HAL open science

Kernel methods for high dimensional data analysis

Alba Chiara de Vitis

To cite this version:

Alba Chiara de Vitis. Kernel methods for high dimensional data analysis. Computational Geometry [cs.CG]. COMUE Université Côte d'Azur (2015 - 2019), 2019. English. NNT : 2019AZUR4034 . tel-02419727v2

HAL Id: tel-02419727

https://theses.hal.science/tel-02419727v2

Submitted on 2 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Méthodes du noyau pour l'analyse des données de grande dimension

Alba Chiara De Vitis

INRIA SOPHIA ANTIPOLIS

Présentée en vue de l'obtention du grade de docteur en STIC **d'Université Côte d'Azur**

Dirigée par : Jean-Daniel Boissonnat
Co-encadrée par : David Cohen-Steiner
Soutenue le : 28/05/2019

Devant le jury, composé de :
Jean-Daniel Boissonnat, DR, INRIA
Frédéric Cazals, DR, INRIA
David Cohen-Steiner, CR, INRIA
Ilaria Giulini, MCF, Université Paris Diderot
Bertrand Michel,PR, Ecole Centrale de Nantes
Marco Pettini, PR, Centre de physique théorique Luminy

Inria
inventeurs du monde numérique

Méthodes du noyau pour l'analyse des données de grande dimension

Jury :

Rapporteurs:

Bertrand Michel, PR, Ecole Centrale de Nantes
Marco Pettini, PR, CPT Luminy

Examinateurs:

Ilaria Giulini, MCF, Université Paris Diderot
Frédéric Cazals, DR, INRIA Sopia Antipolis
Jean-Daniel Boissonnat, DR, INRIA Sophia Antipolis
David Cohen-Steiner, CR, INRIA Sophia Antipolis

Résumé

Méthodes du noyau pour l'analyse de données de grande dimension

Résumé. Les nouvelles technologies permettant la collecte de données dépendant d'un nombre de plus en plus important de paramètres, les ensembles de données voient leur dimension devenir de plus en plus grande. Les *problèmes théoriques*, qui dépendent notamment de la *dimension intrinsèque* de l'ensemble des données, et les *problèmes de calcul*, liés à la dimension de l'espace où vivent les données, affectent l'analyse de données en grandes dimensions. Dans cette thèse, nous étudions le problème de l'analyse de données en grandes dimensions en nous plaçant dans le cadre des espaces métriques mesurés. Nous utilisons la concentration de la mesure pour produire des outils capables de décrire la structure des ensembles de données de grandes dimensions. Nous visons à introduire un nouveau point de vue sur l'utilisation des distances et des mesures de probabilité définies sur les données. Plus précisément, nous montrons que les méthodes de noyau, déjà utilisées en petites dimensions intrinsèques pour réduire la dimensionnalité, peuvent etre utilisées en grandes dimensions et appliquées à des cas non traités dans la littérature.

Mots clés: Apprentissage de données, Concentration de la mesure, Méthodes de noyau, Grandes dimensions, Analyse des données

Kernel Methods for High Dimensional Data Analysis

Abstract. Since data are being collected using an increasing number of features, datasets are of increasingly high dimension. *Computational problems*, related to the *apparent* dimension, i.e. the dimension of the vectors used to collect data, and *theoretical problems*, which depends notably on the *effective* dimension of the dataset, the so called *intrinsic dimension*, have affected high dimensional data analysis.

In order to provide a suitable approach to data analysis in high dimensions, we introduce a more comprehensive scenario in the framework of metric measure spaces.

The aim of this thesis, is to show how to take advantage of high dimensionality phenomena in the pure high dimensional regime. In particular, we aim at introducing a new point of view in the use of distances and probability measures defined on the data set. More specifically, we want to show that *kernel methods*, already used in the intrinsic low dimensional scenario in order to reduce dimensionality, can be investigated under purely high dimensional hypotheses, and further applied to cases not covered by the literature.

Key words: Learning Mixtures, Concentration of Measure, Kernel Methods, High Dimensions, Data Analysis.

Contents

Introduction **1**

Denoising metric measure spaces. 2
Clustering high dimensional mixtures. 3
Description of chapters. 4

1 High Dimensional Data Analysis **8**

1.1 Datasets: Collection and Models 9
1.1.1 Models for the Data 10
Manifold Models. 10
Mixtures of Distributions. 10
1.2 Intrinsic Dimension 11
Embedding Dimension. 11
Intrinsic Dimension. 12
1.3 Use of Distances in Data Analysis 13
1.3.1 Divergences 13
1.3.2 Density Based Distances 14
1.4 From Distances to Topological Invariants: Topological Data Analysis 15
1.5 Nearest neighbour search 17
Local Sensitive Hashing. 17
Partition Trees. 17
Dynamic Continuous Indexing. 17
1.6 Dimensionality Reduction 18
1.6.1 Spectral Methods 18
PCA. 18
Multidimensional Scaling (MDS). 19
Kernel PCA. 19
Isomap. 20
Stochastic Neighbour Embedding. 21
Concentration of distances. 21
1.7 Clustering Methods 21
Partition methods. 21
Hierarchical Algorithms. 22
Density Based Clustering, Model Based and Grid-Based. 23

2 Metric Measure Spaces and Concentration of Measure **24**
2.1 Metric Measure Spaces 25
Isomorphism of mm-spaces. 26
2.1.1 Examples 26
Unit sphere. 26
Gaussian Spaces. 26
Molecular Datasets. 26
2.2 Concentration of Measure 27
Concentration function. 27
Concentration and Lipschitz functions. 27
2.2.1 From Isoperimetric Inequalities to Concentration 28
Concentration Functions. 29
Concentration on the Sphere. 30
Concentration for the Gaussian Space. 30
Concentration of Measure for Log-concave Functions 31
2.3 Emptiness of Space 32
Volume of the Unit Sphere 32
Volume of the Cube 33

3 The Distance Transform of a Metric Measure Space **34**
3.1 Distances Between Metric Measures Spaces: $\mathbb{D}$-distance 35
Isomorphism of mm-spaces. 35
Wasserstein Distances. 35
$\mathbb{D}$-distance. 36
3.2 Distance Between Metric Measure Spaces: Gromov-Wasserstein Distance 37
Correspondences. 37
Gromov-Hausdorff and Gromov-Wasserstein distance 37
3.3 Denoising 38
3.3.1 A transform on metric measure spaces 39
3.3.2 Stability 39
3.3.3 Behavior with respect to products 40
3.3.4 Distortion bound for simple spaces 42

4 Spectral Properties of Radial Kernels and Clustering in High Dimensions **44**
4.1 Introduction 44
4.2 Kernels in high dimensions 47
4.2.1 Main result 49
4.2.2 Distance matrices 50
4.3 Positive definite kernels and clustering 52
4.4 Covariance based clustering 54
4.5 Proofs 57
4.5.1 Proof of Theorem 4.1 57
4.5.1.1 A property of $\phi_{\bar{\mu}}$ 59
4.5.1.2 Decomposition of $\Phi_h(X)$ 63
4.5.1.3 Sample size 65

4.5.2 Proof of Proposition 4.3 66
4.5.3 Proof of Corollary 4.6 69
4.5.4 Proof of Corollary 4.5 69
4.5.5 Proof of Theorem 4.6 71
4.5.6 Proof of Theorem 4.7 74

Bibliography **80**

Introduction

Since new technologies enabled the collection of data using a larger and larger number of features, datasets have become more and more *high dimensional*. To give only a few examples: a larger number of camera sensors gives higher resolution in pictures, and consequently datasets of high dimension; large memory supplies enable to store internet datasets using a large number of features, medical data using a large number of parameters, etc [1], [2]; genome investigation [3],[4], proteins interaction analysis [5] and molecular simulations produce large datasets in high dimensional spaces. More prominently, high dimensionality can be considered as a feature of Big Data [6]. Usually, high dimensionality is not included among the characterizing features of Big Data, *i.e. Volume, Variety, Velocity, Veracity, Value*, but it appears as a consequence of the variety requirement and the parallel evolution of the size of data and of the number of data parameters. In fact, in order to obtain high variety, many features of the data must be collected. Conversely, a sample in high dimension is supposed to have a large enough size, which implies that quite often high dimensional datasets have a very large volume. However, relatively small datasets can have a high dimension. e.g. short-time experimental data from neurosciences [7], [8].

High dimensionality can be either an intrinsic feature of the dataset or due to redundant features and noise. To decide in which regime one should put the problem under consideration, usually the *intrinsic dimension* of the data space has to be taken into account. Here *intrinsic dimension* refers to the actual dimension of the dataset, i.e. the number of independent parameters needed to represent the data. More specifically the intrinsic dimension of a random vector X can be defined as the topological dimension of the support of X. For a metric space, the intrinsic dimension can be defined in terms of the *doubling constant*

of the space: the minimum λ such that every ball can be covered by λ balls of half the radius, and corresponds to the *doubling dimension* of X, defined as $dim(X) = \log_2 \lambda$. This value gives also a lower bound on the minimum dimension for which a metric space can be embedded in a normed space, with low distortion [9]. If the dataset lies on a manifold, the intrinsic dimension of the dataset is the *topological dimension* of the manifold [10], i.e. for a dataset it can be estimated in term of the neighbourhood structure. In general, if the dataset has constant dimension, the intrinsic dimension of the space can be defined *globally* for the entire dataset, e.g. the *topological dimension* for a dataset sampled from a manifold, otherwise it has to be defined on *local* neighbourhoods for which the dimension can be considered constant. Intrinsically high dimensional spaces experience a class of high dimensional phenomena, *e.g. concentration of distances and emptiness of space*, that make many low dimensional methods not effective. In fact, as observed *e.g.* by Beyer et al [11] nearest neighbor do not carry meaningful information for typical high-dimensional data, as the ratio between the variance of the distance between points and the mean value of the distance generally tends to 0. From a practical point of view, for dimensions as low as 10, methods based on *k-NN search* may be severely affected by high dimensionality. Many methods have been proposed to reduce the dimensionality of the data, in order to avoid problems related to the high dimensionality of the ambient space. Classical *spectral methods* have been proved to be successful in providing target embedding spaces, e.g. principal component analysis (PCA) for linear embeddings, Isomap and Kernel PCA for non linear datasets [12], [13], [14]. However such methods have their shortcomings and do not allow to successfully address all the problems that arise in high dimensions.

In this Thesis, we study the problem of high dimensional data analysis in the framework of metric measure spaces [15]. We take advantage of concentration of measure to produce tools able to describe the structure of datasets in high dimensions, using suitably defined Lipschitz maps. We will describe two contributions:

Denoising metric measure spaces. In chapter 3, we suggest an approach, which turns out to be related to the algorithm proposed in [16], for reducing noise

in metric measure spaces. The procedure, which we call the distance transform, consists in embedding the metric measure space into the space of L^p functions defined on it by means of distance functions. A first property of this transform is its stability with respect to perturbations of the input metric measure space. Such perturbations may be quantified using a distance on the set of metric measure spaces called the Gromov-Wasserstein distance. We show that the distance transform is 2-Lipschitz for this distance. The main purpose of this transform however is its noise reduction properties. Corruption by noise may be defined in several inequivalent ways for a metric measure space. We consider the simple case where the noisy space is obtained from the ideal one by taking its product with a space enjoying the concentration of measure property. In this case, we show that the effects of noise are considerably reduced by the distance transform, especially if the "noise space" is high dimensional. While the transform does however affect the ideal space as well, we argue that these effects remain small when the ideal space has simple enough geometry.

Clustering high dimensional mixtures. A fundamental problem in data analysis is to cluster mixture models. For that problem to be well-defined, one needs to make a priori assumptions on the components of mixture. In the context of high dimensional data, a natural assumption is that the components satisfy the concentration of measure property. Indeed, many classical basic distributions are known to satisfy it, and according to the famous KLS conjecture [17], the large class of log-concave distributions also does, assuming a bound on their covariance matrix. For this clustering problem, one case is solved in a relatively easy way: When the centers of the components of mixtures are sufficiently far apart, it is enough to perform a PCA-based dimensionality reduction step, and then to apply an off-the-shelf clustering algorithm such as k-means. The case where the centers of the components are close or even equal is more difficult. The best algorithm for that problem [18] uses tensor decomposition algorithms on the degree 6 moments of the mixture. While its guarantees are strong, the correctness of this algorithm crucially depends on Isserli's theorem, a specific identity between moments that holds only for Gaussian components. However, assuming that the components are exactly Guassian probably isn't very realistic

in practice. Furthermore, the computational complexity is at least the dimension n to the sixth power, making it impratical when n exceeds, say a hundred.

We consider the approach to the clustering based on radial kernels. Our main technical result is that for a mixture of distributions that concentrate, such kernel matrices can be written as the sum of a blockwise row constant matrix and a blockwise column constant matrix, up to a small error term. For distance matrices, in the case of single component, this result implies that the ratio between the first and the second singular values is of the order of the dimension n, rather than $\sqrt{n}$ as one might naively expect from a basic application of concentration. This "second order" concentration phenomenon can be used to show guarantees for kernel PCA based clustering approaches. We further show that for points lying on a sphere, the conclusions of our first result can be strengthened, in the sense that kernel matrices are now well approximated by block constant matrices. We introduce a specific radial kernel, and an associated spectral algorithm that is able to cluster mixtures of concentric distributions provided their covariance matrices are far enough. Analysis shows that the required angular separation between the components covariance matrices tends to 0 as the dimension goes to infinity. To the best of our knowledge, this is the first polynomial time algorithm for clustering such mixtures beyond the Gaussian case.

Description of chapters. We now describe in more detail the contents of each chapter. In Chapter 1 we give an overview of basic geometric techniques in data analysis and discuss their limits for data with high intrinsic dimension. Chapter 2 focuses on *metric measure spaces*, as a suitable framework for high dimensional data analysis, and on concentration of measure phenomena. In Chapter 3 we present our denoising approach for metric measure spaces, and study some of its theoretical properties. The last chapter of the thesis is about our work on radial kernel matrices and high-dimensional clustering, which in fact came out as an elaboration of the ideas in Chapter 3. It is a verbatim retranscription of a dedicated article to be submitted for publication.

Acknowledgements

My acknowledgments are due to many people who contributed in many ways to this thesis.
First of all my thanks go to my advisors, Jean-Daniel Boissonnat and David Cohen-Steiner. They trusted in this project since the very beginning. I wish to thank Jean-Daniel Boissonnat for having given me the opportunity to work on this subject. We spent so much time searching intensely at the crossroads between geometry and data analysis, and he gave me all his support for the choice that led to this thesis, giving me the possibility to find a topic that perfectly matched my research interests. I wish to thank David Cohen-Steiner for having accepted to follow me in this project. We spent these years in an intense research exploring connections between concentration of measure phenomena and high dimensional data analysis. Many times we encountered problems, many times we have found a way out. His tenacity and lively intellectual tension made it possible to overcome the problems we faced. Not only my acknowledgements, but also my gratitude goes to Jean-Daniel and David, since their support never faded, and led to the intense research journey of this thesis. And my personal debt goes beyond the actual results of the PhD program, because they conveyed something about scientific research that is now part of my personal way of being.
I would like also to thank Michel Bertrand and Marco Pettini for having accepted to be referees for this thesis, and Frédéric Cazals and Ilaria Giulini for having accepted to be part of the jury de thèse.

Many other people have to be mentioned here. I want to thank again Frédéric Cazals for having given me the opportunity to enter in contact with INRIA, having taken part to the first stages of this thesis, and having spent time in very

interesting conversations about applications of data analysis to molecular problems; Mariette Yvinec for having taken part to the analysis of the thesis' topics in my first year; Nathalie Gayraud and Emanuela Merelli for our interesting conversations about application of high dimensional data analysis to neurosciences; Clément Maria, for our interesting conversations about persistent homology in high dimensions; Mathijs Wintraecken and Siargey Kachanovich for our interesting conversation about high dimensional problems.
My thanks to them go in fact far beyond. I want to thanks Clément, Ross, Manish and Deepesh for having been such great PhD fellows in my first year at INRIA, and Mathijs, Siargey and Mael for having being such great friends in our years spent together.
I wish to thank all the people that made these years so interesting: Remy, Romain, Simon, Alix, Pierre, Yuliya, Mathieu, David, Dorothy, Dorian, Alfredo, Simon, Clément, Kunal, Pawel, Harry, Nathalie, Siddarth, Arijit and Ramsey.
Many thanks go also to Florence Barbara, for her support and friendship in my period spent at Inria.
I would also like to thank all the Geometrica, ABS, Titane and DataShape teams, and INRIA for having provided such a greatly active research environment. This thesis was partially supported by the ECR Advanced Grant GUDHI.

My personal thanks goes also to all the people that supported me in these intense part of my life, my family, my mother, my sister Eliana, my grandmother Donata, my friends, my collaborators.
Special thanks to Maria, Clarissa and my mother for stressing me so much about this Thesis; Fabio, Lory, Marco, Giulio, Elisa, Emilia e Francesco, Chiara, Sara, Diana, Cinzia, Marianna, Giovanni, Eleonora, Ginevra, Ilaria, Marco, Giorgio, Paolo, Marina, Leonardo e Marina, Leonardo e Kora, Maria Consiglia, Elvira, Maria Antonietta, Francesca.

I wish to thank all the people that in these years never let me give up. Nice and Paris for having made me feel home. And again, all the lights, white stones, sand and the immortal sea of my hometown for giving me the strength for changing language and country, again and again. Florence, for being so beautiful and

timeless.

Thanks to Gabriele, for having made possible to me to be myself, since ever, and always more. And for being so special every day.

Chapter 1

High Dimensional Data Analysis

Since computer power enables massive computations, data analysis has been driven by the necessity to produce algorithms able to recover the structure of datasets from input points, e.g. using manifold reconstruction and metric approximation [19]. Many methods developed in the last decades use extensively distances and metrics in order to produce structures able to capture the manifold underlying the dataset, e.g. Delaunay triangulations [20], marching cubes [21], manifold reconstruction [22]. Moreover, many of those methods strongly rely on a partition of the space, e.g. Voronoi diagrams in order to produce Delaunay triangulations [20], kd-trees [23] and for nearest neighbour search. These methods and techniques (e.g. use of distances, partition of the space, nearest-neighbour search) are affected in high-dimensions by the *curse of dimensionality* [24], and consequently, algorithms based on them may be not efficient. The term *curse of dimensionality*, introduced by Bellman in 1961 [25], [26] is nowadays used to refer to the class of phenomena that occur in high dimensions in contrast with the low dimensional scenario. Important examples are the tendency of data to become very sparse in high dimensions [24], [27], and the concentration of distances. Usually dimensions $d \leq 6$ are considered low. A high dimensional regime has to be considered when dimension $d \geq 10$ [11].

In general, *high dimensional data analysis* relies on the strong hypothesis that, in practice, datasets have low *intrinsic dimension*, largely supported by observation (§ 1.2), and that their dimension is only *apparently* high. As a consequence, many dimensionality reduction methods have been proposed. These methods aim at reducing the dimensionality of the dataset by embedding the original dataset into a lower dimensional space [28], [29], [12]. Anytime that a simplicial complex structure can be defined on the data, the use of topological tools, like Topological Data Analysis (TDA), has also been proved to be effective in order to recover the structure of the dataset. In this chapter, we give a brief overview of the methods and techniques mostly used in high dimensional data analysis in order to reduce dimensionality.

Toward the blessing of dimensionality. On the other hand, when the *actual* dimension is high, these methods may not be applied directly, or may give only approximation of the structure of the dataset. This happens, for instance, in presence of very large noise, or for datasets that are described by a very large number of parameters (e.g. molecular datasets).
Our aim in this Thesis, is to show how to take advantage of high dimensionality phenomena in the high dimensional regime. We aim at introducing a new point of view in the use of distances and probability measures defined on the data set. This approach, i.e. the possibility of using in a good way high dimensional phenomena, is sometimes referred to as *blessing of dimensionality*. More specifically, we want to show that *kernel methods*, already used in the intrinsic low dimensional scenario in order to produce dimensionality reduction, can be investigated under purely high dimensional hypotheses, and further applied to cases not covered in the literature.

1.1 Datasets: Collection and Models

A *dataset* of dimension n and size N is a collection of N items (points) with relations, often structured in matrices $N \times n$, where the rows represent the data entries and each column represents a variable corresponding to a feature used to collect the data. In the *probabilistic and statistical* approach, an n-dimensional

dataset is the result of *sampling* N points from a *distribution* in $\mathbb{R}^n$, which gives relations between points, where each entry of the dataset is a sample point. In the *geometric model*, relations between points can be given in terms of *distances*, an the dataset can be described as a finite metric space of size N and dimension n. Data analysis relies on the fact that not all the features used to collect the data are relevant, and that the relations between the data points can be given in terms of a lower number of features. Usually this is formalized by introducing the definition of *intrinsic dimension* of the dataset (§ 1.2).

1.1.1 Models for the Data

A dataset is usually modelled according to the mathematical structure that better encodes its properties.

Manifold Models. In the geometric model, the most relevant criterion to produce a dataset description takes into account the existence of a manifold that approximate the dataset. In this case, several manifold recovering methods have been proposed, which strongly rely on the use of distances defined on the datasets. Many methods used extensively distances and metrics to produce structures able to capture the manifold underlying the dataset, e.g. Delaunay triangulations [20, 22], marching cubes [21]. These methods strongly rely on a partition of the space, e.g. Voronoi diagrams in order to produce the Delaunay triangulation [20] or kd-tree[23].
In most cases, the data points cannot be assumed to lie on a manifold due, for example, to the presence of noise that may destroy the geometrical structure. However, if the amplitude of the noise is small and the data points remain close enough to a manifold, the data set can be represented as complexes (*e.g. simplicial*) from which the topological and possibly geometrical features of the data can be extracted.

