})$, it holds that*

$$\mathfrak{d}(fg) = g(\xi)\mathfrak{d}(f) + f(\xi)\mathfrak{d}(g) \ . \tag{2.1}$$

Image of this operator over all $f \in C^\infty(M)$ is called tangent space of $\mathcal{M}$ at ξ, denoted by $\mathcal{T}_\xi\mathcal{M}$. An element of $\mathcal{T}_\xi\mathcal{M}$ is called a tangent vector.

Further, differential operator $\mathfrak{d}$ on $\mathcal{M}$ induces a mapping $\mathrm{d}F$ *defined on smooth maps between two manifolds $\mathcal{M}, \mathcal{N}$. Note that for any smooth function $f \in C^\infty(\mathcal{N})$, $f \circ F \in C^\infty(\mathcal{M})$. This mapping* $\mathrm{d}F$ *sends tangent vectors in $\mathcal{T}_\xi\mathcal{M}$ to elements in $\mathcal{T}_{F(\xi)}\mathcal{N}$ in the way that*

$$(\mathrm{d}F)(f) = \mathfrak{d}(f \circ F) \quad \forall f \in C^\infty(\mathcal{N}), g \in C^\infty(\mathcal{M}) \ .$$

$\mathrm{d}F$ *is called the* differential *of F. The rank of* $\mathrm{d}F$ *is called* rank *of F.*

This is a formal definition. Intuitively, one can think of tangent vectors as all possible directions in Euclidean space. Derivation is then directional derivatives in the usual calculus sense. One can check that differential is a linear map between tangent spaces and satisfy the chain rule as standard derivatives.

There are special cases of smooth mappings between two manifolds.

Definition 2.4 (Embedding)**.** *A smooth map $F : \mathcal{M} \to \mathcal{N}$ is an immersion if rank F equals the dimension of $\mathcal{M}$. An immersion F is an embedding if it is injective and homeomorphism of its image $F(\mathcal{M}) \subset \mathcal{N}$ in subspace topology.*

Of particular interest is the case $\mathcal{M} \subset \mathcal{N}$; if the inclusion $i : \mathcal{M} \to \mathcal{N}$ is an embedding, then $\mathcal{M}$ is said to be a submanifold of $\mathcal{N}$. Commonly in statistics, the high dimensional data that lie initially $\mathcal{N} = \mathbb{R}^D$, and we model them by $\mathcal{M}$ a submanifold of $\mathbb{R}^d$ to be estimated. Then D is called the ambient dimension (of the data)

A good thing about embedding is avoiding using multiple charts to describe a manifold. Instead, one can find a global mapping $F : \mathcal{M} \to \mathcal{N}$ where $\mathcal{N}$ is easier to understand. Whitney's embedding theorem states that every $d-$dimensional manifold can be embedded into $\mathbb{R}^{2d}$. Therefore, if one can find a valid embedding, a significant dimension reduction can be achieved (from D to $O(d)$). Such an embedding is one of the major targets of manifold learning algorithms.

In differential geometry, a global rank theorem shows that the property of an embedding can be given by rank theorem.

Theorem 2.1 (Global rank theorem)**.** *Let $\mathcal{M}, \mathcal{N}$ be two differential manifolds, and $F : \mathcal{M} \to \mathcal{N}$ is a smooth map with constant rank.*

- *If F is surjective, it is a submersion.*
- *If F is injective, it is an immersion.*
- *If F is bijective, it is a diffeomorphism.*

2.1.2 Riemannian metric

A scientist may be interested in the distance between two molecular configurations ξ_1, ξ_2, seen as points of $\mathcal{M} \subset \mathbb{R}^D$. Their Euclidean distance $||\xi_1 - \xi_2||$ is readily available without requiring additional statistics. However, this value may not be of physical interest since most putative configurations along the segment ξ_1 to ξ_2 in $\mathbb{R}^D$ are not physically possible. To deform from state ξ_1 to ξ_2, the ethanol molecule must follow a path contained in (or

near) the manifold $\mathcal{M}$ of possible configurations, and the distance $d_{\mathcal{M}}(\xi_1, \xi_2)$ shall naturally be defined as the shortest possible length of such a path (and is defined as the *geodesic distance*). Just like in $\mathbb{R}^d$, the distance between two points is independent of the choice of basis and invariant if $\mathbb{R}^d$ is a subspace of a larger Euclidean space, distances along curves in a manifold $\mathcal{M}$ can be defined solely based on the coordinate charts (U, φ), hence *intrinsically*, without reference to the ambient space $\mathbb{R}^D$, and are purely geometric, hence are independent of the choices of charts.

In $\mathbb{R}^D$, the scalar product $\langle v, u\rangle = v^\top u$ is sufficient to define both distances, by $||v-u||^2 = \langle v-u, v-u\rangle$, and angles, by their cosine value $\langle v, u\rangle/(||v||||u||)$. Moreover, any positive definite matrix $\mathbf{A} \in \mathbb{R}^{D\times D}$ can define a scalar product by $\langle v, u\rangle_{\mathbf{A}} = v^\top \mathbf{A} u$; in this case, $\mathbf{A}$ is often called a *metric* on $\mathbb{R}^D$.

Riemann took up this idea and introduced the following definition to study the intrinsic geometry of a manifold. The *Riemannian metric* defined below plays the same role as $\mathbf{A}$ above. However, this metric is allowed to vary from point to point.

Definition 2.5. *A Riemannian metric* $\mathbf{g}$ *of a manifold* $\mathcal{M}$ *associates to each point* $\xi \in \mathcal{M}$ *an inner product* $\langle\cdot,\cdot\rangle_{\mathbf{g}(\xi)}$ *on the tangent space* $\mathcal{T}_\xi\mathcal{M}$*, which varies smoothly in the following sense: let* (U, φ) *be a coordinate chart containing* ξ *with components functions* $(x^1, \cdots, x^d)$*, then the function* $\mathbf{g}_{ij}(x^1, \cdots, x^d) = \langle \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\rangle_{\mathbf{g}(\xi)}$ *is smooth for all pairs of* i, j *in* $1, \cdots, d$*. The pair* $(\mathcal{M}, \mathbf{g})$ *is called a Riemannian manifold.*

This inner product defines various geometric quantities of $\mathcal{M}$. These expressions are invariant to the choice of bases in $\mathcal{T}\mathcal{M}$, hence to the choice of coordinate charts on $\mathcal{M}$.

- norm $\|v\|_{\mathbf{g}} = \sqrt{\langle v, v\rangle_{\mathbf{g}}}$
- distance $||u - v||$
- angle $\langle u, v\rangle_{\mathbf{g}}/(\|u\|_{\mathbf{g}}\|v\|_{\mathbf{g}})$
- line element $\mathrm{d}l = \sum_{i,j=1}^{d} \mathbf{g}_{ij}\mathrm{d}x^i\mathrm{d}x^j$
- volume element $\mathrm{d}V = \sqrt{\det(\mathbf{g})}\mathrm{d}x^1 \cdots \mathrm{d}x^d$

A smooth embedding F from a Riemannian manifold $(\mathcal{M}, \mathbf{g})$ to a smooth manifold $\mathcal{N}$ induces a metric $F_*\mathbf{g}$ called *push-forward metric* through

$$\langle u, v\rangle_{F_*\mathbf{g}} = \langle (\mathrm{d}F)^{-1}u, (\mathrm{d}F)^{-1}v\rangle_{\mathbf{g}}, \quad \text{for all} \;\; u, v \in \mathcal{T}_{F(\xi)}\mathcal{N} \; . \tag{2.2}$$

Ideally, we would expect the learned embedding $\hat{F}$ to converge to a mapping that preserves the metric of $\mathcal{M}$. Mathematically, this property is defined as follows.

Definition 2.6 (Isometry)**.** *A smooth embedding $F : \mathcal{M} \to \mathcal{N}$ between two Riemannian manifolds $(\mathcal{M}, \mathbf{g}), (\mathcal{N}, \mathbf{h})$ is isometric if*

$$\langle u, v\rangle_{\mathbf{g}(\xi)} = \langle \mathrm{d}F_\xi(u), \mathrm{d}F_\xi(v)\rangle_{\mathbf{h}(F(\xi))}, \quad \textit{for all} \;\; u, v \in \mathcal{T}_\xi\mathcal{M} \; .$$

An isometry F preserves local geometric quantities such as angles, distances, path lengths, and volumes. For manifold learning, we almost always assume that the manifold where data lie is locally isometrically embedded in the ambient space $\mathbb{R}^D$. Since as we will see, these algorithms use local geometry in $\mathbb{R}^D$ to approximate the intrinsic geometry.

2.1.3 Gradients and Laplace-Beltrami operator

Using the Riemannian metric, we can define differential operators as in the study of calculus in Euclidean space. We first review the following definitions.

Definition 2.7 (Gradients in $\mathbb{R}^D$)**.** *The gradient of a function $f \in C^1(\mathbb{R}^D)$ is given by*

$$\nabla f = \left[\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_D}\right] \; .$$

Under the Riemannian manifold setting, these differential operators are defined as follows.

Definition 2.8 (Gradients on $\mathcal{M}$)**.** *Let $\mathcal{M}$ be a Riemannian manifold with metric $\mathbf{g}$. The* gradient *of a function $f \in C^1(\mathcal{M})$ is a collection of tangent vectors $\operatorname{grad}_{\mathcal{M}} f(\xi)$, one at each point ξ, such that for all $\xi \in \mathcal{M}$ and all $v \in \mathcal{T}_\xi\mathcal{M}$*

$$\langle \operatorname{grad}_{\mathcal{M}} f(\xi), v\rangle_{\mathbf{g}} = \mathrm{d}f(v)|_\xi \; . \tag{2.3}$$

We can verify that when the Riemannian metric $\mathbf{g} = \mathbf{1}_{i=j}$, these operators degenerate to those in Euclidean space. When $\mathcal{M}$ is a $d-$dimensional manifold embedded in $\mathbb{R}^D$ with inherited metric. $\mathcal{T}_\xi\mathcal{M}$ can be identified as a $d-$dimensional linear subspace of $\mathcal{T}_\xi\mathbb{R}^D$, whose basis can be represented by an orthogonal $D \times d$ matrix $\mathbf{T}_\xi$. Let f be a smooth real-valued function defined on an open neighborhood of $\mathcal{M}$. There are two points of view for f when it is restricted on $\mathcal{M}$: (i) as a function on $\mathbb{R}^D$ and has gradient ∇f as usual. (ii) as a function on $\mathcal{M}$ and one can show that the gradient field $\text{grad}_\mathcal{M} f$ given by the coordinate representation $\text{grad} f := \mathbf{T}_\xi^\top \nabla f$ satisfies (2.3) [Lee, 2003].

More generally, consider a map $F = (f_1, \cdots, f_s) : \mathcal{M} \mapsto \mathbb{R}^s$. The differential $\mathrm{d}F = (\mathrm{d}f_1, \cdots, \mathrm{d}f_s)$ is then defined to be a linear mapping from $\mathcal{T}_\xi\mathcal{M} \to \mathcal{T}_\xi\mathbb{R}^s$. Under basis $\mathbf{T}_\xi$, a coordinate representation of $\mathrm{d}F$ is $\mathbf{T}_\xi^\top \nabla F$, where ∇F is a $D \times s$ matrix, constructed by column-wise stacking the gradients $\nabla f_1, \cdots, \nabla f_s$. The matrix ∇F is the transpose of the usually defined Jacobian matrix.

Another useful differential operator is the Laplace operator.

Definition 2.9 (Laplace operator in $\mathbb{R}^D$)**.** *The Laplace operator of a function $f \in C^2(\mathbb{R}^D)$ is defined by*

$$\Delta f = \sum_{i=1}^{D} \frac{\partial^2 f}{\partial x_D^2} \ .$$

We fix a point $\xi \in \mathcal{M}$ in the following definition and a local coordinate chart (U, φ). The generalization of the Laplace-Beltrami operator on manifold $\mathcal{M}$ is defined through the local coordinate function.

Definition 2.10 (Laplace-Beltrami operator on $\mathcal{M}$)**.** *Using local coordinate $\varphi(\xi) : (x_1(\xi), \cdots, x_d(\xi))$, Laplace-Beltrami operator is represented by*

$$\Delta_\mathcal{M} f = \frac{1}{\sqrt{\det \mathbf{g}}} \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\partial}{\partial x_i} \left(\sqrt{\det \mathbf{g}} \mathbf{g}_{ij}(\xi) \frac{\partial f}{\partial x_j} \right) \ .$$

Lastly, one can connect Laplace-Beltrami operator with Riemannian metric with a simpler representation through the following proposition.

Proposition 2.1. *Let $\mathcal{M}$ be the manifold with Riemannian metric $\mathbf{g}$. Under local coordinate $\varphi(\xi) : (x_1(\xi), \cdots, x_d(\xi))$, the ij-th entry of $\mathbf{h} = \mathbf{g}^{-1}$ is*

$$\mathbf{h}_{ij} = \frac{1}{2}\Delta_{\mathcal{M}}(x_i - x_i(\xi))(x_j - x_j(\xi))\big|_{x_i = x_i(\xi), x_j = x_j(\xi)} .$$

2.1.4 Regularity of manifolds

In the results of statistical estimation of manifolds, the regularity of a manifold also plays an important role in controlling the manifold's shape from being degenerated. These difficulties often come from the geometric features of a manifold. We provide examples of them in this section.

Curvature A direct difference between data in Euclidean space and on a manifold is that the space is now curved. Most manifold learning algorithms now solve this problem by assuming that the data lies locally near a linear subspace. This assumption is hard to satisfy if the curvature of $\mathcal{M}$ is large. Then this approximation can only be valid in a small neighborhood, which requires that the sample is dense enough.

Reach and injectivity radius The *reach* is a global descriptor that quantifies the regularity of a manifold. [Aamari et al., 2019] defines the reach $\tau_{\mathcal{M}}$ of a manifold $\mathcal{M} \subset \mathbb{R}^D$ to be the maximal distance from $\mathcal{M}$ of a point $\xi \notin \mathcal{M}$ for which the projection onto $\mathcal{M}$ is well defined (i.e., unique). For this reason, reach is also called (maximal) *injectivity radius*.

Equivalently, the injectivity radius can be defined imagining a closed ball $\bar{B}_r \subset \mathbb{R}^D$ of radius r, that "rolls" over $\mathcal{M}$, so that it touches $\mathcal{M}$ at a single point. $\tau_{\mathcal{M}}$ is the supremum over all values of r for which such a ball exists at all $\xi \in \mathcal{M}$. Intuitively, $\tau_{\mathcal{M}}$ is an "inverse curvature", hence larger $\tau_{\mathcal{M}}$ guarantees that the manifold curvature cannot be too large; it also guarantees that $\mathcal{M}$ cannot "fold back" near itself closer than $2\tau_{\mathcal{M}}$. A subspace has $\tau_{\mathcal{M}} = \infty$ (and zero curvature). Hence, in regions of large injectivity radius, one $\mathcal{M}$ can be estimated from fewer samples.

2.2 Manifold learning algorithms

2.2.1 Paradigms of manifold learning

In this section, we review classical manifold learning algorithms and some statistical results. Suppose we are given data $\{\xi_i\}_{i=1}^n$ where each data entry $\xi_i \in \mathbb{R}^D$. A key assumption for manifold learning is data are sampled from a distribution $\mathbb{P}$ that is supported close to a d dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$. Manifold learning algorithms $\hat{\phi}$ maps $\xi_i \in \mathbb{R}^D$ to $y_i \in \mathbb{R}^m$, where m is usually much smaller than D, but could be higher than the intrinsic dimension d.

Depending on the different output types, there are three different paradigms of manifold learning algorithms: local principal component analysis(PCA), principal curves and surfaces, and embedding algorithms.

Local PCA Local PCA considers dimension reduction for data points close to a point $\xi_i \in \mathcal{M}$. It estimates the tangent space $\mathcal{T}_\xi\mathcal{M}$ as a $d-$dimensional linear subspace in $\mathbb{R}^D$ and providing local coordinates of neighboring data point $\xi_{i'}$ by projecting the difference vector $\xi_{i'} - \xi_i$ onto this estimated tangent space. Therefore the output of local PCA is only a local coordinate since it will change when the reference point ξ_i varies.

Principal curves and surfaces PCS is not technically used for dimension reduction. Usually, it will map data point $\xi_i \in \mathbb{R}^D$ to $y_i \in \mathbb{R}^D$, but the image is constructed to lie on $\mathcal{M}'$, which is typically a curve (one-dimensional manifold) or a surface(a two-dimensional manifold). It was proposed to remove noise and generate a non-parametric summary of data as higher dimensional generalization of the median of one-dimensional distributions.

embedding is the major paradigm of manifold learning this dissertation discusses. These methods map all $\xi_i \in \mathbb{R}^D$ to $y_i \in \mathbb{R}^m$, where usually $m \ll D$. In the regime that $\mathbb{P}$ is supported exactly on $\mathcal{M}$, and sample size $n \to \infty$, a good manifold learning algorithm $\hat{\phi}$ should recover a smooth embedding function. This requires that the algorithm be guaranteed to recover the embedding as long as $\mathcal{M}$ is not singular (as will be defined later), regardless of the shape of $\mathcal{M}$.

Table 2.1 shows these three paradigms of manifold learning algorithms.

Table 2.1: Three different paradigms of manifold learning

name	mapping	notes
Local PCA	$\xi_i \in \mathcal{M} \subset \mathbb{R}^D \overset{\phi}{\rightarrow} y_i \in \mathcal{T}_{\xi_i}\mathcal{M} \in \mathbb{R}^d$	local coordinates only
Principal Curves and Surfaces	$\xi_i \in \mathcal{M} \subset \mathbb{R}^D \overset{\phi}{\rightarrow} y_i \in \mathcal{M}' \subset \mathbb{R}^D$	global coordinates
Embedding algorithms	$\xi_i \in \mathcal{M} \subset \mathbb{R}^D \overset{\phi}{\rightarrow} y_i \in \phi(\mathcal{M}) \subset \mathbb{R}^m, m \ll D$	global coordinates

2.2.2 Neighborhood graphs

Almost all manifold learning algorithms start with the same steps: constructing a neighborhood graph and computing a corresponding matrix.

Neighborhood graph Every data point ξ_i represents a node in this graph, and an edge connects two nodes if their corresponding data points are neighbors. In this thesis, we use $\mathcal{N}_i$ as the indices of ξ_i's neighbors and $k_i = |\mathcal{N}_i|$ be the number of neighbors of ξ_i. The matrix $\mathbf{N}_i \in \mathbb{R}^{k_i \times D}$ is the matrix with each row representing a neighbor of ξ_i, and $\mathbf{\Xi}_i$ denotes the local data set $\mathbf{\Xi}_i = (\xi_j - \xi_i)_{j \in \mathcal{N}_i}$. There are different ways to define neighbors. In a *radius-neighbor graph*, ξ_j is a neighbor of ξ_i iff $||\xi_i - \xi_j|| \leq r$. Here r is a parameter that controls the neighborhood scale, similar to a bandwidth parameter in kernel density estimation. Consistency of manifold learning algorithms is usually established assuming an appropriately selected neighborhood size [Bernstein et al., 2000, Coifman and Lafon, 2006, Singer and Wu, 2012]. In *$K-$nearest neighbor (K-NN)graph*, ξ_j is the neighbor of ξ_i iff ξ_j is among the closest K points to ξ_i. The K-NN graph has many computational advantages since it is regular (each node has exactly K neighbors, including itself) and connected for any $K > 1$. More software is available to construct (approximate) K-NN graphs fast for large data. But theoretically, it is much more difficult to establish any consistency of manifold learning algorithms.

Distance matrix Typically, the distances between neighbors matter, and they are stored in the distance matrix $\mathbf{A}$ defiined as

$$\mathbf{A}_{ij} = \begin{cases} ||\xi_i - \xi_j|| & \xi_j \in \mathcal{N}_i, \\ \infty & \text{otherwise .} \end{cases} \tag{2.4}$$

Similarity matrix In some algorithms, the neighborhood graph is represented by an $n \times n$ matrix of weights that are decreasing with distances. This is called the *similarity matrix*. The weights are given by a *kernel function* Belkin and Niyogi [2002], Coifman and Lafon [2006], Singer and Wu [2012], which is almost universally the Gaussian kernel, defined as

$$\mathbf{K}_{ij} := \begin{cases} \exp\left(-\frac{||\xi_i - \xi_j||^2}{h^2}\right), & \xi_j \in \mathcal{N}_i, \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

In the above, h, the kernel width, is another hyperparameter that must be tuned. Note that, even if $\mathcal{N}_i$ would trivially contain all the data, the similarity $\mathbf{K}_{ij}$ would be vanishingly small for far-away data points. Therefore, (2.5) effectively defines a radius-neighbor graph with $r \propto h$.

When k is selected to be an indicator function, i.e., the similarity matrix is defined as

$$\mathbf{K}_{ij} := \begin{cases} 1, & \xi_j \in \mathcal{N}_i, \\ 0, & \text{otherwise.} \end{cases} \tag{2.6}$$

Then $\mathbf{K}$ is the unweighted adjacency matrix of the neighborhood graph. By construction, the matrix $\mathbf{K}$ is usually a sparse matrix.

2.2.3 Examples of manifold embedding algorithms

Isomap Isomap [Bernstein et al., 2000] is a multidimensional scaling(MDS) generalization that recovers global coordinates from a pairwise squared Euclidean distance matrix. The idea is to generate global low-dimensional representations that preserve geodesic distance on the manifold. To achieve this, Isomap first uses the shortest path distance computed from pairwise distance matrix $\mathbf{A}$ to approximate geodesic distance on a manifold. Then invoke MDS to output low dimensional representations from the squared shortest path distance

matrix. Isomap comes with a consistency guarantee when the manifold $\mathcal{M}$ is flat and data space is convex. It can learn the manifold up to the choice of origin, rotation, and mirror imaging.

Laplacian eigenmaps Laplacian eigenmaps or diffusion maps are two closely related manifold embedding techniques. Laplacian eigenmaps first appeared in spectral clustering [Belkin and Niyogi, 2002] and was then proved to be deeply associated with the underlying geometry. The core step is to use eigenvectors of graph Laplacian of a neighborhood similarity matrix $\mathbf{K}$ as the low dimensional representation. Theoretically, diffusion maps or Laplacian eigenmaps are also appealing in multiple ways. When sample size $n \to \infty$, it is shown that Laplacian eigenmaps are related to the Laplace-Beltrami operator $\Delta_{\mathcal{M}}$, which plays an essential role in modern differential geometry. Let q be the sampling density on $\mathcal{M}$, then [Coifman and Lafon, 2006] showed that eigenfunctions of graph laplacian converge to those of the operator $\Delta_{\mathcal{M}} - \Delta_{\mathcal{M}} q/q$. Graph Laplacian can eliminate the bias term $\Delta_{\mathcal{M}} q/q$ and converge to the Laplace-Beltrami operator regardless of sampling density using a renormalization trick. The eigenfunctions of Laplace-Beltrami operators can constitute a valid smooth embedding of manifold $\mathcal{M}$.

Local tangent space alignment LTSA assumes that the observed data set $\mathcal{D}\{\xi_i\}_{i=1}^n$ is the image of a smooth function on low dimensional vectors. That is, there exists a function $\hat{\phi}$ and low dimensional representations $\{y_i\}_{i=1}^n$ with $y_i \in \mathbb{R}^m$ such that $\xi_i = F(y_i) + e_i$. Based on this, LTSA proposed a two-step procedure for recovering a global low dimensional representation: first, for each data, LTSA finds a local representation of its neighbors through projections on the tangent space; then it finds the best global low dimensional coordinates and local affine transformations so that the approximation has a minimal reconstruction error. This optimization problem can be smartly transformed into another eigenvalue problem. More details of LTSA can be found in [Zhang and Zha, 2004]. When the neighborhood radius used for local PCA is $O(r)$ and $\hat{\phi}$ is non-singular, it can be shown that the reconstruction error of LTSA is $O(||\mathbf{E}||+r^2)$, where $\mathbf{E}$ is the error matrix consisting of $[e_1, \cdots, e_n]$.

Local linear embedding (LLE) Local linear embedding [Roweis and Saul, 2000] is a heuristic method that utilizes the following assumption: locally a data point can be viewed as the weighted average of its neighbors, and this weights should be preserved for global coordinates. However, it is not yet known whether LLE provides valid manifold embedding.

2.3 Other geometric estimation problems and statistical results

This section discusses related geometric estimation problems and some statistical results. Suppose we are given $\mathcal{D} = \{\xi_i\}_{i=1}^n \in \mathbb{R}^D$. This section reviews methods to estimate other related geometric quantities given $\mathcal{D}$.

2.3.1 Tangent space

As $\mathcal{M}$ is embedded in $\mathbb{R}^D$, it is possible to estimate the tangent space at each point on $\mathcal{M}$ as a $D \times d$ orthogonal matrix, which forms the orthogonal basis of a $d-$dimensional linear subspace of $\mathbb{R}^D$.

Estimating tangent spaces at each point is usually achieved by local PCA procedure. If ξ_j is close to ξ, then $\xi_j - \xi$ is close to its projection on the tangent space (see e.g., [Bernstein et al., 2000, Coifman and Lafon, 2006, Singer and Wu, 2012]). In a local neighborhood of ξ, the difference vectors $\xi_j - \xi$ mainly lie on the tangent space centered at ξ. Hence it is possible to use the PCA procedure to approximate the tangent space at ξ as algorithm TangentSpaceBasis(Unweighted).

Algorithm 1 TangentSpaceBasis(Unweighted)

1: **Input:** A point $\xi_i \in \mathcal{D}$ and its neighbor $\{\xi_j\}_{j\in\mathcal{N}_i}$, embedding dimension m.
2: Construct difference data $\mathbf{\Xi}_i = [\xi_j - \xi_i]_{j\neq i} \in \mathbb{R}^{k_i \times D}$.
3: Perform PCA on the difference data $\mathbf{\Xi}$.
4: **Output:** Represent $\xi_j \in \mathcal{N}_i$ by $y_j = \mathrm{Proj}_{\mathbf{T}_i}(\xi_j - \xi_i) = \mathbf{T}_i^\top(\xi_j - \xi_i)$.

In step 2, PCA on $\mathbf{\Xi}_i$ is achieved by SVD of $\mathbf{\Xi}_i$ directly or by spectral decomposition of $\mathbf{\Xi}_i^\top \mathbf{\Xi}_i = \sum_{j=1}^n (\xi_j - \xi_i)(\xi_j - \xi_i)^\top \overline{\mathbf{1}}_{j\in\mathcal{N}_i}$.

It is often the case [Singer and Wu, 2012, Chen et al., 2013, Aamari and Levrard, 2018] that a weighted version of local PCA is used. When computing local covariance matrices, one may weight different points. These weights of each ξ_j in $\mathcal{N}_i$ can be proportional to the kernel function used to construct the similarity matrix $\mathbf{K}$ as in 2.5. Given these weights $\mathbf{K}_{ij}$ for ξ_js, the local weighted mean and weighted covariance at ξ_i can be estimated, and singular value decomposition is used to find the basis. This weighted version of the tangent space estimation algorithm is displayed in algorithm TangentSpaceBasisWeighted. In this algorithm $\overline{\mathbf{1}}$ means a vector with all entries being one.

Algorithm 2 TangentSpaceBasis(Weighted)

1: **Input:** Local dataset $\mathbf{\Xi}_i$, intrinsic dimension d, kernel parameter ϵ_N.
2: Compute local kernel weights $\mathbf{K}_{i,\mathcal{N}_i} = (\mathbf{K}_{ij})_{j\in\mathcal{N}_i} \in \mathbb{R}^{k_i}$.
3: Compute weighted mean $\bar{\xi}_i = (\mathbf{K}_{i,\mathcal{N}_i}^\top \overline{\mathbf{1}}_{k_i})^{-1}\mathbf{K}_{i,\mathcal{N}_i}^\top \mathbf{\Xi}_i$.
4: Compute weighted local difference matrix $\mathbf{Z}_i = (\mathbf{K}_{i,\mathcal{N}_i}^\top \overline{\mathbf{1}}_{k_i})^{-1} K_{i,\mathcal{N}_i}^\top .(\mathbf{\Xi}_i - \overline{\mathbf{1}}_{k_i}\bar{\xi}_i)$.
5: Compute $\mathbf{T}_i, \mathbf{\Lambda} \leftarrow \mathrm{SVD}(\mathbf{Z}_i^\top \mathbf{Z}_i, d)$.
6: **Output:** $\mathbf{T}_i$.

[Aamari and Levrard, 2018, 2019] studied the minimax rate of the tangent space estimation problem. For all C^k manifold with minimum reach $\tau_{\min} > 0$, we have

$$\inf_{\hat{\mathbf{T}}} \sup_{\mathcal{M}} \mathbb{E} \max_{1\le i\le n} \angle(\mathcal{T}_{\xi_i}\mathcal{M}, \hat{\mathbf{T}}_i) \asymp n^{-\frac{k-1}{d}} ,$$

where $\angle$ is the principal angle between two spaces. The faster rate when the manifold is smoother is given by local polynomial regression type approximation of the manifold instead of local PCA.

2.3.2 Laplace-Beltrami operator

Laplacian eigenmaps start from a neighborhood similarity matrix $\mathbf{K}$, which is defined as 2.5. Let $d_i = \sum_{j\in\mathcal{N}_i} \mathbf{A}_{ij}$ represent the degree and $\mathbf{D} = \mathrm{diag}\{d_1, \cdots, d_n\}$, then multiple choices of graph Laplacian exist [Ng et al., 2001]:

- Unnormalized Laplacian: $\mathbf{L}^{un} = \mathbf{D} - \mathbf{K}$

- Normalized Laplacian: $\mathbf{L}^{nor} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{K}\mathbf{D}^{-1/2}$

- Random-walk Laplacian: $\mathbf{L}^{rw} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{K}$

Laplacian eigenmaps or diffusion maps do not have the nice property to preserve geodesic distances. However, they are computationally less challenging when compared with Isomap. First, computing graph Laplacian only requires computing distance to neighbors. With the techniques that approximate nearest neighbors, computation complexity can be reduced to $O(n^{1+\gamma})$, where $\gamma < 1$ is a positive constant. Second, graph Laplacian is a sparse matrix. Hence multiple computation and storage tricks for sparse matrix can be adopted. Theoretically, there are some established consistency result of estimating Laplace-Beltrami operator. Readers can consult e.g.,[Belkin and Niyogi, 2007, Coifman and Lafon, 2006, Hein et al., 2005, 2007, Ting et al., 2010, Berry and Harlim, 2016].

2.3.3 Push-forward metric of embedding coordinates

For each $\hat{\phi}(\xi_i) : \mathcal{M} \to \mathbb{R}^m$, the pushforward Riemannian metric $\hat{\phi}_*\mathbf{g}$ expressed in the coordinates of $\mathbb{R}^m$ is a symmetric, semi-positive definite $m \times m$ matrix $\mathbf{G}_i$ of rank d. The scalar product $\langle u, v\rangle_{\mathbf{g}}$ takes the form $u^\top \mathbf{G}_i v$. The matrices $\mathbf{G}_i$ can be estimated by the algorithm RMetric of Perraul-Joncas and Meila [2013]. The algorithm uses only local information and thus can be run efficiently using the Laplacian, the neighborhood graph, and local embedding coordinate matrices. The computation is within the neighbors of each ξ_i. Recall that $\mathcal{N}_i$ is neighbors of ξ_i and $k_i = |\mathcal{N}_i|$ is the number of neighbors of ξ_i (including ξ_i itself).

Algorithm 3 RMetric

1: **Input:**Laplacian row $\mathbf{L}_{i,\mathcal{N}_i}$, embedding coordinates $y_i = \hat{\phi}(\xi_i)$, intrinsic dimension d.

2: Compute centered local embedding coordinates $\tilde{y}_i = y_i - \overline{\mathbf{1}}_{k_i} y_i^\top$.

3: Form matrix $\mathbf{H}_i$ by

$\mathbf{H}_i \leftarrow [\mathbf{H}_{i,k,k'}]_{k,k' \in 1:m}$ with $\mathbf{H}_{i,k,k'} = \sum_{i' \in \mathcal{N}_i} \mathbf{L}_{i,i'} \tilde{y}_{i,i',k} \tilde{y}_{i,i',k'}$ for $k, k' = 1 : m$.

4: Compute $\mathbf{V}_i, \mathbf{\Lambda}_i \leftarrow \mathrm{SVD}(\mathbf{H}_i, d)$.

5: $\mathbf{G}_i \leftarrow \mathbf{V}_i \mathbf{\Lambda}_i^{-1} \mathbf{V}_i^\top$.

6: **Output:** $\mathbf{G}_i, \mathbf{V}_i$.

Note that $\mathbf{G}_i$ is of rank d. The first d eigenvectors of $\mathbf{G}_i$ forms a orthonormal basis of tangent space $\mathcal{T}_{\hat{\phi}(\xi)} \hat{\phi}(\xi)$.

We can further estimate the gradient of each embedding function in the tangent space of $\mathcal{M}$. The high-level idea is to utilize the observation that the gradient of embedding functions in embedding coordinates have a simple expression that $\nabla \phi = \mathbf{I}_m$. When ξ_j is a neighbor of ξ_i, consider the tangent vector formed by project $\xi_j - \xi_i$ onto the tangent space $\mathcal{T}_{\xi_i} \mathcal{M}$. The differential $\mathrm{d}\hat{\phi}$ sends it approximately to the projection of difference of the embedded image $y_j - y_i$ onto tangent space $\mathcal{T}_{\hat{\phi}(\xi_i)} \hat{\phi}(\mathcal{M})$.

Hence

$$\langle \mathrm{Proj}_{\mathcal{T}_{\hat{\phi}(\xi)} \hat{\phi}(\mathcal{M})} (y_j - y_i), \mathrm{grad}_{\hat{\phi}(\mathcal{M})} \hat{\phi} \rangle \approx \langle \mathrm{Proj}_{\mathcal{T}_\xi \mathcal{M}} (\xi_j - \xi_i), \mathrm{grad}_{\mathcal{M}} \hat{\phi}(\xi_i) \rangle_{\mathbf{g}} \,, \tag{2.7}$$

Since $\mathrm{grad}_{\hat{\phi}(\mathcal{M})} \hat{\phi} = \hat{\mathbf{T}}_i^\top \mathbf{I}_m$, one can choose multiple neighbors of ξ_i and stack equations 2.7 for each neighbor together to form a linear system and obtain the estimation of $\mathrm{grad}_{\mathcal{M}} \hat{\phi}$ through least squares. The procedure is given in algorithm PullBackD.

Algorithm 4 PullBackD

Input:local data $\boldsymbol{\Xi}_i$, embedding coordinates $\mathbf{Y}_i = (\hat{\phi}(\xi_j))_{j \in \mathcal{N}_i}$, basis of tangent space $\mathbf{T}_i$), Laplacian rows $\mathbf{L}_{i,\mathcal{N}_i}$, intrinsic dimension d.

Compute pushforward metric eigendecomposition $\mathbf{G}_i, \hat{\mathbf{T}}_i \leftarrow \text{RMetric}(\mathbf{L}_{i,\mathcal{N}_i}, y_i, d)$.

Compute $\mathbf{B}_i \leftarrow (\mathbf{Y}_i^\top - y_i \overline{\mathbf{1}}_{k_i}^\top)$.

Compute $\mathbf{A}_i \leftarrow (\mathbf{B}_i)^\top (\boldsymbol{\Xi}_i^\top - \xi_i \overline{\mathbf{1}}_{k_i}^\top)$.

Output $\text{grad}_{\mathcal{M}} \hat{\phi} \leftarrow \mathbf{A}_i^\dagger \mathbf{B}_i^\top \hat{\mathbf{T}}_i \hat{\mathbf{T}}_i^\top$.

Chapter 3

DICTIONARY-BASED MANIFOLD LEARNING

In this chapter, we discuss our main contribution to manifold learning. As a motivation example, consider the unit circle $\mathbb{S}^1$ embedded in $\mathbb{R}^2$. A parametrization form of $\mathbb{S}^1$ is the image of

$$\begin{aligned} \alpha(\theta) : [0, 2\pi) &\to \quad \mathbb{R}^2 \\ \theta \qquad &\to (\cos\theta, \sin\theta) \, . \end{aligned} \tag{3.1}$$

Here each angle parameter θ uniquely determines a point on the circle. Using a parametrization form of a manifold simplifies the study of a circle in many problems since it reduces the number of variables that determines the system's state. The parametrization of a manifold may not be unique. In the circle example, $\lambda\theta$ for any $\lambda \in \mathbb{R} \setminus \{0\}$ is also a valid parametrization of the circle. Among all valid parametrizations, we prefer the minimal one (to be defined soon in Section 3.1) with simple forms and interpretable meanings.

Back in the scenario of manifold learning, we assume that data $\mathcal{D} = \{\xi_i\}_{i=1}^n$ are sampled from a d-dimensional connected smooth[1] submanifold $\mathcal{M}$ embedded in the Euclidean space $\mathbb{R}^D$, where typically $D \gg d$. Assume that the intrinsic dimension d is known. $\mathcal{M}$ has the Riemannian metric induced from $\mathbb{R}^D$.

To introduce domain knowledge in the manifold learning paradigm, we assume the existence of a dictionary $\mathcal{G} = \{g_1, \cdots, g_p\}$ where each dictionary function g_i is a smooth function from $\mathcal{M}$ to $\mathbb{R}$. We require that they are smooth on $\mathcal{M}$ (as a subset of $\mathbb{R}^D$), and have analytically computable gradients in $\mathbb{R}^D$. It is common practice in scientific research that scientists propose several related variables and tries to recover the most related ones to describe the object studied, i.e., the manifold data.

There are two high-level goals in this chapter:

---

[1] In this chapter, by *smooth* manifold or function we mean of class C^ℓ, $\ell \geq 2$.

- Find a subset of functions from the dictionaries such that these dictionary functions can parametrize the manifold.

- Given embedding coordinates from other standard manifold embedding algorithms, find a subset of dictionary functions that forms a minimal parametrization of the manifold that interprets given embedding.

As the motivational application of this chapter, we consider molecular dynamic simulation (MDS) data. In chemistry, a common problem is to discover so-called *collective coordinates* describing the evolution of molecular configurations at long time scales, which correspond to macroscopically interesting transformations of the molecule, and can explain some of its properties [Clementi et al., 2000, Noé and Clementi, 2017]. The molecular configuration is represented by the $3N_a$ vector of spatial locations of the N_a atoms comprising the molecule. A *molecular dynamics (MD) simulation* produces a sample of molecular configurations; the distribution of this sample describes the molecule's behavior in the given experimental conditions. It has been shown empirically that manifolds approximate these high-dimensional distributions [Dsilva et al., 2013]. Figure 3.1a shows the toluene molecule, consisting of $N_a = 15$ atoms, and 3.1d shows the mapping of an MD simulated trajectory into $m = 2$ dimensions (the embedding coordinates) by a manifold learning algorithm. Visual inspection shows that this configuration space is well-approximated by a one-dimensional manifold parametrized by a geometric quantity, the *torsion* g_1 of the methyl bond, which is the angle formed by the planes inscribing the first three and last three atoms of the orange lines joining four atoms in Figure 3.1d. Thus, g_1 is a collective coordinate which explains the large scale data manifold by the rotation of the CH_3 methyl group relative to the plane of the other carbon atoms, filtered out from the faster modes of vibration by the manifold learning algorithm. Similarly, as shown in Figures 3.1e, 3.1h 3.1f, and 3.1i, the large scale geometry of the ethanol and malonaldehyde MD data is explained by two torsion angles each.

In Section 3.1, we will make the previous two targets mathematically rigorous. We present the rank conditions of existence of a minimal parametrization in a given dictionary. This serves as the mathematical foundation of the algorithms presented in the following

sections. In Section 3.2, we present algorithm TSLasso that directly finds a subset of dictionary functions to parametrize the manifold. Meanwhile in section 3.3, we present algorithm ManifoldLasso that finds dictionary functions that are most related to the existing embedding coordinates. In section 3.5, we discuss theoretical properties of the proposed algorithms. Section 3.6 shows experimental results on synthetic and molecular dynamics datasets. Section 3.7 discusses related work and concludes this chapter.

3.1 Parametrization and interpretations

3.1.1 Definitions

As a generalization to the example of $\mathbb{S}^1$, we define the parametrization of a manifold as follows:

Definition 3.1 (parametrization)**.** *A parametrization of a $d-$ dimensional manifold $\mathcal{M}$ is a mapping $F : \mathcal{M} \to \mathbb{R}^m$ such that:*

1. *$F : \mathcal{M} \to F(\mathcal{M})$ is injective.*

2. *F is smooth (at least C^1) and locally a diffeomorphism almost everywhere[2] on $\mathcal{M}$.*

Example: $\mathbb{S}^1$ Consider the function defined for $(x, y) : |x^2 + y^2 - 1| \leq 1/2$, then

$$F : (x, y) \mapsto \begin{cases} \arcsin \frac{y}{\sqrt{x^2+y^2}} & x \geq 0 \ ; \\ \pi - \arcsin \frac{y}{\sqrt{x^2+y^2}}, & x < 0 \ , \end{cases} \tag{3.2}$$

is a parametrization for $\mathcal{M}$.

Example: Embeddings By definition, if F is an embedding of $\mathcal{M}$ to $\mathbb{R}^m$, then F is a parametrization as in definition 3.1.

Before we proceed, we make several remarks here for our definition of parametrization since modifications to standard definitions in differential geometry are made.

---

[2] with respect to Hausdorff measure

First, in differential geometry terminology, $(U \subseteq \mathcal{M}, \varphi)$ locally is a *coordinate chart* for $\mathcal{M}$ and φ^{-1} is called a *parameterization* of U. Here we refer to φ as the 'parameterization', as φ, φ^{-1} are diffeomorphisms and are both representative. We argue that φ is of more immediate interest since in our case, this map will consist of interpretable and analytically computable dictionary functions, and φ^{-1}, while guaranteed to exist on $\varphi(U)$, is defined only implicitly in many scenarios.

Second, since a manifold may require multiple charts, we relax the requirement that F is locally a diffeomorphism everywhere to *almost everywhere*. In the circle example, since the manifold $\mathcal{M}$ is compact, it is impossible to find a single smooth function that can locally be a diffeomorphism everywhere. This relaxation allows us to find d functions parametrizing a $d-$dimensional compact manifold in our definition.

Given two different parametrizations of the same manifold, we could establish a comparison relation through the following definition.

Definition 3.2 (Interpretation)**.** *For two parametrizations $F_1 : \mathcal{M} \to \mathbb{R}^{m_1}$ and $F_2 : \mathcal{M} \to \mathbb{R}^{m_2}$, we refer to F_1 as a smaller interpretation of F_2 if $m_1 \leq m_2$ and there exists a function τ that is almost everywhere C^1 in $\mathbb{R}^{m_1}$ such that $F_2 = \tau \circ F_1$ for all $\xi \in \mathcal{M}$.*

3.1.2 Rank conditions

The global rank theorem in Section 2.1 shows that we can characterize the global properties of a smooth map with its global property, such as rank. In this section, we show that similar conditions based on rank can be obtained for our definition 3.1. We summarize them as follows.

Proposition 3.1 (Existence of parametrization)**.** *Let $\mathcal{M}$ be a $d-$dimensional manifold. Then a smooth (at least C^1) injective map F is a parametrization as in definition 3.1 iff* $\operatorname{rank} F = d$ *almost everywhere on $\mathcal{M}$.*

The next proposition shows that, in fact, the function composition condition in the definition of *smaller parametrization* is redundant. Therefore, comparing two parametrizations on the set of parametrizations is a partial order relation. Given a set of parametrizations,

we can define a *minimal* parametrization—also, a minimal parametrization *interprets* all other parametrizations.

Proposition 3.2 (Criterion of interpretation)**.** *Let $\mathcal{M}$ be a $d-$dimensional manifold and $F_1 : \mathcal{M} \to \mathbb{R}^{m_1}$ is a parametrization. Then another parametrization $F_2 : \mathcal{M} \to \mathbb{R}^{m_2}$ is a smaller parametrization to F iff $m_2 \leq m_1$.*

3.2 TSLasso*: Parametrization of manifold with physical meanings*

Given the dictionary $\mathcal{G} = \{g_1, \cdots, g_p\}$, our first task is to find a subset of dictionary functions that form a (minimal) parametrization of $\mathcal{M}$. Starting this section, we denote S to be a subset of indices $\{1, \cdots, p\}$. g_S represents the mapping consisting of $g_i, i \in S$. From the previous section, g_S is a parametrization iff $\operatorname{rank} g_S = d$ a.e. on $\mathcal{M}$. Denote $\mathbf{X}_{\xi,S} \in \mathbb{R}^{d\times|S|}$ to be the matrix representation of the linear map $\mathrm{d}g_S$ (recall definition in 2.1). It holds that $\operatorname{rank} \mathbf{X}_{\xi,S} = d$. Therefore, there exists some matrices $\mathbf{B}_{\xi,S} \in \mathbb{R}^{|S|\times d}$ such that for a.e. $\xi \in \mathcal{M}$

$$\mathbf{I}_d = \mathbf{X}_{\xi,S}\mathbf{B}_{\xi,S} \, . \tag{3.3}$$

The idea of the TSLasso algorithm is then to express the orthonormal bases $\mathbf{T}_\xi \in \mathbb{R}^{D\times d}$ of the manifold tangent spaces $\mathcal{T}_\xi\mathcal{M}$ as sparse linear combinations of dictionary function gradient vector fields. This simplifies the non-linear problem of selecting a best functional approximation to $\mathcal{M}$ to the linear problem of selecting best local approximations in the tangent bundle.

For notation simplicity, we will write $\mathbf{X}_{iS}, \mathbf{B}_{iS}, \mathcal{T}_i\mathcal{M}$ as the corresponding quantities at point ξ_i when we are discussing finite sample. We can select $S = [p]$, and simplify the notation of $\mathbf{X}_{iS}, \mathbf{B}_{iS}$ to $\mathbf{X}_i \in \mathbb{R}^{d\times p}, \mathbf{B}_i \in \mathbb{R}^{p\times d}$, but crucially, if we do not have colinear gradients, then we can restrict all but $|S|$ rows of $\mathbf{B}_i$ to be zeros. We can also select $s = \{j\}$, and define $\mathbf{B}_{\cdot j} \in \mathbb{R}^{nd}$ as the vector formed by concatenating $\mathbf{B}_{i\{j\}}$. Stacking $\mathbf{B}_{\cdot j}$ together forms $\mathbf{B} \in \mathbb{R}^{p\times nd}$.

We now seek a subset $S \subset [p]$ such that (1) only the corresponding $n|S|$ vectors $\mathbf{B}_{\cdot j} : j \in S$ have non-zero entries and (2) each submatrix $\mathbf{X}_{iS}$ forms a rank d matrix. The previous

observation inspires minimizing Frobenius norm $\mathbf{I}_d - \mathbf{X}_i\mathbf{B}_i$ with joint sparsity constraints over rows of $\mathbf{B}_i$. This sparsity is also induced jointly over all data points.

$$J_{\lambda_n}(\mathbf{B}) = \frac{1}{2}\sum_{i=1}^{n}||\mathbf{I}_d - \mathbf{X}_i\mathbf{B}_i||_F^2 + \frac{\lambda_n}{\sqrt{dn}}\sum_{j=1}^{p}||\mathbf{B}_{\cdot j}||_2. \tag{3.4}$$

Note that this optimization problem is a variant of Group Lasso [Yuan and Lin, 2006] that forces group of coefficients of size dn to be zero simultaneously in the regularization path. It can be shown this loss function is invariant to local tangent space rotation.

We present the full TSLasso approach in algorithm TSLasso. Following the above logic, we transform our non-linear manifold parameterization support recovery problem into a collection of sparse linear problems in which we express coordinates of individual tangent spaces as linear combinations of gradients of functions from our dictionary. Tangent spaces at each point are estimated in step 10, enabling utilizing gradients of dictionary functions in $\mathcal{T}_\xi\mathcal{M}$ by projecting the gradient $\nabla g_j(\xi_i) \in \mathbb{R}^D$ on to estimated tangent spaces $\mathbf{T}_i$. Finally we input these gradients into objective function (3.4) to solve for the support.

3.3 ManifoldLasso*: Explain existing embedding coordinates from dictionary functions*

In this section we further consider the problem of finding a interpretation of a given embedding coordinate ϕ. We follow the notations used in the previous section: the dictionary being $\mathcal{G} = \{g_1, \cdots, g_p\}$, denote S to be a subset of indices $\{1, \cdots, p\}$, g_S represents the mapping consisting of $g_i, i \in S$.

Recall that given the definition in section 3.1, an interpretation parametrization must also have rank d and further there is a smooth a.e. function τ such that $\phi = \tau \circ g_S$.

The main idea of our approach is to exploit the chain rule $\mathrm{d}F = \mathrm{d}\tau \circ \mathrm{d}g_S$. Each differential can be written in terms of gradients $\operatorname{grad}_{\mathcal{M}} \phi_{1:m}$ and $\operatorname{grad}_{\mathcal{M}} g_{1:p}$. Identifying the tangent vector $\operatorname{grad}_{\mathcal{M}} \phi_{1:m}, \operatorname{grad}_{\mathcal{M}} g_{1:p}$ with their coordinate representation in a basis of $\mathcal{T}_\xi\mathcal{M}$, chain rule essentially confirms that $\operatorname{grad}_{\mathcal{M}} \phi_{1:m}$ is a sparse linear combination of $\operatorname{grad}_{\mathcal{M}} g_{1:p}$. Following the same notation in section 3.2, still denote $\mathbf{X}_{\xi,S} \in \mathbb{R}^{d\times|S|}$ as the matrix representation of the linear map $\mathrm{d}g_S$. Further, use $\mathbf{Y}_{\xi,1:m} \in \mathbb{R}^{d\times m}$ to represent the

Algorithm 5 TSLasso

Input: Dataset $\mathcal{D}$, dictionary $\mathcal{F}$, intrinsic dimension d, regularization parameter λ_n, radius parameter r_n, kernel parameter h_n.
for $j = 1, 2, \dots p$ **do**
  Compute $\nabla g_j(\xi_i)$ for $i = 1, \dots n$.
  Compute ζ_j^2 by (3.8) and normalize $\nabla g_j(\xi_i) \leftarrow (1/\zeta_j)\nabla g_j(\xi_i)$ for $i = 1, \dots n$.
end for
for $i = 1, 2, \dots n$ (or subset $I \subset [n]$) **do**
  Compute $\mathcal{N}_i$ and $\mathbf{\Xi}_i$ using $\mathcal{D}, r_n$.
  Compute the orthonormal tangent space basis $\mathbf{T}_i \leftarrow$TangentSpaceBasis$(\mathbf{\Xi}_i, d, h_n)$.
  Compute $\nabla g_j(\xi_i)$ for $j \in [p]$.
  Project onto tangent space $\mathbf{X}_i = \mathbf{T}_i^\top [\nabla g_j(\xi)]_{j\in[p]}$.
end for
Solve for $\mathbf{B}$ by minimizing $J_{\lambda_n}(\mathbf{B})$ in (3.4).
Output: $S(\mathbf{B}) = \{j \in [p] : ||\mathbf{B}_{\cdot j}||_2 > 0\}$.

matrix representation of $\mathrm{d}\phi$. Then there exist matrices $\mathbf{B}_{\xi,S} \in \mathbb{R}^{|S|\times m}$ such that for a.e. $\xi \in \mathcal{M}$[3],

$$\mathbf{Y}_{\xi,1:m} = \mathbf{X}_{\xi,S}\mathbf{B}_{\xi,S} \ . \tag{3.5}$$

Therefore, similar to TSLasso , it is natural to use a group lasso based algorithm again to select the support set. Using the same notation that $\mathbf{X}_{iS}, \mathbf{B}_{iS}, \mathbf{Y}_i$ denotes $\mathbf{X}_{\xi_i,S}, \mathbf{B}_{\xi_i,S}, \mathbf{Y}_{\xi,1:m}$ and simplify $\mathbf{X}_{i[p]} \in \mathbb{R}^{d\times p}, \mathbf{B}_{i[p]} \in \mathbb{R}^{p\times m}, \mathbf{Y}_{i,1:m} \in \mathbb{R}^{d\times m}$ to be $\mathbf{X}_i, \mathbf{B}_i, \mathbf{Y}_i$. Again, by merely restricting all but $|S|$ rows of $\mathbf{B}_i$ to be all zeros, (3.5) also holds. Denote $\mathbf{B}_{\cdot j} \in \mathbb{R}^{nm}$ as the vector formed by concatenating all $\mathbf{B}_{i\{j\}}, j \in [p]$. Row-wise stacking $\mathbf{B}_{\cdot j}$ forms $\mathbf{B} \in \mathbb{R}^{p\times nm}$. Then the previous idea can be transferred to minimizing

[3]Note that in our notation, $\mathbf{X}_\xi, \mathbf{Y}_\xi$ are the transpose of a Jacobian matrix, hence after applying the chain rule we obtain $\mathbf{X}_\xi\mathbf{B}_\xi$ instead of $\mathbf{B}_\xi\mathbf{X}_\xi$.

$$J_{\lambda_n}(\mathbf{B}) = \frac{1}{2}\sum_{i=1}^{n}||\mathbf{I}_d - \mathbf{X}_i\mathbf{B}_i||_F^2 + \frac{\lambda_n}{\sqrt{dn}}\sum_{j=1}^{p}||\mathbf{B}_{\cdot j}||_2 \ . \tag{3.6}$$

The ManifoldLasso algorithm, the main algorithm of this paper, implements this idea. It takes as input data $\mathcal{D}$ sampled from an unknown manifold $\mathcal{M}$, a dictionary $\mathcal{F}$ of functions defined on $\mathcal{M}$ (or alternatively on an open subset of the ambient space $\mathbb{R}^D$ that contains $\mathcal{M}$), and an embedding $\phi(\mathcal{D})$ in $\mathbb{R}^m$. The output of ManifoldLasso is a set S of indices in $\mathcal{F}$, representing the functions in $\mathcal{F}$ that explain $\mathcal{M}$.

The first part of the algorithm contains preparatory steps for geometric analysis covered in previous chapter. Steps 2 and **??** construct the neighborhood graph and the Laplacian matrix used for manifold learning and tangent space estimation.

The second part of ManifoldLasso calculates the necessary gradients; this comprises Steps 10–12. In Step 10, we estimate orthogonal bases of tangent subspaces by the TangentSpaceBasis algorithm. The gradients of the dictionary w.r.t. the manifold are then obtained as columns of the $d \times p$ matrix $\mathbf{X}_i$ in Steps 6, 7, and 11. These operations takes advantage of existing work reviewed in 3.4. In Step 12, the gradients at ξ_i of the coordinates $\phi_{1:m}$, also w.r.t. $\mathcal{M}$, are calculated as columns of the $d \times m$ matrix $\mathbf{Y}_i$ by the PullBackD algorithm described in Section 3.4.

In the last part of ManifoldLasso, Step 15 finds the support S by solving the sparse regression. A group lasso algorithm is called to perform the sparse regression of the manifold coordinates' gradients $\mathbf{Y}_i$ on the gradients of the dictionary functions, represented by $\mathbf{X}_i$. The indices of those dictionary functions whose $\mathbf{B}_i$ coefficients are not identically null represent the row-wise support set $\operatorname{supp}\mathbf{B}$. Scaling of functions is addressed through normalization in Steps 7 and 14; this procedure is described in more detail in Section 3.4.

There are several optional steps and substitutions in our algorithm. An embedding can be computed in Step 4, or input separately by the user - we denote this step generically as EmbeddingAlg. Finally, although we explicitly describe tangent space estimation methods of both $\mathcal{T}_{\xi_i}\mathcal{M}$ and $\mathcal{T}_{\phi(\xi_i)}\phi(\mathcal{M})$ in our algorithms, other approaches to estimate them may be used.

Algorithm 6 ManifoldLasso

Input: Dataset $\mathcal{D}$, dictionary $\mathcal{G}$, embedded coordinates $\hat{\phi}(\mathcal{D})$, intrinsic dimension d, kernel bandwidth h_n, neighborhood cutoff size r_n, regularization parameter λ.
Construct $\mathcal{N}_i$ for $i = 1 : n$; $i' \in \mathcal{N}_i$ iff $||\xi_{i'} - \xi_i|| \leq r_n$, and local data matrices $\mathbf{\Xi}_{1:n}$
Construct kernel matrix $\mathbf{K}$ and Laplacian matrix $\mathbf{L}$ as 2.2.
[Optionally compute embedding: $\hat{\phi}(\xi_{1:n}) \leftarrow$EmbeddingAlg$(\mathcal{D}, \mathcal{N}_{1:n}, m, \ldots)$.]
for $j = 1, 2, \ldots p$ **do**
    Compute $\nabla g_j(\xi_i)$ for $i = 1, \ldots n$.
    Compute ζ_j^2 by (3.8) and normalize $\nabla g_j(\xi_i) \leftarrow (1/\zeta_j)\nabla g_j(\xi_i)$ for $i = 1, \ldots n$.
end for
for $i = 1, 2, \ldots n$ **do**
    Compute basis $\mathbf{T}_i \leftarrow$TangentSpaceBasis$(\mathbf{\Xi}_i, \mathbf{K}_{i,\mathcal{N}_i}, d)$.
    Project $\mathbf{X}_i \leftarrow (\mathbf{T}_i)^\top \nabla g_{1:p}$.
    Compute $\mathbf{Y}_i \leftarrow$PullBackD$(\mathbf{\Xi}_i, \Phi_i, \mathbf{T}_i, \mathbf{L}_{i,\mathcal{N}_i}, d)$.
end for
Compute $\zeta_k^2 \leftarrow \frac{1}{n}\sum_{i=1}^n \|\mathbf{Y}_{ik}\|^2$. (i.e. (3.7)), for $k = 1, \ldots m$ and normalize $\mathbf{Y}_i \leftarrow \mathbf{Y}_i \,\mathrm{diag}\{1/\zeta_{1:m}\}$, for $i = 1, \ldots n$.
$\mathbf{B} \leftarrow$ Goup Lasso problem (3.6).
Output $S(\mathbf{B}) = \{j \in [p] : ||\mathbf{B}_{\cdot j}||_2 > 0\}$.

