On Manifold Dimension Estimation

by **Noah Bergam**

Senior Thesis in Mathematics
Advisor: Professor Andrew J. Blumberg

COLUMBIA UNIVERSITY
March 31, 2024

On Manifold Dimension Estimation

Noah Bergam

Abstract

This thesis is a review of algorithms and statistical complexity results for the manifold intrinsic dimension (ID) estimation problem. The task is as follows: *given an independent and identically distributed sample of points from a low-dimensional submanifold embedded in high-dimensional Euclidean space, determine the dimension of the submanifold.* This problem is of key interest in data science, as many algorithms can be made to depend on the intrinsic dimension of data, rather than the dimension of its ambient space. We pay close attention to the linear case of this problem, which reduces to principle component analysis (PCA). In the general manifold case, the kinds of approaches become much more diverse. We distinguish two very different kinds of methods: (1) those which isolate a local statistic (e.g. number of neighbors within a certain radius) and analyze its scaling behavior in varying neighborhood sizes; and (2) those which analyze a global statistic and its scaling behavior independent of local information (e.g. the Wasserstein distance between two independently-formed empirical distributions, and how it scales with the size of their samples). We then compare lower bounds on the sample complexity of ID estimation, in a model with noise and a model without.

Contents

Chapter 1

Introduction

High-dimensional, high-volume data is increasingly abundant in the natural and social sciences. Finding the right representation or encoding of such data is of fundamental interest in unsupervised machine learning and data science writ large. Ideally, this representation is a *reduction* in some sense, making the dataset smaller while preserving important information. For instance, to reduce the number of data points, one could apply k-means and represent the dataset in terms of cluster centers; or to reduce dimension, one could pursue a principal component analysis (PCA) or Johnson-Lindenstrauss (JL) transform to project the data onto a low-dimensional linear subspace. These kinds of reductions are useful for both algorithmic applications (e.g. nearest neighbor search) and scientific interpretation (e.g. data visualization).

Making such data reductions often comes with (heavy) assumptions regarding the generative process behind the dataset. For instance, the original expectation-maximization (EM) algorithm is designed for data generated from a mixture of Gaussians; and PCA is best-suited for data which lies on a linear subspace (or, better yet, data which is sampled from a probability distribution supported on a linear subspace).

As a generalization of PCA, one might consider data generated from a "nonlinear subspace," or more precisely, an embedded submanifold of Euclidean space. Assuming a dataset has such a structure is often called the *manifold hypothesis*, and the task of learning this structure is often called *manifold learning.* One major problem in manifold learning—the topic of this paper—is estimating the dimension of a data-generating manifold. We refer to this as the intrinsic dimension (ID) estimation problem, and we introduce some general parameters for the problem as follows:

Definition 1 (ID estimation problem)**.** *Let $\mathcal{M} \subset \mathbb{R}^D$ be a d-dimensional Riemannian manifold embedded in $\mathbb{R}^D$ (with $d \ll D$ presumably), such that:*

- *The manifold is bounded, say $\mathcal{M} \subset [0,1]^D$,*
- *the reach of the manifold (a proxy for curvature, defined in Section 2) is bounded, i.e. $\tau(\mathcal{M}) < T$, and*
- *the volume is bounded $\mathrm{vol}(\mathcal{M}) < V$.*

Let μ be a probability distribution supported on $\mathcal{M}$. Design an efficient algorithm that takes in n independent samples from μ and returns an estimator $\hat{d}_n \in [D]$ such that $\hat{d}_n = d$ (or is sufficiently close) with high probability.

This paper reviews major results and frames important open questions regarding this problem of ID estimation. We proceed in three steps:

- **Background:** First, we review the necessary background in dimensionality reduction and differential geometry literature. We transition from the data scientific problem motivating manifold learning to the mathematical foundations needed to rigorously discuss and prove guarantees for manifold learning.

- **Algorithms:** Then, we review various estimators, focusing on their intuition, implementation, and algorithmic guarantees. We split methods into two rough categories: local methods, which examine the behavior of the data in neighborhoods, and global methods, which make use of the whole structure of points.

- **Complexity:** Then, we discuss the statistical complexity of intrinsic dimension estimation. We investigate how the presence of noise has a substantial difference on the convergence rate of dimension estimators.

This paper assumes a strong grasp of linear algebra and basic statistics. Background in topics like differential geometry, probability theory (empirical process theory), and machine learning is generally helpful but not entirely necessary.

Chapter 2

Review of Manifold Learning

In this section we give context for the manifold dimension estimation problem. We provide background in differential geometry, statistical learning theory, and where these topics intersect in "manifold learning." This sets up what we can expect from the dimension estimation methods we design and analyze in later sections.

2.1 What is a manifold?

A manifold is a special kind of topological space. A topological space is, roughly speaking, a set with some notion of "closeness." One can endow a topological space with additional structure, such as a metric, norm, or inner product. Euclidean space is a particularly well-endowed topological space (a Hilbert space, in fact). A manifold is a topological space that is *locally* similar to Euclidean space.

Definition 2 (topological space)**.** *A topological space is a pair (X, τ) with X an arbitrary set and $\tau \subset \mathcal{P}(X)$ a collection of ("open") subsets of X such that (1) $\emptyset \in \tau$, (2) τ is closed under arbitrary unions, and (3) τ is closed under finite intersections.*

Two notions of well-behavedness for a topological space are:

- *Hausdorff*, meaning for all $u \neq v$ there exist open neighborhoods U of u and V of v that are disjoint. (Importantly, this guarantees the uniqueness of limits).
- *Second countable*, meaning there exists a countable set of open sets $\mathcal{U} \subset \tau$ such that any open set can be expressed as a union of elements from $\mathcal{U}$.

It is easy and instructive to verify that Euclidean space satisfies both of these conditions; the discrete topology ($\tau = \mathcal{P}(X)$) is Hausdorff but not second countable; and the indiscrete topology ($\tau = \{\emptyset, X\}$) is trivially second countable but not Hausdorff. In light of these examples, observe how the second countable condition ensures that there aren't too many open sets, while the Hausdorff condition ensures there aren't too few.

We now define a manifold in a topological sense.

Definition 3 (topological manifold)**.** *A manifold M of dimension n is a Hausdorff, second-countable topological space such that for all $p \in M$, there exists (U, ϕ) where U is an open neighborhood containing p and $\phi : U \to \phi(U) \subset \mathbb{R}^n$ is a homeomorphism (i.e. bijective with continuous inverse). We call (U, ϕ) a coordinate chart.*

In order to have a more expansive theory of calculus on manifolds, we would like to impose some differentiability conditions on the coordinate charts.

Definition 4 (smooth manifold)**.** *A smooth or C^∞ manifold is a manifold with a smooth atlas, i.e. a collection of charts $\{U_\alpha, \phi_\alpha\}$ such that:*

- *The coordinate neighborhoods cover the manifold: $M = \bigcup_\alpha U_\alpha$.*
- *The coordinate charts are smoothly compatible, meaning: for (U, ϕ) and (V, ψ), the following two "transition maps" are smooth (as functions $\mathbb{R}^n \to \mathbb{R}^n$):*

$$\psi \circ \phi^{-1} : \phi(U \cap V) \to \psi(U \cap V) \qquad \phi \circ \psi^{-1} : \psi(U \cap V) \to \phi(U \cap V)$$

Definition 5 (smooth map)**.** *Let M and N be m and n-dimensional manifolds, respectively. A map $F : M \to N$ is **smooth** if, for all $p \in M$, there exists charts (U, ϕ) about p and (V, ψ) about $F(p) \in N$ such that $\phi^{-1} \circ F \circ \psi : U \to V$ is a smooth map.*

The graph of a function, say the curve $\{e^x\}_{x \in \mathbb{R}}$ in $\mathbb{R}^2$, is a canonical example of a submanifold. There are various notions of submanifold, but two will be particularly important for us:

- An **immersed submanifold** of M is the image of an immersion map $f : N \to M$, i.e. if the pushforward $f_{*,x} : T_xN \to T_xM$ is injective for all x.
- An **embedded** or **regular submanifold**), is an immersed submanifold for which the inclusion map is a topological embedding. That is, the submanifold topology on S is the same as the subspace topology.

The tangent space is a vector space associated to every point of a manifold which effectively encodes the local linear structure of a manifold. Here is one way of defining the tangent space, explained in [Tu11].

Definition 6 (tangent space)**.** *Let the $C_p^\infty(M)$ denote the set of **smooth germs** at p, i.e. smooth functions $M \to \mathbb{R}$ modulo $\sim$ where $f \sim g$ if f, g agree on a neighborhood of p. The **tangent space of** M **at** p, denoted T_pM, is the set of point-derivations at p, i.e. linear maps $D : C_p^\infty(M) \to \mathbb{R}$ satisfying the so-called Leibniz rule:*

$$D(fg) = (Df)g(p) + f(p)Dg.$$

This perspective of tangent spaces acting on germs of real-valued functions on manifold will be crucial in our next step: using the tangent space to conceptualize a first notion of differentiation on manifolds.

Definition 7 (differential)**.** *Let $f : M \to N$ be a smooth map between manifolds. Then the **differential** or **pushfoward** $f_* : T_pM \to T_pN$ is defined as follows: for $X_p \in T_pM$ and $g \in C_{f(p)}^\infty(N)$, we have:*

$$\Big(f_*(X_p)\Big)(g) = \Big(X_p\Big)(g \circ f)$$

where $g \circ f$ belongs to $C_p^\infty(M)$.

Definition 8 (tangent bundle, informal)**.** *A smooth manifold M's* ***tangent bundle*** *is the set $TM = \{(x, y) : x \in M, y \in T_pM\}$ equipped with a natural topology and smooth structure (see [Tu11], page 131) that makes it a smooth manifold itself. If M is of dimension n, TM is of dimension $2n$.*

Though it may seem unnatural or contrived at first glance, the tangent bundle provides us an excellent notation for thinking about certain geometric objects. Two examples:

- Given a smooth map between smooth manifolds $f : M \to N$, the derivative map is most compactly written as a map between tangent bundles:

$$Df : TM \to TN \qquad Df(p, X_p) = (f(p), f_{*,p}(X_p))$$

- A smooth vector field on M (a.k.a. smooth section of TM) is a smooth map between manifolds $X : M \to TM$ where $X(p) = (p, X_p)$.

A **geodesic** is a shortest path along a manifold. In order for a geodesic metric to be well-defined, it turns out to be crucial to have a local sense of angle. The mathematical manifestation of this is a smoothly varying inner product over on the tangent spaces of a manifold, formally known as a Riemannian metric.