Mixtures of Distributions. When data are sampled from a probability distribution (e.g. configuration space of a molecule), the existence of a manifold approximating the dataset is not always satisfied or the sample cannot be regular

enough to guarantee for manifold reconstruction.
In the case of data sampled from several probability distributions, where no claim can be made on the underlying geometrical structure or on the support of the distributions, data are described as *mixtures*. The dataset is, in fact, described as the union of subsets (components), where each component is sampled from a different distribution. The most effective approach, in this case, is to classify the points according to the component they are sampled from.

The two approaches are not necessary distinct, and the metric and probability approach can be combined in the framework of *metric measure spaces* (Chapter 2).

1.2 Intrinsic Dimension

The *ambient dimension* n of the dataset is given by the number of variables used to collect the data, and the *size* N is given by the number of entries in the dataset. The aim of high dimensional data analysis is to define the *effective dimension* of a dataset, in terms of the relevant features that describe it.

In the high dimensional regime, dimensionality reduction methods aim at finding a lower dimensional target space in order to provide an embedding for the dataset. In this section, we recall some results on *metric* dimension and *intrinsic* dimension with the aim of formalizing the dimensionality reduction problem.

Embedding Dimension. For two metric spaces (X, d_X),(Y, d_Y), an injective map $f : X \to Y$ is an embedding of X into Y. The *distortion* is defined as $max_{(u,v)\in X} \frac{dist_Y(f(u),f(v))}{d_X(u,v)}$. The *metric dimension* refers to the dimension of the target real normed space in which a dataset can be mapped with low distortion. In particular, since datasets can be seen as metric spaces, the dimension of the embedding in Euclidean metric spaces has been proven to depend on the number of points in the datasets. For small data set (N points) in high dimensional spaces, a lower dimensional embedding can be provided [9]. Bourgain [30] proved that every metric space with N points can be embedded in $O(\log^2 N)$-dimensional Euclidean space with a distortion of $O(\log N)$. The

Johnson-Lindenstrauss lemma [31] states that when the input metric space is Euclidean, a $O(\log N)$ embedding dimension is enough to achieve arbitrarily small distortion.

Intrinsic Dimension. The *intrinsic dimension* refers to the actual dimension of the dataset, i.e. the number of independent parameters needed to represent the dataset. More specifically, the intrinsic dimension of a random vector X can be defined as the topological dimension of the support of X. If the dataset lies on a manifold, the intrinsic dimension of the dataset can be defined as the topological dimension of the manifold. For a metric space X, the intrinsic dimension can be defined in terms of the *doubling constant* of the space: i.e. the minimum λ such that every ball can be covered by λ balls of half the radius. Then the *doubling dimension* of X, is defined as $dim(X) = \log_2 \lambda$. A small doubling dimension allows low distortion embeddings in smaller spaces. In fact, $O(dim(X))$ dimensions are enough to get $O(\log^{1+\epsilon} n)$ distortion for any $\epsilon > 0$[9]. Several attempts have been made to generalize this definition of intrinsic dimension [32]. Moreover, the connection between sparsity and intrinsic dimensionality of the data has been addressed in research on nearest neighbour search [33].

A common assumption is that although datasets lie in high dimensional ambient spaces, their intrinsic dimension is often low. This assumption is largely supported by observation. As a consequence, many dimensionality reduction methods have been proposed. These methods aim at reducing the dimensionality of the dataset by embedding the original dataset in a lower dimensional space. Methods in this class have very good performance when the above assumption holds [28], [29], [12]. Such methods avoid problems related to high dimensionality such as the exponential size of space subdivisions and the sparsity of data in high dimensions. Topological tools, like those used in Topological Data Analysis (TDA), can also be effective in recovering the low dimensional structure of the dataset.

1.3 Use of Distances in Data Analysis

Distances are ubiquitous in data analysis and are the most natural descriptors for a dataset. Often datasets are collected as matrices of coordinates of points and the Euclidean distance or Minkowski distances are used to measure the distance between points. When data are collected with reference to the distribution they are sampled from, we may prefer to use *pseudo distances*, such as KL-divergences (pseudo-distance on the space of probability distributions), or other similarity measures (pseudo-distance on the dataset) to evaluate the similarity between datasets or data points.

1.3.1 Divergences

If the dataset can be modelled in terms of mixtures of distributions, effective methods that produce good classification are based on *divergences* between distributions instead of distance between points. KL-divergence gives a measure of how a distribution diverges from an expected distribution. In the continuous case, for two distributions P and Q, the Kullback-Leibler divergence can be defined as

$$D_{KL}(P\|Q) = \int_{-\infty}^{+\infty} p(x)\frac{p(x)}{q(x)}dx$$

and, for the discrete case,

$$D_{KL}(P\|Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)}.$$

The KL-divergence is not symmetric and does not satisfy the triangular inequality. It is therefore not a distance. Nevertheless, it has found many applications and several generalization have been proposed. A notable application of KL-divergences can be found in Stochastic Neighbour Embedding (SNE) methods for dimensionality reduction where it is defined to define the cost function to be optimized.

1.3.2 Density Based Distances

Density based distances provide a similarity measure that can be used to perform dimensionality reduction and clustering. They take into account the distribution of the points and not only the distance. As an example, we can consider a dataset sampled from a mixture of two distributions that results in two groups of points, clearly separated by an empty strip. Points that are at the same Euclidean distance from a query point but belong to regions with different densities may be expected to be assigned to different classes.

Definition 1.1. Let $f(x)$ be a probability density function in $\mathbb{R}^n$, a *density based measure* in $\mathbb{R}^n$ may be defined as a path-length measure that assigns short lengths to paths through high density regions and longer lengths to path passing through low density regions:

$$\mathcal{M}_f(x_1 \rightsquigarrow^{\gamma} x_2) = \int_0^1 g(f(\gamma(t))\|\gamma'(t)\|_p dt$$

where $\gamma : [0.1] \to \mathbb{R}^d$ is a continuous path from $\gamma(0) = x_1$ to $\gamma(1) = x_2$ and $g : \mathbb{R}_+ \to \mathbb{R}$ is a monotonically decreasing function (e.g. $g(u) = 1/u$).

$\mathcal{M}_f$ provides a density based measure of path length, and a **density based distance** may be defined as

$$\mathcal{D}_f(x_1, x_2) = \inf_{\gamma} \mathcal{M}_f(x_1 \rightsquigarrow^{\gamma} x_2)$$

If the density function is not given, a natural way to define a density function is in term of the inverse of distance to the k-nearest neighbours around a query point.

Density based distances allow to take into account probability distributions. By associating a weight to each point of the dataset, they provide a more complete description compared to methods that rely only on metric properties of the space. As a result, they allow to solve cases that are not accessible to a simple metric analysis.

Density based methods and low dimensional hypothesis. However, density based methods quite often rely on the assumption that the dataset is of low intrinsic dimension. Indeed, low dimensional density based methods strongly rely on distances to compute k-nearest neighbours and estimate density, which may cause difficulties. On the one hand, computing nearest neighbours become intrinsically difficult (§ 1.9) and, on the other hand, the number of points needed to perform calculations grows very rapidly with n. In practice, data may be very sparse and methods like *Kernel Density Estimator*(KDE) or *Parzen Windows* may fail in high dimensions [34] and widely used algorithms, such as DBSCAN [35], may experience problems.

Deep Learning and Density Estimation. Recently, high dimensional density estimation has been performed using *deep learning algorithms*. These algorithms may recover the density distribution on the observed dataset by minimizing the KL-divergence between the empirical distribution mapped in the new spaces (representation space) and a good distribution chosen *a priori*. In this approach, minimizing the entropy of the density distribution may provide a factorization in the latent space [36]. However, for these methods and related algorithms, a reliable proof cannot always be given.

1.4 From Distances to Topological Invariants: Topological Data Analysis

Computational topology offers methods that are complementary to the aforementioned methods based on distances. Topological Data Analysis provides new descriptors of datasets able to encode high dimensional properties and to describe the structure of the datasets. In particular, TDA focuses on the *homological properties* of a dataset and topological invariants are used in order to recover the intrinsic dimensionality of the dataset.

In topology, spaces are classified according to the following equivalence relations: two spaces are *homeomorphic* if there exist a continuous invertible transformation

with continuous inverse between them; a weaker condition is to have the same *homotopy* type, i.e. one space can be continuously deformed into the other; an even weaker notion, which is a kind of linearization of homotopy, is the notion of *homology* that classifies topological spaces in terms of the number of holes. For a space X, computing the number of holes (of any dimension) reduces to computing the dimension of its homology groups.[1] Being the weaker equivalence, the homological classification is the coarsest one and the easiest to achieve.

Persistent homology [37] is a multi-scale variant of homology. It aims at counting holes and recording how long they live when the space is looked upon at increasing resolution. More precisely, one associates to a given dataset a simplicial complex K (which is a topological space) together with a filtration, i.e. a nested sequence of simplicial complexes associated to various values of a scale parameter: $K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$. A *barcode* (or persistent diagram) registers the appearance and the disappearance of the generators of the persistent homology with respect to the value of the scale parameter. TDA is based on the principle that the topological features of the dataset that are associated to long lasting holes are meaningful and distinguish from the short ones that are related to noise. Bar codes are a new type of descriptors that can be used to further analyze the data [38], [39], [40], [41],[42]. TDA hence provides a general framework to analyze data in a manner that is insensitive to the particular metric chosen and provides dimensionality reduction and robustness to noise. Moreover, the use of Vietoris-Rips complexes and efficient computational tools [43] made the algorithms efficient.

The drawback of these methods, however, is that, even if they are robust to small noise, they may not resolve cases in which the noise is larger or comparable to the intrinsic geometrical features of the dataset. In those cases, due to intrinsically high dimensional spaces, the TDA approach is not effective.

[1] Homology groups with coefficients in a field are vector spaces and by dimension here we consider the dimension of the vector spaces.

1.5 Nearest neighbour search

Nearest neighbour search is a fundamental problem whose exact solution can easily be obtained by comparing the distance of the query point to all the points in the data set. Beating this trivial algorithm is not easy in high dimensions and in fact reporting the true nearest neighbour in sublinear time requires a data structure of size exponential in the dimension of the space. This fact leads to consider various kinds of approximations.

Local Sensitive Hashing. *Local Sensitive Hashing* (LSH) is an approach to report all points that are within a distance r from a query point using hash functions [44], [45]. Hash functions are defined so that nearby points have a high probability of receiving the same hash code (the value assigned by the hash function) and to be collected in the same bucket. For a query point s, all points in a bucket are retrieved as *near* points to s. Since computing hash codes turns to be fast, and the number of points in a bucket is much smaller than the total number of points, LSH is quite efficient: for datasets in $\mathbb{R}^n$, the algorithm depends only polynomially on the ambient dimension n. The main issue in LSH methods is to define suitable hash functions [46], [47], [48], [49]. Particularly efficient LSH methods can be obtained when the intrinsic dimension is low. However, high intrinsic dimensionality results in difficulty in indexing, and poor performance [46] (§ 1.9), [50].

Partition Trees. An effective partition of the space has been given by Das-Gupta at al. [32]. The proposed structure, called the *random projection tree* (RP tree), is inspired from the *kd*-tree. It is produced by splitting the space along random directions instead of the coordinate directions and allowing splitting points that are not median points. RP trees have the notable property to adapt to the intrinsic dimension of the space.

Dynamic Continuous Indexing. In recent work, Li and Malik introduced a new approach named *dynamic continuous indexing* (DCI) [50] which improves

on the k-nearest neighbour search in high dimensions. DCI reduces the dependence on the dimension of the ambient space from exponential to linear. In the *prioritized DCI*, they manage to further improve on the query time.

1.6 Dimensionality Reduction

In general, reducing dimensionality is the best approach to tame high dimensional datasets when the intrinsic dimension is low enough to allow for good low dimensional embedding. It consists in reducing the number of parameters that represent a dataset in order to obtain a faithful and meaning representation of the dataset in a lower dimensional space. Dimensionality reduction has been addressed from various points of view. Geometrical aspects of the subject include *topological data analysis*, *manifold reconstruction* and *data embedding*. In statistics, dimension recovery is related to *multivariate density estimation*. Dimensionality reduction is also related to *feature extraction* in *pattern recognition*, and to *data compressing* and *encoding* in information theory [12], [13].

Linear and non-linear techniques have been proposed for dimensionality reduction. Linear methods, including principal component analysis (PCA), factor analysis and classical scaling, are the most used, while non-linear methods have the capability to adapt to non linear data. The latter class includes isometric mapping, kernel PCA, multi-dimensional scaling (MDS), locally linear embedding and its variants, Laplacian eigenmaps, diffusion maps [13].

1.6.1 Spectral Methods

This class of methods includes some of the most effective dimensionality reduction methods. In practice, spectral methods provide dimensional reduction using projection on eigenspaces associated to eigenvalues of a suited designed matrix.

PCA. PCA gives a low dimensional representation of the data, using a linear basis that consists of the directions that maximize the variance of the data. PCA

computes the covariance matrix $cov(X)$ of the dataset and solves the eigenproblem

$$cov(X)v = \lambda v$$

The eigenvalues of $\mathcal{C}$ provide a measure of the variance of the high dimensional data set along the principal axes (eigenvectors). The top m eigenvectors of the matrix, $(e_1, ...e_m)$, provide a basis for the embedding space. This space turns out to minimize the reconstruction error between the input dataset and the projection in the m-dimensional subspace

$$E_{PCA} = \sum_i \|x_i - \sum_{\alpha=1}^{m} (x_i \cdot e_\alpha) e_\alpha\|^2$$

The output of PCA is the projection of the input dataset on the m-dimensional subspace spanned by the top m eigenvectors of $\mathcal{C}$. PCA is the most widely used and effective method for dimensionality reduction.

Multidimensional Scaling (MDS). In multidimensional scaling the input is the distance matrix M of the pairwise Euclidean distances between points in the dataset. The core of the method uses PCA on M in order to provide the low dimensional embedding.

Kernel PCA. PCA provides a best *linear* embedding of the dataset. However, the hypothesis that the dataset lies on a linear subspace of the observation space is actually too strong in many cases. A notable example is a curve in a high dimensional space that lie on a 2-dimensional plane. The curve can be efficiently embedded in a 2-dimensional space but not in a 1-dimensional space. To recover the intrinsic 1-dimensional structure of the curve, one can consider a non-linear embedding such as the one provided by kernel PCA [10].

Let $\{x_1, ...x_N\}$ be a set of N n-dimensional vectors in $\mathbb{R}^n$ (the data points) and let k be a positive kernel function. The kernel function defines a $N \times N$ symmetric matrix M

$$M_{ij} = k(x_i, x_j)$$

called the kernel matrix. Kernel PCA consists in applying PCA to the positive kernel matrix.In order to analyse precisely kernel PCA in high dimensions, it is necessary to understand the behaviour of kernel matrices and of their spectra as the dimension increases. While the literature on eigenvalues of random matrices is vast and growing rapidly, the knowledge about random kernel matrices is not growing at the same rate although important contributions have been made [51], [52].

Isomap. When the dataset live on a non linear submanifold, using the standard Euclidean distance in the ambient space may not allow to correctly identify the neighbours of a data point on the submanifold. To overcome this difficulty, Isomap, as well as other graph-based methods, constructs a graph structure from the dataset where the *nodes* are the input data points and the *edges* represent neighbourhood relations. The pairwise distances between the points are used as an approximation of the geodesic distance between the points on the manifold. The length of the shortest path joining two nodes on the graph then provides an approximation of the geodesic distance on the manifold. The algorithm consists of three parts:

- Construct a proximity graph whose nodes are the points of the dataset and whose edges connect points that are close. Assign weights to the edges based on the Euclidean distances between the points.
- For those pairs of points that are not in the proximity graph, compute their distance as the length of the shortest path between the points in the proximity graph (Dijkstra's Algorithms),
- Applying MDS on the computed interpoint distances.

A problem of Isomap is its instability against noise. Indeed, noise and outliers may produce erroneous links in the graph that may lead to topological errors. Also, while its correctness has been proved for submanifolds isometric to convex subsets of Euclidean spaces, it remains a heuristic in more general cases.

Stochastic Neighbour Embedding. Stochastic neighbour embedding and its variants compute an embedding of the data in a lower dimensional space so as to minimize the similarity between the distribution of the pairwise distances between the points in the dataset in the high dimensional space and the corresponding distribution measured in the lower dimensional space [53]. The similarity between the two distributions is usually measured using the KL-*divergence* [54].

Concentration of distances. Besides the computational difficulty of computing nearest neighbors in high dimensions, it turns out their relevance become increasingly unclear as the intrinsic dimension of the data grows. Indeed, it was noticed by practitioners that the ratio distances to the nearest neighbor and distances to the furthest point often tend to 1, implying that the nearest neighbor might be determined by the sample's randomness. This phenomenon is referred to as *concentration of distances* and strongly affects searching and indexing high dimensional datasets [55], [56]. Concentration of distances is part of a large class of higher dimensional phenomena usually referred to as *concentration of measure*, which we will investigate in more detail in Chapter 2.

1.7 Clustering Methods

A first step in recovering the structure of a dataset is to identify clusters. We give in the following section a brief overview of clustering methods. As is common, we distinguish between partition methods and hierarchical methods [57].

Partition methods. In partition algorithms, a set of N data points is partitioned in k clusters. The number of clusters k is given as part of the input and remains fixed. At each iteration, the algorithm reassign datapoints among clusters. The algorithm starts with an initial partition which is further improved so as to optimize an objective function. The celebrated k-means algorithms is one of the most widely used clustering algorithm. Each cluster is represented by its center of gravity (or mean), i.e. the point that minimizes the sum of squared

distances to the elements in the cluster. Given a set $X = \{x_1, ..., x_n\}$ of means $\{m_1, ..., m_k\}$ the algorithm consists of the two following steps

- *Expectation:* at step t, each data point is assigned to the cluster whose mean is closest. Hence, the clusters $S_1^t, ..., S_k^t$ are defined as the Voronoi partition

$$S_i^t = \{x \in X : \|x - m_i^t\|^2 \leq \|x - m_j^t\|^2 1 \leq j \leq k\}$$

- *Maximisation:* new means $\{m_1^{t+1}, ..., m_k^{t+1}\}$ are computed according to the new partition

$$m_i^{t+1} = \frac{1}{|S_i^t|} \sum_{x_j \in S_i^t} x_j$$

While the algorithm converges to a local optimum, there is no guarantee that it will converge to a global optimum. Moreover, if one uses Euclidean distances as described above, the cost functional will minimize the standard cluster variance, implying that the algorithm may fail to disciver anisotropic clusters.

Hierarchical Algorithms. In hierarchical algorithms, the clusters are produced by grouping clusters in bottom-up fashion (agglomerative clustering), at each step combining two clusters that are similar. At the beginning of the process, each element is in a cluster of its own. The clusters are then sequentially combined into larger clusters until all elements end up being in the same cluster. The result of the clustering can be visualized as a dendrogram, which shows the sequence of cluster fusion and the distance at which each fusion took place.

The grouping rule can be guided by several similarity measures. Single-linkage clustering consists in merging the two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other. In complete linkage clustering, the distance between two clusters is given by the greatest distance between any points in the clusters. In average linkage clustering, similarity is measured as the average distance between the points of the two clusters.

Density Based Clustering, Model Based and Grid-Based. Since clustering is based on grouping objects according to some common features, two classes of measures are usually used: distance based measures and similarity measures. Han and Kamber (2001) further proposed to use three approaches: density based methods, model based clustering and grid-based methods. In model-based clustering, a theoretical model is provided and optimized so that it best fits the data. In grid based models, the space is partitioned into a finite number of cells. All these methods are affected by high dimensionality and experience concentration of volume and emptiness of the space, leading to poor performance.

Chapter 2

Metric Measure Spaces and Concentration of Measure

Metric spaces are a natural structure used to describe datasets. More specifically, sample points in a dataset often come with a natural notion of distance, turning it into a metric space. Associating (for example equal) weights to the data points further turns the dataset into a *metric measure space* [58], [15]. The arsenal of mathematical results and techniques for metric measure spaces can then be leveraged to unveil and investigate the hidden structure of the dataset. In particular, high dimensional datasets can be analysed by exploiting well-known phenomena that seem to govern high-dimensional metric spaces. These phenomena, known as *concentration of measure*, have been extensively studied for several decades. Concentration of measure refers to the fact that regular functions tend to become nearly constant as the dimension increases. This phenomenon was first observed by Milman, and further developed by Gromov [58–61]. It describes for example the behaviour of *distance functions* in high dimension, i.e. *concentration of distance* (which results in the nearest and the furthest point to have comparable distances,§ 1.9) and in general of *Lipschitz functions* (i.e. *concentration of functions*). Another example is *concentration of volumes* [62], which is responsible for the so called *emptiness of space*, along with the sparsity of the data in high dimension.

Our aim is to take advantage of concentration of measure in order to give a description of the dataset. As we will see, Lipschitz functions can be used as tools to analyse the structure of the dataset, by leveraging the concentration of measure property.

In this chapter we give a brief overview of the mathematical framework of metric measure spaces, and we present some theorems and examples that will be further refined in the last chapter.