3.4 Other considerations

Normalization As with many sparse regression methods, normalization is necessary to balance the relative influence of dictionary elements and embeddings coordinates. Multiplying g_j by a non-zero constant and dividing its corresponding $\mathbf{B}_{\cdot j}$ by the same constant leaves the reconstruction error of all y's invariant, but affects the norm $||\mathbf{B}_{\cdot j}||$. Therefore, the relative scaling of the dictionary functions g_j can influence the recovered support S, by favoring the dictionary functions whose columns have larger norm. A similar effect is present if a particular embedding coordinate ϕ_k is rescaled by a constant. For example, multiplying a certain ϕ_k by a number close to zero will cause the penalty accrued by learned coefficients

for that coordinate to be smaller than for the other coefficients, and for that ϕ_k to dominate support recovery.

We therefore normalize all $\text{grad}_{\hat{\phi}(\mathcal{M})}\, \phi_{1:m}$ and $\text{grad}_{\mathcal{M}}\, g_{1:p}$ as follows. Denote f a function on $\mathcal{M}$, which can be either a coordinate function or a dictionary function. When f is defined on $\mathcal{M}$, but not outside $\mathcal{M}$, we calculate the *normalizing constant*

$$\zeta^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} \|\,\text{grad}_{\mathcal{M}}\, f(\xi_i)\|^2; \tag{3.7}$$

then we set $f \leftarrow f/\zeta$. The above ζ is the finite sample version of $\|\,\text{grad}_{\mathcal{M}}\, f\|_{L_2(\mathcal{M})}$, integrated w.r.t. the data density on $\mathcal{M}$. We apply this normalization to coordinate functions ϕ_k, but it could also be applied to functions g_j when they are defined only on $\mathcal{M}$. A similar approach was used in Haufe et al. [2009].

When function f is defined on a neighborhood around $\mathcal{M}$ in $\mathbb{R}^D$, we compute the normalizing constant with respect to ∇f. That is,

$$\zeta^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(\xi_i)\|^2. \tag{3.8}$$

Then, once again, we set $f \leftarrow f/\zeta$. We apply this normalization to our dictionary functions g_j. This approximates normalization by $||\nabla g_j||_{L_2(\mathcal{M})}$. Since $|\nabla g_j(\xi_i)|^2 = |\text{grad}_{\mathcal{M}}\, g_j(\xi_i)|^2 + |\nabla g_j^{\perp}(\xi_i)|^2$, where $\nabla g_j^{\perp}$ denotes the component of ∇g_j orthogonal to $\mathcal{M}$, normalization prior to projection penalizes functions with large $\nabla g_j^{\perp}$ and favors functions whose gradients are more parallel to the tangent space of $\mathcal{M}$. Note that, in the high-dimensional setting, we expect random functions to have gradient perpendicular to $\mathcal{T}\mathcal{M}$, and so these will be penalized by our normalization strategy.

Tuning Tuning parameters are often selected by cross-validation in Lasso-type problems. However, in our setting, the recovered support generally span the tangent space, and as discussed in Section 3.5, we are theoretically motivated to identify a size d support. Since the cardinality of the support decreases as the tuning parameter λ is increased, we base our choice of λ on matching the cardinality of the support to d. Sufficient conditions for this estimation strategy are given in Section 3.5. To identify this λ, which we call λ_0, we perform a simple binary search over λ in the range $[0, \lambda_{\max}]$ where $\lambda_{\max}$, the theoretical

maximum λ value, is $\lambda_{\max} = \max_j(\sum_{i=1}^n (\| \operatorname{grad}_{\mathcal{M}} g_j(\xi_i))\|_2^2)^{1/2}$ for TSLasso and $\lambda_{\max} = \max_j(\sum_{i=1}^n \sum_{k=1}^m (\operatorname{grad}_{\mathcal{M}} g_j(\xi_i))^\top (\operatorname{grad}_{\mathcal{M}} \phi_k(\xi_i)))^{1/2}$ in ManifoldLasso

3.5 Support Recovery Guarantee

In this section, we discuss the behavior of our proposed algorithms theoretically. When minimal parametrization exists and is unique, we provide sufficient conditions so that TSLasso correctly selects this group with high probability w.r.t. sampling on the manifold and this probability converges to one if sample size tends to infinity. Meanwhile, it is more difficult to generate end-to-end support recovery result for ManifoldLasso , as it is almost impossible to perform probabilistic analysis of an arbitrary embedding algorithm. However, we are still able to discuss conditions for successful support recovery given any embedding coordinates by repeating very similar arguments presented in this section. Hence those theoretical results of ManifoldLasso are omitted.

Assumption 3.1. *Throughout this section, we assume the followings to be true.*

1. *$\mathcal{M}$ is a d-dimensional $C^\ell, \ell \geq 1$ compact manifold with reach $\tau > 0$ embedded in $\mathbb{R}^D$ with inherited Euclidean metric.*

2. *Data $\{\xi_i\}_{i=1}^n$ are sampled from some probability measure P on the manifold that has a Radon-Nikodym derivative $\pi(\xi)$ with respect to the Hausdorff measure. There exist two positive constants $\pi_{\min}, \pi_{\max}$ such that $0 < \pi_{\min} \leq \pi(\xi) \leq \pi_{\max}$ for all $\xi \in \mathcal{M}$.*

3. *Dictionary $\mathcal{F} = \{g_j(\xi) : j \in [p]\}$ contains p C^1 functions defined on a neighborhood of $\mathcal{M}$ in $\mathbb{R}^D$. Further assume that $\delta := \inf_{\xi \in \mathcal{M}} \min_{j \in [p]} ||\nabla g_j(\xi)|| > 0$ and denote $\Gamma := \sup_{\xi \in \mathcal{M}} \max_{j \in [p]} ||\nabla g_j(\xi_i)||$.*

4. *$S \subset [p], |S| = d$ is the only subset such that $\operatorname{rank} g_S = d$ a.e. on $\mathcal{M}$ w.r.t. Hausdorff measure.*

Assumption 1 on manifold and 2 on sampling are common in the manifold estimation literature (e.g. Aamari and Levrard [2018]). The positive reach in 1 will avoid extreme

curvature and bizarre behavior of the manifold, and the assumption 2 on the density enforces the uniformity of sampling. Assumption 3 restricts the smoothness of all dictionary functions and ensures that all dictionary functions do not have critical points on $\mathcal{M}$ as a function on $\mathbb{R}^D$. One should also notice that $\Gamma < \infty$ by the compactness assumption of $\mathcal{M}$.

Now we are ready to prove support recovery consistency under suitable conditions. Let $\hat{\mathbf{B}}$ be the solution of problem (3.4) and $S(\hat{\mathbf{B}})$ be the nonzero rows of $\hat{\mathbf{B}}$. We will show that the probability of $S(\hat{\mathbf{B}}) = S$ converges to 1 as n increases. We start by defining

$$b_S = \inf_{\xi:\text{rank d}g_S(\xi)=d} \min_{j\in S} ||\mathbf{B}_{\xi,\{j\}}||_2 \ . \tag{3.9}$$

Larger b_S is an indicator of higher strength of signal. Further consider the matrix $\tilde{\mathbf{X}}_\xi$ whose j-th column is $\mathbf{X}_{\xi,\cdot j}/||\nabla g_j(\xi)||$. Correspondingly we can define $\tilde{\mathbf{X}}_{\xi,S}$ as the submatrix of $\tilde{\mathbf{X}}$ with columns in S. Let $\mathbf{G}_{\xi,S} = \text{diag}\{||\nabla g_j(\xi)||\}_{j\in S}$ and define

$$\mu_S = \sup_{\xi\in\mathcal{M}, j\in S, j'\notin S} |\tilde{\mathbf{X}}_{\xi,\cdot j}^\top \tilde{\mathbf{X}}_{\xi,\cdot j'}| \ , \tag{3.10}$$

$$\nu_S = \sup_{\xi\in\mathcal{M}} ||(\tilde{\mathbf{X}}_{\xi,S}^\top \tilde{\mathbf{X}}_{\xi,S})^{-1} - \mathbf{G}_{\xi,S}^2|| \ . \tag{3.11}$$

Here ν_S is finite if $\mu_S < 1/(d-1)$, guaranteed by the Gershgorin circle theorem. The parameter μ_S can be thought of as a renormalized incoherence between the functions in S and those not in S; ν_S is a internal colinearity parameter, which is small when the columns of $\mathbf{X}_S(\xi)$ are closer to orthogonality and the gradient of functions in S are more parallel to the tangent space. We also define

$$\phi_S = \sup_{\xi\in\mathcal{M}} \max_{j\in S} ||\nabla g_j(\xi)||_2 \ , \tag{3.12}$$

which upper bounds the Euclidean gradient of functions in S.

Theorem 3.1. *Suppose Assumptions 3.1 hold. In algorithm 5, suppose tangent spaces are estimated by WL-PCA in using Gaussian kernel and bandwidth parameter choice* $\epsilon_n = r_n = C((\log n/(n-1))^{1/d})$ *with large enough constant* C*, and normalization on dictionary is performed as described in 3.4. If* $(1+\nu_S/\delta^2)\mu_S\phi_S\Gamma d < 1$ *and* $\lambda_n(1+\nu_S/\delta^2) < b_S\sqrt{n}/2$*, then there is a constant* N *depending only on* $\mathcal{M}, \pi_{\min}, \pi_{\max}$ *such that when* $n > N$*, it holds that*

$$Pr(S(\widehat{\mathbf{B}}) = S) \geq 1 - 4(\frac{1}{n})^{\frac{2}{d}} \ . \tag{3.13}$$

The proof is contained in the supplementary material. The main idea is first to find a sufficient condition so that given correct gradient of each function TSLasso can find the correct support, assuming correct estimation of the tangent space. Then we consider this condition in the case where gradient is estimated from data and obtain the guarantee by the fact that tangent spaces can be consistently estimated with larger sample size.

There are some differences to be noted of this recovery result compared with classical recovery guarantees in Group Lasso type problems in e.g. Wainwright [2009], Obozinski et al. [2011], Elyaderani et al. [2017]. First, we cannot adopt directly the usual assumption in Lasso literature that each column of $\mathbf{X}$ has unit norm, considering the normalization in Section 3.4. Also, the asymptotic regime we are considering here is only $n \to \infty$. Although we are using a Group Lasso type optimization problem, the dimension p is fixed since we only consider the fixed dictionary. There is no other conditions between p and n in our result, as required in many literature. Third, the noise structure is not the same as a general Group Lasso problem since the source of noise is estimation of tangent space. Since we are sampling *on the manifold*, there is no noise level parameter that appears in standard Lasso literature. In a simulation experiment, we also explore the behavior of our method on noisy settings.

3.6 Experiments

We illustrate the behavior of TSLasso and ManifoldLasso on both synthetic and real data. Our synthetic data sets include a swiss roll in $\mathbb{R}^{49}$ and a rigid ethanol data in $\mathbb{R}^{50}$ and our real datasets are data molecular dynamics simulation (MDS) for three different molecules (Ethanol , Malonaldehyde and Toluene). Experiments were performed in Python on a 16 core Linux Debian Cluster with 768 gigabytes of RAM. Code is available at `https://github.com/sjkoelle/montlake` and implementation is completed by collaborator Samson Koelle.

For all of the experiments, the data consist of n data points in D dimensions. TSLasso and ManifoldLasso are applied to a uniformly random subset of size $n' = |\mathcal{I}|$ using p dictio-

nary functions, and this process is repeated ω number of times. Note that the entire data set is used for tangent space estimation. In our experiments, the intrinsic dimension d is assumed known, but could be estimated by a method such as in Levina and Bickel [2004]. The local tangent space kernel bandwidth h_n is estimated using the algorithm of Joncas et al. [2017] for molecular dynamics data. Parameters are summarized in Table 3.1. More details on the MDS data preprocessing can be found in the appendix of this chapter 3.8.4.

3.6.1 Validation of TSLasso

Results on Swiss Roll synthetic data We begin our experimental study by demonstrating that TSLasso is invariant to the choice of embedding algorithm on the classic unpunctured SwissRoll dataset. This dataset consists of points sampled from a two dimensional rectangle and rolled up along one of the two axes aFigure 3.2a shows the SwissRoll dataset in $\mathbb{R}^3$, then randomly rotated in $D = 49$ dimensions.

The dictionary $\mathcal{F}$ consists of $g_{1,2}$, the two intrinsic coordinates, as well as $g_{j+2} = \xi_j$, for $j = 1, \cdots 49$, the coordinates of the feature space. Applying TSLasso to the embeddings identifies the set $S = \{g_1, g_2\}$ as the manifold parametrization. This successful recovery of parametrizing functions is observed in each replicate. Figure 3.2b shows the regularization path in one replicates.

Results on Rigid Ethanol Dataset We construct an ethanol skeleton composed of the atoms shown in Figure 3.7a. We then sample configurations as we rotate the atoms around the C-C and C-O bonds. In contrast with the MD trajectories, which are simulated according to quantum dynamics, these two angles are distributed uniformly over a grid, and Gaussian noise ($N_D(0, \sigma^2 I_D)$) is added to the position of each atom. We call the resultant dataset RigidEthanol. As expected given our two a priori known degrees of freedom, Figures 3.3a, 3.3b and 3.3c show that the estimated manifold is a two-dimensional surface with a torus topology similar to that observed for the MD Ethanol in Figure 3.8a. In particular, it is parameterized by bond torsions g_1 and g_2. The dictionary contains the 12 torsions implicitly defined by the bond diagram, the same as the MDS real data experiment. The function pattern is also the same as the real ethanol dataset.

Figure 3.4a-3.4c show the result of experiments on regid ethanol without any noise. We can tell from the result that over all 25 replicates, TSLasso successfully recover the true support, one function from each colinear group.

With the increase in noise, we display the watch plot in figure 3.5a-3.5d. With the increase in the noise, it is possible that TSLasso do not recover the correct support. For example when noise level is $\sigma = 0.1$, in all replicates, TSLasso selects two functions in the same group. Interestingly, when we look at the embedding given by Diffusion Maps at this noise level, we observe that the torus topology is broken, as shown in figure 3.6a and 3.6b.

Results on MDS data We display the experimental results on molecular dynamic simulation data. Original data are available at https://`figshare.com/s/fbd95c10b09f1140389d`.

Dataset	n	N_a	D	d	h_n	m	n'	p	ω
SwissRoll	10000	NA	49	2	.18	2	100	51	1
RigidEthanol	10000	9	50	2	3.5	3	100	12	25
Ethanol	50000	9	50	2	3.5	3	100	12	25
Malonaldehyde	50000	9	50	2	3.5	3	100	12	25
Toluene	50000	16	50	1	1.9	2	100	30	25

Table 3.1: Summary of experiments. **SwissRoll** and **RigidEthanol** are toy data, while **Toluene**, **Ethanol**, and **Malonaldehyde** are from quantum molecular dynamics simulations by Chmiela et al. [2017a]. The columns list the following experimental parameters: n is the sample size for manifold embedding, N_a is the number of atoms in the molecule, D is the dimension of ξ, d is the intrinsic dimension, h_n is the kernel bandwidth, m is the embedding dimension used for ManifoldLasso , n' is the size of the subsample used for TSLasso and ManifoldLasso, p is the dictionary size, and ω is the number of independent repetitions of TSLasso and ManifoldLasso.

These simulations dynamically generate atomic configurations which, due to interatomic interactions, exhibit non-linear, multiscale, non-i.i.d. noise, as well as non-trivial topology and geometry. That is, they lie near a low-dimensional manifold Das et al. [2006]. Such simulations are reasonable application for TSLasso because there is no sparse parameterization of the data manifold known a priori. Such parameterizations are useful. They provide scientific insight about the data generating mechanism, and can be used to bias future simulations. However, these parameterizations are typically are detected by a trained human expert manually inspecting embedded data manifolds for covariates of interest. Therefore, we instead apply TSLasso to identify functional covariates that parameterize this manifold.

We obtain a Euclidean group-invariant featurization of the atomic coordinates as a vector of planar angles $a_i \in \mathbb{R}^{3\binom{N_a}{3}}$: the planar angles formed by triplets of atoms in the molecule [Chen et al., 2019]. We then perform an SVD on this featurization, and project the data onto the top $D = 50$ singular vectors to remove linear redundancies. Note that this represents a particular metric on the molecular *shape space*.

The dictionaries we considered are constructed on *bond diagram*, a priori information about molecular structure garnered from historical work. Building a dictionary based on this structure is akin to many other methods in the field [Krenn et al., 2020, Xie et al., 2019]. Specifically, this dictionary consist of all equivalence classes of 4-tuples of atoms implicitly defined along the molecule skeletons.

Since original angular data featurization is an overparameterization of the shape space, one cannot use automatically obtained gradients in TSLasso. We therefore project the gradients prior to normalization on the tangent bundle of the shape space as it is embedded in $\mathbb{R}^D$.

For TSLasso, the regularization parameter λ_n ranges from 0 to the value for which $||\mathbf{B}_{\cdot j}||_2 = 0$ for all j. The last d surviving dictionary functions are chosen as the parameterization for the manifold.

The toluene case is a manifold with $d = 1$. We observe that in all replicates, TSLassosuccessfully select one of the six torsions associated with the peripheral methyl group bond, which shows the ability of our algorithm to automatically select appropriate parametrizing functions.

We plot the incoherence for Ethanol and Malonaldehyde as the heatmap in figure 3.7b and 3.7f, which present two groups of highly linearly dependent torsions, corresponding to the two bonds between heavy atoms in the molecules. Therefore, we expect to select a pair of incoherent torsions out of these dictionaries. In figure 3.7h and 3.7d, support recovery frequencies for sets of size $d = s = 2$ using TSLasso on ethanol and malonaldehyde data respectively. As we expected, TSLasso select one function from the two groups of highly colinear functions in most replicates. These results shows that our approach is able to identify embedding coordinates that are comparable or preferable to the a priori known functional support.

Also we point out that in our experiments, the subsampled size $n' = 100$ is only around 1% of the whole dataset and in almost all replicates this subsample is sufficient to obtain a valid parametrization. Tangent space estimation is only needed for these points. Therefore bypassing the usual manifold embedding procedure (on the whole dataset) we are able to obtain interpretable embeddings with fewer samples and in a shorter time.

Comparison with embeddings and TSLasso The of nonlinear dimension reduction usually are generated subsequently to running a non-parametric manifold learning algorithm, either through visual or saliency-based analyses, but we are able to achieve comparable results without the use of such an algorithm. These results also suggest that the local denoising property of the tangent space estimation, coupled with the global regularity imposed by the assumption that the manifold is parameterized by the same functions throughout, is sufficient to replicate the denoising effect of a manifold learning algorithm. Plus, with the help of domain functions, our embeddings come with good interpretability.

The comparisons with Diffusion maps of Toluene are shown in the introduction. Here we display some comparison of TSLasso with Diffusion maps on real MDS data, which are widely used for dimension reduction. Figure 3.8a and 3.8b shows that the two functions selected from the TSLasso indeed parameterize the structure of the data. As the values are roughly varying along with two circles of the torus. Figure 3.9a and 3.9b shows a pair of functions selected by TSLasso .

3.6.2 Validation of ManifoldLasso

In this subsection we illustrate experimental results for ManifoldLasso . For all of the following experiments, the data consist of n data points in D dimensions, as well an embedding $\phi_{1:m}(\mathcal{D})$. We assume access to the manifold dimension d, a kernel bandwidth h_n used in the estimation of the tangent spaces, and p dictionary functions. Except where otherwise specified, m and ϵ_M are used in the preliminary step of generating embeddings $\phi_{1:m}$ using the diffusion maps algorithm as EmbeddingAlg. ManifoldLasso is applied to a uniformly random subset of size $n' = |\mathcal{I}|$ and this process is repeated ω number of times. These parameters are passed to the Laplacian, TangentSpaceBasis, RMetric, and PullBackD algorithms, and are summarized in Table 3.1. The regularization parameter λ ranges over $[0, \lambda_{\max}]$ as described in Section 3.4.

Results on Swiss Roll synthetic data The construction of Swiss Roll data is the same as in validation of TSLasso with the same dictionary. We learn the manifold using three techniques: local tangent space alignment, diffusion maps, and isomap, shown in Figures 3.10c, 3.10e and 3.10g. For comparison, we also analyze the "trivial embedding" $\phi_1^{Internal} = g_1$, $\phi_2^{Internal} = g_2$ (Figure 3.10a). These rectilinear coordinates are colored in red and blue, and show clear associations with the other embedding coordinates.

Applying ManifoldLasso to the embeddings identifies the set $S = \{g_1, g_2\}$ as the manifold explanation, and identifies the association of the recovered support with individual embedding coordinates $\phi_{1,2}$. By visual inspection of Figures 3.10a, 3.10c, 3.10e, and 3.10g, we see that all embedding algorithms recover the original manifold, although the embeddings $\phi^{\text{Iso}}, \phi^{DM}, \ldots$ are not isometric (this is particularly noticeable with diffusion maps), and sign changes are possible. Figures 3.10b, 3.10d, 3.10f and 3.10h demonstrate that ManifoldLasso recovers the two manifold-specific coordinate functions in each case, while the coefficients $\beta_{3:51}$ decay rapidly to 0 with λ. Furthermore, each of g_1 and g_2 is always mapped to the correct embedding coordinate. The regularization paths are virtually identical for all embeddings, even though the embeddings are not isometric.

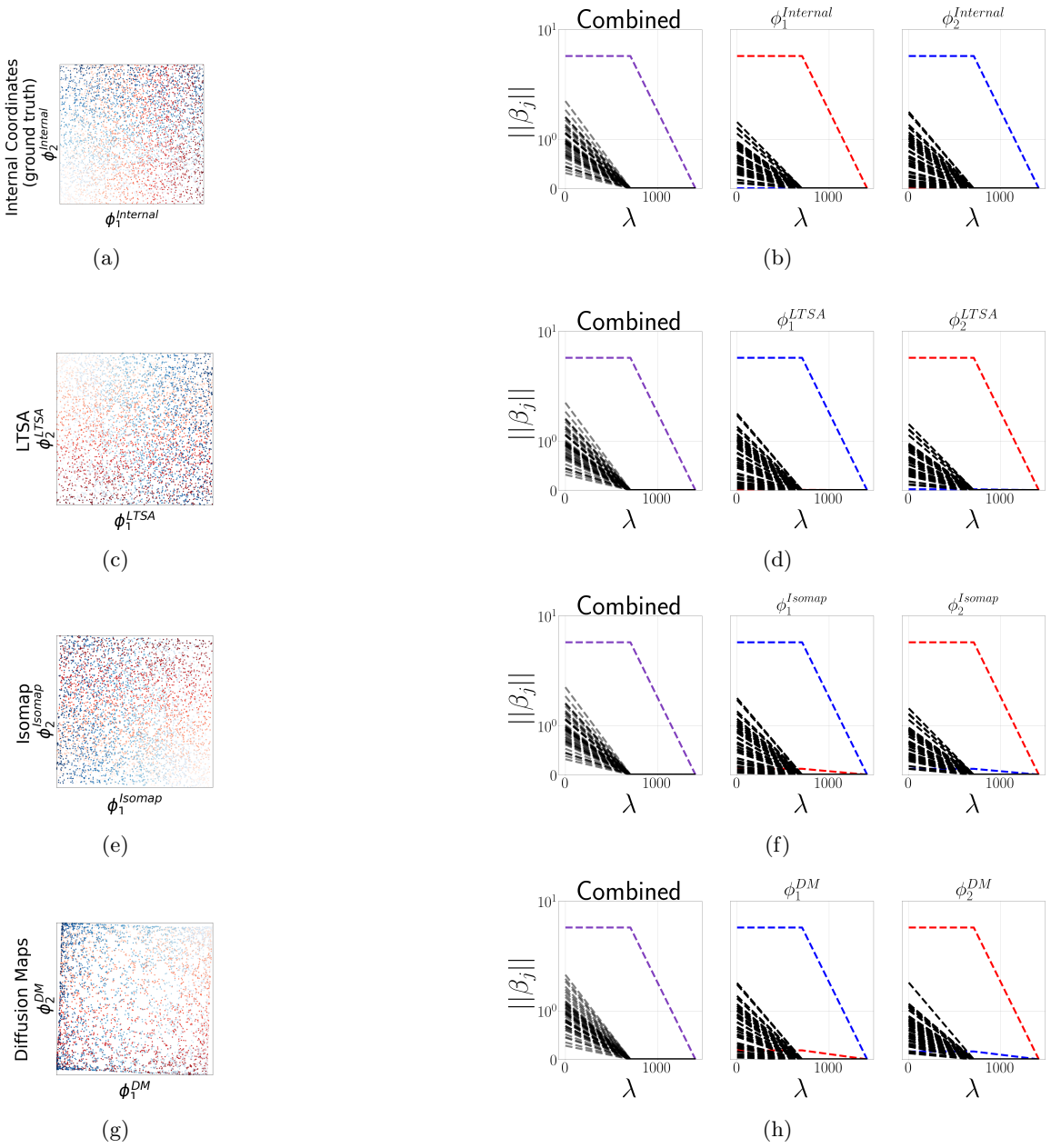


Figure 3.10: Results for **SwissRoll** embedded using a variety of manifold learning algorithms. Figure 3.10a shows the data mapped w.r.t. the edges of the rectangle colored by g_1 in red and g_2 in blue. Figures 3.10c, 3.10e, and 3.10g display embeddings of **SwissRoll** generated by several different manifold learning methods, colored by the rectilinear coordinates in red and blue. Figures 3.10b, 3.10d, 3.10f, and 3.10h display the regularization paths of ManifoldLasso for these embeddings. The combined norms $\|\beta_j\|$ used in ManifoldLasso are given on the left, and the norms for the individual embedding coordinates $\|\beta_{jk}\|$ on the right.

Results of ManifoldLasso **on a Rigid Ethanol Data** We also validate ManifoldLasso on a rigid ethanol skeleton data set, constructed the same as the one for TSLasso . As expected given our two a priori known degrees of freedom, Figures 3.11b and 3.11c show that the estimated manifold is a two-dimensional surface with a torus topology parameterized by bond torsions g_1 and g_2 similar to that observed for the MD **Ethanol** in Figure 3.1.

Results of ManifoldLasso **on MDS data** Bond diagrams are based on a priori information about molecular structure garnered from historical work. Building a dictionary based on this structure is akin to many other methods in the field [Krenn et al., 2020, Xie et al., 2019]. As in the case of **RigidEthanol**, our dictionaries consist of all equivalence classes of 4-tuples of atoms implicitly defined by bond diagrams, and the incoherence plots for **Ethanol** and **Malonaldehyde** in Figures 3.12a and 3.12d show two groups of highly dependent torsions, corresponding to the two bonds between heavy atoms in the molecules. We have labelled these by their central bond. For example, g_1 of ethanol is described by 9 functionally dependent torsions, since each central carbon has three peripheral atoms, while g_2 of ethanol is described by only 3 functionally dependent torsions, since, by the diagram, the oxygen atom only has one peripheral atom. Therefore, success means recovering a pair of incoherent torsions out of these dictionaries. For **Toluene**, the manifold dimension is $d = 1$ and success means recovering one of the 6 torsions associated with the peripheral methyl group bond. For this molecule, there are also $p-6 = 24$ torsions that do not explain the data manifold. These correspond to bonds within the main benzene ring. We apply ManifoldLasso with these dictionaries to the embeddings shown in Figure 3.1.

As Figure 3.12 shows, ManifoldLasso is always able to identify torsions corresponding to the expected labelled bonds. Figures 3.12b, 3.12e, and 3.12g show combined regularization paths for single replicates of ManifoldLasso, and Figures 3.12c, 3.12f and 3.12h show frequencies of support recovery of sets of size d over $w = 25$ replicates. ManifoldLasso finds that the toroidal **Ethanol** manifold is explained by pairs of torsions from the C-O and C-C bonds, while **Malonaldehyde** is explained by one of each of the two central bonds. **Toluene** is explained by the torsion of the peripheral methyl group. These agree

with our domain-expert validated parameterizations in figure 3.1.

3.7 Discussion

In this section, we discuss compare this chapter with related works.

Nonparametric methods of obtaining embeddings We draw a firm distinction between our approach and purely non-parametric methods that attempt to learn a parameterization of $\mathcal{M}$. For example, the early works of Saul and Roweis [2003] and Teh and Roweis [2002] (and references therein) propose parametrizing the manifold by finite mixtures of local linear models, aligned so as to provides global coordinates, in a way reminiscent of Local Tangent Space Alignment [Zhang and Zha, 2004]. Another idea is to use d eigenfunctions of the Laplace-Beltrami operator $\Delta_{\mathcal{M}}$ as a parametrization of $\mathcal{M}$. Hence, the Diffusion Maps coordinates could be considered such a parametrization [Coifman and Lafon, 2006, Coifman et al., 2005, Gear, 2012]. However, these are not in and of themselves interpretable, and it is not clear how many such coordinates are needed [Chen and Meilă, 2019]. In Mohammed and Narayanan [2017], it was shown that principal curves and surfaces can provide an approximate manifold parametrization. These methods can often be used as embedding algorithms in our approach, but make no attempts at synergizing with an interpretable dictionary. Dsilva et al. [2018] tackle the related problem of choosing among the infinitely many Laplacian eigenfunctions d which provide a d-dimensional parametrization of the manifold. Their approach is to solve a sequence of Local Linear Embedding [Roweis and Saul, 2000] problems, each aiming to represent an eigenfunction as a combination of the preceding ones. Similarly, Chen and Meilă [2019] is another method for reducing the number of "covarying" eigenfunctions. However, these methods fail to provide physical meaning for the selected functions.

Our work differs from the above entirely non-parametric methods in two key ways: (1) the explanations we obtain are endowed with the meaning of the domain specific dictionaries, (2) less obviously, descriptors like principal curves or Laplacian eigenfunctions are generally still non-parametric (i.e exist in infinite dimensional function spaces), while the parameterizations by dictionaries we obtain (e.g. the torsions) are in finite dimensional spaces. This

distinction is mirrored in comparison with the many so-called *dictionary learning* methods in which a low-dimensional transformation is learned simultaneously with its inverse. We note that our method is not dictionary learning per se, but rather sparse coding, in which the dictionary is given [Szabó et al., 2011].

Symbolic regression The *symbolic regression* methods of Brunton et al. [2016], Rudy et al. [2019], and Champion et al. [2019] for estimating governing laws of dynamical systems are perhaps most similar to this work. These methods use sparse regression with respect to a dictionary and the idea of differential composition. Their goal is to identify the functional equations of non-linear dynamical systems by regressing the time derivatives of the state variables on a subset of functions in the dictionary selected using a sparsity inducing penalty. This provides a natural interpretability. However, although these methods can loosely be considered univariate analogs, they do not consider the multidimensional data-manifold, and their synergies with dimension-reduction algorithms are developed in separate directions.

Sparse regression With respect to sparse regression, the seminal Group Lasso paper of Yuan and Lin [2006] and support recovery analyses of Elyaderani et al. [2017], Wainwright [2009] are central to our approach. However, our use of replicates in experiments is reminiscent of the Stability Selection method of Meinshausen and Bühlmann [2010]. Such methods address instabilities of the variable selection, in particular, when restrictive theoretical conditions are violated [Zhao and Yu, 2006, Huan Xu et al., 2012]. Some attractive alternate approaches to this problem that we do not pursue are the use of non-convex penalties such as SCAD [Fan and Li, 2001, Breheny and Huang, 2011] and weighted data points in the Adaptive Lasso [Zou, 2006]. We note the method of Haufe et al. [2009], which applies group lasso to analyze sparse decomposition of vectors fields, albeit in a different setting.

We also draw several distinctions between the TSLasso method and the ManifoldLasso method presented in this chapter. First, ManifoldLasso uses the same essential idea of sparse linear regression in gradient space, but in order to explain individual embedding coordinate functions. In contrast, we have no consistent matching between unit vectors in $\mathbf{I}_d$, and so can only provide an overall regularization path, rather than one corresponding

to individual tangent basis vectors. The tangent bases are not themselves gradients of a known function, and, indeed it may not be the case that such a function even exists. Second, TSLasso method dispenses *with the entire Embedding algorithm*, Riemannian metric estimation, and pulling back the embedding gradients steps in ManifoldLasso , while providing almost everything a user can get from ManifoldLasso . Apart from simplification, TSLasso can be run on $n' \ll n$ data points, about 1/500 of the data in our experiments, while the algorithm in ManifoldLasso computes an embedding from all data points. Hence, all operations before the actual GroupLasso are hundreds of times faster than in ManifoldLasso .

Although in this chapter the dictionary consists of functions with physical meaning, our general principle of finding parametric geometrically-motivated approximations of learned representations is relevant to a range of machine learning contexts. Examining functions in embedding coordinates is quite typical in genomics [Amir et al., 2013], and much deep learning work also makes use of explicit traversal of a latent space [Lin et al., 2020, Shukla et al., 2018]. It is also known in a range of settings that learned gradients provide interpretable [Adebayo et al., 2018] or otherwise statistically-useful information [Wu et al., 2010, Constantine et al., 2014, Yang, 2020]. Our approach relies on the classical weighted local PCA method for tangent space estimation [Joncas et al., 2017, Aamari and Levrard, 2019]. Improvement of this estimator in the presence of noise is an active area of research [Puchkin and Spokoiny, 2019].

3.8 Proof of results in chapter 3

In this section we will provide additional proofs to the theoretical results in the main text.

3.8.1 Proofs for Section 3.1

Proof of proposition 3.1. If F is a parametrization for a.e. $\xi \in \mathcal{M}$, F is a local diffeomorphism, i.e., there exists a neighborhood $V_0 \subset \mathcal{M}$ such that F is a diffeomorphism on V_0. Then on V_0 there is an inverse F^{-1} such that $F \circ F^{-1} = \mathbf{Id}$. By chain rule we have $\operatorname{rank} F = d$ on $\mathcal{M}^\circ$. On the other hand if $\operatorname{rank} F = d$ almost everywhere on $\mathcal{M}$, let $\mathcal{M}^\circ = \{\xi : \operatorname{rank} F = d\}$, then at each $\xi \in \mathcal{M}^\circ$, consider the coordinate representation of the

linear map dF: it must contains a $d \times d$ invertible submatrix. Since F is C^1, in a local neighborhood V_0 this submatrix has positive determinant. Hence $V_0 \subset \mathcal{M}^\circ$. Then by the global rank theorem F is a local diffeomorphism on V_0 to $F(V_0)$ and hence a parametrization. $\square$

Proof of proposition 3.2. Since both F, F' are parametrizations, then a.e. on $\mathcal{M}$, their ranks are d. it then suffices to prove that there exists a function τ that $F_1 = \tau \circ F_2$ and τ is smooth almost everywhere in $\mathbb{R}^{m_2}$.

Define $\mathcal{M}^\circ = \{\xi : \operatorname{rank} F_2 = d\}$. We first show that there exists a smooth function τ defined on $F_2(\mathcal{M}^\circ)$ such that $F_1 = \tau \circ F_2$ on $\mathcal{M}^\circ$. Once this is achieved, by Sard's theorem [Lee, 2003], $F_2(\mathcal{M} \setminus \mathcal{M}^\circ)$ has Lebesgue measure zero in $\mathbb{R}^{m_2}$ so that we can assign arbitrary value of τ on $F_2(\mathcal{M} \setminus \mathcal{M}^\circ)$ and the desired result is proved.

To show the claim, we first fix a local point $\xi \in \mathcal{M}^\circ$ with a local chart (U, φ). Denote the $i-th$ component function of F_1, F_2 to be F_1^i, F_2^i respectively. Lemma 3.1 implies that for each $\xi \in \mathcal{M}^\circ$, there exists some neighborhood $U_\xi \in \mathcal{M}^\circ$ of ξ and C^1 functions $\tau_\xi^i : \mathbb{R}^{m_2} \to \mathbb{R}, i = 1, 2, \cdots, m_1$ such that

$$F_1^i(\xi) = (F_1^i \circ \varphi^{-1})(\varphi(\xi)) = \tau_\xi^i(F_2 \circ \varphi^{-1}(\varphi(\xi))) = \tau_\xi^i(F_2(\xi)), \text{ for } i = 1, 2, \cdots, m_1, \ \xi \in U_\xi. \tag{3.14}$$

Here we should notice that τ_ξ^i is defined only on $F_2(U_\xi)$. Since this holds for every $\xi \in \mathcal{M}^\circ$, we can find an open cover $\{U_\xi\}$ of the original manifold $\mathcal{M}^\circ$. By partition of unity in Lemma 3.2, namely that $\mathcal{M}^\circ$ admits a smooth partition of unity subordinate to the cover $\{U_\xi\}$. We denote this partition of unity by $\psi_\xi(\cdot)$.

Hence we can define

$$\tau^i(y) = \sum_{\xi \in \mathcal{M}^\circ} \psi_\xi(F_2^{-1}(y)) \tau_\xi^i(y), \quad y \in F_2(\mathcal{M}^\circ). \tag{3.15}$$

where τ^i is a function mapping from $F_2(\mathcal{M}^\circ) \to \mathbb{R}$. For each fixed $\xi \in \mathcal{M}^\circ$, the function $y \to \psi_\xi(F_2^{-1}(y)) \tau_\xi^i(y)$ for $y \in F_2(\mathcal{M}^\circ)$ is C^1. According to the properties of partition of unity, in a local neighborhood of each point, this is a summation of finitely many smooth functions. Then this τ^i will be a C^1 function on $F_2(\mathcal{M}^\circ)$. Also, by $1 = \sum_x \psi_x(\xi)$, it holds that $\tau^i(g_{S'}(\xi)) = F_2^i(\xi)$ for any $i = 1, \cdots, m_1$.

Therefore, globally in U we have

$$F_1^i(\xi) = \tau^i(F_2(\xi)), \text{ for } i = 1, \cdots, m_1, \ \xi \in \mathcal{M}^\circ. \tag{3.16}$$

Hence the desired result holds. □

Lemma 3.1 (Remark 2 after Zorich [2004] Theorem 2 in Section 8.6.2)**.** *Let $f : U \to \mathbb{R}^m$ be a mapping defined in an open neighborhood $U \subset \mathbb{R}^d$ of a point $x^\star \in \mathbb{R}^d$. Suppose $f \in C^\ell$, the rank of the mapping f is k at every point in U, and $k < m$. Moreover, assume that the principal minor of order k of the matrix Df is not zero at $x^\star$. Then in some neighborhood $U_{x^\star} \subset U$ there exist $m - k$ C^ℓ functions $g_i, i = k+1, \cdots, m$ such that for any $x = (x_1, \cdots, x_d) \in U(x^\star)$,*

$$g_i(x_1, x_2, \cdots, x_d) = g_i(g_1(x_1, x_2, \cdots, x_d), g_2(x_1, x_2, \cdots, x_d), \cdots, g_k(x_1, x_2, \cdots, x_d)). \tag{3.17}$$

Partitions of unity enable us to expand the above lemma from local to global. Mathematically, a *smooth partition of unity subordinate to* $\{U_\alpha\}$ is an indexed family $(\psi_\alpha)_{\alpha \in A}$ of smooth functions $\psi_\alpha : \mathcal{M} \to \mathbb{R}$ with the following properties:

1. $0 \leq \psi_\alpha(\xi)$ for all $\alpha \in A$ and all $\xi \in \mathcal{M}$;

2. supp $\psi_\alpha \subset U_\alpha$ for each $\alpha \in A$;

3. Every $\xi \in \mathcal{M}$ has a neighborhood that intersects supp ψ_α for only finitely many values of α;

4. $\sum_{\alpha \in A} \psi_\alpha(\xi) = 1$ for all $\xi \in \mathcal{M}$.

Lemma 3.2 (Lee [2003] Theorem 2.23)**.** *Suppose that $\mathcal{M}$ is a smooth manifold, and $\{U_\alpha\}_{\alpha \in A}$ is any indexed open cover of $\mathcal{M}$. Then there exists a smooth partition of unity subordinate to $\{U_\alpha\}$.*

3.8.2 Proofs of basis-invariance property of TSLasso

Proposition 3.3. *Consider alternative bases* $\mathbf{T}_i' = \mathbf{T}_i\mathbf{\Gamma}_i$ *where* $\mathbf{\Gamma}_i$ *are* $d \times d$ *orthonormal matrices. If* $\{\mathbf{B}_i\}_{i=1}^n$ *minimizes* (3.4)*, then in the new tangent bases,* $\{\mathbf{B}_i\mathbf{\Gamma}_i\}_{i=1}^n$ *minimizes the corresponding loss function, which is constructed through replacing* $\mathbf{X}_i$ *by* $\mathbf{\Gamma}_i\mathbf{X}_i$ *in* (3.4)*. Furthermore, the selected support* S *is independent of the basis chosen for each tangent space.*

Proof of Proposition 3.3. It suffices to show that the loss in (3.4) does not change under orthogonal transformation of individual tangent bases. As long as this holds, $\mathbf{B}_i\mathbf{\Gamma}_i$ must minimize the loss since otherwise one could argue that $J_{\lambda_n}(\mathbf{B})$ is not a minimum value for the original tangent space bases. Note that the norm $||\mathbf{B}_{\cdot j}||_2$ is unitary invariant. This is because $\mathbf{B}_{\cdot j} = (j - \text{th row of } \mathbf{B}_i)_{i=1}^n$ is constructed by stacking the $j-$th row of each $\mathbf{B}_i$. Hence the new norm is given by the norm of $(j - \text{th row of } \mathbf{B}_i\mathbf{\Gamma}_i)_{i=1}^n$; therefore the Group Lasso penalty doesn't change after changing $\mathbf{B}_i$ to $\mathbf{B}_i\mathbf{\Gamma}_i$ for each i. Finally, it holds that $||\mathbf{I}_d - \mathbf{\Gamma}_i^\top\mathbf{X}_i\mathbf{B}_i\mathbf{\Gamma}_i||_F^2 = ||\mathbf{\Gamma}_i^\top(\mathbf{I}_d - \mathbf{X}_i\mathbf{B}_i)\mathbf{\Gamma}_i||_F^2 = ||\mathbf{I}_d - \mathbf{X}_i\mathbf{B}_i||_F^2$, so the ℓ_2-loss is not changed under orthonormal transformation of the tangent bases. These rotation invariances guarantee the same support S. □

3.8.3 Proof of section 3.5

We start with stating the following lemma, which gives the sufficient and necessary condition of certain matrices $\mathbf{B}_i$ to be the solution to problem (3.4). It also provides conditions on unique support recovery and unique solutions. The proof is standard in convex analysis literature; we follow a procedure as in [Wainwright, 2009].

Lemma 3.3. *1. Matrix* $\mathbf{B}$ *is the optimal solution to problem* (3.4) *if and only if there exists an matrix* $Z = (z_1^\top, z_2^\top, \cdots, z_p^\top)^\top \in \mathbb{R}^{p\times nd}$ *such that*

$$z_j = \begin{cases} \frac{\beta_i}{||\beta_i||} & \beta_i \neq 0\,, \\ \in \mathbb{R}^{nd} \text{ with } ||z_j||_2 \leq 1, & \text{otherwise}\,; \end{cases} \tag{3.18}$$

and

$$\Big(\mathbf{X}_1^\top(\mathbf{I}_d - \mathbf{X}_1\mathbf{B}_1), \mathbf{X}_2^\top(\mathbf{I}_d - \mathbf{X}_2\mathbf{B}_2), \cdots, \mathbf{X}_n^\top(\mathbf{I}_d - \mathbf{X}_n\mathbf{B}_n)\Big) = \frac{\lambda_n}{\sqrt{nd}}\mathbf{Z}\,. \tag{3.19}$$

.

2. *If under the setting of (a), further in* (3.18)*, we have* $||z_i|| < 1$ *whenever* $\beta_i = 0$*, then all optimal solutions* $\widetilde{\mathbf{B}}$ *of Tangent Lasso problem will have support* $S(\widetilde{\mathbf{B}}) \subset S(\mathbf{B})$.

3. *Under setting of (a) and (b). Let* $\mathbf{X}_{iS(\mathbf{B})}$ *be the submatrix constructed by the* $S(\mathbf{B})$ *columns of of* $\mathbf{X}_i$*. If all* $\mathbf{X}_{iS(\mathbf{B})}^\top \mathbf{X}_{iS(\mathbf{B})}$ *are invertible, then the TSLasso solution is unique.*

Proof of lemma 3.3. Before we further explore the result, we transform the problem (3.4). We stack the matrices at each point together. We will now write

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \cdots \\ \mathbf{X}_n \end{pmatrix} \in \mathbb{R}^{nd \times p}, \quad \mathbf{B} = \Big(\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_n\Big) \in \mathbb{R}^{p \times nd} .$$

Then β_j is the $j-th$ row for $\mathbf{B}$. Further let matrix

$$\mathbf{E}_i = \Big(\mathbf{0}, \cdots, \mathbf{0}, \mathbf{I}_d, \mathbf{0}, \cdots, \mathbf{0}\Big)^\top \in \mathbb{R}^{nd \times d}$$

be the block matrix with the $i-th$ block being identity matrix and the other blocks are all zeros. Then the loss function of TSLasso can be rewritten as

$$J_{\lambda_n}(\mathbf{B}) = \frac{1}{2} \sum_{i=1}^{n} ||\mathbf{E}_i^\top (\mathbf{I}_{nd} - \mathbf{X}\mathbf{B})\mathbf{E}_i||_F^2 + \frac{\lambda_n}{\sqrt{nd}} ||\mathbf{B}||_{1,2} . \tag{3.20}$$

where $||\mathbf{B}||_{1,2}$ is the norm defined by $\sum_{j=1}^{p} ||\beta_j||_2$.

The proof of this lemma is standard technique in convex analysis. Define $h_i(\mathbf{B}) = ||\mathbf{E}_i^\top (\mathbf{I}_{nd} - \mathbf{X}\mathbf{B})\mathbf{E}_i||_F^2$ penalty part and g is the group lasso penalty.

The first step is to compute the gradient of $h_i(\mathbf{B})$ with respect to $\mathbf{B}$. For any $\mathbf{H} \in \mathbb{R}^{p \times nd}$, compute

$$\begin{aligned}
&h_i(\mathbf{B} + \mathbf{H}) - h_i(\mathbf{B}) \\
&= \operatorname{trace}(\mathbf{E}_i^\top (\mathbf{I}_{nd} - \mathbf{X}(\mathbf{B} + \mathbf{H}))\mathbf{E}_i)^\top (\mathbf{E}_i^\top (\mathbf{I}_{nd} - \mathbf{X}(\mathbf{B} + \mathbf{H}))\mathbf{E}_i) - \operatorname{trace}(\mathbf{E}_i^\top (\mathbf{I}_{nd} - \mathbf{X}\mathbf{B})\mathbf{E}_i)^\top (\mathbf{E}_i^\top (\mathbf{I}_{nd} - \mathbf{X}\mathbf{B})\mathbf{E} \\
&= -2 \operatorname{trace}(\mathbf{H}^\top \mathbf{X}^\top \mathbf{E}_i \mathbf{E}_i^\top (\mathbf{I}_{nd} - \mathbf{X}\mathbf{B})\mathbf{E}_i \mathbf{E}_i^\top) + O(||\mathbf{H}||_F^2) \\
&= -2 \Big\langle \mathbf{H}, \mathbf{X}^\top \mathbf{E}_i \mathbf{E}_i^\top (\mathbf{I}_{nd} - \mathbf{X}\mathbf{B})\mathbf{E}_i \mathbf{E}_i^\top \Big\rangle_F + O(||\mathbf{H}||_F^2) .
\end{aligned}$$

Hence we can conclude that $\nabla_{\mathbf{B}} h_i(\mathbf{B}) = -2\mathbf{X}^\top \mathbf{E}_i \mathbf{E}_i^\top (\mathbf{I}_{nd} - \mathbf{X}\mathbf{B})\mathbf{E}_i \mathbf{E}_i^\top = -2\mathbf{X}^\top \mathbf{E}_i (\mathbf{I}_d - \mathbf{X}_i \mathbf{B}_i)\mathbf{E}_i^\top$, and therefore

$$\begin{aligned}\nabla_{\mathbf{B}} \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{I}_d - \mathbf{X}_i\mathbf{B}_i||_F^2 &= \sum_{i=1}^{n} -\mathbf{X}^\top \mathbf{E}_i(\mathbf{I}_d - \mathbf{X}_i\mathbf{B}_i)\mathbf{E}_i^\top \\ &= -\left(\mathbf{X}_1^\top(\mathbf{I}_d - \mathbf{X}_1\mathbf{B}_1), \mathbf{X}_2^\top(\mathbf{I}_d - \mathbf{X}_2\mathbf{B}_2), \cdots, \mathbf{X}_n^\top(\mathbf{I}_d - \mathbf{X}_n\mathbf{B}_n)\right) .\end{aligned} \tag{3.21}$$

Recall that we use β_i to denote the $i-th$ row of $\mathbf{B}$. We use a similar argument in proof of lemma 2 of [Obozinski et al., 2011] and notice that the original optimization problem is convex and strictly feasible (hence strong duality holds). The primal problem is

$$\begin{aligned}\min_{\substack{\mathbf{B}\in\mathbb{R}^{p\times nd}\\ b\in\mathbb{R}^p}} \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{E}_i^\top(\mathbf{I}_{nd} - \mathbf{X}\mathbf{B})\mathbf{E}_i||_F^2 + \frac{\lambda_n}{\sqrt{nd}}\sum_{j=1}^{p} b_j \\ s.t.\ (\beta_j, b_j) \in \mathcal{K}, 1 \le j \le p\ ,\end{aligned}$$

where $\mathcal{K}$ is the second-order cone as usually defined. The dual problem is given by

$$\begin{aligned}\max_{\substack{\mathbf{Z}\in\mathbb{R}^{p\times nd}\\ t\in\mathbb{R}^p}} \min_{\substack{\mathbf{B}\in\mathbb{R}^{p\times nd}\\ b\in\mathbb{R}^p}} L(\mathbf{B}, b, \mathbf{Z}, t) = \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{E}_i^\top(\mathbf{I}_{nd} - \mathbf{X}\mathbf{B})\mathbf{E}_i||_F^2 + \frac{\lambda_n}{\sqrt{nd}}\sum_{j=1}^{p} b_j + \sum_{j=1}^{p}\langle (z_j, t_j), (\beta_j, b_j)\rangle \\ s.t.\ (z_j, t_j) \in \mathcal{K}^\circ\ ,\end{aligned}$$

where $z_j \in \mathbb{R}^{nd}$ is the $j-$th row of $\mathbf{Z}$. Note that $\mathcal{K}^\circ$ is the polar cone of $\mathcal{K}$ and second order cone is self-dual. Hence we have $(z_i, -\mathbf{T}_i) \in \mathcal{K}$.

Since the primal problem is strictly feasible, strong duality holds. For any pair of $(\mathbf{B}^*, b^*)$ and $(\mathbf{Z}^*, t^*)$ primal and dual solutions, they have to satisfy the KKT condtion that

$$||\beta_j^*||_2 \le b_j^*, \qquad 1 \le j \le p\ , \tag{3.22a}$$

$$||z_j^*||_2 \le -t_j^*, \qquad 1 \le j \le p\ , \tag{3.22b}$$

$$z_j^{*T}\beta_j^* + t_j^* b_j^* = 0, \qquad 1 \le j \le p\ , \tag{3.22c}$$

$$\nabla_B \left[\frac{1}{2}\sum_{i=1}^{n} ||\mathbf{I}_d - \mathbf{X}_i\mathbf{B}_i||_F^2\right] + \mathbf{Z}^* = 0\ , \tag{3.22d}$$

$$\frac{\lambda_n}{\sqrt{nd}} + t_j^* = 0\ . \tag{3.22e}$$

Note that (3.22c)implies that $t_j^* = -\frac{\lambda_n}{\sqrt{nd}}$. Then by (3.22a) and (3.22b) we have $||z_j^{*T}\beta_j^*|| \leq \frac{\lambda_n}{\sqrt{nd}}||\beta_j||_2$. Notice that the equality holds in (3.22c), there fore $||z_j^*|| = \frac{\sqrt{nd}}{\lambda_n}$ and $b_j^* = ||\beta_j^*||$. Renormalize $z_j^* = \frac{\sqrt{nd}}{\lambda_n} z_j^*$ and part (a) holds. For part b, for any $j, z_j^{*T}\beta_j = ||\beta_j||_2$. Then $\beta_j = 0$ must hold for any $||z_j|| < 1$. For part (c) note that in this case the loss function is strictly convex when the original problem is restricted to minimizing over $\mathbf{B} : \beta_i = 0, \quad \forall i \notin S(\mathbf{B})$. This strict convexity implies the uniqueness of solution. □

The previous lemma provides a tool for understanding the support recovery consistency of TSLasso.

For any arbitrary $S \subset [p]$ such that $\operatorname{rank} \mathbf{X}_{iS} = d$ holds for all $i \in [n]$, we establish a sufficient condition on $\mathbf{X}_{iS}$ such that they can be discovered by the TSLasso. Suppose at each data point i, we decompose the matrix $\mathbf{I}_d$ by

$$\mathbf{I}_d = \mathbf{X}_{iS}\mathbf{B}_{iS}^* + \mathbf{W}_{iS} \ , \tag{3.23}$$

where $\mathbf{B}_{iS}^*$s are $|S| \times d$ matrices that only has non zero entries in rows in S and minimizes the loss $||\mathbf{I}_d - \mathbf{X}_i\mathbf{B}_i||_F^2$. In fact, since $\mathbf{X}_{iS}$ is full rank, there exists a unique $\mathbf{B}_{iS}^*$ for each i such that $\mathbf{W}_{iS} = 0$. Denote $\mathbf{B}_{i,j\cdot}^*$ be the $j-$th row in $\mathbf{B}_{iS}^*$ and define

$$b_S = \min_{i\in[n]} \min_{j\in S} ||\mathbf{B}_{i,j\cdot}^*|| \ . \tag{3.24}$$

This is a sample version of b_S defined in (3.9).

The following lemma shows a sufficient condition on $\mathbf{X}_i$ so that the true support can be found. We first define several derived quantities of $\mathbf{X}_i$. Denoting the $j-$th column of matrix $\mathbf{X}_i$ by x_{ij}, we define

$$\text{S-incoherence} \quad \tilde{\mu}_S = \max_{i=1:n, j\in S, j'\notin S} \frac{|x_{ij}^\top x_{ij'}|}{||\nabla g_j(\xi_i)||||\nabla g_{j'}(\xi_i)||} \ . \tag{3.25a}$$

$$\text{internal-colinearity} \quad \tilde{\nu}_S = \max_{\imath=1:n} ||(\tilde{\mathbf{X}}_{iS}^\top \tilde{\mathbf{X}}_{iS})^{-1} - \mathbf{G}_S(\xi_i)^2|| \ . \tag{3.25b}$$

$$\text{maximal gradient norm} \quad \tilde{\phi}_S = \max_{i=1:n} \max_{j\in S} ||\nabla g_j(\xi_i)|| \ . \tag{3.25c}$$

These are sampled version of μ_S,ν_S and ϕ_S defined on the whole manifold from (3.10),(3.11) and (3.12).

Now we are ready to prove theorem 3.1. We start with some lemmas in linear algebra.

Lemma 3.4. *Let* $\mathbf{A}, \mathbf{B}$ *be* $d \times d$ *positive definite matrices. Then* $||\mathbf{A}^{-1} - \mathbf{B}^{-1}|| \leq ||\mathbf{B}^{-1}||||\mathbf{A} - \mathbf{B}||$.

Lemma 3.5. *Let* $\mathbf{A}, \mathbf{B}$ *be two* $d \times d$ *matrices.* $\mathbf{A}$ *is positive semidefinite. Denote* $||\mathbf{A}||_{\infty,2}$ *be the maximum* ℓ_2 *norm of the rows of* $\mathbf{A}$*. Then* $||\mathbf{AB}||_{\infty,2} \leq \sqrt{d}||\mathbf{A}||||\mathbf{B}||_F$.

Proof of lemma 3.5. Write $\mathbf{A} = (a_{ij})_{d\times d}, \mathbf{B} = (b_{ij})_{d\times d}$, then by definition

$$\begin{aligned}
||\mathbf{AB}||_{\infty,2}^2 &= \max_{i=1:d} \sum_{j=1}^{d} \left(\sum_{k=1}^{d} a_{ik} b_{kj} \right)^2 \\
&\leq \max_{i=1:d} \sum_{j=1}^{d} \left(\sum_{k=1}^{d} a_{ik}^2 \right) \left(\sum_{k=1}^{d} b_{kj}^2 \right) \\
&\leq \left(\max_{i=1:d} \sum_{k=1}^{d} a_{ik}^2 \right) \left(\sum_{j=1}^{d} \sum_{k=1}^{d} b_{kj}^2 \right) \\
&= d||\mathbf{A}||_{\infty,2}^2 ||\mathbf{B}||_F^2 \ .
\end{aligned}$$

Since $\mathbf{A}$ is positive semidefinite, we have

$$||\mathbf{A}||_{\infty,2}^2 = \max_{i=1:d} (\mathbf{AA})_{ii} \leq ||\mathbf{A}^2|| = ||\mathbf{A}||^2 \ .$$

Hence we conclude the desired result. □

Lemma 3.6. *Let* $\delta = \min_{\xi \in \mathcal{M}} \min_{j=1:p} ||\nabla g_j(\xi)||$*, then* $||(\mathbf{X}_{iS}^\top \mathbf{X}_{iS})^{-1}||_2 \leq 1 + \frac{\tilde{\nu}_S}{\delta^2}$.