Definition 9 (Riemannian metric)**.** *g is a Riemannian metric on a smooth manifold M if for each $p \in M$ there exists $g_p : T_pM \times T_pM \to \mathbb{R}$ such that:*

- *g_p is an inner product on T_pM (positive definite, symmetric, and bilinear).*
- *g is smoothly varying, i.e. $p \to g_p(X_p, Y_p)$ is a smooth function for every smooth vector field $X, Y \in \mathcal{X}(M)$.*

Proposition 1 ([Lee12], Prop. 13.3)**.** *Every C^∞ manifold admits a Riemannian metric.*

With this additional structure on the tangent space, we are able to reason about paths and curvature on a manifold.

Definition 10 (geodesic)**.** *Given a Riemannian manifold (M, g),* ***the geodesic distance*** *between $p, q \in M$ is as follows:*

$$d_M(p, q) = \inf_{\gamma \in \mathsf{Paths(p,q)}} \mathsf{Length}(\gamma)$$

where $\mathsf{Length}(\gamma) = \int_0^1 g_{\gamma(t)}(\gamma'(t), \gamma'(t))dt$*, and*

$\mathsf{Paths(p, q)} = \{\gamma : [0, 1] \to \mathsf{M}$ *such that* $\gamma(0) = \mathsf{p}, \gamma(1) = \mathsf{q}, \gamma$ *piecewise smooth curve in* $\mathsf{M}\}$

If there exists $\gamma^* \in \mathsf{Paths(p, q)}$ such that $d_M(p, q) = \mathsf{Length}(\gamma^*)$, we call γ^* a ***geodesic path between*** *p* ***and*** *q.*

This upgrades the manifold as a whole into a metric space (note that the Riemannian metric only turns the tangent space into an inner product space).

2.1.1 Notions of Regularity

Working with general manifolds can be extremely difficult. In particular, they can have high curvature, nearly self-intersecting themselves. This poses a substantial problem for intrinsic dimension estimators, as it can make them overestimate the dimension of the manifold-sampled data (interpreting the "folds" of the manifold as volume). In this section, we describe some common notions of regularity used in the manifold learning literature.

Definition 11 (reach and injectivity radius)**.** *Let M be a compact submanifold of $\mathbb{R}^D$. Define the* ***medial axis*** *to be the set of points with at least two projections onto M, i.e.*

$$\mathit{Med}(M) = \{x \in \mathbb{R}^D \ : \ \exists\ p \neq q \in S \ \textit{such that}\ d(x,p) = d(x,q) = d(x,S)\}$$

The ***reach*** *of M is the largest real number τ such that all points within a distance τ of M have a unique projection onto M. The simplest way to say it is that it is the Euclidean distance between the manifold and its medial axis, i.e.*

$$\tau(M) = d_{\mathit{Euclidean}}(M, \mathit{Med}(M))$$

The ***injectivity radius*** *of M is defined similarly, but with the manifold's intrinsic geodesic distance metric: namely, it is the largest r such that for all $p \in M$, if $d_M(p,q) < r$, then the geodesic path from p to q is unique.*

$$\iota(M) = \sup\{r \in \mathbb{R} : \forall\ p \in M, d_M(p,q) < r \implies \exists!\ \gamma^* \quad d_M(p,q) = \mathsf{Length}(\gamma^*)\}$$

The existence of certain geodesic paths makes for an interesting and rather important technical property of manifolds known as geodesic completeness.

Fact 1. *For (M,g) a Riemannian manifold, and any $p \in M$ and $X_p \in T_pM$, there exists a unique curve $\gamma = \gamma^{p,X_p} : S \to M$ for $S \subset \mathbb{R}$ such that:*

$$\gamma(0) = p \quad \gamma'(0) = X_p$$

It is natural for us then to define the following set:

$$E = \{(p, X_p) : \gamma^{p,X_p}\textit{'s domain can be extended to all of } R\}$$

Definition 12 (geodesic completeness)**.** *A Riemannian manifold is geodesically complete if $E = TM$, i.e. the unique curve passing through each point can be extended indefinitely.*

Oftentimes we care about a manifold having bounded volume, with respect to its intrinsic volume measure. Indeed, this volume measure is crucial to the development of probability theory on the manifold. This requires an understanding of integration of Riemannian manifolds. We discuss briefly the concept of a volume form, which we define briefly below:

Definition 13. *If (M,g) is a Riemannian manifold, then an n-form ω is called a volume form of m if it is canonically defined, i.e. $\omega = \theta^1 \wedge ... \wedge \theta^n$ is independent of the choice of a positively oriented orthonormal frame. Then the* ***volume of the manifold*** *M is given by $\mathrm{vol}(M) = \int_M \omega$.*

2.2 What is (statistical) learning?

Learning is a subjective and complex phenomenon. It can seem surprising at first that it can be rigorously studied at all. In this section, we discuss a relatively general notion of statistical learning, roughly based on the probably approximately correct (PAC) learning framework of [Val84].

The basic ingredients of statistical learning are:

- A sample space X and a set of probability distributions over this sample space $\mathcal{P} \subset \Delta(X)$, where $\Delta(X)$ denotes the set of all distributions over X. It is convenient for us to be able to *sample* a distribution from $\mathcal{P}$, in which case we assume it is a distribution over distributions, i.e. $\mathcal{P} \in \Delta(\Delta(\Omega)))$.
- A statistic $\theta(P)$ attached to each distribution $P \in \mathcal{P}$. This is encoded in a function $\theta : \mathcal{D} \mapsto \Theta$. We call Θ the parameter space; we typically think of it as having a metric, call it $d : \Theta \times \Theta \to \mathbb{R}_+$.
- An algorithm $\mathcal{A} : X^n \to \Theta$ which presumably takes in n i.i.d. samples from a distribution $P \in \mathcal{P}$ and outputs an estimate of the desired statistic of that distribution.

The combination (X, θ) constitutes a **learning problem**. We say an algorithm $\mathcal{A}$ (ε, δ)-learns (X, θ) if for all $\mathcal{P} \subset \Delta(X)$, with probability $1-\delta$ over the choice of distribution $P \sim \mathcal{P}$ followed by the choice of training data $x_{\text{train}} \sim P^{\times n}$,

$$d(\mathcal{A}(x_{\text{train}}), \theta(P)) \leq \varepsilon$$

As computational learning theorists, we are concerned with the existence and construction of $\mathcal{A}$. The major considerations here are:

- **Sample Complexity**: how many $n = n(\varepsilon, \delta)$ are needed to (ε, δ)-learn the statistic θ of a distribution $\mathcal{P}$.
- **Computational Complexity**: how efficiently can we implement $\mathcal{A}$, in terms of number of atomic operations (runtime) and space usage?

To make this more concrete, consider the example of estimating the bias of a coin.

Suppose $X = \{0, 1\}$ and $\mathcal{P}$ is a set of distributions over X (i.e. Bernoullis). These distributions are parametrized by $[0, 1]$. Let $\theta : \mathcal{P} \to [0, 1]$ encode that parametrization. The simplest learning algorithm, call it $\mathcal{A}_{\text{mean}}$, simply returns the mean of the n i.i.d. samples.

To illustrate the basic workflow of a computational learning theorist, we demonstrate the sample complexity of $\mathcal{A}_{\text{mean}}$. Computational complexity, in this case, is not super interesting, since $\mathcal{A}_{\text{mean}}$ is as efficient as it gets (adding n numbers takes time roughly linear in n).

Theorem 1. *$n = \Omega(\log(1/\delta)/\varepsilon^2)$ samples is sufficient for $\mathcal{A}_{mean}$ to (ε, δ)-learn the Bernoulli parameter estimation problem.*

Proof. Let s be the parameter of a given input distribution P. If $s_n = \frac{1}{n} \sum_{i=1}^{n} x_i$ is the average of the i.i.d. samples from P, we know that, by a standard Hoeffding bound, $\mathbb{P}(|s_n - s| \geq \varepsilon) \leq \exp(-2\varepsilon^2 n)$. □

It turns out that, as a function of δ and ε, you cannot do any better than $\mathcal{A}_{\text{mean}}$ for this problem. We show this rigorously in Section 4.

We can map the basic objects of statistical learning onto the manifold dimension estimation problem, as well as other manifold learning problems.

- Let X be the space of nice manifolds. Let $\mathcal{P} \subset \Delta(X)$ be the set of nice distributions on these nice manifolds. We could specify that $\mathcal{P}$ only consists of uniform distributions, for instance.
- In the case of **manifold dimension estimation**, $\theta = \theta_{\text{dim}}$ takes in a distribution over a manifold and outputs the dimension of that underlying distribution.
- One could formulate many other interesting learning problems over manifolds, such as estimating the reach [Aam+19], coordinate charts, or isometric maps into low dimension [Ver12] (naturally, for these cases where the statistic for each distribution is a function, determining a suitable notion of distance between statistics is delicate). [FMN16] considers the learning problem of determining whether data was sampled from a manifold to begin with!

It is important to have this general statistical learning framework in mind when we prove upper and lower bounds on the sample complexity of our dimension estimation algorithms. Zooming out, it is also important to note the limitations of this statistical learning framework—in particular, how realistic is it that our data points are sampled independently? One can certainly imagine learning frameworks where the learning algorithm can query for samples and adjust their queries based on what they have seen so far. This falls under the umbrella of active learning [Hsu10]. In this thesis, we stick to the most basic i.i.d. setting, but we note that active learning on manifolds is an interesting and relatively unexplored research direction.

2.3 Perspectives in manifold learning

The central practical purpose of manifold learning is dimension reduction: given manifold data embedded in a high ambient dimension, I want to work with it in its intrinsic manifold dimension. With this application comes three basic questions:

- How is this dimension reduction done?
- What are the reasonable desiderata of a manifold learning dimension reduction method? How hard is it to achieve these desiderata?
- How do I know which dimension should I embed into?

In this section, we answer the first two questions. We note that the last question is the key practical motivation for dimension estimation!

2.3.1 Dimension Reduction: It's (almost) all PCA

The story of dimension reduction begins with principal component analysis (PCA). PCA tells us that the top eigenvectors of the covariance matrix encode most of the relevant information in our data. It is an inherently linear method. In order to upgrade it to non-linear dimension reduction, we use a kernel map. This is called *Kernelized PCA*. We sketch out the simplest form of Kernelized PCA, known as Isomap [**isomap**].

Throughout this section, we typically write $X = [x_1, ..., x_n] \in \mathbb{R}^{D\times n}$ to denote a dataset. We usually let d denote the target dimension of our low-dimensional embeddings, where we assume $d \ll D$.