2.1 Metric Measure Spaces

Metric measure spaces are mathematical structures that combine the metric and the measure defined on a set, e.g. a dataset in the discrete case.

Definition 2.1.

- Let (X, d) be a *complete separable* metric space. A *measure* on X is a measure on the space $(X, \mathcal{B}(X))$, with $\mathcal{X}$ the Borel σ-algebra of X (generated by the opens balls of X).
- The *push forward* of ν under a measurable map $f : X \to Y$ into another metric space Y is the probability measure $f_*\nu$ on Y given by

$$(f_*\nu)(A) := \nu(f^{-1}(A))$$

for all measurable $A \subset X$.

Metric measure spaces. With reference to the definition given in [15], a *metric measure space* (mm-space) is a triple (X, d, μ) where

- (X, d) is a *complete separable* metric space with distance d,
- μ is a *measure* on $(X, \mathcal{B}(X))$ which is locally finite i.e. $\mu(B_r(x)) < \infty$ for all $x \in X$ and sufficiently small r. $\mathcal{B}(X)$ is the Borel algebra on X induced by d, and $B_r(x)$ the ball of radius r centered at x.

From now on, we restrict our attention to metric measure spaces for which $\mu(X) = 1$, *i.e.* μ is a probability measure.

Isomorphism of mm-spaces.

Definition 2.2. Let (X, d_X, μ_X) and (Y, d_Y, μ_Y) be metric measure spaces. An *isomorphism of mm-spaces* between (X, d_X, μ_X) and (Y, d_Y, μ_Y) is a map which is an isometry on the support of the measures, $\phi : supp[\mu_X] \to supp[\mu_Y]$, such that $\mu_X(\phi^{-1}(B)) = \mu_Y(B)$ for all $B \subset Y$ measurable.

2.1.1 Examples

Unit sphere. An example of a metric measure space is $(S^{n-1}, g, \sigma_{n-1})$, the unit sphere S^{n-1}, endowed with the geodesic distance and the uniform measure [60], [62].

Gaussian Spaces. A notable example of metric measure spaces is given by Gaussian spaces. A Gaussian space is of the form $G_n = (\mathbb{R}^n, \|\cdot\|_2, \gamma^n_{c,\Sigma})$, i.e. $\mathbb{R}^n$ endowed with the Euclidean distance and a Gaussian measure $\gamma^n_{c,\Sigma}$. The Gaussian measure is the multivariate Gaussian distribution $\mathcal{N}(c, \Sigma)$, where c is the mean value vector and Σ is the covariance matrix. For isotropic Gaussian spaces, $\Sigma = \sigma^2 I_n$ and

$$\gamma^n_{c,\sigma} = \frac{1}{(\sqrt{2\pi}\sigma)^n} \int_{\mathbb{R}^n} e^{-\frac{\|x-c\|^2}{2\sigma^2}} dx,$$

and, for general Gaussian spaces,

$$\gamma^n_{c,\Sigma} = \frac{1}{\sqrt{(2\pi)^n}\sqrt{\det\Sigma}} \int_{\mathbb{R}^n} e^{-\frac{1}{2}(x-c)^T\Sigma^{-1}(x-c)} dx$$

with $\det(\Sigma)$ the determinant of Σ.

Molecular Datasets. Molecular dataset can be described as metric spaces where the measure is the Boltzmann distribution. Data from molecular dynamics are collected in the configuration space X, with $X \subset \mathbb{R}^{3N}$, where N is the number of atoms (possibly a selection e.g. the number of carbons). Data are generated using an effective potential E_{tot} involving parameters such as the torsion angles between covalent bonds or the distance btween specific atoms,

from which a force field is derived to perform the simulation. The associated Boltzmann distribution can then be written as

$$\mu \propto e^{-\frac{E_{tot}}{kT}}$$

The potential energy expression usually contains a large number of variables, i.e. it is written in terms of a large number of coordinates. Often one uses a small number of coordinates N_R, the so called *reaction* or *active* coordinates to describe the system, which results in a reduced Boltzmann distribution μ_R obtained by pushforward from μ. The space $(\mathbb{R}^{N_R}, \|\cdot\|, \mu_R)$ is an example of metric measure space describing the behavior of the molecule at thermal equilibrium.

2.2 Concentration of Measure

Concentration of measure is a property of metric measure spaces that roughly says that regular functions tend to be nearly constant [59],[60],[58],[61]. It can be observed in many spaces, typical examples being high dimensional spheres with the uniform measure, or general Gaussian spaces.

Concentration function. We say that a measure μ on some metric measure space (X, d, μ) has *σ-concentration* if there exists a constant a_c, such that for any set A with $\mu(A) \geq \frac{1}{2}\mu(X)$, for any $\epsilon \geq 0$ we have:

$$\mu(A_\epsilon) \geq 1 - a_c e^{-\frac{\epsilon^2}{\sigma^2}} \tag{2.1}$$

where $A_\epsilon = \{x \in X, d(x, A) < \epsilon\}$ is the ϵ-offset of A.

Concentration and Lipschitz functions. Concentration in a metric measure space (X, d, μ) can equivalently be stated in terms of Lipschitz functions. For f a real function on X, with median $M(f)$, let

$$\alpha_f(\epsilon) = \mu\{x : |f(x) - Mf| \geq \epsilon\}$$

be the concentration function for f on X. We say that f has σ-concentration, for some $\sigma > 0$, if for any $\epsilon > 0$:

$$\mu\{x : |f(x) - M(f)| \geq \epsilon\} \leq e^{-\frac{\epsilon^2}{2\sigma^2}} \tag{2.2}$$

A metric measure space X is said to have σ-concentration if all 1-Lipschitz functions have σ-concentration on X. In particular, if $\epsilon = O(\sigma)$, the right term in equation 2.2 is constant, so that $f(x)$ differs from $M(f)$ by at most $O(\sigma)$, with constant probability.

2.2.1 From Isoperimetric Inequalities to Concentration

Concentration of measure appears as a natural consequence of isoperimetric inequalities.
Let (X, d, μ) be a metric measure space and $\mu^+ = \lim_{r\to 0} \inf \frac{1}{r}\mu(A_r/A)$ be the boundary measure for a Borel set $A \in X$. Then the *isoperimetric function* of μ is the largest function I_μ on $[0, \mu(X)]$ so that

$$\mu^+(A) \geq I_\mu(\mu(A)) \tag{2.3}$$

holds for every Borel set A, with $\mu(A) \leq \infty$. B is said to be an *extremal set* if the equality holds, i.e. $\mu^+(B) = I_\mu(\mu(B))$.

As a notable example, for the sphere, the isoperimetric inequality states that the spherical caps minimize the boundary measure at fixed volume. While isoperimetric inequalities deal explicitly with *extremal sets*, *concentration* is related to non-infinitesimal neighbourhoods [60], and can be extended to situation not covered by the isoperimetric formalization. Moreover isoperimetric functions are known only for few example, and usually are very difficult to compute.

We present in this section the link between isoperimetric inequalities and concentration properties [60]. In fact, concentration of measure, for important classes of metric measure spaces, can be formalized in this framework.

Concentration Functions. A notable example, for which isoperimetric function can be computed, is the class of constant curvature metric measure spaces, namely, manifolds with constant curvature endowed with their canonical measure. Letting $v(r)$ be the volume of a ball of radius r, the isoperimetric function can be expressed as

$$I_v = v' \circ v^{-1}$$

In general the following result holds:

Proposition 2.3. *[60] Assume $I_\mu \geq v' \circ v^{-1}$ for some strictly increasing differentiable functions. The, for every r*

$$v^{-1}(\mu(A_r)) \geq v^{-1}(\mu(A)) + r.$$

In fact, the condition in proposition 2.3, for a Borel set A of finite measure, reduces to

$$\mu^+(A) \geq v' \circ v^{-1}(\mu(A)). \tag{2.4}$$

For spaces of constant curvature this condition is satisfied by geodesic balls, and enables to express the above proposition as

$$\mu(A_r) \geq \mu(B_r), \tag{2.5}$$

as soon as with $\mu(A) = \mu(B)$, and where B is a ball. Specifically, for the sphere, among all measurable sets $A \subset S^{n-1}$, for a given measure, spherical caps minimize the measure of the ϵ-neighborhood $\mu(A_\epsilon)$.
From the previous conditions 2.4, 2.5, one can recover the concentration function of the space X using the following proposition:

Proposition 2.4. *[60] Let (X, d, μ) be a metric measure space, for which proposition 2.3 applies, and assume $I_\mu \geq v' \circ v^{-1}$, then*

$$\alpha_{(X,d,\mu)}(r) \leq 1 - v\left(v^{-1}\left(\frac{1}{2}\right) + r\right), \quad r > 0 \tag{2.6}$$

Concentration on the Sphere. Let us consider the case of the unit sphere S^n in $\mathbb{R}^{n+1}$, endowed with its uniform measure σ. For $0 < r < \pi$,

$$v(r) = \frac{1}{\int_0^\pi \sin^{n-1}\theta d\theta} \int_0^r \sin^{n-1}\theta d\theta,$$

From calculations, it follows that:

Proposition 2.5. *[60] For the unit sphere S^n, we have:*

$$\alpha_{S^n} \leq e^{-(n-1)r^2/2}, \quad r > 0.$$

From the definition of concentration function, we get:

Theorem 2.6 (Concentration on the sphere)**.** *For an arbitrary measurable set A on the sphere, with $\sigma(A) \geq \frac{1}{2}$,*

$$\sigma(A_\epsilon) \geq 1 - 2e^{-\frac{(n-1)\epsilon^2}{2}} \tag{2.7}$$

In general, for a sphere of radius R:

$$\alpha_{RS^n} \leq e^{-\frac{(n-1)r^2}{2R^2}} \tag{2.8}$$

When compared to equation 2.3, it is important to note that the concentration properties 2.7,2.6,2.1 can be expressed with no dependence on any extremal set.

Concentration for the Gaussian Space. The isoperimetric inequality for the sphere implies the isoperimetric inequality for Gaussian spaces. In fact, by Poincaré lemma, it is known that the uniform measure on a sphere of radius $\sqrt{n}$ approximates, after projection on a finite number of coordinates, a Gaussian distribution[1].

For a Gaussian measure γ_n in R^n, with associated Euclidean distance, concentration function for Gaussian spaces, can be found by performing the limit in 2.8, setting $R = \sqrt{n}$,

[1]Specifically, the extremal sets for the Gaussian isoperimetric inequality are the half spaces, and starting from the distribution of the half spaces, one can recover the same concentration function.

$$\alpha_{(\mathbb{R}^n, \|\cdot\|_2, \gamma_n)} \leq e^{-\frac{r^2}{2}}. \tag{2.9}$$

It is important to note that there is no explicit dependence on the dimension.

In fact, it is easy to see that condition 2.9 implies several concentration properties for the Gaussian space. They became explicit in the functional version, the *Levy's lemma* [61], which states concentration bounds in terms of L-Lipschitz functions defined on the isotropic Gaussian space G_n.

Theorem 2.7 (Levy's Lemma)**.** *Let f be a Lipschitz function of constant L defined on the isotropic Gaussian space $G_n = (\mathbb{R}^n, \|\cdot\|_2, \gamma_n)$. Then f has σL concentration.*

This result applies to anisotropic Gaussians if one takes σ^2 to be the maximum variance of the distribution, i.e. $\sigma^2 = \max_j \sigma_j^2$, over all principal directions, $j = \{1, .., n\}$. Levy's lemma implies that, for high dimensional Gaussian spaces, most of the points are at about the same distance from the center.

Proposition 2.8. *Almost all the mass of an isotropic Gaussian is concentrated in a spherical shell of radius $\sigma\sqrt{n}$ and width $O(\sigma)$.*

Indeed, for an isotropic Gaussian vector x,

$$\mathbb{E}(\|x\|^2) = \mathbb{E}(x)^2 + \mathrm{Var}(\mathcal{N}(0, \sigma^2 I_n) = \sigma^2 n.$$

As distance functions are 1-Lipschitz, by Levy's lemma, they have σ-concentration. This implies that the distance of every point from the center differs by at most $O(\sigma)$ from $\sigma\sqrt{n}$,

$$\|x\| \sim \sigma\sqrt{n} \pm O(\sigma).$$

Concentration of Measure for Log-concave Functions Other examples of distributions that conjecturally exhibit similar concentration properties are uniform distributions over isotropic convex bodies, and more generally isotropic measures with log-concave densities, meaning that the logarithm of the density is

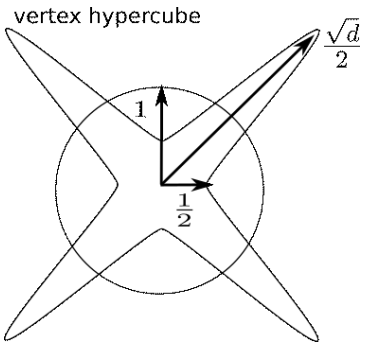

FIGURE 2.1: Growing ratio between the radius of the sphere and the diagonal of the cube when the dimension d increase.

concave. Specifically, for this class of distributions, a weaker form of concentration of measure called exponential concentration would be implied by the *KLS conjecture* [17].

2.3 Emptiness of Space

Another consequence of high dimensionality is the so called *emptiness of space phenomenon*. The picture that we associate to *higher dimensional* cubes, and in general to *high dimensional* shapes and volumes, should change drastically when the dimension increases [62] [27]. In fact, doing simple computations, it is easy to see that when the dimension n increases the diagonal of the unit cube tends to increase, while the radius of the sphere remains constant (by definition).

Moreover, along with the dimension of the space the size N of the dataset is supposed to increase. In order to keep the same distance between the points in the sample from a uniform distribution, when the dimension of the ambient space increases, the size of the sample has to increase exponentially with the dimension, too. This implies that, in general, data tend to be sparse in high dimensions.

Volume of the Unit Sphere To give another example, consider the volume of the unit ball in the Euclidean space. It tends to zero when the dimension of the space tends to infinity. In fact, since the volume of the ball is given by

$$V_n = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}$$

it is easy to see that $\lim_{n\to\infty} V_n = 0$.

Moreover, when the dimension increases the mass of the ball becomes concentrated in a thin shell. Indeed, the ratio between the volume of a ball of radius $(1-\epsilon)$ and the one of a ball of radius 1 is $(1-\epsilon)^n$, which goes to 0 for $\epsilon \gg 1/n$ when n goes to ∞.

Volume of the Cube As a remark, we note that, for large enough n, the volume of the cube tends to be concentrated around its the vertices (figure 2.1). Indeed, if we consider the ball of radius $c\sqrt{n}$, for any $c < (2\pi)^{-1/2}$, we see using Stirling's formula that its volume tends to zero as the dimension tends to infinity. This implies that for n large enough, almost all the volume of the unit cube will lie outside that of that sphere.

Chapter 3

The Distance Transform of a Metric Measure Space

Noise is one of the reason for high dimensionality of datasets: a low-dimensional signal corrupted by isotropic ambiant noise will appear high-dimensional, although most of its variance represents unwanted information. Under some specific hypotheses, concentration of measure can be used to suitably shrink the noise, in order to make the data consistent with the signal. We present an approach designed for denoising data in the framework of metric measure spaces. In order to do so, we use some distances defined on the space of metric measure spaces.

The possibility of defining a distance between metric spaces, the *Gromov-Hausdorff distance* [58], resulted in powerful tools for computational purposes. The *GH-based methods* have been used to compare datasets and shapes [63], [64], [65]. For specific classes of datasets, in particular low dimensional and smooth manifolds, this approach led to important results. The GH-distance has been further generalized, by Memoli, to a distance on the space of metric measure spaces: the *Gromov-Wasserstein distance*, defined taking into account both metric and measure features of a data set [66].
We first introduce some technical tools in order to define distances between metric measure spaces [15], following Sturm's approach, in the formalism of Chap. 2.

Then we introduce Memoli's definition of Gromov-Wasserstein distance inspired by the Gromov-Hausdorff distance between metric spaces. We then suggest a simple way to "denoise" metric measure spaces, which we will refine in the last chapter of the thesis.

3.1 Distances Between Metric Measures Spaces: $\mathbb{D}$-distance

In this section we present the definition of a distance on the space of isomorphism classes of mm-spaces, the $\mathbb{D}-$*distance*, by Sturm. In order to give the definition of $\mathbb{D}$-distance, we introduce some fundamental tools such as the *Wasserstein distance*, *coupling of measures*, and *coupling of metrics*.

We consider below isomorphism classes of metric measure spaces, and the distances are defined on the space of isomorphism classes of mm-spaces:

Isomorphism of mm-spaces.

Definition 3.1. Let (X, d_X, μ_X) and (Y, d_Y, μ_Y) be metric measure spaces. An *isomorphism of mm-spaces* between (X, d_X, μ_X) and (Y, d_Y, μ_Y) is a map which is an isometry on the support of the measures, $\phi : supp[\mu_X] \to supp[\mu_Y]$, such that $\mu_X(\phi^{-1}(B)) = \mu_Y(B)$ for all $B \subset Y$ measurable.

Wasserstein Distances.

Definition 3.2 (Coupling of measures)**.** Given two metric measure spaces (X, μ_X), (Y, μ_Y) a measure μ on $X \times Y$ is a *coupling* of μ_X and μ_Y if its marginals are μ_X and μ_Y, that is, if

$$\mu(A \times Y) = \mu(A),\ q(X \times A') = \nu(A')$$

for all measurable sets $A \subset X$, $A' \subset Y$.

An example of coupling, the most obvious but not always the most suitable, is the product measure $\mu_X \times \mu_Y$. Couplings μ of μ_X and μ_Y are also called *transportation plans* from μ_X to μ_Y, and may be described as the plans to transport some products whose locations are distributed according to μ_X to consumer distributed according to μ_Y. The set of all couplings for given measures μ_X and μ_Y is denoted by $\mathcal{M}(\mu_X, \mu_Y)$. The problem of comparing measures can be stated as a *mass-transportation problem*, i.e. to find the best way to move a certain mass from producers distributed according to a law μ_X to customers, distributed according to μ_Y. This leads to the definition of *Wasserstein* distances, which are metrics on the space of Borel probability measures μ on X with $\int d(x_0, x)^p d\mu(x) < \infty$ for some $x_0 \in X$:

Definition 3.3. (Wasserstein distance). Let (X, d) be a metric space, for $p \geq 1$, the L_p-*Wasserstein distance* between μ_1 and μ_2 is defined as

$$d_{W,p}(\mu_1, \mu_2) = \left(\inf_{\mu \in \mathcal{M}(\mu_1, \mu_2)} \int d^p(x, y) \mathrm{d}\mu(x, y) \right)^{1/p}$$

$\mathbb{D}$-distance. The problem of defining distances between two metric measure spaces is formalized by Sturm [15] who introduced the $\mathbb{D}$-distance. To define it, we first need to introduce *couplings of metrics*:

Definition 3.4. (Coupling of metrics). Given two metric (measure) spaces (X, d_X, μ_X) (Y, d_Y, μ_Y), a pseudometric $\hat{d}$ on the disjoint union $X \sqcup Y$ is a *coupling* of d_X and d_Y iff

$$\hat{d}(x, y) = d(x, y) \text{ and } \hat{d}(x', y') = d'(x', y')$$

for all $x, y \in supp[m] \subset X$ and all $x', y' \in supp[m'] \subset Y$, *i.e.* $\hat{d}$ extends d and d' on $M \sqcup M'$.

Definition 3.5 ($\mathbb{D}$-distance)**.** Given $p \geq 1$, the $\mathbb{D}$-distance between two metric measure spaces is defined as

$$\mathbb{D}_p((X, d_X, m_X), (Y, d_Y, m_Y)) = \inf_{\hat{\mathrm{d}}, \mu} \left(\int_{M \times M'} \hat{d}^p(x, y) \mathrm{d}\mu(x, y) \right)^{1/p}$$

where $\hat{d}$ is a coupling of d_X and d_Y, and μ is a coupling of μ_X and μ_Y.

The $\mathbb{D}_p$-distance is a complete and separable metric on the family of all isomorphism classes of normalised metric measure spaces. While it enjoys certain useful properties from a theoretical point of view, it isn't very easy to handle from a computational perspective, because the optimization with respect to the metric coupling is difficult to perform in general.

3.2 Distance Between Metric Measure Spaces: Gromov-Wasserstein Distance

A modified version of Sturm's definition 3.5 is given by Memoli [66], and is known as the *Gromov-Wasserstein* distance. While similar, the definitions are quite independent. More specifically, Memoli's definition of Gromov-Wasserstein distance is inspired by the Gromov-Hausdorff distance between metric spaces. In fact, while Sturm's approach originates from the formalism of metric measure spaces, Memoli definitions and techniques are more data analysis oriented. This makes Gromov-Wasserstein distances also more amenable for numerical computation ([66], section 7).

Correspondences. In order to give the definition of Gromov-Wasserstein distance we introduce the notion of *correspondences* and the related *Gromov-Hausdorff distance*.

Definition 3.6 (Correspondence)**.** Let A and B be sets. A subset $R \in A \times B$ is a correspondence between A and B if

- for each $a \in A$, there exist a $b \in B$ such that $(a, b) \in R$.
- for each $b \in B$ there exist $a \in A$ such that $(a, b) \in R$.

We denote by $\mathcal{R}(A, B)$ the set of all possible correspondence between A and B.

Gromov-Hausdorff and Gromov-Wasserstein distance

Definition 3.7. Let (X, d) be a metric space. The *Hausdorff distance* between any two sets $A, B \in X$ can be expressed as

$$d_H(A, B) = \inf_R \sup_{(a,b)\in R} d(a, b)$$

where the infimum is taken over all $R \in \mathcal{R}(A, B)$.