Proof of 3.6. Recall that $\mathbf{G}_S(\xi_i) = \text{diag}\{||\nabla g_j(\xi_i)||\}_{j,j' \in S}$. We first consider that

$$\begin{aligned}
&||(\mathbf{X}_{iS}^\top \mathbf{X}_{iS})^{-1} - \mathbf{I}_d||_2 \\
=&||\mathbf{G}_S^{-1}(\xi_i)(\tilde{\mathbf{X}}_{iS}^\top \tilde{\mathbf{X}}_{iS})^{-1} \mathbf{G}_S(\xi_i)^{-1} - \mathbf{G}_S^{-1}(\xi_i)\mathbf{G}_S(\xi_i)^2 \mathbf{G}_S^{-1}(\xi_i)|| \\
\leq&||(\tilde{\mathbf{X}}_{iS}^\top \tilde{\mathbf{X}}_{iS})^{-1} - \mathbf{G}_S(\xi_i)^2||||\mathbf{G}_S^{-1}(\xi_i)||^2 \\
\leq&\frac{\tilde{\nu}_S}{\delta^2} \ .
\end{aligned}$$

And the desired results come from triangular inequality. □

Lemma 3.7. *Let* $\{\xi_i\}_{i=1}^n$ *be fixed data points on* $\mathcal{M}$*. Let* $\tilde{\delta} = \min_{\xi \in \mathcal{M}} \min_{j=1:p} ||\nabla g_j(\xi)||$ *and* $\Gamma = \max_{\xi \in \mathcal{M}} \max_{j=1:p} ||\nabla g_j(\xi)||$ *Let* $\tilde{\mu}_S, \tilde{\nu}_S, \tilde{\phi}_S$ *defined from* $\mathbf{X}_{iS}$ *according to* (3.25a),(3.25b)

and (3.25c) *respectively. Then Tangent Lasso problem* (3.4) *has a unique solution* $\widehat{\mathbf{B}} = [\widehat{\mathbf{B}}_1, \widehat{\mathbf{B}}_2, \cdots, \widehat{\mathbf{B}}_n] \in \mathbb{R}^{p\times nd}$ *with support* $S(\widehat{\mathbf{B}})$ *included in the true support* S *if* $(1+\frac{\tilde{\nu}_S}{\delta^2})\tilde{\mu}_S\tilde{\phi}_S\Gamma d < 1$. *Furthermore, if* $\lambda_n(1+\tilde{\nu}_S/\delta^2) < \tilde{b}_S\sqrt{n}/2$, *then* $S(\widehat{\mathbf{B}}) = S$.

Proof of lemma 3.7. We follow the procedure of Primal-Dual witness method (see e.g.Wainwright [2009], Obozinski et al. [2011], Elyaderani et al. [2017]).

Still considering the reformulated optimization problem (3.20), we first find $\widehat{\mathbf{B}}$ from minimizing a restricted optimization problem

$$\min_{S(\mathbf{B})\subset S} J_{\lambda_n}(\mathbf{B}) = \frac{1}{2}\sum_{i=1}^{n}||\mathbf{E}_i^\top(\mathbf{I}_{nd} - \mathbf{X}\mathbf{B})\mathbf{E}_i||_F^2 + \frac{\lambda_n}{\sqrt{nd}}||\mathbf{B}||_{1,2} \ . \tag{3.26}$$

We then construct a dual solution $\widehat{\mathbf{Z}}$ and show that $\widehat{\mathbf{B}}$ is the solution to the original optimization problem. We write z_j as the $j-$th row of $\widehat{\mathbf{Z}}$ and decompose each $\widehat{z}_j = [\widehat{z}_{j,1}, \widehat{z}_{j,2}, \cdots, \widehat{z}_{j,n}]$. According to lemma 3.3, we can solve for $\widehat{\mathbf{Z}}$ from those optimality conditions.

First, notice that

$$\widehat{\mathbf{B}}_{iS} - \mathbf{B}^*_{iS} = -\frac{\lambda_n}{\sqrt{nd}}(\mathbf{X}_{iS}^\top\mathbf{X}_{iS})^{-1}\widehat{\mathbf{Z}}_{S,i} \ .$$

where $\widehat{\mathbf{Z}}_S$ is constructed by concatenating the $j \in S$ row of $\widehat{\mathbf{Z}}_i$.

For an $d \times d$ matrix $\mathbf{A}$, we write $||\mathbf{A}||_{\infty,2} = \max_{i=1}^{d}||a_i||_2$, where a_i is the $i-$th row of A. Then it holds that from lemma 3.5

$$||(\mathbf{X}_{iS}^\top\mathbf{X}_{iS})^{-1}\widehat{\mathbf{Z}}_{S,i}||_{\infty,2} \leq ||(\mathbf{X}_{iS}^\top\mathbf{X}_{iS})^{-1}||||\widehat{\mathbf{Z}}_{S,i}||_F \ .$$

Therefore recall that $||\widehat{\mathbf{Z}}_S||_{\infty,2} = 1$ we conclude that $||\widehat{\mathbf{Z}}_{S,i}||_F \leq \sqrt{d}\tilde{\nu}_S$. And adopting lemma 3.6 we have

$$||(\mathbf{X}_{iS}^\top\mathbf{X}_{iS})^{-1}\widehat{\mathbf{Z}}_{S,i}||_{\infty,2} \leq \sqrt{d}(1 + \frac{\tilde{\nu}_S}{\delta^2}) \ .$$

According to (3.8.3) if $\lambda_n\sqrt{d}(1+\frac{\tilde{\nu}_S}{\delta^2})/\sqrt{nd} < \frac{1}{2}\tilde{b}_S$, then $||\widehat{\mathbf{B}}_{iS,j\cdot}|| \geq \frac{1}{2}\tilde{b}_S$ for each row $j \in S$. And this condition is satisfied by the assumption.

On the other hand, for any $j' \notin S$, we have

$$\widehat{z}_{j',i} = x_{ij'}^\top\mathbf{X}_{iS}(\mathbf{X}_{iS}^\top\mathbf{X}_{iS})^{-1}\widehat{\mathbf{Z}}_{S,i} \ .$$

It suffices to verify that $||\widehat{z}_j|| < 1$ for all $j' \notin S$. For any i, we have

$$||x_{ij'}^\top \mathbf{X}_{iS}(\mathbf{X}_{iS}^\top \mathbf{X}_{iS})^{-1}||_2 \leq (1+\frac{\tilde{\nu}_S}{\delta^2})||x_{ij'}^\top \mathbf{X}_{iS}||_2 \leq \sqrt{d}(1+\frac{\tilde{\nu}_S}{\delta^2})\tilde{\mu}_S||\nabla g_{j'}(\xi_i)||\max_{j\in S}||\nabla g_j(\xi_i)|| \leq \sqrt{d}(1+\frac{\tilde{\nu}_S}{\delta^2})\tilde{\mu}_S\tilde{\phi}_S\Gamma \ .$$

Directly compute that

$$\begin{aligned}
||\widehat{z}_{j'}||^2 &\leq \sum_{i=1}^n ||x_{ij'}^\top \mathbf{X}_{iS}(\mathbf{X}_{iS}^\top \mathbf{X}_{iS})^{-1}\widehat{\mathbf{Z}}_{S,i}||_2^2 \\
&\leq \sum_{i=1}^n ||x_{ij'}^\top \mathbf{X}_{iS}(\mathbf{X}_{iS}^\top \mathbf{X}_{iS})^{-1}||_2^2 ||\widehat{\mathbf{Z}}_{S,i}||_F^2 \\
&\leq d(1+\frac{\tilde{\nu}_S}{\delta^2})^2\tilde{\mu}_S^2\tilde{\phi}_S^2\Gamma^2 \sum_{i=1}^n ||\widehat{\mathbf{Z}}_{S,i}||_F^2 \\
&\leq (1+\frac{\tilde{\nu}_S}{\delta^2})^2\tilde{\mu}_S^2\tilde{\phi}_S^2\Gamma^2 d^2 < 1 \ .
\end{aligned}$$

□

This lemma is the recovery result if the tangent space is estimated without any noise. Note that this conditions also implies further results on the 'isometric' property of TSLasso. If there are two different subsets S, S' such that $|S| = |S'| = d$ and both has rank d at each data point. Then for both subsets, $\mathbf{X}_{iS}^\top \mathbf{X}_{iS}$ are invertible, and the lemma also implies that $\tilde{\mu}_S\tilde{\nu}_S d < 1$ cannot hold at the same time for both subsets. The one picked by TSLasso (usually) has a lower value of $\tilde{\nu}_S$, and will be closer to isometry to some extent.

This recovery result does not involve the tuning parameter for false inclusion. Therefore, it justifies our selection of tuning parameter that force the support has cardinality less than d. If we do observe d functions selected and they have rank d everywhere, then under incoherence condition they must be a right parameterization. To avoid false exclusion, the tuning parameter λ_n cannot be too large.

Now we connect these support recovery results inherent to our optimization approach with the tangent space estimation algorithm. Let $\mathbf{T}_i, \widehat{\mathbf{T}}_i$ be the orthogonal basis in $\mathbb{R}^{D\times d}$ for true and estimated tangent space respectively, and write

$$e = \max_{i=1}^n ||\mathbf{T}_i\mathbf{T}_i^\top - \widehat{\mathbf{T}}_i\widehat{\mathbf{T}}_i^\top||_2 \ . \tag{3.27}$$

We have the following recovery result in the setting that gradient is estimated with some noise.

Lemma 3.8. *Let $\xi_i, i = 1:n$ be fixed data points on manifold $\mathcal{M} \subset \mathbb{R}^D$. Given S a subset of functions in dictionary $\mathcal{F} = \{g_j, j \in [p]\}$ with $|S| = d$. Suppose* $\operatorname{rank} \operatorname{grad} g_S = d$ *at each data point. Fix $\mathbf{T}_i$ as an orthonormal basis of tangent space at ξ_i, and $\widehat{\mathbf{T}}_i$ a basis for the estimated tangent space. And further define $\mathbf{X}_i = \mathbf{T}_i^\top[\nabla g_j], \widehat{\mathbf{X}}_i = \widehat{\mathbf{T}}_i^\top[\nabla g_j], j \in [p]$ where ∇ is the ambient gradient. Define $\mathbf{B}^*_{iS}, \tilde{b}_S$ the same as lemma 3.7. Assume that $||\nabla g_j|| = 1$ for all $\xi_i, i \in [n], j \in [p]$. Define $\tilde{\mu}_S, \tilde{\nu}_S$ from* (3.25a) *and* (3.25b) *and e from* (3.27)*. Then let $\widehat{\mathbf{B}}$ be the solution of TSLasso problem*

$$J_{\lambda_n}(\mathbf{B}) = \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{I}_d - \widehat{\mathbf{X}}_i \mathbf{B}_i||_F^2 + \frac{\lambda_n}{\sqrt{nd}} ||\mathbf{B}||_{1,2} \, . \tag{3.28}$$

If $(1 + \tilde{\nu}_S/\delta^2)\tilde{\mu}_S\tilde{\phi}_S\Gamma d < 1$ and $\lambda_n(1 + \tilde{\nu}_S/\delta^2) < \tilde{b}_S\sqrt{n}/2$, there exists a positive constant c_0 such that if $e < c_0$ then $S(\widehat{\mathbf{B}}) = S$.

Proof of lemma 3.8. The proof is direct by identifying the new $\tilde{\mu}'_S, \tilde{\nu}'_S$ parameters under noisy estimation of tangent space. The other parameters $\tilde{\phi}_S, \Gamma, \delta$ are not related with tangent spaces and thus remains unchanged.

Denote $\widehat{x}_i j$ the $j-$th column of $\widehat{\mathbf{X}}_i$. Similarly, to (3.25a), we first bound

$$\begin{aligned}\widehat{x}_{ij}^\top \widehat{x}_{ij'} &= \nabla g_j(\xi_i)^\top [\widehat{\mathbf{T}}_i \widehat{\mathbf{T}}_i^\top - \mathbf{T}_i \mathbf{T}_i^\top] \nabla g_j(\xi_i) + \nabla g_j(\xi_i)^\top \mathbf{T}_i \mathbf{T}_i^\top \nabla g_j(\xi_i) \\ &\leq ||\widehat{\mathbf{T}}_i \widehat{\mathbf{T}}_i^\top - \mathbf{T}_i \mathbf{T}_i^\top||_2 ||\nabla g_j(\xi_i)|| ||\nabla g_{j'}(\xi_i)|| + \tilde{\mu}_S ||\nabla g_j(\xi_i)|| ||\nabla g_{j'}(\xi_i)||, \quad \text{for all } \; j \in S, j' \notin S, i \in [n] \, .\end{aligned}$$

So $\tilde{\mu}'_S \leq \tilde{\mu}_S + e$.

By definition, let

$$\tilde{\widehat{\mathbf{X}}}_{iS} = \left[\frac{\widehat{\mathbf{T}}_i^\top \nabla g_j(\xi_i)}{||\nabla g_j(\xi_i)||}\right]_{j \in S} = \widehat{\mathbf{X}}_{iS} \mathbf{G}(\xi_i)^{-1}$$

where $\mathbf{G}(\xi_i) = \operatorname{diag}\{||\nabla g_j(\xi_i)||\}_{j \in S}$ and then we have

$$\tilde{\nu}'_S = ||(\tilde{\widehat{\mathbf{X}}}_{iS}^\top \tilde{\widehat{\mathbf{X}}}_{iS})^{-1} - \mathbf{G}(\xi_i)^{-2}|| \leq \tilde{\nu}_S + ||(\tilde{\widehat{\mathbf{X}}}_{iS}^\top \tilde{\widehat{\mathbf{X}}}_{iS})^{-1} - (\tilde{\mathbf{X}}_{iS}^\top \tilde{\mathbf{X}}_{iS})^{-1}|| \, .$$

It suffices to upper bound the second term. We can apply lemma 3.4, the perturbation bound of inverse of positive definite matrices. It suffice to compute

$$||(\tilde{\mathbf{X}}_{iS}^\top \tilde{\mathbf{X}}_{iS})^{-1}|| \leq ||(\tilde{\mathbf{X}}_{iS}^\top \tilde{\mathbf{X}}_{iS})^{-1} - \mathbf{G}_S(\xi_i)^2 + \mathbf{G}_S(\xi_i)^2|| \leq \tilde{\phi}_S^2 + \tilde{\nu}_S \, .$$

And since for any $j, j' \in S$, it holds that

$$|(\tilde{\hat{\mathbf{X}}}_{iS}^{\top}\tilde{\hat{\mathbf{X}}}_{iS})_{jj'} - (\tilde{\mathbf{X}}_{iS}^{\top}\tilde{\mathbf{X}}_{iS})_{jj'}| \leq ||\mathbf{T}_i\mathbf{T}_i^{\top} - \widehat{\mathbf{T}}_i\widehat{\mathbf{T}}_i^{\top}|| \leq e \ .$$

$$||\tilde{\hat{\mathbf{X}}}_{iS}^{\top}\tilde{\hat{\mathbf{X}}}_{iS} - \tilde{\mathbf{X}}_{iS}^{\top}\tilde{\mathbf{X}}_{iS}|| \leq d\sqrt{\max_{j,j'}|(\tilde{\hat{\mathbf{X}}}_{iS}^{\top}\tilde{\hat{\mathbf{X}}}_{iS})_{jj'} - (\tilde{\mathbf{X}}_{iS}^{\top}\tilde{\mathbf{X}}_{iS})_{jj'}|} \leq de \ .$$

And thus we have

$$||(\tilde{\hat{\mathbf{X}}}_{iS}^{\top}\tilde{\hat{\mathbf{X}}}_{iS})^{-1} - (\tilde{\mathbf{X}}_{iS}^{\top}\tilde{\mathbf{X}}_{iS})^{-1}|| \leq (\frac{1}{\delta^2} + \tilde{\nu}_S)||\tilde{\hat{\mathbf{X}}}_{iS}^{\top}\tilde{\hat{\mathbf{X}}}_{iS} - \tilde{\mathbf{X}}_{iS}^{\top}\tilde{\mathbf{X}}_{iS}|| \leq (\tilde{\phi}_S^2 + \tilde{\nu}_S)de \ .$$

Hence $\tilde{\nu}'_S \leq \tilde{\nu}_S + (\tilde{\phi}_S^2 + \tilde{\nu}_S)de$.

Let $\tilde{c}_1 := -B/2A + \sqrt{B^2 - 4AC}/2A$, where

$$A = \frac{\tilde{\phi}_S^2 + \tilde{\nu}_S}{\delta^2}d, \quad B = \frac{\tilde{\mu}_S d(\tilde{\phi}_S^2 + \tilde{\nu}_S)}{\delta^2} + \frac{\nu_S}{\delta^2} + 1, \quad C = \tilde{\mu}_S(1 + \frac{\nu_S}{\delta^2}) - \frac{1}{\tilde{\phi}_S\Gamma d} \ . \tag{3.29}$$

If $e < \tilde{c}_1$, then $(1 + \frac{\tilde{\nu}'_S}{\delta^2})\tilde{\mu}'_S\tilde{\phi}_S\Gamma d < 1$.

Let

$$\tilde{c}_2 := \frac{\delta^2}{d(\tilde{\phi}_S^2 + \tilde{\nu}_S)}\left[\frac{\tilde{b}_S\sqrt{n}}{2} - \lambda_n(1 + \frac{\tilde{\nu}_S}{\delta^2})\right] \ . \tag{3.30}$$

If $e < \tilde{c}_2$, then $\lambda_n(1 + \frac{\tilde{\nu}'_S}{\delta^2})/\sqrt{n} < \frac{1}{2}\tilde{b}_S$.

The conditions of guarantees that c_1, c_2 are both positive. Hence take $c_0 = \min\{c_1, c_2\}$, lemma 3.7 guarantees exact recovery when $e < c_0$. □

Proof of theorem 3.1. With probability one, the following comparisons between sample based quantities and whole manifold versions holds:

$$\tilde{\mu}_S \leq \mu_S, \quad \tilde{\nu}_S \leq \nu_S, \quad \tilde{\phi}_S \leq \phi_S, \quad \tilde{b}_S \geq b_S \ . \tag{3.31}$$

Let c_1, c_2 be the same as $\tilde{c}_1, \tilde{c}_2$ defined in the proof of theorem 3.8, replacing all sample version quantities $\mu_S, \nu_S, \phi_S, b_S$ with their global manifold counterparts $\mu_S, \nu_S, \phi_S, b_S$.

Then the assumptions of the proposition guarantees that $c_1, c_2 > 0$ and further it holds that $c_1 \leq \tilde{c}_1, c_2 \leq \tilde{c}_2$, hence with probability one when $e < c_0 = \min\{c_1, c_2\}$, $e < \tilde{c}_0 = \min\{\tilde{c}_1, \tilde{c}_2\}$. Note that c_0 is a constant determined by the manifold and the dictionary. It suffices to notice that

$$P(S(\hat{\mathbf{B}}) = S) \leq P(e < c_0) \geq 1 - 4\left(\frac{1}{n}\right)^{\frac{2}{d}} \tag{3.32}$$

given by lemma 3.9. □

Lemma 3.9 (Proposition in Aamari and Levrard [2018])**.** *For sufficiently large C, let $r_N = C(\log n/(n-1))^{1/d}$, tangent spaces $\hat{\mathbf{T}}_i$ estimated by WL-PCA in section* **??** *with linear kernel satisfy that with probability at least $1 - 4(1/n)^{2/d}$*

$$\max_{i=1:n} ||\mathbf{T}_i\mathbf{T}_i^\top - \hat{\mathbf{T}}_i\hat{\mathbf{T}}_i^\top|| = O(r_N) = O((\frac{\log n}{n-1})^{\frac{1}{d}}) \ . \tag{3.33}$$

Remark 3.1. *Note that in this lemma, the hidden constant in big-O notation is determined by the manifold and sampling density.*

3.8.4 Backgrounds on molecule dynamic simulation data

The method of MD simulations is one of the principal tools in the study of molecular systems. Such simulations provide detailed information on the fluctuations and conformational changes of the system, and are now routinely used to investigate the structure, dynamics and thermodynamics of biological macromolecules and their complexes. In such simulations, the positions of atoms within a molecule are sampled as they proceed through time from some initial conditions according to interatomic effects. The distribution of this sample describes the molecule's behavior in the given experimental conditions. It has been shown empirically that manifolds approximate these high-dimensional distributions [Dsilva et al., 2013]. Accordingly, application of manifold learning to find the collective coordinates has achieved great success [Das et al., 2006, Tribello et al., 2012, Noé and Clementi, 2017, Sidky et al., 2020]. Even though the vector of atomic coordinates can take any value, due to interatomic interactions, the relative positions of atoms within the molecule lie near a low-dimensional *slow manifold*. Performing manifold learning on these data separates the conformational changes, modeled by the manifold, from the fluctuations represented by the "noise" around the manifold.

Representing molecular configurations

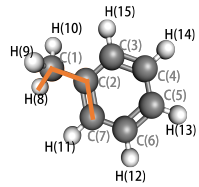
Our MD data are quantum-simulations from Chmiela et al. [2017a]. The raw data consists of X, Y, Z coordinates for each of the N_a atoms of the chosen molecule. For a single observation,

we denote these by $r_i \in \mathbb{R}^{3N_a}$. The first step in our data analysis pipeline is to featurize the configuration in a way that is invariant to rotation and translation. In the present experiments, we follow Chen et al. [2019] and represent a molecular configuration as a vector $a_i \in \mathbb{R}^{3\binom{N_a}{3}}$ of the *planar angles* formed by triplets of atoms. We then perform an SVD on this featurization, and project the data onto the top $D = 50$ singular vectors to remove linear redundancies; we denote the new data points by $\xi_{1:n}$. The EmbeddingAlg and TangentSpaceBasis algorithms work directly with ξ in dimension D. Other possible representations such as applying a Procrustes transform to each configuration to align it with the first one give similar results, and no matter which low level representation we choose, large-scale conformational changes are described by the relative rotations of groups of atoms - the bond torsions illustrated in Figure 3.1 [Chen et al., 2019].

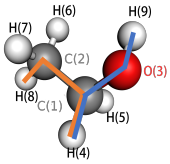
Dictionaries for MD Data

Therefore, in the **RigidEthanol**, **Ethanol**, **Malonaldehyde**, and **Toluene** MD datasets, we construct dictionaries consisting of bond *torsions*. We then apply ManifoldLasso to select combinations of these higher-level torsion features that explain the manifold in the lower-level planar angle feature space. Given an ordered 4-tuple of atoms $ABCD$, the torsion g_{ABCD} is the angle of the planes defined by the locations of ABC and BCD. Note that $g_{ABCD} \equiv g_{DBCA} \equiv g_{DCBA} \equiv g_{ACBD}$. Any torsion g is expressible in closed form as functions of the planar angles feature vector a. In particular, a torsion g_{ABCD} is a function of the angles of the triangles inscribing atoms ABC, ABD, ACD, and BCD. We compute the gradients of the torsions by automatic differentiation [Paszke et al., 2019].

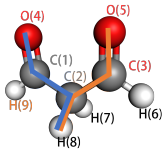
One cannot use the obtained gradients directly in ManifoldLasso, since the angular features overparameterize the molecular *shape space* $\Sigma_3^{N_a}$ [Addicoat and Collins, 2010, Kendall, 1989] of dimension $D' = 3N_a - 7$, and off-manifold gradients are therefore not well-defined. For example, whether one chooses to use triangles ABC, ABD, and ACD, or ABC, ABD, and BCD to compute g_{ABCD} has no effect on the value of g_{ABCD}, but changes the value of its partial derivatives in $\mathbb{R}^D$. We therefore project the gradients on the tangent bundle of the shape space as it is embedded in $\mathbb{R}^D$.



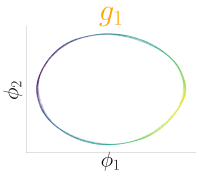
(a) Toluene



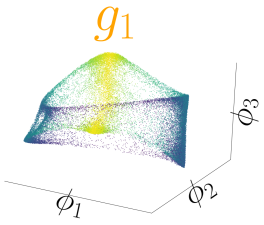
(b) Ethanol



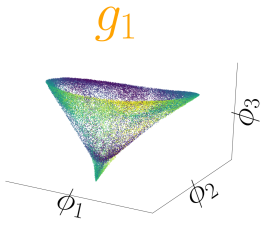
(c) Malonaldehyde



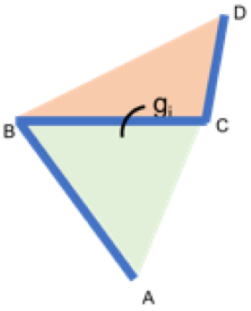
(d)



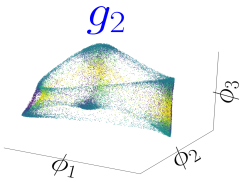
(e)



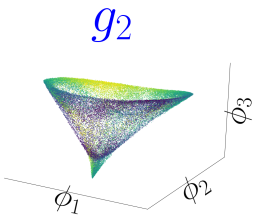
(f)



(g) Torsion example

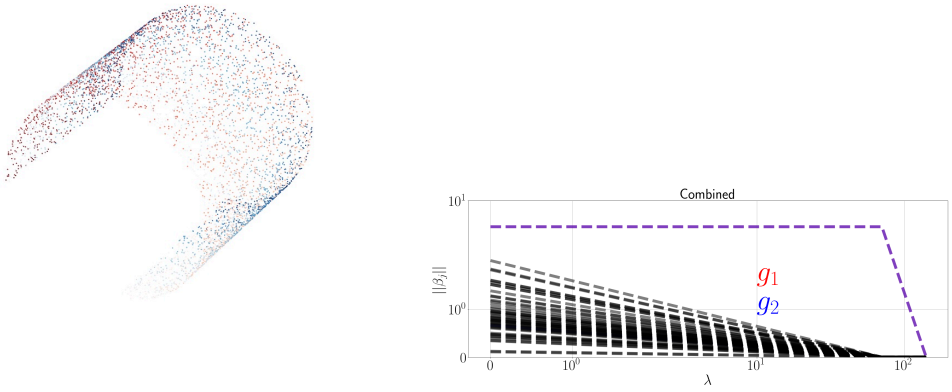


(h)



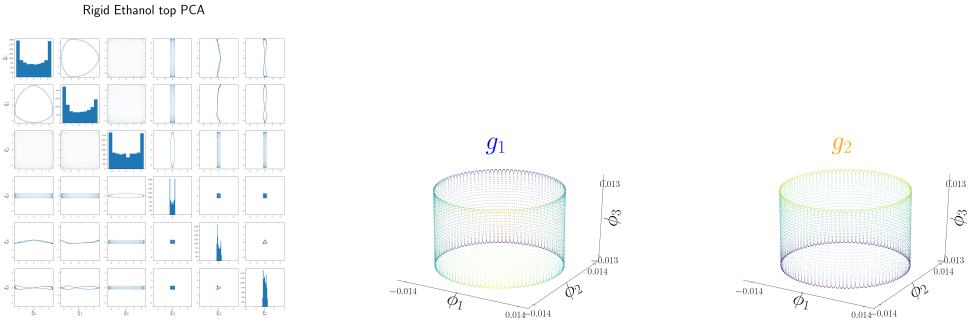
(i)

Figure 3.1: Manifold coordinates with physical meaning in molecular dynamics (MD) simulations. 3.1a-3.1c Diagrams of the toluene (C_7H_8), ethanol (C_2H_5OH), and malonaldehyde ($C_3H_4O_2$) molecules, with the carbon (C) atoms in grey, the oxygen (O) atoms in red, and the hydrogen (H) atoms in white. Bonds defining important torsions g_j are marked in orange and blue. The bond torsion is the angle of the planes inscribing the first three and last three atoms on the line (3.1g). 3.1d Embedding of the configurations of toluene into $m = 2$ dimensions, showing a manifold of $d = 1$. The color corresponds to the values of the orange torsion g_1. 3.1e, 3.1h Embedding of the configurations of the ethanol in $m = 3$ dimensions, showing a manifold of dimension $d = 2$, respectively colored by the blue and orange torsions in Figure 3.1b. 3.1f, 3.1i. Embedding of the configurations of malonaldehyde in $m = 3$ dimensions, showing a manifold of dimension $d = 2$, respectively colored by the blue and orange torsions in Figure 3.1c.



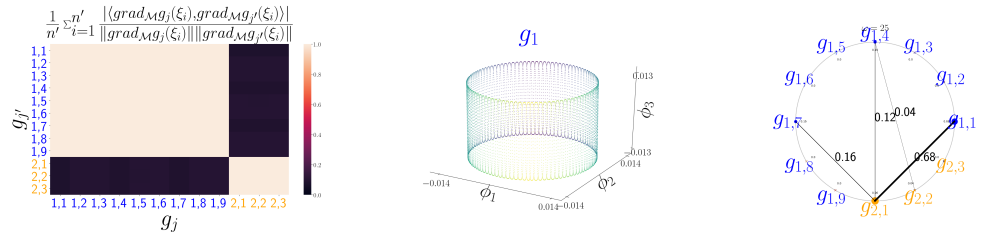
(a) (b)

Figure 3.2: Swiss Roll data and result. **Left:** Unrotated swiss roll dataset in $\mathbb{R}^3$. This dataset is then randomly rotated into $\mathbb{R}^{49}$. **Right:** The regularization path of TSLasso on SwissRoll datatset in one replicate. Note that in fact there are two functions selected and their regularization path added together.



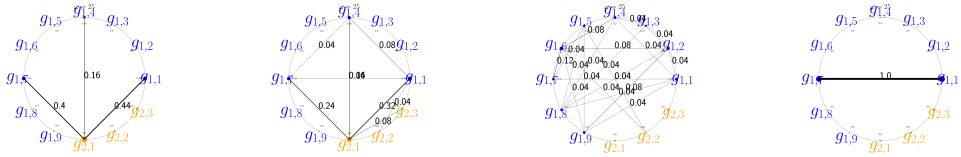
(a) (b) (c)

Figure 3.3: PCA features and Diffusion Map embedding features of rigid ethanol data without noise.



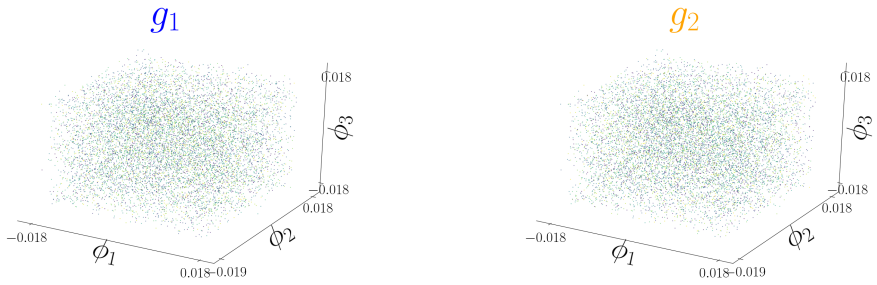
(a) (b) (c)

Figure 3.4: Results of Rigid Ethanol Experiment with no noise. **Left:** cosine plots of dictionary functions, showing the existence of two groups of highly colinear functions. **Middle:** regularization path in one replicate. **Right:** The frequency of each pair of function selected in all 25 replicates.



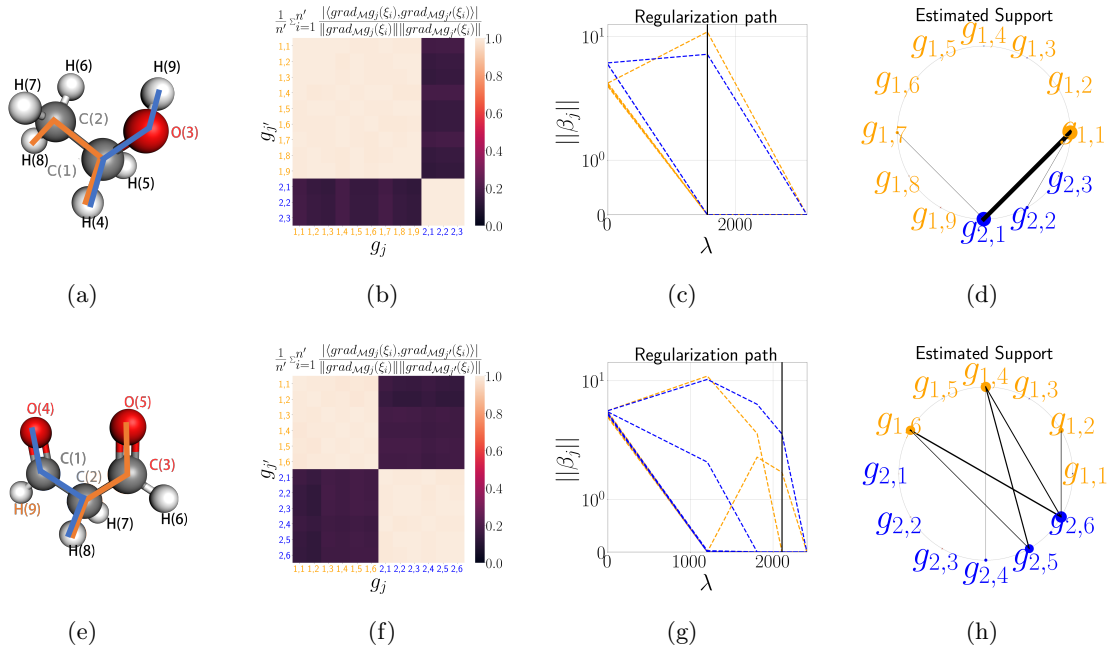
(a) $\sigma = 0.0001$ (b) $\sigma = 0.001$ (c) $\sigma = 0.01$ (d) $\sigma = 0.1$

Figure 3.5: Watch plot of support recovery frequencies under different noise levels.



(a) (b)

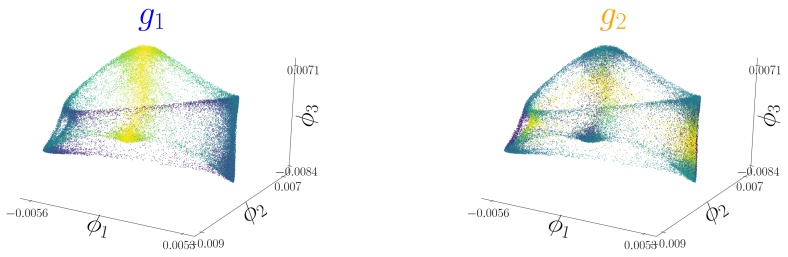
Figure 3.6: Diffusion map embedding for synthetic rigid ethanol data. Data points are colored by the true torsion g_1 and g_2 respectively.



(a) (b) (c) (d)

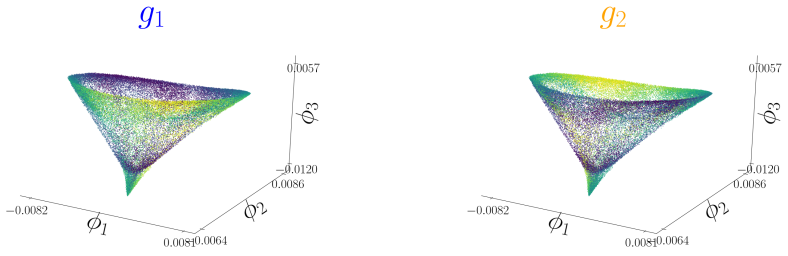
(e) (f) (g) (h)

Figure 3.7: Results from molecular dynamics data. 3.7a, 3.7e show bond diagrams for ethanol and malonaldehyde, respectively. 3.7b and 3.7f show the heatmap of cosines (incoherences) of dictionary functions. The color is darker when there is more colinearity. 3.7c, 3.7g are regularization paths for a single replicate of ethanol and malonaldehyde. Note that in both figures there are a redundant trajectory of two functions that are added together. 3.7d, 3.7h Selection of pairs of functions for ethanol and malonaldehyde over replicants using TSLasso . The node point on the circles represents all functions in the dictionary and the number along the lines are frequencies of each pairs selected over 25 repetitions. 3.7d means in all 25 repetitions, TSLasso selects $g_{1,1}$ and $g_{2,1}$, which are the bond torsions around C-C bond and C-O bond respectively. 3.7h show that in 24 out of 25 replicates, TSLasso is able to select one function from each highly colinear function group.



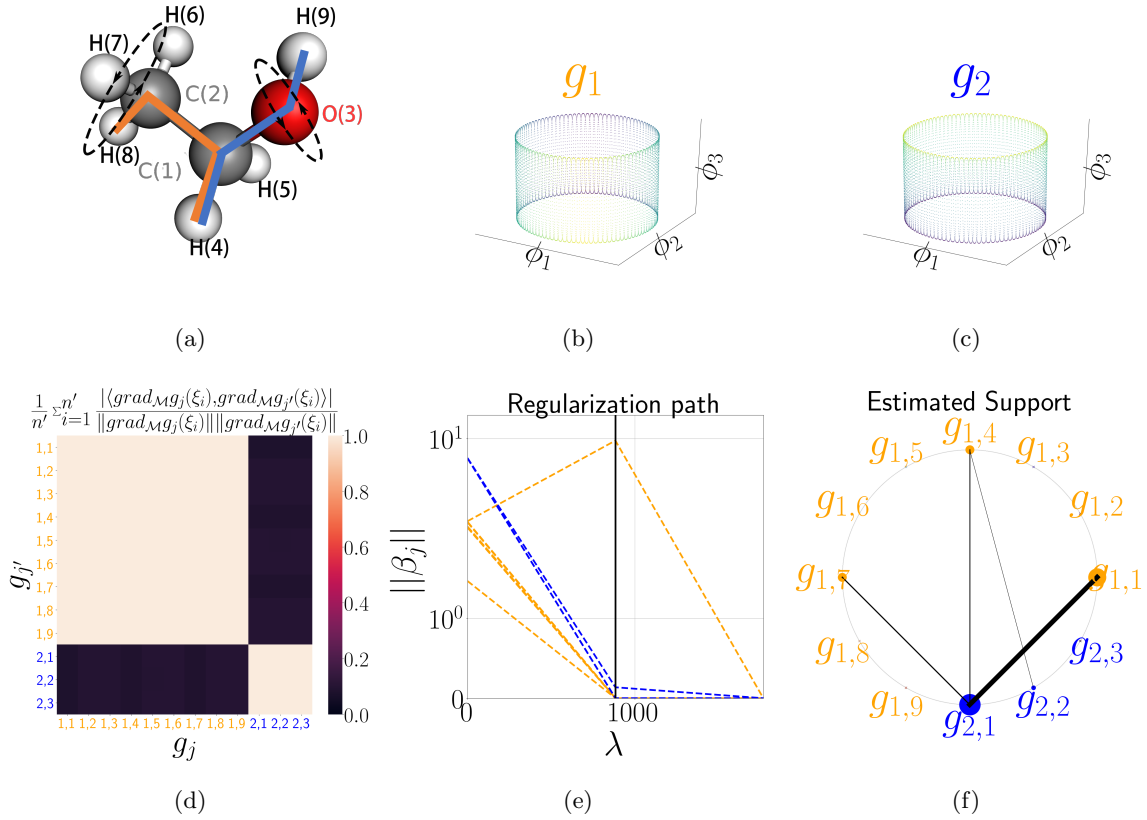
(a) (b)

Figure 3.8: Diffusion map embedding for real ethanol data. Data points are colored by the two torsion functions g_0, g_9 found by TSLasso respectively.



(a) (b)

Figure 3.9: Diffusion map embedding for Malonaldehyde data. Data points are colored by the two torsion functions found by TSLasso respectively.



(a) (b) (c)

(d) (e) (f)

Figure 3.11: Results of ManifoldLasso for **RigidEthanol**. Figure 3.11a shows the simplified dynamics of our rigid molecular simulation. Atoms in the rigid ethanol skeleton are articulated around the C-O and C-C bonds by a torus of rotations. Figure 3.11b shows the learned torus, colored by C-C torsion g_1 from Figure 3.1. Figure 3.11c shows the same torus, colored by the C-O torsion g_2 from Figure 3.1. Figure 3.11d displays the incoherences, i.e. pairwise collinearities of dictionary gradients; C-C torsions functionally dependent on g_1 are in orange, C-O torsions functionally dependent on g_2 are in blue. Figure 3.11e shows combined regularization paths $\|\beta_j\|$ vs. λ for a single replicate. The tuning parameter at which $|S| = d$ is indicated by the vertical black line. The chord diagram in Figure 3.11f represents the frequency of selecting each pair of torsions in replicate experiments. The frequencies with which individual torsions are selected are given by the sizes of the perimeter dots corresponding to each dictionary element, while the frequencies with which pairs of torsions are selected are given by the line widths connecting the dots.

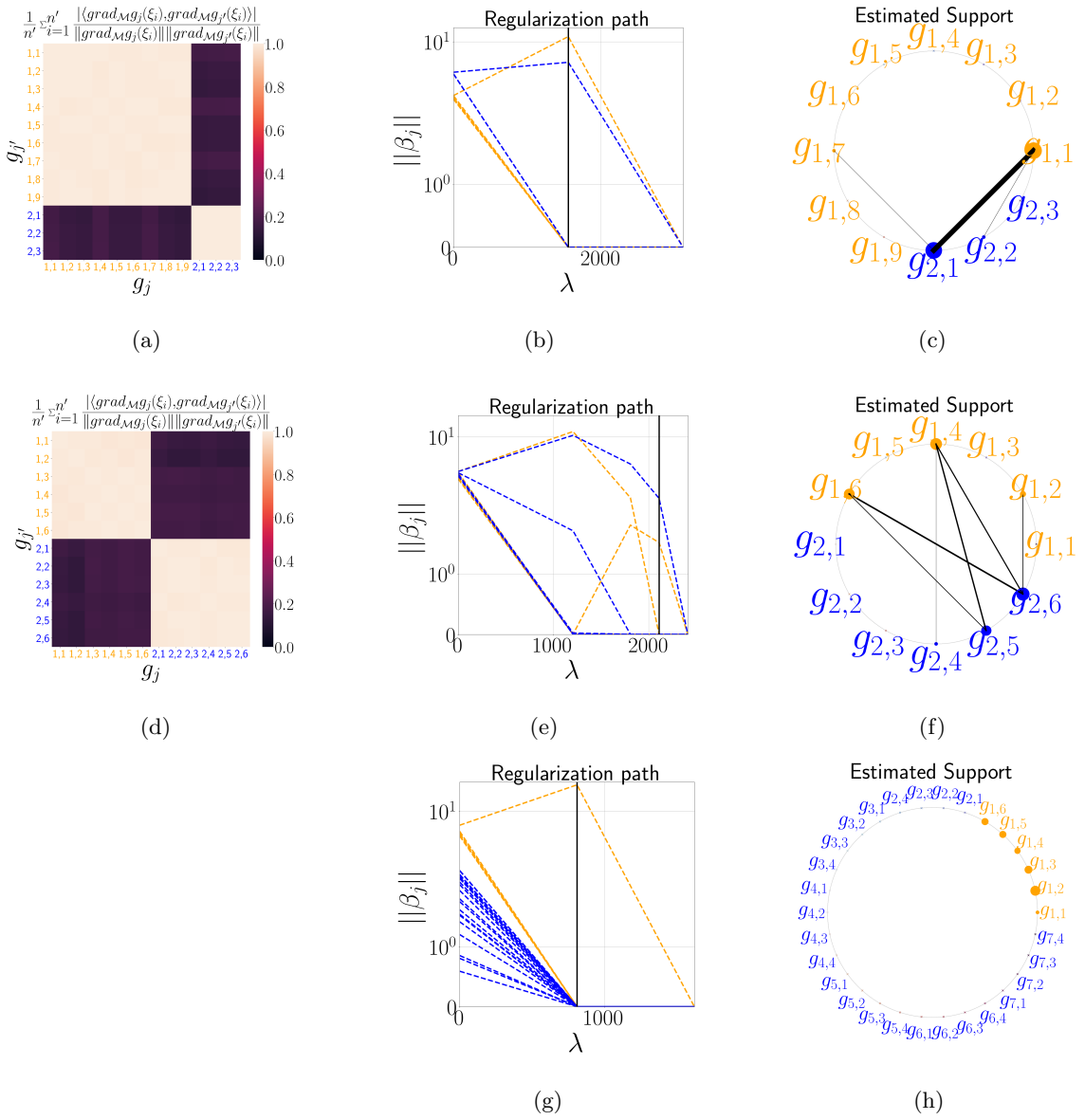


Figure 3.12: Results for MD data with a priori dictionaries given by the bond diagrams in Figure 3.1. The three rows correspond to **Ethanol**, **Malonaldehyde**, and **Toluene**, respectively. Figures 3.12a and 3.12d display pairwise collinearities of dictionary gradients, colored by bond as in Figure 3.1. Toluene, a $1-d$ manifold, has trivial cosines, and so these are not shown. Figures 3.12b, 3.12e, and 3.12g show combined regularization paths of $\|\beta_j\|$ for single replicates. Vertical black lines indicate the tuning parameter at which $|S| = d$. Figures 3.12c, 3.12f, and 3.12h show chord diagrams displaying frequency of support recovery of sets of size d for 25 replicates. As for **RigidEthanol**, two-dimensional support recovery frequency is denoted by chord width, and one-dimensional support recovery frequency is denoted by size of perimeter dot. Note that blue in toluene corresponds to torsions in the benzene ring.

Part II

CLUSTERING WITH STABILITY GUARANTEES

Chapter 4

BACKGROUNDS ON CLUSTERING WITH GUARANTEES

In chapter 5 and 6, we will explore clustering problems. Clustering algorithms often suffer from instability issues as they involve hard optimization problems. In this chapter, we review non-exhaustively on recent developments on clustering with guarantees on Euclidean data and with known number of clusters, denoted by K. There are non-parametric clustering algorithms that estimate K automatically (e.g. density-based clusterings, Dirichlet process mixtures,etc) but they are out-of-scope of this dissertation.

The organization of this chapter is as follows: in section 4.1, we distinguish the two clustering paradigms we discuss in this dissertation: hard-clustering (loss-based) and soft-clustering (model-based). Afterwards, in section 4.2 we use K-means clustering as an example of loss-based clustering and review recent progress on K-means clustering. in section 4.3 we discuss recent development on Gaussian mixture model, which is the most widely used probabilistic model for model-based clustering.

4.1 Clustering problem formulation

Given dataset $\mathcal{D} = \{x_i\}_{i=1}^n$, a clustering $\mathcal{C} = \{C_1, \cdots, C_K\}$ is a partition of the indice set $[n]$ into K non-empty subsets. Each subset $C_1, \cdots, C_K$ is called a cluster.

We often require that the subsets $C_1, \cdots, C_K$ are mutually disjoint, or equivalently, each data point can belong to only one cluster. Such clustering paradigm is called hard clustering. We could use a set of indicator function $\{\mathbf{\Gamma}_{ik} = \mathbf{1}_{i \in C_k}\}$ to represent this clustering. As we will see, such hard clusterings are often obtained by minimizing a loss function. A *loss function* $\text{Loss}(\mathcal{D}, \mathcal{C})$ specifies what kind of clusters the user is interested in via the optimization problem below.

$$\text{Clustering problem:} \quad L^{\text{opt}} = \min_{\mathcal{C} \in \mathbf{C}_K} \text{Loss}(\mathcal{D}, \mathcal{C}), \quad \text{with solution } \mathcal{C}^{\text{opt}} . \tag{4.1}$$

The majority of interesting loss functions result in combinatorial optimization problems (4.1) and known to be hard in the worst case. Here are two examples.

Example: K-Means clustering K-means clustering is one of the most classical clustering methods that dates back to 1967. It minimizes the total within-cluster variances, i.e.

$$\min \mathrm{Loss}_{\mathrm{Km}}(\mathcal{D}, \mathcal{C}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2, \quad \text{with} \quad \mu_k = \frac{1}{n_k} \sum_{i \in C_k} x_i \text{ for all } \quad k \in [K] \, . \quad (4.2)$$

Example: K-medoids clustering This method requires that in each cluster, there is one data point that is the most representative of all data in that cluster, then its loss function can be written as

$$\min \mathrm{Loss}_{Kc}(\mathcal{D}, \mathcal{C}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k|| \quad \text{with} \quad \mu_k = \arg \min_{x_i : i \in C_k} \sum_{j \in C_k} ||x_j - x_i|| \text{ for all } \quad k \in [K] \, . \quad (4.3)$$

In contrast, we could also provide a degree of membership for each pair of data point and cluster. In this case we relax $\mathbf{\Gamma}_{ik}$ from $\{0, 1\}$ to $[0, 1]$. Often we associate the data with a probabilistic model and estimate the degree of membership $\mathbf{\Gamma}_{ik}$ as the posterior probability of each x_i in cluster k. We call the clustering paradigm that generates degree of memberships *soft clustering*.

Example: Gaussian mixture model One of the most widely used probabilistic model for model-based clustering is Gaussian mixture model.

Definition 4.1 (Gaussian mixture model)**.** *Gaussian mixture model assume observed data* $\mathcal{D} = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ *are generated with an unobserved latent label* y_i *following:*

$$y_i \sim \text{Multinomial}(w_1, \cdots, w_K), \quad x_i | y_i \sim \mathcal{N}_D(\mu_{y_i}, \mathbf{\Sigma}_{y_i}) \, .$$

The parameter for a general Gaussian mixture model is $(w_1, \ldots, w_K, \mu_1, \cdots, \mu_K, \mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_K)$ *with the constraint that* $w_k \geq 0, \sum_{k=1}^K w_k = 1$*. The density function is given by*

$$f(x) = \sum_{k=1}^{K} w_k p(x | \mu_k, \mathbf{\Sigma}_k), \text{ with } p(x | \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi |\det \mathbf{\Sigma}|)^{\frac{d}{2}}} \exp\left(-\frac{1}{2} (x - \mu)^\top \mathbf{\Sigma}^{-1} (x - \mu) \right)$$

Given a gaussian mixture model, the clustering assignment is given by γ_{ik} representing the posterior probability of x_i is generated with hidden label $y_i = k$, defined as

$$\gamma_{ik} = \frac{w_k p(x_i|\mu_k, \mathbf{\Sigma}_k)}{\sum_{k'=1}^{K} w_{k'} p(x_i|\mu_{k'}, \mathbf{\Sigma}_{k'})} . \tag{4.4}$$

Essentially, when the model used are parametric, model-based clustering are estimation problems of model parameters. Ideally, parameter estimation methods can all be used for estimating a model such as maximum likelihood estimation of methods of moment estimation. For Gaussian mixture models, we will discuss in detail in section 4.3 of how to implement these standard statistical procedure in Gaussian mixture model context.

4.2 K-means clustering

In this section, we review recent results on K-means clustering with more details. We will first display Lloyd's algorithm, the standard algorithm to solve K-means clustering problem. Then we turn to modern developments of convex relaxations of K-means clustering. Afterwards we will discuss comparing clusterings and evaluating clustering results.

4.2.1 Solve K-means clustering problem

To establish K-means loss functions in a compact form, we introduce matrix representation of clusterings. There are several standard matrix representations of clustering results, in the forms of a mapping from $\mathbf{C}_K$ to some matrix space.

- $\mathbf{\Gamma} : \mathbf{C}_K \to \{0,1\}^{n\times K}, \mathbf{\Gamma}_{ik} = \mathbf{1}_{i\in C_k}$.
- $\mathbf{Z} : \mathbf{C}_K \to [0,1]^{n\times K}, \mathbf{Z}_{ik} = \frac{\mathbf{1}_{i\in C_k}}{\sqrt{n_k}}$.
- $\mathbf{S} : \mathbf{C}_K \to \{0,1\}^{n\times n}, \mathbf{S}_{ij} = \begin{cases} 1, & \text{if } i,j \in C_k \text{ for some } k \in [K] \\ 0 & \text{otherwise} \end{cases}$.
- $\mathbf{X} : \mathbf{C}_K \to [0,1]^{n\times n}, \mathbf{X}_{ij} = \begin{cases} \frac{1}{n_k}, & \text{if } i,j \in C_k \text{ for some } k \in [K] \\ 0 & \text{otherwise} \end{cases}$.

Such representations are very useful in recent clustering literature. It simplifies the presentation of many loss functions in loss based clustering and enables the development of convex relaxation techniques on hard clustering. For example, one can show that, let $\tilde{\mathbf{A}} : \tilde{\mathbf{A}}_{ij} = ||x_i - x_j||^2$ be the pairwise squared distance matrix of dataset $\mathcal{D}$, K-means clustering is equivalent to minimizing the loss function

$$\min \mathrm{Loss}_{\mathrm{Km}}(\mathcal{D}, \mathcal{C}) = \langle \tilde{\mathbf{A}}, \mathbf{X} \rangle \ . \tag{4.5}$$

Directly optimizing loss functions such as in (4.2) and (4.3) is hard combinatorial optimization problem. For any given dimension d and number of clusters K, enumeration of all possible clusterings can be done in $O(n^{dK+1})$ [Inaba et al., 1994]. Further, finding exact solution of K-means is NP-hard even when $d = 2, K \geq 2$ [Aloise et al., 2009] or $K = 2, d \geq 3$ [Mahajan et al., 2012].

Iterative algorithms that find local optimizers are developed. The standard method for solving K-means clustering is Lloyd–Forgy algorithm [Lloyd, 1982]. Its major steps are displayed in algorithm LLOYD. Although empirically, better initialization techniques such as K-means++ performs better [Arthur and Vassilvitskii, 2007], this algorithm does not have theoretical guarantee of convergence to global optimum.

Algorithm 7 LLOYD

1: **Input:**Dataset $\mathcal{D} = \{x_i\}_{i=1}^n$, number of clusters K.
2: Initialize: K centers $\mu_1^{(0)}, \cdots, \mu_K^{(0)} \in \mathbb{R}^d$.
3: **while** not converged **do**
4: Update cluster assignment: $C_k^{(t)} \leftarrow \{i : k = \arg\min_{k \in [K]} ||x_i - \mu_k^{(t)}||\}$.
5: Update cluster mean: $\mu_k^{(t+1)} \leftarrow \frac{1}{|C_k^{(t)}|} \sum_{i \in C_k^{(t)}} x_i$.
6: **end while**

4.2.2 Clustering with guarantee by convex relaxations

A *convex relaxation* of the problem (4.1) is an optimization problem defined as follows. Let $\mathcal{X}$ be a convex set in a matrix space such that $\mathcal{X} \supset \{\mathbf{X}(\mathcal{C}), \mathcal{C} \in \mathbf{C}_K\}$. Extend $\mathrm{Loss}_{\mathrm{Km}}(\mathcal{D}, \mathcal{C})$

to $\text{Loss}_{\text{Km}}(\mathcal{D}, \mathbf{X})$, convex in $\mathbf{X}$ for all $\mathbf{X} \in \mathcal{X}$. Then,

$$L^* = \min_{\mathbf{X} \in \mathcal{X}} \text{Loss}_{\text{Km}}(\mathcal{D}, \mathbf{X}), \quad \text{with solution } \mathbf{X}^* \tag{4.6}$$

is a convex relaxtion for the clustering problem (4.1). In the above, the representation $\mathbf{X}(\mathcal{C})$ can also be changed to other representation matrices ($\mathbf{S}, \mathbf{Z}$,etc), or a different injective mapping of $\mathbf{C}_K$ into a Euclidean space. Because $\mathcal{X} \supset \mathbf{C}_K$, we have $L^* \leq L^{\text{opt}}$ and $\mathbf{X}^*$ is generally not a clustering matrix.

For example, multiple convex relaxations exist for K-means clustering. These observations are based on the following proposition.

Proposition 4.1. *For any clustering $\mathcal{C}$, its clustering matrix $\mathbf{X} = \mathbf{X}(\mathcal{C})$ has the following property: $\mathbf{X}_{ij} \leq \mathbf{X}_{ii}, \forall i, j \in [n]$ $\mathbf{X}\mathbf{1} = \mathbf{1}, \text{Tr}\,\mathbf{X} = K, ||\mathbf{X}||_F^2 = K, \mathbf{X}$ is positive semidefinite.*

Then convex relaxation for K-means clustering can be established through

- LP relaxation [Awasthi et al., 2015a]:

$$\begin{aligned} \min_{\mathbf{X} \in \mathbb{R}^{n \times n}} & \langle \tilde{\mathbf{A}}, \mathbf{X} \rangle \\ s.t.\ & \mathbf{X}_{ij} \in [0, 1], \forall i, j \in [n] \\ & \mathbf{X}_{ij} \leq \mathbf{X}_{ii}, \forall i, j \in [n] \\ & \mathbf{X}\mathbf{1} = \mathbf{1} \\ & \text{Tr}\,\mathbf{X} = K\ . \end{aligned}$$

- SDP relaxation [Awasthi et al., 2015a]:

$$\begin{aligned} \min_{\mathbf{X} \in \mathbb{R}^{n \times n}} & \langle \tilde{\mathbf{A}}, \mathbf{X} \rangle \\ s.t.\ & \mathbf{X}_{ij} \geq 0, \forall i, j \in [n] \\ & \mathbf{X}_{ij} \succcurlyeq 0 \\ & \mathbf{X}\mathbf{1} = \mathbf{1} \\ & \text{Tr}\,\mathbf{X} = K\ . \end{aligned}$$

Under stochastic ball model, such convex relaxations can provide recovery guarantee for K-means clustering. Suppose that data are generated from a population $P = K^{-1}\sum_{i=1}^{K} P_k$, where each P_k is supported rotaionally symmetrically on a unit ball centered at μ_k. Assume that the minimal distance of centers of these balls is $c := \min_{i\neq j}||\mu_i - \mu_j||$. Then the correct clustering assignment can be achieved if $c > 4$ for LP relaxation and can be achieved for $c > 2\sqrt{2}(1 + 1/\sqrt{d})$. Later in [Iguchi et al., 2017], this separation condition is relaxed to $c > 2 + K^2\,\mathrm{Cond}(\mu_1, \cdots, \mu_K)/d$, where Cond characterizes the ratio between maximum center separation and minimum center separation.

Also [Awasthi et al., 2015a] provides a pathetic example such that there exists pathetic examples of stochastic ball model that the K-means++ with constant probability will converge to a bad local minimum, even if the sample size tends to infinity and separation between ball centers are large. In such pathetic cases, however, method based on convex relaxations can still guarantee exact recovery.

Convex relaxations can also be established for other clustering problems. For graph partitioning problems, Xing and Jordan [2003] introduced two relaxations based on Semi-Definite Programs (SDP). Correlation clustering, a graph clustering problem appearing in image analysis, has been given an SDP relaxation in Swamy [2004] and Ahmadian and Swamy [2016]. For community detection under the Stochastic Block Model [Holland et al., 1983] several SDP relaxations have been recently introduced by Chen and Xu [2016], Vinayak et al. [2014] and Jalali et al. [2016] as well as Sum-of-Squares relaxations for finding hidden cliques, in Deshpande and Montanari [2015]. For centroid based clustering, we have Linear Program (LP) based relaxations for K-medians by Charikar and Guha [1999] and K-means Awasthi et al. [2015b] and more recent, tighter relaxations via SDP in Awasthi et al. [2015a]. Relaxations exist also for exemplar-based clustering [Zhu et al., 2014]. For hierarchical clustering in the cost-based paradigm introduced by Dasgupta [2016], we have LP relaxations introduced by Roy and Pokutta [2016], Charikar and Guha [1999], Charikar and Chatziafratis [2017].