PCA: Best-Fitting Subspace The most natural and arguably most well-understood objective for dimension reduction is the task of finding the best-fitting linear subspace of some fixed lower dimension. An linear subspace of $\mathbb{R}^d$ is a subset spanned by some collection of elements. We typically model our data as being mean-centered or re-center it to fit this formulation.

We can formulate the best-fitting subspace problem as follows: let $\mathcal{A}_{D\times d}$ denote the set of $D \times d$ orthogonal matrices, i.e. $A^T A = I_d$. The idea here is that the columns of A provide an orthonormal basis for our d-dimensional subspace. It is easy to check that AA^T is the matrix that projects our data onto this d-dimensional subspace; the algebraic property of projection follows simply from $(AA^T)^2 = A(A^T A)A^T = AA^T$. So our subspace approximation problem becomes:

$$\text{PCA}(X) = \underset{A\in\mathcal{A}_{D\times d}}{\arg\min} \sum_{i=1}^{n} \|AA^T x_i - x_i\|_2^2 = \underset{A\in\mathcal{A}_{D\times d}}{\arg\min}\ \|AA^T X - X\|_F$$

where $\|\cdot\|_F$ denotes Frobenius norm. It turns out that there is an efficient algorithm for computing such an orthoprojector A, which comes down to eigendecomposition. It goes by two names.

- **Principal Component Analysis**: Take A where the d columns are the top d eigenvectors (i.e. highest eigenvalues) of the covariance matrix $\frac{1}{n}XX^T \in \mathbb{R}^{D\times D}$.
- **Singular Value Decomposition**: Take A where the d columns are the top d left singular vectors of X, i.e. if $X = \sum_{i=1}^n \sigma_i u_i v_i^T$ with σ_i in descending order then $A = [u_1, ..., u_d]$.

Here is a simple way of arguing that SVD and PCA recover the best-fitting subspace (albeit one that relies on some heavy machinery). Rewrite the loss as follows:

$$\sum_{i=1}^{n} \|UU^T x_i - x_i\|_2^2 = \|UU^T X - X\|_F$$

where $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$ denotes the Frobenius norm of a matrix. By the *Eckart-Young theorem*[1], the best rank-d approximation of X is given by the d-rank truncation of the singular value decomposition of X, call this $\hat{X}_d$. Our rank-d approximation is precisely

[1] Stated formally: Take $A \in \mathbb{R}^{n\times m}$ and let A_k be the kth order SVD truncation of A. Then $\|A - A_k\|_F \leq \|A - B\|_F$ for all $B \in \mathbb{R}^{n\times m}$.

the projected AA^TX. Setting these equal, we have:

$$\hat{X}_d = \sum_{i=1}^{d} \sigma_i u_i v_i^T = \sum_{i=1}^{D} \sigma_i \Big(AA^T u_i\Big) v_i^T = AA^T X$$

Equality is achieved for $AA^T u_i = u_i \cdot \mathbf{1}(i \le d)$. This is accomplished precisely when AA^T projects onto $\text{span}(u_1, ..., u_d)$, i.e. when $A = [u_1, ..., u_d]$.

Observe that the left singular vectors of X are precisely the eigenvectors of XX^T and the singular values of X are the square root of the eigenvalues of XX^T. This follows from $XX^T = U\Sigma V^T(U\Sigma V^T)^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T$. This factorization is precisely an eigendecomposition. *With this, we have shown how linear subspace approximation reduces to an eigenvalue problem.*

The benefit of PCA, aside from its algorithmic and algebraic straightforwardness, is its interpretability. By looking at the magnitude of the eigenvalues, we can measure exactly how much of the interpoint distance information we keep in our embedding.

The downside of PCA is that some interpoint distances may be affected more dramatically than others. One may wonder if there are algorithms which approximately preserve *every single interpoint distance.* Miraculously, a simple random linear projection into roughly $\log(n)$ dimensions does the job.

Theorem 2 (Johnson-Lindenstrauss Lemma, [DG03]). *For any $\{x_i\}_{i\in[n]} \subset \mathbb{R}^D$ there exists $R \in \mathbb{R}^{d\times D}$ with $d = \Omega(\log(n)/\varepsilon^2)$ such that for all distinct $i, j \in [n]$,*

$$(1-\varepsilon)\|x_i - x_j\|_2 \le \|Rx_i - Rx_j\|_2 \le (1+\varepsilon)\|x_i - x_j\|_2. \tag{2.1}$$

Furthermore, if R is a matrix with i.i.d. Gaussian entries, it satisfies the above property with high probability.

Random projections also have favorable properties for manifold-sampled data, in that they approximately preserve the geodesic distances [Ver11].

Kernelized PCA There are many ways of writing the PCA objective. The following form will be very useful for us as we generalize PCA to manifolds. Borrowing notation from earlier, let $Y = A^TX \in \mathbb{R}^{d\times n}$ be the PCA projection corresponding to an orthoprojector $A \in \mathbb{R}^{D\times d}$. Observe that

$$\arg\min_A \|AA^TX - X\|_F = \arg\min_A \|X^TAA^TX - X^TX\|_F = \arg\min_Y \|Y^TY - X^TX\|_F.$$

Intuitively, this is saying that we want to find a low-dimensional embedding whose inner products (and therefore interpoint distances) match X's as much as possible.

It will be convenient for us to rewrite the new formulation of PCA in terms of the interpoint distances of X. The following lemma gives us this.

Lemma 1. *Let $D \in \mathbb{R}^{n\times n}$. Then $D_{ij} = \|x_i - x_j\|^2$ for $\{x_i\} \subset \mathbb{R}^k$ if and only if the Gram matrix $B = -\frac{1}{2}HDH$ is positive semidefinite (PSD). Furthermore, if D is Euclidean embeddable, then $B = X^TX$ where $X = [x_1, ..., x_n] \in \mathbb{R}^{d\times n}$ is a point set such that $\|x_i - x_j\| = d_{ij}$.*

If $D = D_X$ is the interpoint distance matrix of X, then we can once again rewrite the PCA objective as:

$$\mathrm{PCA}(X) = \arg\min_{Y} \|Y^T Y + \frac{1}{2} H D_X H\|_F$$

If we are only given the interpoint distance matrix of the high-dimensional points, rather than the high-dimensional points themselves, this problem is known as **multi-dimensional scaling**.

This idea of changing what we put in place of $X^T X$ lends itself to a powerful generalization of PCA known as **kernelized PCA**. Let K_X be a kernel matrix, which is effectively just a proxy for $X^T X$. The kernelized PCA of X according to kernel K_X is

$$\mathrm{Kernel\,PCA}(X) = \arg\min_{Y} \|Y^T Y - K_X\|_F.$$

We typically imagine K_X as being the inner product matrix of some feature expansion of the dataset $[\phi(x_1), ..., \phi(x_n)]$, i.e. $[K_X]_{ij} = \phi(x_i) \cdot \phi(x_j)$, though this is not crucial for our purposes. For multidimensional scaling, we have that $K_X = X^T X = -\frac{1}{2} HDH$. The idea of practically all standard manifold learning methods—Laplacian eigenmaps [BN01], locally linear embedding [SR00], maximum variance unfolding [WS06], and so on—is to change K_X so that it reflects the intrinsic manifold geometry of the data. We discuss one of the first and simplest instantiations of kernelized PCA.

Isomap [TSL00], a portmanteau for isometric mapping, is the simplest such manifold learning method. The idea of Isomap is to use the shortest path distance between the input points as a proxy for the *geodesic distance* between points on the manifold. We then use this pseudo-geodesic distance as input for multidimensional scaling. In the language of Kernel PCA, we use $K_X = -\frac{1}{2} H D_{(\text{geodesic})} H$ where:

$$D_{\text{geodesic}} = \min_{P \in \mathcal{P}} \sum_{(i,j) \in P} \|x_i - x_j\|^2 \qquad \mathcal{P} = \{\text{paths in adjacency matrix of } X\}$$

One can show that, with a large and tight enough sample over a sufficiently well-behaved manifold, the shortest path distance matches the geodesic distance.

Theorem 3 ([Ber+00])**.** *Let M be a compact submanifold of $\mathbb{R}^D$ and $\{x_i\}$ be a finite set of data points on M. Let G be a nearest-neighbors graph on $\{x_i\}$. Pick any $\varepsilon \in (0, 1)$. If M is geodesically convex and the graph G and the dataset $\{x_i\}$ satisfy suitable conditions (which get stricter for smaller ε), then for large enough sample size n we have:*

$$(1 - \varepsilon) d_M(x, y) \leq d_G(x, y) \leq (1 + \varepsilon) d_M(x, y)$$

This theorem implies that, if the geodesic metric on the manifold is Euclidean embeddable, then the MDS step of Isomap produces an approximately isometric embedding of the input data in the manifold geodesic metric. It is known that this asymptotic result can be converted into a finite-sample method, but it has not been done in the literature. We consider this an excellent exercise for the reader.

2.3.2 Desiderata and the Nash Embedding Theorems

We have guarantees that linear dimension reduction methods can preserve the metric structure of the input data, either in an average-case sense, as in the case of PCA, or a worst-case sense, as in the case of the Johnson-Lindenstrauss transform. Isomap seeks to create an approximate isometry for the *interpoint geodesic distances* using the shortest path metric on the sampled points. The consistency theorem shows that this is possible on a restricted class of manifolds—namely, those which are geodesically convex and isomorphic to Euclidean space. The following theorem from differential geometry shows us that we can hope for better.

Theorem 4 (Nash Embedding Theorem)**.** *Any compact d-dimensional Riemannian manifold (M, g) can be isometrically C^1 embedded in Euclidean space of dimension $2d + 1$ and C^∞ embedded in dimension $O(d^2)$.*

This is a remarkably general theorem with a remarkably simple proof, see [Nas56]. Note that it does not even require that the input manifold is embedded in Euclidean space! The proof essentially works by working with a smooth embedding of the manifold into $\mathbb{R}^{2d+1}$ (guaranteed by the Whitney embedding theorem, see [Lee12]). This embedding is then scaled down such that all the interpoint distances are less than the true geodesic distances. Then the manifold distances are corrected by "twisting" the manifold, increasing the geodesic distances such that asymptotically, they reach their true desired values.