The *Gromov-Hausdorff* distance function d_{GH} and the *Gromov-Wasserstein* distance function d_{GW} can be defined, using the previous definitions.

Definition 3.8. Let (X, d_X), (Y, d_Y) be metric spaces and $\Gamma(x, x', y, y') = |d_X(x, x') - d_Y(y, y')|$. The Gromov-Hausdorff distance between the two metric spaces is:

$$d_{GH}(X, Y) = \frac{1}{2} \inf_{R\in\mathcal{R}(X,Y)} \|\Gamma\|_{L^\infty(R\times R)}$$

Recall the definition of coupling of measures 3.2 in the previous section. Then, analogously to the Gromov-Hausdorff distance, the Gromov-Wasserstein distance can be defined. In fact the Gromov-Wasserstein distance can be seen as a *relaxed* form of the Gromov-Hausdorff distance [66].

Definition 3.9 (Gromov-Wasserstein)**.** Let (X, d_X, μ_X), (Y, d_Y, μ_y) be metric measure spaces and $\Gamma(x, x', y, y') = |d_X(x, x') - d_Y(y, y')|$. For $1 \leq p \leq \infty$, the Gromov-Wasserstein distance between the two mm-spaces is:

$$d_{GW}(X, Y) = \frac{1}{2} \inf_{\mu\in\mathcal{M}(\mu_X,\mu_Y)} \|\Gamma\|_{L^p(\mu\times\mu)}$$

It can be shown that the $\mathbb{D}$-distance and the d_{GW}-distance coincide for $p = \infty$.

3.3 Denoising

The term *denoising* or *noise-reduction* usually refers to a general procedure aimed at eliminating the *non-signal* components from a a dataset. For data in Euclidean space, if the noise is assumed to be additive, denoising can be formulated as a deconvolution problem. Equivalently, data in this model are

a mixture of components describing the structure of the noise corrupting each signal point, and the goal is to retrieve the parameters of this mixture.

We propose an idea that takes advantage of the properties of Lipschitz functions in high dimensional spaces, to perform denoising on a dataset. This algorithm operates at the level of metric measure spaces and is thus not limited to data in Euclidean spaces. Although this idea is essentially already proposed by Dubnov et al. in 2002 [16], the analysis we provide sheds new light on its properties in the context of high dimensional data.

3.3.1 A transform on metric measure spaces

We define the following map on the set of metric measure spaces, which we call the *distance transform*:

Definition 3.10. For $\{X, d, \mu\}$ a metric measure space, define $\phi_X : X \to \mathbb{R}^X$ by $\phi_X(x) = d(x, .)$ for all $x \in X$. Given $p \geq 1$, if X is such that the image of ϕ_X is in $L^p(X)$, the distance transform of X is the metric measure space $\Phi(X) = \phi_X(X)$.

It is clear from the triangle inequality that ϕ_X is always non expansive, so that distances in $\Phi(X)$ are at most the ones in X. Also, for $p = \infty$, Φ reduces to the identity map, since $|d(x, y) - d(y, y)| = d(x, y)$.

3.3.2 Stability

We first check that the distance transform is robust with respect to perturbations of the input mm-space. More precisely:

Lemma 3.11. *The map Φ is 2-Lipschitz for any $p \geq 1$:*

$$d_{GW}(\Phi(X), \Phi(Y)) \leq 2 d_{GW}(X, Y)$$

for any two mm-spaces X and Y for which Φ is defined.

Proof. We consider for simplicity the case of Gromov-Wasserstein distances with exponent 1. The GW-distance between the images of $X = (X, d_X, \mu_X)$ and $Y = (Y, d_Y, \mu_Y)$ can be written

$$
\begin{aligned}
d_{GW}(\Phi(X),\Phi(Y)) &= \frac{1}{2}\inf_{\mu\in\mathcal{M}(\mu_X,\mu_Y)}\int |\|d_X(x,.)-d_X(x',.)\|_{L^p(\mu_X)} \\
&\qquad -\|d_Y(y,.)-d_Y(y',.)\|_{L^p(\mu_Y)}|d\mu(x,y)d\mu(x',y') \\
&\leq \frac{1}{2}\inf_{p_X,p_Y}\int |\|d_X(p_X(a),p_X(.))-d_X(p_X(a'),p_X(.))\|_{L^p(\Omega)} \\
&\qquad -\|d_Y(p_Y(b),p_Y(.))-d_Y(p_Y(b'),p_Y(.))\|_{L^p(\Omega)}|dada'dbdb' \\
&\leq \frac{1}{2}\inf_{p_X,p_Y}\int \Bigg|\left(\int (d_X(p_X(a),p_X(z))-d_X(p_X(a'),p_X(z)))^p dz\right)^{1/p} \\
&\qquad -\left(\int (d_Y(p_Y(b),p_Y(t))-d_Y(p_Y(b'),p_Y(t)))^p dt\right)^{1/p}\Bigg| dada'dbdb' \\
&\leq \frac{1}{2}\inf_{p_X,p_Y}\int \left(\int (d_X(p_X(a),p_X(z))-d_Y(p_Y(b),p_Y(t)))^p dz\right)^{1/p} \\
&\qquad +\left(\int (d_X(p_X(a'),p_X(z))-d_Y(p_Y(b'),p_Y(t)))^p dt\right)^{1/p}\Bigg| dada'dbdb' \\
&\leq 2d_{GW}(X,Y)
\end{aligned}
$$

In the second line we reformulated couplings as measure preserving parametrizations p_X and p_Y of both measures μ_X and μ_Y over a common probability space Ω. In the fourth line we swap terms between the two L_p norms using the triangle inequality, which gives the desired claim.

□

3.3.3 Behavior with respect to products

Given two mm-spaces $X = (X, d_X, \mu_X)$ and $N = (N, d_N, \mu_N)$ their product is the mm-space $X \times N = (X \times N, d_{X\times N}, \mu_X \otimes \mu_N)$, with $d_{X\times N}((x,n),(x',n')) = d_X(x,x') + d_N(n,n')$. Thinking of X as the "signal", for example a low dimensional manifold, and of N as the "noise", for example a Gaussian space, the product $X \times N$ can be thought of as a noise corrupted version of X. In this model of noise, if N is a high dimensional Gaussian space, generally $X \times N$ will be very far from X in the GW-distance, as distances will be typically increase by the variance of the Gaussian, which is proportional to its dimension. We observe below that applying the distance transform has the effect of considerably reducing that effect if N satisfies the concentration of measure property:

Lemma 3.12. *If N has σ-concentration, then*

$$d_{GW}(\Phi(X), \Phi(X \times N)) \leq O(\sigma)$$

for any fixed finite p, the Gromov-Wasserstein distance being computed with exponent p.

Proof.

$$\begin{aligned}
d(\phi(x,n), \phi(x',n')) &= \|d_{(x,n)} - d_{x',n')}\|_{L^p(X \times N)} \\
&= \left(\int |d_X(x,y) + d_N(n,m) - d_X(x',y) - d_N(n',m)|^p d\mu_X(y) d\mu_N(m) \right)^{1/p} \\
&\leq \left(\int |d_X(x,y) - d_X(x',y)|^p d\mu_X(y) \right)^{1/p} \\
&\quad + \left(\int |d_N(n,m) - d_N(n',m)|^p d\mu_N(m) \right)^{1/p}
\end{aligned}$$

Hence

$$|d(\phi(x,n), \phi(x',n')) - d(\phi(x), \phi(x'))| \leq \left(\int |d_N(n,m) - d_N(n',m)|^p d\mu_N(m) \right)^{1/p}$$

But $(n, n', m) \mapsto d_N(n,m) - d_N(n',m)$ is 4-Lipschitz and has zero mean, hence by concentration its norm in $L_p(N^3)$ is $O(\sigma)$. Therefore, using the projection $X \times N \to X$ as coupling, we have

$$\begin{aligned}
d_{GW}(\Phi(X), \Phi(X \times N))^p &\leq \int |d(\phi(x,n), \phi(x',n')) - d(\phi(x), \phi(x'))|^p d\mu_X(x) d\mu_X(x') d\mu_N(n) d\mu_N(n') \\
&\leq \int |d_N(n,m) - d_N(n',m)|^p d\mu_N(m) d\mu_N(n) d\mu_N(n') \\
&\leq O(\sigma^p)
\end{aligned}$$

which is the desired claim. □

3.3.4 Distortion bound for simple spaces

While the above paragraph shows that the distance transform has noise reduction properties, one should bear in mind that it does affect the signal as well. However, the distortion can be controlled for low dimensional spaces, for example as follows:

Lemma 3.13. *Let $X = (X, d, \mu)$ be a metric measure space. If there exists $c > 0$ such that*

(i) the balls of radius $c/2$ have measure at least λ
(ii) any ball of radius $2c$ can be mapped to an Euclidean d-ball by a measure-preserving bijection that changes pairwise distances by at most ϵ

then X and $\Phi(X)$ are approximately bilipschitz equivalent, in the sense that

$$d(x, y) \geq d_{\Phi(X)}(\phi(x), \phi(y)) \geq C(d, p, \lambda)(d(x, y) - \epsilon)$$

for all x, y in X.

Proof. The first inequality always holds. For the second one, we first consider the case where $d(x, y) \geq 2c$. Function $d(x, .) - d(y, .)$ is 2-Lipschitz and equals $d(x, y)$ at y. Hence it is larger than $d(x, y) - c \geq d(x, y)/2$ on $B(y, c)$. As a consequence the L_p norm of $d_x - d_y$ is at least $\lambda^{1/p} d(x, y)/2$. If $d(x, y) < 2c$, we use the fact that $B(x, 2c)$ is close to a Euclidean d-ball B. Let ψ be the measure preserving bijection given by assumption (ii), and let $\mu_x = \mu_{|B(x,2c)}/\mu(B(x, 2c))$ and ν be the uniform probablity distribution of B. By change of variable, we see that

$$\|d_x - d_y\|_{L_p(\mu_x)} \geq \|d_B(\psi(x), .) - d_B(\psi(y), .)\|_{L_p(\nu)} - \epsilon$$

By elementary geometry, it is easy to see that the first term in the right hand side is at least $C d_B(\psi(x), \psi(y))$, where C depends on d and p. Hence

$$\|d_x - d_y\|_{L_p(\mu_x)} \geq C d(x, y) - (C + 1)\epsilon$$

And since $B(x, 2c)$ has measure at least λ by assumption (i), we obtain

$$\|d_x - d_y\|_{L_p(\mu)} \geq \lambda^{1/p}(Cd(x,y) - (C+1)\epsilon)$$

which proves the desired claim. $\square$

The constant C in the above lemma intuitively depends on the dimension and on the "geometric complexity" of X. Indeed, parameter λ is related to the number of "simple" patches that are required to cover X. If X is for example a Riemannian d-manifold, there exists a radius c such that X looks ϵ-close to Euclidean space at scale c, and $1/\lambda$ indicates how small that scale is with respect to the size of the whole manifold. Using differential geometric estimates, it can be shown for example that λ is controlled by a function of the volume of the manifold, its maximum absolute sectional curvature and its injectivity radius. It seems likely that in the Riemannian case, X and its distance transform are guaranteed to be bilipschitz equivalent rather than approximately so.

Regarding the application of the distance transform to data, we note that it is trivially implementable given data in the form of a finite input metric measure space. However, rather than further developping this initial idea in a broad context, we refine it in the context of Euclidean spaces in the next chapter of thesis.

Chapter 4

Spectral Properties of Radial Kernels and Clustering in High Dimensions

4.1 Introduction

Given a set of data points drawn from a mixture of distributions, a basic problem in data analysis is to cluster the observations according to the component they belong to. For this to be possible, it is clearly necessary to impose separation conditions between the different components in the mixture.

Many approaches have been proposed to solve the problem of clustering mixtures of distributions. We give below a brief historical account of the algorithms that come with theoretical guarantees, focusing on the high dimensional situation. Unlike in the low dimensional dimensional case, approaches based *e.g.* on single linkage or spectral clustering cannot be employed, because such methods require dense samples which would have an unreasonably large cardinality. The first result in this field, due to Dasgupta, used random projection onto a low dimensional subspace [67]. It was shown that a mixture of Gaussians with unit covariance in dimension n could be provably well clustered if the separation between the means of the components was $O(\sqrt{n})$. The result was later improved by Dasgupta and Schulman [68] using a variant of EM for unit covariance Gaussians, and by Arora

and Kannan [69], using a distance-based algorithm, for Gaussians with at most unit covariance. These methods, to correctly classify the components, require a $O(n^{1/4})$ separation between the centers of the Gaussian. For mixtures of unit covariance Gaussians, Vempala and Wang [70] used PCA to obtain a dimension-free separation bound, which depends only on the number of the components. Their method is based on the fact that the space spanned by the k top singular vectors of the mixture's covariance matrix contains the centers of the components. Projecting to this space has the effect of reducing the variance of each component while maintaining the separation between the centers. Kannan et al. [71] extended this idea to mixtures of log concave distributions with at most unit covariance, also requiring a separation between the centers that depends only on the number of the components. Achlioptas and McSherry [72] improved further the dependency of the separation bound on the number of components. A combination of PCA with a reweighting technique was proposed by Brubaker and Vempala [73]. This method is affine invariant and can deal with highly anisotropic inputs as a result. When applied to a sample from a mixture of two Gaussians, the algorithm correctly classifies the sample under the condition that there exists a half space containing most of the mass of one Gaussian and almost none of the other. Finally, a different family of approaches uses the moments of the mixture to learn the parameters of the components. Strong results have been obtained in this direction (see *e.g.* [18, 74]). These methods do not require any separation assumption, however their downside is that they require a priori knowledge of a small parametric family containing the component's distributions. They also become inefficient when applied to high dimensional data, since the number of moments involved grows rapidly with the dimension. For example, the currently fastest algorithm [18] for learning mixtures of Gaussians runs in time $O(n^6)$.

Another possible approach to the analysis of mixtures uses kernel matrices. On a dataset $\{x_1, ...x_N\}$ of N points in $\mathbb{R}^n$ a kernel function $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defines a $N \times N$ kernel matrix whose ij entry is $k(x_i, x_j)$. An important class of kernels are positive definite kernels, which are those for which the associated kernel matrix is positive definite for any dataset. The use of such kernel matrices, and

in particular of their spectral decomposition as in the popular kernel PCA algorithm, has long become commonplace in data analysis. Still, surprisingly little is known regarding theoretical justifications for kernel based clustering methods. Notably, the analysis in [75] implies that a kernel PCA type algorithm will correctly cluster mixtures when the components are sufficiently separated. However, the arguments used in this paper follow the low (or constant) dimensional intuition and the required separation between the components is of the order of the width of the kernel, which typically leads to a separation that grows like the square root of the dimension.

In order to improve the above analysis of kernel PCA, it is necessary to better understand the behavior of kernel matrices and of their spectra as the dimension increases. Again, while the literature on eigenvalues of random matrices is vast and growing rapidly, the knowledge about random kernel matrices is much scarcer. A notable exception is [52], which gives an asymptotic description of radial kernel matrices of the form $k(x_i, x_j) = h(\|x_i - x_j\|^2/n)$ as the dimension n tends to infinity, for a fixed function h. In the case of distributions whose coordinates are independent after some linear change of coordinate, *e.g.* Gaussians, it is shown that the kernel matrices converge in the operator norm to a certain matrix related to the covariance of the data. Under the weaker condition that the distribution enjoys concentration properties, the corresponding convergence result is proved to hold at the level of spectral distributions, but no result is derived for individual eigenvalues.

In this paper, we prove new results about radial kernel matrices of mixtures of high dimensional distributions. Unlike [52], we do not assume independence of coordinates. Rather, we only assume that the components in the mixtures have exponential concentration. Specifically, we show that such matrices can be very well approximated by the sum of a matrix that is row constant within each component and a matrix that is column constant within each component. For distance matrices of mixtures with a single component, the result implies a large spectral gap between the two largest eigenvalues: The ratio between these eigenvalues is of the order of the dimension, rather than that of the square root of the dimension, as one might naively expect from basic concentration results. When the input distributions are supported on a sphere, this "double

concentration" phenomenon is enhanced and large eigenvalue gaps arise for kernel matrices more general than distance matrices. The proof technique is geometric and very different from the one used in [52].

For positive kernels, a consequence of the above result is that kernel PCA is a valid clustering method as long as the Gram matrix of the mixture's components, when viewed as elements of the corresponding Reproducing Kernel Hilbert Space (RKHS), is sufficiently well conditioned. In particular, this allows to check that kernel PCA allows to correctly clusters mixtures of two Gaussians with a required separation between centers that does not depend on the dimension.

In the case of even distributions supported on a sphere and satisfying a Poincaré inequality, we further show that our main result can be strengthened, so that kernel matrices are well approximated by block constant matrices, provided the kernel, which may or may not be positive, is smooth enough. We also design a specific non positive kernel for which this result can be extended to non necessarily even (and non necessarily centered) distributions. This kernel is not of the form studied in [52], so the results of this paper do not apply even for Gaussian mixtures. Based on this kernel, we derive a simple spectral algorithm for clustering mixtures with possibly common means. This algorithm will succeed if the angle between any two covariance matrices in the mixture (seen as vectors in $\mathbb{R}^{n^2}$) is larger than $O(n^{-1/6}\log^{5/3} n)$. In particular, the required angular separation tends to 0 as the dimension tends to infinity. To the best of our knowledge, this is the first polynomial time algorithm for clustering such mixtures beyond the Gaussian case.

4.2 Kernels in high dimensions

Our analysis of kernel matrices for high dimensional data hinges on the concentration of measure phenomenon. Concentration of measure is a property of metric measure spaces that roughly says that regular functions are nearly constant [58–60]. It can be observed in many spaces, typical examples being Gaussian spaces or manifolds with Ricci curvature bounded below. We give precise definitions below for a probability measure μ on $\mathbb{R}^n$. We say that $f : \mathbb{R}^n \to \mathbb{R}$ has

exponential σ-concentration, or σ-concentration for short, for some $\sigma > 0$, if for any $\varepsilon > 0$:

$$\mu\{x : |f(x) - M(f)| \geq \varepsilon\} \leq O(e^{-\frac{\varepsilon}{\sigma}})$$

where $M(f)$ is a median of f. The measure μ is said to have σ-concentration if all 1-Lipschitz functions have σ-concentration. In particular, we have that f equals $M(f)$ plus or minus $O(\sigma)$ with high probability.

Levy's lemma [61] states that an isotropic Gaussian with covariance $\sigma^2 I$ has Gaussian concentration, which is a stronger property implying $O(\sigma)$-concentration. This result is also true for anisotropic Gaussians if one takes σ^2 to be the maximum eigenvalue of the covariance matrix. In particular, it implies that for high dimensional Gaussian spaces, most of the points are at about the same distance from the center. More precisely, almost all the mass of an isotropic Gaussian is concentrated in a spherical shell of radius $\sigma\sqrt{n}$ and thickness $O(\sigma)$. Indeed, for an isotropic Gaussian vector x, $\mathbb{E}(\|x\|^2) = \sigma^2 n$. As distance functions are 1-Lipschitz, by Levy's lemma, they have σ-concentration. Hence the distance from a random point to the center differ by at most $O(\sigma)$ from $\sigma\sqrt{n}$, with high probability.

A stronger form of concentration that we will also consider is based on Poincaré inequality. We will say that a probability measure μ satisfies a Poincaré inequality if for any Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ whose mean is zero with respect to μ, we have

$$\int f^2 d\mu \leq O(1) \int \|\nabla f\|^2 d\mu$$

A probability measure that satisfies a Poincaré inequality necessarily has $O(1)$-concentration [76]. Gaussians distributions whose covariance have $O(1)$ eigenvalues are known to satisfy a Poincaré inequality. The famous KLS conjecture [17] states that uniform distributions over isotropic convex bodies, and more generally isotropic measures with log-concave densities also do.

4.2.1 Main result

We consider a mixture μ of k distributions μ_i in $\mathbb{R}^n$, with weights w_i, which we treat as numerical constants. We assume that each component μ_i has $O(1)$-concentration. Drawing a sample of N points independently from the mixture gives a point set X that is, with probability 1, the disjoint union of subsets X_i, corresponding to each component. The *radius* of μ_i is the quantity $(\mathbb{E}_{\mu_i}||x - \mathbb{E}_{\mu_i}x||^2)^{1/2}$ for a random variable x with law μ_i, and we denote by R the smallest radius of the μ_i. We consider a function $h : \mathbb{R}_+ \to \mathbb{R}$ and the associated radial kernel. This defines a kernel matrix $\Phi_h(X)$ whose entries are $h(\|x_i - x_j\|)/N$, for x_i, x_j in X. We assume that the indices are ordered in such a way that the components form contiguous intervals ; in particular, we have a natural block structure (doubly)-indexed by the components.