4.2.3 External evaluation: comparing two clusterings

Although as an unsupervised method, there is no ground truth of clustering on the whole dataset, it is often the case that labels based on domain knowledge are accessible. Thus, it is necessary to compare clustering obtained from clustering algorithm and the "true" label. Such evaluation procedure is called *external* evaluation of clusterings. Therefore, several methods of comparing two clusterings appear to determine if the algorithm output is consistent with domain knowledge, i.e. ground truth clustering.

Confusion matrix Confusion matrix is a fundamental tool used to compare two hard clusterings. When $\mathcal{C} = \{C_k\}_{k=1}^{K}, \mathcal{C}' = \{C'_{k'}\}_{k'=1}^{K'}$ are two clusterings of the same data set $\mathcal{D} = \{x_i\}_{i=1}^{n}$, we introduce the confusion matrix $\mathbf{M}$ to be

$$\mathbf{M} \in \mathbb{N}_{+}^{K \times K'}, \text{with } \mathbf{M}_{kk'} = |C_k \cap C'_{k'}| \, .$$

We further denote

$$\mathbf{M}_{\cdot k'} = \sum_{k=1}^{K} \mathbf{M}_{kk'}, \quad \mathbf{M}_{k\cdot} = \sum_{k'=1}^{K'} \mathbf{M}_{kk'} \, .$$

Earth mover's distance (misclassification error rate) between two clusterings $\mathcal{C}, \mathcal{C}'$ over the same set of n points is

$$d^{EM}(\mathcal{C}, \mathcal{C}') \; = \; 1 - \frac{1}{n} \max_{\pi \in \mathbb{S}_K} \sum_{k=1}^{K} \left(\sum_{i \in C_k \cap C'_{\pi(k)}} 1 \right), \tag{4.7}$$

where π ranges over the set of all permutations of K elements $\mathbb{S}_K$, and $\pi(k)$ indexes a cluster in $\mathcal{C}'$.

Using the $\mathbf{X}$ representation of clustering, the earth-mover distance can be upper bounded through matrix distance through the following theorem.

Theorem 4.1 (Meilă [2012], Theorem 9)**.** *For two clusterings $\mathcal{C}, \mathcal{C}'$ with the same number of clusters K, denote $\mathcal{C} = \{C_1, \ldots C_K\}$, and $w_{\min} = \frac{1}{n} \min_{[K]} |C_k|$, $w_{\max} = \frac{1}{n} \max_{[K]} |C_k|$. Then, for any $\epsilon \leq w_{\min}$, if $\frac{1}{2} \|\mathbf{X}(\mathcal{C}) - \mathbf{X}(\mathcal{C}')\|_F^2 \leq \frac{\epsilon}{w_{\max}}$, then $d^{EM}(\mathcal{C}, \mathcal{C}') \leq \epsilon$.*

There exist other metrics to compare two clusterings, examples include rand index, adjusted rand index, Fowlkes-Mallows scores, normalized mutual information, etc. Among them, the most popular one is adjusted rand index, given by

$$AdjRand(\mathcal{C},\mathcal{C}') = \frac{\sum_{k=1}^{K}\sum_{k'=1}^{K'}\binom{\mathbf{M}_{kk'}}{2} - \left[\sum_{k=1}^{K}\binom{\mathbf{M}_{k\cdot}}{2}\right]\left[\sum_{k'=1}^{K'}\binom{\mathbf{M}_{\cdot k'}}{2}\right]\Big/\binom{n}{2}}{\left[\sum_{k=1}^{K}\binom{\mathbf{M}_{k\cdot}}{2} + \sum_{k'=1}^{K'}\binom{\mathbf{M}_{\cdot k'}}{2}\right]\Big/2 - \left[\sum_{k=1}^{K}\binom{\mathbf{M}_{k}}{2}\right]\left[\sum_{k'=1}^{K'}\binom{\mathbf{M}_{\cdot k'}}{2}\right]\Big/\binom{n}{2}} .$$

4.2.4 Internal validation and resampling stability

Without ground truth clustering, it is then important to develop internal clustering metrics that measure the quality of clusterings. Such metrics include siloutte score [Rousseeuw, 1987], Gap statistics [Tibshirani et al., 2001], etc. Mostly are developed to select number of clusters. One closely related concept to this dissertation is *resampling stability*, which are detailed reviewed in [Ben-David et al., 2006, Ben-David and von Luxburg, 2008, Ben-David et al., 2007, von Luxburg, 2009]. The idea is the assumption that if the number of clusters is correct, then the clustering should be stable with respect to resampling of data. The procedure is summarized as follows:

- Construct $b = 1, 2, \cdots, B$ bootstrapped samples $\mathcal{D}_b^* = \{x_i^*\}_{i=1}^n$ and a different K-means clustering $\mathcal{C}_b$.

- Estimate the instability metric defined as [1]

$$\widehat{Instab}(K,\mathcal{D}) = \frac{2}{B(B-1)}\sum_{1\le b<b'\le B} d^{EM}(\mathcal{C}_b,\mathcal{C}_{b'}) \ . \tag{4.8}$$

If we further assume that $\mathcal{D}$ is i.i.d. sample from some distribution P, then there are more detailed results. We define $Instab(K,n) = \mathbb{E}\widehat{Instab}(K,n)$

- [Bubeck et al., 2009]: when P is a Gaussian mixture in one dimension with $K = 2$ well separated components, then with high probability over sampling procedure, K-means

[1] Essentially $\mathcal{C}_b, \mathcal{C}_{b'}$ are obtained on different data. However since they are both K-means clustering, they are clustering induced by two sets of centers. Then the distance can be computed based on the union of data $\mathcal{D}_b^* \cup \mathcal{D}_{b'}^*$ with the label induced by the centers.

output (with K-means++ initialization) when setting $K = 2$ is stable; With constant probability is instable when $K = 3$.

- [Ben-David et al., 2006, Ben-David et al., 2007]: denote $L^{\infty}(\mathcal{C}) = \lim_{n\to\infty} \mathbb{E}\,\mathrm{Loss}_{\mathrm{Km}}(\mathcal{D}, \mathcal{C})$. Then $\lim_{n\to\infty} Instab(K, n) = 0$ if and only if L^{∞} has a unique global minimum.

- [Shamir and Tishby, 2008a,b, 2009]: When L^{∞} has a unique global minimum solution, the renormalized instablity criterion $\sqrt{n}Instab(K, n)$ converges in probability to a random variable depending on K and the center of the unique global minimum.

It is worthwhile to notice that, in most cases P has a unique global minimum no matter what value K is. This is conjectured to be an artifact of finding solutions of K-means on population level [von Luxburg, 2009]. As concluded in [Ben-David and von Luxburg, 2008], resampling stability is not ideal for selecting number of K, due to multiple pathetic counter examples.

4.2.5 Large sample properties of K-means loss

Finally, we briefly discuss large sample property of K-means loss under mild assumptions of data generating distribution P.

Non-asymptotic Glivenko-Cantelli result Depending on different assumptions of P, Glivenko-Cantelli type results have been established in the form that with probability at least $1 - \delta$ over resampling size n sample $\mathcal{D}$ from P

$$\sup_{\mathcal{C}\in\mathbf{C}_K(\mathcal{D})} |L^{\infty}(\mathcal{C}) - \mathrm{Loss}_{\mathrm{Km}}(\mathcal{D}, \mathcal{C})| \leq \Psi(n, \delta) \ .$$

- [Maurer and Pontil, 2010]: assume P is compactly supported in the sense that $P(||X_i|| \leq R) = 1$ for some $R > 0$, then it holds for

$$\Psi(n, \delta) = R^2 K \sqrt{\frac{18\pi}{n}} + R^2 \sqrt{\frac{8\log(1/\delta)}{n}} \ . \tag{4.9}$$

- [Telgarsky and Dasgupta, 2013]: assume $p-$the order moment of P is bounded by M, where $p \geq 4$ and is a multiple of 4, then it holds for

$$\Psi(n,\delta) = n^{-\frac{1}{2}+\min\{\frac{1}{4},\frac{2}{p}\}} \left(4 + (72c_1^2 + 32M_1^2)\sqrt{\frac{1}{2}\log\left(\frac{3(nN_1)^{dK}}{\delta}\right)} + \sqrt{\frac{2^{p/4}ep}{8n^{1/2}}}\left(\frac{6}{\delta}\right)^{\frac{4}{p}}\right) . \tag{4.10}$$

where

$$c_1 = (2M)^{1/p},\ M_1 = M^{1/(p-2)} + M^{2/p},\ N_1 = 2 + 576(c_1 + c_1^2 + M_1 + M_1^2) . \tag{4.11}$$

Consistency and Asymptotic normality When P is such that L^∞ has a unique global minimizer at centers $\mu = (\mu_1^*, \cdots, \mu_K^*)$, let $\hat{\mu} = (\hat{\mu}_1, \cdots, \hat{\mu}_K)$ be the global minimizer of $\text{Loss}(\mathcal{C}, \mathcal{D})$, then

- [Pollard, 1981]: $\{\hat{\mu}_1, \cdots, \hat{\mu}_K\} \to \{\mu_1^*, \cdots, \mu_K^*\}, a.e.$

- [Pollard, 1981]: $\sqrt{n}(\hat{\mu} - \mu)$ is asymptotically normal distributed.

4.3 Model-based clustering based on Gaussian mixture model

In this section, we introduce background knowledge of recent development of Gaussian mixture model. As discussed earlier, essentially this is a parameter estimation problem. Denote the parameter space

$$\Theta = \left\{\theta : \theta = (w_1, \cdots, w_K, \mu_1, \cdots, \mu_K, \mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_K), \sum_{k=1}^{K} w_k = 1, w_k \geq 0\right\}$$

Ideally this could be solve through maximum likelihood estimation over parameter space

$$\max_{\theta\in\Theta} L(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log(\sum_{k=1}^{K} w_k p(x_i|\mu_k, \mathbf{\Sigma}_k)) , \tag{4.12}$$

where $p(x|\mu_k, \mathbf{\Sigma}_k)$ is the density function of $d-$dimensional multivariate Gaussian distribution with mean μ_k and covariance $\mathbf{\Sigma}_k$.

However, the likelihood function has several undesired property.

- Parameter space Θ has high dimension $(K + dK + d^2K - 1)$.

- Likelihood function $L(\theta)$ is unbounded: this is the case even with $K = 2$. The pathetic case is one of the Gaussian components tend to degenerate to a point mass.

- Optimization problem is neither convex nor concave. In fact, finding exact solution is NP-hard [Tosh and Dasgupta, 2018].

4.3.1 Expectation-Maximization algorithm and recent results

The current standard way of estimating GMM parameters is expectation-maximization(EM) algorithm. The major steps of EM is showed in algorithm EM.

Algorithm 8 EXPECTATION-MAXIMIZATION

1: **Input:** Dataset $\mathcal{D} = \{x\}_{i=1}^n$, intializing parameters $\theta^{(0)} = (w_1^{(0)}, \cdots, w_K^{(0)}, \mu_1^{(0)}, \cdots, \mu_k^{(0)}, \mathbf{\Sigma}_1^{(0)}, \cdots, \mathbf{\Sigma}_K^{(0)}$.

2: Initialize $\theta^{(0)} \leftarrow (w_1^{(0)}, \cdots, w_K^{(0)}, \mu_1^{(0)}, \cdots, \mu_k^{(0)}, \mathbf{\Sigma}_1^{(0)}, \cdots, \mathbf{\Sigma}_K^{(0)})$.

3: **while** not converged **do**

4: **E-Step** Compute expected complete-data likelihood through $\gamma_{ik}^{(t)}$ as equation 4.4 and

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik}^{(t)}(\log w_k^{(t)} + \log p(x|\mu_k^{(t)}, \mathbf{\Sigma}_k^{(t)})) \ .$$

5: **M-Step** Maximize $Q(\theta|\theta^{(t)})$ and obtain

$$w_k^{(t+1)} = \frac{\sum_{i=1}^n \gamma_{ik}^{(t)}}{n},\ \mu_k^{(t+1)} = \frac{\sum_{i=1}^n \gamma_{ik}^{(t)} x_i}{\sum_{i=1}^n \gamma_{ik}^{(t)}},\ \mathbf{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^n \gamma_{ik}^{(t)}(x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^\top}{\sum_{i=1}^n \gamma_{ik}^{(t)}} \ .$$

6: **end while**

There are multiple variants of EM updates. For example, in generalized EM, it is not necessarily to maximize Q function in the M-step, instead, it is sufficient to increase the value of it. In first order EM [Balakrishnan et al., 2017], select a step size α, then the parameters are updated through $\theta^{(t+1)} = \theta^{(t)} + \alpha\nabla_\theta Q(\theta^{(t)}|\theta^{(t)})$. When Gaussian mixtures are in high dimension, with sparsity assumptions some algorithms specially designed for high dimensional scenarios are developed.

Although empirically, EM performs very well, there is still pathetic cases that EM will converge to bad local minima [Jin et al., 2016a] with constant probability under random initialization even if the sample size tends to infinity. Therefore, the recent progress on EM's (in Gaussian mixture model especially) theoretical guarantees rely on additional assumptions of the distribution.

A key property [mcl, 2000] that guarantees EM convergence is called *self-consistency property*: that is if θ^* is the global optimizer of the likelihood function 4.12, then it holds that $\theta^* = \arg\max Q(\theta|\theta^*)$.

- **separation of Gaussian means**: [Dasgupta and Schulman, 2007] assumes that when the mixtures of Gaussians are well separated[2], EM will converge to the true parameters in merely two iterations.

- **local concavity**: [Balakrishnan et al., 2017] considers first order EM, and showed that population EM converges to global minimum if the initial points are close to the true parameters and certain concavity assumptions are imposed on $q(\theta) = Q(\theta|\theta^*)$.

- **sparsity** [Wang et al., 2015] (Liu Han High-dimensional EM) considers the statistical guarantee of two-Gaussian mixtures with the same spherical covariances and the mean vector of each component is sparse. Similar settings are explored in [] to introduce guarantees for regularized EM.

- **overspecified model** Recently, [Dwivedi et al., 2018] discussed the statistical guarantee of EM performed on a component number K that is larger than the true model.

4.3.2 Separation-based Estimation

A different line of research regarding fitting Gaussian mixture model relies on the separation assumption on the mean of the Gaussian means. These algorithms all assumes that the minimal separation

$$c := \min_{i \neq j} \frac{||\mu_i - \mu_j||}{\max\{\lambda_{\max}(\mathbf{\Sigma}_1), \lambda_{\max}(\mathbf{\Sigma}_2)\}} \tag{4.13}$$

[2]We will define this well-separation later

is sufficiently large. Then statistical guarantee can be obtained based on their different estimation strategies. Table 4.1 shows a few examples of research work in this line. In this table, we are considering estimating a Gaussian mixture with K components in $\mathbb{R}^d$. The minimal component weight is $w_{\min} := \min_{k\in[K]} w_k$.

Table 4.1: GMM estimation with guarantee based on separations

Reference	Separation condition c	Mixture model class
[Dasgupta, 1999]	$c = O(\sqrt{d})$	General Gaussian
[Vempala and Wang, 2004]	$c = O(d^{\frac{1}{4}})$	Spherical Gaussian
[Arora and Kannan, 2001]	$c = \begin{cases} O(d^{\frac{1}{4}}), & \text{Same variance,} \\ 0, & \text{One is sufficiently larger.} \end{cases}$	General Gaussian
[Achlioptas and McSherry, 2005]	$c = O(\sqrt{1/w_{\min}})$	General Gaussian; Log-concave and concentrated models.
[Kannan et al., 2008]	$c = O(K^{\frac{3}{2}}/w_{\min}^2)$	Log-concave models
[Dasgupta and Schulman, 2007]	$c = O(d^{\frac{1}{4}})$	Spherical Gaussian, $\epsilon-$accuracy in mean
[Regev and Vijayaraghavan, 2017]	$c = O(\sqrt{\log K})$	Spherical Gaussian

Other results show that it is necessary to assume separations on the components to achieve general recovery guarantee of Gaussian mixtures. [Kalai et al., 2010] shows that there exists two GMMs with very small total variation distance ($O(e^{-k/30})$) but are not close $\Omega(1)$ in parameter distance without any separation conditions. [Regev and Vijayaraghavan, 2017] constructs examples such that two GMMs both with $o(\sqrt{\log K})$ separation that has large parameter distance and small total variation distance.

4.3.3 Method-of-moments based estimation

Another recently developed series of estimation procedure is based on moments estimates.

- [Hsu and Kakade, 2013]: Estimate mixtures of spherical Gaussians, assuming that the space spanned by the component centers has dimension K. This method requires estimation of up to third order moments and use a spectral algorithm to find each components' parameters. Consistency guarantees can be achieved under correctly specified model.

- [Ge et al., 2015] generallizes [Hsu and Kakade, 2013] by extending spherical mixtures of Gaussians to general covariance matrices. This method used tensor decomposition techniques and requires estimation of up to $6-$th moments. This method provide guarantee that a smoothed Gaussian mixture can be learned with polynomially dependence on n and K.

- [Wu and Yang, 2020]: Estimate one dimensional Gaussian location mixture model (i.e. assuming the variances of each component to be the same). This method requires estimation of up to $2K$ order of moments, and then use a SDP criterion to perform a denoise procedure on the estimated moments. When the number of components $K = O(\log n / \log \log n)$, the authors show that this method reaches the optimal rate.

- [Doss et al., 2020]:: Estimate Gaussian location mixture model in high dimension. When K is bounded, the centers are bounded and dimension d to be as large as n, then this paper proposed a method that estimate the location parameters within time

complexity $O(nd^2 + n^{5/4})$. This method estimate projections of mean parameters into one dimension through the method in [Wu and Yang, 2020] and recover the multi-dimension parameters by a minimal W_1 distance estimator.

4.3.4 Identifiability under total variation distance

Finally, we shall present recent results that studies the parameter identifiability of GMM. In these results two GMM are distinguished through their total variation distance, i.e. for two distributions P, P' with density p, p' respectively, define

$$TV(P, P') = \frac{1}{2} \int |p - p'| \mathrm{d}x \ .$$

We first consider $TV(P, P') \to 0$. Let $G = \sum_{k=1}^{k} w_k \delta_{(\mu_k, \mathbf{\Sigma}_k)}, G' = \sum_{k'=1}^{K'} w_{k'} \delta_{(\mu_{k'}, \mathbf{\Sigma}_{k'})}$, we can view each GMM as an instance of atomic distributions on the parameter space. Hence, we can consider the W_r distance of parameters defined by

$$W_r(G, G') = \left(\inf_{\Pi} \sum_{i,j} q_{ij} (||\mu_i - \mu'_j|| + ||\mathbf{\Sigma}_i - \mathbf{\Sigma}'_j||)^r \right)^{\frac{1}{r}} \ ,$$

where $\Pi = (q_{ij})_{i=1,\cdots,K, j=1,\cdots,K'}$ is a joint distribution over pairs (i, j).

Write GMM associate with G to be P_G, then following results have been established in [Ho and Nguyen, 2016]:

- **Exact fitted** Given G_0 an atomic distribution with K_0 atoms, there are positive constants ϵ_0, C_0 (depending on G_0) such that as long as G has K_0 atoms and $W_1(G, G_0) \leq \epsilon_0$, it holds that $TV(P_G, P_{G_0}) \geq C_0 W_1(G, G_0)$.

- **Over fitted** Given G_0 with an atomic distribution with K_0 atoms, there are positive constants ϵ_0, C_0 depending on G_0 such that as long as G has $K \geq K_0 + 1$ atoms and $W_2(G, G_0) \leq \epsilon_0$, it holds that $TV(P_G, P_{G_0}) \geq C_0 W_2^2(G, G_0)$.

A second scenario is $TV(P, P')$ being a positive constant. Recent results on robust learning of Gaussian mixture models [Kane, 2021, Liu and Moitra, 2021, Bakshi et al., 2022] establish the following identifiability result.

Theorem 4.2. *Consider two gaussian mixtures* $P = \sum_{k=1}^{K} w_k \mathcal{N}_d(\mu_k, \mathbf{\Sigma}_k), P' = \sum_{k'=1}^{K'} w'_{k'} \mathcal{N}_d(\mu'_k, \mathbf{\Sigma}'_k)$. *Then there exists a partition (could be trivial) of* $[K]$ *into sets* $R_0, \cdots, R_l$ *and a parition of* $[K']$ *into sets* $S_0, S_1, \cdots, S_l$ *such that*

1. *There exists a sufficiently small* $g \in (0,1)$ *such that or any* $i \in \{1, \cdots, l\}$,

$$|\sum_{j \in R_i} w_j - \sum_{j \in S_i} w'_j| = O(\epsilon^{g^{\max\{K,K'\}}}) ;$$
$$TV(\mathcal{N}_d(\mu_k, \mathbf{\Sigma}_k), \mathcal{N}_d(\mu_{k'}, \mathbf{\Sigma}_{k'})) = O(\epsilon^{g^{\max\{K,K'\}}}) ;$$

2. $|\sum_{j \in R_0} w_j - \sum_{j \in S_0} w'_j| = O(\epsilon^{g^{\max\{K,K'\}}})$.

When the components of P, P' to be assumed to be separated by certain total variation distance and $\min_{k \in [K]} w_k, \min_{k' \in [K'] w_{k'}}$ is lower bounded, then one further conclude that $K = K'$ and the each component in partition in the previous theorem contains exactly one components. In other words, there exist a one-to-one correspondence between components of P, P' such that their parameter distance is close.

Chapter 5

OPTIMALITY INTERVAL OF K-MEANS CLUSTERING

5.1 Introduction

As we reviewed in the background chapter 4, there are many results for K-means clustering with guarantee both empirically or theoretically. However, these remarkable results remain disconnected from the practice of clustering because (i) they are asymptotic and depend on unspecified constants, (ii) they assume that the global optimum is found. Most importantly, these results do not distinguish between "clusterable" data and data that are not. The property of having clusters is not generic. We focus on the best cases, when the data are clusterable, as described shortly.

In this chapter, we show that resampling stability is not sufficient to characterize the property of a clustering result. We then turn to a related characterization of clustering stability in the form of the following theorem.

Stability Theorem $(*)$**.** *Given a clustering* $\mathcal{C}$ *of data set* $\mathcal{D}$*, a loss function* Loss*, there is a pair* (γ, ϵ) *such that* $d^{EM}(\mathcal{C}, \mathcal{C}') \leq \epsilon$ *whenever* $\text{Loss}(\mathcal{D}, \mathcal{C}') \leq \text{Loss}(\mathcal{D}, \mathcal{C}) + \gamma$*.*

A clustering satisfying the Stability Theorem is called (γ, ϵ)-*stable* (or simply *stable*), and a data set that admits an ϵ-stable clustering is said to be (γ, ϵ)-*clusterable* (or simply *clusterable*). The major loss function we shall study in this chapter is the K-means loss.

Here the data set $\mathcal{D}$ can be a population or a finite sample from that population. When we are working with population, such stability notion gives a more accurate description of how K-means loss behaves near global minimizers. Afterwards, we will establish a reliability result to connect population stability and sample stability.

We also provide an algorithm that provides stability guarantee of finite sample clustering. Given a sample $\mathcal{D}$ and a clustering $\mathcal{C}$ of this sample, we propose a framework for providing guarantees for a given clustering $\mathcal{C}$ of a given set of points, without making untestable

assumptions about the data generating process. In the simplest terms, the algorithm enables a user to tell, with no prior knowledge, if the clustering $\mathcal{C}$ returned by a clustering algorithm is meaningful? or correct? or optimal?

In the above definition $d^{EM}(\mathcal{C}, \mathcal{C}')$ is the widely used *earth mover's distance (EM)* distance between partitions of a set of n objects. Theorem $*$ would be trivial if ϵ was arbitrarily large; hence, we shall always require that ϵ be small, for instance, smaller than the relative size of the smallest cluster in $\mathcal{C}$.

A Stability Theorem states that any way to partition the data which is very different from $\mathcal{C}$ will result in higher Loss. Hence, the data supports only one way to be partitioned with low Loss, and small perturbations thereof. It should also be evident that it is not possible to obtain such guarantees in general; they can only exist for specific data sets and clusterings, as illustrated in Figure 5.1.

From Theorem $(*)$ it follows immediately that C^{opt}, the clustering minimizing Loss on $\mathcal{D}$, together with the entire *sublevel set* $\{\mathcal{C}', \text{Loss}(\mathcal{C}') \leq \text{Loss}(\mathcal{C})\}$ is contained in the ball of radius ϵ centered at $\mathcal{C}$. We call this ball an *optimality interval (OI)* for Loss on $\mathcal{D}$. Somehow abusively, we will occasionally call ϵ itself an OI. Thus, from a practitioner's point of view, Stability Theorem gives a guarantee that the clustering found, $\mathcal{C}$, is no more than ϵ-away from the optimum, and more importantly, that $\mathcal{C}$ is a set of similar clusterings, which are the only possible groupings of the data to attain low Loss. Thus the theorem provides an internal guarantee of (almost) correctness: given a Loss as an implicit model for "good clustering", the theorem can confirm that the sample $\mathcal{D}$ and the found clustering $\mathcal{C}$, fit the model well.

From a broader perspective, this chapter sits firmly within descriptive statistics. By replacing distributional assumptions with quantities computed from data, we can obtain post-inference guarantees in a model-free, finite-sample framework. The main technical contribution that enables these results is the innovative use of convex analysis. By formulating the stability theorem statement as a convex problem, we are able to obtain explicity ϵ values tractably. Even though the ϵ bounds we obtain are distribution free, worst case, empirically they can tighter than the best known model-based bounds.

The organization of this chapter is as follows: In 5.2, we show how to compute a stability

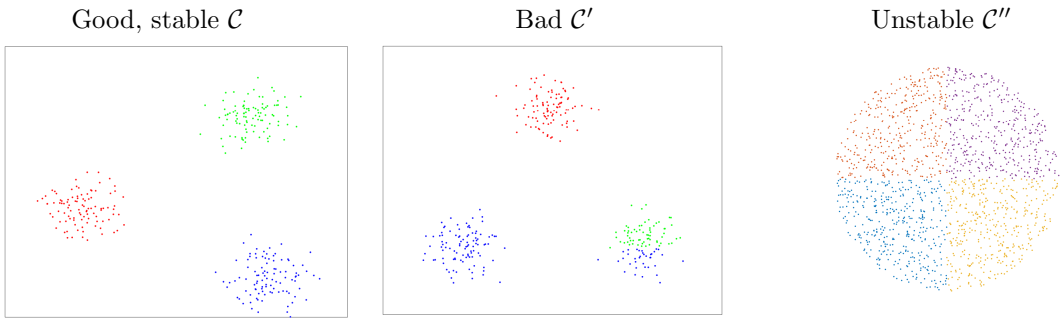


Figure 5.1: Left: a clusterable data set. Stability Theorem applied to these data results in $\epsilon = 10^{-3}$; since $\epsilon < 1/n = 1/200$, this guarantee implies that $\mathcal{C}$ is optimal w.r.t. Loss. Middle: the same data, with a clustering $\mathcal{C}''$ which is not stable. Right: a data set that is not clusterable. The clustering $\mathcal{C}''$ shown is nearly optimal, but it is not stable, as the data admits other clusterings with similar Loss, but very different from $\mathcal{C}''$.

guarantee for K-means clustering with a given random sample. Such stability guarantee can be understood as an optimality interval. Section 5.3 provides the precise definition of stability in the population setting and discuss its relation with the resampling stability. Afterwards in section 5.4 we provide preliminary discussion on stability of a population and of a random sample from that population. Experiments are then shown in 5.5 to demonstrate the algorithm proposed in 5.2.

5.2 Finite sample stability guarantee by convex relaxation

5.2.1 Sublevel set problems and the generic Sublevel Set method

In this section, we introduce a framework to provide (γ, ϵ) stability guarantee of a clustering on finite sample. Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$ are a given data set on $\mathbb{R}^D$. $\mathcal{C}$ is a clustering that partition the indice set $[n]$ to K mutually disjoint sets $C_1, \cdots, C_K$. Still let $w_{\min}, w_{\max}$ be the proportion of data in the smallest and largest cluster, respectively. We consider $\mathbf{X}(\mathcal{C})$ defined in section 4.2 as a matrix representation of a clustering. Let $\mathbf{C}_K(\mathcal{D})$ be the space of all clusterings with K clusters on this data set.

Ideally, (γ, ϵ)-stability of a generic loss function Loss is defined as follows.

$$\text{SS Problem: Original} \quad \epsilon = \max_{\mathcal{C}' \in \mathbf{C}_K(\mathcal{D})} d^{EM}(\mathcal{C}, \mathcal{C}'), \quad \text{s.t. } \text{Loss}(\mathcal{D}, \mathcal{C}') \leq \text{Loss}(\mathcal{D}, \mathcal{C}) + \gamma \ . \tag{5.1}$$

To simplify the presentation, in the remaining of this section we focus on the case $\gamma = 0$, but the procedure applies to $\gamma > 0$ with little modification. When $\gamma = 0$, the solution also has an interpretation as *optimality interval* (OI). As a byproduct of the procedure, once a local optimizing clustering is found, an OI can be computed to show how much difference this local optimizer is with the global minimizer.

Unfortunately, this original SS problem is usually an untractable optimization problem. Instead of providing an exact solution to this problem, we propose to consider a conservative upper bound: we will compute an ϵ' such that ϵ as the solution of problem 5.1 must be upper bounded by ϵ'.

The solution is to use an existing convex relaxation to obtain guarantees of the form $(*)$ for clustering. Given a Loss, its clustering problem (4.1), and a convex relaxation (4.6) for it we proceed as follows:

Step 1: We use the convex relaxation to find a set of good clusterings that contains a given $\mathcal{C}$. This set is $\mathcal{X}_{\leq l} = \{\mathbf{X} \in \mathcal{X}, \text{Loss}(\mathcal{D}, \mathbf{X}) \leq l\}$, the *sublevel set* of Loss, at the value $l = \text{Loss}(\mathcal{D}, \mathcal{C})$. This set is convex when Loss is convex in $\mathbf{X}$.

Step 2: We show that if $\mathcal{X}_{\leq l}$ is sufficiently small, then all clusterings in it are contained in the d^{EM} ϵ-ball $\{\mathcal{C}', \ d^{EM}(\mathcal{C}, \mathcal{C}') \leq \epsilon\}$. This ball is an optimality interval for $\mathcal{D}, \text{Loss}$ and K.

In more detail, consider a dataset $\mathcal{D}$, with a clustering $\mathcal{C} \in \mathbf{C}_K$. Assume for the given Loss a convex relaxation exists, with feasible set $\mathcal{X}$, and let $\mathbf{X}(\mathcal{C})$ be the image of $\mathcal{C}$ in $\mathcal{X}$. We modify the relaxed optimization problem (4.6) to define optimization problems such as the one below, which we call (Relaxed) *Sublevel Set (SS)* problems.

$$\text{SS Problem: Relaxed} \quad \epsilon' = \max_{\mathbf{X}' \in \mathcal{X}} \|\mathbf{X}(\mathcal{C}) - \mathbf{X}'\|, \quad \text{s.t. } \text{Loss}(\mathcal{D}, \mathbf{X}') \leq \text{Loss}(\mathcal{D}, \mathbf{X}(\mathcal{C})) \ . \tag{5.2}$$

The norm $||\,||$ can be chosen conveniently and in the case presented here it will be the Frobenius norm $||\,||_F$ defined in Proposition 4.1. The feasible set for (5.2) is $\mathcal{X}_{\leq \mathrm{Loss}(\mathcal{D},\mathcal{C})}$, a convex set. The convexity and tractability of this SS problem depends on its objective, and we will show that the mapping $\mathbf{X}(\mathcal{C})$ in section 4.2, along with the Frobenius norm, always leads to tractable SS problems. When the SS problem is tractable, then, by solving it we obtain that $||\mathbf{X}(\mathcal{C}') - \mathbf{X}(\mathcal{C})|| \leq \epsilon'$ for all clusterings $\mathcal{C}'$ with $\mathrm{Loss}(\mathcal{D}, \mathcal{C}') \leq \mathrm{Loss}(\mathcal{D}, \mathbf{X}(\mathcal{C}))$.

5.2.2 Optimality intervals for the K-means loss

This section instantiate how our framework work for K-means clustering. The objective is to minimize the *squared error loss*, also known as the *K-means loss*

$$\mathrm{Loss}_{\mathrm{Km}}(\mathcal{D}, \mathcal{C}) \;=\; \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k} ||x_i - \mu_k||^2, \quad \text{with } \mu_k \;=\; \frac{1}{n_k}\sum_{i\in C_k} x_i, \quad \text{for } k \in [K]\,. \tag{5.3}$$

Define the *squared distances matrix* $\tilde{\mathbf{A}}$ by

$$\tilde{\mathbf{A}} \;=\; [\tilde{\mathbf{A}}_{ij}]_{i,j\in[n]}, \qquad \tilde{\mathbf{A}}_{ij} \;=\; ||x_i - x_j||^2\,. \tag{5.4}$$

where $||x||$ denotes the Euclidean norm of x. Furthermore, let $\langle \mathbf{A}, \mathbf{B}\rangle \stackrel{def}{=} \mathrm{trace}(\mathbf{A}^\top\mathbf{B})$ denote the Frobenius scalar product and recall that $||\mathbf{A}||_F = \langle \mathbf{A}, \mathbf{A}\rangle^{1/2}$. It can be shown that $\mathrm{Loss}_{\mathrm{Km}}$ is a function of the matrices $\mathbf{X}$ and $\tilde{\mathbf{A}}$.

$$\mathrm{Loss}_{\mathrm{Km}}(\mathcal{D}, \mathcal{C}) \;\equiv\; \mathrm{Loss}_{\mathrm{Km}}(\tilde{\mathbf{A}}, \mathbf{X}(\mathcal{C})) \;=\; \frac{1}{2n}\langle \tilde{\mathbf{A}}, \mathbf{X}(\mathcal{C})\rangle\,. \tag{5.5}$$

This formulation inspired Peng and Wei [2007] to propose the following convex relaxation of the K-means problem.

$$\min_{\mathbf{X}\in\mathcal{X}} \langle \tilde{\mathbf{A}}, \mathbf{X}\rangle\,. \tag{5.6}$$

where $\mathcal{X} = \{\mathbf{X} \in \mathbb{R}^{n\times n}$, $\mathrm{trace}\,\mathbf{X} = K$, $\mathbf{X}\mathbf{1} = \mathbf{1}$, $\mathbf{X}_{ij} \geq 0$, for $i, j \in [n]$, $\mathbf{X} \succeq 0\}$ is the set of matrices satisfying the conditions in Proposition 4.1. In [Peng and Wei, 2007] it was shown that problem (5.6) can be cast as a *Semidefinite Program (SDP)*.

We use the relaxation (5.6) to obtain OI for K-means. We shall assume that a data set $\mathcal{D}$ is given, and that the user has already found a clustering $\mathcal{C}$ of this data set (by e.g. running the K-means algorithm).

The SDP below corresponds to the SS problem (5.2). The main difference from equation (5.2) is that we used the identity $\|\mathbf{X}(\mathcal{C}) - \mathbf{X}'\|^2 = 2K - 2\langle \mathbf{X}(\mathcal{C}), \mathbf{X}' \rangle$ to obtain a convex minimization objective instead of a norm maximization.

$$(\mathrm{SS}_{\mathrm{Km}}) \quad \kappa(\mathcal{C}) \;=\; \min_{\mathbf{X}' \in \mathcal{X}} \langle \mathbf{X}(\mathcal{C}), \mathbf{X}' \rangle \quad \text{s.t.} \langle \tilde{A}, \mathbf{X}' \rangle \;\leq\; \langle \tilde{\mathbf{A}}, \mathbf{X}(\mathcal{C}) \rangle \;. \tag{5.7}$$

Our main result below states that when the value $\kappa(\mathcal{C})$ is near K, it controls the maximum deviation from $\mathcal{C}$ of any other good clustering.

Theorem 5.1. *Let $\mathcal{D}$ be represented by its squared distance matrix D, let $\mathcal{C}$ be a clustering of $\mathcal{D}$, with $K, w_{\min}, w_{\max}$ as in theorem 4.1, and let $\kappa(\mathcal{C})$ be the optimal value of problem (SS_{Km}). Then, if $\epsilon = (K - \kappa(\mathcal{C}))w_{\max} \leq w_{\min}$, any clustering $\mathcal{C}'$ with $\mathrm{Loss}(\mathcal{C}') \leq \mathrm{Loss}(\mathcal{C})$ is at distance $d^{EM}(\mathcal{C}, \mathcal{C}') \leq \epsilon$.*

When ϵ defined by Theorem 5.1 is smaller than the relative size of the smallest cluster, then $\mathcal{C}$, even though not necessarily optimal, is a representative of a small set that contains the optimal clustering $\mathcal{C}^{\mathrm{opt}}$ as well as all the other clusterings that are as good as $\mathcal{C}$. Sometimes, when $\epsilon < \frac{1}{n}$, as in Figure 5.1, Theorem 5.1 also implies that $\mathcal{C} = \mathcal{C}^{\mathrm{opt}}$. With $\mathcal{C}$ and D known, a user can solve this SDP in practice and obtain an OI defined by ϵ. We summarize this procedure below.

Input Data set with $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ defined as in (5.4), clustering $\mathcal{C}$ with K clusters, $w_{\min}, w_{\max}$, and clustering matrix $\mathbf{X}(\mathcal{C})$.

1. Solve problem ($\mathrm{SS}_{\mathrm{Km}}$) numerically (by e.g. calling a SDP solver); let κ be the optimal value obtained.

2. Set $\epsilon = (K - \kappa)w_{\max}$.

3. **If** $\epsilon \leq w_{\min}$ **then**

   Theorem 5.1 holds: ϵ gives an OI for $\mathcal{C}$.

   else no guarantees for $\mathcal{C}$ by this method.

The above method exemplifies the goals set forth in the Introduction; it depends only on observed and computable quantities, and does not relie on assumptions about the data generating process. These bounds exist only when the data is clusterable. Currently we cannot show that all the clusterable cases can be given guarantees; this depends on the tightness of the relaxation, as well as on the tightness of Step 2 of the SS method.

The SDP relaxation (5.6) is not the only way to obtain a SS problem for Loss_{Km}, and we further illustrate the versatility of the SS method by constructing a second SS problem for this clustering loss. In Awasthi et al. [2015a] the following relaxation to the K-means problem is presented.

$$\min_{\mathbf{X}\in\mathcal{X}_{\text{LP}}} \langle \tilde{\mathbf{A}}, \mathbf{X}\rangle \ . \tag{5.8}$$

In the above, the mapping $\mathcal{C} \to \mathbf{X}(\mathcal{C})$ is the same as in section 4.2; the convex set $\mathcal{X}_{\text{LP}}$ is $\mathcal{X}_{\text{LP}} = \{\text{trace}\,\mathbf{X} = K,\, \mathbf{X}\mathbf{1} = \mathbf{1},\, \mathbf{X}_{ij} \le \mathbf{X}_{ii} \text{ for all } i,j \in [n],\, \mathbf{X}_{ij} \in [0,1] \text{ for all } i,j \in [n]\}$. Relaxation (5.8) can be cast as a Linear Program, making it more attractive from the computational point of view. It is straightforward to state the Sublevel Set problem SS corresponding to (5.8), which is also an LP.

$$\kappa_{\text{LP}}(\mathcal{C}) \ = \ \min_{\mathbf{X}'\in\mathcal{X}_{\text{LP}}} \langle \mathbf{X}(\mathcal{C}), \mathbf{X}'\rangle, \quad \text{s.t. } \langle \tilde{\mathbf{A}}, \mathbf{X}'\rangle \ \le \ \langle \tilde{\mathbf{A}}, \mathbf{X}\rangle \ . \tag{5.9}$$

Since $\kappa_{\text{LP}}(\mathcal{C})$ bounds the same quantity $\langle \mathbf{X}(\mathcal{C}), \mathbf{X}'\rangle$, Theorem 5.1 applies. When multiple OI can be obtained, the tightest one bounds the distance $d^{EM}(\mathcal{C}, \mathcal{C}')$. In Awasthi et al. [2015a] it is shown that the SDP relaxation is strictly tighter than the LP relaxation, for data generated from separated balls. This suggests that the OI from the LP will not be as tight as the SDP OI.

5.3 $(\gamma, \epsilon)-$population stability under K-means clustering loss

We turn our focus to population level in this section. Suppose that the data $\mathcal{D} = \{x_1, \cdots, x_n\}$ are sampled i.i.d. from P, a distribution supported on a subset of $\mathbb{R}^d$. Our ultimate goal is to ask whether the population P is "clusterable" with cluster number K for a certain clustering diagram. In general, a clustering $\mathcal{C}$ is a partition of the space $\mathbb{R}^d = \sqcup_k C_k$, where each C_k is a Borel set with positive Lebesgue measure.

The performance of such partition can be measured by different loss functions. Similar to the finite sample case, we can define a K-means population loss as the weighted sum of within cluster sum-of-square loss, i.e.

$$\text{Loss}_{\text{Km}}(P, \mathcal{C}) = \sum_{k=1}^{K} P(X \in C_k)\mathbb{E}\left[||X - \mathbb{E}[X|X \in C_k]||^2 | X \in C_k\right] . \tag{5.10}$$

For any clustering $\mathcal{C}$, one can identify an initialization set of centers given by $c_k = \mathbb{E}[X|X \in C_k]$ for all $k \in [K]$, one can then perform a population K-means similarly to the Lloyd's algorithm.

- Update cluster assignment: $\tilde{C}_k = \left\{x : k = \arg\min_{k\in[K]}||x - \mu_k||\right\}$,
- Update cluster mean: $\tilde{\mu}^{(k)} = \mathbb{E}[X|X \in \tilde{C}_k]$.

Hence any partition of $\mathbb{R}^d$ that minimizes loss Loss_{Km} must be induced by K *Voronoi centers* $\mu_1, \cdots, \mu_K \in \mathbb{R}^d$, in the sense that any $x \in \mathbb{R}^d$ is assigned to the cluster with closest center.[1] We then identify each clustering $\mathcal{C}$ with the set of its Voronoi centers. This ensures that $\text{Loss}_{\text{Km}}(P, \mathcal{C})$ is well defined as

$$\text{Loss}_{\text{Km}}(P, \mathcal{C}) = \int_{\mathbb{R}^d} \min_{k\in[K]} ||x - \mu_k||^2 P(\mathrm{d}x) . \tag{5.11}$$

This loss functions is exactly the loss function denoted by $L^\infty(\mathcal{C})$ in 4.2. Here we introduce the new notation to emphasize that the loss function also takes P as an input.

Denote by $\mathbf{C}_K(\mathbb{R}^d)$, $\mathbf{C}_K(\mathcal{D})$ the set of all clusterings of $\mathbb{R}^d$, respectively of a fixed data set $\mathcal{D}$, defined by K *distinct* Voronoi centers.[2] With a slight abuse of notation, we will use $\mathcal{C} = \{C_1, \ldots C_K\}$ for clusterings in either set. Minimizing $\text{Loss}_{\text{Km}}(\mathcal{D}, \mathcal{C})$, as defined in (5.11), when we view $\mathcal{D}$ as a population with *finite* support leads to the previous definition of K-means loss in (5.3). Note however that, for an arbitrary $\mathcal{C} \in \mathbf{C}_K(\mathbb{R}^d)$ or $\mathbf{C}_K(\mathcal{D})$, the

[1] $\mathcal{C}$ is defined only up to a zero measure set, but we can ignore such distinctions since Loss_{Km} and d_P^{EM} as defined in this section are invariant to them.

[2] The reader will note that $\mathbf{C}_K(\mathcal{D})$ is only a subset of $\mathbf{C}_K(\mathbb{R}^d)$. We need this restriction to ensure that $\mathbf{C}_K(\mathbb{R}^d)$ has finite VC-dimension. Moreover, the mapping from Voronoi centers to partitions of $\mathbb{R}^d$ is not injective; however, this does not affect the results in this paper.

Voronoi centers do not coincide with the means of the clusters, unless $\mathcal{C}$ is a fixed point of the K-means algorithm.

The earth mover's distance can be directly generalized to $\mathbf{C}_K(\mathbb{R}^d)\times\mathbf{C}_K(\mathbb{R}^d)$ as $\mathbb{E}d^{EM}(\mathcal{C},\mathcal{C}')$, where expectation is over i.i.d. resampling of data from P. We will denote with $d_P^{EM}(\mathcal{C},\mathcal{C}')$ the distance of two clusterings $\mathcal{C},\mathcal{C}' \in \mathbf{C}_K(\mathbb{R}^d)$, and $d^{EM}(\mathcal{C},\mathcal{C}')$ as before for distances of two clusterings in $\mathbf{C}_K(\mathcal{D})$.

A clustering $\mathcal{C} \in \mathbf{C}_K(\mathbb{R}^d)$ is called (γ,ϵ) stable if any clustering $\mathcal{C}' \in \mathbf{C}_K(\mathbb{R}^d)$ with $\mathrm{Loss}(P,\mathcal{C}') \leq \mathrm{Loss}(P,\mathcal{C}) + \gamma$ is at distance $d_P^{EM}(\mathcal{C},\mathcal{C}') \leq \epsilon$. A similar definition holds for (γ,ϵ)-stable clusterings in $\mathbf{C}_K(\mathcal{D})$. If $\mathcal{C}$ is not (γ,ϵ) stable then it is called (γ,ϵ) unstable. Note that (γ,ϵ)-stability (or instability) for any clustering $\mathcal{C}$ implies (weaker) stability (or instability) statements for any other clustering $\mathcal{C}'$ in the $l = \mathrm{Loss}(\mathcal{C}) + \gamma$ sublevel set. For example, if $\mathcal{C}$ is (γ,ϵ) stable, and $\mathrm{Loss}(\mathcal{C}') = \mathrm{Loss}(\mathcal{C}) + \gamma'$, with $\gamma' < \gamma$, then $\mathcal{C}'$ is $(\gamma-\gamma', 2\epsilon)$-stable.

One can further define for any P,

$$\epsilon_P^*(\gamma) = \sup\{\epsilon > 0 : \exists \text{ a global minimizer of } \ \mathrm{Loss}_{\mathrm{Km}}(P,\mathcal{C}) \text{ is } (\gamma,\epsilon) - \text{stable}\} \tag{5.12}$$

then ϵ_P^* serve as the population version of problem $(\mathrm{SS}_{\mathrm{Km}})(\gamma)$.

In the end of this section, we provide some examples and discuss our concept with the resampling stability framework that is reviewed in the previous chapter (see Section4.2). Recall that a sufficient and necessary condition of a population being resampling stable is $\mathrm{Loss}_{\mathrm{Km}}(P,\mathcal{C})$ has a unique minimizer. In the end of this section, we illustrate and compare the two stability notions with specific examples.

Stochastic Ball Model Recall the popular generative model used to study K-means recovery guarantee (Section 4.2). Let $\{\mu_i\}_{i=1}^{K_0}$ be the centers and $\tilde{P}$ be a rotationally invariant distribution supported in unit ball in $\mathbb{R}^d$. Suppose the separation between each unit ball in $\mathbb{R}^d$ is $s > 2 + 2\sqrt{2}$. Then one can show that the unique global optimizer of K-means with cluster number $K = K_0$ is simply given by cluster each ball as a single cluster; and $K > K_0$ is a further refinement. Hence due to the rotational invariance property, $\epsilon_P^*(0) > 0$ when $K > K_0$ and further $\lim_{\gamma\to 0}\epsilon_P^*(\gamma)/\gamma = +\infty$. When the separation is not so large, it may

be possible that even $K > K_0$ the global minimizer of Loss_{Km} is unique. However, one can still show that $\lim_{\gamma \to 0} \epsilon_P^*(\gamma)/\gamma = \Omega(K)$.

Uniform distribution on $[0, 1]$ When P is uniform distribution on the line segment $[0, 1]$, a direct Cauchy-Schwartz's inequality shows that the for any $K \geq 2$, the optimal K-means clustering is given by the equal spaced partition. Therefore for any $K \geq 2$, the global minimizer of population loss is unique and K-means is hence resampling stable. This is not ideal, since intuitively we don't expect a uniform distribution on $[0, 1]$ to have any meaningful clustering structure. With our notation, we can further show that $\lim_{\gamma \to 0} \epsilon_P^*(\gamma)/\gamma = +\infty$ for any $K \geq 2$.

As a contrast, consider uniform distribution on the union of two line segments $[-\frac{s+1}{2}, -\frac{s}{2}] \cup [\frac{s}{2}, \frac{s+1}{2}]$. This is a separated uniform distribution. But in this case one can show that $\lim_{\gamma \to 0} \epsilon_P^*(\gamma)/\gamma = O(1/s) < +\infty$.

5.4 Reliability of (γ, ϵ)*-stability*

Finally, given stability on finite sample and on population, it is then natural to ask if stability of a clustering $\mathcal{C}$ on a sample $\mathcal{D}$ can allow us to infer something about the distribution that generated the sample. Recall that for any $\mathcal{D}, \mathcal{C}$ and γ, one can obtain an optimality interval from $(\text{SS}_{\text{Km}})(\gamma)$ whenever the resulting $\epsilon = (K - \kappa(\gamma))w_{\max}$ is no larger than $w_{\min}$.

This section shows that population $(\gamma, \epsilon)-$stability of its global optimizer can be connected to similar property on finite sample with only generic Glivenko-Cantelli type assumptions on P and K-means loss Loss_{Km}. The following assumption is imposed on P.

Assumption 5.1 (Uniform Convergence of Loss_{Km})**.** *There exists a function* Ψ *such that, for any* n *sufficiently large and any* $\delta \in (0, 1]$*, with probability* $1 - \delta$ *over resampling the size* n *sample* $\mathcal{D}$ *from* P*,*

$$\sup_{\mathcal{C} \in \mathbf{C}_K(\mathcal{D})} |\text{Loss}_{\text{Km}}(P; \mathcal{C}) - \text{Loss}_{\text{Km}}(\mathcal{D}, \mathcal{C})| \leq \Psi(n, \delta)\,, \tag{5.13}$$

In the above, the supremum is taken over all sets of distinct Voronoi centers, which allows an identification of a $\mathcal{C} \in \mathbf{C}_K(\mathbb{R}^d)$ from $\mathcal{C} \in \mathbf{C}_K(\mathcal{D})$. Intuitively, equation (5.13) bounds the difference between $\text{Loss}_{\text{Km}}(\mathcal{D}, \mathcal{C})$ and the Loss_{Km} of any clustering of $\mathbb{R}^n$ that

is consistent with $\mathcal{C}$. Assumption 5.1 holds, for instance, when P has compact support [Maurer and Pontil, 2010] or finite higher order moments [Telgarsky and Dasgupta, 2013]. We now view $\Psi(n, \delta)$ as a known function of n, δ.

The following theorem shows how stability guarantees obtained from $(\mathrm{SS}_{\mathrm{Km}})(\gamma)$ in a sample $\mathcal{D}$ can support stability inferences in the distribution P.

Theorem 5.2. *Suppose P satisfies Assumptions 5.1. For any $\delta \in (0, 1]$, if optimal clustering $\mathcal{C}^{opt}$ on P is (γ, ϵ) unstable for some $\gamma > 0$, then with probability $1 - \delta$ over samples $\mathcal{D}$, with $|\mathcal{D}| = n$, any optimal clustering $\widehat{\mathcal{C}}^{opt}$ of $\mathcal{D}$ is $(\gamma + 2\Psi(n, \delta/2), \epsilon/2 - \sqrt{\log(4/\delta)/2n})$ unstable.*

This result opens the way for inferences on the (γ, ϵ) clusterability of P that could be framed as a family of hypothesis tests parametrized by γ. Select $\delta \in (0, 1]$ and γ a tolerance of excess loss. Then consider null hypothesis

$$H_0(\epsilon) : \text{Any optimal K-means clustering on } P \text{ is } (\gamma, \epsilon) \text{ instable.}$$

Let κ be the optimal value of solving $(\mathrm{SS}_{\mathrm{Km}})(\gamma + 2\Psi(n, \delta/2))$ on the sample $\mathcal{D}$. We reject the null hypothesis $H_0(\epsilon)$ with $\epsilon = 2((K - \kappa)w_{\max} + \sqrt{\frac{\log(4/\delta)}{2n}})$ when $(K - \kappa)w_{\max} \leq w_{\min}$. Supposing $H_0(\epsilon)$ is true, by Theorem 5.2 the probability of type I error is at most δ. Thus, one can interpret an OI from $(\mathrm{SS}_{\mathrm{Km}})(\gamma + 2\Psi(n, \delta/2))$ as sufficient for rejecting (γ, ϵ) instability with a p-value at most δ. Moreover, the inference above remain valid, albeit weaker, when instead of the optimal $\widehat{\mathcal{C}}^{\mathrm{opt}}$ only a sub-optimal clustering of the sample is known. While this particular test would be overly conservative, and not necessarily practical, it serves to alert to the possibility of inferring stability in the population from finite sample stability.

Previously people have proposed and studied different paradigm of clustering stability [Ben-David et al., 2006] for model selection. However those notions fail to associate instability on sample with instability on population. The key difference between the assumptions we make and those of the previous papers, is the uniform bound $\Psi(n, \delta)$ for $\mathrm{Loss}_{\mathrm{Km}}$. Our result, on the other hand, shows that we could provide probability guarantee for the (γ, ϵ) stability in our framework for finite samples.

The next result, show in another direction that if the population stability holds for some (γ, ϵ), then with high probability finite sample global optimum enjoys stability.

Theorem 5.3. *There exists a universal constant C such that for any P that satisfies assumption 5.1 and that has a unique global optimal solution $\mathcal{C}^*$ that is (γ, ϵ) stable, for any $\delta > 0$, when n is sufficiently large such that $2\Psi(n, \delta/3) < \gamma$, with probability at least $1 - \delta$, the global optimal solution on finite sample from P is (γ', ϵ') stable, with*

$$\gamma' = \gamma - 2\Psi(n, \frac{\delta}{3}),, \quad \epsilon' = 2\epsilon + 2\sqrt{\frac{\log(6/\delta)}{2n}} + C\sqrt{\frac{8Kd(4Kd + 2K + 1)}{n}} .$$

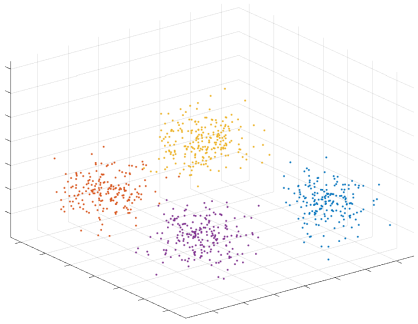
5.5 Experiments

We implemented the SS_{Km} problem using the SDP solver SDPNAL+Zhao et al. [2010], Yang et al. [2015]. We also implemented the spectral bound of Meilă [2006a], the only other method offering optimality intervals for K-means. The main questions of interest were (1) do our OI exist for realistic situations? (2) how tight are the bounds obtained? The implementation is completed by co-author Marina Meila. The author is very thankful for her contribution.

Synthetic Data of Stochastic Ball Model For this setting we sampled data uniformly from K balls in dimension d with unit radius. Let c be the minimal distance between the centers of the K balls, then the SDP relaxation of K-Means we adopted in this paper can guarantee the exact recovery of K clusters with high probability when $c > 2 + \epsilon(d)$, where $\epsilon(d) \to 0$ as $d \to \infty$. This means that under this specific stochastic ball model, the k-means guarantee can only be obtained when the seperation between balls are large enough so that they don't touch each other.

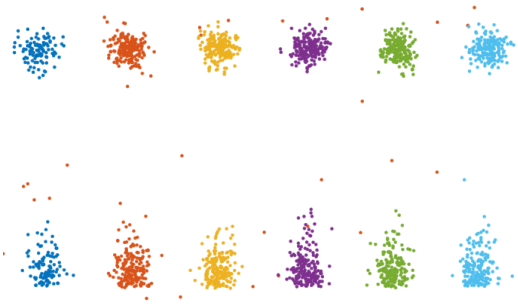
We sampled $n = 500$ data from the stochastic ball model with $K = 4, d = 2$ and c ranging from 1.4 - 3.2. The centers of each ball are aligned on one line segment with equal space between. Then we perform K-means clustering with the initialization of correct labels, since we are only interested in understanding the behaviour of our method's ability to obtain a guarantee for clustering result. Under this setting, the theoretical bound is trivial (larger than 4). Theoretically we can say nothing about how good the SDP relaxation is for

$K = 4$, separation ≈ 5.66

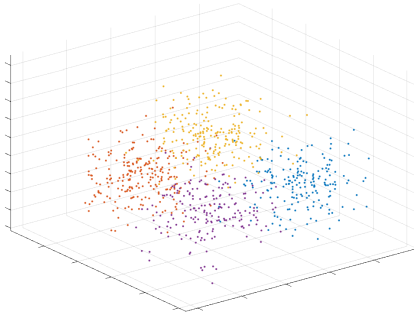


$\sigma = .6$

$K = 6$, separation 1



$\sigma = .1$



$\sigma = 1.1$

$K = 6$ clusters

	normal	non-normal
σ	$n = 525$	$n = 525$
0.06	0.00(0.00)	0.005(0.001)
0.08	0.01(0.00)	0.006(0.001)
0.1	0.01(0.00)	0.009(0.003)

Figure 5.2: Some data used in the experiments. In the first three plots, the clusters are sampled from mixtures of spherical Gaussians. In the last, one of the 15 coordinates is from a Gamma$(2, .4)$ distribution and rescaled by σ. Separation is the distance between the Gaussian means, and σ is the standard deviation of the Gaussians. The $K = 6$ data sets are designed to be hard for the spectral bounds but not for the SDP bounds. Bottom, left: optimality intervals ϵ for data sampled from normal and non-normal mixtures with $K = 6$ (mean and standard deviation over 10 replications). The values of ϵ_{Sp} were much larger than $w_{\min}$ and are shown.