This procedure can be converted into a finite-sample algorithm, which learns the isometric embedding into $O(d)$ space. Note that this is the output dimension even for arbitrarily precise $(1 \pm \varepsilon)$-approximately isometric embeddings—unlike, say, the Johnson-Lindenstrauss transform, which requires dimension scaling like $\Omega(1/\varepsilon^2)$. The dependence on and the ambient dimension D instead show up

Theorem 5 (Approximate Nash Embedding Algorithm, informal, see [Ver12])**.** *Given a sufficiently tight finite sample of n points from X a C_M-regular, volume V, m-dimensional compact, connected manifold of global reach τ embedded in $\mathbb{R}^D$, one can compute efficiently a map $\mathcal{A} : \mathbb{R}^D \mapsto \mathbb{R}^d$ such that:*

- *$\mathcal{A}$ produces a $(1 \pm \varepsilon)$-isometric embedding of M into $\mathbb{R}^d$,*
- *$d = \Omega(m + \log(V/\tau^m))$ (assuming $d \leq D$), and*
- *$n = \Omega(V(\frac{D}{\tau\varepsilon d})^d)$.*

Note that this algorithm only works in the noiseless case. We discuss the difficulties of noise briefly in the next section.

2.3.3 Difficulty of Manifold Learning in the Presence of Noise

Suppose you are given n-point i.i.d. samples from one of two distributions, either

- a uniform distribution on a unit sphere $S^d \subset \mathbb{R}^{d+1}$, or
- a uniform distribution on an equator of this sphere $\cong S^{d-1}$,

and your task is to distinguish where the sample came from. If there is no noise, this problem is easily solvable with $d + 1$ samples, by simply checking if all the points are co-equatorial. However, the moment that you add an arbitrarily small amount of independent Gaussian noise to all of these points, the decision problem suddenly requires $\Omega(e^d)$ samples to solve [Wei14]. This is follows from the so-called Gaussian Annulus Theorem and is representation of a broader pattern in high-dimensional statistics called the concentration of measure phenomenon [Weg21]. Complexity results like these pose a formidable challenge to the prospect of dimension estimation in the presence of noise. They can be outmaneuvered using adaptive algorithms, assuming certain special structure on the manifold like bounded curvature, or restricting our attention to the noiseless case of the problem.

Chapter 3

Algorithms

Most of the aforementioned manifold learning methods require, as a hyperparameter, the dimension of the output embedding. Without an accurate guess of the true manifold dimension of the data, most algorithmic guarantees are moot. This is one key motivation for intrinsic dimensionality estimation. Other motivations are complexity-oriented: the runtime of many important machine learning algorithms (e.g. density estimation, kd-trees, nearest-neighbor search) have exponential dependence on the intrinsic dimension of data. It is useful to know these finite-sample convergence rates ahead of time.

The most common strategy for manifold intrinsic dimension estimation is as follows: analyze some **local statistic** that scales with dimension, and then use the observed scaling behavior to reverse-engineer a guess for the intrinsic dimension. In this function we discuss three local statistics that can be leveraged to construct estimators.

- **Local covariance structure**, i.e. PCA in a neighborhood [DF08; Lit+09].
- Number of **nearest neighbors** within a given radius r [LB04; FSA07].
- **Covering number**, i.e. number of boxes or balls of size r needed to cover the manifold [Kég02].

We also discuss two notable **global methods**, i.e. estimators which observe the scaling behavior of statistics relating to the entire dataset.

- **Minimal subgraphs**, e.g. length of the traveling salesman path [KRW16] or minimum spanning tree [CH06] on the interpoint distance graph.
- **Convergence rate of empirical measure to the true measure**, e.g. under the Wasserstein metric [Blo+22]. They analyze the scaling of this metric with respect to the size of the sample.

Another flavor of global methods involved running dimension reduction algorithms like multidimensional scaling and compare the errors of the output. Naturally, the observed error should decrease as the dimension increases, but the idea here would be that the true dimension is at the ”elbow” of this curve, where the marginal benefits of using more dimensions for the output embedding start to decay significantly. This kind of “elbow method” happens to have a provable guarantee in the linear dimension estimation case (Theorem 6).

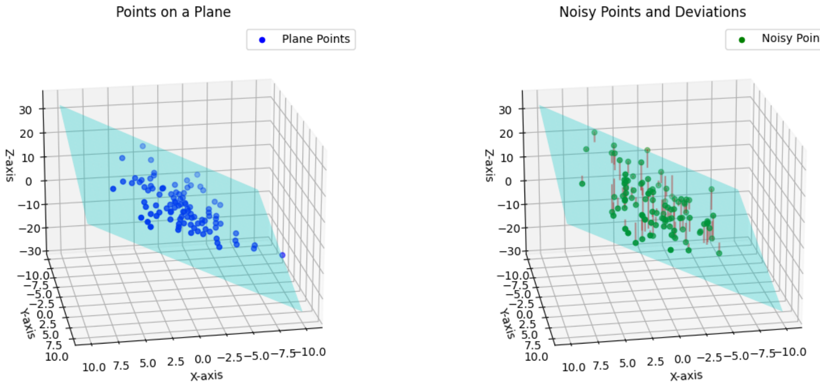

Figure 3.1: **Linear Case**. Example of Gaussian data generated on a two-dimensional plane in a three-dimensional space, noiselessly (left) and noisily (right).

3.1 Dimension Estimation: Linear Case

Probabilistic PCA In order to appreciate the manifold intrinsic dimension estimation problem, it is essential that we have a good grasp on the case where the manifold of interest has no curvature. In other words, we would like to estimate the dimension of a linear subspace, based on finite samples. The solution, as we will see, depends crucially on the PCA method, discussed at length in the previous section.

As per [TB99], PCA is known to arise from maximum likelihood estimation on the following generative model:

$$y = Wx + \mu + \varepsilon \in \mathbb{R}^D \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_D) \quad x \sim \mathcal{N}(0, I_d) \tag{3.1}$$

with $\mu \in \mathbb{R}^D$, $W \in \mathbb{R}^{D \times d}$ an orthogonal transformation, and $d \ll D$ presumably. We assume isotropic noise. Note that in the noiseless regime ($\sigma = 0$) the problem of dimension estimation is trivial:

Remark 1. *If $\sigma^2 = 0$, then the following algorithm computes the dimension of the manifold with exactly $n = d + 1$ samples (almost surely):*

Algorithm 1 Linear ID Estimation: Noiseless Case

```
Require: Samples $\{y_1, ..., y_n\} \subset \mathbb{R}^d$, i.i.d. according to (3.1).
  for $t \in [n]$ do
    If $\{y_1, ..., y_t\}$ are linearly dependent, terminate and output $t - 1$.
  end for
```

For $\sigma^2 > 0$, we can develop a simple maximum likelihood estimator. Due to the additivity of Gaussian distributions, it is easy to analyze y conditioned on the value x.

$$y \mid x \sim \mathcal{N}(Wx + \mu, \sigma^2 I_D)$$

If you marginalize, taking $p(y) = \int_x p(y|x)p(x)dx$, then indeed y is still a multivariate normal distribution. If $\Sigma = WW^T + \sigma^2 I_D$, then $y \sim \mathcal{N}(\mu, \Sigma)$. Since W is an orthogonal matrix, Σ is a diagonal matrix with all entries σ^2 except for d entries of size $1 + \sigma^2$. We set up maximum-likelihood estimate in the standard manner.

$$d_{\text{MLE}} = \operatorname*{arg\,min}_{d' \in [D]} \min_{W,\mu} \mathbb{P}\Big(y_1, ..., y_n \;\Big|\; W, \mu\Big)$$

We may take the logarithm and simplify the density of the multivariate Gaussian to obtain:

$$\operatorname*{arg\,min}_{d' \in [D]} \min_{W,\mu} -\frac{N}{2}\Big(d' \ln(2\pi) + \ln(\det \Sigma) + \operatorname{tr}\Big(\Sigma^{-1} \cdot \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)(y_i - \mu)^T\Big)\Big)$$

It is shown in [TB99] that, for fixed dimension this objective recovers the usual PCA method discussed in Section 2. If we know the noise ahead of time, we can use this information to filter out the signal principal components from the ones that arise by noise, and thereby detect the true dimension of the data. The algorithm is as follows:

Algorithm 2 Linear ID Estimation: Noisy Case, **Covariance-based Estimator**

```
Require: Samples $\{y_1, ..., y_n\} \subset \mathbb{R}^D$, i.i.d. according to (3.1).
Require: Cutoff parameter $\eta$.
  Compute sample covariance matrix $\Sigma_n = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)(y_i - \mu)^T$ where $\mu = \frac{1}{n}\sum_i y_i$.
  Return $\widehat{d}$ as the number of eigenvalues of $\Sigma_n$ of size $\geq \eta$.
```

In order for this algorithm to work, we need to make sure the signal is strong enough that it does not get drowned out by the noise. This is captured in the condition we set in the following theorem. Let $\sigma_i(X)$ denote the ith largest singular value of a matrix X and $\lambda_i(Y)$ denote the ith largest eigenvector of a symmetric matrix Y. Let $\operatorname{srank}(X) = \sum_i \lambda_i(X)/\max_j \lambda_j(X)$ denote the stable rank of a symmetric PSD matrix X. Stable rank is a lower bound on the true rank and is typically a good proxy for it. For Σ we know $\operatorname{srank}(\Sigma) = (d + \sigma^2 D)/(1 + \sigma^2)$. We prove the following guarantee in our simple Gaussian generative model.

Theorem 6. *With cutoff parameter $\eta = (1/2) + \sigma^2$, $\sigma^2 < 1$, and sample size*

$$n = \tilde{O}(\operatorname{srank}(\Sigma)) = \tilde{O}(d + \sigma^2 D),$$

Algorithm 2 outputs $\widehat{d} = d$ with constant success probability.

In order to prove this, we invoke the following theorem regarding the convergence of the empirical covariance matrix to its true value. Once the estimate is close enough, the signal and noise principal components are separated and our estimator gets exactly the right dimension.

Theorem 7 (see [Ver10], Corollary 5.52)**.** *Let X_i denote independent samples from a distribution in $\mathbb{R}^D$ with Σ. Suppose the distribution is such that $\mathbb{E}\max_{i \in [n]} \|X_i\|_2^2 \leq m$.*

Let $\varepsilon \in (0,1)$, $t \geq 1$. Then with probability $\geq 1 - D^{-t^2}$,

$$n \geq C(t/\varepsilon)^2 \|\Sigma\|^{-1} m \log D \implies \mathbb{E}\|\Sigma_n - \Sigma\| \leq \varepsilon \|\Sigma\|$$

where $\Sigma_n = \frac{1}{n}\sum_{i=1}^n X_i \otimes X_i$, the sample covariance matrix, $\|\cdot\|$ is the operator norm, and C is a universal constant.