Theorem 4.1. *If the number of samples N is drawn according to the Poisson distribution with mean N_0, then with arbitrarily high probability, we have:*

$$\|\Phi_h(X) - A\| \leq O\left(c_h + \|h\|_\infty \sqrt{\frac{n \log N_0}{N_0}}\right)$$

where the entries of A in the ij block are given by

$$A_{xy} = \frac{1}{N}\left(\int h(\|x - z\|)d\mu_j(z) + \int h(\|y - z'\|)d\mu_i(z') - \int h(\|z - z'\|)d\mu_i(z)d\mu_j(z')\right)$$

and with

$$c_h = \sup_{r \geq R/2}\left(|h''(r)| + \frac{1}{r}|h'(r)|\right) + \|h'\|_\infty \exp\left(-\Theta(R)\right)$$

Furthermore, if the components μ_i are supported on the sphere centered at 0 with radius $\sqrt{n}$, and have mean at distance $O(1)$ from the origin, the conclusions above hold with c_h replaced by

$$c'_h = \sup_{r \geq R/\Theta(\log(R))}\left(\frac{\log^2(R)|h''(r)| + |h'(r)|}{r}\right) + \|h'\|_\infty / R$$

The proof of Theorem 4.1 follows from the analysis of the map sending each point x in $\mathbb{R}^n$ to its kernel function $h(\|x - .\|)$ in $L^2(\mathbb{R}^n, \mu)$ or, more precisely,

of a finite sample version of this map. That analysis crucially depends on the fact that in Euclidean spaces, the cross derivative of the distance $\frac{\partial^2}{\partial x \partial y}\|x-y\|$ is upper bounded by $O(1/\|x-y\|)$. A first consequence of Theorem 4.1 is the following result about the spectrum of $\Phi_h(X)$, which follows directly from the variational characterization of eigenvalues:

Corollary 4.2. *Under the assumptions of Theorem 4.1, the spectrum of $\Phi_h(X)$ has at most k eigenvalues larger than $O\left(c_h + \|h\|_\infty \sqrt{\frac{n \log N_0}{N_0}}\right)$, and at most k eigenvalues smaller than $-O\left(c_h + \|h\|_\infty \sqrt{\frac{n \log N_0}{N_0}}\right)$, with arbitrarily high probability.*

4.2.2 Distance matrices

To illustrate Theorem 4.1, setting for example $h(r) = r$ gives a description of distance matrices. Consider the case of a sample drawn from a mixture of k Gaussians with unit covariance. If x_i and x_j are drawn independently from two Gaussians in the mixture, $x_i - x_j$ is a Gaussian with covariance $2I$. Concentration of measure then implies that the entries $||x_i - x_j||$ of each block concentrate around their mean value, i.e. they differ by at most $O(1)$ from the mean of the block with high probability:

$$\Phi_h(X) = \left(\begin{array}{c|c|c} \Phi_{11} & \cdot\cdot & \Phi_{1k} \\ \hline \vdots & \cdot\cdot & \vdots \\ \hline \Phi_{k1} & \cdot\cdot & \Phi_{kk} \end{array}\right) = \frac{1}{N}\left(\begin{array}{c|c|c} m_1 \pm O(1) & \cdot\cdot & m_{1k} \pm O(1) \\ \hline \vdots & \cdot\cdot & \vdots \\ \hline m_{k1} \pm O(1) & \cdot\cdot & m_{kk} \pm O(1) \end{array}\right) \tag{4.1}$$

A finer description of $\Phi_h(X)$ is given by Theorem 4.1. For an isotropic Gaussian, the radius R is $\Theta(\sqrt{n})$, and from $|h'| = 1$, $|h''| = 0$ we get $c_h = \Theta\left(1/\sqrt{n}\right)$.

The dependancy on the average number of samples N_0 in Theorem 4.1 involves $\|h\|_\infty$, which is unbounded. However, assuming for example that the centers of the components are at distance $O(1)$, then the fraction of pairs of sample points whose distance is larger than an appropriate constant times $\sqrt{n}$ is exponentially small by concentration. Hence we can first modify h by thresholding such that

$\|h\|_\infty$ becomes $O(\sqrt{n})$, with an exponentially small change in $\Phi_h(X)$. Furthermore, by making the transition between the linear part and the constant part smooth enough, we can ensure that the second derivatives of the modified kernel g are $O(1/\sqrt{n})$, so that $c_g = O(1/\sqrt{n})$. Applying the theorem to g implies that with a polynomial number of samples ($N_0 = \Omega(n^3 \log n)$ suffices), with arbitrarily high probability, each block of $\Phi_h(X)$ has the following structure

$$\Phi_{ij} = \frac{1}{N}\begin{pmatrix} a_1 & a_2 & \cdot\cdot & a_{N_i} \\ a_1 & a_2 & \cdot\cdot & a_{N_i} \\ \vdots & \vdots & \cdot\cdot & \vdots \\ a_1 & a_2 & \cdot\cdot & a_{N_i} \end{pmatrix} + \frac{1}{N}\begin{pmatrix} b_1 & b_1 & \cdot\cdot & b_1 \\ b_2 & b_2 & \cdot\cdot & b_2 \\ \vdots & \vdots & \cdot\cdot & \vdots \\ b_{N_j} & b_{N_j} & \cdot\cdot & b_{N_j} \end{pmatrix} + B$$

with $\|B\| = O(1/\sqrt{n})$. Note that the error term B is now much smaller than the one in (4.1), which is a priori up to $O(1)$ in the operator norm.

Furthermore, for each block the vectors (a_s) and (b_t) are, up to a constant, averages of the columns of the distance matrix. As a result these vectors are 1-Lipschitz and thus have $O(1)$-concentration. Also, we can assume they have the same mean, namely half the average distance m_{ij} within the block, that is, at least $\Omega(\sqrt{n})$. So we can write $a_s = m_{ij}(1+\varepsilon_s)/2$ and $b_t = m_{ij}(1+\delta_t)/2$ with ε_s and δ_t in $O(1/\sqrt{n})$ with high probability. This implies that each block is very well approximated by a rank one matrix. Indeed

$$a_s + b_t = m_{ij}(2+\varepsilon_s+\delta_t)/2 = m_{ij}((1+\varepsilon_s/2)(1+\delta_t/2) + O(1/n))$$

In particular, the normalized distance matrix of points drawn according to a single Gaussian has only one eigenvalue that is larger than $O(1/\sqrt{n})$, this top eigenvalue being $\Theta(\sqrt{n})$. This observation, which we stated for isotropic Gaussians for concreteness, applies to any distribution with $O(1)$-concentration and variance $\Theta(n)$ as well.

We also remark that in the case of distributions on the sphere with $O(1)$-concentration and variance $\Theta(n)$, the contribution of h'' in the error bound in Theorem 4.1 is divided by $\Theta(\sqrt{n}/\log^3 n)$, which makes it possible to extend the above discussion to kernels other than distance functions. We do not elaborate

further as the spherical case will be studied in more detail in the sequel of the paper.

4.3 Positive definite kernels and clustering

For radial kernels that are positive definite, *i.e.* that define positive definite kernel matrices, Corollary 4.2 implies that there are at most k significant eigenvalues for mixtures of k probability measures that concentrate. We can use this result to provide guarantees for a simple clustering algorithm. First, assuming a certain gap condition, we can relate eigenspaces of the kernel matrix to the space of piecewise constant vectors, *i.e.* vectors that are constant on each component in the mixture.

The required gap condition can be conveniently formulated in terms of *kernel distances* [77, 78]. Recall that kernel distances are Hilbertian metrics on the set of probability measures, which are obtained by embedding the ambiant Euclidean space into a universal RKHS. More precisely, given two probability measures μ_1 and μ_2 on $\mathbb{R}^n$, the expression

$$\langle \mu_1, \mu_2 \rangle = \int h(||x-y||)\, d\mu_1(x) d\mu_2(y)$$

is a positive definite kernel and the kernel distance is the associated distance.

Proposition 4.3. *Assume h defines a positive definite kernel, and that the conditions of Theorem 4.1 are satisfied. Let*

$$G_h = (\langle \mu_i, \mu_j \rangle)_{i,j=1\ldots k}$$

be the Gram matrix of the components in the kernel distance.

If the smallest eigenvalue of G_h is at least Kc_h, then the maximum angle formed by the space spanned by the top k eigenvectors of $\Phi_h(X)$ and the space of piecewise constant vectors is at most $O(1/\sqrt{K})$, with arbitrarily high probability, provided $N_0 \geq N_1$, with:

$$N_1 = O\left(\frac{\|h'\|_\infty^2}{c_h^2} + \frac{n\|h\|_\infty^2}{c_h^2} \log\left(\frac{n\|h\|_\infty^2}{c_h^2}\right)\right)$$

Under these assumptions we can provide a guarantee for the following basic kernel PCA clustering algorithm. First, we perform a spectral embedding using the k top eigenvectors of $\Phi_h(X)$. Namely, each data point x is mapped to $(\phi_1(x), \ldots, \phi_k(x))$, $\phi_1, \ldots, \phi_k$ being the k dominant eigenvectors of $\Phi_h(X)$. In order to have the right dependency on the total number of points, these eigenvectors are scaled to have norm $\sqrt{N}$. By the above proposition, this will give a point cloud that is $O(\sqrt{1/K})$ close in the transportation distance W_2 to a point cloud obtained using the embedding provided by an orthogonal basis of piecewise constant vectors, scaled to have norm $\sqrt{N}$. Note that in the latter point cloud, each component becomes concentrated at a single location, the distance between any two such locations being $\Omega(1)$. In such a situation, any constant factor approximation algorithm for the k-means problem will find a clustering with a fraction of at most $O(1/K)$ misclassified points. We just proved:

Corollary 4.4. *If the assumptions of Proposition 4.3 are satisfied, kernel PCA allows to correctly cluster a $1 - O(1/K)$ fraction of the mixture, with arbitrarily high probability.*

As an example, we consider the case of a mixture of two Gaussians using a Gaussian kernel $h(r) = \exp(-r^2/(2\tau^2))$. In this case, matrix G_h can be computed in closed form, so that the conditions of Proposition 4.3 can be checked explicitly.

Corollary 4.5. *Consider a mixture of two Gaussians with $O(1)$-concentration in $\mathbb{R}^n$. Assuming that the variance of each Gaussian is $\Theta(n)$, for $\tau = \Theta(\sqrt{n})$, Gaussian kernel PCA allows to correctly cluster a $1 - O(1/K)$ fraction of the mixture if the distance between the centers is K.*

The choice of variance for the components in the above corollary is to fix ideas, similar conclusions would hold with other behaviors. The above guarantee matches the dimension-independent separation required by the PCA-based algorithms described in [71, 72] for example. Finally, the results in this section are in fact not strongly tied to the Hilbertian nature of positive kernels. More precisely, they may be easily extended to conditionally positive kernels, by simply restricting the involved quadratic forms to the space of zero mean functions. We omit further details.

4.4 Covariance based clustering

As shown in the above section, the approximation of kernel matrices provided by Theorem 4.1 is sufficient to conclude that their top eigenvectors are nearly constant on the clusters if the kernel is positive, which allows to correctly cluster the data. Unfortunately, while we showed that positive kernels could allow to cluster *e.g.* mixtures of Gaussians with different enough centers, the range of cases that can be successfully clustered using positive kernels remains unclear at this stage. In this section we show that by relaxing the positivity constraint, one can design kernels that can deal with more difficult situations, such as mixtures of distributions with common centers but different covariances. While Theorem 4.1 alone is insufficient for this purpose, we show that stronger conclusions can be obtained assuming that the components of the mixtures are supported on the sphere S with radius $\sqrt{n}$ and centered at the origin, and satisfy a Poincaré inequality. Namely, kernel matrices can then be approximated by block constant matrices, rather than a sum of column and row constant matrices within each block. We state below such a result for general kernels, assuming the input distributions are even. We also consider the case of non necessarily even distributions with small enough means. Similar conclusions can then be drawn for the kernel

$$h_t(r) = \cos\left(\frac{t}{\sqrt{n}}(n - r^2/2)\right)$$

where t is a parameter. The argument is more direct and avoids the use of Poincaré inequality. A more transparent way to write this kernel is to remark that for x and y on S,

$$h_t(\|x - y\|) = \cos\left(\frac{t}{\sqrt{n}} < x, y >\right)$$

Note that h_t has a perhaps non intuitive behavior compared to the most commonly used kernels as it oscillates $\Theta(\sqrt{n})$ times over the sphere S for $t = \Theta(1)$ for example.

Theorem 4.6. *Assume measures μ_i are supported on S, even, and satisfy a Poincaré inequality. Let $\tilde{h}(r) = h'(r)/r$. If the number of samples N is drawn according to the Poisson distribution with mean N_0, then with arbitrarily high*

probability, we have:

$$\|\Phi_h(X) - B\| \leq O\left(c'_h + \sqrt{n}c'_{\tilde{h}} + \|h\|_\infty \sqrt{\frac{n \log N_0}{N_0}}\right)$$

where the entries of B in the ij block are all equal to

$$G_h(i,j)/N = \frac{1}{N}\left(\int h(\|z - z'\|) d\mu_i(z) d\mu_j(z')\right)$$

For the kernel h_t, if measures μ_i are supported on S, have $O(1)$-concentration and if their means are at distance $O(1)$ from the origin, then:

$$\|\Phi_{h_t}(X) - B\| \leq O\left(\frac{t \log^3 n}{\sqrt{n}} + \sqrt{\frac{n \log N_0}{N_0}}\right)$$

with arbitrarily high probability for $t = O(1)$.

In particular, in the case of even distributions satisfying a Poincaré inequality, letting the sample size go to infinity, expliciting the upper bound in the first part of the theorem implies that for any fixed bounded function h with bounded derivatives up to the third order, the radial convolution operator from $L^2(\mathbb{R}^n, \mu_i)$ to $L^2(\mathbb{R}^n, \mu_j)$ has at most one singular value larger than $O(\log^3 n/\sqrt{n})$. It seems likely that the logarithmic factor can in fact be removed, by replacing the Lipschitz extension argument by a Dirichlet energy estimate in the proof of Theorem 4.1.

We now show that the second part of the above theorem can be used to cluster high dimensional mixtures based on the components covariance matrices. We assume that the components μ_i have $O(1)$-concentration and variance $\Theta(n)$. As the PCA algorithm of [71] allows to separate components whose means are at distance at least $\Omega(1)$ from the other means, it is sufficient to consider the case where all means are at distance $O(1)$ from the origin. We denote by Σ_i the non centered covariance matrix of μ_i. Given $s > 0$ and a symmetric matrix M, we define $f_s(M)$ to be the matrix having the same eigenvectors as M, eigenvalues

being transformed by function $\lambda \mapsto f_s(\lambda)$, with $f_s(\lambda) = \max(0, |\lambda| - s)$. Let

$$\Delta = \sqrt{n} \min_{u \neq v} \left\| \frac{\Sigma_u}{\text{trace}\Sigma_u} - \frac{\Sigma_v}{\text{trace}\Sigma_v} \right\|_2$$

As covariance matrices have trace $\Theta(n)$, they have Frobenius norm $\Theta(\sqrt{n})$, so that $\Delta = \Omega(\alpha_{min})$, α_{min} being the minimum angle between any two covariance matrices. Let further C_1, C_2 be two appropriate universal constants. The algorithm we propose is the following:

Algorithm 1 CovarianceClustering(X)

$\tilde{X}$ = data points projected on S

$\Phi = \Phi_{h_t}(\tilde{X})$, with $t = C_1\Delta$

Approximately solve the k-means problem for the columns of $f_{C_2\Delta^4}(\Phi)$

To prove that this algorithm succeeds, we apply Theorem 4.6 to the data projected on S, which tells us that $\Phi_{h_t}(\tilde{X})$ is well approximated by block constant matrix B. We then show that under our separation assumptions, matrices G_{h_t} are well-conditioned in the case of mixtures of two components. Using this fact, we show that the columns $f_{C_2\Delta^4}(B)$ corresponding to different components are sufficiently far apart. Applying a perturbation bound then allows to conclude, and obtain the following guarantee:

Theorem 4.7. *If $\Delta \geq Kn^{-1/6}\log^{5/3} n$, the above algorithm allows to correctly cluster a $O(1/K^6)$ fraction of the mixture with arbitrarily high probability, provided $N_0 \geq N_1$, with:*

$$N_1 = O\left(\log(n/\Delta)n^2/\Delta^2\right)$$

Hence clustering will succeed if the minimum angle α_{min} between the components covariances is larger than $O(n^{-1/6}\log^{5/3} n)$. First note that one case is not covered by this algorithm, namely the case where different components have covariance matrices differing only by a scaling. This situation can be dealt with easily by clustering the data according to the distance to the origin. A second remark can be made about the sample size. The guarantee given above aims for the smallest angular separation, and as a result requires a number of points that is

more than quadratic in the dimension. While it is possible that a better analysis would give smaller sample sizes in this regime, we remark that if $\alpha_{min} = \Omega(1)$, the proof can be modified to show that correct clustering will require only $O(n \log n)$ points. Indeed, in this situation, the error bound in Theorem 4.6 is dominated by the contribution of the sample size, and having $O(n \log n)$ points will make it small enough so that the rest of the analysis can be applied.

To conclude, we give some numerical results on specific examples of equal weight mixtures of two Gaussian distributions μ_1 and μ_2 with mean zero on $\mathbb{R}^n$, with n even. The covariances Σ_1 and Σ_2 are both diagonal in the standard basis. For a parameter $s > 0$, the eigenvalues of Σ_1 are $1+s$ on the first $n/2$ coordinates, and $1-s$ on the last $n/2$ coordinates. Eigenvalues of Σ_2 are reversed, so that $\Sigma_1+\Sigma_2 = 2I$, meaning that the whole distribution is isotropic. Under the assumptions of Theorem 4.7, as shown in the proof, the spectral soft thresholding operation used in the algorithm will leave at most 2 non zero eigenvalues. Rather than implementing the whole algorithm, we just plot the second dominant singular vector of Φ, as the first one turns out not to separate the components. Figure 4.1 shows it for $s = 0.9$, $n = 10$, $s = 0.6$, $n = 100$, $s = 0.33$, $n = 1000$ and $s = 0.2$, $n = 10000$, with $t = 0.1$. In all cases each Gaussian has n sample points. We see that the clusters are easily detected. Note that in the latter case, the Gaussians are nearly spherical, the relative error being of roughly 10% in terms of standard deviation.

4.5 Proofs

We give the proof of Theorem 4.1 in section 5.1, of Proposition 4.3 in Section 5.2, and of Corollaries 4.2 and 4.5 in Sections 5.3 and 5.4. Theorems 4.6 and 4.7 are proved in Sections 5.5 and 5.6.

4.5.1 Proof of Theorem 4.1

For technical reasons we will not work directly with the input measure μ, but rather with its empirical measure $\bar{\mu} = \sum_i w_i \bar{\mu}_i$, the number of samples being drawn according to a Poisson distribution with appropriately large mean M_0.

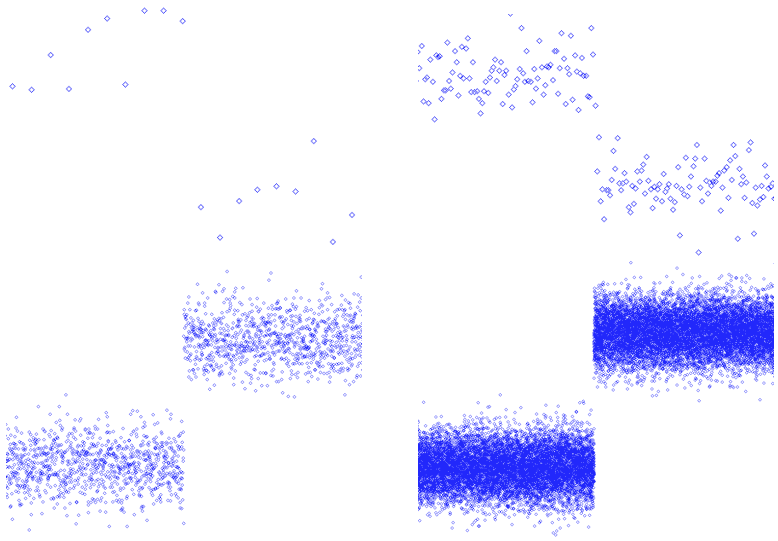

FIGURE 4.1: Second singular vector of Φ for isotropic mixtures of centered Gaussians.

Since the μ_i have $O(1)$-concentration, a vector X with law μ_i satisfies $\mathbb{E}(|X - \mathbb{E}X\|^q)^{1/q} = O(\sqrt{n})$ for constant $q \geq 1$, which implies (see *e.g.* [79]) that

$$\mathbb{E}(W_l(\mu_i, \bar{\mu}_i)) = O(nM_0^{-1/n})$$

where W_l are the transportation distances for $l = 1$ or 2. By Markov inequality, for any $\delta > 0$, these distances are at most δ with probability at least $1 - p$, with

$$p = O\left(\frac{nM_0^{-1/n}}{\delta}\right)$$

Consider the map

$$\begin{aligned} \phi_{\bar{\mu}} : \mathbb{R}^n &\to L^2(\mathbb{R}^n, \bar{\mu}) \\ x &\mapsto \phi_{\bar{\mu}}(x) = h(\|x - \cdot\|) \end{aligned}$$

The gist of our proof of Theorem 4.1 is as follows. We first observe that the directional derivatives of $\phi_{\bar{\mu}}$ at each point satisfy a Lipschitz condition with a small constant. More precisely, this is true after modifying them in a small region, which is enough for our purposes. Using concentration of measure, this implies that these derivatives, modulo piecewise constant functions on the components, are small. This can be further reinterpretated as saying that $\phi_{\bar{\mu}}$, after centering

on each component, has a small Lipschitz constant. Because each component has constant concentration by assumption, this implies that the image of each component by $\phi_{\bar{\mu}}$, after centering on each component, has small concentration. The desired claim on the block structure of $\Phi_h(X)$ can then be deduced.