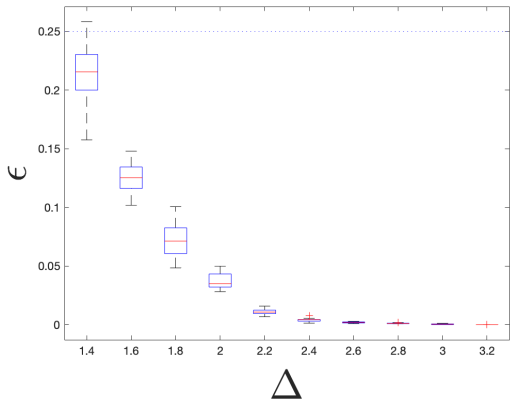


Figure 5.3: ϵ bound obtained for data sampled from stochastic ball model. The procedure is repeated for 10 times under each setting. The ϵ bound are meaningful when they are smaller than $p_{\min} \approx 0.25$.

the clustering. However as shown in figure 5.3 our method can provide some guarantees on a particular clustering result. In all settings, initializing with the true label, the distance between true label and the K-means solutions are close, with earth mover distance approximately $\sim 10^{-6}$. We see that in the model touching cases, our SS method provides insights on the distance between the K-means global optimal solution to the true label. On the other hand, all $c \geq 2.2$ are within the regime where previous theoretical guarantee does not work. The results shows that K-means global optimizers approximately achieve the exact recovery of stochastic ball models.

Synthetic Data of Gaussian Mixture We sampled data from a mixture of $K = 4$ normal distributions with equal spherical covariances $\sigma^2\mathbf{I}$, in $d = 15$ dimensions. The cluster sizes n_k were approximately equal to $\lfloor n/K \rfloor$. The cluster means were at the corners of a regular tetrahedron with center separation $||\mu_k - \mu_{k'}|| = 4\sqrt{2} \approx 5.67$. The data was clustered by K-means with random initialization, then the bounds ϵ, ϵ_{Sp} corresponding respectively to the SS method and to the spectral method of Meilă [2006a] were computed. In the experiments we also performed *outlier removal*, as follows. For each x_i, we computed the sum of the distances to its $w_{\min}/2$ nearest neigbors. We then removed the n_0 data

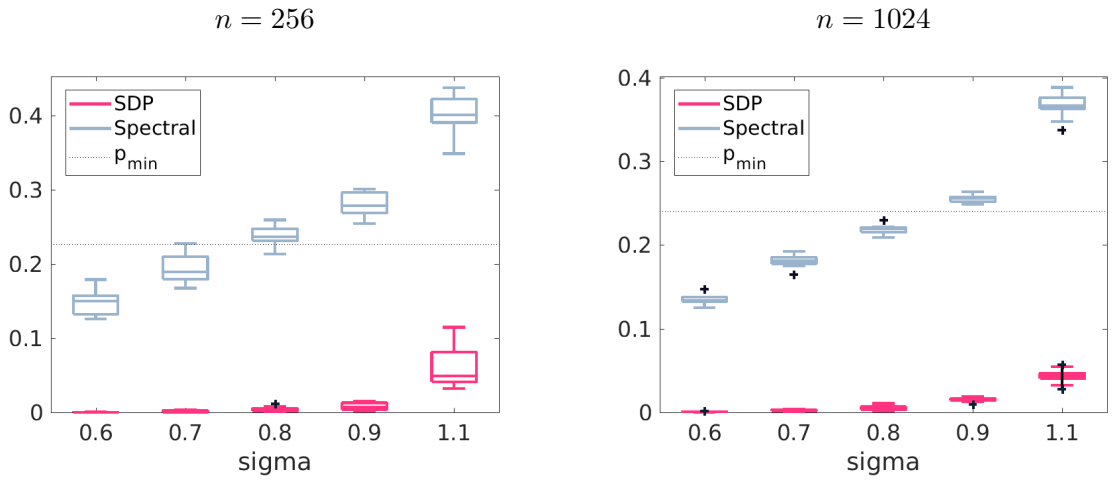


Figure 5.4: The optimality intervals ϵ and ϵ_{Sp} for data sampled from mixtures of normal distributions with $K = 4$, $n = 256$ and 1024 and various σ values (over 10 replications). The values of ϵ_{Sp} exceeding $w_{\min}$ are not valid. Note that the ϵ bounds are near 0 even though the clusters are not separated for $\sigma > 0.8$.

points with the largest values for this sum. For good measure, we first added 20 outliers, then removed $n_0 = 4\%n$ respectively $n_0 = 2\%n$ points (so that n_0 is slightly larger than 20), before computing the bounds ϵ, ϵ_{Sp}. Consequently, these bounds do not refer to clusterings of the original $\mathcal{D}$, but to the "cleaned" dataset. Note that the outlier removal does not depend on the cluster labels; it is performed before clustering the data. Figure 5.4 displays the bounds ϵ, ϵ_{Sp} for these data, while Figure 5.2, top and bottom, left, displays some representative samples. The ϵ optimality interval is much tighter than the spectral one ϵ_{Sp}, and, surprisingly enough, holds even when the clusters "touch", i.e when there is no region of low density between the clusters. Figure 5.5 (left) shows that, when $\sigma > 0.8$, the minimal spheres containing the clusters intersect; on the right we see that there are points which are almost equidistant from two cluster centers. Otherwise put, the *distribution free* bounds hold even when the data are not contained in non-intersecting balls, which is the best known condition for clusterability *under model assumptions* Awasthi et al. [2015a, 2014].

Next, we performed experiments with unequal cluster sizes $p_{1:4} = 0.1, 0.2, 0.3, 0.4$, and

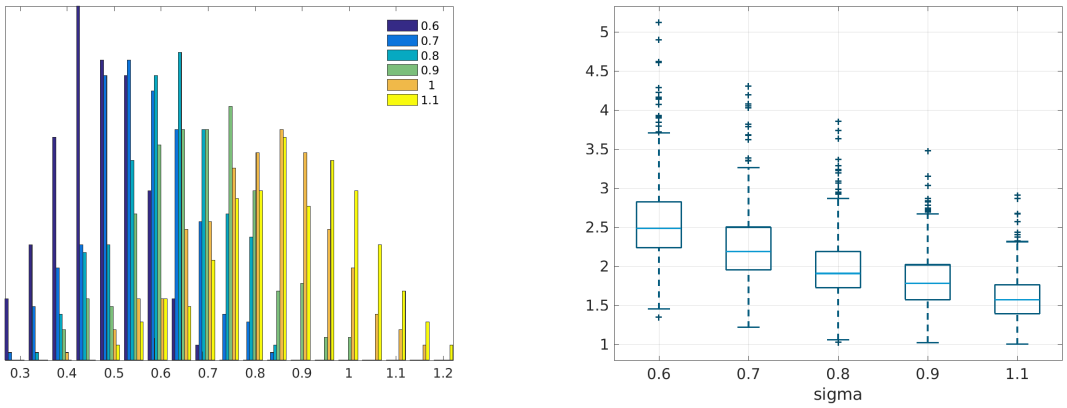


Figure 5.5: Separation statistics for the $K = 4$ data, $n = 1024$, all σ values. Left: histogram of $\min_k ||x_i - \mu_k|| / \min_{k,k'} ||\mu_k - \mu_{k'}||$ (i.e. distance of point to its center over minimum center separation) colored by σ. Note that when the clusters are contained in equal non-intersecting balls this ratio is strictly smaller than 0.5. Right: boxplot of distance to second closest center over distance to own center, versus σ.

we also generated non-gaussian clusters (details in the Appendix). We also performed experiments with $K = 6$ clusters, with $p_{1:6} = 0.1, 0.18, 0.18, 0.18, 0.18, 0.18$. For $K = 6$ we placed cluster centers along a line, as shown in Figure 5.2, top and bottom, right. This hurts the spectral bound which depends on a stable $K-1$-subspace, but does not hurt, and may even help the SDP bound ϵ.

The results are shown in Table 5.1, with run times for all experiments in the Appendix. The spectral bound ϵ_{Sp} was much larger than ϵ for all $K14$ experiments, and was never valid for $K = 6$; therefore, it is omitted. The table also shows that ϵ takes similar values in the case of equal and unequal clusters. However, in the latter case, the condition $\epsilon \leq w_{\min}/w_{\max}$, is more stringent, hence some of the bounds obtained are not valid.

Over all experiments, we have found that the values of ϵ are virtually insensitive to the sample size n, and degrade slowly when $w_{\min}$ decreases. The main limitation in these experiments is the requirement that $\epsilon \leq w_{\min}$. This requirement can be traced back to Theorem 4.1. The bound in this Theorem, even though state of the art, is not tight,

Table 5.1: The OI ϵ for $K = 4$ clusters of unequal sizes (mean and standard deviation over 10 replications). The values in gray are not valid, owing to the fact that $\epsilon w_{\max} > w_{\min}$ in these cases. Bounds for smaller σ values were essentially zero and are ommitted.

	Unequal normal clusters			Unequal non-normal clusters		
σ	$n = 200$	$n = 400$	$n = 800$	$n = 200$	$n = 400$	$n = 800$
0.6	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.001(0.001)	0.001(0.000)	0.002(0.007)
0.8	0.01(0.01)	0.01(0.01)	0.01(0.01)	0.006(0.004)	0.004(0.002)	0.007(0.003)
1.0	0.09 (0.05)	0.06 (0.01)	0.07 (0.02)	0.04 (0.02)	0.03 (0.01)	0.03 (0.01)
1.2	0.28 (0.08)	0.21 (0.05)	0.21 (0.03)	0.16 (0.06)	0.14 (0.03)	0.13 (0.03)

suggesting that some the ϵ values marked in gray are valid OI even though we cannot prove it at this time.

Real data: configurations of the aspirin ($C_9H_8O_4$) molecule These $n = 2118$ samples (see Figure 5.6) were obtained via Molecular Dynamics (MD) simulation at $T = 500$ degrees Kelvin by Chmiela et al. [2017b] and represent 3D positions of the 21 atoms of aspirin. It was discovered recently that aspirin's potential energy surface has two energy wells. The purpose of clustering of MD simulations is to label the data by energy well; this allows chemists to identify energy wells, on one hand, and on the other to identify states around on the transition path between the energy wells. These states describe the mechanism of a transition, and, being rare events in the contex of many simulations, are of great interest. Currently, the labeling is done by ad hoc algorithms. Having guarantees of (almost) correctness for the grouping saves the time needed to validate the clustering by human inspection. Hence, we cluster these data into $K = 2$ clusters, after having removed $n_0 = 0.5\%n = 106$ outliers. The clusters found have relative sizes $w_{\min} = .26, w_{\max} = .74$, and the OI is $\epsilon = .065$, an informative bound.

Figure 5.6: Left: the aspirin molecule. Middle: The first two principal components of the 57-dimensional aspirin data. The data is a Molecular Dynamics sequence of 211,762 configurations. We sample every 100-th point of the data for clustering. The axes are represented *at scale*.

5.6 Discussion

Distribution free cluster validation in context A researcher who wants to discover cluster structure in data must perform several inference tasks. This paper has focused on post-clustering validation, which happens to be the least studied of these inferences. When the data is clusterable, we have shown that validating the clusters is possible, without providing sharp thresholds. In Section **??** we have also cited works that prove that finding the clusters is tractable, under the assumption of clusterability. Hence, the loop is about to close, and we hope that in future work, to integrate the SS method with a clustering algorithm, providing thus a complete "clustering with guarantees" methodology. Furthermore, our distributional results show that for sufficiently large n, the SS method can be the basis of a test for clusterability, under generic Glivenko-Cantelli assumptions.

Proofs of instability What happens when the Stability Theorem does not hold for a clustering $\mathcal{C}$? In this case, the researcher can try to certify that $\mathcal{C}$ is *unstable*. This task is comparatively easier, since a single counterexample $\mathcal{C}'$ with $\text{Loss}(\mathcal{C}') \approx \text{Loss}(\mathcal{C})$ and $d^{EM}(\mathcal{C}', \mathcal{C}) > w_{\min}$ suffices. This, again, is a well studied area. As an alternative worth exploration, one could use the output of the SS method to find a *witness* of instability. More precisely, when the SS method fails to produce a valid OI, the matrix X' with $\text{Loss}(\mathcal{D}, X') \leq \text{Loss}(\mathcal{D}, \mathcal{C})$ is far from $X(\mathcal{C})$ in Frobenius norm. One could try to find a clustering $\mathcal{C}'$ by e.g. rounding X', which would not differ much from X' in either Loss or

distance to $\mathcal{C}$.

Stability and the choice of K Throughout the paper, we have assumed that K is fixed. We now remark that the SS method implicitly solves the problem of selecting K, and even that of selecting Loss. Indeed, a clustering $\mathcal{C}$ that is found stable, with *any* K and *any* loss function, is a "correct" clustering of the data under our paradigm. In the beginning, we presented the informed user as selecting the Loss; with the concept of stability, one can take another view, of a user lucky enough to find a stable clustering while searching over loss functions and K values. The SS method does not preclude the existence of more than one stable clustering for a data set $\mathcal{D}$. For example, if clusters are hiearchically nested, it is possible to find stable clusterings at several levels of the hierarchy.

In practice, K is not known, and it is *chosen after* a set of clusterings $\mathcal{C}^{(K)}$, with $K = 1, 2, \ldots K_{max}$ have been obtained. With the SS method, one could dispense with the (more or less ad-hoc) methods for selecting K in loss-based clustering. Indeed, by our initial argument, if $\mathcal{C}^{(K)}$ is proved to be stable for some K, this implicitly validates K itself, as well as the loss function used. It is also possible to select more than one K, when the data supports meaningful partitions with different numbers of clusters.

Silhouette and other cluster quality indices We contrast the framework proposed here with the existing literature on *internal cluster validation*; see e.g Maria Halkidi [2015], Arbelaitz et al. [2013], Hennig and Liao [2013] by indices such as the *silhouette* Rousseeuw [1987]. As it is well known, these indices *are not associated with a clustering paradigm*, whereas the present paper argues for paradigm specific validation, as a way to assure that the same criterion is used to find the clusters and to validate them. These indices could be potentially used as goodness measures, if their relations to specific clustering loss functions became better understood; the works cited above take steps in this direction.

Comparison with VC bounds It is extremely rare in statistical inference to have worst case error bounds that are relevant in practice. For instance, the well known VC bounds for the 0-1 classification loss (see e.g. Vapnik [1998]) typically take values above 1 (hence are

completely uninformative from a practical standpoiht) and depend on the VC-dimension, a property of the hypotheses class that is usually intractable to compute.

In contrast, with the SS method, the OI ϵ is always informative when it exists. With SDP relaxations, we obtain bounds that are not only informative, they are near 0 in non-trivial situations. To appreciate how far these guarantees can extend, recall than when $\sigma \approx \frac{1}{6}$ of the center separation, two spherical normal densities start to touch – no region of low density is left between them. Several of the informative, valid bounds in Section 5.5 are obtained near or even above these critical values. Moreover, an optimality interval is a distribution free, worst case bound. Thus, we believe that the computational demands of the SDP solver are justified by the guarantees offered.

5.7 Proofs of results in Chapter 5

We first state several helpful propositions needed for our proofs. To simplify notation, we would suppress $\mathrm{Loss}_{\mathrm{Km}}(P, \mathcal{C})$ to $L(C)$ for a fixed population.

Proposition 5.1. *For any* $\mathbf{X} \in \mathcal{X}$, $||\mathbf{X}||_F^2 \leq K$.

Proposition 5.2. *For any fixed two clusterings* $\mathcal{C}, \mathcal{C}'$ *it holds that*

$$|d_P^{EM}(\mathcal{C}, \mathcal{C}') - d_{\widehat{P}}^{EM}(\mathcal{C}, \mathcal{C}')| \leq \sqrt{\frac{\log(4/\delta)}{2n}}\ . \tag{5.14}$$

with probability $1 - \delta/2$

Proof of Proposition 4.1. $X_{ij} \in [0, 1]$ is obvious from the definition in section 4.2.

$$\operatorname{trace} \mathbf{X} \ = \sum_{k=1}^{K} \sum_{i \in C_k} \mathbf{X}_{ii} \ = \sum_{k=1}^{K} \sum_{i \in C_k} \frac{1}{n_k} \ = \sum_{k=1}^{K} n_k \frac{1}{n_k} \ = \ K. \tag{5.15}$$

Denote by k_0 the cluster containing data point i.

$$(X\mathbf{1})_i \ = \sum_{j=1}^{n} \mathbf{X}_{ij} \ = \sum_{k=1}^{K} \sum_{i \in C_k} \mathbf{X}_{ij} \sum_{j \in C_{k_0}} \frac{1}{n_k} \ = \ 1. \tag{5.16}$$

Moreover, $\mathbf{X} = \mathbf{Z}\mathbf{Z}^\top$, hence $\mathbf{X} \succeq 0$. □

Proof of Theorem 5.1. Note that for any clustering $\mathcal{C}$, $||\mathbf{X}(\mathcal{C})||_F^2 = \sum_{i,j=1}^n \mathbf{X}_{ij}^2 =, = \sum_{k=1}^K n_k^2 \left(\frac{1}{n_k}\right)^2 = K$. Moreover, from proposition 5.1 we have $||\mathbf{X}||_F^2 \leq K$.

Note also that $||\mathbf{X} - \mathbf{X}'||_F^2 = ||\mathbf{X}||_F^2 + ||\mathbf{X}'||_F^2 - 2\langle \mathbf{X}, \mathbf{X}' \rangle = 2K - 2\langle \mathbf{X}, \mathbf{X}' \rangle$. Hence, the optimization problem ($\mathrm{SS}_{\mathrm{Km}}$) finds the feasible $\mathbf{X}'$ which is furthest away from $\mathbf{X}$. This completes Step 1. For Step 2 we can apply Theorem 9 of Meilă [2012], which bounds the earthmover distance d^{EM}. □

Proof of Proposition 5.1. Denote by $\lambda_1, \dots \lambda_n$ the eigenvalues of $\mathbf{X}$. Since $\mathbf{X}$ has non-negative elements, and $\mathbf{X}\mathbf{1} = \mathbf{1}$, by the Frobenius Theorem, $|\lambda_i| \leq 1$, and because $\mathbf{X} \succeq 0$, $\lambda_i \geq 0$ for all $i \in [n]$. Hence, $\lambda_i \in [0, 1]$, for all i, and $||\mathbf{X}||_F^2 = \operatorname{trace} \mathbf{X}^2 = \sum_{i=1}^n \lambda_i^2 \leq \sum_{i=1} \lambda_i = \operatorname{trace} \mathbf{X} = K$. □ □

Proof of Theorem 5.2. On the sample with probability $1 - \delta/2$ it holds that

$$\sup_{\mathcal{C} \in \mathbf{C}_K(\mathcal{D})} |L(\mathcal{C}) - \widehat{L}(\mathcal{C})| \leq \Psi(n, \frac{\delta}{2}). \tag{5.17}$$

Now we condition on the event that (5.17) holds.

If there exists a population minimizer $\mathcal{C}^{opt}$ such that $d_P^{EM}(\mathcal{C}^{opt}, \widehat{\mathcal{C}}^{opt}) \leq \epsilon_0/2$. Note that by assumption of instability, there exists another clustering $\mathcal{C}^*$ such that $d_P^{EM}(\mathcal{C}^*, \mathcal{C}^{opt}) > \epsilon_0$ and $L(\mathcal{C}^*) < L(\mathcal{C}^{opt}) + \eta$, then $d_P^{EM}(\mathcal{C}^*, \widehat{\mathcal{C}}^{opt}) > \epsilon_0/2$. We can bound

$$\widehat{L}(\mathcal{C}^*) \leq L(\mathcal{C}^*) + \Psi(n, \frac{\delta}{2}) \leq L(\mathcal{C}^{opt}) + \Psi(n, \frac{\delta}{2}) + \eta \leq L(\widehat{C}^{opt}) + \Psi(n, \frac{\delta}{2}) + \eta \leq \widehat{L}(\widehat{\mathcal{C}}) + 2\Psi(n, \frac{\delta}{2}) + \eta \tag{5.18}$$

and take $\widehat{\mathcal{C}}' = \mathcal{C}^*$.

If such a clustering $\mathcal{C}^{opt}$ does not exist, then note that for any optimal solution of the population clustering $\mathcal{C}^{opt}$,

$$\widehat{L}(\mathcal{C}^{opt}) \leq L(\mathcal{C}^{opt}) + \Psi(n, \frac{\delta}{2}) \leq L(\widehat{\mathcal{C}}) + \Psi(n, \frac{\delta}{2}) \leq \widehat{L}(\widehat{\mathcal{C}}) + 2\Psi(n, \frac{\delta}{2}) \tag{5.19}$$

and take $\widehat{\mathcal{C}}' = \mathcal{C}^{opt}$.

In both cases, applying proposition 5.2 it holds that $d_P^{EM}(\widehat{\mathcal{C}}', \widehat{\mathcal{C}}^{opt}) \geq \epsilon_0/2 - \sqrt{\log(4/\delta)/2n}$. Therefore $\widehat{\mathcal{C}}^{opt}$ is $(\gamma + 2\Psi(n, \frac{\delta}{2}), \epsilon_0/2 - \sqrt{\log(4/\delta)/2n})$ instable. □

Proof of Proposition 5.2. For fixed permutation π, from Hoeffding's inequality

$$\left| \mathbb{E}_P \sum_{i=1}^{K} \mathbf{1}_{X \in C_i \cap C_{\pi(i)}} - \mathbb{E}_{\widehat{P}} \sum_{i=1}^{K} \mathbf{1}_{X \in C_i \cap C_{\pi(i)}} \right| \leq \sqrt{\frac{\log(2/\delta)}{2n}} \tag{5.20}$$

with probability $1-\delta$. Now let $\pi^*, \widehat{\pi}^*$ be the permutation maximizing $\mathbb{E}_P \sum_{i=1}^{K} \mathbf{1}_{X \in C_i \cap C_{\pi(i)}}$ and $\mathbb{E}_{\widehat{P}} \sum_{i=1}^{K} \mathbf{1}_{X \in C_i \cap C_{\pi(i)}}$ respectively. Then

$$\mathbb{E}_P \sum_{i=1}^{K} \mathbf{1}_{X \in C_i \cap C_{\pi^*(i)}} \leq \mathbb{E}_{\widehat{P}} \sum_{i=1}^{K} \mathbf{1}_{X \in C_i \cap C_{\pi^*(i)}} + \sqrt{\frac{\log(2/\delta)}{2n}} \leq \mathbb{E}_{\widehat{P}} \sum_{i=1}^{K} \mathbf{1}_{X \in C_i \cap C_{\widehat{\pi}^*(i)}} + \sqrt{\frac{\log(2/\delta)}{2n}} \tag{5.21}$$

$$\mathbb{E}_{\widehat{P}} \sum_{i=1}^{K} \mathbf{1}_{X \in C_i \cap C_{\widehat{\pi}^*(i)}} \leq \mathbb{E}_P \sum_{i=1}^{K} \mathbf{1}_{X \in C_i \cap C_{\widehat{\pi}^*(i)}} + \sqrt{\frac{\log(2/\delta)}{2n}} \leq \mathbb{E}_P \sum_{i=1}^{K} \mathbf{1}_{X \in C_i \cap C_{\pi^*(i)}} + \sqrt{\frac{\log(2/\delta)}{2n}} \tag{5.22}$$

Therefore one concludes that $|d_P^{EM}(\mathcal{C}, \mathcal{C}') - d_{\widehat{P}}^{EM}(\mathcal{C}, \mathcal{C}')| \leq \sqrt{\log(2/\delta)/2n}$ with probability $1-\delta$. □

Proof of theorem 5.3. Since $\widehat{C}$ is the optimal solution on the sample, we have with probability $1-\delta/3$

$$L(\widehat{C}) - \Psi(n, \frac{\delta}{3}) \leq \widehat{L}(\widehat{C}) \leq \widehat{L}(C^*) \leq L(C^*) + \Psi(n, \frac{\delta}{3}) \, . \tag{5.23}$$

Rearranging this inequality, we have $L(\widehat{C}) \leq L(C^*) + 2\Psi(n, \frac{\delta}{3})$. Since $\gamma_P > 2\Psi(n, \frac{\delta}{3})$, by the stability assumption of C^* we conclude that $d_P(C^*, \widehat{C}) \leq \epsilon$. Hence combining lemma 2.1, we have

$$d_{\widehat{P}}(C^*, \widehat{C}) \leq \epsilon + \sqrt{\log(6/\delta)/2n} \tag{5.24}$$

with probability at least $1-\delta/3$.

Now let $\widehat{C'}$ be any different clustering such that $\widehat{L}(\widehat{C'}) \leq \widehat{L}(\widehat{C}) + \gamma - 2\Psi(n, \frac{\delta}{3})$. Then with probability at least $1-\delta/3$

$$L(\widehat{C'}) \leq \widehat{L}(\widehat{C'}) + \Psi(n, \frac{\delta}{3}) \tag{5.25}$$

$$\leq \widehat{L}(\widehat{C}) + \gamma - \Psi(n, \frac{\delta}{3}) \tag{5.26}$$

$$\leq L(C^*) + \gamma \tag{5.27}$$

again by stability it holds that $d_P(\widehat{C}', C^*) \leq \epsilon$. We will now derive an upper bound of $d_{\widehat{P}}(\widehat{C}', \widehat{C})$ that holds with probability at least $1 - \delta/3$.

To show this uniform bound, we start by introducing an operator **IndNear**. Given K centers, **IndNear** returns the index of the closest center of $\mu_1, \cdots, \mu_K$ to each point x. And if there are ties, return the smallest index. Then we further consider the family of classifiers

$$\mathcal{F} = \left\{ f(x; \boldsymbol{\theta}) = \mathbf{1}_{\boldsymbol{IndNear}(x; a_1, \cdots, a_K) \neq \boldsymbol{IndNear}(x; b_1, \cdots, b_K)} : \boldsymbol{\theta} = \mathrm{vec}(a_1, \cdots, a_K, b_1, \cdots, b_K) \in \mathbb{R}^{2Kd} \right\} \tag{5.28}$$

Given parameters $\boldsymbol{\theta} = \mathrm{vec}(a_1, \cdots, a_K, b_1, \cdots, b_K)$, one can determine a function $f(x; \boldsymbol{\theta})$ mapping from $\mathbb{R}^d$ to $\{0, 1\}$. Hence this is a well-defined classifier. This is a parametric form of predictable functions with dimension of parameters being $p = 2Kd$. Furthermore, to predict the label of a point x, one needs following operations:

- Compute distance to a center: $\mathbb{R}^d$ differences, $(d-1)$ sums and 1 square on real numbers.
- Find the index of nearest center: $K - 1$ comparisons.
- Compare two indexes from two different sets of centers: 1 comparison.

In total to predict the label of a point x, one needs $t = 2K(d + d - 1 + 1) + 2(K - 1) + 1 = 4Kd + 2K - 1$ operations. Therefore by lemma 5.1, we conclude that the VC dimension of $\mathcal{F}$, denoted by V, can be upper bounded by

$$V \leq 4p(t + 2) = 8Kd(4Kd + 2K + 1) \tag{5.29}$$

With this class of classifiers in mind, we now focus on the set of switching labels between $\widehat{C}, \widehat{C}'$. namely the set $S_{\widehat{C}'} = d - \cup_{i=1}^{K}(\widehat{C}_i \cap \widehat{C}'_i)$. Then the centers inducing $\widehat{C}$ and $\widehat{C}'$ together as a parameter $\boldsymbol{\theta}$ leads to a classifier in $\mathcal{F}$ such that recovers this set of switching labels $S_{\widehat{C}'}$ exactly.

Now consider all possible $\widehat{C}'$. By lemma 5.2, we conclude that with probability at least $1 - \delta/3$,

$$\sup_{\widehat{C}'} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \in S_{\widehat{C}'}} - \mathbb{E}\mathbf{1}_{X_i \in S_{\widehat{C}'}} \right| \leq C \sqrt{\frac{8Kd(4Kd + 2K + 1)}{n}} + \sqrt{\frac{\log(6/\delta)}{2n}} \tag{5.30}$$

where C is a universal constant (as in lemma).

Note that we do not regard different labeling of $\widehat{C}'$ as the same clustering here and in the definition of $\mathcal{F}$, we conclude that

$$\begin{aligned} d_{\widehat{P}}(\widehat{C}, \widehat{C}') &\leq d_P(\widehat{C}, \widehat{C}') + C\sqrt{\frac{8Kd(4Kd+2K+1)}{n}} + \sqrt{\frac{\log(6/\delta)}{2n}} \\ &\leq \epsilon + C\sqrt{\frac{8Kd(4Kd+2K+1)}{n}} + \sqrt{\frac{\log(6/\delta)}{2n}} \,. \end{aligned} \tag{5.31}$$

holds for any $\widehat{C}'$ with probability $1 - \delta/3$ over resampling of data. Therefore, the desired result is obtained by put the upper bound in (5.24),(5.27) and (5.31) together. □

Here we states the two lemmas we used in the proof.

Lemma 5.1 (Theorem 8.4 from [Anthony and Bartlett, 1999])**.** *Suppose f is a function from $\mathbb{R}^d \times \mathbb{R}^p$ to $\{0,1\}$ and let $\mathcal{F} = \{x \to f(x,a) : a \in \mathbb{R}^p\}$ be the class determined by f. Suppose that f can be computed by an algorithm that takes as input the pair $(x,a) \in \mathbb{R}^d \times \mathbb{R}^p$ and returns $f(x,a)$ after no more than t operations of the following types:*

- *the arithmetic operations $+, -, \times, /$ on real numbers,*
- *jumps conditioned on $>, \geq, <, \leq, =, \neq$ comparisons of real numbers*
- *output 0 or 1.*

Then the VC dimension of $\mathcal{F}$ is upper bounded by 4p(t+2).

Lemma 5.2 (Generalization Bounds via VC dimension)**.** *Let $\mathcal{S}$ be a family of sets in $\mathbb{R}^d$. Let $x_1, \cdots, x_n$ are i.i.d. sample from distribution P on $\mathbb{R}^d$. Suppose that the VC dimension of $\mathcal{S}$ is bounded by $V < \infty$, then with probability $1 - \delta$ we have*

$$\sup_{S \in \mathcal{S}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \in S} - \mathbb{E}\mathbf{1}_{X_i \in S} \right| \leq C\sqrt{\frac{V}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}} \tag{5.32}$$

where C is a universal constant given by $24\sqrt{3\log 2 + 2}\int_0^1 \sqrt{1 + \log(\frac{1}{r})}dr$

Proof. We start by thereom 1 in [Haussler, 1995] that for every $x_1, \cdots, x_n \in \mathbb{R}^d$ and $0 \leq r \leq 1$ we have

$$N(r, \mathcal{S}(x_1^n)) \leq e(V+1)\left(\frac{2e}{r^2}\right)^V. \tag{5.33}$$

where

$$\mathcal{S}(x_1^n) = \{b = (b_1, \cdots, b_n) \in \{0,1\}^n : \exists S \in \mathcal{S} : b_i = \mathbf{1}_{x_i \in S}, i \in [n]\} \tag{5.34}$$

is the set of bit vectors b such that it is describing whether each element of x_i is in S or not; And $N(r, \mathcal{S}(x_1^n))$ is the covering number on this space induced by the squared root of normalized Hamming distance. Theorem 3.2 in [Devroye and Lugosi, 2001] shows that

$$\mathbb{E}[\sup_{S \in \mathcal{S}} |\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i \in S} - \mathbb{E}\mathbf{1}_{X_i \in S}|] \leq \frac{24}{\sqrt{n}} \max_{x_1, \cdots, x_n \in \mathbb{R}^d} \int_0^1 \sqrt{\log 2N(r, \mathcal{S}(x_1^n))} dr \tag{5.35}$$

Hence we can further bound

$$\mathbb{E}[\sup_{S \in \mathcal{S}} |\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i \in S} - \mathbb{E}\mathbf{1}_{X_i \in S}|] \leq \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\log(V+1) + (V+1)(\log 2 + 1) + 2V\log(\frac{1}{r})} dr \tag{5.36}$$

$$\leq \left[24\sqrt{3\log 2 + 2} \int_0^1 \sqrt{1 + \log(\frac{1}{r})} dr\right] \sqrt{\frac{V}{n}} \tag{5.37}$$

Hence $\mathbb{E}[\sup_{S \in \mathcal{S}} |\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i \in S} - \mathbb{E}\mathbf{1}_{X_i \in S}|] \leq C\sqrt{V/n}$ where $C = 24\sqrt{3\log 2 + 2}\int_0^1 \sqrt{1 + \log(\frac{1}{r})} dr$ is a universal constant.

Finally recall the simple consequence of the McDiarmid's inequality, we have

$$P\left(\left|\sup_{S \in \mathcal{S}} |\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i \in S} - \mathbb{E}\mathbf{1}_{X_i \in S}| - \mathbb{E}\left[\sup_{S \in \mathcal{S}} |\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i \in S} - \mathbb{E}\mathbf{1}_{X_i \in S}|\right]\right| > t\right) \leq 2e^{-2nt^2} \tag{5.38}$$

The desired result follows directly. □

Chapter 6

STABILITY OF GAUSSIAN MIXTURE MODELS

6.1 Introduction

In this chapter, we turn our focus to the problem of fitting spherical Gaussian Mixture Models (sGMM) to an unknown distribution Q. Without assuming knowledge about the target Q, what kind of guarantees can we give about an estimated sGMM? And under what condition are guarantees possible?

Previous work (e.g.,[Dasgupta, 1999, Sanjeev and Kannan, 2001, Achlioptas and McSherry, 2005, Kannan et al., 2008], etc) established estimation guarantees for the mixture parameters under model assumptions about the data source Q (being, e.g., a mixture of well separated Gaussians). Specifically, W.r.t. the scope of this chapter, when Q is a mixture of K spherical Gaussians, satisfying two criteria: (i) non-vanishing component proportions, and (ii) sufficient separation between components, polynomial run-time estimation algorithms exist.

This chapter handles the question: what can be said without such precise knowledge of Q? We assume instead *indirect* knowledge of Q, namely that a well-separated mixture model P was fit to Q, and that *the model fit is good*. The main difference is that now Q is not required to belong to the model class, but to be "close" to it, in a way to be defined (specifically, in this paper, the model fit is measured by the *total variation distance* between P and Q, $TV(P,Q)$). We also assume that P is a mixture of well-separated Gaussians, with component proportions bounded below. As it turns out, knowing that Q is close to a "good" model class, is almost as useful as knowing Q is in the model class. Under these conditions on P and Q, we prove that P's parameters are unique up to perturbations that we upper bound. In summary, we aim to prove a statment the following form.

Theorem 6.1 (Generic $(\gamma, \mathsf{d})-$Stability)**.** *For distribution Q and a class $\mathfrak{M}$ of sGMM, a given model $P \in \mathfrak{M}$ is (γ, d)-stable if for any $P' \in \mathfrak{M}$ such that $TV(Q,P') \leq TV(Q,P)+\gamma$,*

it holds that $d_{param}(P, P') \leq \mathsf{d}$*, where* $d_{param}(\cdot, \cdot)$ *is a divergence defined in parameter space*

.

This type of statement was proposed earlier by [Meilă, 2006b] in the context of loss-based clustering, but to date it has not been instantiated for any parametric model fitting problem.

A distribution P that satisfies Theorem 6.1 will be called stable. Obviously, such a distribution is generally not unique or distinguished in $\mathfrak{M}$, hence stability should be defined as a property of (a subset of) $\mathfrak{M}$. This also follows from setting $Q = P^* \in \mathfrak{M}$. Therefore, we define parametric stability of a model class as follows.

Definition 6.1. *Let* $\mathfrak{M}$ *be a class of spherical Gaussian Mixtures and* $d_{param}(\cdot, \cdot)$ *be a divergence between parameters; for sufficiently small* $\epsilon > 0$*,* P *is* $(\epsilon, \mathsf{d})-$*stable in* $\mathcal{M}$ *if any model* $P' \in \mathcal{M}$ *such that* $d_{TV}(P, P') \leq 2\epsilon$ *satisfies that* $d_{param}(P, P') \leq \mathsf{d}$*.*

The technical contribution of our paper is to formulate conditions on $\mathfrak{M}$ and establish upper bounds $\mathsf{d}(\epsilon, \mathfrak{M})$ for any $P \in \mathfrak{M}$, in the population regime. These are given in Theorem 6.2. The upper bounds on $\mathsf{d}(\epsilon, \mathfrak{M})$ we obtain are *tractable*, with explicit constants depending on the model class only. As a statistical procedure, the bound $\mathsf{d}(\epsilon, \mathfrak{M})$ provides quantitative post-estimation evaluation or diagnostic analysis for fitting a Gaussian Mixture Model, without any prior knowledge. For example, for an arbitrary population Q (with density q) on $\mathbb{R}^d$, let $\hat{P} = \sum_{i=1}^{K} \hat{w}_i \mathcal{N}_d(\hat{\mu}_k, \hat{\sigma}_k^2 \mathbf{I}_d)$ be a learned sGMM. Figure 6.1 shows an example of stable and unstable distribution sGMMs, and further for the stable sGMM it constructs good-fit region such that all sGMM within small total variation distance to the population must have parameters located in these regions. These regions can be reported as the usual confidence regions are reported in a regular parametric estimation problem.

Furthermore, Definition 6.1 concerns only the model class $\mathcal{M}$, a subset of spherical Gaussian mixtures. Hence, one can obtain distribution free stability results generically, by studying the stability of distributions inside a model class, such as that of sGMM. While this remark is nearly obvious, suprisingly little work has attended to the possiblity of obtaining distribution free guarantees as a side effect of consistency or identifiability proofs. We hope

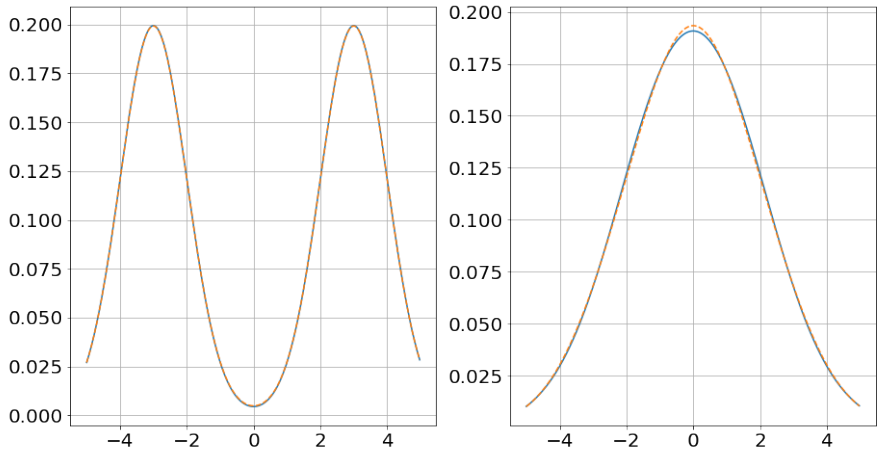


Figure 6.1: Stable and unstable spherical Gaussian Mixtures in 1D. **Left:** A well separated mixture of $K = 2$ Gaussians, $P = 0.5N(-3,1) + 0.5N(3,1)$, with separation $c = 3$ (as defined in Section 6.2, A3), superimposed on a distribution Q such that $TV(P,Q) \leq 0.001$. Our Theorem 6.2 in Section 6.2 guarantees that any mixture P' with $K = 2$ components, minimal weights 0.45, and separation constants 3 that fits Q equally well must have parameters close to P; namely (see Section 6.2 for the parameter definitions) μ_i' within 0.0200 of -3 and 3, $\max\{1/\sigma_i'^2, \sigma_i'^2\} \leq 1.034$, and $|0.5 - w_i'| \leq 0.004$. **Right:** Two spherical Gaussian mixtures P, P' which are unstable, in the sense that while they are close in TV distance, they have very different parameters; $P = 0.0625N(-3,\sigma^2) + 0.4375N(-1,\sigma^2) + 0.4375N(1,\sigma^2) + 0.0625N(3,\sigma^2)$, $P' = 0.0078125N(-4,\sigma^2) + 0.21875N(-2,\sigma^2) + 0.546875N(0,\sigma^2) + 0.21875N(2,\sigma^2) + 0.0078125N(4,\sigma^2)$, with $\sigma^2 = 2.25$. In this example, the mixture components are less separated, and some of the mixture proportions are small as well.

that one contribution of this paper be at the conceptual level, in drawing attention to this possibility, which remains the primary motivator for this work.

In Section 6.2 we define model classes $\mathfrak{M}$ of interest, and instantiate $d_{param}(\cdot,\cdot)$ and all other parameters, then state our main stability result in Theorem 6.2. In the following

sections 6.3 and 6.4 we present the main result in different scenarios. Section 6.5 illustrates the result with some numeric examples. Section 6.7 provides the detailed proof of our main result Theorem 6.2, with additional lemmas proved in Section 6.8.

6.2 Problem Formulation

We first pose the problem in formal terms. As in the definition 6.1, we identify three key components: (i) model class $\mathfrak{M}$, (ii) distance or divergence between the models in parameter space $d_{param}(P, P')$, and (iii) goodness of fit measure.

Model Class A spherical Mixture of Gaussians P over $\mathbb{R}^d$ can be written in the form

$$P = \sum_{k=1}^{K} w_k \mathcal{N}_d(\mu_k, \sigma_k^2 \mathbf{I}_d), \quad \text{with } \sum_{k=1}^{K} w_k = 1, \quad w_k \geq 0. \tag{6.1}$$

In the above, we have adopted the standard notation, whereby $\mathcal{N}_d(\mu_k, \sigma_k^2 \mathbf{I}_d)$, called *mixture components*, are normal distributions with means $\mu_{1:K} \in \mathbb{R}^d$ and diagonal covariance matrices $\sigma_{1:K}^2 \mathbf{I}_d \in \mathbb{R}^{d \times d}$, while $w_{1:K}$ are called *mixture proportions*.

Further on, we assume the number of components K is fixed, and we add restrictions on the smallest component proportion and the component separation. Thus the model classes we consider are denoted $\mathfrak{M}(K, w_{\min}, w_{\max}, c)$.

Definition 6.2. *An sGMM P is in model class $\mathfrak{M}(K, w_{\min}, w_{\max}, c)$ iff the following holds:*

A1 $K \geq 2$.

A2 $\min_{k \in [K]} w_k \geq w_{\min}$, $\max_{k \in [K]} w_k \leq w_{\max}$.

A3 $||\mu_i - \mu_j|| > c(\sigma_i + \sigma_j)$ *for any* $i, j \in [K]$ *and* $i \neq j$.

For $P \in \mathfrak{M}(K, w_{\min}, w_{\max}, c)$, the maximal proportion is always less than or equal to $1 - (K-1)w_{\min}$. Therefore, we abbreviate the model class notation to be $\mathfrak{M}(K, w_{\min}, c)$ when $w_{\max} = 1 - (K-1)w_{\min}$. The separation constant c is necessary. Theorem 3.1 in [Regev and Vijayaraghavan, 2017] points out that as K increase, when separation constant c is $o(\sqrt{\log K})$, it is possible to find two sGMMs so that their parameters are $\Omega(1)$ different but their total variation distance is superpolynomially small in K. We rule out this difficulty by deriving assumptions on the separation constant c, which regularizes the model class.

The divergence of model parameters between any two models P, P' with K components is

$$d_{param}(P,P') = \min_{\text{perm}\in\mathfrak{S}_K} \max_{i\in[K]} \frac{|w_i - w'_{\text{perm}(i)}|}{\min\{w_i, w'_{\text{perm}(i)}\}} + \frac{||\mu_i - \mu'_{\text{perm}(i)}||}{\max\{\sigma_i, \sigma'_{\text{perm}(i)}\}} + \frac{|\sigma_i^2 - \sigma'^2_{\text{perm}(i)}|}{\min\{\sigma_i^2, \sigma'^2_{\text{perm}(i)}\}} \ . \tag{6.2}$$

In the above, perm $\in \mathfrak{S}_K$ is a permutation of the set $[K]$; d_{param} is not necessarily a metric[1]. Theorem 6.2 presented below provides upper bounds on each of the three terms of (6.2) separately.

Consider $G = \sum_{k=1}^K w_k \delta_{(\mu_k,\sigma_k^2)}$ with δ being the point mass. One can also view each sGMM distribution $P_G = \sum_{k=1}^K w_k N(\mu_k, \sigma_k^2 \mathbf{I}_d)$ as a smoothed distribution of discrete probability on the parameter space. Let G, G' be two such distributions on the parameter space, then Wasserstein-1 distance defined by

$$W_1(G,G') = \inf_{\Pi} \int d((\mu_k,\sigma_k^2),(\mu'_k,\sigma'^2_k))\Pi(G,G')$$

is considered. This is shown to be a natural parameter divergence in several previous papers (e.g.,[Ho and Nguyen, 2016, Heinrich and Kahn, 2018, Wu and Yang, 2020, Doss et al., 2020]). Our d_{param} does not capture the same topology as the W_1 distance on parameters. Consider two sequences $P_n = 0.5N(-1,(n+1)^2) + 0.5N(1,(n+1)^2)$ and $Q_n = 0.5N(-1,n^2) + 0.5N(1,n^2)$. Then, as $n \to \infty$, $d_{param}(P_n, Q_n)$ converges to zero, but $W_1(n^2,(n+1)^2)$ diverges. On the other hand it can also be shown that for a fixed G and corresponding P_G, when $W_1(G,G') \to 0$, it holds that

$$d_{param}(P_G, P_{G'}) \asymp W_1(G,G') \asymp \min_{\text{perm}\in\mathfrak{S}^K} \sum_{i=1}^K |w_i - w_{\text{perm}(i)}| + ||\mu_i - \mu'_{\text{perm}(i)}|| + |\sigma_i^2 - \sigma'^2_{\text{perm}(i)}| \ . \tag{6.3}$$

Hence our definition of d_{param} is in better agreement with the model fit criterion, does not need the assumption that the parameter space is compact [Ho and Nguyen, 2016] and there will be a direct way to construct good-fit regions or confidence regions since we provide upper bounds on the terms in d_{param} separately. Also, d_{param} is invariant w.r.t. rescaling, and this will be useful in, e.g. model-based clustering.

---

[1] d_{param} is also not a divergence in the information geometric sense.

Goodness of fit can be measured by various criteria (e.g. likelihood or KL divergence, W_2 distance), but in this paper total variation distance [Gibbs and Su, 2002] is used for its convenience. For two probability distributions P, Q on $\mathbb{R}^d$, the *total variation distance* is given by

$$TV(P,Q) = \sup_{A\in\mathcal{B}} |P(A) - Q(A)| \,, \tag{6.4}$$

where $\mathcal{B}$ is all Borel sets on $\mathbb{R}^d$.

6.3 Symmetric reuslt

In this section, we formulate Theorem 6.1 under the assumption that both P, P' are assumed in a model class $\mathfrak{M}(K, w_{\min}, c)$. We will show that if a priori it is known that P, P' are both well-separated Gaussian mixtures, then parametric stability can hold. More importantly, we provide tractable bound so that such the deviation on parameters is computable.

Before we formulate for the specific case of mixtures of spherical Gaussian, we introduce several constants depending on the model class $\mathfrak{M}(K, w_{\min}, c)$ and ϵ, but independent of the actual distributions considered. These parametrize the upper bounds on the terms of d_{param} (6.2) we are about present. The constants η_0, η^*, to be defined below, represent ratios between standard deviations; $\eta^* > 1$ will bound the perturbation in σ_i, and the aim is to bring it as close to 1 as possible. The constant c_0, together with η_0 gives the minimum sufficient value for the separation c, while c^* bounds the perturbation of the components' means.

Denote by $\Phi(x)$ the CDF of the standard normal distribution in one dimension and by F_d the CDF of Gamma$(\frac{d}{2}, \frac{d}{2})$ distribution.

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2w}} \exp\left(-\frac{1}{2}t^2\right) dt \,, \quad F_d(x) = \int_0^x \frac{(\frac{d}{2})^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} t^{\frac{d}{2}-1} \exp\left(-\frac{d}{2}t\right) dt \,. \tag{6.5}$$

Constants c_0, η_0 determined by $w_{\min}, \epsilon, w_{\max}$(or K) are defined by

$$c_0 = 2\Phi^{-1}(1 - \frac{w_{\min} - 2\epsilon}{2}) \,; \tag{6.6}$$

η_0 satisfies $\eta_0 \geq 1$ and

$$1 - \frac{w_{\min} - 2\epsilon}{w_{\max}} + \frac{2(1 - w_{\max})}{w_{\max}} \Phi(-\frac{1}{2}\eta_0 c_0) = F_d(\frac{2\eta_0^2 \log \eta_0}{\eta_0^2 - 1}) - F_d(\frac{2\log \eta_0}{\eta_0^2 - 1}) \,. \tag{6.7}$$

Lemma 6.16 (Section 6.7) will show that the right hand side of (6.7) is in fact $TV(N(0,\mathbf{I}_d), N(0,\eta^2\mathbf{I}_d))$ and is lower bounded by $1-(2\eta/(\eta^2+1))^{d/2}$.

With the values c_0, η_0 one can establish upper bounds for the three terms of equation (6.2); in particular $c_0 \max\{\sigma_i, \sigma'_{\mathrm{perm}(i)}\}$ bounds the first term, representing the means' variation. This is shown in detail in Section 6.7.3. We call c_0, η_0 *initial bounds* because, in fact, they can be further reduced, by a technique that we present in Section 6.7.4. Thus, we obtain *refined* parameter distance upper bounds depending on constants c^*, η^* which satisfy $0 \le c^* \le c_0$, $1 \le \eta^* \le \eta_0$ and the following conditions

$$1-2\Phi(-\frac{c^*}{2}) = \frac{2\epsilon}{w_{\min}} + \frac{2(1-w_{\min})}{w_{\min}}\Phi(-\frac{1}{2}[c(1+\frac{1}{\eta^*})-c^*]) \;, \tag{6.8}$$

$$F_d(\frac{2(\eta^*)^2\log\eta^*}{(\eta^*)^2-1}) - F_d(\frac{2\log\eta^*}{(\eta^*)^2-1}) = \frac{2\epsilon}{w_{\min}} + \frac{2(1-w_{\min})}{w_{\min}}\Phi(-\frac{1}{2}[c(1+\frac{1}{\eta^*})-c^*]) \;. \tag{6.9}$$

Note that through equations (6.8) and (6.9), c^*, η^* are also determined by $w_{\min}, \epsilon$, c and $w_{\max}$. The following section shows that the four constants c_0, η_0, c^*, η^* exist and are unique. Now we are ready to state our first main result.

Theorem 6.2. *Let $P \in \mathfrak{M}(K, w_{\min}, w_{\max}, c)$. Suppose P' is any model in $\mathfrak{M}(K', w_{\min}, w_{\max}, c)$ such that $TV(P,P') \le 2\epsilon$ where $\max\{K,K'\} \le 1/w_{\min}$, $w_{\max} \le 1-(\min\{K,K'\}-1)w_{\min}$. Let c_0, η_0 be defined as in (6.6) and (6.7). Then, if $c \ge c_0\eta_0$ and $w_{\min} > 2\epsilon$, we have $K = K'$ and further, there exists a permutation $\mathrm{perm} \in \mathfrak{S}_K$ and constants $c^* \in [0, c_0], \eta^* \in [1, \eta_0]$ satisfying (6.8) and (6.9), such that for each $i \in [K]$,*

$$TV(P_i, P'_{\mathrm{perm}i}) \le \frac{2\epsilon}{w_{\min}} + \frac{2(1-w_{\min})}{w_{\min}}\Phi(-\frac{1}{2}[c(1+\frac{1}{\eta^*})-c^*]) \;, \tag{6.10}$$

$$||\mu_i - \mu'_{\mathrm{perm}(i)}|| \le c^*\eta^*\sigma_i \;, \tag{6.11}$$

$$\max\{\sigma_i/\sigma'_{\mathrm{perm}(i)}, \sigma'_{\mathrm{perm}(i)}\sigma_i\} \le \eta^* \;, \tag{6.12}$$

$$|w_i - w'_{\mathrm{perm}(i)}| \le 2\epsilon + (1-w_{\min}+w_{\max})\Phi(-C(c,c^*,\eta^*)) \;, \tag{6.13}$$

where $C(c,c^*,\eta^*)$ is defined by

$$C(c,c^*,\eta^*) := \sqrt{\frac{c^2}{2(\eta^*)^2} + \frac{1}{2\eta^*}(c-\frac{c^*}{2})^2 - \frac{(c^*)^2(1+\eta^*)^2}{16(\eta^*)^2}} - \frac{c^*}{2} \;. \tag{6.14}$$

This theorem extends the usual identifiability result. Furthermore, this upper bound is tractable since all the constants are explicit or computable and it does not assume prior knowledge on parameters. Computability opens up the possibility of applying our result to finite samples.

The following proposition gives an estimate of the constants c_0, η_0, c^*, η^* in the asymptotic regime that $K \to \infty$.

Proposition 6.1. *Given $w_{\min} > 2\epsilon > 0$, then c_0, η_0, c^*, η^* defined in equations (6.6)–(6.8) exist and each is unique. Further suppose there are nonnegative constants α, β and sufficiently large positive constant ι such that $\beta \leq 1+\alpha$, $w_{\min} = \Omega(K^{-(1+\alpha)})$,$w_{\max}/w_{\min} = O(K^{\beta})$ and $2\epsilon/w_{\min} \leq \iota \ll 1$. When $K \to \infty$, we have the initial separation condition $c_0 = O(\sqrt{\log K})$, $\eta_0 = O(K^{2\beta/d})$. With any separation $c > c_0\eta_0$, the ultimate upper bound $c^* = O(1)$ and $\eta^* = O(1)$, i.e. they are bounded when $K \to \infty$.*

We make several remarks here. First, specifically, for balanced model classes where $w_{\max}/w_{\min} = 1$, we conclude that the minimal separation condition is $c > c_0\eta_0 = O(\sqrt{\log K})$. With the prior knowledge of ratios between standard deviations, our result matches the sharp separation threshold $O(\sqrt{\log K})$ established in [Regev and Vijayaraghavan, 2017] for learning well-separated mixture of Gaussians. Second, [Regev and Vijayaraghavan, 2017] also shows that for sufficiently large C and $K > C^8$, there exist two mixtures P, P' in the model class $\mathfrak{M}(K, 1/K, C^{-24}\sqrt{\log K})$ with unit variance for every components, such that their parameter distance is at least $C^{-24}\sqrt{\log K}$ but $d_{TV}(P, P') \leq K^{-C}$. However, with larger multipliers of $\sqrt{\log K}$ in the separation constants, our result further shows that the parameter distance is not diverging as $K \to \infty$. That is, the mixtures can be identified componentwisely no matter how many clusters there are under our separation conditions.

Third, when K is fixed, the separation lower bound $c_0\eta_0$ decreases in d. Informally, with the same K, $w_{\min}, w_{\max}$, it is easier to distinguish two Gaussians mixtures in higher dimension. This is an instance of the *blessing of dimensionality* for Gaussian Mixture Models [Anderson et al., 2014].

Fourth, constraints on maximal and minimal standard deviation of each components also provide a valid way of regularization. If the components of P, P' satisfiy the relation

$c > \rho_\sigma c_0$ where $\rho_\sigma = \max_{i,j}\{\sigma_i, \sigma'_j\}/\min_{i,j}\{\sigma_i, \sigma'_j\} \leq \eta_0$, then the one to one correspondance also holds. Finally, when Theorem 6.2 applies, the two mixtures must have same number of components. This further leads to the following corollary showing that a well-separated Gaussian mixtures cannot be close in total variation distance to a single spherical Gaussian. In the proof in Section 6.7, we prove this corollary and specify the constants.

Corollary 6.1. *For any $\epsilon < w_{\min}$, there is a positive constant c_{single} depending on $w_{\min}$ and ϵ such that when $c > c_{\text{single}}$, no spherical Gaussians are within total variation distance 2ϵ from any P in $P \in \mathfrak{M}(K, w_{\min}, c)$.*

We shall point out that the upper bounds in Theorem 6.2 do not tends to 0 as $d_{TV}(P, P') \to 0$. Ideally, for a fixed mixture distribution $P \in \mathfrak{M}(K, w_{\min}, c)$, it holds that $\liminf TV(P, P')/d_{param}(P, P') > 0$ as $d_{param}(P, P') \to 0$, due to a local Taylor expansion analysis proposed in [Ho and Nguyen, 2016]. On the other hand, our ultimate bound c^* and η^* have two parts: the first part coming from the total variation distance between two Gaussians and the second part coming from not-far-enough separation between components. According to the author's knowledge it is not known if one can obtain computable parameter distance bound between P, P' which scales as $O(TV(P, P'))$ as the total variation distance $TV(P, P')$ tends to zero.

6.4 Non-symmetric result

In previous section, we have assumed P, P' are both taken from the same model class $\mathfrak{M}(K, w_{\min}, c)$. The guarantee we obtained will be affected by the selection of parameters $K, w_{\min}, c$ largely. In this section, we will relax the separation assumption on P'. Throughout this section, $w_{\min}$ is a pre-defined constant and $P = \sum_{k=1}^{K} w_k \mathcal{N}_d(\mu_k, \sigma_k^2 I_d)$ be a known sGMM in model class $\mathfrak{M}(K, w_{\min}, c)$.