Proof of Theorem 6. For all $i \in [n]$, we have $\mathbb{E}\|X_i\|^2 = \mathrm{tr}(\Sigma) = \mathrm{srank}(\Sigma)\|\Sigma\|$. This means $\mathbb{E}(\max_{i \in [n]} \|X_i\|_2^2) \leq \mathrm{srank}(\Sigma)\|\Sigma\| \log n = m$. Let $t \geq \sqrt{1/\log D}$ such that $D^{-t^2} \leq 1/2$. If $n \geq \tilde{\Omega}(\mathrm{srank}(\Sigma)/\varepsilon^2)$, then with constant probability,

$$n \geq C(t/\varepsilon^2)\|\Sigma\|^{-1} m \log D \geq \frac{C\|\Sigma\|^{-1}(\mathrm{srank}(\Sigma)\|\Sigma\|)(\log n)\log D}{\varepsilon^2 \log D}$$

and therefore $\mathbb{E}\|\Sigma_n - \Sigma\| \leq \varepsilon\|\Sigma\|$. With constant probability we have $\|\Sigma_n - \Sigma\| \leq \varepsilon\|\Sigma\|$. By definition of the operator norm, this means for all i,

$$|\lambda_i(\Sigma_n) - \lambda_i(\Sigma)| \leq \varepsilon \lambda_{\max}(\Sigma) = \varepsilon(1 + \sigma^2)$$

For $\epsilon < \frac{1}{4(1+\sigma^2)}$ the eigenvalues above and below the cutoff will stay above and below the cutoff respectively. This means our estimator will only count the signal eigenvalues. □

This covariance-based method is very commonly used in practice, since it goes hand-in-hand with computing the actual low-dimensional subspace for PCA. The difficulty in any practical application of this algorithm is determining the right cutoff between signal and noise. There are heuristics one might use to do this: for instance, plotting the reconstruction error of the various principal components and seeing where the marginal benefits of including more principal components seems to start diminishing (some call this the elbow method). One particularly interesting way to do this is to use the predictions of random matrix theory to distinguish the noise directions from the signal directions. In practice, this comes down to fitting a Marchenko-Pastur distribution (the limiting distribution of singular values of a Gaussian random matrix) to the eigenvalues and only taking eigenvalues after the predicted threshold [Apa+20].

Note that one could develop a conceptually simpler and more computationally efficient dimension estimator based on the variance of the data, which is $d + \sigma^2 D$ in expectation. The sample variance should approach this reasonably fast.

Algorithm 3 Linear ID Estimation: Noisy Case, **Variance-based Estimator**

```
Require: Samples $\{y_1, ..., y_n\} \subset \mathbb{R}^D$, i.i.d. according to (3.1).
  Compute unbiased sample variance $V_n = \frac{1}{n-1}\sum_{i=1}^n \|y_i - \mu\|^2$ where $\mu = \frac{1}{n}\sum_i y_i$.
  Return $\hat{d} = V_n - \sigma^2 D$
```

This is not as widely used in practice as the covariance-based estimator, since it does not help compute the actual subspace and its practical success is heavily dependent on a good estimate of σ^2. However, it is quite nice from a theoretical perspective, since it outputs a real-numbered estimate rather than an integer. We leave it as an exercise to the reader to analyze this estimator.

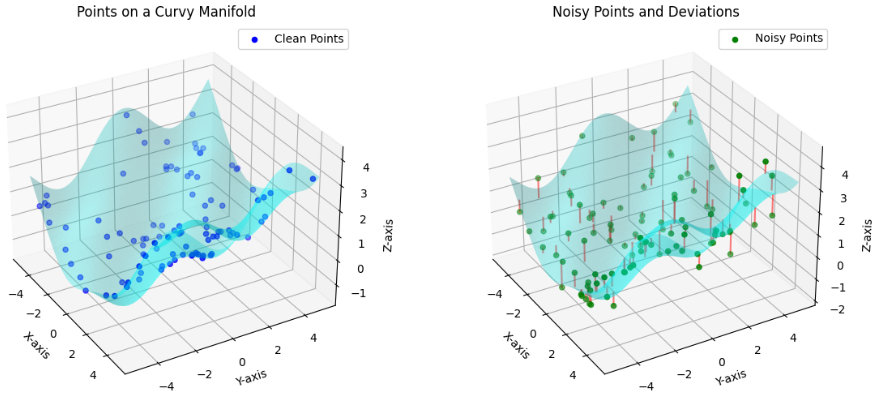

Figure 3.2: **Nonlinear Case**. Illustration of data randomly sampled from a manifold with nontrivial curvature, noiselessly (left) and noisily (right).

3.2 Topological Notions of Dimension

In this section we describe a number of classical, asymptotic notions of dimension, which describe topological spaces more generally than manifolds. Since manifolds are in many regards some of the most well-behaved topological spaces, these notions of dimension almost always coincide with the notion of manifold dimension. While they are often not efficiently computable, they will help us develop intuition about how to learn manifold dimension. We begin with the notion of covering numbers and Minkowksi dimension.

Definition 14. *Let X be a topological space with a measure μ on it. For $S \subset X$, let the **covering number** $\mathcal{N}_\varepsilon(S)$ be the infimum of the number of balls of radius ε needed to cover S;. Similarly, let the **box-covering number** $\mathcal{B}_\varepsilon(S)$ denote the infimum of the number of boxes of side-length ε needed to cover S.*

*The **Minkowski dimension** (a.k.a. capacity dimension) of a set S is given by:*

$$d_M(S) = \limsup_{\varepsilon \to 0} \frac{\log \mathcal{N}_\varepsilon(S)}{\log(1/\varepsilon)}$$

*The **box-counting dimension** of a set S is similar:*

$$d_B(S) = \limsup_{\varepsilon \to 0} \frac{\log_\varepsilon \mathcal{B}_\varepsilon(S)}{\log(1/\varepsilon)}$$

The Minkowski and box-counting dimensions gauge dimension as a sort of scaling process: as ε decreases, the ε-covering numbers increase exponentially in d. This idea of exponential scaling in d is used everywhere in intrinsic dimension estimation.

We get a slightly different perspective through the Hausdorff dimension. While is may seem contrived at first glance, there is a simple intuition behind it all: namely, that an overestimate of the dimension of a set will result in us assigning zero measure to that set (e.g. a square has positive measure in $\mathbb{R}^2$ but zero measure in $\mathbb{R}^3$).

Definition 15. *The* ***Hausdorff dimension*** *of a set $S \subset X$ is given by:*

$$d_H(S) = \inf\{d : \mu_H^{(d)}(S) = 0\}$$

where $\mu_H^{(d)}$ is the d-Hausdorff measure:

$$\mu_H^{(d)} = \lim_{\varepsilon \to 0} \inf \Big\{ \sum_{k=1}^{\infty} r_k^d : S \subset \bigcup_{k=1}^{\infty} B(x_k, r_k),\ r_k \le \varepsilon\ \forall k \Big\}$$

A more intuitive but much less robust related notion of dimension is the local Hausdorff dimension, which depends on the choice of a measure.

Definition 16 (see [CS16])**.** *Let μ be a probability distribution on $S \subset X$. If the following limit exists, it is the pointwise or* ***local Hausdorff dimension****.*

$$d_{lH} = \limsup_{\varepsilon \to 0} \frac{\log \mu(B(x, \varepsilon))}{\ln(\varepsilon)}$$

Definition 17 (Assouad 1983)**.** *The* ***doubling dimension*** *of a set $S \subset \mathbb{R}^D$ is:*

$$d_A(S) = \inf\{d : \forall B(x,r) \subset \mathbb{R}^D,\ \mathcal{N}_{r/2}(B(x,r) \cap S) \le 2^d\}$$

Definition 18. *Let (X, d) be a metric space. A set $V \subset X$ is* ***r-separated*** *if $d(x, y) \ge r$ for all distinct $x, y \in V$. The r-**packing number** $M_\varepsilon(S)$ of a set $S \subset X$ is the maximum cardinality of an ε-separated subset of S.*

Fact 2 (see [Kég02])**.** *A basic inequality between packing and covering numbers holds:*

$$\mathcal{N}_\varepsilon(S) \le M_\varepsilon(S) \le \mathcal{N}_{\varepsilon/2}(S)$$

This implies that the Minkowski dimension can be rewritten in terms of packing numbers:

$$d_M(S) = \limsup_{\varepsilon \to 0} \frac{\log(M_\varepsilon(S))}{\log(1/\varepsilon)}$$

3.3 Local Methods

3.3.1 Estimating Packing Numbers

The idea of [Kég02] is to use packing numbers to estimate Minkowski dimension and thereby manifold dimension. The natural definition—approximating the limit that is as follows:

Definition 19. *The (r_1, r_2)-**scale-dependent capacity dimension** (where $r_2 > r_1$) of a finite set $S = \{x_1, ..., x_n\}$ is defined as follows:*

$$\hat{d} = -\frac{\log M_{r_2}(S) - \log M_{r_1}(S)}{\log r_2 - \log r_1}$$

The next question would be: how do we calculate packing numbers of finite sets? The following result would seem to suggest that this is a hopeless endeavor.

Claim 1. *Computing $M_\varepsilon(S)$ for $S = \{x_1, ..., x_n\} \subset \mathbb{R}^d$ is NP-hard.*

Proof. There is a simple reduction from maximum independent set, an NP-complete problem: take the graph of where the vertices are the points in S and the vertices have edges only if the correspond points are a distance r away. Then the size of the maximum independent set in this graph corresponds to the max subset of points that are all a distance ε from each other. □

To make matters worse, maximum independent set is NP-hard to approximate within a factor of $n^{1-\varepsilon}$ for all $\varepsilon > 0$ [Kég02]. However, there still is hope for approximation: on weighted disk graphs (i.e. the kind of graph that comes up for estimating packing numbers in two dimensions) there are poly-time approximation schemes. This PTAS extends to higher dimensions, at the expense of an exponential dependence on the dimension [EJS05].

To avoid this exponential dependence, [Kég02] implements a greedy algorithm which appears to work well in practice but lacks proven guarantees.

Algorithm 4 Kegl's algorithm to estimate $M_\varepsilon(S)$

```
Samples $S = (y_1, ..., y_n) \subset \mathbb{R}^d$. Set of centers $C = \phi$.
Let $\overline{C} = C \cup \{i \in [n] : \exists c \in C$ s.t. $\|y_i - c\| \leq \varepsilon\}$.
while $\overline{C} \neq S$ do
    Randomly permute $S$.
    Iterate through $S \setminus C$, add up to $|C|$ points to $C$.
end while
```

With the estimate for $M_\varepsilon(S)$ at different scales, one can compute the scale-dependent capacity dimension and average. A more robust handling, in fact, would be to plot the logarithms of the capacity dimension against the radius and use the slope of the least-squares regressor.