4.5.1.1 A property of $\phi_{\bar{\mu}}$

Let $E \in L^2(\mathbb{R}^n, \bar{\mu})$ be the space of functions that are constant on the support of each μ_i, and P_E and $P_{E^\perp}$ denote the orthogonal projectors onto E and $E^\perp$. Further denote by S the sphere with radius $\sqrt{n}$ centered at 0.

Proposition 4.8. *With probability at least $1-p$, for any x_1 and $x_2 \in \mathbb{R}^n$,*

$$\|P_{E^\perp}\phi_{\bar{\mu}}(x_1) - P_{E^\perp}\phi_{\bar{\mu}}\mu(x_2)\| \leq O(c_h(\delta))\|x_1 - x_2\|$$

Furthermore if measures μ_i are supported on S and their mean is $O(1)$, then with probability at least $1-p$, for any x_1 and x_2 in S:

$$\|P_{E^\perp}\phi_{\bar{\mu}}(x_1) - P_{E^\perp}\phi_{\bar{\mu}}(x_2)\| \leq O(c'_h(\delta))\|x_1 - x_2\|$$

with

$$\begin{aligned} c_h(\delta) &= (1+\delta)c_h + \sqrt{\delta}\|h'\|_\infty \\ c'_h(\delta) &= (1+\delta)c'_h + \sqrt{\delta}\|h'\|_\infty \end{aligned}$$

To prove the first part of Proposition 4.8 we argue that

$$\begin{aligned} \|P_{E^\perp}\phi_{\bar{\mu}}(x_1) - P_{E^\perp}\phi_{\bar{\mu}}(x_2)\|_2 &\leq \sup_{x_0, v, \|v\|=1} \left\| \left.\frac{d}{dx}\right|_{v, x=x_0} P_{E^\perp}\phi_{\bar{\mu}} \right\|_2 \|x_1 - x_2\| \\ &\leq \sup_{x_0, v, \|v\|=1} \left\| P_{E^\perp} \left.\frac{d}{dx}\right|_{v, x=x_0} \phi_{\bar{\mu}} \right\|_2 \|x_1 - x_2\| \\ &\leq \sup_{x_0, v, \|v\|=1} \left(\sum_i w_i \left\| f_i - \int f_i d\bar{\mu}_i \right\|_2^2 \right)^{1/2} \|x_1 - x_2\| \end{aligned}$$

where in the last line f_i denotes the directional derivative of $\phi_{\bar{\mu}_i}$ at x_0 in direction v. To conclude, it is sufficient to prove that

$$\sup_{x_0,v,\|v\|=1} \left\| f_i - \int f_i d\bar{\mu}_i \right\|_2 \leq O(c_h(\delta)) \tag{4.2}$$

For the second part, we use a similar argument except that we interpolate between x_1 and x_2 using a great circle on S instead of a straight line. This shows that establishing

$$\sup_{x_0,v,\|v\|=1,\langle v,x_0\rangle=0} \left\| f_i - \int f_i d\bar{\mu}_i \right\|_2 \leq O(c'_h(\delta)) \tag{4.3}$$

suffices to conclude. Proving these two inequalities is the point of the rest of this section. For some $\rho > 0$, let

$$L_{v,\rho} = \{y| \ |\langle y, v\rangle| \leq 1/\rho\}$$

Further define:

$$\begin{aligned} d_h &= \sup_{r\geq\rho R} \left(|h''(r)| + \frac{1}{r}|h'(r)| \right) \\ d'_h &= \sup_{r\geq\rho R} \left(\frac{|h''(r)|}{\rho^2 r} + \frac{1}{r}|h'(r)| \right) \end{aligned}$$

Lemma 4.9. *Function f_i is d_h-Lipschitz outside $B(x_0, \rho R)$ and $|f_i|$ is bounded everywhere by* $\sup|h'|$. *Furthermore, if v is a unit tangent vector at $x_0 \in S$, then f_i is d'_h-Lipschitz on $L_{v,\rho} \setminus B(x_0, \rho R)$.*

Proof. We consider the radial coordinate system (r, θ) centered at x_0, where θ denotes the angle formed by $y - x_0$ and v. A direct calculation shows that

$$f_i(y) = \left.\frac{\mathrm{d}}{\mathrm{d}x}\right|_{v,x=x_0} \phi_{\bar{\mu}_i}(x)(y) = h'(r)\cos\theta$$

Hence

$$\begin{aligned} \frac{d}{dr} f_i(r(y), \theta(y)) &= h''(r)\cos\theta \\ \frac{d}{d\theta} f_i(r(y), \theta(y)) &= -h'(r)\sin\theta \end{aligned}$$

Noticing that r is a 1-Lipschitz function of y, and that $|d\theta/dy| \leq 1/r$ allows to bound the derivatives of f_i in the radial and tangent directions using the chain rule, implying:

$$\begin{aligned} \|\nabla f_i(y)\| &= \left((h''(r(y))\cos\theta(y))^2 + \left(\frac{h'(r(y))}{r(y)} \sin\theta(y) \right)^2 \right)^{1/2} \\ &\leq \max\left(h''(r(y))\,|\cos\theta(y)|, \frac{h'(r(y))}{r(y)} \right) \end{aligned}$$

Using that $|\cos\theta(y)| \leq 1/(\rho^2 r)$ on $L_{v,\rho} \setminus B(x_0, \rho R)$, the conclusion follows. □

Lemma 4.10. *We can write $f_i = \tilde{f}_i + g_i$, where $\tilde{f}_i$ is d_h-Lipschitz, and g_i is supported on $B(x_0, \rho R)$ with $||g_i||_\infty \leq 2\sup_r |h'(r)|$. If v is a unit tangent vector at $x_0 \in S$, then we can find a similar decomposition with $\tilde{f}_i$ d'_h-Lipschitz and g_i supported on $B(x_0, \rho R) \cup \mathbb{R}^n \setminus L_{v,\rho}$ with $||g_i||_\infty \leq 2\sup_r |h'(r)|$.*

Proof. Define $\tilde{f}_i$ to be a d_h-Lipschitz extension of $f_i|_{\mathbb{R}^n \setminus B(x_0,\rho R)}$ to $\mathbb{R}^n$, which exists by Kirszbraun's extension theorem [80]. We choose $\tilde{f}_i$ such that $\sup_{B(x_0,\rho R)} |\tilde{f}_i| = \sup_{\partial B(x_0,\rho R)} |\tilde{f}_i|$, which can be done by thresholding if necessary. The result follows by letting $g_i = f_i - \tilde{f}_i$. The spherical case is proved similarly. □

Lemma 4.11. *With probability at least $1-p$, we have $\mathrm{Var}_{\bar{\mu}_i}(f_i) = O(c_h(\delta)^2)$. If measures μ_i are supported on S with mean $O(1)$, then with probability at least $1-p$, for v a unit tangent vector at $x_0 \in S$, $\mathrm{Var}_{\bar{\mu}_i}(f_i) = O(c'_h(\delta)^2)$.*

Proof. For the first claim, we write

$$\begin{aligned} \sqrt{\mathrm{Var}_{\bar{\mu}_i}(f_i)} &\leq \sqrt{\mathrm{Var}_{\bar{\mu}_i}(\tilde{f}_i)} + \sqrt{\mathrm{Var}_{\bar{\mu}_i}(g_i)} \\ &\leq \sqrt{\mathrm{Var}_{\bar{\mu}_i}(\tilde{f}_i)} + ||g_i||_2 \\ &\leq \sqrt{\mathrm{Var}_{\bar{\mu}_i}(\tilde{f}_i)} + \sup|g_i|\bar{\mu}_i(B(x_0,\rho R))^{1/2} \end{aligned}$$

Because $\tilde{f}_i$ is d_h-Lipschitz, the pushforwards of μ_i and $\bar{\mu}_i$ satisfy

$$W_2(\tilde{f}_{i\sharp}\bar{\mu}_i, \tilde{f}_{i\sharp}\mu_i) \leq d_h W_2(\bar{\mu}_i, \mu_i) \leq d_h \delta$$

And since μ_i has $O(1)$-concentration, $\tilde{f}_{i\sharp}\mu_i$ has at most $O(d_h^2)$ variance. As a result

$$\mathrm{Var}_{\bar{\mu}_i}(\tilde{f}_i) = \mathrm{Var}\tilde{f}_{i\sharp}\bar{\mu}_i \leq O((1+\delta^2)d_h^2)$$

Also, letting d_{x_0} be the distance function to x_0, we have that

$$W_1(d_{x_0\sharp}(\bar{\mu}_i), d_{x_0\sharp}(\mu_i)) \leq \delta$$

since distance functions are 1-Lipschitz. Consider an optimal coupling (X, Y) between $d_{x_0\sharp}(\bar{\mu}_i)$ and $d_{x_0\sharp}(\mu_i)$. By Markov inequality, the probability that $X \leq \rho R$ and $Y \geq \rho R + 1$ is at most δ. This implies that

$$\bar{\mu}_i(B(x_o, \rho R)) \leq \delta + \mu_i(B(x_o, \rho R + 1))$$

Since d_{x_0} $O(1)$-concentrates on μ_i, its median is $O(1)$ close to $(\int d_{x_0}^2 d\mu_i)^{1/2}$. As the latter quantity is at least R, we have by concentration

$$\mu_i(B(x_0, \rho R + 1)) \leq \exp\left(-\Omega(1-\rho)R + O(1)\right)$$

As a consequence

$$\sqrt{\mathrm{Var}_{\bar{\mu}_i}(f_i)} \leq O\left((1+\delta^2)^{1/2} d_h + \sup_r |h'(r)| \left(\delta + \exp\left(-\Omega(1-\rho)R\right)\right)^{1/2}\right)$$

The first claim follows by setting $\rho = 1/2$. The spherical case is proved similarly, except that we use the inequalities

$$\bar{\mu}_i\left(B(x_0, \rho R) \cup \mathbb{R}^n \setminus L_{v,\rho}\right) \leq \bar{\mu}_i\left(B(x_0, \rho R)\right) + \bar{\mu}_i\left(\mathbb{R}^n \setminus L_{v,\rho}\right)$$

and

$$\begin{aligned}
\bar{\mu}_i\left(\mathbb{R}^n \setminus L_{v,\rho}\right) &\leq \delta + \mu_i\left(\{y \mid |\langle y, v\rangle| \geq 1/\rho - 1\}\right) \\
&\leq \delta + 2\exp\left(-\Omega(1/\rho) + O(1)\right) \\
&\leq \delta + O(\exp(-\Omega(1/\rho)))
\end{aligned}$$

which follows as above from the fact that linear functions $O(1)$-concentrate on μ_i and have mean $O(1)$. Choosing appropriate $\rho = \Theta(\log(R)^{-1})$, the bound above

becomes $\delta + 1/R^2$, hence

$$\begin{aligned}\sqrt{\mathrm{Var}_{\bar{\mu}_i}(f_i)} &\leq O\left((1+\delta^2)^{1/2} d'_h + \sup_r |h'(r)| \left(\delta + 1/R^2\right)^{1/2}\right) \\ &\leq O(c'_h(\delta))\end{aligned}$$

□

This proves (4.2) and (4.3) and concludes the proof of Proposition 4.8.

4.5.1.2 Decomposition of $\Phi_h(X)$

We first show the following variant of Theorem 4.1:

Proposition 4.12. *If the number of samples M is drawn according to the Poisson distribution with mean M_0, then with probability at least $1-p$, we have $\|\Phi_h(X) - A\| = O(e_h(\delta))$ with $e_h(\delta) = c_h(\delta)(1+\delta) + \delta\|h'\|_\infty$, and $\|\Phi_h(X) - A\| = e'_h(\delta)$ with $e'_h(\delta) = c'_h(\delta)(1+\delta) + \delta\|h'\|_\infty$ in the spherical case.*

The argument is the same for the spherical and for the non-spherical case, so we only consider the non spherical case. Let M be the number of samples of $\bar{\mu}$. First decompose the unnormalized kernel matrix $D_h(X) = M\Phi_h(X)$ as follows:

$$D_h(X) = P_E D_h(X) + P_{E^\perp} D_h(X)$$

The first term $P_E D_h(X)$ is column constant within each block. We now focus on the second one.

Lemma 4.13. *With probability at least $1-p$, the centered covariance matrix of the columns of $P_{E^\perp} D_h(X)$ corresponding to any component has eigenvalues at most $O(M c_h(\delta)^2(1+\delta^2))$.*

Proof. The columns of $P_{E^\perp} D_h(X)$ are the images of the sample points by $P_{E^\perp}\phi_{\bar{\mu}}$, expressed in the standard basis. Hence by Proposition 4.8, the map $\bar{\phi}$ associating each sample point with its column in $P_{E^\perp} D_h(X)$ is $O(\sqrt{M} c_h(\delta))$-Lipschitz with probability at least $1-p$. Let $\tilde{\phi}$ be a $O(\sqrt{M} c_h(\delta))$-Lipschitz extension of $\bar{\phi}$ to $\mathbb{R}^n$.

Consider a unit vector $v \in \mathbb{R}^M$ and let U be a random column of $P_{E^\perp} D_h(X)$. Variable $\langle U, v\rangle$ is equal to $\langle \bar{\phi}(V), v\rangle = \langle \tilde{\phi}(V), v\rangle$, where V is drawn according to $\bar{\mu}_i$. Let now W be drawn according to μ_i. Since μ_i has $O(1)$-concentration, $\langle \tilde{\phi}(W), v\rangle$ has variance $O(Mc_h(\delta)^2)$. Because with probability at least $1-p$, $W_2(\bar{\mu}_i, \mu_i) < \delta$, the distributions of $\langle \tilde{\phi}(W), v\rangle$ and $\langle \tilde{\phi}(V), v\rangle$ are $O(\sqrt{M}c_h(\delta)\delta)$ away in the W_2 distance. As a consequence

$$\mathrm{Var}(\langle \tilde{\phi}(V), v\rangle) = O(\mathrm{Var}(\langle \tilde{\phi}(V), v\rangle) + Mc_h(\delta)^2\delta^2) = O(Mc_h(\delta)^2(1+\delta^2))$$

□

Let us further decompose

$$P_{E^\perp} D_h(X) = P_{E^\perp} D_h(X) P_E + P_{E^\perp} D_h(X) P_{E^\perp}$$

as a sum of matrix $P_{E^\perp} D_h(X) P_E$ which is row constant within each block, and a remainder $M.B = P_{E^\perp} D_h(X) P_{E^\perp}$ whose columns are the columns of $P_{E^\perp} D_h(X)$ centered in each block. By Lemma 4.13, the non centered covariance matrix of all the columns of $M.B$ has eigenvalues at most $O(Mc_h(\delta)^2(1+\delta^2))$. As this covariance matrix is $M.BB^t$, this shows that $||B|| = O(c_h(\delta)(1+\delta))$. Thus we get:

$$\Phi_h(X) = P_E \Phi_h(X) + P_{E^\perp} \Phi_h(X) P_E + B$$

Letting $\bar{A} = P_E \Phi_h(X) + P_{E^\perp} \Phi_h(X) P_E$, we see that for $x \in \mathrm{support}(\bar{\mu}_i)$ and $y \in \mathrm{support}(\bar{\mu}_j)$, the xy entry of $\bar{A}$ is given by

$$M.\bar{A}_{xy} = \int h(\|x-z\|) d\bar{\mu}_j(z) + \int h(\|y-z'\|) d\bar{\mu}_i(z') - \int h(\|z-z'\|) d\bar{\mu}_i(z) d\bar{\mu}_j(z')$$

By Kantorovich-Rubinstein theorem,

$$\|A - \bar{A}\| \leq \sup_{xy} |M.\bar{A}_{xy} - M.A_{xy}| \leq O(\delta \|h'\|_\infty)$$

which concludes the proof.

4.5.1.3 Sample size

In order to prove that Theorem 4.1 also holds for small sample size, we use the following result in [81]. For a random variable W, let $E_k W$ denotes the L_k norm of W. For a matrix U, $\|U\|_\infty$ is the maximum entry of U, and $\|U\|_{1,2}$ is the maximum norm of the columns of U.

Theorem. Let Z be a $M \times M$ Hermitian matrix, decomposed into diagonal and off-diagonal parts: $Z = D + H$. Fix k in $[2, \infty)$, and set $q = \max\{k, 2\log M\}$. Then

$$E_k\|RZR\| \leq O\left(qE_k\|RHR\|_\infty + \sqrt{\eta q}E_k\|HR\|_{1,2} + \eta\|H\|\right) + E_k\|RDR\|$$

where R is a diagonal matrix with independent $0-1$ entries with mean η.

Let us apply this theorem to $Z = M(\Phi_h(X) - A_M)$, where X is an iid sample of μ with cardinality M distributed according to a Poisson distribution with mean M_0, and A_M is the matrix specified in Theorem 4.1. In any case $\|RZR\| \leq O(\text{trace}(R)\|h\|_\infty)$, and by Proposition 4.12, with probability at least $1-p$, we have $\|Z\| \leq O(Me_h(\delta))$ (and similarly for the spherical case). Clearly $E_k\|RDR\|$ and $E_k\|RHR\|_\infty$ are both bounded by $O(\|h\|_\infty)$, and $E_k\|HR\|_{1,2}$ is at most $O(\|h\|_\infty E_k\sqrt{M})$. Also $\|H\| \leq \|Z\| + \|D\| \leq O(Me_h(\delta) + \|h\|_\infty)$ with probability at least $1-p$. Hence the theorem above gives:

$$E_k\|RZR\| \leq \|h\|_\infty O(pE_k\text{trace}(R) + q + \sqrt{\eta q}E_k\sqrt{M} + \eta) + O(\eta e_h(\delta)E_k M)$$

Taking $k = 2$ and $\eta = N_0/M_0$, we have $E_k\text{trace}(R) = O(N_0)$, $E_k\sqrt{M} = O(\sqrt{M_0})$ and $E_k M = O(M_0)$. With $q = 2\log M_0$, we get

$$\begin{aligned}
E_2\left(\frac{\|RZR\|}{\text{trace}(R)}\right) &\leq O\left(E_2\left(\frac{\|RZR\|}{N_0}\right)\right) \\
&\leq \|h\|_\infty O\left(p + \frac{\log M_0}{N_0} + \sqrt{\frac{\log M_0}{N_0}} + \frac{1}{M_0}\right) + O(e_h(\delta)) \\
&\leq \|h\|_\infty O\left(\frac{n}{\delta M_0^{1/n}} + \sqrt{\frac{\log M_0}{N_0}}\right) + O(e_h(\delta)) \\
&\leq \|h\|_\infty O\left(\frac{n}{\delta M_0^{1/n}} + \sqrt{\frac{\log M_0}{N_0}} + \delta + (1+\delta)\sqrt{\delta}\right) + (1+\delta)^2 O(c_h)
\end{aligned}$$

assuming $N_0 \geq \log M_0$. Matrix $RZR/\text{trace}(R)$ is simply $\Phi_h(Y) - A_N$, where Y is an iid sample of μ with cardinality N distributed according to a Poisson distribution with mean N_0. Continuing the last equation, taking $M_0 = N_0^{3n/2}$ and $\delta = (n/M_0^{1/n})^{2/3}$ so that $p = \Theta(\sqrt{\delta})$, we have

$$\begin{aligned} E_2(\|\Phi_h(Y) - A_N\|) &\leq \|h\|_\infty O\left(\frac{n}{\delta M_0^{1/n}} + \sqrt{\frac{\log M_0}{N_0}}\right) + O(c_h) \\ &\leq O\left(c_h + \|h\|_\infty \sqrt{\frac{n \log N_0}{N_0}}\right) \end{aligned}$$

The conclusion follows by applying Markov inequality.