We study whether another sGMM P' is sufficiently close to P in total variation distance implies its separation can be lower bounded and further leads to similar parametric guarantee as in the previous section. Again it will be problematic if the minimal component weight of P' is not lower bounded, as there can be as many components as possible to have a same P', we still need to require that the minimal weight proportion $w'_{k'}$ is also lower bounded by $w_{\min}$. We will show that only under the assumption of $w_{\min}$, it is also possible

Algorithm 9 ONE2ONECRITERION

Input: sGMM $P = \sum_{k=1}^{K} w_k \mathcal{N}_d(\mu_k, \sigma_k^2 I_d)$, minimal weight component $w_{\min}$, total variation distance threshold ϵ and test parameter $\rho < w_{\min} - 2\epsilon$.
Compute separation parameter c of P as $c = \min_{i \neq j} ||\mu_i - \mu_j||/(\sigma_i + \sigma_j)$, variance ratio parameter $\kappa = \max_{i \neq j} \sigma_i/\sigma_j$.
Initial $C_s^{(0)} = C_b^{(0)} = 2\Phi^{-1}(1 - \rho/2)$.
while Sequence $\{C_s^{(t)}\}_{t=0,1,2,\cdots}$ not converged **do**
  if $t(0, \frac{4c - 2C_b^{(t)}}{C_s^{(t)} + C_b^{(t)}}) > 1 - \rho$ **then**
    Stop and output True: one-to-one correspondence holds.
  else
    Solve for $C_s^{(t+1)}$ through $t(\frac{C_s^{(t+1)}}{2}, \frac{4c - 2C_b^{(t)}}{C_b^{(t)} + C_s^{(t)}}) = 1 - \rho$.
    Solve for $C_b^{(t+1)}$ through $t(\frac{C_b^{(t+1)}}{2}, \frac{(2c - C_b^{(t)})(1 + \frac{1}{\kappa})}{C_b^{(t)} + C_s^{(t)}}) = 1 - \rho$.
  end if
end while
Output one-to-one correspondence not guaranteed.

to guarantee parametric stability of P as in previous section, under a stricter assumption on separation constant c and total variation distance ϵ.

Let Φ_d be the standard multivariate Gaussian distribution in $\mathbb{R}^d$, and Φ' is $\mathcal{N}_d(C(1 + \eta), \eta^2 I_d)$ distribution with $C \geq 0, \eta \geq 1$. We use $t(C, \eta)$ to denote the total variation distance between P and P'. Lemma 6.3 and 6.18 show that $t(C, \eta)$ is increasing in C and η respectively. With function $t(C, \eta)$, we first construct a sufficient condition such that a one-to-one correspondence between P and any sGMM P' with minimal weight greater than $w_{\min}$ and total variation distance $TV(P, P') \leq 2\epsilon$ can be checked in the sense that only paired components are closed in total variation distance. This procedure is described as algorithm ONE2ONECRITERION.

Note that when P is completely known, one can also utilize information more specific. For example in procedure ONE2ONECRITERION , we introduce a new parameter

$$\kappa = \max_{i \neq j} \sigma_i/\sigma_j$$

, which is readily computed from any known sGMM P.

Theorem 6.3. *Given a minimal component weight $w_{\min}$ and an sGMM $P \in \mathfrak{M}(K, w_{\min}, c)$. If there exists a positive constant $\rho < w_{\min} - 2\epsilon$ such that $c > 2\Phi^{-1}(1-\rho/2)$ and* One2OneCriterion *outputs that one-to-one correspondence holds, then any sGMM $P' = \sum_{k'=1}^{K'} w'_{k'} \mathcal{N}_d(\mu_{k'}, \sigma^2_{k'} I_d)$ such that $TV(P', P) \leq 2\epsilon$ and $\min_{k'} w'_{k'} \geq w_{\min}$ must have $K' = K$ and for each component P'_i of P', there exists a unique component P_i of P such that $TV(P_i, P'_i) \leq 1 - \rho$.*

Further, for each pair of corresponding components P_i, P'_i it holds that $\max\{\sigma_i/\sigma'_i, \sigma'_i/\sigma_i\} \leq \eta_\rho$ where η_ρ is given by equation

$$1 - \frac{w_{\min} - 2\epsilon}{w_K} + \frac{(1 - w_K)\rho}{w_K} = F_d(\frac{2\eta_\rho^2 \log \eta_\rho}{\eta_\rho^2 - 1}) - F_d(\frac{2 \log \eta_\rho}{\eta_\rho^2 - 1}) \,, \tag{6.15}$$

and separation constant of P', denoted as c', satisfies that

$$c' := \min_{i \neq j} \frac{||\mu'_i + \mu'_j||}{\sigma'_i + \sigma'_j} \geq \max\{0, \frac{c}{\eta_\rho} - \frac{C_0(1 + \eta_\rho)}{2\eta_\rho}\} \,, \tag{6.16}$$

where $C_0 = 2\Phi^{-1}(1 - (w_{\min} - 2\epsilon)/2)$.

Specifically, let $c_\rho^{\max}, \eta_\rho^{\max}$ be the solution of $t(c_\rho^{\max}, 1) = 1 - \rho, t(0, \eta_\rho^{\max}) = 1 - \rho$ respectively. if $(c/c_\rho^{\max} - 1/2)(1 + 1/\kappa) > \eta_\rho^{\max}$ or equivalently $c > (\kappa \eta_\rho^{\max}/(\kappa + 1) + 1/2) c_\rho^{\max}$, then procedure One2OneCriterion will stop at the first iteration and hence the one-to-one correspondence between components of P, P' can be established. Therefore, introducing the procedure One2OneCriterion relaxes the requirement on separation in the sinse that it only requires iterations stop in any finite steps. This procedure also leads to another neat sufficient condition as follows.

Proposition 6.2. *Given a minimal component weight $w_{\min}$ and an sGMM $P \in \mathfrak{M}(K, w_{\min}, c)$. If there exists a positive constant $\rho < w_{\min} - 2\epsilon$ such that*

$$t\left(\frac{2(\kappa + 1)c}{2\kappa\eta + \kappa + 1}, \eta\right) > 1 - \rho$$

holds for all $\eta \geq 1$, then the One2OneCriterion *will output True for this $P, w_{\min}, \epsilon$ and ρ.*

Ratewise, under the same setting of 6.1, one can show that the separation requirement is $O(K^{\frac{2(1+\alpha)}{d}} \sqrt{\log K})$ when $w_{\min} = \Omega(K^{-(1+\alpha)})$. In conclusion, this section provides a

prelimary result showing that any sGMM close to a really well-separated sGMM is also well-separated globally.

6.5 Numerical Examples

The upper bounds as well as the conditions in Theorem 6.2 are computable and here we evaluate them on a variety of examples.

Minimum Separation Recall that for Theorem 6.2 to apply, both Gaussian mixtures P, P' need to have well separated components, with relative separation $c \geq c_0\eta_0$. Here we calculate the minimal separation $c_0\eta_0$ in the limit case $\epsilon = 0$. This will give a view of the domain of applicability of the theorem.

We consider mixtures of spherical Gaussians in $d = 5, 20, 35$ dimensional Euclidean spaces with the number of components K from 2 to 40. For each d and K, the ratio $\eta_\pi = w_{\max}/w_{\min}$ is set to 1,2,4,8, and 16. From K and η_π, we set $w_{\min}, w_{\max}$ so that $w_{\max} = 1-(K-1)w_{\min}$. Figure 6.2 illustrates what the minimal separation requirements are so that Theorem 6.2 can be applied. We observe that the heterogeneity of mixture proportions η_π indeed has a large effect on the separation requirement; with η_π as large as 16, one may need separation constant $c \geq 12$ to apply Theorem 6.2 in dimension 5 even with 2 clusters. On the other hand, for balanced mixtures, the requirement is not so severe. For mixtures in $\mathcal{M}(2, 1/2, c)$ namely with two equal components, even for $d = 1$ the lower bound for c is 2.29. As expected [Anderson et al., 2014], when d increases, this requirement decreases as low as $c = 1.55$ when $d = 35$. When K increases, $c_0\eta_0$ increases at the rate of approximatevely $\sqrt{\log K}$.

Upper Bounds for parameter divergence Here we present the numerical values of the upper bounds c^*, η^* as well as of the upper bound on the relative difference $|wi - \pi'_{\mathrm{perm}(i)}|/w_{\min}$ from equation (6.13), for different levels of model fit ϵ. We consider $P \in \mathcal{M}(K, 1/(K+1), c)$ on $\mathbb{R}^{20}$, and vary $K = 2, 5, 10$, and $c = 3, 4, 5, 6$. Figure 6.3 shows the values of the three bounds in this scenario. For example, a mixture of $K = 2$ components in $d = 20$ dimensions requires a separation $c \approx 3$ for $\epsilon \approx 0.01$ by Theorem 6.2. Once this

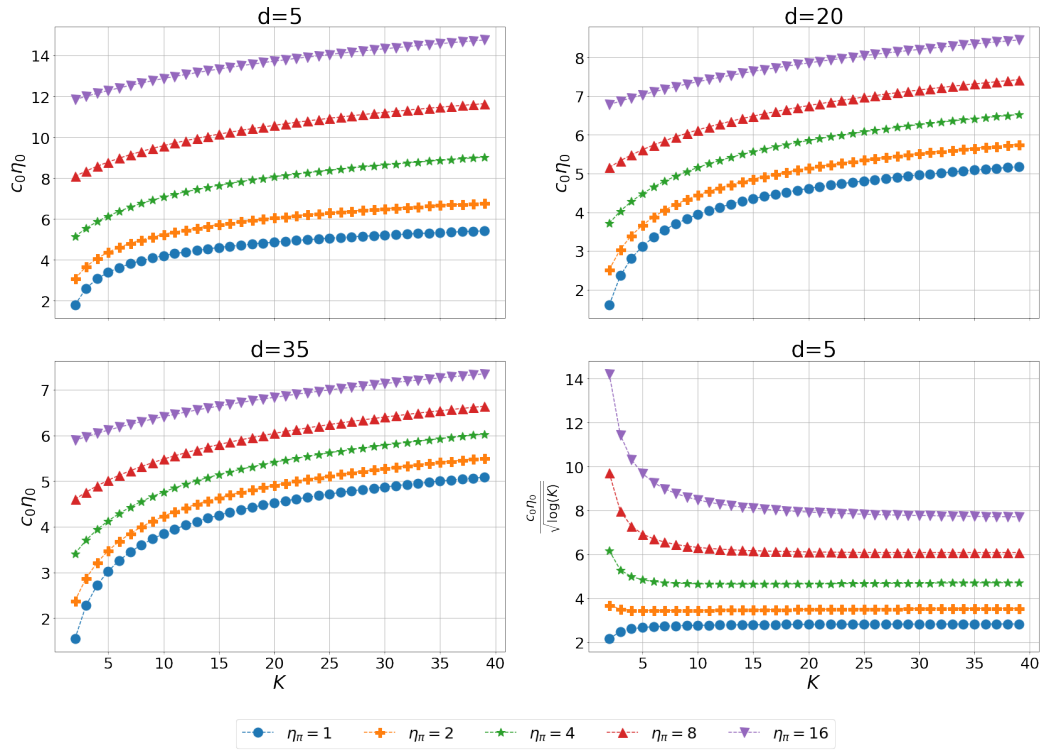


Figure 6.2: **Sufficient minimal separation** $c_0\eta_0$ in Theorem 6.2 under different settings. **Top left, Top right, Bottom Left** show the dependence of $c_0\eta_0$ on K and $\eta_\pi = w_{\max}/w_{\min}$ in dimensions $d = 5, 20, 35$, respectively. **Bottom right** shows that the dependence of $c_0\eta_0$ on K asymptotes to $\sqrt{\log K}$.

condition holds, we have good guarantees: for each pair of corresponding components,

$$||\mu_i - \mu_i'|| \leq 0.151 \max\{\sigma_i, \sigma_i'\}, \quad \frac{\max\{\sigma_i, \sigma_i'\}}{\min\{\sigma_i, \sigma_i'\}} \leq 1.035, \quad |w_i - w_i'| \leq 0.02 \approx 0.06 w_{\min} .$$

From Figure 6.3 we see that when ϵ is small, all bounds are dominated by the separation c. As ϵ increases, we observe that all three bounds are dominated by ϵ. Note the relation between these graphs and the orange curve $\eta_\pi = 2$ from Figure 6.2.

6.6 Discussion

Comparisons with recent robust identifiability results Theorem 6.2 is the first result to our knowledge to provide computable *global* parameter stability upper bounds of

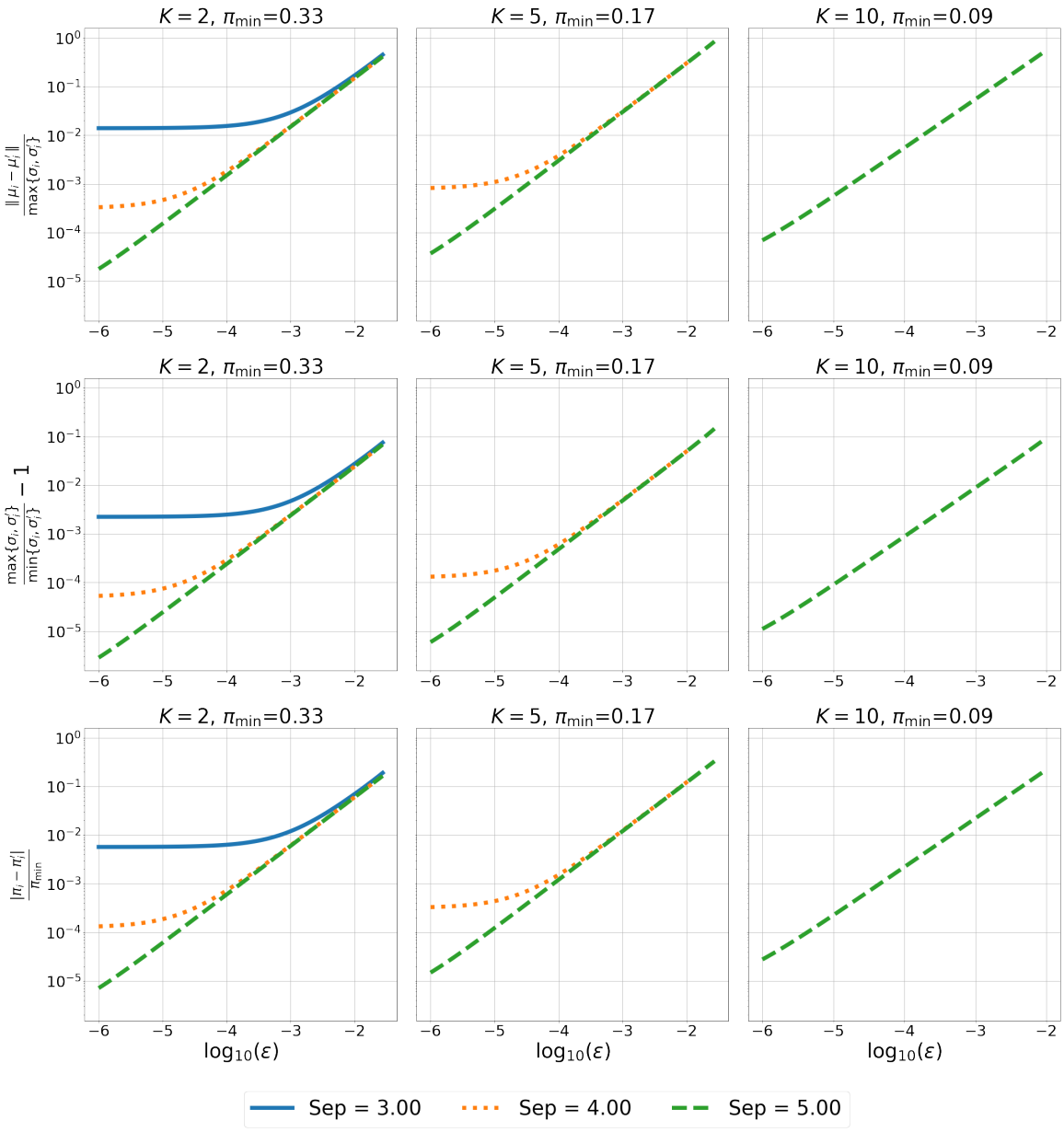


Figure 6.3: **Upper Bounds** on the distance between corresponding means c^* (top row), ratio of standard deviations η^* minus one (middle row) and the difference in mixture proportions measured in multiples of $w_{\min}$ (bottom row) from Theorem 6.2 for different values of K, and of separation Sep= c, in $d = 20$ dimensions. Some curves don't appear because the separation is not large enough to obtain the upper bound, as indicated by the Top, Left panel of Figure 6.2 (orange curve).

mixtures of well-separated sGMMs. This result can also be seen as a *robust identifiability* result for $\mathfrak{M}(K, w_{\min}, c)$.

When $TV(P, P') \to 0$, some identifiability result exists using W_1 distance of parameters: [Ho and Nguyen, 2016],[Heinrich and Kahn, 2018], etc, show that for two sets of parameter distributions G, G' with $W_1(G, G')$ small, $TV(P_G, P'_{G'})$ is asymptotically greater than the $W_1(G, G')$ distance (without assuming separation). This result is almost the converse of our Theorem 6.2.

The closest related works are the parameter identifiability theorems in robust learning of Gaussian mixtures, namely Theorem B.1 in [Diakonikolas et al., 2017], Theorem 8.1 in [Liu and Moitra, 2021], Theorem 9.1 in [Bakshi et al., 2020], which are based on newly developed methods of moments proof techniques.

Consider two mixtures of general Gaussians P, P' with maximal number of components being K. The assumptions made are that $TV(P, P') \leq \epsilon$, minimal weights of P, P' are greater than ϵ^{b_1} and different components of P, P' have separation in total variation distance greater than ϵ^{b_2}, for sufficiently small constants $b_1, b_2 > 0$ depending only on K. The key observation used in the proofs is that the parameters distances of $P_k, P'_{\text{perm}(k)}$ are bounded by $poly(\epsilon)$ if the Hermite moment polynomials is bounded by $poly(\epsilon)$, which is guaranteed by the assumption that the total variation distance between P, P' is upper bounded by ϵ. Their proof technique can be directly modified into proving that P is $(\epsilon, \tilde{\mathsf{d}}(\epsilon))$ stable with $\tilde{\mathsf{d}}(\epsilon) = O_k(\epsilon^{b_3^K})$, where b_3 is another sufficiently small positive constant.

When P is fixed (the case of interest both for us and for robust identifiability in general), if $\epsilon \to 0$ these results are tighter than ours as the bound $\tilde{\mathsf{d}} \to 0$. In this case, the assumptions on P become more relaxed than ours. However, it is not known how to determine the unspecified constants in $\tilde{\mathsf{d}}$ from their proof techniques, and the dependence on K and exponent of ϵ are not optimal (Regev and Vijayaraghavan [2017], respectively in the discussion in Section 6.2).

For larger, more realistic ϵ, the rate $\tilde{\mathsf{d}}(\epsilon) \sim \epsilon^{b_3^K}$ is much slower than what our numerical simulations achieve, which appears as $\mathsf{d}(\epsilon, \mathfrak{M}) \sim \epsilon$. In this case, the assumptions on $w_{\min}$ and separation of Theorem 6.2 remain fixed, while the assumptions in Liu and Moitra [2021], Bakshi et al. [2020] become more restrictive (or may even not apply) with larger ϵ.

Hence, our results are more useful for robust recovery, where extending stability to larger ϵ is desired, while [Liu and Moitra, 2021, Bakshi et al., 2020] are useful in the limit $\epsilon \to 0$ and for *algorithmic* stability (the main goal of these works). This growth of $\mathsf{d}(\epsilon, \mathfrak{M}) \sim \epsilon$ matches the sharp threshold of [Ho and Nguyen, 2016], which is obtained asymptotically for $\epsilon \to 0$, suggesting that [Ho and Nguyen, 2016] may hold for larger ϵ and that our worst case bound may match it at least under certain conditions.

Regarding the dependence on K, a sharp threshold for sGMM was obtained by [Regev and Vijayaraghavan, 2017], who show that $\Omega(\sqrt{\log K})$ separation is necessary and sufficient to recover this class of GMM in polynomial sample size and time. Our Theorem 6.2 is approximately matching this threshold.

Other related problems Solving for exact MLE in Spherical Gaussian Mixture Models has been proved to be NP-hard [Tosh and Dasgupta, 2018]. The current mainstream approach to fitting GMM is the *Expectation-Maximization* (EM) algorithm [Dempster et al., 1977]; One cannot prove that EM will converge to the global maximum of log-likelihood function without further assumptions, and in fact, EM can converge to bad local maxima with high probability [Jin et al., 2016a], even for well-separated Gaussians; the same paper also confirms the existence of local maxima in the population likelihood function. This negative result makes it the more necessary to have a post-processing validation stage, in which to be able to reject bad local optima. The results in our paper can fulfill this role, in the population sense.

A second approach to estimating GMM parameters with consistency guarantees relies on *Methods of Moments* (MOM) [Pearson, 1894], and is exemplified in [Hsu and Kakade, 2013, Wu and Yang, 2020]. [Hsu and Kakade, 2013] used moments up to the third order to learn an sGMM. They don't require separation conditions, but need the means of different components to be linear independent. Therefore, this method cannot be applied to low d and large K scenario. [Wu and Yang, 2020] used an semi-definite programming based denoising procedure of moments up to order $O(K)$. When the cluster components are well-separated, a different line of research proposed in the seminal paper [Dasgupta, 1999] and refined in [Sanjeev and Kannan, 2001, Achlioptas and McSherry, 2005, Kannan et al., 2008],etc, and

[Dasgupta and Schulman, 2007], provides an accurate estimation of GMM parameters with high probability. All these results are predicated on the data being sampled from a mixture of well separated Gaussians.

Also, previously very few results have been established on evaluating the result of a Gaussian Mixture fit. In these works, e.g. [Drton and Plummer, 2017], [Huang et al., 2017]), the main target is model selection, especially consistently estimating the number of components in a Gaussian Mixture population.

In conclusion, this chapter obtains the first *computable* robust identifiability bounds for spherical Gaussian mixtures, in the population setting. Our bounds can be extended to Gaussian mixtures with full covariance matrices Σ_k, and bounded excentricity. The bounds $\mathsf{d}(\epsilon, \mathfrak{M})$ match the known sharp threshold w.r.t K from Regev and Vijayaraghavan [2017], and are uniform over the model class $\mathfrak{M}(K, w_{\min}, c)$.

In the chapter we introduce an iterative approach for tightening the bounds which is original, to our knowledge. Several other results of our Lemmas can be of independent interest, such as a tighter bounds on parameter variation for a single Gaussian, that remain informative for larger perturbations in TV distance than the previous bound of Devroye et al. [2020].

6.7 Proof of Results in Chapter 6

6.7.1 Proof of Proposition 6.1

Proof of Proposition 6.1. First c_0 exists and is unique from (6.6). The definition of c_0 shows that

$$\Phi(-c_0/2) = 1 - \Phi(c_0/2) = 1 - \Phi(\Phi^{-1}(1 - \frac{w_{\min} - 2\epsilon}{2})) = \frac{1}{2}(w_{\min} - 2\epsilon) \ . \tag{6.17}$$

Consider $R(\eta)$ to be a function defined by the the R.H.S. of (6.7) tending to zero, i.e.

$$R(\eta) = F_d\left(\frac{2\eta^2 \log \eta}{\eta^2 - 1}\right) - F_d\left(\frac{2 \log \eta}{\eta^2 - 1}\right) \tag{6.18}$$

and $R(\eta) \to 0$ when $\eta \to 1^+$. According to Lemma 6.18, $R(\eta)$ is increasing in η and tends to one when $\eta \to \infty$. Denote

$$L(\eta) = 1 - \frac{w_{\min} - 2\epsilon}{\pi} + \frac{1 - \pi}{\pi} 2\Phi(-\frac{\eta_0 c_0}{2}) \ . \tag{6.19}$$

Then the L.H.S. of (6.7) can both be written as $L(\eta)$. Since

$$L(1) = 1 - \frac{w_{\min} - 2\epsilon}{\pi} + \frac{1-\pi}{\pi}(w_{\min} - 2\epsilon) \tag{6.20}$$

$$= 1 - (w_{\min} - 2\epsilon)\left(-\frac{1}{\pi} + \frac{1}{\pi} - 1\right) \tag{6.21}$$

$$= 1 - w_{\min} + 2\epsilon > 0 \tag{6.22}$$

and when $\eta \to \infty$, $L(\eta)$ tends to $1 - (w_{\min} - 2\epsilon)/\pi < 1$. Therefore since $L(\eta)$ is decreasing in η and $R(\eta)$ is increasing, there exists a unique η_0 that satisfies equation (6.7).

For the second part of this lemma, we now consider the case $w_{\min} = \Omega(K^{-(1+\alpha)})$ for some $\alpha \geq 0$,$w_{\max}/w_{\min} = O(K^{\beta})$ for some $0 \leq \beta \leq 1 + \alpha$ and ϵ small enough such that $w_{\min} - 2\epsilon = \Omega(K^{-(1+\alpha)})$. By (6.6), it holds that

$$\Phi(-\frac{1}{2}c_0) = \Omega\left(\frac{1}{2K^{1+\alpha}}\right) . \tag{6.23}$$

Note that for any constant $\zeta > 0$,

$$\Phi(-\frac{1}{2}2\sqrt{2\log(2K^{1+\alpha}/\zeta)}) \leq \exp\left(-\frac{1}{8}(2\sqrt{2\log(2K^{1+\alpha}/\zeta)})^2\right) = \frac{\zeta}{2K^{1+\alpha}} \lesssim \Phi(-\frac{1}{2}c_0) \tag{6.24}$$

where $a \lesssim b$ means there exists a constant $A > 0$ such that $a \leq A \cdot b$. Hence there exists some $\zeta_0 > 0$ such that

$$c_0 \leq 2\sqrt{2\log(\frac{2K^{1+\alpha}}{\zeta_0})} = O(\sqrt{\log K})$$

.

Now we upper bound η_0. Therefore since $\Phi(-c_0/2) \leq 1/2K \leq 1/4$ or $c_0/2 \geq 0.67 > 0.5$

$$\frac{1}{2K} \geq \Phi(-\frac{1}{2}c_0) \geq \frac{1}{5\sqrt{2\pi}\frac{1}{2}c_0}\exp(-\frac{1}{8}c_0^2) , \tag{6.25}$$

i we conclude that $\exp(-\frac{1}{8}c_0^2) = O(\sqrt{\log K}/K)$. Therefore,

$$L(\eta) = 1 - \frac{w_{\min} - 2\epsilon}{w_{\max}} + \frac{1 - w_{\max}}{w_{\max}} 2\Phi(-\frac{1}{2}\eta c_0) \tag{6.26}$$

$$\leq 1 - \frac{w_{\min} - 2\epsilon}{w_{\max}} + 2K\exp\left(-\frac{1}{8}\eta^2 c_0^2\right) \tag{6.27}$$

$$= 1 - \frac{w_{\min} - 2\epsilon}{w_{\max}} + O\left(K\left[\frac{\sqrt{\log K}}{K}\right]^{\eta^2}\right) . \tag{6.28}$$

On the other hand the R.H.S. of (6.7) is lower bounded by

$$R(\eta) > 1 - \left(\frac{2\eta}{\eta^2+1}\right)^{\frac{d}{2}} > 1 - \left(\frac{2}{\eta}\right)^{\frac{d}{2}}. \tag{6.29}$$

Hence take $\tilde{\eta} = 2(2w_{\max}/(w_{\min}-2\epsilon))^{2/d} = O(K^{2\beta/d})$, it holds that

$$R(\tilde{\eta}) - L(\tilde{\eta}) > \frac{1}{2}\frac{w_{\min}-2\epsilon}{w_{\max}} - \Omega\left(K\left[\frac{\sqrt{\log K}}{K}\right]^{\tilde{\eta}^2}\right). \tag{6.30}$$

However by the selection of $\tilde{\eta}$, $R(\tilde{\eta}) - L(\tilde{\eta})$ has to be positive for sufficiently large K. Therefore $\eta_0 < \tilde{\eta}$ and we conclude that for sufficiently large K, it holds that $\eta_0 = O(K^{2\beta/d})$.

To estimate c^*, η^*, we consider two difference cases:

- When $\eta_0 \leq 47$, apparently by definition $\eta^* \leq \eta_0$, and this leads to a bounded pairwise total variation distance (≤ 0.95).

- When $w_{\min} - 2\epsilon > 2\epsilon$, it is impossible to have $\eta_0 > 47$ from definition of η_0, (6.7). It remains to discuss the case $w_{\min} - 2\epsilon < 0.05$. Recall that c_0 is defined so as

  $$\Phi(-\frac{1}{2}c_0) = \frac{w_{\min}-2\epsilon}{2}.$$

  By the same argument as previously, we have $c_0 \leq 2\sqrt{2(\log(2/(w_{\min}-2\epsilon)))}$. Then, the R.H.S. of (6.8) is always upper bounded by

  $$\begin{aligned}
  \frac{2\epsilon}{w_{\min}} + \frac{2(1-w_{\min})}{w_{\min}}\Phi(-\frac{1}{2}[c(1+\frac{1}{\eta^*}) - c^*]) &\leq \iota + \frac{2}{w_{\min}}\Phi(-\frac{1}{2}c_0\eta_0) \\
  &\leq \iota + \frac{2}{w_{\min}}\exp\left(-\frac{1}{8}\eta_0^2 c_0^2\right) \\
  &\leq \iota + \frac{2}{w_{\min}}\left[\frac{5}{4}\sqrt{2\pi}c_0(w_{\min}-2\epsilon)\right]^{\eta_0^2} \\
  &\leq \iota + \frac{2}{w_{\min}}\left[5\sqrt{\pi}\sqrt{\log(\frac{2}{w_{\min}-2\epsilon})}(w_{\min}-2\epsilon)\right]^{\eta_0^2}
  \end{aligned}$$

  Note that for any $\rho \in (0, 0.05)$, the function

  $$h(\rho) = \frac{2}{\rho}\left[5\sqrt{\pi}\rho\sqrt{\log(\frac{2}{\rho}}\right]^{\eta_0^2}$$

  is upper bounded by 0.028. Hence pairwise total variation distance between matched component pairs is upper bounded by $\iota + 0.028 < 1$.

As a conclusion, for any matched components P_i, P_i', their total variation distance is $O(1)$ and smaller than 1 for sufficiently large K. Then we further conclude that $c^* = O(1)$ and $\eta^* = O(1)$. □

6.7.2 Technical Tools for Proving Theorem 6.2

Instead of directly diving into the proof of Theorem 6.2, we start by some observations in mixture models and analysis on total variation distances between spherical Gaussians. It is worthwhile noticing that in the lemmas in this subsection, we don't require the two mixture models P, P' to have the same number of components.

Total Variation Distance Between Spherical Gaussians

We develop a lemma upper bounding parameter distances between spherical Gaussians given their total variation distance. For this, one can use Hellinger distance, which includes an anlytical expression as a natural lower bound Gibbs and Su [2002], but they are usually not tight in constants and therefore they cannot separate the bound in mean parameter and variance parameter. Devroye et al. [2020] proposes another lower bound starting from the definitions but their results are only meaningful when the total variation distance is sufficiently small.

Lemma 6.1. *Suppose* $P_1 = N_d(\mu_1, \sigma_1^2 I_d)$ *and* $P_2 = N_d(\mu_2, \sigma_2^2 I_d)$. *Let* $C_0(\rho) = 2\Phi^{-1}(1 - \frac{\rho}{2})$ *and* $\eta_0(\rho)$ *be the solution of*

$$1 - \rho = F_d\left(\frac{2\eta^2 \log \eta}{\eta^2 - 1}\right) - F_d\left(\frac{2 \log \eta}{\eta^2 - 1}\right) . \tag{6.31}$$

Then the following holds:

- *If* $||\mu_1 - \mu_2|| \geq C_0(\rho)(\sigma_1 + \sigma_2)/2$, *then* $TV(P_1, P_2) \geq 1 - \rho$. *Equality holds iff* $\sigma_1 = \sigma_2$ *and* $||\mu_1 - \mu_2|| = C_0(\rho)(\sigma_1 + \sigma_2)/2$.

- *If* $\max\{\sigma_1/\sigma_2, \sigma_2/\sigma_1\} \geq \eta_0(\rho)$, *then* $TV(P_1, P_2) \geq 1 - \rho$. *Equality holds iff* $\mu_1 = \mu_2$ *and* $\max\{\sigma_1/\sigma_2, \sigma_2/\sigma_1\} = \eta_0(\rho)$.

We separate the Gaussian total variation lower bound into three different lemmas and prove them.

Lemma 6.2. *Suppose $P_1 = N_d(\mu_1, \sigma_1^2 I_d)$ and $P_2 = N_d(\mu_2, \sigma_2^2 I_d)$. Let Φ be the CDF of standard normal distribution and define for each $0 < \rho < 1$,*

$$C_0(\rho) = 2\Phi^{-1}(1 - \frac{\rho}{2}) \tag{6.32}$$

If $C = ||\mu_1 - \mu_2||/\max\{\sigma_1, \sigma_2\} \geq C_0(\rho)$, then there exists a set A such that $P_1(A) - P_2(A) \geq 1 - \rho$. Equality holds iff $\sigma_1 = \sigma_2$ and $C = C_0(\rho)$.

Proof of Lemma 6.2. Let p_1, p_2 be the density of P_1, P_2 and random variables X_1, X_2 follow P_1, P_2 respectively.

WLOG assume that $\sigma_1 \leq \sigma_2$. If the statement holds under this case, then when $\sigma_1 > \sigma_2$, one can select $\overline{A}$ such that $P_2(\overline{A}) - P_1(\overline{A}) > 1 - \rho$. Then take $A = \overline{A}^{\complement}$ and the desired result holds.

To start with, we consider the case $d = 1$ and assume $\mu_2 - \mu_1 = C\sigma_2 = C\eta\sigma_1 > 0$. When $\sigma_2 > \sigma_1$, let A be the set $\{x \in \mathbb{R} : x_{0,-} \leq x \leq x_{0,+}\}$ where

$$x_{0,\pm} = \frac{\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \pm \sqrt{\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 \sigma_2^2} + 2(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}) \log \frac{\sigma_2}{\sigma_1}}}{\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}} \tag{6.33}$$

$x_{0,\pm}$ are the two roots of the equation that the two normal densities equal, i.e.

$$\frac{1}{\sqrt{2\pi}\sigma_1} \exp\{-\frac{(x - \mu_1)^2}{2\sigma_1^2}\} = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\{-\frac{(x - \mu_2)^2}{2\sigma_2^2}\} \tag{6.34}$$

Denote $\eta = \sigma_2/\sigma_1 \geq 1$, when $\eta > 1$ it holds that

$$x_{1,\pm} = \frac{x_{0,\pm} - \mu_1}{\sigma_1} = \frac{-C\eta \pm \eta\sqrt{C^2\eta^2 + 2(\eta^2 - 1)\log\eta}}{\eta^2 - 1} \tag{6.35}$$

$$x_{2,\pm} = \frac{x_{0,\pm} - \mu_2}{\sigma_2} = \frac{-C\eta^2 \pm \sqrt{C^2\eta^2 + 2(\eta^2 - 1)\log\eta}}{\eta^2 - 1} \tag{6.36}$$

Let φ be the density function of standard normal distribution then $\varphi(x_{2,+}) = \eta\varphi(x_{1,+}), \varphi(x_{2,-}) = \eta\varphi(x_{1,-})$ by the selection of $x_{0,\pm}$. Since the desired expression

$$P_1(A) - P_2(A) = Pr(x_{1,-} \leq \frac{x - \mu_1}{\sigma_1} \leq x_{1,+}) - Pr(x_{2,-} \leq \frac{x - \mu_2}{\sigma_2} \leq x_{2,+}) \tag{6.37}$$

$$= \Phi(x_{1,+}) - \Phi(x_{1,-}) - \Phi(x_{2,+}) + \Phi(x_{2,-}) \tag{6.38}$$

is implicitly a function of C, η, we denote it as $h_+(C,\eta)$. Note that since

$$x_{0,\pm} = \frac{\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)^2 - \frac{(\mu_1-\mu_2)^2}{\sigma_1^2\sigma_2^2} - 2(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2})\log\frac{\sigma_2}{\sigma_1}}{\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \mp \sqrt{\frac{(\mu_1-\mu_2)^2}{\sigma_1^2\sigma_2^2} + 2(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2})\log\frac{\sigma_2}{\sigma_1}}\right)} = \frac{\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} - 2\log\frac{\sigma_2}{\sigma_1}}{\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \mp \sqrt{\frac{(\mu_1-\mu_2)^2}{\sigma_1^2\sigma_2^2} + 2(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2})\log\frac{\sigma_2}{\sigma_1}}} \tag{6.39}$$

Consider the limit $\sigma_2 \to \sigma_1$ with σ_1 fixed and recall that $\mu_2 - \mu_1 = C\sigma_2 > 0$, it holds that $\lim_{\sigma_2\to\sigma_1} x_{0,-} = -\infty$ and $\lim_{\sigma_2\to\sigma_1} x_{0,+} = (\mu_1+\mu_2)/2$. And further

$$\lim_{\sigma_2\to\sigma_1^+} x_{1,+} = \frac{\mu_2 - \mu - 1}{2\sigma_1} = \frac{C}{2}, \quad \lim_{\sigma_2\to\sigma_1^+} x_{1,-} = -\infty \tag{6.40}$$

$$\lim_{\sigma_2\to\sigma_1^+} x_{2,+} = \frac{\mu_1 - \mu_2}{2\sigma_2} = -\frac{C}{2}, \quad \lim_{\sigma_2\to\sigma_1^+} x_{2,-} = -\infty \tag{6.41}$$

which implies, since Φ is continuous,

$$\lim_{\eta\to 1^+} h_+(C,\eta) = 1 - 2\Phi(-\frac{C}{2}) \tag{6.42}$$

Observe that $x_{1,\pm} - \eta x_{2,\pm} = C\eta$ or $x_{2,\pm} = x_{1,\pm}/\eta - C$. Then

$$\frac{\partial x_{1,\pm}}{\partial\eta} - \eta\frac{\partial x_{2,\pm}}{\partial\eta} = \frac{\partial x_{1,\pm}}{\partial\eta} - \eta\left[-\frac{1}{\eta^2}x_{1,\pm} + \frac{1}{\eta}\frac{\partial x_{1,\pm}}{\partial\eta}\right] = \frac{1}{\eta}x_{1,\pm}\,. \tag{6.43}$$

(6.34) shows that $\eta\varphi(x_{1,\pm}) = \varphi(x_{2,\pm})$ Then, by taking partial derivative with η, it holds that

$$\frac{\partial h_+(C,\eta)}{\partial\eta} = \varphi(x_{1,+})\frac{\partial x_{1,+}}{\partial\eta} - \varphi(x_{1,-})\frac{\partial x_{1,-}}{\partial\eta} - \varphi(x_{2,+})\frac{\partial x_{2,+}}{\partial\eta} + \varphi(x_{2,-})\frac{\partial x_{2,-}}{\partial\eta} \tag{6.44}$$

$$= \varphi(x_{1,+})\left[\frac{\partial x_{1,+}}{\partial\eta} - \eta\frac{\partial x_{2,+}}{\partial\eta}\right] - \varphi(x_{1,-})\left[\frac{\partial x_{1,-}}{\partial\eta} - \eta\frac{\partial x_{2,-}}{\partial\eta}\right] \tag{6.45}$$

$$= \frac{1}{\eta}(x_{1,+}\varphi(x_{1,+}) - x_{1,-}\varphi(x_{1,-})) \tag{6.46}$$

$$> 0\,. \tag{6.47}$$

where in the last inequality we use the observation that $x_{1,-} < 0 < x_{1,+}, |x_{1,-}| > |x_{1,+}|$. Therefore $h_+(C,\eta)$ is increasing with η when C is fixed. And when $C > C_0(\rho)$ it holds that $h_+(C,\eta) > 1 - 2\Phi(-C_0(\rho)/2) = 1 - \rho$.

When $\mu_1 - \mu_2 = C\sigma_2 > 0$ again define

$$x_{1,\pm} = \frac{x_{0,\pm} - \mu_1}{\sigma_1} = \frac{C\eta \pm \eta\sqrt{C^2\eta^2 + 2(\eta^2-1)\log\eta}}{\eta^2 - 1} \tag{6.48}$$

$$x_{2,\pm} = \frac{x_{0,\pm} - \mu_2}{\sigma_2} = \frac{C\eta^2 \pm \sqrt{C^2\eta^2 + 2(\eta^2-1)\log\eta}}{\eta^2 - 1}\,. \tag{6.49}$$

Consider $h_-(C,\eta) = P_1(A) - P_2(A) = \Phi(x_{1,+}) - \Phi(x_{1,-}) - \Phi(x_{2,+}) + \Phi(x_{2,-})$. First we observe that $\lim_{\eta\to 1^+} x_{1,+} = +\infty$ and $\lim_{\eta\to 1^+} x_{1,-} = -C/2$. This shows that

$$\lim_{\eta\to 1^+} h_-(C,\eta) = 1 - 2\Phi(-\frac{C}{2}) \tag{6.50}$$

still holds. Note that we still have $x_{2,\pm} = x_{1,\pm}/\eta + C$. Taking derivatives w.r.t. η with same argument we have

$$\frac{\partial h_-(C,\eta)}{\partial \eta} = \frac{1}{\eta}(x_{1,+}\varphi(x_{1,+}) - x_{1,-}\varphi(x_{1,-}))\ . \tag{6.51}$$

Since $x_{1,-} < 0 < x_{1,+}$, the partial derivative is still positive. Therefore the result holds with similar arguments. Combining the two cases we complete the proof for the case $d = 1$.

Finally when $d \geq 2$, we consider the set of x projected on to the one dimensional space spanned by $\mu_2 - \mu_1$.

When $\sigma_2 = \sigma_1$, the set $A = \{x \in \mathbb{R}^d : p_1(x) \geq p_2(x)\}$ is the half space

$$A = \left\{ x \in \mathbb{R}^d : \langle x, \frac{\mu_2 - \mu_1}{||\mu_2 - \mu_1||} \rangle \leq \tfrac{\mu_1+\mu_2}{2} \right\}$$

Therefore $P_1(A) - P_2(A) = 1 - 2\Phi(-\frac{C}{2}) \geq 1 - \rho$ since $C \geq C_0(\rho) = 2\Phi^{-1}(1-\rho/2)$. Equality holds iff $C = C_0(\rho)$.

When $\sigma_2/\sigma_1 > 1$, consider $\tilde{P}_1 = N(\tilde{\mu}_1, \sigma_1^2)$ with $\tilde{\mu}_1 = \langle \mu_1, \frac{\mu_2-\mu_1}{||\mu_2-\mu_1||} \rangle$ and $\tilde{P}_2 = N(\tilde{\mu}_2, \sigma_2^2)$ with $\tilde{\mu}_2 = \langle \mu_2, \frac{\mu_2-\mu_1}{||\mu_2-\mu_1||} \rangle$. Compute $\tilde{x}_{0,\pm}$ by equation (6.33) with $\tilde{P}_1, \tilde{P}_2$ and let set

$$A = \left\{ x \in \mathbb{R}^d : \tilde{x}_{0,-} \leq \langle x, \frac{\mu_2 - \mu_1}{||\mu_2 - \mu_1||} \rangle \leq \tilde{x}_{0,+} \right\}\ , \quad \tilde{A} = \{x \in \mathbb{R} : \tilde{x}_{0,-} \leq x \leq \tilde{x}_{0,+}\}\ . \tag{6.52}$$

Note that $|\tilde{\mu}_2 - \tilde{\mu}_1| = ||\mu_2 - \mu_1|| = C\max\{\sigma_1, \sigma_2\}$ and thus

$$P_1(A) - P_2(A) = \tilde{P}_1(\tilde{A}) - \tilde{P}_2(\tilde{A}) > 1 - \rho\ . \tag{6.53}$$

□

Lemma 6.3. *Under the same setting of lemma 6.2, if $C = ||\mu_1 - \mu_2||/(\sigma_1 + \sigma_2) \geq C_0(\rho)/2$, then there exists a set A such that $P_1(A) - P_2(A) \geq 1 - \rho$. Equality holds iff $\sigma_1 = \sigma_2$ and $C = C_0(\rho)/2$.*

Proof of lemma 6.3. The proof is mostly unchnaged compared with lemma 6.2. We still consider the case $d = 1$ and assume $\sigma_1 \leq \sigma_2$ first. Once this is established, the remaining can be obtained from similar argument as in the proof of lemma 6.2. Let $x_{0,\pm}$ be the same as in proof of lemma 6.2.

When $\mu_2 - \mu_1 = C(\sigma_1 + \sigma_2) > 0$, now we should define

$$x_{1,\pm} = \frac{x_{0,\pm} - \mu_1}{\sigma_1} = \frac{-C(\eta+1) \pm \eta\sqrt{C^2(\eta+1)^2 + 2(\eta^2-1)\log\eta}}{\eta^2 - 1} \tag{6.54}$$

$$x_{2,\pm} = \frac{x_{0,\pm} - \mu_2}{\sigma_2} = \frac{-C\eta(\eta+1) \pm \sqrt{C^2(\eta+1)^2 + 2(\eta^2-1)\log\eta}}{\eta^2 - 1} , \tag{6.55}$$

Similarly to the proof of Lemma 6.2, it holds that

$$h_+(C,\eta) = \Phi(x_{1,+}) - \Phi(x_{1,-}) - \Phi(x_{2,+}) + \Phi(x_{2,-}) , \tag{6.56}$$

and

$$\lim_{\eta\to 1^+} h_+(C,\eta) = 1 - 2\Phi(-C) . \tag{6.57}$$

Since $x_{1,\pm} - \eta x_{2,\pm} = C(\eta+1)$ or $x_{2,\pm} = x_{1,\pm}/\eta - C(1+1/\eta)$, then

$$\frac{\partial x_{1,\pm}}{\partial\eta} - \eta\frac{\partial x_{2,\pm}}{\partial\eta} = \frac{\partial x_{1,\pm}}{\partial\eta} - \eta\left[-\frac{1}{\eta^2}x_{1,\pm} + \frac{1}{\eta}\frac{\partial x_{1,\pm}}{\partial\eta} + \frac{C}{\eta^2}\right] = \frac{1}{\eta}x_{1,\pm} - \frac{C}{\eta} . \tag{6.58}$$

By the same computation as in equation (6.47) we have

$$\frac{\partial h_+(C,\eta)}{\partial\eta} = \frac{1}{\eta}\left((x_{1,+} - C)\varphi(x_{1,+}) - (x_{1,-} - C)\varphi(x_{1,-})\right) . \tag{6.59}$$

Note that

$$x_{1,+} - C = \frac{-C(\eta+1) - C(\eta^2-1) + \eta\sqrt{C^2(\eta+1)^2 + 2(\eta^2-1)\log\eta}}{\eta^2-1} \tag{6.60}$$

$$= \frac{-C\eta(\eta+1)}{\eta-1} + \frac{\eta\sqrt{C^2(\eta+1)^2 + 2(\eta^2-1)\log\eta}}{\eta^2-1} > 0 \tag{6.61}$$

$$x_{1,-} - C = \frac{-C\eta(\eta+1)}{\eta-1} - \frac{\eta\sqrt{C^2(\eta+1)^2 + 2(\eta^2-1)\log\eta}}{\eta^2-1} < 0 \tag{6.62}$$

Hence the derivative in (6.59) is positive. By the same argument, it holds that $h_+(C,\eta) \geq 1 - 2\Phi(-C) > 1 - \rho$.

A similar argument can be applied to the case $\mu_1 > \mu_2$ and is omitted. For the equality case, when $\sigma_1 = \sigma_2$ and $C = C_0(\rho)/2$, it is the same as in the proof in Lemma 6.2. □

Lemma 6.4. *Suppose* $P_{1,2} = N_d(\mu_{1,2}, \sigma_{1,2}^2 I_d), 0 < \rho < 1$. *Let* F_d *be the CDF of Gamma*$(\frac{d}{2}, \frac{d}{2})$, *and* $\eta_0(\rho)$ *is the solution of*

$$1 - \rho = F_d\left(\frac{2\eta^2 \log \eta}{\eta^2 - 1}\right) - F_d\left(\frac{2 \log \eta}{\eta^2 - 1}\right) . \tag{6.63}$$

If $\eta := \max\{\sigma_1/\sigma_2, \sigma_2/\sigma_1\} > \eta_0(\rho)$, *then there exists a set* A *such that* $P_1(A) - P_2(A) > 1 - \rho$.

To prove Lemma 6.4, we need to invoke a generalization of Reynolds' transportation theorem. The following lemma and required definitions can be found as equation 7.2 in Flanders [1973]

Lemma 6.5. *Let* Ω_t *be an* $\ell-$*dimensional time-variant domain of integration in* $\mathbb{R}^d$, *and can be given by the image of a smooth map* $(u, t) \to x(u, t)$, *where* u *runs over a fixed domain in a* $\mathbb{R}^\ell$. *Let* ω *be an exterior* $\ell-$*form that can be represented in local coordinates as*

$$\omega = \sum_H a_H(x, t) dx^H, \quad dx^H = dx^{h_1} \wedge \cdots \wedge dx^{h_\ell} . \tag{6.64}$$

where H *is the multi-index and* $1 \le h_1 < h_2 < \cdots < h_\ell \le n$. *Then*

$$\frac{d}{dt} \int_{\Omega_t} \omega = \int_{\Omega_t} i_{\boldsymbol{v}}(d_x \omega) + \int_{\Omega_t} \dot{\omega} + \int_{\partial \Omega_t} i_v \omega \tag{6.65}$$

where $v = \partial x/\partial t$, i_v *denotes the interior product with* v, $d_x\omega$ *is the exterior derivative of* ω *with respect to* x *and* $\dot{\omega} = \sum_H \frac{\partial a_H(x,t)}{\partial t} dx^H$

Proof of lemma 6.4. The proof is with the similar but more complicated technique as in the proof of lemma 6.2. WLOG assume that $\sigma_2 = \eta\sigma_1 \ge \eta_0(\rho)\sigma_1 > \sigma_1$ and

$$\mu_1 = (\mu_{11}, \cdots, \mu_{1d}), \quad \mu_2 = (C\sigma_2 + \mu_{21}, \cdots, \mu_{2d}). \tag{6.66}$$

By symmetricity, we can further assume $C \ge 0, \mu_{22} = \mu_{12}, \cdots, \mu_{2d} = \mu_{1d}$.

Let p_1, p_2 be the densities of P_1, P_2 respectively. Note that p_1 is just the density of standard $d-$dimensional Gaussian, which we will denote as φ_d. Then algebraic computation shows that the set

$$A = \{x \in \mathbb{R}^d : p_1(x) \ge p_2(x)\} = \left\{x \in \mathbb{R}^d : ||x - \frac{\eta^2\mu_1 - \mu_2}{\eta^2 - 1}|| \le \frac{\sigma_2}{\eta^2 - 1}\sqrt{C^2\eta^2 + 2d(\eta^2 - 1)\log\eta}\right\} . \tag{6.67}$$

Note that A is a ball in $d-$dimension. Write

$$A_1(C,\eta) := \left\{ x \in \mathbb{R}^d : ||x - (-\frac{C\eta}{\eta^2-1}, 0, \cdots, 0)|| \leq \frac{\eta}{\eta^2-1}\sqrt{C^2\eta^2 + 2d(\eta^2-1)\log\eta} \right\} , \tag{6.68}$$

we can compute the probability directly by

$$\begin{aligned} P_1(A) &= \int_{x:||x-\frac{\eta^2\mu_1-\mu_2}{\eta^2-1}|| \leq \frac{\sigma_2}{\eta^2-1}\sqrt{C^2\eta^2+2d(\eta^2-1)\log\eta}} \frac{1}{(2\pi)^{\frac{d}{2}}\sigma_1^d} \exp\left(-\frac{||x-\mu_1||^2}{2\sigma_1^2}\right) dx \\ &= \int_{y:||y+\frac{C\eta}{\eta^2-1}|| \leq \frac{\eta}{\eta^2-1}\sqrt{C^2\eta^2+2d(\eta^2-1)\log\eta}} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{||y||^2}{2}\right) dy \qquad (6.69) \\ &= \int_{A_1(C,\eta)} \varphi_d(x)dx , \qquad (6.70) \end{aligned}$$

where φ_d is the density of $N_d(0, I_d)$ distribution, and Similarly let $A_2(C,\eta) = \{x \in \mathbb{R}^d : ||x - (-C\eta^2/(\eta^2-1), 0, \cdots, 0)|| \leq \sqrt{C^2\eta^2 + 2d(\eta^2-1)\log\eta}/(\eta^2-1)\}$ we have

$$P_2(A) = \int_{A_2(C,\eta)} \varphi_d(x)dx . \tag{6.71}$$

Let $h(C,\eta) = P_1(A) - P_2(A)$, as the latter is only determined by C and η. We now prove that for any fixed $\eta \geq \eta_0(\rho)$, $h(C,\eta)$ is monotonically increasing in $C \geq 0$.

To show this, we consider taking partial derivative of $h(C,\eta)$ with respect to C.

$$\frac{\partial}{\partial C}h(C,\eta) = \frac{\partial}{\partial C}\int_{A_1(C,\eta)} \varphi_d(x)dx - \frac{\partial}{\partial C}\int_{A_2(C,\eta)} \varphi_d(x)dx , \tag{6.72}$$

We first tackle $\frac{\partial}{\partial C}\int_{A_1(C,\eta)} \varphi_d(x)dx$. $A_1(C,\eta)$ is a hyperball. Let n be the normal vector given in polar coordinates, then one can perform the following change of variable

$$\begin{aligned} x : (n, r) &\to \mathbb{R}^d \qquad (6.73) \\ x_1 &= o_1 + r\cos\theta_1 \qquad (6.74) \\ x_2 &= r\sin\theta_1\cos\theta_2 \qquad (6.75) \\ &\cdots \qquad (6.76) \\ x_{d-2} &= r\sin\theta_1\cdots\sin\theta_{d-2}\cos\theta_{d-1} \qquad (6.77) \\ x_d &= r\sin\theta_1\cdots\sin\theta_{d-2}\sin\theta_{d-1} \qquad (6.78) \end{aligned}$$

where $o_1 = -C\eta/(\eta^2-1)$ and $0 \le r \le R_1$ with

$$R_1 = \frac{\eta}{\eta^2-1}\sqrt{C^2\eta^2 + 2d(\eta^2-1)\log\eta}\,. \tag{6.79}$$

For each $C > 0$, there is a small neighborhood of $C \in (C_a, C_b)$ such that $x = (x_1, \cdots, x_d)$ is a continuously differentiable map on $[0, R_1] \times [0,\pi]^{d-2} \times [0, 2\pi] \times (C_a, C_b) \to \mathbb{R}^d$. Here is where lemma 6.5 kicks in. Consider the $d-$form $\omega = \varphi_d(x) dx_1 \wedge \cdots dx_d$. Then since ω is a $d-$form on $\mathbb{R}^d$, its exterior derivative w.r.t. $x \in \mathbb{R}^d$ is zero. And since φ_d is independent of C, $\dot{\omega} = 0$.

Therefore equation (6.65) is simplified to

$$\frac{\partial}{\partial C}\int_{A_1(C,\eta)} \omega = \int_{\partial A_1(C,\eta)} i_{v_1}(\omega). \tag{6.80}$$

where $v_1 = \partial x/\partial C$.

We shall now identify $i_{v_1}(\omega)$. As $\partial A_1(C,\eta)$ is a hypersphere with dimension $d-1$, it is orientable. Further, consider the following bases of tangent space of $\partial A_1(C,\eta)$

$$e_i = \frac{\partial}{\partial \theta_i}, \quad 1 \le i \le d-1 \tag{6.81}$$

and outward pointing unit normal $n = \partial/\partial r$. Then for any vector field $t = \sum_{1\le i\le d-1} t_i e_i + t_d n$, by the definition of interior product and the fact that ω is an alternating $d-$linear tensor, it holds that

$$\begin{aligned} i_t(\omega)(e_1, \cdots, e_{d-1}) =& \omega(t, e_1, \cdots, e_{d-1}) && (6.82)\\ =& \sum_{i=1}^{d-1} t_i \omega(e_i, e_1, \cdots, e_{d-1}) + t_d \omega(n, e_1, \cdots, e_{d-1}) && (6.83)\\ =& \varphi_d(x) t_d (dx_1 \wedge \cdots dx_d)(n, e_1, \cdots, e_{d-1}) && (6.84)\\ =& \varphi_d(x)\langle t, n\rangle d\Sigma(n)(e_1, \cdots, e_{d-1}) && (6.85) \end{aligned}$$

where we denote $d\Sigma(n) = i_n(dx_1 \wedge \cdots \wedge dx_n)$ to be the area element. In other words, $d\Sigma(n)$

is the differential form such that

$$d\Sigma(n)(e_1,\cdots,e_{d-1}) = (dx_1\wedge\cdots\wedge dx_d)(n,e_1,\cdots,e_{d-1}) \tag{6.86}$$

$$= R_1^{d-1}\sin^{d-2}\theta_1\cdots\sin\theta_{d-2}(dr\wedge d\theta_1\wedge\cdots\wedge d\theta_{d-1})(n,e_1,\cdots,e_{d-1}) \tag{6.87}$$

$$= R_1^{d-1}\sin^{d-2}\theta_1\cdots\sin\theta_{d-2}\ , \tag{6.88}$$

which shows that $d\Sigma(n) = R_1^{d-1}\sin^{d-2}\theta_1\cdots\sin\theta_{d-2}d\theta_1\wedge\cdots\wedge d\theta_{d-1}$. Then since $i_{v_1(\omega)} = \varphi_d(x)\langle v_1,n\rangle d\Sigma(n)$,

$$\frac{\partial}{\partial C}\int_{A_1(C,\eta)}\omega = \int_{\mathbb{S}^{d-1}} R_1^{d-1}\varphi_d(x(n,R_1))\langle v_1,n\rangle d\Sigma(n)\ . \tag{6.89}$$

Similarly $A_2(C,\eta)$ is given by

$$y:(n,r)\to\mathbb{R}^d \tag{6.90}$$

$$y_1 = o_2 + r\cos\theta_1 \tag{6.91}$$

$$y_2 = r\sin\theta_1\cos\theta_2 \tag{6.92}$$

$$\cdots \tag{6.93}$$

$$y_{d-1} = r\sin\theta_1\cdots\sin\theta_{d-2}\cos\theta_{d-1} \tag{6.94}$$

$$y_d = r\sin\theta_1\cdots\sin\theta_{d-2}\sin\theta_{d-1} \tag{6.95}$$

with $o_2 = -C\eta^2/(\eta^2-1)$ and $0\le r\le R_2$, where

$$R_2 = \frac{1}{\eta^2-1}\sqrt{C^2\eta^2 + 2d(\eta^2-1)\log\eta}\ . \tag{6.96}$$

By the same argument, we have

$$\frac{\partial}{\partial C}\int_{A_2(C,\eta)}\omega = \int_{\partial A_2(C,\eta)}\varphi_d(y(n,R_2))\langle v_2,n\rangle d\Sigma(n) = \int_{\mathbb{S}^{d-1}} R_2^{d-1}\varphi_d(y(n,R_2))\langle v_2,n\rangle d\Sigma(n)\ . \tag{6.97}$$

where $v_2 = \partial y/\partial C$.