3.3.2 Correlation Dimension

The idea of [GP83]'s method is to use a dimension estimator more directly adapted to the setting of estimating manifold dimension given a stream of finite samples. Note the similarity to the local Hausdorff dimension, except instead of looking pointwise we consider all pairwise distances.

Definition 20 (Correlation Dimenson)**.** *Let $\{y_i\}_{i\in\mathbb{N}}$ be a sequence of elements sampled i.i.d. from some metric space (X, d). The correlation integral is:*

$$C(\varepsilon) = \lim_{l\to\infty} \frac{1}{\binom{l}{2}} \sum_{i=1}^{l} \sum_{j=i}^{l} \mathbf{1}[d(y_i, y_j) \leq \varepsilon]$$

The ***correlation dimension*** *of X is given by:*

$$d_{Corr} = \lim_{\varepsilon\to 0} \frac{\ln(C(\varepsilon))}{\ln(\varepsilon)}$$

Unlike the packing number, the correlation integral is easy to compute: one only needs to iterate over every pairwise distance and check if it small enough. A natural problem with this kind of approach is: what is the appropriate scale to look at? [HA05] address this issue using a fact about U-statistics (aside: a U-statistic is a class of statistics defined as the average over the application of a given function applied to all tuples of a fixed size).

Definition 21. *Let $\{y_i\}_{i=1}^n$ be sampled from a d-dimesional submanifold of $\mathbb{R}^D$. The empirical Hein-Audibert correlation dimension is given by:*

$$U_{l,d}(\varepsilon) = \frac{1}{\binom{l}{2}} \sum_{i=1}^{l} \sum_{j=i}^{l} \varepsilon^{-d} K(\|y_i - y_j\|^2/\varepsilon^2)$$

where K is a generic non-negative function.

The main insight of [HA05] is that there is a "correct" bandwidth ε to look at, in the sense that U_l only converges if $l\varepsilon^d \to \infty$. The algorithm involves looking at different values of l (up to the size of the dataset, of course). Here is an overview:

- Fix a scaling of $\varepsilon = \varepsilon_d(l)$ as a function of l and d.
- Break data into subsamples of varying sizes $N_1, ..., N_k = [n]$.

  Compute for each $d' \in [D]$

  $$S_{d'} = \{(\log \varepsilon_{d'}(N_i), \log U_{N_i,d'}(\varepsilon_{d'}(N_i)))\}_{i \in [k]}$$

- Choose d^* minimizing the least-squares estimated slope through $\{U\}$

3.3.3 Local PCA

The most naive approach to manifold intrinsic dimension estimation is to attempt a sort of *local PCA*. This comes from the mathematical understanding that the dimension of the tangent space at a point in a manifold is equal to the dimension of the manifold itself. Here is one formalization of such a concept.

Definition 22 (Definition 2 in [DF08])**.** *Set $S \subset \mathbb{R}^D$ has* ***local covariance dimension*** *(d, ε, r) if its restriction to any ball of radius r has covariance matrix whose largest d eigenvalues satisfy $\sum_{i\in[d]} \sigma_i^2 \geq (1-\varepsilon) \sum_{i\in[D]} \sigma_i^2$.*

Note that this definition applies to arbitrary subsets of Euclidean space, not necessarily manifolds. It was shown in [DF08] that random projection trees—a variant of spatial decision trees or *kd*-trees, where partitions of the space are chosen randomly rather than axis-aligned—is able to adapt to this particular notion of intrinsic dimension of data. What these means is that the depth of the random projection tree required to sufficiently bucket all of the point scales with this intrinsic local covariance dimension rather than the ambient dimension of the space.

Ultimately, like other ID estimation techniques, local PCA suffers from a scaling problem: what is an appropriate neighborhood size r to take these restricted eigenvalue estimates? This motivates [Lit+09]'s multi-scale approach to local PCA, where these eigenvalue estimates are taken at multiple resolutions and the scaling of these estimates are used to determine intrinsic dimension. We sketch this approach below. Let $\sigma_i^{(r)}(z)$ denotes the ith singular value ($i \in [D]$) of the covariance matrix of the points $S \cap B_r(z)$. The idea is to compute $\sigma_i^{(r)}(z)$ for some representative points $z \in M$ and a range of radii $r > 0$. Based on scaling of singular values with respect to r, we classify that singular value as signal or non-signal. The split is roughly as follows:

- The singular value directions corresponding to noise are the ones whose singular values do not scale with r.
- The singular value directions corresponding to curvature or self-intersections are those whose singular values scale quadratically (or higher) with r.
- The singular value directions corresponding to tangent directions (signal) are those whose singular values scale linearly with r.

This analysis of the behavior at different resolutions helps determine an appropriate neighborhood size at different parts of the manifold (e.g. we can model flatter regions with larger neighborhood sizes than curvier regions).

3.3.4 Nearest Neighbors

If points are evenly sampled on a D-dimensional manifold, then it indeed the case that a radius r ball should expect to contain $O(r^D)$ points. For finite data, we understand this as the rate at which nearest neighbors appear in a growing ball. This is the derivation behind, for instance, the popular MLE estimator of [LB04]. The estimator, for a fixed data point y, is given by:

$$\hat{d}_k(y) = \Big[\frac{1}{k-2}\sum_{j=1}^{k-1} \log \frac{T_k(y)}{T_j(y)}\Big]^{-1}$$

where $T_k(y)$ is the Euclidean distance from a data point y to its kth nearest neighbor.

The derivation of this estimator is largely heuristic and involves modeling the sampling from the manifold in a small enough neighborhood as a homogenous Poisson process. While asymptotically consistent, it is relatively difficult to establish finite-sample guarantees. Instead, we present a very similar estimator by [FSA07].

The analysis proceeds as follows: define

$$\eta(\mu, r) = r^{-d} \cdot \mathbb{P}(y_i \in B(\mu, r))$$

It turns out, if we sample points uniformly on a manifold with standard regularity assumptions, then $\eta(x, \cdot)$ is slowly varying for small enough r. This gives rise the following (approximate) relationship between the rank of a nearest neighbor and its distance.

$$k/n = \eta_0 \cdot [T_k(x)]^d$$

The trick is to take the logarithm of the above and see it as a function that is linear in d. We can calculate the slope of this function given two points.

$$\ln(k/n) = \ln(\eta_0) + d\ln(T_k(x))$$

Noticing the linear relationship, we can relatively easily solve for d and use this as the basis of our estimator.

$$\hat{d}(x) = \frac{\ln(2)}{\ln(T_k(x)) - \ln(T_{\lceil k/2 \rceil}(x))}$$

There are two straightforward ways in which we can combine the estimator at different points in order to give a holistic estimate of intrinsic dimension: the authors call this "averaging" versus "voting."

$$\hat{d}_{\text{avg}} = \frac{1}{n}\sum_{i=1}^{n} \hat{d}(x_i) \wedge D \qquad \hat{d}_{\text{vote}} = \operatorname*{arg\,max}_{d' \in \mathbb{N}^+} \sum_{i=1}^{n} \mathbf{1}[\hat{d}(x_i) = d']$$

Starting with a guarantee on the individual point-estimates of intrinsic dimension, and combining these using McDiarmid's inequality and a counting argument relying on the covering of a manifold by cones, the authors arrive at the following exponential rates of convergence of the estimators.

Theorem 8. *For constants $c_1, c_2, c > 0$ we have:*

$$\mathbb{P}(\hat{d}_{vote} \neq d) \leq \exp\left(\frac{-c_1 n}{(c^d k)^2}\right) \qquad \mathbb{P}(\hat{d}_{avg} \neq d) \leq \exp\left(\frac{-c_2 n}{(Dc^d k)^2}\right)$$

In particular, for $n \geq O(k^2 c^{2d} \log(1/\delta)/c_1)$, we have that $\hat{d}_{vote}$ is a correct estimate of the dimension with probability $\geq 1 - \delta$.

3.4 Global Methods

Local methods generally focus on estimating the dimension of the tangent space. They suffer from a certain adaptivity problem: one must deduce the right neighborhood size for which the manifold has this approximate linearity. Global methods, on the other hand, look at statistics that somehow depend on the entire dataset. We point out two such methods.

3.4.1 Minimal Subgraphs

A metric on points is essentially a weighted complete graph. Certain subgraphs of this graph can be used to detect the intrinsic dimension the data.

Traveling Salesman Path Given an undirected complete weighted graph, the traveling salesman path is a tour through the graph (i.e. a cycle going through all vertices of the graph) of minimum weight. The idea of [KRW16] was to use a traveling salesman path, weighted by interpoint distance, to estimate intrinsic dimension.

$$\text{TSP}(X_{1:n}; d_1) = \min_{\sigma \in S_n} \left\{ \sum_{i=1}^{n} \|X_{\sigma(i+1)} - X_{\sigma(i)}\|_{\mathbb{R}^D}^{d_1} \right\} \tag{3.2}$$

Consider a binary version of the ID estimation problem, where the learner is given samples from one of two distributions, with manifold dimension d_1, d_2 respectively.

Lemma 2. *If* $\mathrm{TSP}(X_{1:n}; d_1) \leq O(\max\{1, \tau_g^{d_1 - D}\})$*, return* $\hat{d} = d_1$*. Otherwise, return* $\hat{d} = d_2$*. For* $1 \leq d_1 < d_2 \leq D$ *and* $\tau_g = \Theta(1)$ *the global reach of the manifold, then*

$$\mathbb{P}(\hat{d} \neq d) = O(n^{-n(d_2/d_1 - 1 - \epsilon)})$$

where ϵ *is an arbitrarily small positive constant and the probability is over the randomness in any sufficiently well-behaved sample (see the assumptions in [KRW16]).*

Note that this estimator is presumably intractable to compute since traveling salesman is an NP-hard problem [Zam06]. Note as well that the larger the ratio d_2/d_1, the faster the convergence of the error rate.

Minimum Spanning Tree Following [CH03], we consider the use of the *minimum spanning tree* (MST) of an undirected weighted graph as a proxy for intrinsic dimension. Their method is considerably different from the TSP method and proceeds as follows:

- Compute a nearest neighbor graph with distance-weighted edges. Construct a shortest-path metric over this nearest neighbor graph (in the spirit of Isomap), call it D_n. Let $\mathcal{T}_n$ denote the set of spanning trees over D_n. Consider the γ-power-scaled minimum spanning tree cost:

$$L_\gamma(D) = \min_{T \in \mathcal{T}_n} \sum_e |D_e|^\gamma$$

- This quantity scales asymptotically with some function of the dimension of the manifold d. In particular, if the data is a conformal mapping from $\mathbb{R}^d$ to $[0,1]^D$, we have

$$\lim_{n \to \infty} L_\gamma(D)/n^{(d' - \gamma)/d'} = \begin{cases} \infty & d' < d \\ \text{[constant]} & d' = d \\ 0 & d' > d. \end{cases} \tag{3.3}$$

  Looking at the empirical scaling of the quantity on the left hand side with respect to n allows for us to estimate the true data dimension.