4.5.2 Proof of Proposition 4.3

We want to show that for a positive kernel, the space spanned by the k top eigenvectors of $\Phi_h(X)$ is close to the space of piecewise constant functions E. We first observe that for a large enough number of samples, matrix G_h is close to its finite sample version $\widehat{G_h}$, whose ij entry is the average of the kernel over $X_i \times X_j$:

Lemma 4.14. *For any $c > 0$, we have:*

$$P\left(||G_h - \widehat{G_h}|| \geq c\right) \leq 1 - O\left(N_0 \exp\left(-N_0 \Omega\left(\min\left(\frac{c}{\|h'\|_\infty}, \frac{c^2}{\|h'\|_\infty^2}\right)\right)\right)\right)$$

Proof. The desired operator norm can be bounded using entries magnitude as follows:

$$\begin{aligned} P\left(||G_h - \widehat{G_h}|| \geq c\right) &\leq P\left(||G_h - \widehat{G_h}||_2^2 \geq c^2\right) \\ &\leq \max_{ij} P\left(|G_h(i,j) - \widehat{G_h}(i,j)| \geq c/k\right) \end{aligned} \tag{4.4}$$

In order to control the error on entry ij, we write:

$$\begin{aligned} G_h(i,j) - \widehat{G_h}(i,j) &= \frac{1}{N_i N_j} \sum_{x \in X_i, y \in X_j} h(||x-y||) - \int h(||x-y||) d\mu_i(x) d\mu_j(y) \\ &= \frac{1}{N_i} \sum_{x \in X_i} \frac{1}{N_j} \sum_{y \in X_j} \left(h(||x-y||) - \int h(||x-y||) d\mu_j(y) \right) \\ &\quad + \frac{1}{N_i} \sum_{x \in X_i} \left(\int h(||x-y||) d\mu_j(y) - \int h(||x-y||) d\mu_i(x) d\mu_j(y) \right) \end{aligned}$$

Since $\|h'\|_\infty$ is the Lipschitz constant of $||h(x-.)||$, we see by concentration that for fixed x and for y distributed according to μ_j:

$$\|h(||x-y||) - \int h(||x-y||) d\mu_j(y)\|_{\psi_1} = O(\|h'\|_\infty)$$

where for a random variable U, $\|U\|_{\psi_1} = \sup_{p \geq 1} p^{-1} \left(E\|U\|^p\right)^{1/p}$ is its Orlicz ψ_1 norm. As a consequence, conditionally to N_j, this implies (Corollary 5.17 in [82]) that for any $\varepsilon > 0$:

$$P\left(|S_x| \geq \varepsilon\right) \leq 2 \exp\left(-N_j \Omega \left(\min\left(\frac{\varepsilon}{\|h'\|_\infty}, \frac{\varepsilon^2}{\|h'\|_\infty^2} \right) \right) \right)$$

with

$$S_x = \frac{1}{N_j} \sum_{y \in X_j} \left(h(||x-y||) - \int h(||x-y||) \right)$$

Hence by the union bound:

$$\begin{aligned} P\left(\left| \frac{1}{N_i} \sum_{x \in X_i} S_x \right| \geq \varepsilon \right) &\leq 2N_i \exp\left(-N_j \Omega \left(\min\left(\frac{\varepsilon}{\|h'\|_\infty}, \frac{\varepsilon^2}{\|h'\|_\infty^2} \right) \right) \right) \\ &\leq O\left(N_0 \exp\left(-N_0 \Omega \left(\min\left(\frac{\varepsilon}{\|h'\|_\infty}, \frac{\varepsilon^2}{\|h'\|_\infty^2} \right) \right) \right) \right) \end{aligned}$$

Similarly, as the Lipschitz constant of $\int h(||. - y||) d\mu_j(y)$ is at most $\|h'\|_\infty$ as well, we get:

$$P\left(|U| \geq \varepsilon\right) \leq 2 \exp\left(-N_0 \Omega \left(\min\left(\frac{\varepsilon}{\|h'\|_\infty}, \frac{\varepsilon^2}{\|h'\|_\infty^2} \right) \right) \right)$$

with

$$U = \frac{1}{N_i} \sum_{x \in X_i} \left(\int h(||x-y||) d\mu_j(y) - \int h(||x-y||) d\mu_i(x) d\mu_j(y) \right)$$

The last two inequalities together with (4.4) imply the desired claim. □

Let now $\widehat{M_h}$ be the matrix obtained from $\widehat{G_h}$ by multiplying the ij entry by $\sqrt{w_i w_j}$. Applying the above lemma with $c = c_h$, its smallest eigenvalue can be lower bounded as follows:

$$\lambda_1(\widehat{M_h}) = \Omega(\lambda_1(\widehat{G_h})) = \Omega(\lambda_1(G_h) - c_h) = \Omega(Kc_h)$$

with arbitrarily high probability, assuming $N_0 = \Omega(\|h'\|_\infty^2 / c_h^2)$.

Now, note that $\widehat{M_h}$ is the matrix of the quadratic form $\Phi_h(X)$ restricted to E. More precisely, the indicator functions of the clusters, normalized to have unit L_2-norm, form an orthornormal basis of E, and writing that quadratic form in this basis gives $\widehat{M_h}$. Let λ be the smallest eigenvalue of $\widehat{M_h}$. By the variational characterization of eigenvalues, there exist at least k eigenvalues of $\Phi_h(X)$ that are at least λ. Let H denote the space spanned by the k-top eigenvectors of $\Phi_h(X)$, and let L denote the space spanned by the remaining $N-k$. We show using a perturbation argument that the maximum of the principal angles between space E and space H is small.

Let $x \in E^\perp$ be a unit vector. We may write $x = \alpha x_L + \beta x_H$ with $\alpha^2 + \beta^2 = 1$, and x_L and x_H are unit vectors belonging respectively to L and H. Then:

$$x^t \Phi_h(X) x = \alpha^2 x_L^t \Phi_h(X) x_L + \beta^2 x_H^t \Phi_h(X) x_H$$

Since $x \in E^\perp$, we have $x^t A x = 0$, where A is the matrix defined in Theorem 4.1. Hence by Theorem 4.1, with arbitrarily high probability:

$$x^t \Phi_h(X) x \leq O(c_h)$$

provided

$$N_0 = \Omega\left(\frac{n\|h\|_\infty^2}{c_h^2} \log\left(\frac{n\|h\|_\infty^2}{c_h^2} \right) \right)$$

Also, by assumption:

$$x_H^t \Phi_h(X) x_H \geq \lambda \geq K\Omega(c_h)$$

As a consequence:

$$d(x, L) = \beta \leq O(1/\sqrt{K})$$

That is, the maximum angle between the $(N-k)$-flats $E^\perp$ and L is $O(1/\sqrt{K})$. Hence, so is the maximum angle between their orthogonals E and H, which is the desired claim.

4.5.3 Proof of Corollary 4.6

Let $E^\perp \in \mathbb{R}^N$ be the space of vectors whose mean is zero on each block. This space has codimension k. Now, for any vector $x \in E^\perp$, we easily see that $x^t A x = 0$, where A is the matrix from Theorem 4.1. As a result, the quadratic form $\Phi_h(X)$ is at most $O\left(c_h + \|h\|_\infty \sqrt{\frac{n \log N_0}{N_0}}\right) I$ on $E^\perp$ with arbitrarily high probability, implying that $\Phi_h(X)$ has at least $(N-k)$ eigenvalues that are at most $O\left(c_h + \|h\|_\infty \sqrt{\frac{n \log N_0}{N_0}}\right)$. Applying the same argument to $-\Phi_h(X)$, the result follows.

4.5.4 Proof of Corollary 4.5

Matrix G_h has entries

$$G_h(i,j) = \mathbb{E} h(||x_i - x_j||)$$

where x_i are independant random variables with law $\mathcal{N}(\mu_i, \Sigma_i)$, where μ_i and Σ_i are the means and covariances of the two Gaussians in the mixture.

Lemma 4.15. *If u is a centered Gaussian random variable with covariance Σ, then:*

$$\mathbb{E}(h(||u||)) = \det\left(I + \frac{1}{\tau^2}\Sigma\right)^{-\frac{1}{2}}$$

Proof.

$$\begin{aligned}
\mathbb{E}(h(||u||)) &= \int \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}x^t\left(\Sigma^{-1}+\frac{1}{\tau^2}I\right)x\right)dx \\
&= \frac{\det\left(\left(\Sigma^{-1}+\frac{1}{\tau^2}I\right)^{-1}\right)^{1/2}}{\det(\Sigma)^{1/2}} \\
&= \det\left(I+\frac{1}{\tau^2}\Sigma\right)^{-\frac{1}{2}}
\end{aligned}$$

□

By standard algebraic manipulations, shifting the center amounts to scaling the expectation by a certain factor:

Lemma 4.16. *If u is a Gaussian random variable with covariance Σ and mean μ, then:*

$$\mathbb{E}(h(||u||)) = \exp\left(-\frac{1}{\tau^2}\mu^t(I-(I+\tau^2\Sigma^{-1})^{-1})\mu\right)\ \det\left(I+\frac{1}{\tau^2}\Sigma\right)^{-\frac{1}{2}}$$

In particular, letting B_h be the 2×2 matrix with entries

$$B_h(i,j) = \mathbb{E}h(||y_i - y_j||)$$

where y_i are independent random variables with law $\mathcal{N}(0,\Sigma_i)$, we see that G_h is obtained from B_h by scaling the off diagonal entries by a factor λ that is at most

$$\begin{aligned}
\exp\left(-\frac{1}{\tau^2}\mu^t(I-(I+\tau^2(\Sigma_1+\Sigma_2)^{-1})^{-1})\mu\right) &\leq \exp\left(-\frac{1-(1+2\tau^2)^{-1}}{\tau^2}||\mu_1-\mu_2||^2\right) \\
&= 1-\Theta\left(\frac{||\mu_1-\mu_2||^2}{n}\right)
\end{aligned}$$

Because $\det B_h$ is non negative and the entries of B_h are $\Theta(1)$, we deduce that

$$\begin{aligned}
\det G_h &= \det B_h + (B_h)_{12}^2 - \lambda^2(B_h)_{12}^2 \\
&\geq (1-\lambda^2)(B_h)_{12}^2 \\
&= \Theta\left(\frac{||\mu_1-\mu_2||^2}{n}\right)
\end{aligned}$$

Now, the largest entries of G_h are the same as for B_h, that is, $\Theta(1)$, which implies that the maximal eigenvalue of G_h is $\Theta(1)$ as well. From this we see that:

$$\lambda_1(G_h) = \Theta\left(\frac{||\mu_1 - \mu_2||^2}{n}\right)$$

To conclude, it suffices to check that for our choice of kernel and assumptions on the variance of the Gaussians, $c_h = \Theta(1/n)$.

4.5.5 Proof of Theorem 4.6

We first show that constant functions are sent to nearly constant functions by the convolution operator with kernel h from $L^2(\mathbb{R}^n, \mu_i)$ to $L^2(\mathbb{R}^n, \mu_j)$.

Lemma 4.17. *Let $f_i(x) = \int h(\|y-x\|)d\mu_i(y)$, and $\bar{f}_i(x) = \int \tilde{h}(\|y-x\|)d\mu_i(y)$. If μ_i and μ_j are supported on the sphere S, even, and satisfy a Poincaré inequality, then:*

$$\mathrm{Var}_{\mu_j} f_i = O\left(nc_{\tilde{h}}'^2\right)$$

Proof. The gradient of f_i is as follows:

$$\nabla f_i(x) = \int (x-y)\tilde{h}(\|x-y\|)d\mu_i(y)$$

For $x \in S$, the gradient of the restriction of f_i to S is

$$\nabla f_{i|S}(x) = P_{T_xS}\nabla f_i(x) = -P_{T_xS}\left(\int y\tilde{h}(\|x-y\|)d\mu_i(y)\right) \tag{4.5}$$

Denoting by $M : L^2(\mathbb{R}^n, \mu_i) \to L^2(\mathbb{R}^n, \mu_j)$ the operator defined by

$$Mg(x) = \int g(y)\tilde{h}(\|x-y\|)d\mu_i(y)$$

From the structure of blocks described in Theorem 4.1, and letting the sample size go to infinity, we get that $\|M - M'\| = O(c'_{\tilde{h}})$, where

$$M'g(x) = \int g(y)M'_{xy}d\mu_i(y)$$

and

$$M'_{xy} = \int \tilde{h}(\|y - z'\|)d\mu_i(z') + \int \tilde{h}(\|x - z\|)d\mu_j(z) - \int \tilde{h}(\|z - z'\|)d\mu_i(z)d\mu_j(z')$$

Calling y the coordinate vector of S, that is, the identity map of S, the above equation expresses $M'y$ as the sum of two terms T_1 and T_2. The first one is

$$T_1 = \int y \left(\int \tilde{h}(\|y - z'\|)d\mu_i(z') \right) d\mu_i(y) = \int y\bar{f}_i(y)d\mu_i(y)$$

we see that as μ_i is even, $\bar{f}_i(y)$ is an even function of y. Hence multiplying it by y gives an odd function whose integral against μ_j must be be zero as μ_j is even as well. Hence T_1 vanishes. The second term T_2 is

$$T_2 = \left(\int y d\mu_i(y) \right) \left(\int \tilde{h}(\|x - z\|)d\mu_j(z) - \int \tilde{h}(\|z - z'\|)d\mu_i(z)d\mu_j(z') \right)$$

As μ_i is even, it has zero mean so T_2 cancels. From (4.5), the above discussion gives:

$$\begin{aligned} \|\nabla f_{i|S}\|^2 &\leq \|My\|^2 \\ &\leq \|(M - M')y\|^2 \\ &\leq O\left(c'^2_{\tilde{h}}\|y\|^2\right) \\ &\leq O\left(nc'^2_{\tilde{h}}\right) \end{aligned}$$

The desired claim follows using Poincaré inequality. □

Lemma 4.18. *Taking $h = h_t$, we have:*

$$\mathrm{Var}_{\mu_j} f_i \leq O(t^2/n)$$

assuming μ_i and μ_j are supported on S, have $O(1)$ means and $O(1)$-concentration.

Proof. For any x, y in S we can write

$$h_t(\|x - y\|) = \mathrm{Re}\ \exp\left(\frac{it}{\sqrt{n}} < x, y > \right)$$

Hence

$$f_i(x) \quad = \quad \text{Re}\left(\widehat{\mu_i}(-tx/\sqrt{n})\right)$$

As a consequence, for any unit vector u:

$$\begin{aligned} |<\nabla f_i(x), u>| &\leq \frac{t}{\sqrt{n}}|<\nabla\widehat{\mu_i}(-tx/\sqrt{n}), u>| \\ &\leq \frac{t}{\sqrt{n}}|\widehat{\mu_i^u}(-tx/\sqrt{n})| \\ &\leq \frac{t}{\sqrt{n}}O(\|\mu_i^u\|_1) \\ &\leq O(t/\sqrt{n}) \end{aligned}$$

where μ_i^u is μ_i multiplied by function $x \mapsto < x, u >$, the last line using the fact that μ_i has $O(1)$-concentration and $O(1)$ mean. Hence f_i is $O(t/\sqrt{n})$-Lipschitz. The lemma follows since μ_j has $O(1)$-concentration. □

To prove the first part of Theorem 4.6, using Theorem 4.1, it is sufficient to show that with arbitrarily high probability $\|A - B\| = O(U)$ where

$$U = \sqrt{n}c'_{\tilde{h}}$$

and A is the matrix given by Theorem 4.1. By definition of A, and after a small manipulation, we see that the entries of $A - B$ in the ij block are given by

$$\begin{aligned} (A-B)_{xy} &= \frac{1}{N}\left(\int h(\|x-z\|)d\mu_j(z) - \int h(\|x-z\|)d\mu_j(z)d\mu_i(x)\right) \\ &\quad + \frac{1}{N}\left(\int h(\|y-z'\|)d\mu_i(z') - \int h(\|y-z'\|)d\mu_i(z')d\mu_j(y)\right) \\ &= \frac{1}{N}\left(\left(f_j(x) - \int f_j(x)d\mu_i(x)\right) + \left(f_i(y) - \int f_i(y)d\mu_j(y)\right)\right) \end{aligned}$$

Hence by Lemma 4.17, the entries of $A - B$ have, conditionally to N, variance $O(U^2/N^2)$. In particular, $A - B$ has expected squared Frobenius norm at most $O(U^2)$. Bounding the operator norm by the Frobenius norm and applying Markov inequality proves the desired bound on $\|A - B\|$ and concludes the proof of the first part of the theorem. For the second part of Theorem 4.6, the argument is the same except one uses the bound given in Lemma 4.18 instead

of Lemma 4.17. Expliciting the constant $c'_{h_t} = O(t \log^3 n/\sqrt{n})$ then gives the desired bound.

4.5.6 Proof of Theorem 4.7

Since the desired conclusions are unchanged by scaling the components by a constant factor, and as we assume their variance is $\Theta(n)$, we can assume that their variance is n. Let $\widetilde{\mu_i}$ be the pushforwards of μ_i by the closest point projection on S. The following lemma is easily proved:

Lemma 4.19. *Measure $\widetilde{\mu_i}$ has $O(1)$-concentration and mean $O(1)$.*

Proof. Let $f : S \to \mathbb{R}$ be a 1-Lipschitz function. To prove that $\widetilde{\mu_i}$ has $O(1)$-concentration, we prove that for X distributed according to $\widetilde{\mu_i}$, there exists a number c such that $\|f(X) - c\|_{\psi_1} = O(1)$. The range of f on S is contained in an interval of length $2\sqrt{n}$. By shifting f if necessary, we can assume that $\|f\|_\infty = O(\sqrt{n})$. We also assume f is smooth, which is sufficient. Define

$$\begin{aligned} g : \mathbb{R}^n &\to \mathbb{R} \\ x &\mapsto f\left(\frac{x}{\|x\|}\right) \quad \text{if } \|x\| \geq \sqrt{n}/2 \\ x &\mapsto \frac{2\|x\|}{\sqrt{n}} f\left(\frac{x}{\|x\|}\right) \quad \text{else} \end{aligned}$$

We have:

$$\begin{aligned} \nabla g(x) &= \frac{\sqrt{n}}{\|x\|} \nabla f\left(\frac{x}{\|x\|}\right) \quad \text{if } \|x\| \geq \sqrt{n}/2 \\ &= \frac{2}{\sqrt{n}} \left(\frac{x}{\|x\|} f\left(\frac{x}{\|x\|}\right) + \sqrt{n} \nabla f\left(\frac{x}{\|x\|}\right)\right) \quad \text{else} \end{aligned}$$

As a consequence function g is $O(1)$-Lipschitz, hence by concentration, for Y distributed according to μ_i, there exists a number c such that by $\|g(Y) - c\|_{\psi_1} = O(1)$. Letting now $\bar{f} : x \mapsto f(x/\|x\|)$, we have that $P(g(Y) \neq \bar{f}(Y)) \leq \exp(-\Theta(1)\sqrt{n})$ since g and $\bar{f}$ only differ on $B(0, \sqrt{n}/2)$, which has exponentially small measure by concentration. Also clearly $\|g(Y) - \bar{f}(Y)\|_\infty \leq O(\sqrt{n})$. As a consequence, the ψ_1 norm of $g(Y) - \bar{f}(Y)$ is at most $O(\sqrt{n})$ times the ψ_1 norm of a Bernoulli variable with expectation $\exp(-\Theta(1)\sqrt{n})$. Since the ψ_1

norm of such variables is $O(1/\sqrt{n})$, $\|g(Y) - \bar{f}(Y)\|_{\psi_1} = O(1)$, from which we get $\|\bar{f}(Y) - c\|_{\psi_1} = O(1)$. This is what we wanted to prove, as $\bar{f}(Y)$ and $f(X)$ have the same distribution.

To relate the means of μ_i and $\widetilde{\mu}_i$, we notice that by concentration of the distance to the origin, the 1-transportation distance between both measures is $O(1)$. In particular the means of μ_i and $\widetilde{\mu}_i$ differ by $O(1)$, hence the mean of $\widetilde{\mu}_i$ is $O(1)$. □

The above lemma shows that we can apply Theorem 4.6 to the projected point cloud $\tilde{X}$: With arbitrarily high probability, matrix $\Phi_{h_t}(\tilde{X})$ is $\delta = O(t \log^3 n/\sqrt{n})$ close to B in the operator norm, assuming N_0 is $\Omega(\log(n/t)n^2/t^2)$.

We now would like to argue that B retains enough information about the components so that we can separate them. To do so, we restrict B to the subspace $E_{u,v}$ of piecewise constant vectors supported on the two components $\tilde{X}_u$ and $\tilde{X}_v$, for some indices u and v. In the orthornormal basis formed by the normalized indicator vectors of the two components, the ij entry ($i, j \in \{u, v\}$) matrix of this restriction is $(\bar{w}_i\bar{w}_j)^{-1/2}\widehat{G}_{h_t}(i, j)$, $\widehat{G}_{h_t}$ being the 2×2 matrix associated with $\widehat{\mu}_u$ and $\widehat{\mu}_v$, and $\bar{w}_i$ being the fraction of data points in the ith component. As the $\bar{w}_i$'s are $\Theta(1)$, the singular values of B restricted to $V_{u,v}$ are within a constant factor of those of $\widehat{G}_{h_t}$.

Now, using the power series expansion of h_t, one can show the following lower bound on the smallest singular value of the 2×2 matrix G_{h_t} associated with μ_u and μ_v, based on the difference between their covariance matrices:

Lemma 4.20. *There exists $C_1 = \Theta(1)$ such that if $t \leq C_1\|\Sigma_u - \Sigma_v\|_2/\sqrt{n}$, the smallest singular value of G_{h_t} is at least $\Omega(t^2\|\Sigma_u - \Sigma_v\|_2^2/n)$. Furthermore:*

$$\|\widehat{G}_{h_t} - G_{h_t}\| = O(t/\sqrt{n})$$

Proof. By Taylor's theorem, for $i, j \in \{u, v\}$, we have:

$$\begin{aligned} G_{h_t}(i, j) &= \int \cos\left(\frac{t}{\sqrt{n}} < x, y >\right) d\mu_i(x)d\mu_j(y) \\ &= \sum_{l=0}^{\infty} \int (-1)^l \frac{(t/\sqrt{n})^{2l}}{(2l)!} < x, y >^{2l} d\mu_i(x)d\mu_j(y) \end{aligned}$$

Let x and y be two independent random vectors distributed respectively according to μ_i and μ_j. Conditioned to $x = x_0 \in \mathbb{R}^n$, $< x, y >$ has $O(\|x_0\|)$-concentration and mean $O(\|x_0\|)$, so its ψ_1 norm is $O(\|x_0\|)$. Hence

$$\| < x, y > \|_{\psi_1} \leq O(\mathbb{E}\|x\|) \leq O(\sqrt{n})$$

As a consequence the distribution of $| < x, y > /\sqrt{n}|$ decays exponentially. Hence its lth moment is controlled by the lth moment of an exponential distribution with mean $\Theta(1)$, that is, $\Theta(1)^l l!$. This implies

$$\begin{aligned} |G_{h_t}(i,j) - 1 + \int \frac{t^2}{2n} < x, y >^2 d\mu_i(x) d\mu_j(y)| &\leq \sum_{l=2}^{\infty} \frac{t^{2l}}{(2l)!} \Theta(1)^{2l} (2l)! \\ &\leq O(t^4) \end{aligned}$$

for t less than some numerical constant. Now

$$\begin{aligned} \int < x, y >^2 d\mu_i(x) d\mu_j(y) &= \int y^t \Sigma_i y \, d\mu_j(y) \\ &= \int \text{trace } \Sigma_i y y^t d\mu_j(y) \\ &= \text{trace } \Sigma_i \Sigma_j \end{aligned}$$

We may thus expand the determinant of G_{h_t} as follows:

$$\begin{aligned} \det G_{h_t} &= G_{h_t}(u,u) G_{h_t}(v,v) - G_{h_t}(u,v)^2 \\ &= \left(1 - \frac{t^2}{2n} < \Sigma_u, \Sigma_u > + O(t^4)\right) \left(1 - \frac{t^2}{2n} < \Sigma_v, \Sigma_v > + O(t^4)\right) \\ &\quad - \left(1 - \frac{t^2}{2n} < \Sigma_u, \Sigma_v > + O(t^4)\right)^2 \\ &= -\frac{t^2}{2n} \|\Sigma_u - \Sigma_v\|_2^2 + O(t^4) \end{aligned}$$

Hence by assumption, for well chosen C_1, the first term in the expansion above dominates, so $|\det G_{h_t}|$ satisfies the desired lower bound. Since the entries of G_{h_t} have absolute value less than 1, the lower bound also holds for the smallest singular value of G_{h_t}.