We observe two identites. First, for the same

$$n = (\cos\theta_1,\sin\theta_1\cos\theta_2,\cdots,\sin\theta_1\cdots\sin\theta_{d-2}\cos\theta_{d-1},\sin\theta_1\cdots\sin\theta_{d-1})$$

it holds that

$$
\begin{aligned}
&\frac{\varphi_d(x(n,R_1))}{\varphi_d(y(n,R_2))} \\
=&\frac{\exp\left(-\frac{1}{2}||x(n,R_1)||^2\right)}{\exp\left(-\frac{1}{2}||y(n,R_2)||^2\right)} && (6.98)\\
=&\frac{\exp\left(-\frac{1}{2}\left[o_1^2+R_1^2+2o_1R_1\cos\theta_1\right]\right)}{\exp\left(-\frac{1}{2}\left[o_2^2+R_2^2+2o_2R_2\cos\theta_1\right]\right)} && (6.99)\\
=&\frac{\exp\left(-\frac{1}{2}\left[\frac{C^2\eta^2}{(\eta^2-1)^2}+\frac{\eta^2}{(\eta^2-1)^2}(C^2\eta^2+2d(\eta^2-1)\log\eta)-2\frac{C\eta^2}{(\eta^2-1)^2}\sqrt{C^2\eta^2+2d(\eta^2-1)\log\eta}\cos\theta_1\right]\right)}{\exp\left(-\frac{1}{2}\left[\frac{C^2\eta^4}{(\eta^2-1)^2}+\frac{1}{(\eta^2-1)^2}(C^2\eta^2+2d(\eta^2-1)\log\eta)-2\frac{C\eta^2}{(\eta^2-1)^2}\sqrt{C^2\eta^2+2d(\eta^2-1)\log\eta}\cos\theta_1\right]\right)} && (6.100)\\
=&\exp\left(-\frac{1}{2}2d\log\eta\right) && (6.101)\\
=&\frac{1}{\eta^d} && (6.102)
\end{aligned}
$$

Second, for same $\boldsymbol{n}$, it holds that $\boldsymbol{x}(\boldsymbol{n},R_1)-\eta\boldsymbol{y}(\boldsymbol{n},R_2)=(C\eta,0,\cdots,0)$. This shows $\boldsymbol{v}_1-\eta\boldsymbol{v}_2=(\eta,0,\cdots,0)$

Therefore

$$
\begin{aligned}
&\frac{\partial h(C,\eta)}{\partial C} && (6.103)\\
=&\int_{\mathbb{S}^{d-1}}\left(R_1^{d-1}\varphi_d(\boldsymbol{x}(\boldsymbol{n},R_1))\langle\boldsymbol{v}_1,\boldsymbol{n}\rangle-R_2^{d-1}\varphi_d(\boldsymbol{y}(\boldsymbol{n},R_1))\langle\boldsymbol{v}_2,\boldsymbol{n}\rangle\right)d\Sigma(\boldsymbol{n}) && (6.104)\\
=&\int_{\mathbb{S}^{d-1}}\left(R_1^{d-1}\varphi_d(\boldsymbol{x}(\boldsymbol{n},R_1))\langle\boldsymbol{v}_1,\boldsymbol{n}\rangle-\left(\frac{R_1}{\eta}\right)^{d-1}\eta^d\varphi_d(\boldsymbol{x}(\boldsymbol{n},R_1))\langle\boldsymbol{v}_2,\boldsymbol{n}\rangle\right)d\Sigma(\boldsymbol{n}) && (6.105)\\
=&\int_{\mathbb{S}^{d-1}}R_1^{d-1}\varphi_d(\boldsymbol{x}(\boldsymbol{n},R_1))\langle\boldsymbol{v}_1-\eta\boldsymbol{v}_2,\boldsymbol{n}\rangle d\Sigma(\boldsymbol{n}) && (6.106)\\
=&\int_{\mathbb{S}^{d-1}}R_1^{d-1}\eta\cos\theta_1\varphi_d(\boldsymbol{x}(\boldsymbol{n},R_1))d\Sigma(\boldsymbol{n}) && (6.107)\\
=&R_1^{d-1}\eta\int_0^\pi\cdots\int_0^\pi\int_0^{2\pi}\cos\theta_1\frac{1}{(2\pi)^{\frac{d}{2}}}\exp\left(-\frac{1}{2}[o_1^2+R_1^2+2o_1R_1\cos\theta_1]\right)\sin^{d-2}\theta_1\cdots\sin\theta_{d-2}d\theta_1d\theta_2\cdots d\theta_{d-1} && (6.108)\\
=&\Xi\int_0^\pi\cos\theta_1\exp(-o_1R_1\cos\theta_1)\sin^{d-2}\theta_1d\theta_1\ , && (6.109)
\end{aligned}
$$

where Ξ is a positive constant. Finally,

$$\int_0^{\pi} \cos\theta_1 \exp(-o_1 R_1 \cos\theta_1) \sin^{d-2}\theta_1 d\theta_1 \tag{6.110}$$

$$= \int_0^{\frac{\pi}{2}} \cos\theta_1 \exp(-o_1 R_1 \cos\theta_1) \sin^{d-2}\theta_1 d\theta_1 + \int_{\frac{\pi}{2}}^{\pi} \cos\theta_1 \exp(-o_1 R_1 \cos\theta_1) \sin^{d-2}\theta_1 d\theta_1 \tag{6.111}$$

$$= \int_0^{\frac{\pi}{2}} \cos\theta_1 \exp(-o_1 R_1 \cos\theta_1) \sin^{d-2}\theta_1 d\theta_1 - \int_0^{\frac{\pi}{2}} \cos\theta_1 \exp(o_1 R_1 \cos\theta_1) \sin^{d-2}\theta_1 d\theta_1 \tag{6.112}$$

$$= \int_0^{\frac{\pi}{2}} \cos\theta_1 \sin^{d-2}\theta_1 \left[\exp(-o_1 R_1 \cos\theta_1) - \exp(o_1 R_1 \cos\theta_1)\right] d\theta_1 > 0 \tag{6.113}$$

where the last inequality is due to the fact that $o_1 < 0$ and for any $x > 0, e^x > 1 > e^{-x}$. This shows that when $C > 0$, $h(C, \eta)$ is increasing in C. Therefore as long as $h(0, \eta) > 1 - \rho$, $P_1(A) - P_2(A) > 1 - \rho$. It remains to show that $h(0, \eta) \geq 1 - \rho$ as long as $eta \geq \eta_0$. This is direct, when $C = 0$,

$$A_1 = \left\{ x \in \mathbb{R}^d : ||x|| \leq \sqrt{\frac{2d\eta^2 \log\eta}{\eta^2 - 1}} \right\} \tag{6.114}$$

$$A_2 = \left\{ x \in \mathbb{R}^d : ||x|| \leq \sqrt{\frac{2d \log\eta}{\eta^2 - 1}} \right\} \tag{6.115}$$

Let $X \sim N_d(0, I_d)$, then $||X||^2/d \sim Gamma(\frac{d}{2}, \frac{d}{2})$. Then

$$h(0,\eta) = Pr\left(||X||^2 \leq \frac{2d\eta^2 \log\eta}{\eta^2 - 1} \right) - Pr\left(||X||^2 \leq \frac{2d \log\eta}{\eta^2 - 1} \right) = F_d(\frac{2\eta^2 \log\eta}{\eta^2 - 1}) - F_d(\frac{2\log\eta}{\eta^2 - 1}) \,. \tag{6.116}$$

By Lemma 6.18, $h(0, \eta)$ is increasing in η. Therefore $h(0, \eta) \geq h(0, \eta_0(\rho)) = 1 - \rho$, and this completes the proof. □

Lemma 6.6 (Gaussian tail bound)**.** *For any $C \geq 1/2$, it holds that*

$$\frac{1}{5\sqrt{2\pi}C} \exp\left(-\frac{1}{2}C^2\right) \leq 1 - \Phi(C) \leq \exp\left(-\frac{1}{2}C^2\right) \tag{6.117}$$

Proof of lemma 6.6. The upper bound is the standard Chernoff's bound. The moment generation function of a random variable $X \sim N(0, 1)$ is given by

$$\mathbb{E}\exp[tX] = \exp\left(\frac{1}{2}t^2\right) \tag{6.118}$$

Then by Markov's inequality,

$$1-\Phi(C)=P(X>C)\leq \inf_{t:t>0}\frac{\mathbb{E}e^{tX}}{e^{t}C}=\exp\left(-\frac{1}{2}t^{2}\right) \tag{6.119}$$

To show the lower bound, consider

$$g(C)=1-\Phi(C)-\frac{1}{\sqrt{2\pi}}\frac{C}{C^{2}+1}\exp(-\frac{1}{2}C^{2}). \tag{6.120}$$

Since $g'(C)=-2e^{-C^2/2}/(C^2+1)^2$, $g(C)$ is strictly decreasing. Since $\lim_{C\to\infty}g(C)=0$, we have $g(C)\geq 0$ for all positive C, i.e.,

$$1-\Phi(C)\geq\frac{1}{\sqrt{2\pi}}\frac{C}{C^{2}+1}\exp(-\frac{1}{2}C^{2})\ . \tag{6.121}$$

For any $C\geq 1/2$, it holds that $C/(C^2+1)>1/5C$, hence the lower bound in 6.6 holds. □

Lemma 6.7 (Estimate of $\Phi^{-1}(1-\rho/2)$)**.** *For any constant ρ such that $0<\rho\leq 1/2$, $c_\rho^{\max}$ defined as $c_\rho^{\max}=\Phi^{-1}(1-\rho/2)$ satisfies that*

$$\sqrt{2\log\frac{2}{\rho}-\log(10\sqrt{\pi\log\frac{2}{\rho}})}\leq c_\rho^{\max}\leq\sqrt{2\log\frac{2}{\rho}} \tag{6.122}$$

Proof of lemma 6.7. From the upper bound in lemma 6.6, it is true that

$$1-\Phi(\sqrt{2\log\frac{2}{\rho}})\leq\exp(-\frac{1}{2}2\log\frac{2}{\rho})=\frac{\rho}{2}=1-\Phi(c_\rho^{\max})\ , \tag{6.123}$$

and hence from the monotonically increasing of Φ function, $c_\rho^{\max}\leq\sqrt{2\log\frac{2}{\rho}}$.

On the other hand, since $\rho\leq 1/2$ and $c_\rho^{\max}=\Phi^{-1}(1-\rho/2)\geq\Phi^{-1}(3/4)\approx 0.67>1/2$, hence

$$\frac{\rho}{2}=1-\Phi(c_\rho^{\max})\geq\frac{1}{5\sqrt{2\pi}c_\rho^{\max}}\exp\left(-\frac{1}{2}(c_\rho^{\max})^{2}\right)\geq\frac{1}{5\sqrt{4\pi\log\frac{2}{\rho}}}\exp\left(-\frac{1}{2}(c_\rho^{\max})^{2}\right)\ , \tag{6.124}$$

where we use the estimate $c_\rho^{\max}\leq\sqrt{2\log(2/\rho)}$. After rearrangement we have

$$c_\rho^{\max}\geq\sqrt{2\log\frac{2}{\rho}-\log(10\sqrt{\pi\log\frac{2}{\rho}})}\ . \tag{6.125}$$

Hence the proof is complete. □

Lemma 6.8 (Lower bound of TV of two Gaussians with same mean)**.** *Suppose* $P = \mathcal{N}_d(0, I_d)$ *and* $P' = \mathcal{N}_d(0, \eta^2 I_d)$ *where* $\eta > e^{1/2}$*, then*

$$TV(P, P') \geq 1 - \frac{\exp(d\sqrt{\log \eta})}{\eta^d} - \left(\frac{de}{\pi(e-1)}\right)^{\frac{d}{2}} \frac{(\log \eta)^{\frac{d}{2}}}{\eta^d} \tag{6.126}$$

Lemma 6.9 (Chi-square tail bound, Lemma 1 in [Laurent and Massart, 2000])**.** *Let* X *be a random variable following chi-square distribution with degree of freedom* d*, then*

$$P(X - d \geq x) \leq \exp\left(-\frac{x^2}{2(d+x) + 2\sqrt{d^2 + 2dx}}\right) \tag{6.127}$$

Proof of lemma 6.8. As in the proof of lemma 6.18, consider the set $B(\eta)$ defined by

$$B(\eta) := \{x : p(x) \geq p'(x)\} = \left\{x \in \mathbb{R}^d : ||x|| \leq \sqrt{\frac{2d\eta^2 \log \eta}{\eta^2 - 1}}\right\} . \tag{6.128}$$

Then $TV(P, P') = P(B(\eta)) - P'(\eta)$. Note that by a change-of-variable argument, the probability of P' on this set can be upper bounded by

$$\begin{aligned}
P'(B_\eta) &= \int_{\{x:||x|| \leq \sqrt{\frac{2d\eta^2 \log \eta}{\eta^2-1}}\}} \frac{1}{(2\pi\eta^2)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\eta^2}||x||^2\right) dx \\
&= \int_{\{x:||x|| \leq \sqrt{\frac{2d \log \eta}{\eta^2-1}}\}} \varphi_d(x) dx \\
&\leq \varphi_d(0) \left(\frac{2d \log \eta}{\eta^2 - 1}\right)^{\frac{d}{2}} \\
&= \left(\frac{1}{2\pi} \frac{2d \log \eta}{\eta^2 - 1}\right)^{\frac{d}{2}} \qquad (6.129) \\
&\leq \left(\frac{de}{\pi(e-1)}\right)^{\frac{d}{2}} \frac{(\log \eta)^{\frac{d}{2}}}{\eta^d} \qquad (6.130)
\end{aligned}$$

On the other hand, when $\eta > e^{1/2}$, we have $2d\eta^2 \log \eta/(\eta^2 - 1) \geq d \log \eta^2 \geq d$. Therefore,

by lemma 6.9, it holds that

$$\begin{aligned} P(B_\eta) &\geq P\left(\left\{x \in \mathbb{R}^d : ||x|| \leq d\log\eta^2\right\}\right) \\ &\geq 1 - \exp\left(-\frac{d^2(\log\eta^2 - 1)^2}{2(d\log\eta^2) + 2\sqrt{d^2(2\log\eta^2 - 1)}}\right) \\ &= 1 - \exp\left(-\frac{d}{2}\left[2\log\eta - \sqrt{4\log\eta - 1}\right]\right) \qquad (6.131) \\ &= 1 - \frac{1}{\eta^d}\exp\left(\frac{d}{2}\sqrt{4\log\eta - 1}\right) \qquad (6.132) \\ &\geq 1 - \frac{\exp(d\sqrt{\log\eta})}{\eta^d} \qquad (6.133) \end{aligned}$$

Therefore the conclusion of lemma 6.8 holds. □

Lemma 6.10 (Estimate of $\eta_\rho^{\max}$)**.** *For any $\rho \in (0, 1/4)$, it holds that*

$$(\frac{1}{2\rho})^{1/d} \leq \eta_\rho^{\max} \leq \max\left\{e^{1/2}, \min\left\{\frac{2}{\rho^{\frac{2}{d}}}, \frac{\exp\left(\sqrt{\log 2 + \frac{2}{d}\log\frac{1}{\rho}}\right)}{\rho^{\frac{1}{d}}}(1 + \sqrt{\frac{d}{\pi e(e-1)}})\right\}\right\} \qquad (6.134)$$

Proof of lemma 6.10. We first show a courser upper bound of $\eta_\rho^{\max}$ Consider two distributions $P = \mathcal{N}_d(0, I_d)$ and $P' = (0, (\eta_\rho^{\max})^2)$ and denote their densities as p, p' respectively. By definition of $\eta_\rho^{\max}$, the total variation distance between P, P' is $1-\rho$, which can be lower bounded by squared Hellinger distance , which in this case has a analytical form as

$$H^2(P, P') = \frac{1}{2}\int_{\mathbb{R}^d}(\sqrt{p} - \sqrt{p'})^2 \mathrm{d}x = 1 - \left(\frac{2\eta_\rho^{\max}}{(\eta_\rho^{\max})^2 + 1}\right)^{\frac{d}{2}} \leq 1 - \rho\,. \qquad (6.135)$$

Since $\eta_\rho^{\max} > 1$ we have the following inequalities

$$\rho \leq \left(\frac{2\eta_\rho^{\max}}{(\eta_\rho^{\max})^2 + 1}\right)^{\frac{d}{2}} < \left(\frac{2}{\eta_\rho^{\max}}\right)^{\frac{d}{2}}\,, \qquad (6.136)$$

which gives us the upper bound that $\eta_\rho^{\max} < 2/\rho^{2/d}$,i.e.$d\log\eta_\rho^{\max} < d\log 2 + 2\log(1/\rho)$.

Since, from lemma 6.8, we conclude that

$$\begin{aligned} 1 - \rho &\geq 1 - \frac{\exp(d\sqrt{\log\eta_\rho^{\max}})}{(\eta_\rho^{\max})^d} - \left(\frac{de}{\pi(e-1)}\right)^{\frac{d}{2}}\frac{(\log\eta_\rho^{\max})^{\frac{d}{2}}}{(\eta_\rho^{\max})^d} \\ &> 1 - \frac{\exp(d\sqrt{\log\eta_\rho^{\max}})}{(\eta_\rho^{\max})^d}\left[1 + \left(\sqrt{\frac{e}{\pi(e-1)}}\frac{\sqrt{d\log\eta_\rho^{\max}}}{\exp(\frac{\sqrt{d}}{d}\sqrt{d\log\eta_\rho^{\max}})}\right)^d\right] \qquad (6.137) \end{aligned}$$

Consider a utility function $f(x) = \exp(x/\sqrt{d})/\sqrt{e/(\pi(e-1))}x$. Then its derivative

$$f'(x) = \frac{\pi(e-1)}{ex^2}[\sqrt{\frac{e}{\pi(e-1)}}\exp(\frac{x}{\sqrt{d}})(\frac{x}{\sqrt{d}} - 1) \tag{6.138}$$

Since $f'(x) < 0$ when $x < \sqrt{d}$ and $f'(x) > 0$ when $x > \sqrt{d}$, $f(x)$ is decreasing on $(0, \sqrt{d}]$ and increasing on $(\sqrt{d}, +\infty)$ and hence it has a minimal value at $f(\sqrt{d}) = \sqrt{\pi e(e-1)/d}$. Therefore, we can further lower bound after equation 6.137 that

$$\begin{aligned} 1 - \rho &> 1 - \frac{\exp(d\sqrt{\log \eta_\rho^{\max}})}{(\eta_\rho^{\max})^d}\left[1 + (\sqrt{d/\pi e(e-1)})^d\right] \\ &> 1 - \frac{\exp(d\sqrt{\log \eta_\rho^{\max}})}{(\eta_\rho^{\max})^d}\left[1 + \sqrt{d/\pi e(e-1)}\right]^d \end{aligned} \tag{6.139}$$

This further shows that

$$\eta_\rho^{\max} < \frac{1}{\rho^{\frac{1}{d}}}\left(1 + \sqrt{\frac{d}{\pi e(e-1)}}\right)\exp\left(\sqrt{\log \eta_\rho^{\max}}\right) < \frac{\exp\left(\sqrt{\log 2 + \frac{2}{d}\log\frac{1}{\rho}}\right)}{\rho^{\frac{1}{d}}}(1+\sqrt{\frac{d}{\pi e(e-1)}}) \tag{6.140}$$

For the lower bound, we consider upper bound the total variation distance by using the following inequality in

$$(1-\rho)^2 = TV(P, P')^2 \leq 1 - \exp(-D_{KL}(P \parallel P'))\,, \tag{6.141}$$

or after rearrangement,

$$\exp(-D_{KL}(P \parallel P') \leq 1 - (1-\rho)^2 = 2\rho - \rho^2 < 2\rho \tag{6.142}$$

where the $D_{KL}(P \parallel P')$ is KL divergence defined as

$$D_{KL}(P \parallel P') = \int_{\mathbb{R}^d} p \log \frac{p}{q} \mathrm{d}x = \frac{d}{2}\left[\log(\eta_\rho^{\max})^2 - 1 + \frac{1}{(\eta_\rho^{\max})^2}\right] \tag{6.143}$$

Combining the previous two inequalities, we have

$$\frac{d}{2}\log\frac{1}{(\eta_\rho^{\max})^2} + \frac{d}{2} - \frac{d}{2(\eta_\rho^{\max})^2} \leq \log 2\rho \tag{6.144}$$

Since $\eta_\rho^{\max} \geq 1$, the left hand side of the previous inequality can be further lower bounded by $d\log(1/\eta_\rho^{\max})$ and leads to $\eta_\rho^{\max} > (1/2\rho)^{1/d}$. □

Component-wise comparison theorem

The second key ingredient for proving our result is a group of component-wise comparison lemmas on two mixture distributions with small total variation distance. In the following lemmas, consider two mixture distributions $P = \sum_{i=1}^{K} w_i P_i$ and $P' = \sum_{j=1}^{K'} w'_j P_j$ and denote $w_{\min} = \min_{i,j}\{w_i, w'_j\}$ and $w_{\max} = \max_{i,j}\{w_i, w'_j\}$. The proofs can be found in Section 6.7; these lemmas essentially show that (1) the weighted sum of component-wise total variation difference is upper bounded if the two mixtures are close and (2) when two well-separated mixtures have same number of components, under particular correspondence conditions, their mixture proportions are close.

Lemma 6.11. *Suppose $TV(P, P') \leq 2\epsilon$. For any $j \in [K']$, if there exists a collection of sets $\{A_{ij}\}_{i\in[K]}$ such that $\rho_{ij} = 1 - P_i(A_{ij}) + P'_j(A_{ij}) \in [0, 1]$ for all $i \in [K]$, then*

$$\sum_{i=1}^{K} \max\{w_i, w'_j\}\rho_{ij} \geq w'_j - 2\epsilon \ , \quad \sum_{i=1}^{K} w_i \rho_{ij} \geq w_{\min} - 2\epsilon \ . \tag{6.145}$$

Similarly, if there exists a collection of sets $\{B_{ij}\}_{i\in[K]}$ such that $\rho_{ij} = 1 - P'_j(B_{ij}) + P_i(B_{ij}) \in [0, 1]$ for all $i \in [K]$, equation also holds.

This lemma stems from the fact that there cannot be a set that has a large probability mass of a component in P_i of P but has small probability mass from every component P'_j of P'. A_{ij} is the set that has a large probability mass under P_i but small under P'_j, while B_{ij} vice versa. A symmetric result can be obtained by switching the role of P and P'.

Lemma 6.11 is particularly useful when, for some component P'_j, all components P_i, $i \neq j$ are far away from P'_j in total variation distance; then by Lemma 6.11 one can upper bound the total variation distance between P'_j and the remaining component P_j of P.

Lemma 6.12. *Suppose $TV(P, P') \leq 2\epsilon$ and denote $w_{\min} = \min_{i,j}\{w_i, w'_j\}$. Consider two components P_{j_0}, P'_j with $w_{j_0} \leq w'_j$. If for all $i \in [K], i \neq j_0$, there is a set A_{ij} such that $P_i(A_{ij}) - P'_j(A_{ij}) \geq 1 - \rho$ with some constant ρ satisfying $0 \leq \rho \leq (w_{\min} - 2\epsilon)/(1 - w_{\min})$, it holds that*

$$TV(P_{j_0}, P'_j) \leq \frac{2\epsilon}{w_{\min}} + \frac{1 - w_{\min}}{w_{\min}}\rho \ . \tag{6.146}$$

Now, we obtain a component-wise comparison result for the differences in mixture proportions between corresponding components.

Lemma 6.13. *Suppose* $TV(P,P') \leq 2\epsilon$. *Denote* $w_{\min} = \min_{i,j}\{w_i, w'_j\}$ *and* $w_{\max} = \max_{i,j}\{w_i, w'_j\}$ *If there is a pair of components* P_i, P'_i *and a set* A *such that for some* $\rho \in (0,1)$, $\min\{P_i(A), P'_i(A)\} \geq 1-\rho$ *and* $\max_{j\neq i}\{P_j(A), P'_j(A)\} \leq \rho$, *then*

$$|w_i - w'_i| \leq 2\epsilon + (1 + w_{\max} - w_{\min})\rho \tag{6.147}$$

Proof of Lemma 6.11. Let $Q = (P+P')/2$. Then $TV(P,Q) \leq \epsilon$, $TV(P',Q) \leq \epsilon$. Let P'_j be an arbitrary component of P'. We will first show by contradiction the first half of (6.145) holds. Suppose that $\sum_{i=1}^K \max\{w_i, w'_j\}\rho_{ij} < w'_j - 2\epsilon$. Now consider the set $A = \cap_{i=1}^K A_{ij}^{\complement}$. On one hand,

$$Q(A) \geq w'_j P'_j((\cup_{i=1}^K A_{ij})^{\complement}) - \epsilon \geq w'_j(1 - \sum_{i=1}^K P'_j(A_{ij})) - \epsilon = w'_j - w'_j \sum_{i=1}^K P'_j(A_{ij}) - \epsilon := LB \tag{6.148}$$

Meanwhile,

$$Q(A) \leq \sum_{i=1}^K w_i P_i(A_i^{\complement}) + \epsilon = 1 - \sum_{i=1}^K w_i P_i(A_i) + \epsilon := UB \tag{6.149}$$

Since for every $i \in [K]$ it holds that $P_i(A_{ij}) - P'_j(A_{ij}) = 1 - \rho_{ij}$, then by Lemma 6.17, $w_i P_i(A_{ij}) - w'_j P'_j(A_{ij}) \geq w_i - \max\{w_i, w'_j\}\rho_{ij}$. Therefore, applying Lemma 6.17 to all components $i \in [K]$, it holds that

$$LB - UB = w'_j - 1 - 2\epsilon + \sum_{i=1}^K \big(w_i P_i(A_{ij}) - w'_j P'_j(A_{ij})\big) \tag{6.150}$$

$$\geq w'_j - 1 - 2\epsilon + \sum_{i=1}^K \big(w_i - \max\{w_i, w'_j\}\rho_{ij}\big) \tag{6.151}$$

$$= w'_j - 1 - 2\epsilon + 1 - \sum_{i=1}^K \max\{w_i, w'_j\}\rho_{ij} > 0\,. \tag{6.152}$$

This is a contradiction since $Q(A) \geq LB > UB \geq Q(A)$. So the first half of (6.145) holds for P'_j.

For the second half of (6.145), if $\sum_{i=1}^{K} w_i\rho_{ij} < w_{\min} - 2\epsilon$, then consider a different lower bound of $Q(A)$

$$Q(A) \geq w'_j P'_j((\cup_{i=1}^{K} A_{ij})^{\complement}) - \epsilon \geq w_{\min} P'_j((\cup_{i=1}^{K} A_{ij})^{\complement}) - \epsilon \geq w_{\min} - w_{\min} \sum_{i=1}^{K} P'_j(A_{ij}) - \epsilon := LB';, \tag{6.153}$$

then notice that

$$LB' - UB \geq w_{\min} - 1 - 2\epsilon + \sum_{i=1}^{K} w_i P_i(A_{ij}) - w_{\min} P'_j(A_{ij}) \tag{6.154}$$

$$\geq w_{\min} - 1 - 2\epsilon + \sum_{i=1}^{K} w_i(1 - \rho_{ij}) \tag{6.155}$$

$$= w_{\min} - 1 - 2\epsilon + 1 - \sum_{i=1}^{K} w_i\rho_{ij} > 0\ , \tag{6.156}$$

which is a contradiction and completes the proof of the second half of (6.145). □

Proof of Lemma 6.12. Define

$$\rho_1 = \frac{w_{\min} - 2\epsilon}{w_{\min}} - \frac{1 - w_{\min}}{w_{\min}}\rho \tag{6.157}$$

Note that $0 \leq \rho \leq (w_{\min} - 2\epsilon)/(1 - w_{\min})$ implies that $0 \leq \rho_1 \leq 1$. We will show by contradiction that the total variation distance $TV(P_{j_0}, P'_j) \leq 1 - \rho_1$. If otherwise $TV(P_{j_0}, P'_j) > 1 - \rho_1$, one can select a set $A_{j_0 j}$ such that $P_{j_0}(A_{j_0 j}) - P'_j(A_{j_0 j}) > 1 - \rho_1$. Define function

$$f(x) = x(\rho_1 - 1) + \sum_{\substack{i \in [K] \\ i \neq j_0}} \max\{w_i, x\}\rho\ , \tag{6.158}$$

then $f(x)$ is a piecewise linear function, with its slope upper bounded by

$$\rho_1 - 1 + \sum_{i: i \neq j} \rho = \rho_1 + (K-1)\rho - 1 \leq \rho_1 + \left(\frac{1 - w_{\min}}{w_{\min}}\right)\rho - 1 = \frac{w_{\min} - 2\epsilon}{w_{\min}} - 1 < 0\ . \tag{6.159}$$

Hence $f(x)$ is decreasing in x. Since $w_{j_0} \leq w'_j$, according to the remark after Lemma 6.11,

we have

$$f(w'_j) = w'_j\rho_1 + \sum_{\substack{i\in[K]\\ i\neq j_0}} \max\{w_i, w'_j\}\rho - w'_j \tag{6.160}$$

$$> \sum_{i=1}^{K} \max\{w_i, w'_j\}(1 - P_i(A_{ij}) + P'_j(A_{ij})) - w'_j \tag{6.161}$$

$$\geq -2\epsilon\ . \tag{6.162}$$

Then we conclude that $f(w_{\min}) \geq f(w'_j) > -2\epsilon$. But we can explicitly compute that

$$f(w_{\min}) = w_{\min}\left(\frac{w_{\min} - 2\epsilon}{w_{\min}} - \frac{1 - w_{\min}}{w_{\min}}\rho - 1\right) + \sum_{\substack{i\in[K]\\ i\neq j_0}} \max\{w_i, w_{\min}\}\rho \tag{6.163}$$

$$= -2\epsilon - (1 - w_{\min})\rho + \rho \sum_{\substack{i\in[K]\\ i\neq j_0}} w_i \tag{6.164}$$

$$\leq -2\epsilon - (1 - w_{\min})\rho + (1 - w_{\min})\rho \tag{6.165}$$

$$= -2\epsilon\ . \tag{6.166}$$

This contradiction completes the proof of (6.146). □

Proof. Proof of Lemma 6.13] Note that $w_i(A^{\complement}) \leq \rho$ and $w'_i(A^{\complement}) \leq \rho$

$$|w_i - w'_i| = |w_iP_i(A) - w'_iP'_i(A) + w_iP_i(A^{\complement}) - w'_iP'_i(A^{\complement})| \tag{6.167}$$

$$\leq |P(A) - \sum_{j\neq i} w_jP_j(A) - P'(A) + \sum_{j\neq i} w'_jP'_j(A)| + |w_iP_i(A^{\complement}) - w'_iP'_i(A^{\complement})| \tag{6.168}$$

$$\leq 2\epsilon + \max\{\sum_{j\neq i} w_jP_j(A), \sum_{j\neq i} w'_jP'_j(A)\} + \max\{w_iP_i(A^{\complement}), w'_iP'_i(A^{\complement})\} \tag{6.169}$$

$$\leq 2\epsilon + (1 - w_{\min} + w_{\max})\rho\ . \tag{6.170}$$

□

6.7.3 Initialization

Let c_0 be defined as in (6.6), then the following theorem guarantees that for each component of P', there exists a component of P that is close to it in total variation distance. Starting from here and in the remaining of Section 6.7, we will assume the following:

B1 $P \in \mathcal{M}(K, w_{\min}, w_{\max}, c)$ and $P' \in \mathcal{M}(K', w_{\min}, w_{\max}, c)$.

B2 There exists a distribution Q such that $TV(P, Q \leq \epsilon$ and $TV(P', Q) \leq \epsilon$.

B3 $\max\{K, K'\} \leq 1/w_{\min}$, $w_{\max} \leq 1 - (\min\{K, K'\} - 1)w_{\min}$.

Theorem 6.4. *Suppose B1-B3 hold. For any component P'_j, there must be a component P_i such that $TV(P_i, P'_j) \leq 1 - w_{\min} + 2\epsilon$*

Proof of Theorem 6.4. WLOG we prove the statement for P'_1. Assume that the conclusion of the theorem does not hold, that for any component P_i, it holds that $TV(P_i, P'_1) > 1 - w_{\min} + 2\epsilon$, implying the existence of a set A_i such that $P_i(A_i) - P'_1(A_i) > 1 - w_{\min} + 2\epsilon$, or $w_{\min} - 2\epsilon > 1 - P_i(A_i) + P'_1(A_i)$. Then applying Lemma 6.11 to the collection of sets $\{A_i\}_{i\in[K]}$, we have

$$w_{\min} - 2\epsilon > \sum_{i=1}^{K} w_i(1 - P_i(A_i) + P_1(A'_i)) \geq w_{\min} - 2\epsilon \; . \tag{6.171}$$

Therefore, the desired result holds. □

The second theorem states that, if both P and P' are well-separated and have nearly balanced components, then the matching established in theorem 6.4 is one-to-one. Specifically, for any component in P', there is one and only one component in P such that their centers are close.

Theorem 6.5. *Suppose B1-B3 hold, If $c > \eta_0 c_0$, where c_0, η_0 are defined in* (6.6) *and* (6.7) *respectively, then $K = K'$ and there is a permutation $\theta : [K] \to [K]$ such that $||\mu_i - \mu'_{\theta(i)}|| \leq c_0 \max\{\sigma_i, \sigma'_{\theta(i)}\}$ and $\max\{\sigma_i/\sigma'_{\theta(i)}, \sigma'_{\theta(i)}/\sigma_i\} \leq \eta_0$.*

Proof of Theorem 6.5. According to Theorem 6.4 and Lemma 6.2 each P'_j must be matched to at least one component P_i in the sense that their center distance $||\mu_i - \mu'_j|| \leq c_0 \max\{\sigma_i, \sigma'_j\}$. Note that Theorem 6.4 is symmetric on both P and P'. For each P_i, there must be a P'_j such that their centers are close.

We now considering the following removing procedure. For a component P_j, if there is a unique P'_j that is matched to P_j, and P_j is the only component in P that is matched to

P'_j, then remove both P_j and P'_j. For a removed pair P_j, P'_j, one can upper bound the ratio between their standard deviations by $\max\{\sigma_j, \sigma'_j\} \leq \eta_0$. To see this WLOG we assume that $\sigma_j \leq \sigma'_j$. Then for any $i \neq j$, let P'_i be a component matched with P_i. The distances between centers are

$$||\mu'_i - \mu_j|| \geq ||\mu'_i - \mu'_j|| - ||\mu'_j - \mu_j|| \geq c(\sigma'_i + \sigma'_j) - c_0\sigma'_j = c\sigma'_i + (c - c_0)\sigma'_j \,. \tag{6.172}$$

- If $\sigma'_i \geq \sigma_j$, it holds that $||\mu'_i - \mu_j|| \geq c \max\{\sigma'_i, \sigma_j\}$. By Lemma 6.2, one can select a set A_{ij} such that $P'_i(A_{ij}) - P_j(A_{ij}) > 1 - 2\Phi(-c/2) \geq 1 - 2\Phi(-c_0\eta_0/2)$

- If $\sigma'_i < \sigma_j$, then since P'_i is not matched with P_j

$$||\mu'_i - \mu_j|| \geq c_0 \max\{\sigma'_i, \sigma_j\} = c_0\sigma_j \,. \tag{6.173}$$

  Adding (6.172) and (6.173), we have

$$||\mu'_i - \mu_j|| \geq \frac{1}{2}\left(c\sigma'_i + (c - c_0)\sigma'_j + c_0\sigma_j\right) \tag{6.174}$$

$$\geq \frac{1}{2}c(\sigma'_i + \sigma_j) \,. \tag{6.175}$$

  By Lemma 6.3, one can select a set A_{ij} such that $P'_i(A_{ij}) - P_j(A_{ij}) \geq 1 - 2\Phi(-c_0\eta_0/2)$.

Hence in both cases, for any $i \neq j$ there $P'_i(A_{ij}) - P_j(A_{ij}) \geq 1 - \rho_2$ with $\rho_2 = 2\Phi(-c_0\eta_0/2) \in (0, 1)$.

Now suppose $\max\{\sigma_j, \sigma'_j\}/\min\{\sigma_j, \sigma'_j\} > \eta_0$, by Lemma 6.4, one can select a set A_{jj} such that $P'_j(A_{ij}) - P_j(A_{ij}) > 1 - \rho_1$, where ρ_1 is defined as

$$\rho_1 := \frac{w_{\min} - 2\epsilon}{w_{\max}} - \frac{2(1 - w_{\max})}{w_{\max}}\Phi(-\frac{1}{2}\eta_0 c_0) \,. \tag{6.176}$$

Note that $1 - \rho_1$ is the left hand side of equation (6.7). Further, $\rho_1 < w_{\min}/w_{\max} < 1$ and

$$\rho_1 > (w_{\min} - 2\epsilon)\left(\frac{1}{w_{\max}} - \frac{1 - w_{\max}}{w_{\max}}\right) = w_{\min} - 2\epsilon > \rho_2 \tag{6.177}$$

According to Lemma 6.11,

$$w'_j\rho_1 + \sum_{i:i\neq j} w'_i\rho_2 > \sum_{i=1}^{K} w'_i(1 - P'_i(A_{ij}) + P_j(A_{ij})) \geq w_{\min} - 2\epsilon; \,. \tag{6.178}$$

However the L.H.S of proceeding (6.178) satisfies that

$$w'_j\rho_1 + \sum_{i:i\neq j} w'_i\rho_2 = w'_j(\rho_1-\rho_2)+\rho_2 \leq w_{\max}\rho_1 + (1-w_{\max})\rho_2 = w_{\min} - 2\epsilon \qquad (6.179)$$

The contradiction shows that $\max\{\sigma_j/\sigma'_j, \sigma'_j/\sigma_j\} \leq \eta_0$.

Suppose now we complete this removing procedure and there are still remaining components, without loss of generosity we can assume that P_1 is the component that has smallest standard deviation among all remaining components in both P and P'. Since it is not removed, there is a P'_1 and a P_2 such that both P_1 and P_2 are matched to P'_1. Since we assumed that $\sigma_1 \leq \sigma_2$, by triangle inequality

$$c(\sigma_1+\sigma_2) \leq ||\mu_1-\mu_2|| \leq ||\mu'_1-\mu_1|| + ||\mu'_1-\mu_2|| \leq c_0\max\{\sigma_1,\sigma'_1\} + c_0\max\{\sigma_2,\sigma'_1\}. \quad (6.180)$$

Note that $c > c_0\eta_0 > c_0$ by the assumption of the theorem, we consider three cases:

- $\sigma'_1 \leq \sigma_1 \leq \sigma_2$. Then $c_0(\sigma_1+\sigma_2) < c(\sigma_1+\sigma_2) \leq c_0(\sigma_1+\sigma_2)$ and this is impossible.

- $\sigma_1 < \sigma'_1 \leq \sigma_2$. Then $c\sigma_1 + c_0\sigma_2 < c(\sigma_1+\sigma_2) \leq c_0\sigma'_1 + c_0\sigma_2$, implying that $\sigma'_1/\sigma_1 \geq c/c_0 > \eta_0$.

- $\sigma_1 \leq \sigma_2 < \sigma'_1$. Then $2c\sigma_1 \leq c(\sigma_1+\sigma_2) \leq 2c_0\sigma'_1$, also implying that $\sigma'_1/\sigma_1 \geq c/c_0 > \eta_0$.

We conclude that $\sigma'_1/\sigma_1 \geq c/c_0$. By Lemma 6.4 and definition of η_0 in (6.7), the total variation distance of P_1, P'_1 can be lower bounded by

$$TV(P_1,P'_1) > F_d\left(\frac{2\eta_0^2\log\eta_0}{\eta_0^2-1}\right) - F_d\left(\frac{2\log\eta_0}{\eta_0^2-1}\right) = 1-\rho_1\ , \qquad (6.181)$$

where ρ_1 is the same as in (6.176).

On the other hand, notice that P_1 cannot be matched with any components in P' other than P'_1. This is because if there is a P'_2 that is not removed and is matched to P_1, by the same argument above one can show that $\sigma'_2 < \sigma_1 \leq \sigma'_1$. This contradicts the minimal variance selection of σ_1.

Therefore, for any P'_j unremoved with $j \geq 2$, denote P_j be a different component (not P_1) in P that matched to P'_j, then

$$||\mu'_j - \mu_1|| \geq ||\mu'_j - \mu'_1|| - c_0 \max\{\sigma_1, \sigma'_1\} \tag{6.182}$$

$$\geq c(\sigma'_j + \sigma'_1) - c_0\sigma'_1 \tag{6.183}$$

$$\geq c\sigma'_j + (c - c_0)\sigma'_1 \tag{6.184}$$

$$> c\sigma'_j \tag{6.185}$$

Also again by selection of σ_1 $||\mu'_j - \mu_1|| > c\sigma'_j \geq c\sigma_1$. Hence $||\mu'_j - \mu_1|| > c \max\{\sigma'_j, \sigma_1\}$.

For any P'_j that has been removed, still denote P_j to be the unique component matched to P'_j, (6.185) still holds. We now show that $||\mu'_j - \mu_1|| > c \max\{\sigma_1, \sigma'_j\}$. If $\sigma_1 \leq \sigma'_j$, then it trivially holds. When $\sigma'_j < \sigma_1$, if otherwise $||\mu'_j - \mu_1|| \leq c\sigma_1$, then

$$c(\sigma_j + \sigma_1) \leq ||\mu_j - \mu_1|| \leq ||\mu_j - \mu'_j|| + ||\mu'_j - \mu_1|| \leq c_0 \max\{\sigma_j, \sigma'_j\} + c\sigma_1 \ , \tag{6.186}$$

implying that $\max\{\sigma_j, \sigma'_j\}/\sigma_j \geq c/c_0 > \eta_0$. This is impossible, therefore for any P'_j that has been removed, $||\mu_1 - \mu'_j|| \geq c \max\{\sigma_1, \sigma'_j\}$.

Combining the two cases together, again we can find sets A_i such that

- For $i \neq 1$, whether P'_i is removed or not, $P'_i(A_i) - P_1(A_i) > 1 - \rho_2$.
- For $i = 1$, $P'_1(A_1) - P_1(A_1) > 1 - \rho_1$.

The same arguments in (6.178) and (6.179) lead to a contradiction, further showing that all components should have been removed, i.e. when $c > c_0\eta_0$ the match is one-to-one and hence $K = K'$. The upper bounds on center distances and standard deviation ratios have been established earlier. □

6.7.4 Iterative Refinements on Mean and Standard Deviations

Theorem 6.5 provides conditions on separation, minimal proportion, and maximal proportion such that one can establish a one-to-one correspondence between components of P and

P'. In this section, we are going to show that once this is done $K = K'$, we can further iteratively improve the bounds. Hence we can complete the proof of Theorem 6.2. We assume all assumptions in Theorem 6.5 hold.

Proof of Theorem 6.2: Upper bounds on mean and standard deviations. By $TV(P, P') \leq 2\epsilon$, take $Q = (P + P')/2$, we have $TV(P, Q) \leq \epsilon, TV(P', Q) \leq \epsilon$, therefore we can apply Theorem 6.5, which confirms a one-to-one correspondence in the sense that centers are close. Out of simplicity in notation, we re-order components of μ'_j so that P_i, P'_i are correspondence. Specifically, for any pair P_i, P'_i, according to the assumption $||\mu_i - \mu'_i|| \leq c_0 \max\{\sigma_i, \sigma'_i\}$ and $\max\{\sigma_i/\sigma'_i, \sigma'_i/\sigma_i\} \leq \eta_0$.

Now consider a pair P_i, P'_i. WLOG assume that $w_i \leq w'_i$, otherwise one can switch the role of P and P'. For all $j \neq i$,

$$||\mu_j - \mu'_i|| \geq c(\sigma'_j + \sigma'_i) - c_0 \max\{\sigma_j, \sigma'_j\} \geq c\sigma'_i + \left(\frac{c}{\eta_0} - c_0\right) \max\{\sigma'_j, \sigma_j\} \geq c\sigma'_i + \left(\frac{c}{\eta_0} - c_0\right) \sigma_j \; ; \tag{6.187}$$

$$||\mu_j - \mu'_i|| \geq c(\sigma_j + \sigma_i) - c_0 \max\{\sigma_i, \sigma'_i\} \geq c\sigma_j + \left(\frac{c}{\eta_0} - c_0\right) \max\{\sigma_i, \sigma'_i\} \geq c\sigma_j + \left(\frac{c}{\eta_0} - c_0\right) \sigma'_i \; . \tag{6.188}$$

Therefore adding these two equation together we have

$$||\mu_j - \mu'_i|| \geq \frac{1}{2} \left(c + \frac{c}{\eta_0} - c_0\right) (\sigma_j + \sigma'_i) \; . \tag{6.189}$$

Note that $c > \eta_0 c_0$, and hence $||\mu_j - \mu'_i|| \geq \frac{1}{2} c_0 \eta_0 (\sigma_j + \sigma'_i)$. According to Lemma 6.3, there exists a set A_j such that $P_i(A_j) - P'_j(A_j) > 1 - \rho$, where

$$\rho = 2\Phi\left(-\frac{1}{2}(c + \frac{c}{\eta_0} - c_0)\right) < 2\Phi(-\eta_0 c_0/2) < 2\Phi(-c_0/2) = w_{\min} - 2\epsilon$$

. Then by Lemma 6.12, the total variation distance between P_i and P'_i is upper bounded by

$$TV(P_i, P'_i) \leq UB(c_0, \eta_0) := \frac{2\epsilon}{w_{\min}} + \frac{2(1 - w_{\min})}{w_{\min}} \Phi\left(-\frac{1}{2}(c + \frac{c}{\eta_0} - c_0)\right) < 1 - w_{\min} + 2\epsilon \; . \tag{6.190}$$

This further implies, by Lemma 6.2, $||\mu_i - \mu_i'|| \leq c_1 \max\{\sigma_i, \sigma_i'\}, c_1 = 2\Phi^{-1}(1 - \frac{1-UB(c_0,\eta_0)}{2})$. Note that $UB(c_0, \eta_0) < 1 - w_{\min} + 2\epsilon$, therefore $c_1 < c_0$. Also by Lemma 6.4, $\max\{\sigma_i/\sigma_i', \sigma_i'/\sigma_i\} \leq \eta_1$, where η_1 solves

$$F_d(\frac{2\eta_1^2 \log \eta_1}{\eta_1^2 - 1}) - F_d(\frac{2 \log \eta_1}{\eta_1^2 - 1}) = UB(c_0, \eta_0) \tag{6.191}$$

Since $1 - (K-1)w_{\min} \geq w_{\min}$,

$$F_d(\frac{2\eta_0^2 \log \eta_0}{\eta_0^2 - 1}) - F_d(\frac{2 \log \eta_0}{\eta_0^2 - 1}) \tag{6.192}$$

$$= 1 - \frac{w_{\min} - 2\epsilon}{1 - (K-1)w_{\min}} + \frac{2(K-1)w_{\min}}{1 - (K-1)w_{\min}} \Phi(-\frac{1}{2}\eta_0 c_0) \tag{6.193}$$

$$> 1 - \frac{w_{\min} - 2\epsilon}{1 - (K-1)w_{\min}} + \frac{2(K-1)w_{\min}}{1 - (K-1)w_{\min}} \Phi\left(-\frac{1}{2}(c + \frac{c}{\eta_0} - c_0)\right) \tag{6.194}$$

$$= 1 - \frac{w_{\min} - 2\epsilon - [1 - (1 - (K-1)w_{\min})]\,\rho}{1 - (K-1)w_{\min}} \qquad (\text{recall } \rho = 2\Phi(-(c + c/\eta_0 - c_0)/2)) \tag{6.195}$$

$$= 1 - \frac{w_{\min} - 2\epsilon - \rho}{1 - (K-1)w_{\min}} - \rho \tag{6.196}$$

$$\geq 1 - \frac{w_{\min} - 2\epsilon - \rho}{w_{\min}} - \rho \tag{6.197}$$

$$= 1 - \frac{w_{\min} - 2\epsilon}{w_{\min}} + \frac{1 - w_{\min}}{w_{\min}} \rho \tag{6.198}$$

$$= F_d(\frac{2\eta_1^2 \log \eta_1}{\eta_1^2 - 1}) - F_d(\frac{2 \log \eta_1}{\eta_1^2 - 1})\,, \tag{6.199}$$

it holds that $\eta_1 < \eta_0$. To conclude, for now starting with c_0, η_0, we are able to provide refined upper bounds c_1, η_1. Note that by the same argument of (6.190), one can use c_1, η_1 to refine the upper bound by

$$TV(P_i, P_i') \leq UB(c_1, \eta_1) := \frac{2\epsilon}{w_{\min}} + \frac{2(1 - w_{\min})}{w_{\min}} \Phi\left(-\frac{1}{2}(c + \frac{c}{\eta_1} - c_1)\right)\,. \tag{6.200}$$

And similarly $||\mu_i - \mu_i'|| \leq c_2 \max\{\sigma_i, \sigma_i'\}, c_2 = 2\Phi^{-1}(1 - \frac{1-UB(c_1,\eta_1)}{2})$. $\max\{\sigma_i/\sigma_i', \sigma_i'/\sigma_i\} \leq \eta_2$, where η_2 solves

$$F_d(\frac{2\eta_2^2 \log \eta_2}{\eta_2^2 - 1}) - F_d(\frac{2 \log \eta_2}{\eta_2^2 - 1}) = UB(c_1, \eta_1)\,. \tag{6.201}$$

As $UB(c_1, \eta_1) < UB(c_0, \eta_0)$, we have $c_2 < c_1, \eta_2 < \eta_1$. This procedure can be repeated. Let

$$UB(c_t, \eta_t) := \frac{2\epsilon}{w_{\min}} + \frac{2(1 - w_{\min})}{w_{\min}} \Phi\left(-\frac{1}{2}(c + \frac{c}{\eta_t} - c_t)\right) \tag{6.202}$$

Then one can find

$$c_{t+1} = 2\Phi^{-1}(1 - \frac{1 - UB(c_t, \eta_t)}{2}) = 2\Phi^{-1}(\frac{1}{2} + \frac{1}{2}UB(c_t, \eta_t)) \tag{6.203}$$

and η_{t+1} solves

$$F_d(\frac{2\eta_{t+1}^2 \log \eta_{t+1}}{\eta_{t+1}^2 - 1}) - F_d(\frac{2 \log \eta_{t+1}}{\eta_{t+1}^2 - 1}) = UB(c_t, \eta_t) \tag{6.204}$$

such that

$$||\mu_i - \mu_i'|| \leq c_{t+1} \max\{\sigma_i, \sigma_i'\}, \quad \max\{\sigma_i'/\sigma_i, \sigma_i/\sigma_i'\} \leq \eta_{t+1} \ . \tag{6.205}$$

The L.H.S. of (6.204) is a strictly increasing function in η_{t+1} and it has a continuous inverse. Therefore we conclude that there is a continuos mapping $g : [0, 1] \to [0, c_0] \times [1, \eta_0]$ such that

$$(c_{t+1}, \eta_{t+1}) = g(c_t, \eta_t) \tag{6.206}$$

By induction one shows that $UB(c_t, \eta_t) < UB(c_{t-1}, \eta_{t-1})$ for $t \geq 1$, therefore, $c_{t+1} < c_t, \eta_{t+1} < \eta_t$. Hence $\{c_t\}, \{\eta_t\}$ are two decreasing sequences, both lower bounded. Therefore their limits exist. Denote the limits as (c^*, η^*), then it must be a fixed point of g, namely they solve (6.8). Further, $||\mu_i - \mu_i'|| \leq c^* \max\{\sigma_i, \sigma_i'\}, \max\{\sigma_i'/\sigma_i, \sigma_i/\sigma_i'\} \leq \eta^*$.

The upper bounds on proportions $|w_i - w_i'|$ are established with the components c^*, η^* will be proved in the next section, where we establish additional techniques to complete the proof of Theorem 6.2. □

Before we display the proof on difference in proportions, we provide the proof to corollary 6.1.

Proof of Corollary 6.1. Suppose P' is a single spherical Gaussians such that $TV(P, P') \leq 2\epsilon$. Consider two new mixtures of spherical Gaussians

$$\tilde{P} = \frac{1}{2}P + \frac{1}{2}Q \tag{6.207}$$

$$\tilde{P}' = \frac{1}{2}P' + \frac{1}{2}Q \tag{6.208}$$

where Q is a sufficiently large far away Gaussian components. To be specfic, for any component $N_d(\mu_i, \sigma_i^2 I_d)$ in P or P', $Q = N_d(\mu, \sigma^2 I_d)$ is selected such that $||\mu - \mu_i|| \geq c \max\{\sigma, \sigma_i\}$.

Therefore $\tilde{P} \in \mathcal{M}(K+1, w_{\min}/2, c)$ and $\tilde{P}' \in \mathcal{M}(2, w_{\min}/2, c)$. However since $TV(\tilde{P}, \tilde{P}') \geq 2\epsilon/2 = \epsilon$, and the maximal proportion of $\tilde{P}, \tilde{P}'$ is upper bounded by $1/2$. Let η_0' be defined as the solution to

$$1 - w_{\min} + \epsilon + 2\Phi(-\frac{1}{2}\eta_0' c_0') = F_d(\frac{2\eta_0'^2 \log \eta_0'}{\eta_0'^2 - 1}) - F_d(\frac{2 \log \eta_0'}{\eta_0'^2 - 1}) \tag{6.209}$$

where $c_0' = 2\Phi^{-1}(1 - \frac{w_{\min} - \epsilon}{4})$, then if $c > c_0'\eta_0'$, according to Theorem 6.5, $K+1 = 2$, which implies that $K = 1$ and this is a contradiction. □

6.7.5 Difference in Proportions

Finally given η^*, c^* established previously, we will complete the proof of Theorem 6.2 by upper bounding the differences in proportions. For a corresponding pair w_i, w_i' where $||\mu_i - \mu_i'|| \leq c^* \max\{\sigma_i, \sigma_i'\}$ and $\max\{\sigma_i', \sigma_i\}/\min\{\sigma_i', \sigma_i\} \leq \eta^*$, we upper bound $|w_i - w_i'|$ by constructing the set A such that we can apply Lemma 6.13.

To start with, we introduce a geometric lemma.

Lemma 6.14. *Let x_1, x_2, y_1, y_2 be for different points in $\mathbb{R}^d$, where*

$$\tilde{x} = \alpha x_1 + (1-\alpha)x_2 \,, \tag{6.210}$$

$$\tilde{y} = \beta y_1 + (1-\beta)y_2 \,. \tag{6.211}$$

Then

$$\begin{aligned} ||\tilde{x} - \tilde{y}||^2 = {} & \alpha\beta||x_1 - y_1||^2 + (1-\alpha)(1-\beta)||x_2 - y_2||^2 + (1-\alpha)\beta||x_2 - y_1||^2 + \alpha(1-\beta)||x_1 - y_2||^2 \\ & - \alpha(1-\alpha)||x_1 - x_2||^2 - \beta(1-\beta)||y_1 - y_2||^2 \end{aligned}$$

The proof of this lemma is postponed to Section 6.8. Then consider two pairs of components P_i, P_i' and P_j, P_j', one can find a hyperplane that is far away from all four centers.

Lemma 6.15. *Let $P_i = N_d(\mu_i, \sigma_i^2 I_d), i = 1, 2$ and $P_i' = N_d(\mu_i', \sigma_i'^2 I_d)$ be two pairs of spherical Gaussian distributions such that with three constants $c^* \geq 0, \eta^* \geq 1, c \geq c^*\eta^*$ such*

that

$$\max\{\frac{\sigma_1}{\sigma_1'},\frac{\sigma_1'}{\sigma_1}\}\le\eta^*,\qquad \max\{\frac{\sigma_2}{\sigma_2'},\frac{\sigma_2'}{\sigma_2}\}\le\eta^* \tag{6.212}$$

$$||\mu_1-\mu_1'||\le\frac{c^*}{2}(\sigma_1+\sigma_1'),\quad ||\mu_2-\mu_1||\ge c(\sigma_1+\sigma_2) \tag{6.213}$$

$$||\mu_2-\mu_2'||\le\frac{c^*}{2}(\sigma_2+\sigma_2'),\quad ||\mu_1'-\mu_2'||\ge c(\sigma_1'+\sigma_2') \tag{6.214}$$

There exists a hyperplane H such that the distance

$$dist(\mu_i,H)\ge C(c^*,\eta^*,c)\sigma_i,\quad dist(\mu_i',H)\ge C(c^*,\eta^*,c)\sigma_i',\quad i=1,2 \tag{6.215}$$

where

$$C(c^*,\eta^*,c)=\sqrt{\frac{c^2}{2(\eta^*)^2}+\frac{1}{2\eta^*}(c-\frac{c^*}{2})^2-\frac{(c^*)^2(1+\eta^*)^2}{16(\eta^*)^2}}-\frac{c^*}{2} \tag{6.216}$$

Proof of Lemma 6.15. Consider two points

$$\tilde\mu_i=\frac{\sigma_i}{\sigma_i+\sigma_i'}\mu_i+\frac{\sigma_i'}{\sigma_i+\sigma_i'}\mu_i',\quad i=1,2\ , \tag{6.217}$$

Note that for both $i=1,2$,

$$||\tilde\mu_i-\mu_i||=\frac{\sigma_i'}{\sigma_i+\sigma_i'}||\mu_i'-\mu_i||\le\frac{c^*\max\{\sigma_i,\sigma_i'\}}{2},\quad ||\tilde\mu_i-\mu_i'||=\frac{\sigma_i}{\sigma_i+\sigma_i'}||\mu_i'-\mu_i||\le\frac{c^*\max\{\sigma_i,\sigma_i'\}}{2} \tag{6.218}$$

that is, μ_i,μ_i' both lies in a ball centered at $\tilde\mu_i$ with radius $c^*\max\{\sigma_i,\sigma_i'\}/2$. The hyperplane we consider is

$$H:\left\{x\in\mathbb{R}^d:\langle x-\tilde\mu_1,\tilde\mu_2-\tilde\mu_1\rangle=\frac{\sigma_1+\sigma_1'}{\sigma_1+\sigma_1'+\sigma_2+\sigma_2'}||\tilde\mu_2-\tilde\mu_1||\right\} \tag{6.219}$$

By Lemma 6.14, the distance

$$\begin{aligned}||\tilde\mu_1-\tilde\mu_2||^2=&\underbrace{\frac{\sigma_1\sigma_2}{(\sigma_1+\sigma_1')(\sigma_2+\sigma_2')}||\mu_1-\mu_2||^2+\frac{\sigma_1'\sigma_2'}{(\sigma_1+\sigma_1')(\sigma_2+\sigma_2')}||\mu_1'-\mu_2'||^2}_{LB_1}\\&+\underbrace{\frac{\sigma_1\sigma_2'}{(\sigma_1+\sigma_1')(\sigma_2+\sigma_2')}||\mu_1-\mu_2'||^2+\frac{\sigma_1'\sigma_2}{(\sigma_1+\sigma_1')(\sigma_2+\sigma_2')}||\mu_1'-\mu_2||^2}_{LB_2}\\&-\underbrace{\frac{\sigma_1\sigma_1'}{(\sigma_1+\sigma_1')^2}||\mu_1-\mu_1'||^2-\frac{\sigma_2\sigma_2'}{(\sigma_2+\sigma_2')^2}||\mu_2-\mu_2'||^2}_{UB}\end{aligned}$$

We will establish a lower bound on $||\tilde{\mu}_1 - \tilde{\mu}_2||$ by separately lower bounding LB_1, LB_2 and UB.