It would be interesting to develop guarantees for MST-based ID estimators which more closely resemble the simple TSP-based estimator (3.2). In other words, can we threshold the raw unscaled MST cost to distinguish between distributions of different dimension? This would provide an immediate improvement over the TSP-based algorithm since MST is poly-time computable (via Prim's or Kruskal's algorithm) dasgupta2006algorithms.

3.4.2 Convergence of Empirical Measure

Let $\mathbb{P}$ be a probability measure on a manifold M. With n independent samples supported on $\mathbb{P}$, one can construct an empirical distribution with a sum of Dirac delta functions:

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

A natural theoretical question is: in what sense does P_n converge to $\mathbb{P}$, and how fast? A priori, this may seem like a question that is completely unrelated to intrinsic dimension estimation. But, as pointed out by [Blo+22], the convergence rates depend on the dimension of the support and hence can be used to reverse-engineer the intrinsic dimension.

First of all, the most natural notion of convergence between distributions is known as weak convergence, i.e. convergence in distribution. It is well-known that the empirical measure converges to the true measure weakly.

Theorem 9 (Glivenko-Cantelli)**.** *$P_n \to \mathbb{P}$ in distribution (a.k.a. weakly), i.e. for all bounded continuous functions $f : \operatorname{supp}(\mathbb{P}) \to \mathbb{R}$ we have:*

$$\int_S f(x)\ dP_n(x) \to \int_S f(x)\ d\mathbb{P}(x)$$

Note that we can metrize weak convergence through the Wasserstein-p distance. This will be crucial to allow us to calculate convergence rates.

Definition 23. *Let μ, ν be two measures on a metric space (M, d). Let $\Gamma(\mu, \nu)$ be the set of couplings of the two measures (i.e. measures on the product space where μ, ν are the marginal distributions). Then the Wasserstein-p distance between μ and ν is given by:*

$$W_p^M(\mu, \nu)^p = \inf_{(X,Y)\sim\Gamma(\mu,\nu)} \mathbb{E}[d_G(X, Y)^p]$$

Theorem 10 (see [Vil+09], section 6)**.** *For μ_n distributions on a metric space $(\mathcal{X}, d)$ and $p \in [1, \infty)$, the following are equivalent:*

- *$\mu_n \to \mu_0$ weakly, and $\int_{\mathcal{X}} d(0, x)^p \mu(dx)$ for all $i \in \mathbb{N}_0$.*
- *$W_p(\mu_n, \mu_0) \to 0$ as $n \to \infty$.*

We are interested in the *rate* of convergence. The first result to this effect was [Dud69], later sharpened by [MNW24] under particular conditions. The main idea is that the rate of convergence is $W_1(\mathbb{P}, P_n) = \Theta(n^{-1/d})$, where d is the dimension of the support. The curse of dimensionality we observe here is actually turned into a blessing by [Blo+22], who use it to fashion the following estimator.

$$\hat{d}_n = \frac{\log \alpha}{\log W_1^G(P_n, P_n') - \log W_1^G(P_{\alpha n}, P_{\alpha n}')}$$

where W_1^G is the graph metric approximation of the Wasserstein-1 distance, and α is a suitably large natural number.

Note that they use a sort of symmetrization trick here: we do not have access to $\mathbb{P}$ to plug into the Wasserstein metric, but we can take an independent sample and its corresponding empirical measure P_n' and the convergence rate of $W_1(P_n, P_n')$ is asymptotically equivalent to that of $W_1(P_n, \mathbb{P})$.

Under suitable assumptions, they derive the following lower-bound:

$$n \geq \Omega\Big(\Big(\frac{1}{\tau}\Big)^d \vee \Big(\frac{\operatorname{vol}(M)}{\omega_d}\Big)^{\frac{d+2}{2\gamma}} \vee \Big(\log 1/\rho\Big)^3\Big)$$

where $\vee$ denotes maximum. A notable weakness of their approach is its susceptibility to noise. Even the presence of the smallest full-dimensional noise makes the estimator break down, as we lose the the convergence rates for the empirical measures.

Chapter 4

Complexity

We measure the hardness of statistical problems in terms of sample complexity: how many samples does one need to have any hope of high-accuracy estimation? By designing a specific estimator, one can provide only an upper bound on this quantity. In this chapter, we are interested in lower bounds, which characterize the number of samples necessary for any estimator to succeed. More specifically, we are interested in bounding the minimax risk, which is the error rate of the strongest estimator on its most challenging data distribution. We provide background on minimax theory before discussing two models in which minimax theory has been applied in ID estimation. The quick summary is as follows:

- In the noisy model of [Kol00], the minimax risk is *exponential* in the sample size, i.e. $R_n = \Theta(q^n)$ for some $q \in (0,1)$.
- In the noiseless model of [KRW16], the minimax risk is *superexponential* in the sample size, i.e. $\Omega(n^{-2n}) \leq R_n \leq O(n^{-\frac{n}{m-1}})$.

4.1 On Minimax Theory

Fix a probability space $(\Omega, \mathcal{F})$ and a set of probability measures $\mathcal{P}$ supported on this space. Let (Θ, d) be a metric space which we refer to as the **parameter space**. We call $\theta : \mathcal{P} \to \Theta$ a **statistic** and the metric $d : \Theta \times \Theta \to \mathbb{R}_{\geq 0}$ the **loss function**. An **estimator** $\widehat{\theta}_n : \mathbb{R}^n \to \Theta$ takes the observations (i.e. $X_{1:n}$) and outputs a prediction for the statistic.

Definition 24. *Let $X_{1:n} = (X_1, ..., X_n)$ be an i.i.d. sample from a probability measure $P \in \mathcal{P}$. The **minimax risk** is defined as follows:*

$$R_n = \inf_{\widehat{\theta}_n} \sup_{P \in \mathcal{P}} \mathbb{E}^P \left[d(\widehat{\theta}_n(X_{1:n}), \theta(P)) \right]$$

In this section we review the statistical hardness of manifold dimension estimation, following [KRW16]. The notion of hardness we consider is a worst-case metric known as **minimax rate**. We define this below and then explore how upper and lower bounds to this metric work.

To upper bound the minimax risk, it suffices to isolate an estimator $\widehat{\theta}_n$ and then upper bound its worst-case expected loss over all $P \in \mathcal{P}$. Lower-bounding minimax risk is generally much harder, and often requires the use of inequalities like that of Le Cam, Fano, and Assouad [LCY00; YA97]. We recall Le Cam's lemma because it is the simplest and it will be crucial for the lower bound set up by [KRW16].

Theorem 11 (Le Cam)**.** *Let $\mathcal{P}$ be a set of probability measures on $(\Omega, \mathcal{F})$ and $\theta : \mathcal{P} \to \Theta$ be a statistic. Let $S_1, S_2 \subset \Theta$ and define $\mathcal{P}_1, \mathcal{P}_2$ via preimage: $\mathcal{P}_i = \theta^{-1}(S_i)$ for $i \in \{1, 2\}$. Let $Q_i \in \delta(\mathcal{P}_i)$ be a convex combination of distributions in $\mathcal{P}_i$. Then*

$$\begin{aligned} R_n = \min_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}^P \left[d(\widehat{\theta}, \theta(P)) \right] &\geq \frac{1}{2} d(S_1, S_2) \Big(1 - \|Q_1 - Q_2\|_{TV} \Big) \\ &= \frac{1}{2} d(S_1, S_2) \int [q_1(x) \wedge q_2(x)] d\nu(x) \end{aligned}$$

where q_i is the density of Q_i with respect to underlying measure ν.

Proof, adapted from [YA97]. For any $P_1 \in \mathcal{P}_1$ and $P_2 \in \mathcal{P}_2$, we have:

$$\begin{aligned} M &:= 2 \sup_{P \in \mathcal{P}} \mathbb{E}^P [d(\widehat{\theta}, \theta(P))] \\ &\geq \mathbb{E}^{P_1} [d(\widehat{\theta}, \theta(P_1))] + \mathbb{E}^{P_2} [d(\widehat{\theta}, \theta(P_2))] \\ &= \mathbb{E}^{P_1} [d(\widehat{\theta}, S_1)] + \mathbb{E}^{P_2} [d(\widehat{\theta}, S_2))] \end{aligned}$$

The sample inequality holds replacing P_i with $Q_i \in \mathrm{conv}(\mathcal{P}_i)$.

$$\begin{aligned} M &\geq \mathbb{E}^{Q_1} [d(\widehat{\theta}, S_1)] + \mathbb{E}^{Q_2} [d(\widehat{\theta}, S_2)] \\ &= \int d(\widehat{\theta}, S_1) q_1(x) d\nu(x) + d(\widehat{\theta}, S_2) q_1 d\nu(x) \\ &\geq \int \Big(d(\widehat{\theta}, S_1) + d(\widehat{\theta}, S_2) \Big) [q_1(x) \wedge q_2(x)] \; d\nu(x) \\ \text{(triangle inequality)} \;\; &\geq \int [d(S_1, S_2)] (q_1(x) \wedge q_2(x)) \; d\nu(x) \end{aligned}$$

Substituting for M gives the desired formula. □

Since the inequality holds $\forall \widehat{\theta}$, the lower-bound applies to $\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E} \left[l(\widehat{\theta}, \theta(P)) \right]$ which makes this a handy method for lower-bounding the minimax rate.

Armed with Le Cam's lemma, we can return to the example of Bernoulli estimation and show the minimax optimality of the mean estimator.

Theorem 12. *$n = \Omega(1/\varepsilon^2)$ samples is necessary for any algorithm to $(\varepsilon, 0.3)$-learn the Bernoulli parameter estimation problem (i.e. with probability at least $1 - 0.3$, estimate the true parameter within an error of ε).*

Proof. Let $S_1 = \{(1 - \varepsilon)/2\}$ and $S_2 = \{(1 + \varepsilon)/2\}$, such that $Q_1 = \mathcal{P}_1 = \mathrm{Bern}(1/2 - \varepsilon)^{\times n}$ and $Q_2 = \mathcal{P}_2 = \mathrm{Bern}(1/2 + \varepsilon)^{\times n}$ (i.e. product distributions over n samples). Let $d(\theta_1, \theta_2) =$

$\mathbf{1}[|\theta_1 - \theta_2| \geq \varepsilon]$. Note that the left hand side of Le Cam's lemma becomes

$$\min_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{P}(|\widehat{\theta} - \theta(P)| \geq \varepsilon)$$

We will show that this quantity is lower-bounded by 0.3 for $n \ll 1/\varepsilon^2$. Therefore, under this few-sample regime, for all estimators methods, there exists a distribution such that with at least 30% probability, the estimated statistic is $\geq \varepsilon$ away from the true statistic.