To relate matrices G_{h_t} and $\widetilde{G}_{h_t}$, we let δx (resp. δy) be the difference between x (resp. y) and its projection on S, so that $x - \delta x$ (resp. $y - \delta y$) is distributed

according to $\widetilde{\mu_i}$ (resp. $\widetilde{\mu_j}$). We can write

$$\widetilde{G}_{h_t}(i,j) = \int \cos\left(\frac{t}{\sqrt{n}} < x - \delta x, y - \delta y >\right) d\mu_i(x) d\mu_j(y)$$

Also

$$< x - \delta x, y - \delta y > = < x, y > - < \delta x, y > - < \delta y, x > + < \delta x, \delta y >$$

By concentration and since μ_j has $O(1)$ mean, $| < \delta x, y > |$ has expectation $O(\|\delta x\|)$ conditioned to δx. Since $\mathbb{E}\|\delta x\| = O(1)$ by concentration of the distance to the origin, we have $\mathbb{E}| < \delta x, y > | = O(1)$. The last two terms above can be dealt with similarly, yielding that the distributions of $< x - \delta x, y - \delta y >$ and of $< x, y >$ are at 1-transportation distance $O(1)$. Since $\cos(t./\sqrt{n})$ is $O(t/\sqrt{n})$-Lipschitz, we see that

$$|\widetilde{G}_{h_t}(i,j) - G_{h_t}(i,j)| = O(t/\sqrt{n})$$

which concludes the proof. □

In particular, choosing $t = C_1\|\Sigma_u - \Sigma_v\|_2/\sqrt{n} = C_1\Delta$, we see that for any u, v, the smallest singular value of B restricted to $E_{u,v}$ is at least $\Omega(\Delta^4 - O(\Delta/\sqrt{n}))$, which by assumption on Δ is also $\Omega(\Delta^4)$.

Lemma 4.21. *For sufficiently small $C_2 = \Theta(1)$, the columns of $f_{C_2\Delta^4}(B)$ with indices i and j are equal if i and j belong to the same component. If i and j belong to different components, their distance is $\Omega(\Delta^4/\sqrt{N})$.*

Proof. Eigenvectors of B with non zero eigenvalue are piecewise constant, so the first part is clear. Assume indices i and j respectively belong to distinct components u and v. The distance between their columns is $\|f_{C_2\Delta^4}(B)e_{uv}\|$, where e_{uv} has entries $1/\sharp X_u$ (resp. $-1/\sharp X_v$) at indices corresponding to component u (resp. v), and 0 else.

Vector e_{uv} is in E_{uv} and has norm $\Theta(1/\sqrt{N})$. From the singular value lower bound, there must exist a unit vector x such that $| < e_{uv}, Bx > | = \Omega(\Delta^4/\sqrt{N})$.

Denote by $E_{2C_2\Delta^4}$ the vector space generated by the singular vectors of B with singular values at least $2C_2\Delta^4$, and write $x = \alpha y + \beta z$, where y and z are unit vectors respectively lying in $E_{2C_2\Delta^4}$ and in $E^{\perp}_{2C_2\Delta^4}$, and $\alpha^2 + \beta^2 = 1$. We have

$$\begin{aligned} | < e_{uv}, Bx > | &= |\alpha < e_{uv}, By > + \beta < e_{uv}, Bz > | \\ &= O(| < e_{uv}, By > |) + O(C_2\Delta^4/\sqrt{N}) \\ &\leq \max(C_3| < e_{uv}, By > |, C_4C_2\Delta^4/\sqrt{N}) \end{aligned}$$

for some constant C_3 and C_4, where in the second line we used the fact that the largest singular value of B is $O(1)$. Hence for small enough $C_2 = \Theta(1)$, we will have $C_4C_2\Delta^4/\sqrt{N} < | < e_{uv}, Bx > |$, implying $| < e_{uv}, By > | \geq | < e_{uv}, Bx > |/C_3 = \Omega(\Delta^4/\sqrt{N})$. Now because $y \in E_{2C_2\Delta^4}$, as $f_{C_2\Delta^4}$ modifies eigenvalues by a factor at most 2 in that range, there exists a matrix F with the same eigenvectors as B, and with singular values between 1/2 and 2, such that $FBy = f_{C_2\Delta^4}(B)y$. Hence

$$\begin{aligned} | < f_{C_2\Delta^4}(B)F^{-1}e_{uv}, y > | &= | < F^{-1}e_{uv}, f_{C_2\Delta^4}(B)y > | = | < F^{-1}e_{uv}, FBy > | \\ &= | < e_{uv}, By > | = \Omega(\Delta^4/\sqrt{N}) \end{aligned}$$

In particular $f_{C_2\Delta^4}(B)F^{-1}e_{uv}$ has norm at least $O(\Delta^4/\sqrt{N})$. But that vector equals $F^{-1}f_{C_2\Delta^4}(B)e_{uv}$, and as F^{-1} doesn't change distances by more than a factor of 2, we see that $\|f_{C_2\Delta^4}(B)e_{uv}\| = \Omega(\Delta^4/\sqrt{N})$, as claimed. □

Now, as $f_{C_2\Delta^4}$ is 1-Lipschitz, the perturbation inequality proved in [83] states that

$$\begin{aligned} \|f_{C_2\Delta^4}(B) - f_{C_2\Delta^4}(\Phi_{h_t}(\tilde{X}))\| &\leq O\left(\log \frac{\|B\| + \|\Phi_{h_t}(\tilde{X})\|}{\|\Phi_{h_t}(\tilde{X}) - B\|} + 2\right)^2 \|\Phi_{h_t}(\tilde{X}) - B\| \\ &\leq O(\delta \log^2 \delta) \end{aligned}$$

Our assumption on Δ is chosen so that $\Delta^4/(\delta \log^2 \delta) = \Omega(K^3)$. Hence we may assume in particular that $\|\Phi_{h_t}(\tilde{X}) - B\| < C_2\Delta^4$. By Weyl's theorem on eigenvalue perturbations, $f_{C_2\Delta^4}(\Phi_{h_t}(\tilde{X}))$ thus has at most $k = \Theta(1)$ non zero eigenvalues. As a result $\|f_{C_2\Delta^4}(B) - f_{C_2\Delta^4}(\Phi_{h_t}(\tilde{X}))\|_2^2 \leq O(\delta^2 \log^4 \delta) = O(\Delta^8/K^6)$.

This means that within each component, the columns of $f_{C_2\Delta^4}(\Phi_{h_t}(\tilde{X}))$ have variance $O(\Delta^8/(NK^6))$ with respect to the column of $f_{C_2\Delta^4}(B)$ associated with that component. By Lemma 4.21, this implies that the ratio between the maximum variance of the components in $\phi_{C_2}(\tilde{X})$ and the minimum squared distance between their centers in an optimal solution to the k-means problem is $O(K^{-6})$. Applying any constant factor approximation algorithm for the k-means problem will thus cluster the data with the claimed error rate.

Bibliography

[1] D. Francois. *High dimensional data analysis.* Verlag, 2008.

[2] R. Xu X.Li. High-dimensional data analysis in cancer research. *High-Dimensional Data Analysis in Oncology*, 2009.

[3] S. Selvaraj and J. Natarajan. Microarray data analysis and mining tools. *Bioinformation. 2011; 6(3): 95-99*, 2011.

[4] D.M. Mutch, A. Berger, R. Mansourian, A. Rytz, and M.A. Roberts. Microarray data analysis: a practical approach for selecting differentially expressed genes. *Genome Biol. 2001;2(12)*, 2001.

[5] E. Marchiori and J. H. Moore. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics: 6th European Conference, EvoBIO 2008, Naples, Italy, March 26-28, 2008, Proceedings.* Springer, 2008.

[6] A.S. Shirkhorshidi, S. Aghabozorgi, and T.Y. W. T. Herawan. *Big Data Clustering: A Review - Computational Science and Its Applications.* ICCSA 2014, 2014.

[7] M. Menoret, N. Farrugia, B. Pasdeloup, and V. Gripon. Evaluating graph signal processing for neuroimaging through classification and dimensionality reduction. *IEEE Global Conference on Signal and Information*, 2017.

[8] N. Foy N.T.H. Gayraud and M. Clerc. Systems, man, and cybernetics (smc). *2016 IEEE International Conference on*, 2016.

[9] I.Abraham, Y.Bartaly, and O.Neimanz. Embedding metric spaces in their intrinsic dimension. *ISODA 08 Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, 2008.

[10] Jianzhong Wang. *Geometric Structure of High-Dimensional Data and Dimensionality Reduction.* Springer Publishing Company, Incorporated, 2012. ISBN 364227496X, 9783642274961.

[11] K. Beyer, J. Goldstein, R. Ramakrishnan, and U.Shaft. When is Nearest Neighbour Meaningful? *ICDT '99*, 1999.

[12] Miguel Carreira-Perpin. A review of dimension reduction techniques. Technical report, 1997.

[13] L.J.P. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review, 2008.

[14] D. Engel, L. Huttenberger, and B. Hamann. *A Survey of Dimension Reduction Methods for High-dimensional Data Analysis and Visualization in Proceedings of IRTG 1131-Visualization of Large and Unstructured Data Sets Workshop 2011.* Springer, 2011.

[15] K.T. Sturm. On the geometry of metric measure spaces. *Acta Math. 196*, (1):65–131, 2006.

[16] Shlomo Dubnov, Ran El-Yaniv, Yoram Gdalyahu, Elad Schneidman, Naftali Tishby, and Golan Yona. A new nonparametric pairwise clustering algorithm based on iterative estimation of distance profiles. *Machine Learning*, 2002.

[17] Alexander V. Kolesnikov and Emanuel Milman. The KLS isoperimetric conjecture for generalized Orlicz balls. *Ann. Probab.*, 46(6):3578–3615, 11 2018. doi: 10.1214/18-AOP1257. URL https://doi.org/10.1214/18-AOP1257.

[18] Qingqing Huang Rong Ge and Sham M. Kakade. Learning mixtures of Gaussians in high dimensions. *In Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing, STOC '15*, page 761–770, 2015.

[19] Jose Gomes and Aleksandra Mojsilovic. *A Variational Approach to Recovering a Manifold from Sample Points*, pages 3 – 17. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

[20] Jean-Daniel Boissonnat and Mariette Yvinec. *Algorithmic Geometry.* Cambridge University Press, New York, NY, USA, 1998. ISBN 0-521-56529-4.

[21] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph.*, 21(4):163–169, August 1987. ISSN 0097-8930. doi: 10.1145/37402.37422. URL http://doi.acm.org/10.1145/37402.37422.

[22] Jean-Daniel Boissonnat and Arijit Ghosh. Manifold reconstruction using Tangential Delaunay Complexes. *A. Discrete Comput Geom*, 2014. URL https://doi.org/10.1007/s00454-013-9557-2.

[23] Timothy S. Newman and Hong Yi. A survey of the marching cubes algorithm, 2006.

[24] Eamonn Keogh and Abdullah Mueen. *Curse of Dimensionality*, pages 257–258. Springer US, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_192. URL http://dx.doi.org/10.1007/978-0-387-30164-8_192.

[25] Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

[26] Richard Bellman. *Adaptive control processes: a guided tour*. Princeton University Press, 1961.

[27] Guruswami V. and Kannan R. Geometry of high dimensional spaces. 2012.

[28] Ryan Prescott Adams. High-dimensional probability estimation with deep density models. arxiv preprint arxiv:1302.5125. 2013.

[29] K. Saul, K. Weinberger, F. Sha, J. Ham, and D. Lee. Spectral methods for dimensional reduction. *MIT press*, 2006.

[30] J. Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel J. Math. 52*, pages 46–52, 1985.

[31] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *In Conference in modern analysis and probability (New Haven, Conn., 1982) Amer. Math. Soc., Providence*, pages 189–206, 1984.

[32] Sanjoy Dasgupta and Yoav Freund. Random projection trees and low dimensional manifolds. In *Proceedings of the Fortieth Annual ACM Symposium*

on Theory of Computing, STOC '08, pages 537–546, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-047-0. doi: 10.1145/1374376.1374452. URL http://doi.acm.org/10.1145/1374376.1374452.

[33] Junfeng He, Sanjiv Kumar, and Shih-Fu Chang. On the difficulty of nearest neighbor search. *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, 2012.

[34] Alon Orlitsky. Estimating and computing density based distance metrics. In *ICML05, 22nd International Conference on Machine Learning*, 2005.

[35] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[36] Ryan Prescott Adams. High-dimensional probability estimation with deep density models. arxiv preprint arxiv:1302.5125, 2013.

[37] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete Comput. Geom*, 33:249–274, 2005.

[38] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 2005.

[39] Robert Ghrist. Barcodes: The persistent topology of data. Technical report, 2007.

[40] Herbert Edelsbrunner and John Harer. Persistent homology – a survey, 2008.

[41] F. Chazal, Cohen-Steiner, M. D. Glisse, L. J. Guibas, and S. Y. Oudot. Proximity of persistence modules and their diagrams. *Proceedings of the Twenty-fifth Annual Symposium on Computational Geometry SCG '09. New York, NY*, 2009.

[42] Gunnar Carlsson, Afra Zomorodian, Anne Collins, and Leonidas J. Guibas. Persistence barcodes for shapes. *International Journal of Shape Modeling.*, 2004.

[43] Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. *The Gudhi Library: Simplicial Complexes and Persistent Homology*. Springer Berlin Heidelberg, 2014.

[44] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pages 604–613, New York, NY, USA, 1998. ACM. ISBN 0-89791-962-9. doi: 10.1145/276698.276876. URL http://doi.acm.org/10.1145/276698.276876.

[45] Jingdong Wang, Heng Tao Shen, Jingkuan Song, and Jianqiu Ji. Hashing for similarity search: A survey. *CoRR*, abs/1408.2927, 2014. URL http://arxiv.org/abs/1408.2927.

[46] Mayur Datar and Piotr Indyk. Locality-sensitive hashing scheme based on p-stable distributions. In *In SCG04: Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262. ACM Press, 2004.

[47] Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li. Multi-probe LSH: Efficient indexing for high-dimensional similarity search. In *in Proc. 33rd Int. Conf. Very Large Data Bases*, pages 950–961, 2007.

[48] Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. Practical and Optimal LSH for Angular Distance. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 1225–1233, Cambridge, MA, USA, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969239.2969376.

[49] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, SCG '04, pages 253–262, New York, NY, USA, 2004. ACM. ISBN 1-58113-885-7. doi: 10.1145/997817.997857. URL http://doi.acm.org/10.1145/997817.997857.

[50] Ke Li and Jitendra Malik. Fast k-nearest neighbour search via prioritized DCI. *CoRR*, abs/1703.00440, 2017. URL http://arxiv.org/abs/1703.00440.

[51] C. Williams and M. Seeger. The effect of the input density distributions on kernel based classifiers. *International Conference on Machine Learning 17*, 2000.

[52] N. El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 2010.

[53] Geoffrey Hinton and Sam Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems 15*, pages 833–840. MIT Press, 2003.

[54] John A. Lee, Emilie Renard, Guillaume Bernard, Pierre Dupont, and Michel Verleysen. Type 1 and 2 Mixtures of Kullback-Leibler Divergences As Cost Functions in Dimensionality Reduction Based on Similarity Preservation. *Neurocomput.*, 112:92–108, July 2013. ISSN 0925-2312. doi: 10.1016/j.neucom.2012.12.036. URL http://dx.doi.org/10.1016/j.neucom.2012.12.036.

[55] Thomas Eiter and Leonid Libkin. Database theory - icdt 2005. Springer-Verlag Berlin Heidelberg, 2005.

[56] Kumari, Sushma, Jayaram, and Balasubramaniam. Measuring concentration of distances 2014 an effective and efficient empirical index. *IEEE Trans. on Knowl. and Data Eng.*, 2014.

[57] Lior Rokach and Oded Maimon. Data mining and knowledge discovery handbook, chapter 15 clustering methods, 2010.

[58] M. Gromov. *Metric structures for Riemannian and non-Riemannian spaces*. Birkhäuser Verlag Basel, 1999.

[59] V. Milman. A certain property of functions defined on infinite-dimensional manifolds. *Dokl. Akad. Nauk SSSR 200*, pages 781–784, 1971.

[60] M. Ledoux. *The Concentration of Measure Phenomenon*. AMS Mathematical Surveys & Monographs, Providence, 2001.

[61] P. Lévy and F. Pellegrino. *Problèmes concrets d'analyse fonctionnelle.* Collection de monographies sur la théorie des fonctions Gauthier-Villars Paris, 1951.

[62] Jiri Matousek. *Lectures on Discrete Geometry.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2002. ISBN 0387953744.

[63] Memoli F. On the use of Gromov-Hausdorff distances in shape comparison. In M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, editors, *Symposium on Point Based Graphics*, Prague, Czech Republic, 2007. Eurographics Association. ISBN 978-3-905673-51-7.

[64] Memoli Facundo. Gromov- Hausdorff distances in Euclidean spaces. *NORDIA-CVPR-2008*, 2008.

[65] Facundo Memoli. Some Properties of Gromov-Hausdorff Distances. *Discrete & Computational Geometry*, pages 1–25, 2012. ISSN 0179-5376. URL http://dx.doi.org/10.1007/s00454-012-9406-8. 10.1007/s00454-012-9406-8.

[66] Memoli Facundo. Gromov - Wasserstein distances and the metric approach to object matching. *Found. Comput. Mat.*, (11):417–487, 2011.

[67] S. Dasgupta. Learning mixtures of Gaussians. *In Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science, FOCS 99*, pages 634–644, 1999.

[68] S. Dasgupta and L. Schulman. A two-round variant of EM for Gaussian mixtures. *Uncertainty in Artificial Intelligence*, 2000.

[69] S. Arora and R. Kannan. Learning mixture of separated non-spherical Gaussians. *The Annals of Applied Probability 2005, Vol. 15, Institute of Mathematical Statistics*, 2005.

[70] S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences Special issue on FOCS 2002 archive*, 68(4):841–860, June 2004.

[71] R. Kannan, S. Vempala, and H. Salmasian. The spectral method for general mixture models. *Proc. of the 18th Conference on Learning Theory, 2005.SIAM J. Computing,*, 2008.

[72] D. Achlioptas and F. McSherry. On spectral learning of mixture of distributions. *In Proc. of COLT*, 2005.

[73] S. Brubaker and S.Vempala. Isotropic PCA and affine-invariant clustering. *Building Bridges (Ed.s M. Grotchel and G.O.H.Katona), Bolyai Society Mathematical Studies, Proc. of FOCS 2008.*, 2008.

[74] M. Belkin and K. Sinha. Polynomial learning of distribution families. *FOCS*, 2010.

[75] T. Shi, M. Belkin, and B. Yu. Data spectroscopy: Eigenspaces of convolution operators and clustering. *Annals of statistics*, 2009.

[76] M. Ledoux S. Bobkov. Poincaré's inequality and Talagrand's concentration phenomenon for the exponential distribution. *Probab. Theory Relat. Fields (107):383–400*, 1997.

[77] C. Suquet. Distances euclidiennes sur les mesures signées et application des théorèmes de Berry-Esseen. *Bull. Belg. Math. Soc. Simon Stevin, 2(2):161–181*, 1995.

[78] M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. *In Proceedings of AISTATS 2005*, page 136–143, 2005.

[79] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, page 69–92, 2005.

[80] M.D. Kirzbraun. Uber die zusammenziehende und lipschitzsche transformationen. *Fundamenta Math 22:77–108*, 1934.

[81] Joel A. Tropp. Norms of random submatrices and sparse approximation. *Comptes Rendus Mathematique*, 346(23):1271 – 1274, 2008. ISSN 1631-073X. doi: https://doi.org/10.1016/j.crma.2008.10.008. URL http://www.sciencedirect.com/science/article/pii/S1631073X08003002.

[82] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*, 2012.

[83] Yu. B. Farfarovskaia. An estimate of the norm of f(a) - f(b) for selfadjoint operators a and b (in russian). *Zap. Nauchn. Sem. LOMI, (56):143162*, 1976.