First, by the seperation condition

$$LB_1 \geq c^2 \frac{\sigma_1\sigma_2(\sigma_1+\sigma_2)^2 + \sigma_1'\sigma_2'(\sigma_1'+\sigma_2')^2}{(\sigma_1+\sigma_1')(\sigma_2+\sigma_2')(\sigma_1+\sigma_2+\sigma_1'+\sigma_2')^2}(\sigma_1+\sigma_2+\sigma_1'+\sigma_2')^2 \tag{6.220}$$

Note that $\max\{\sigma_i/\sigma_i', \sigma_i'/\sigma_i\} \leq \eta^*, i = 1, 2$, we denote

$$\alpha = \frac{\sigma_1}{\sigma_1+\sigma_1'}\ , \quad \beta = \frac{\sigma_2'}{\sigma_2+\sigma_2'}\ , \quad \zeta = \frac{\sigma_1+\sigma_2}{\sigma_1+\sigma_1'+\sigma_2+\sigma_2'}\ , \tag{6.221}$$

then since $1/(1+\eta^*) \leq \alpha, \beta, \zeta \leq \eta^*/(1+\eta^*)$, it holds that

$$\frac{LB_1}{c^2(\sigma_1+\sigma_2+\sigma_1'+\sigma_2')^2} = \alpha\beta\zeta^2 + (1-\alpha)(1-\beta)(1-\zeta)^2 \tag{6.222}$$

$$\geq \left(\frac{1}{1+\eta^*}\right)^2 [\zeta^2 + (1-\zeta)^2] \tag{6.223}$$

$$\geq \frac{1}{2}\left(\frac{1}{1+\eta^*}\right)^2 . \tag{6.224}$$

namely

$$LB_1 \geq \frac{c^2}{2}\left(\frac{1}{1+\eta^*}\right)^2 (\sigma_1+\sigma_2+\sigma_1'+\sigma_2')^2 \tag{6.225}$$

For the second term, start by observing that

$$||\mu_1 - \mu_2'|| \geq c(\sigma_1+\sigma_2) - \frac{c^*}{2}(\sigma_2+\sigma_2'), \quad ||\mu_1 - \mu_2'|| \geq c(\sigma_1'+\sigma_2') - \frac{c^*}{2}(\sigma_1+\sigma_1') \tag{6.226}$$

Adding these two equations we obtain that

$$||\mu_1 - \mu_2'|| \geq \frac{1}{2}(c - \frac{c^*}{2})(\sigma_1+\sigma_1'+\sigma_2+\sigma_2')\ . \tag{6.227}$$

and similarly $||\mu_1' - \mu_2|| \geq (c - \frac{c^*}{2})(\sigma_1+\sigma_1'+\sigma_2+\sigma_2')/2$. Hence

$$\frac{LB_2}{(\sigma_1+\sigma_1'+\sigma_2+\sigma_2')^2} \geq \frac{1}{4}(c - \frac{c^*}{2})^2[\alpha(1-\beta) + \beta(1-\alpha)]\ . \tag{6.228}$$

Write $g(\alpha, \beta) = \alpha(1-\beta) + \beta(1-\alpha) = \beta(1-2\alpha) + \alpha$.

- If $\alpha = 1/2$ then $g(\alpha, \beta) = 1/2$.

- If $1/(1+\eta^*) \leq \alpha < 1/2$, then $g(\alpha, \beta) \geq g(\alpha, 1/(1+\eta^*))$ since it is an affine function in β with positive linear term. When $\beta = 1/(1+\eta^*) < 1/2$, by same argument we know that $g(\alpha, \beta)$ is an affine function in α with positive linear term. Hence the minimum is

$$g(\frac{1}{1+\eta^*}, \frac{1}{1+\eta^*}) = \frac{2\eta^*}{(1+\eta^*)^2} \ . \tag{6.229}$$

- If $1/2 < \alpha \leq \eta^*/(1+\eta^*)$, the same arguments show that the minimum is

$$g(\frac{\eta^*}{1+\eta^*}, \frac{\eta^*}{1+\eta^*}) = \frac{2\eta^*}{(1+\eta^*)^2} \ . \tag{6.230}$$

As a conclution we have

$$LB_2 \geq \frac{1}{4}\frac{2\eta^*}{(1+\eta^*)^2}(c - \frac{c^*}{2})^2(\sigma_1 + \sigma_1' + \sigma_2 + \sigma_2')^2 \tag{6.231}$$

Finally, we establish an upper bound for UB,

$$UB = \frac{\sigma_1\sigma_1'}{(\sigma_1+\sigma_1')^2}||\mu_1 - \mu_1'||^2 - \frac{\sigma_2\sigma_2'}{(\sigma_2+\sigma_2')^2}||\mu_2 - \mu_2'||^2 \tag{6.232}$$

$$\leq \frac{(c^*)^2}{16}(\sigma_1 + \sigma_1')^2 + \frac{(c^*)^2}{16}(\sigma_2 + \sigma_2')^2 \tag{6.233}$$

$$\leq \frac{(c^*)^2}{16}(\sigma_1 + \sigma_1' + \sigma_2 + \sigma_2')^2 \tag{6.234}$$

Therefore

$$||\tilde{\mu}_1 - \tilde{\mu}_2|| \geq \sqrt{\frac{c^2}{2}\left(\frac{1}{1+\eta^*}\right)^2 + \frac{1}{4}\frac{2\eta^*}{(1+\eta^*)^2}(c - \frac{c^*}{2})^2 - \frac{(c^*)^2}{16}}(\sigma_1 + \sigma_1' + \sigma_2 + \sigma_2') \tag{6.235}$$

The distance of μ_1, μ_1' to H is then lower bounded by

$$dist(\mu_1, H) \geq ||\tilde{\mu}_1 - \tilde{\mu}_2|| \frac{\sigma_1 + \sigma_1'}{\sigma_1 + \sigma_1' + \sigma_2 + \sigma_2'} - \frac{c^*}{2} \max\{\sigma_1, \sigma_1'\} \tag{6.236}$$

$$\geq \sqrt{\frac{c^2}{2}\left(\frac{1}{1+\eta^*}\right)^2 + \frac{1}{4}\frac{2\eta^*}{(1+\eta^*)^2}(c - \frac{c^*}{2})^2 - \frac{(c^*)^2}{16}}(\sigma_1 + \sigma_1') - \frac{c^*}{2}\max\{\sigma_1, \sigma_1'\} \tag{6.237}$$

$$\geq \sqrt{\frac{c^2}{2}\left(\frac{1}{1+\eta^*}\right)^2 + \frac{1}{4}\frac{2\eta^*}{(1+\eta^*)^2}(c - \frac{c^*}{2})^2 - \frac{(c^*)^2}{16}}(\frac{1}{\eta^*} + 1)\max\{\sigma_1, \sigma_1'\} - \frac{c^*}{2}\max\{\sigma_1, \sigma_1'\} \tag{6.238}$$

$$\geq \left(\sqrt{\frac{c^2}{2(\eta^*)^2} + \frac{1}{2\eta^*}(c - \frac{c^*}{2})^2 - \frac{(c^*)^2(1+\eta^*)^2}{16(\eta^*)^2}} - \frac{c^*}{2}\right)\max\{\sigma_1, \sigma_1'\} \tag{6.239}$$

$$:= C(c^*, \eta^*, c)\max\{\sigma_1, \sigma_1'\} \tag{6.240}$$

Note that since $c > c^*\eta^*$ and $\eta^* \geq 1$,

$$\frac{c^2}{2(\eta^*)^2} + \frac{1}{2\eta^*}(c - \frac{c^*}{2})^2 - \frac{(c^*)^2(1+\eta^*)^2}{16(\eta^*)^2} - \frac{(c^*)^2}{4} \tag{6.241}$$

$$> \frac{(c^*)^2}{2} + \frac{(c^*)^2(\eta^* - \frac{1}{2})^2}{2\eta^*} - \frac{(c^*)^2(1+\eta^*)^2}{16(\eta^*)^2} - \frac{(c^*)^2}{4} \tag{6.242}$$

$$= (c^*)^2\left[\frac{4(\eta^*)^2 + 8\eta^*(\eta^* - \frac{1}{2})^2 - (1+\eta^*)^2}{16(\eta^*)^2}\right] \tag{6.243}$$

$$= \left(\frac{c^*}{4\eta^*}\right)^2\left[8(\eta^*)^3 - 5(\eta^*)^2 - 1\right] > 0 \tag{6.244}$$

The coefficient $C(c^*, \eta^*, c)$ in (6.240) is positive, so

$$dist(\mu_1, H) \geq \left(\sqrt{\frac{c^2}{2(\eta^*)^2} + \frac{1}{2\eta^*}(c - \frac{c^*}{2})^2 - \frac{(c^*)^2(1+\eta^*)^2}{16(\eta^*)^2}} - \frac{c^*}{2}\right)\sigma_1 \tag{6.245}$$

$$dist(\mu_1', H) \geq \left(\sqrt{\frac{c^2}{2(\eta^*)^2} + \frac{1}{2\eta^*}(c - \frac{c^*}{2})^2 - \frac{(c^*)^2(1+\eta^*)^2}{16(\eta^*)^2}} - \frac{c^*}{2}\right)\sigma_1' \tag{6.246}$$

This also shows that μ_1, μ_1' are both with the same side of H as $\tilde{\mu}_1$. μ_2, μ_2' can be similarly shown to be with the same side of H as $\tilde{\mu}_2$, which is different from μ_1, μ_1' and and the other two distance $dist(\mu_2, H), dist(\mu_2', H)$ can be lower bounded similarly. □

Proof of Theorem 6.2: Difference in Proportions. Now we can finish the proof to Theorem

6.2. Let P_i, P_i' be two corresponding pairs such that

$$\max\{\frac{\sigma_i}{\sigma_i'}, \frac{\sigma_i'}{\sigma_i}\} \leq \eta^*, \quad ||\mu_i - \mu_i'|| \leq \frac{c^*}{2}(\sigma_i + \sigma_i') \ . \tag{6.247}$$

Here the center distance is slightly different. In fact for the corresponding pairs, it holds that

$$TV(P_i, P_i') \leq \frac{2\epsilon}{w_{\min}} + \frac{2(1-w_{\min})}{w_{\min}}\Phi(-\frac{1}{2}\left(c + \frac{c}{\eta^*} - c^*\right)) = 1 - 2\Phi(-\frac{c^*}{2}) \ . \tag{6.248}$$

According to Lemma 6.3, it also holds that $||\mu_i - \mu_i'|| \leq c^*(\sigma_i + \sigma_i')/2$. Consider an arbitrary different corresponding pair P_j, P_j', then P_i, P_i', P_j, P_j' satisfy the conditions of Lemma 6.15, hence we can select a hyperplane H such that P_i, P_i' and P_j, P_j' are on different side of H, and their distance are lower bounded. Let A_j be the halfspace determined by H and including P_j, P_j', then

$$\max\{P_i(A_j^{\complement}), P_i'(A_j^{\complement}), P_j(A_j), P_j'(A_j)\} \leq \Phi(-C(c, c^*, \eta^*)) \tag{6.249}$$

where $C(c, c^*, \eta^*)$ is determined in (6.216). Now take $A = \cap_{j \neq i} A_j$, then

$$P_i(A) = P_i((\cup_{j\neq i} A_j^{\complement})^{\complement}) \geq 1 - \sum_{j \neq i} P_i(A_j^{\complement}) \geq 1 - (K-1)\Phi(-C(c, c^*, \eta^*)) \ , \tag{6.250}$$

$$P_i'(A) = P_i'((\cup_{j\neq i} A_j^{\complement})^{\complement}) \geq 1 - \sum_{j \neq i} P_i'(A_j^{\complement}) \geq 1 - (K-1)\Phi(-C(c, c^*, \eta^*)) \ ; \tag{6.251}$$

$$P_j(A) \leq P_j(A_j) \leq \Phi(-C(c, c^*, \eta^*)) \tag{6.252}$$

$$P_j'(A) \leq P_j'(A_j) \leq \Phi(-C(c, c^*, \eta^*)) \tag{6.253}$$

And the result is given by Lemma 6.13. □

6.8 Auxillary Lemmas

6.8.1 Lemmas in Section 6.2

Lemma 6.16. *Let F_d be the cumulative distribution function of a Gamma($\frac{d}{2}, \frac{d}{2}$) distribution, then for any $\eta \geq 1$,*

$$\textit{R.H.S. of } (6.7) = F_d(\frac{2\eta^2 \log \eta}{\eta^2 - 1}) - F_d(\frac{2 \log \eta}{\eta^2 - 1}) \geq 1 - \left(\frac{2\eta}{\eta^2 + 1}\right)^{\frac{d}{2}} \ , \tag{6.254}$$

Proof of Lemma 6.16. Consider two distributions $P_1 = N_d(0, I_d)$ and $P_2 = N_d(0, \eta^2 I_d)$ with $\eta \geq 1$ and the set

$$A = \{x \in \mathbb{R}^d : p_1(x) \geq p_2(x)\} = \left\{ x \in \mathbb{R}^d : ||x|| \leq \frac{\eta}{\eta^2 - 1} \sqrt{2d(\eta^2 - 1) \log \eta} \right\} . \quad (6.255)$$

where $p_i, i = 1, 2$ is the density of P_i. Then $P_1(A) - P_2(A)$ is the R.H.S. expression in (6.7). Further, it is the total variation distance between P_1 and P_2. Well known result states that total variation distance is lower bounded by Hellinger distance [Gibbs and Su, 2002], therefore it holds that

$$\text{R.H.S. of (6.7)} \geq \frac{1}{2} \int (\sqrt{p_1} - \sqrt{p_2})^2 = 1 - \left(\frac{2\eta}{\eta^2 + 1} \right)^{\frac{d}{2}} . \quad (6.256)$$

□

6.8.2 Proof of nonsymmetric result

Proof of theorem 6.3. **One-to-one correspondence** By lemma 6.3, if there is a component P'_i of P' and two different components of P: P_i, P_j such that total variation distance $TV(P_i, P'_i) \leq 1 - \rho, TV(P_j, P'_i) \leq 1 - \rho$. WLOG assume $\sigma_i \leq \sigma_j$. Then with $C_b^{(0)} = C_s^{(0)} := 2\Phi^{-1}(1 - \rho/2)$, it holds that

$$||\mu_i - \mu'_i|| \leq \frac{C_s^{(0)}}{2}(\sigma_i + \sigma'_i), \quad ||\mu_j - \mu'_i|| \leq \frac{C_b^{(0)}}{2}(\sigma_j + \sigma'_i) . \quad (6.257)$$

By the separation condition of P we have

$$c(\sigma_i + \sigma_j) \leq ||\mu_i - \mu_j|| \leq \frac{C_b^{(0)}}{2}(\sigma_i + \sigma'_i) + \frac{C_s^{(0)}}{2}(\sigma_j + \sigma'_i) , \quad (6.258)$$

or

$$\sigma'_i \geq \frac{c - \frac{C_b^{(0)}}{2}}{\frac{C_b^{(0)}}{2} + \frac{C_s^{(0)}}{2}}(\sigma_i + \sigma_j) \geq \frac{4c - 2C_b^{(0)}}{C_s^{(0)} + C_b^{(0)}}\sigma_i , \quad (6.259)$$

$$\sigma'_i \geq \frac{c - \frac{C_b^{(0)}}{2}}{\frac{C_b^{(0)}}{2} + \frac{C_s^{(0)}}{2}}(1 + \frac{1}{\kappa})\sigma_j .$$

Note that $c > 2\Phi^{-1}(1-\rho/2) = C_b^{(0)} = C_s^{(0)}$, then $(4c - 2C_b^{(0)})/(C_b^{(0)} + C_s^{(0)}) > 1$. Since $h(C, \eta)$ is increasing in C and η respectively, if $h(0, (4c - 2C_b^{(0)})/(C_b^{(0)} + C_s^{(0)})) > 1 - \rho$ it will be impossible that $TV(P_i, P_i') \leq 1 - \rho$. This contradiction leads to the existence of one-to-one correspondence.

Now when $h(0, (4c - 2C_b^{(0)})/(C_b^{(0)} + C_s^{(0)})) \leq 1 - \rho$, since $h(C_s^{(0)}, (4c - 2C_b^{(0)})/(C_b^{(0)} + C_s^{(0)})) > h(C_s^{(0)}, 1) = 1 - \rho$, by continuity and monotonicity of h in C there exists $C_s^{(1)} \in [0, C_s^{(0)})$ such that $h(C_s^{(1)}, (4c - 2C_b^{(0)})/(C_b^{(0)} + C_s^{(0)})) = 1 - \rho$. Also since $\kappa \geq 1$,

$$\frac{c - \frac{C_b^{(0)}}{2}}{\frac{C_b^{(0)}}{2} + \frac{C_s^{(0)}}{2}}(1 + \frac{1}{\kappa}) \leq \frac{4c - 2C_b^{(0)}}{C_s^{(0)} + C_b^{(0)}} \, , \tag{6.260}$$

and further $h(0, (1+1/\kappa)(2c - C_b^{(0)})/(C_b^{(0)} + C_s^{(0)})) \leq h(0, (4c - 2C_b^{(0)})/(C_b^{(0)} + C_s^{(0)}) < 1 - \rho$. Again by continuity and monotonicity of h in C there exists $C_b^{(1)} \in [C_s^{(1)}, C_b^{(0)})$ such that $h(C_b^{(1)}/2, (1 + 1/\kappa)(2c - C_b^{(0)})/(C_b^{(0)} + C_s^{(0)})) = 1 - \rho$. And it further holds that

$$||\mu_i - \mu_i'|| \leq \frac{C_s^{(1)}}{2}(\sigma_i + \sigma_i'), \quad ||\mu_j - \mu_i'|| \leq \frac{C_b^{(1)}}{2}(\sigma_j + \sigma_i') \, . \tag{6.261}$$

Again by separation condition of P we have

$$c(\sigma_i + \sigma_j) \leq ||\mu_i - \mu_j|| \leq \frac{C_b^{(1)}}{2}(\sigma_i + \sigma_i') + \frac{C_s^{(1)}}{2}(\sigma_j + \sigma_i') \leq \frac{C_b^{(1)}}{2}(\sigma_i + \sigma_j) + (\frac{C_s^{(1)}}{2} + \frac{C_b^{(1)}}{2})\sigma_i' \, , \tag{6.262}$$

or after rearrangement

$$\begin{aligned} \sigma_i' &\geq \frac{c - \frac{C_b^{(1)}}{2}}{\frac{C_b^{(1)}}{2} + \frac{C_s^{(1)}}{2}}(\sigma_i + \sigma_j) \geq \frac{4c - 2C_b^{(1)}}{C_s^{(1)} + C_b^{(1)}}\sigma_i \, , \\ \sigma_i' &\geq \frac{c - \frac{C_b^{(1)}}{2}}{\frac{C_b^{(1)}}{2} + \frac{C_s^{(1)}}{2}}(1 + \frac{1}{\kappa})\sigma_j \, . \end{aligned} \tag{6.263}$$

Similarly as previous argument, if $h(0, (4c - 2C_b^{(1)})/(C_b^{(1)} + C_s^{(1)})) > 1 - \rho$, it is impossible and we obtain the one-to-one correspondence guarantee. Otherwise we can solve for $C_s^{(2)} \in [0, C_s^{(1)})$ and $C_b^{(2)} \in [C_s^{(2)}, C_b^{(1)})$ such that

$$h(\frac{C_s^{(2)}}{2}, \frac{4c - 2C_b^{(1)}}{C_s^{(1)} + C_b^{(1)}}) = 1 - \rho \, , \quad h(\frac{C_b^{(2)}}{2}, \frac{c - \frac{C_b^{(1)}}{2}}{\frac{C_b^{(1)}}{2} + \frac{C_s^{(1)}}{2}}(1 + \frac{1}{\kappa})) = 1 - \rho \, .$$

And we can further updated the distance upper bound as in equation 6.257 and equation 6.261. This further refines the variance bounds as in equation 6.259 to 6.263.

By induction, when $h(0, (4c - 2C_b^{(t)})/(C_b^{(t)} + C_s^{(t)})) \leq 1 - \rho$, we can always find $C_s^{(t+1)} \in [0, C_s^{(t)})$ and $C_b^{(t+1)} \in [C_s^{t+1}, C_b^{(t)})$ such that equations

$$h(\frac{C_s^{(t+1)}}{2}, \frac{4c - 2C_b^{(t)}}{C_s^{(t)} + C_b^{(t)}}) = 1 - \rho \ , \quad h(\frac{C_b^{(t+1)}}{2}, \frac{c - \frac{C_b^{(t)}}{2}}{\frac{C_b^{(t)}}{2} + \frac{C_s^{(t)}}{2}}(1 + \frac{1}{\kappa})) = 1 - \rho \ .$$

hold. Hence we can construct two decreasing sequences $\{C_b^{(t)}\}_{t=0,1,2,\cdots}$ and $\{C_s^{(t)}\}_{t=0,1,2,\cdots}$. When algorithm One2OneCriterionoutputs true, then the construction of these two sequences stops after $T < \infty$ steps. As we see previously, as long as this procedure stops after finite steps, we get a contradiction since the variance ratio σ_i'/σ_i is so large that the total variation distance between P_i, P_i' must be greater than $1 - \rho$. This means that the existence of two components P_i, P_j are both within total variation distance 2ϵ of P_i' is impossible.

On the other hand, since $\rho < w_{\min} - 2\epsilon$, lemma 6.4 shows that there must exists a component P_i of P such that $TV(P_i, P_i') \leq 1 - w_{\min} + 2\epsilon < 1 - \rho$. Hence there exists a one-to-one correspondence between components of P and P' in the sense that each paired components are within total variation distance $1 - \rho$.

Variance ratio upper bound For any correspondent pairs P_i, P_i', denote total variation distance $TV(P_i, P_i') = 1 - \rho_{ii}$. Since for any other components P_j we have $TV(P_j, P_i') > 1 - rho$, according to lemma 6.11, we have $w_i\rho_{ii} + (1 - w_i)\rho \geq w_{\min} - 2\epsilon$, which further shows that

$$\rho_{ii} \geq \frac{w_{\min} - 2\epsilon}{w_i} - \frac{(1 - w_i)\rho}{w_i} = \frac{1}{w_i}(w_{\min} - 2\epsilon - \rho) + \rho. \tag{6.264}$$

Since $\rho < w_{\min} - 2\epsilon$, the right hand side of the previous equation is decreasing in w_i and its minimal value is reached at $w_i = w_K$. Therefore the total variation distance between P_i, P_i' is upper bounded by

$$1 - \frac{w_{\min} - 2\epsilon}{w_K} + \frac{1 - w_K}{w_K}\rho \tag{6.265}$$

If $\max\{\sigma_i/\sigma_i', \sigma_i'/\sigma_i\} > \eta_\rho$ defined by equation 6.15, it holds that

$$TV(P_i, P_i') \geq h(0, \eta_{\max}) = 1 - \frac{w_{\min} - 2\epsilon}{w_K} + \frac{1 - w_K}{w_K}\rho \ . \tag{6.266}$$

This contradiction leads to the upper bound on variances.

Separation lower bound Finally, we will prove the lower bound of separation constant $c' := \min_{i\neq j} \frac{||\mu_i' - \mu_j'||}{\sigma_i' + \sigma_j'}$ of P'. It suffices to prove the result when $c/\eta_\rho - C_0(1+\eta_\rho)/2\eta_\rho > 0$. For any two different components P_i', P_j' of P', denote $||\mu_i' - \mu_j'|| = c'(\sigma_i' + \sigma_j')$. Under the assumption such that the one-to-one correspondence under total variation distance ρ holds, it holds that there exist different components of P, P_i, P_j such that $TV(P_i, P_i') \leq 1 - w_{\min} + 2\epsilon$ and $TV(P_j, P_j') \leq 1 - w_{\min} + 2\epsilon$. By lemma 6.3, it holds that with $C_0 = 2\Phi^{-1}(1 - (w_{\min} - 2\epsilon)/2)$,

$$||\mu_i - \mu_i'|| \leq \frac{C_0}{2}(\sigma + \sigma_i'), \quad ||\mu_j - \mu_j'|| \leq \frac{C_0}{2}(\sigma_j + \sigma_j') \; . \tag{6.267}$$

Therefore by triangular inequality and separation condition of P we have

$$c(\sigma_i + \sigma_j) \leq ||\mu_i - \mu_j|| \leq c'(\sigma_i' + \sigma_j') + \frac{C_0}{2}(\sigma_i + \sigma_i') + \frac{C_0}{2}(\sigma_j + \sigma_j') \tag{6.268}$$

Or after rearrangement

$$\sigma_i' + \sigma_j' \geq \frac{c - \frac{C_0}{2}}{c' + \frac{C_0}{2}}(\sigma_i + \sigma_j) \; . \tag{6.269}$$

This means at least one of σ_i'/σ_i and σ_j'/σ_j is greater than or equal to $(c - C_0/2)/(c' + C_0/2)$. However as we have shown, this variance ratio is upper bounded by η_ρ, therefore it must be true that

$$\frac{c - \frac{C_0}{2}}{c' + \frac{C_0}{2}} \geq \eta_\rho \; , \quad c' \geq \frac{c}{\eta_\rho} - \frac{C_0(1+\eta_\rho)}{2\eta_\rho} \tag{6.270}$$

Hence the result is proved. □

6.8.3 Lemmas in Section **??**

Lemma 6.17. *If there exists a set A such that $P_1(A) - P_2(A) \geq 1 - \rho \geq 0$. Then for any weights $0 \leq w_1, w_2 \leq 1$:*

$$w_1 - \max\{w_1, w_2\}\rho \geq w_1 P_1(A) - w_2 P_2(A) \leq w_1 \tag{6.271}$$

Proof of Lemma 6.17. Note that

$$w_1 P_1(A) - w_2 P_2(A) \geq w_1(P_2(A) + 1 - \rho) - w_2 P_2(A) = w_1 + (w_1 - w_2)P_2(A) - w_1\rho \; .$$

View this as an affine function in $P_2(A)$. Since $0 \le P_2(A) \le \rho$, we have

$$w_1P_1(A)-w_2P_2(A) \ge \min\{w_1+(w_1-w_2)\cdot 0-w_1\rho, w_1+(w_1-w_2)\rho-w_1\rho\} = w_1-\max\{w_1,w_2\}\rho\ . \tag{6.272}$$

□

Lemma 6.18. *When $x > 1$, $f(x) = x\log x/(x-1)$ is increasing in x and $g(x) = \log x/(x-1)$ is decreasing in x.*

Proof of Lemma 6.18. Take derivative

$$f'(x) = \frac{(x-1)(1+\log x) - x\log x}{(x-1)^2} = \frac{x-1-\log x}{(x-1)^2} \ge 0 \tag{6.273}$$

$$g'(x) = \frac{(x-1)/x - \log x}{(x-1)^2} = \frac{1}{(x-1)^2}[1-\frac{1}{x}+\log\frac{1}{x}] \le 0 \tag{6.274}$$

□

Proof ofLemma 6.14. This lemma can be proved by direct computation. First, note that

$$||y_1-y_2||^2 = ||y_1-x_1||^2+||y_2-x_1||^2-2\langle y_1-x_1, y_2-x_1\rangle \tag{6.275}$$

Then

$$\begin{aligned}
&||x_1-\tilde{y}||^2 \\
&= ||\beta(x_1-y_1)+(1-\beta)(x_1-y_2)||^2 && (6.276)\\
&= \beta^2||x_1-y_1||^2+(1-\beta)^2||x_1-y_2||^2+2\beta(1-\beta)\langle x_1-y_1, x_1-y_2\rangle && (6.277)\\
&= \beta^2||x_1-y_1||^2+(1-\beta)^2||x_1-y_2||^2+2\beta(1-\beta)\frac{||x_1-y_1||^2+||x_1-y_2||^2-||y_1-y_2||^2}{2} && (6.278)\\
&= \beta||x_1-y_1||^2+(1-\beta)||x_1-y_2||^2-\beta(1-\beta)||y_1-y_2||^2\ . && (6.279)
\end{aligned}$$

Similarly we have

$$||x_2-\tilde{y}||^2 = \beta||x_2-y_1||^2+(1-\beta)||x_2-y_1||^2-\beta(1-\beta)||y_1-y_2||^2\ . \tag{6.280}$$

Combining all above together, it holds that

$$
\begin{aligned}
&||\tilde{x} - \tilde{y}||^2 \\
&= \alpha||\tilde{y} - x_1||^2 + (1-\alpha)||\tilde{y} - x_2||^2 - \alpha(1-\alpha)||x_1 - x_2||^2 \qquad (6.281) \\
&= \alpha\beta||x_1 - y_1||^2 + (1-\alpha)(1-\beta)||x_2 - y_2||^2 + (1-\alpha)\beta||x_2 - y_1||^2 + \alpha(1-\beta)||x_1 - y_2||^2 \\
&\quad - \alpha(1-\alpha)||x_1 - x_2||^2 - \beta(1-\beta)||y_1 - y_2||^2 \qquad (6.282)
\end{aligned}
$$

□

Chapter 7

DISCUSSION AND FUTURE WORKS

This thesis studies two significant problems in unsupervised learning: manifold learning and clustering. The motivation of this research is to establish mathematically rigorous methods that enable practitioners to understand what the algorithm is doing better, even if there is no ground truth label for unsupervised learning problems. Specifically, we propose two criteria for a practically meaningful unsupervised learning paradigm:

1. Given the same dataset, the algorithm output should remain (approximately) fixed

2. Given the result, the domain expert should be able to interpret with domain knowledge easily.

In the first part of the thesis, the main focus is on solving the second question in manifold learning paradigms. The first criterion is still needed, but for manifold learning algorithms with mathematical foundation (Diffusion maps, Laplacian Eigenmap, etc.), they are often cast as an eigenvalue problem and can satisfy the first requirement. The major obstacle for practitioners is interpreting such algorithms and introducing domain knowledge. In the application example we provide, on the MDS data set, previous validation is achieved by visual inspection, where our methods can automatically select domain functions that "explains" the algorithm output.

One major drawback of the group lasso based method (including both Manifold-Lasso and TSLasso) is that the selected support does not enjoy the stability requirement. The theory in chapter 3 shows that the group of explanatory domain functions is valid as long as it is full rank and hence not unique. Developing criteria and methods to distinguish equivalent subsets of dictionary functions is still necessary. The incoherence condition could be a strong requirement for real data. Also, it is still under research if the method can be applied to large-scale MDS data on more complicated systems, e.g., on proteins [].

In chapter 5, we discuss the $(\gamma, \epsilon)-$stability on K-means clustering for data in $\mathbb{R}^D$. We define a new stability notion, a refinement of previously studied resampling stability. Our notion of stability (1) more precisely characterizes the behavior of K-means clustering loss and its relationship with clusterability, and (2) can connect finite sample stability with population stability with very mild Glivenko-Cantelli assumption on the population P, which significantly overcomes one major drawback of previously proposed resampling stability. On an algorithmic level, we propose an algorithm based on convex relaxations of K-means clustering that can establish an optimality interval, as a guarantee of clustering.

One of the interesting research directions for the theoretical study of clustering in the future is a more thorough study on the limiting behavior of $\epsilon^*_P(\gamma)/\gamma$ when $\gamma \to 0$. It is still unknown whether this limiting behavior is related to clusterability beyond specific generative models (e.g., stochastic ball model). In practice, $(\gamma, \epsilon)-$stability may serve as a criterion for selecting cluster numbers, especially to avoid fitting with too many clusters.

In chapter 6, we discuss whether the $(\gamma, \epsilon)-$stability is useful in model-based clustering. We achieve this by modifying the definition of (γ, ϵ) stability with a different selection of closeness measurement and loss functions. Our result is the first tractable quantitative upper bound of the parameter distance of two spherical Gaussian mixtures when their total variation distance is small.

There are two potential future directions following this result. First, we must tighten our results and weaken our assumptions to make our bound useful so that it is practically meaningful. For example, our bound does not tend to zero when the total variation distance ϵ tends to zero, therefore losing tightness for an extremely small total variation distance. The difficulty comes from the proof technique: for a fixed Gaussian mixture, however large the separation is, the far-away components will still influence the density. A different proof technique is needed to complete this result. The second potential is to turn the guarantee into a finite sample version. Currently, our result is on the population level and is not directly applicable to a finite sample because estimating total variation distance is only possible with further assumptions on population P.

BIBLIOGRAPHY

Multivariate Normal Mixtures, chapter 3, pages 81–116. John Wiley & Sons, Ltd, 2000. ISBN 9780471721185. doi: https://doi.org/10.1002/0471721182.ch3. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0471721182.ch3.

Eddie Aamari and Clément Levrard. Stability and minimax optimality of tangential delaunay complexes for manifold reconstruction. *Discrete & Computational Geometry*, 59 (4):923–971, 2018. doi: 10.1007/s00454-017-9962-z. URL https://doi.org/10.1007/s00454-017-9962-z.

Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *Ann. Stat.*, 47(1):177–204, February 2019.

Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13 (1):1359 – 1399, 2019. doi: 10.1214/19-EJS1551. URL https://doi.org/10.1214/19-EJS1551.

Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the 18th Annual Conference on Learning Theory*, COLT'05, page 458–469, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3540265562. doi: 10.1007/11503415_31. URL https://doi.org/10.1007/11503415_31.

Matthew A. Addicoat and Michael A. Collins. Potential energy surfaces: the forces of chemistry. In Mark Brouard and Claire Vallance, editors, *Tutorials in Molecular Reaction Dynamics*, chapter 2, pages 28–49. Royal Society of Chemistry Publishing, London, 2010.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference*

on Neural Information Processing Systems, NIPS'18, page 9525–9536, Red Hook, NY, USA, 2018. Curran Associates Inc.

Sara Ahmadian and Chaitanya Swamy. Approximation algorithms for clustering problems with lower bounds and outliers. In *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, pages 69:1–69:15, 2016. doi: 10.4230/LIPIcs.ICALP.2016.69.

Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. Np-hardness of euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009. doi: 10.1007/s10994-009-5103-0. URL `https://doi.org/10.1007/s10994-009-5103-0`.

El-Ad David Amir, Kara L Davis, Michelle D Tadmor, Erin F Simonds, Jacob H Levine, Sean C Bendall, Daniel K Shenfeld, Smita Krishnaswamy, Garry P Nolan, and Dana Pe'er. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.*, 31(6):545–552, June 2013.

Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James R. Voss. The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures. In *COLT*, 2014.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. doi: 10.1017/CBO9780511624216.

Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, JesúS M. PéRez, and IñIgo Perona. An extensive comparative study of cluster validity indices. *Pattern Recogn.*, 46(1):243–256, January 2013. ISSN 0031-3203. doi: 10.1016/j.patcog.2012.07.021.

Sanjeev Arora and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *STOC '01: Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257, New York, NY, USA, 2001. ACM Press. ISBN 1-58113-349-9. doi: http://doi.acm.org/10.1145/380752.380808.

David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*,

SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 9780898716245.

Pranjal Awasthi, Maria-Florina Balcan, and Konstantin Voevodski. Local algorithms for interactive clustering. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 550–558, 2014.

Pranjal Awasthi, Afonso S. Bandeira, Moses Charikar, Ravishankar Krishnaswamy, Soledad Villar, and Rachel Ward. Relax, no need to round: Integrality of clustering formulations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, ITCS '15, pages 191–200, New York, NY, USA, 2015a. Association for Computing Machinery. ISBN 9781450333337. doi: 10.1145/2688073.2688116. URL https://doi.org/10.1145/2688073.2688116.

Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop. The hardness of approximation of euclidean k-means. In *31st International Symposium on Computational Geometry, SoCG 2015, June 22-25, 2015, Eindhoven, The Netherlands*, pages 754–767, 2015b. doi: 10.4230/LIPIcs.SOCG.2015.754.

Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M. Kane, Pravesh K. Kothari, and Santosh S. Vempala. Robustly learning mixtures of k arbitrary gaussians. *CoRR*, abs/2012.02119, 2020. URL https://arxiv.org/abs/2012.02119.

Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M. Kane, Pravesh K. Kothari, and Santosh S. Vempala. Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, page 1234–1247, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392648. doi: 10.1145/3519935.3519953. URL https://doi.org/10.1145/3519935.3519953.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statis-*

tics, 45(1):77 – 120, 2017. doi: 10.1214/16-AOS1435. URL https://doi.org/10.1214/16-AOS1435.

M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

Mikhail Belkin and Partha Niyogi. Convergence of laplacian eigenmaps. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 129–136. MIT Press, 2007. URL http://papers.nips.cc/paper/2989-convergence-of-laplacian-eigenmaps.pdf.

Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

Shai Ben-David and Ulrike von Luxburg. Relating clustering stability to properties of cluster boundaries. In *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 379–390, 2008.

Shai Ben-David, Ulrike von Luxburg, and David Pal. A sober look at clustering stability. In *19th Annual Conference on Learning Theory, COLT 2006*. Springer, 2006.

Shai Ben-David, Dávid Pál, and Hans Ulrich Simon. Stability of k -means clustering. In *Learning Theory, 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA, June 13-15, 2007, Proceedings*, pages 20–34, 2007.

Mira Bernstein, Vin de Silva, John C. Langford, and Josh Tennenbaum. Graph approximations to geodesics on embedded manifolds. http://web.mit.edu/cocosci/isomap/BdSLT.pdf, December 2000.

T. Berry and J. Harlim. Variable bandwidth diffusion kernels. *Applied and Computational Harmonic Analysis*, 40(1):68–96, 2016.

Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.*, 5(1):232–253, January 2011.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1517384113. URL `http://www.pnas.org/content/113/15/3932`.

Sebastien Bubeck, Marina Meilă, and Ulrike von Luxburg. How the initialization affects the stability of the k-means algorithm. Technical Report arXiv:0907.5494v1 [stat.ML], ArXiv, 2009.

Kathleen Champion, Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Data-driven discovery of coordinates and governing equations. *Proc. Natl. Acad. Sci. U. S. A.*, 116 (45):22445–22451, November 2019.

M. Charikar and S. Guha. Improved combinatorial algorithms for the facility location and k-median problems. In *40th Annual Symposium on Foundations of Computer Science*, pages 378–388, 1999.

Moses Charikar and Vaggos Chatziafratis. Approximate hierarchical clustering via sparsest cut and spreading metrics. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pages 841–854, USA, 2017. Society for Industrial and Applied Mathematics.

Guangliang Chen, Anna V. Little, and Mauro Maggioni. *Multi-Resolution Geometric Analysis for Data in High Dimensions*, pages 259–285. Birkhäuser Boston, Boston, 2013. ISBN 978-0-8176-8376-4. doi: 10.1007/978-0-8176-8376-4-13. URL `https://doi.org/10.1007/978-0-8176-8376-4-13`.

Yu-Chia Chen and Marina Meilă. Selecting the independent coordinates of manifolds with large aspect ratios. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 1086–1095. Curran Associates, Inc., 2019.

Yu-Chia Chen, James McQueen, Samson J. Koelle, Marina Meila, Stefan Chmiela, and Alexandre Tkatchenko. Modern manifold learning methods for md data – a step by step

procedural overview. www.stat.washington.edu/mmp/Papers/mlcules-arxiv.pdf, July 2019.

Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *J. Mach. Learn. Res.*, 17(1):882–938, January 2016. ISSN 1532-4435.

Stefan Chmiela, Alexandre Tkatchenko, Huziel Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, March 2017a.

Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5):e1603015, 2017b.

C. Clementi, H. Nymeyer, and J.N. Onuchic. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. *Journal of molecular biology*, 2000. says topology (of protein) more important than energy wells.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 30(1):5–30, 2006.

R. R. Coifman, S. Lafon, A. Lee, Maggioni, Warner, and Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. In *Proceedings of the National Academy of Sciences*, pages 7426–7431, 2005.

P. Constantine, E. Dow, and Q. Wang. Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014. doi: 10.1137/130916138. URL https://doi.org/10.1137/130916138.

P. Das, M. Moll, H. Stamati, L.E. Kavraki, and C. Clementi. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proceedings of the National Academy of Sciences*, 103(26):9885–9890, 2006.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, FOCS '99, page 634, USA, 1999. IEEE Computer Society. ISBN 0769504094.

Sanjoy Dasgupta. A cost function for similarity-based hierarchical clustering. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 118–127, 2016. doi: 10.1145/2897518.2897527.

Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8(7):203–226, 2007. URL `http://jmlr.org/papers/v8/dasgupta07a.html`.

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)*, 39(1):1–22, 1977. doi: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x`.

Yash Deshpande and Andrea Montanari. Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 523–562, Paris, France, 03–06 Jul 2015. PMLR.

Luc Devroye and Gábor Lugosi. *Uniform Deviation Inequalities*, pages 17–26. Springer New York, New York, NY, 2001.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians, 2020.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84, 2017. doi: 10.1109/FOCS.2017.16.

Manfredo do Carmo. *Riemannian Geometry*. Springer, 1992.

Natalie Doss, Yihong Wu, Pengkun Yang, and Harrison H. Zhou. Optimal estimation of high-dimensional gaussian mixtures. 2020.

Mathias Drton and Martyn Plummer. A bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):323–380, Feb 2017. doi: 10.1111/rssb.12187.

Carmeline J. Dsilva, Ronen Talmon, Neta Rabin, Ronald R. Coifman, and Ioannis G. Kevrekidis. Nonlinear intrinsic variables and state reconstruction in multiscale simulations. *The Journal of Chemical Physics*, 139(18):184109, 2013. doi: 10.1063/1.4828457. URL `https://doi.org/10.1063/1.4828457`.

Carmeline J Dsilva, Ronen Talmon, Ronald R Coifman, and Ioannis G Kevrekidis. Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study. *Appl. Comput. Harmon. Anal.*, 44(3):759–773, May 2018.

Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Michael I. Jordan, Martin J. Wainwright, and Bin Yu. Singularity, misspecification and the convergence rate of em. *The Annals of Statistics*, 2018.

Mojtaba Kadkhodaie Elyaderani, Swayambhoo Jain, Jeffrey Druce, Stefano Gonella, and Jarvis Haupt. Group-level support recovery guarantees for group lasso estimator. pages 4366–4370, 2017. doi: 10.1109/ICASSP.2017.7952981.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, 96(456):1348–1360, December 2001.

Harley Flanders. Differentiation under the integral sign. *The American Mathematical Monthly*, 80(6):615–627, 1973. ISSN 00029890, 19300972. URL `http://www.jstor.org/stable/2319163`.

Rong Ge, Qingqing Huang, and Sham M. Kakade. Learning mixtures of gaussians in high dimensions. STOC '15, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335362. doi: 10.1145/2746539.2746616. URL `https://doi.org/10.1145/2746539.2746616`.

C W Gear. Parameterization of non-linear manifolds, August 2012.

Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, 70(3):419–435, 2002. ISSN 03067734, 17515823. URL http://www.jstor.org/stable/1403865.

Stefan Haufe, Vadim V Nikulin, Andreas Ziehe, Klaus-Robert Müller, and Guido Nolte. Estimating vector fields using sparse basis field expansions. In D Koller, D Schuurmans, Y Bengio, and L Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 617–624. Curran Associates, Inc., 2009.

David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217 – 232, 1995. ISSN 0097-3165. doi: https://doi.org/10.1016/0097-3165(95)90052-7. URL http://www.sciencedirect.com/science/article/pii/0097316595900527.

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. From graphs to manifolds - weak and strong pointwise consistency of graph laplacians. In *Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*, pages 470–485, 2005. doi: 10.1007/11503415_32. URL http://dx.doi.org/10.1007/11503415_32.

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8: 1325–1368, 2007. URL http://dl.acm.org/citation.cfm?id=1314544.

Philippe Heinrich and Jonas Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844 – 2870, 2018. doi: 10.1214/17-AOS1641. URL https://doi.org/10.1214/17-AOS1641.

C. Hennig and T. F. Liao. How to find an appropriate clustering for mixed type variables with application to socioeconomic stratification. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 62:309–369, 2013.

Nhat Ho and XuanLong Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726 – 2755, 2016. doi: 10.1214/16-AOS1444. URL https://doi.org/10.1214/16-AOS1444.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical gaussians: Moment methods and spectral decompositions. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, page 11–20, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318594. doi: 10.1145/2422436.2422439. URL https://doi.org/10.1145/2422436.2422439.

Huan Xu, C Caramanis, and S Mannor. Sparse algorithms are not stable: A No-Free-Lunch theorem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(1):187–193, January 2012.

Tao Huang, Heng Peng, and Kun Zhang. Model selection for gaussian mixture models. *Statistica Sinica*, 27(1):147–169, 2017. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/44114365.

Takayuki Iguchi, Dustin G. Mixon, Jesse Peterson, and Soledad Villar. Probably certifiably correct k-means clustering. *Math. Program.*, 165(2):605–642, October 2017. ISSN 0025-5610. doi: 10.1007/s10107-016-1097-0.

Mary Inaba, Naoki Katoh, and Hiroshi Imai. Applications of weighted voronoi diagrams and randomization to variance-based k-clustering: (extended abstract). In *Proceedings of the Tenth Annual Symposium on Computational Geometry*, SCG '94, page 332–339, New York, NY, USA, 1994. Association for Computing Machinery. ISBN 0897916484. doi: 10.1145/177424.178042. URL https://doi.org/10.1145/177424.178042.

A. Jalali, Q. Han, I. Dumitriu, and M. Fazel. Relative density and exact recovery in heterogeneous stochastic block models. In *Proc. of NIPS 2016*, December 2016.

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J. Wainwright, and Michael I. Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results

and algorithmic consequences. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4123–4131, Red Hook, NY, USA, 2016a. Curran Associates Inc. ISBN 9781510838819.

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J. Wainwright, and Michael I. Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. *arXiv*, 1609.00978, 2016b.

Dominique Joncas, Marina Meila, and James McQueen. Improved graph laplacian via geometric Self-Consistency. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4457–4466. Curran Associates, Inc., 2017.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC '10, page 553–562, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300506. doi: 10.1145/1806689.1806765. URL `https://doi.org/10.1145/1806689.1806765`.

Daniel M. Kane. Robust learning of mixtures of gaussians. In *Proceedings of the Thirty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '21, page 1246–1258, USA, 2021. Society for Industrial and Applied Mathematics. ISBN 9781611976465.

Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. *SIAM Journal on Computing*, 38(3):1141–1156, 2008. doi: 10.1137/S0097539704445925.

D G Kendall. A survey of the statistical theory of shape. *Stat. Sci.*, 1989.

Mario Krenn, Florian Häse, Akshatkumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.*, 1(4):045024, October 2020.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. ISSN 00905364. URL http://www.jstor.org/stable/2674095.

John M. Lee. *Introduction to Smooth Manifolds*. Springer-Verlag New York, 2003.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17 NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 777–784, 2004. URL http://papers.nips.cc/paper/2577-maximum-likelihood-estimation-of-intrinsic-dimension.

Zinan Lin, Kiran Koshy Thekumparampil, Giulia C. Fanti, and Sewoong Oh. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *ICML*, pages 6127–6139, 2020. URL http://proceedings.mlr.press/v119/lin20e.html.

Allen Liu and Ankur Moitra. Settling the robust learnability of mixtures of gaussians. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 518–531, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380539. doi: 10.1145/3406325.3451084. URL https://doi.org/10.1145/3406325.3451084.

S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is np-hard. *Theoretical Computer Science*, 442:13–21, 2012. ISSN 0304-3975. doi: https://doi.org/10.1016/j.tcs.2010.05.034. URL https://www.sciencedirect.com/science/article/pii/S0304397510003269. Special Issue on the Workshop on Algorithms and Computation (WALCOM 2009).

Christian Hennig Maria Halkidi, Michalis Vazirgiannis. *Method-Independent Indices for*

Cluster Validation and Estimating the Number of Clusters, chapter 26. CRC Press, 2015. doi: 10.1201/b19706-33.

Andreas Maurer and Massimiliano Pontil. K-dimensional coding schemes in hilbert spaces. *IEEE Trans. Inf. Theor.*, 56(11):5839–5846, November 2010.

Marina Meilă. The uniqueness of a good optimum for K-means. In Andrew Moore and William Cohen, editors, *Proceedings of the International Machine Learning Conference (ICML)*, pages 625–632. International Machine Learning Society, 2006a.

Marina Meilă. The uniqueness of a good optimum for k-means. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 625–632, New York, NY, USA, 2006b. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143923. URL `https://doi.org/10.1145/1143844.1143923`.

Marina Meilă. Local equivalence of distances between clusterings – a geometric perspective. *Machine Learning*, 86(3):369–389, 2012.

Nicolai Meinshausen and Peter Bühlmann. Stability selection: Stability selection. *J. R. Stat. Soc. Series B Stat. Methodol.*, 72(4):417–473, July 2010.

Kitty Mohammed and Hariharan Narayanan. Manifold learning using kernel density estimation and local principal components analysis. *arxiv*, 1709.03615, 2017.

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL `https://proceedings.neurips.cc/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf`.

Frank Noé and Cecilia Clementi. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr. Opin. Struct. Biol.*, 43:141–147, April 2017.

Frank Noé and Cecilia Clementi. Collective variables for the study of long-time kinetics

from molecular trajectories: theory and methods. *Current Opinion in Structural Biology*, 43:141–147, 2017.

Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 2011. ISSN 00905364. URL http://www.jstor.org/stable/29783630.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, High-Performance deep learning library. December 2019.

Karl Pearson. Iii. contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. (A.)*, 185:71–110, 1894.

J Peng and Y Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM journal on optimization*, 2007.

Dominique Perraul-Joncas and Marina Meila. Non-linear dimensionality reduction: Riemannian metric estimation and the problem of geometric discovery, May 2013.

David Pollard. Strong consistency of k-means clustering. *Ann. Statist.*, 9(1):135–140, 01 1981. doi: 10.1214/aos/1176345339. URL https://doi.org/10.1214/aos/1176345339.

Nikita Puchkin and Vladimir Spokoiny. Structure-adaptive manifold estimation. June 2019.

Oded Regev and Aravindan Vijayaraghavan. On learning mixtures of well-separated gaussians. *CoRR*, abs/1710.11592, 2017. URL http://arxiv.org/abs/1710.11592.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987. ISSN 0377-0427. doi: http://dx.doi.org/10.1016/0377-0427(87)90125-7.

Sam Roweis and Lawrence Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, December 2000.

Aurko Roy and Sebastian Pokutta. Hierarchical clustering via spreading metrics. In Isabelle Guyon and Ulrike von Luxburg, editors, *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Samuel Rudy, Alessandro Alla, Steven L Brunton, and J Nathan Kutz. Data-Driven identification of parametric partial differential equations. *SIAM J. Appl. Dyn. Syst.*, 18(2): 643–660, January 2019.

Arora Sanjeev and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, STOC '01, page 247–257, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 1581133499. doi: 10.1145/380752.380808. URL `https://doi.org/10.1145/380752.380808`.

Lawrence K. Saul and Sam T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res.*, 4:119–155, December 2003. ISSN 1532-4435. doi: 10.1162/153244304322972667. URL `https://doi.org/10.1162/153244304322972667`.

Ohad Shamir and Naftali Tishby. Cluster stability for finite samples. In J. Platt, D. Koller, Y. Singer, and S. Rowseis, editors, *Advances in Neural Information Processing Systems (NIPS)*, number 21, Cambridge, MA, 2008a. MIT Press.

Ohad Shamir and Naftali Tishby. Model selection and stability in k-means clustering. In Rocco Servedio and T. Zhang, editors, *Proceedings of the 21rst Annual Conference on Learning Theory (COLT)*, 2008b. same (year) as shamir:08?

Ohad Shamir and Naftali Tishby. On the reliability of clustering stability in the large sample regime. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS)*, number 22, Cambridge, MA, 2009. MIT Press.

Ankita Shukla, Shagun Uppal, Sarthak Bhagat, Saket Anand, and Pavan Turaga. Geometry of deep generative models for disentangled representations. In *Proceedings of the 11th*

Indian Conference on Computer Vision, Graphics and Image Processing, ICVGIP 2018, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450366151. doi: 10.1145/3293353.3293422. URL https://doi.org/10.1145/3293353.3293422.

Hythem Sidky, Wei Chen, and Andrew L Ferguson. Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation. *Mol. Phys.*, 118(5): e1737742, March 2020.

A. Singer and H.-T. Wu. Vector diffusion maps and the connection laplacian. *Communications on Pure and Applied Mathematics*, 65(8):1067–1144, 2012. doi: https://doi.org/10.1002/cpa.21395. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21395.

Chaitanya Swamy. Correlation clustering: maximizing agreements via semidefinite programming. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 526–527, 2004.

Z. Szabó, B. Póczos, and A. L?rincz. Online group-structured dictionary learning. In *CVPR 2011*, pages 2865–2872, 2011. doi: 10.1109/CVPR.2011.5995712.

Yee Whye Teh and Sam T. Roweis. Automatic alignment of local representations. In *NIPS*, 2002.

Matus J Telgarsky and Sanjoy Dasgupta. Moment-based uniform deviation bounds for k-means and friends. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2940–2948. Curran Associates, Inc., 2013.

Robert Tibshirani, Guenther Walther, and Trevor 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(2):411–423, 2001. doi: https://doi.org/10.1111/1467-9868.00293. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00293.

Daniel Ting, Ling Huang, and Michael I. Jordan. An analysis of the convergence of graph laplacians. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1079–1086, 2010. URL `http://www.icml2010.org/papers/554.pdf`.

Christopher Tosh and Sanjoy Dasgupta. Maximum likelihood estimation for mixtures of spherical gaussians is np-hard. *Journal of Machine Learning Research*, 18(175):1–11, 2018. URL `http://jmlr.org/papers/v18/16-657.html`.

Gareth A. Tribello, Michele Ceriotti, and Michele Parrinello. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proceedings of the National Academy of Science, USA*, 109:5196—-201, 2012.

V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

Santosh Vempala and Grant Wang. A spectral algorithm for learning mixtures of distributions. *Journal of Computer Systems Science*, 68(4):841–860, 2004.

Ramya Korlakai Vinayak, Samet Oymak, and Babak Hassibi. Graph clustering with missing data: Convex algorithms and analysis. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2996–3004, 2014.

Ulrike von Luxburg. Clustering stability: An overview. *Foundation and Trends in Machine Learning*, 2(3):235–274, 2009.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

Yali Wan and Marina Meilă. Graph clustering: block-models and model-free results. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional em algorithm: Statistical optimization and asymptotic normality. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2521–2529, Cambridge, MA, USA, 2015. MIT Press.

Q Wu, J Guinney, M Maggioni, and S Mukherjee. Learning gradients: predictive models that infer geometry and statistical dependence. *J. Mach. Learn. Res.*, 2010.

Yihong Wu and Pengkun Yang. Optimal estimation of Gaussian mixtures via denoised method of moments. *The Annals of Statistics*, 48(4):1981 – 2007, 2020. doi: 10.1214/19-AOS1873. URL `https://doi.org/10.1214/19-AOS1873`.

Tian Xie, Arthur France-Lanord, Yanming Wang, Yang Shao-Horn, and Jeffrey C Grossman. Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials. *Nat. Commun.*, 10(1):2667, June 2019.

Eric P. Xing and Michael I. Jordan. On semidefinite relaxation for normalized k-cut and connections to spectral clustering. Technical Report UCB/CSD-03-1265, EECS Department, University of California, Berkeley, Jun 2003.

Greg Yang. Tensor programs II: Neural tangent kernel for any architecture. June 2020.

L.Q. Yang, D.F. Sun, and K.C. Toh. Sdpnal+: a majorized semismooth newton-cg augmented lagrangian method for semidefinite programming with nonnegative constraints. *Mathematical Programming Computation*, 7:331–366, 2015.

M Yuan and Y Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Series B Stat. Methodol.*, 2006.

Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Scientific Computing*, 26(1):313–338, 2004.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7: 2541–2563, 2006.

Xinyuan Zhao, Defeng Sun, and Kim-Chuan Toh. A newton-cg augmented lagrangian method for semidefinite programming. *SIAM J. Optimization*, 20:1737–1765., 2010.

Changbo Zhu, Huan Xu, Chenlei Leng, and Shuicheng Yan. Convex optimization procedure for clustering: Theoretical revisit. In *Advances in Neural Information Processing Systems 27*, pages 1619–1627, 2014.

Vladimir A. Zorich. *Mathematical Analysis I.* Springer-Verlag Berlin Heidelberg, 2004.

Hui Zou. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.*, 101(476): 1418–1429, December 2006.