Note that $d(S_1, S_2) = 1$ and the total variational distance between the distributions is upper-bounded, via Pinsker's inequality, as follows:

$$\|Q_1 - Q_2\|_{TV} \leq \sqrt{2\mathrm{KL}(Q_1||Q_2)} = \sqrt{2n\mathrm{KL}(\mathrm{Bern}(1/2+\varepsilon)||\,\mathrm{Bern}(1/2-\varepsilon))}$$

where the last equality comes from the factorization property of the Kullback-Leibler (KL) divergence. Using the fact KL divergence between Bernoullis with parameter p, q is given by $p\log(p||q) + q\log(q||p)$, we find $\mathrm{KL}(\mathrm{Bern}(1/2+\varepsilon)||\,\mathrm{Bern}(1/2-\varepsilon)) \leq 16\varepsilon^2$. So the RHS is lower-bounded by $\frac{1}{2}(1 - 4\varepsilon\sqrt{2n})$. For $n \ll 1/\varepsilon^2$, this quantity is arbitrarily close to 1/2 and indeed larger than 0.3. □

4.2 Linear Case: Spiked Covariance Model

Consider the minimax rate of the harder problem of subspace estimation. Let $\Sigma \in \mathbb{R}^{D\times D}$ be a symmetric PSD matrix. Let

$$P_{\mathcal{I}}(\Sigma) = \sum_{i \in \mathcal{I}} \lambda_i u_i u_i^T$$

be the projection onto the set of eigenvectors indexed by $\mathcal{I}$. We can define a notion of minimax risk as follows:

$$R_n = \inf_{\hat{P}} \sup_{\Sigma} \mathbb{E}_{\Sigma} \|\hat{P} - P_{\mathcal{I}}(\Sigma)\|_{HS}^2$$

where $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm. [Wah22] shows a simple bound on this value.

Lemma 3. *Let* $0 < \sigma^2 < 1$. *Suppose* $\lambda_i(\Sigma) = \begin{cases} 1+\sigma^2 & i \leq d \\ \sigma^2 & \textit{otherwise} \end{cases}$. *Then we have:*

$$R_n \geq c \min\Big(\frac{d(D-d)\sigma^2(1+\sigma^2)}{n}, d, D-d\Big)$$

So in order for $R_n < \delta$ we need $n = \Omega(dD\sigma^2/\delta)$. This shows that estimating the whole subspace well takes sample size that scales with the ambient dimension D.

4.3 Manifold Case I: Noise Deconvolution Model

Suppose you observe samples $Y_j \in \mathbb{R}^D$ generated as follows:

$$Y_j = X_j + \eta_j$$

where $X_j \sim P$ and $\eta_j \sim \mu$, each sampled i.i.d. from their respective distributions. Suppose we know μ (we think of it as the noise distribution) and we want to recover P (the signal) based on the empirical distribution of Y, call this $\widehat{Q}$. The true distribution is $Q = P \star \mu$, where $\star$ denotes the convolution operator.

Think even more generally than dimension estimation for a moment. Let $\mathcal{P}$ be the class of probability distributions in $\mathbb{R}^m$ with compact support, and let τ be any function from $\mathcal{P}$ to non-negative integers $\mathbb{Z}_+$. [Kol00] show that the best convergence rate of such an estimator that one could hope for is exponential in the sample size.

Proposition 2. *Let $|\tau(\mathcal{P})| \geq 2$. Suppose μ is absolutely continuous with uniformly bounded density, nonzero Fourier transform, and bounded KL-divergence under affine shifts. Then there exists $q \in (0,1)$ such that for all large enough n,*

$$\inf_{\widehat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{P}[\widehat{\tau}_n \neq \tau(P)] \geq q^n$$

The natural question is: can we achieve exponential convergence rate for the intrinsic dimension estimation problem, under this setup? It turns out, the answer is yes. But we must establish the following modeling assumptions.

Assumption 1. *Let $\mathcal{P} = \mathcal{P}(\Theta, C)$ be a distribution on $\mathbb{R}^D$ such that*

- $\text{supp}(P) \subset B(0,1)$
- *The Minkowski dimension* $\dim(E) \in [D]$ *(it is crucial that we only consider integer dimension; if we allow the dimension to be real-valued, as is the case for fractal sets, then at best we can expect a logarithmic rate of convergence).*
- *For $d = \dim(P)$, $|\{B \in \mathcal{N}(\varepsilon) : \text{dist}(B^+, \text{supp}(P)) \leq \varepsilon\}| \leq \Theta\varepsilon^{-d}$, where B^+ denotes the same ball with radius doubled, and $\mathcal{N}(\varepsilon)$ is any ε-cover of* $\text{supp}(P)$.
- *For $\varepsilon > 0$ and for any ball of radius ε, $P(B) \leq C\varepsilon^d$*

Key to their analysis is the idea of a **deconvolving empirical measure** $\widehat{P}_n$. Let Ψ be a symmetric Borel measurable probability measure such that $\Psi = \mathcal{K} \star \mu$, where $\mathcal{K}$ is a signed measure of bounded total variation on $\mathbb{R}^m$.

$$\widehat{P}_{n,\Psi}(A) \coloneqq \frac{1}{n}\sum_{j=1}^{n} \mathcal{K}(A - Y_j)$$

where A is a Borel measurable subset of $\mathbb{R}^m$ and $A - Y_j$ is the translate of that subset by Y_j. It happens that this deconvolving empirical measure is consistent with the measure on Ψ, i.e. $\mathbb{E}\widehat{P}_{n,\Psi}(A) = \mathbb{P}_\Psi(A)$. They use this measure to define first a sort of empirical covering number,

$$\widehat{N}_n = |\{B \in \mathcal{N}(\varepsilon) : \widehat{P}_{n,\Psi}(B) \geq 2\gamma\}|$$

and then they use this to construct an empirical estimator for the dimension,

$$\widehat{d}_n = \Big[\frac{\log \widehat{N}_n}{\log(1/\varepsilon)} + \frac{1}{2}\Big] \tag{4.1}$$

Theorem 13. *Suppose* $\varepsilon < (\Theta^{-1} \wedge (2C)^{-1})^{2/\delta(\mathcal{D})}$ *and* $\gamma < (1/2) \wedge (\varepsilon^D/(12\Theta))$*. Suppose* $\Psi(\{x : |x| \geq \varepsilon\}) \leq \gamma$*. Then there exists* $\Lambda > 0$ *and* $q \in (0, 1)$ *such that:*

$$\sup_{P \in \mathcal{P}} \mathbb{P}\Big[\widehat{d}_n \neq \dim(P)\Big] \leq \Lambda q^n$$

4.4 Manifold Case II: Noiseless Model

It is natural to consider: how much did noise hinder our ability to have fast minimax rate? [KRW16] answer this question precisely.

4.4.1 Problem Formulation

The assumed generative process is a well-behaved probability distribution supported on a d-dimensional manifold embedded in $\mathbb{R}^D$. We define the problem carefully below:

Definition 25 (ID estimation problem with i.i.d. sampling)**.** *Let* $\mathcal{M}^d_{\tau_g,\tau_l,K_I,K_v}$ *be the set of compact d-dimensional manifolds M such that:*

1. *M is suitably bounded, i.e.* $M \subset [-K_I, K_I]^D \subset \mathbb{R}^D$*.*
2. *M has global reach at least* τ_g *and local reach at least* τ_l*.*
3. *M is locally geodesically complete with respect to* τ_g*.*
4. *M is of essential volume dimension d*

Define $\mathcal{P}_{K_p}$ *to be the set of Borel probability distributions P such that:*

1. *P is supported on a d-dimensional manifold* $M \in \mathcal{M}^d_{\tau_g,\tau_l,K_I,K_v}$*.*
2. *P is absolutely continuous with respect to the restriction* vol_M *of the d-dimensional Hausdorff measure with* $\sup_{x \in M} \frac{dP}{dvol_M} \leq K_p$*.*

Given $\{x_i\}_{i \in [N]}$ *sampled independently and identically distributed from* $P \in \mathcal{P}_{K_p}$*, output an estimate for d, the dimension of the supporting manifold.*

4.4.2 Lower Bound

As illustrated by Le-Cam's lemma, we can establish a minimax lower bound if we choose two distributions $\mathcal{P}_1$,$\mathcal{P}_2$ such that:

1. There exists $Q_1 \in \text{conv}(\mathcal{P}_1)$ and $Q_2 \in \text{conv}(\mathcal{P}_2)$ with significant shared support.
2. Their statistics $\theta(P_i)$ map far apart.

These conditions capture a very intuitive criterion: we want to choose distributions that look similar (condition 2) but have different statistics (condition 1). The parameter estimation is as hard as the decision problem of distinguishing this particular pair of cases.

For the ID estimation problem, [KRW16] make use of the fact that a low-dimensional manifold with high curvature can look very similar to a high-dimensional manifold with relatively low curvature. Roughly speaking, they define the following pair of distributions:

$$\mathcal{P}_1 = \{\text{distributions supported on a 1-dimensional space-filling-curve manifold}\}$$

$$\mathcal{P}_2 = \{\text{uniform distributions on } [-K_I, K_I]^{d_2}\}$$

From here, we construct a specific set $T \subset I^n$ such that whenever $X = X_{1:n} \in T$, it is difficult to distinguish whether $X \in \mathcal{P}_1$ or $\mathcal{P}_2$. First we describe T.

Lemma 4. *Fix $\tau_l \in (0, \infty]$, $K_I \in [1, \infty)$, $d_1, d_2 \in \mathbb{N}$ with $1 \leq d_1 \leq d_2$. Suppose $\tau_l \leq K_I$. Then there exist distinct $T_1, ..., T_n \subset [-K_I, K_I]^{d_2}$ such that:*

- *For each T_i, there exists an isometry Φ_i such that:*

  $$T_i = \Phi_i\Big([-K_I, K_I]^{d_1-1} \times [0, a] \times B_{\mathbb{R}^{d_2-d_1}}(0, w)\Big)$$

  where a, w are appropriate constants.

- *There exists $\mathcal{M} : (B_{\mathbb{R}^{d_2-d_1}(0,w)})^n \mapsto \mathcal{M}_{\tau_g,\tau_l,K_I,K_v}$ injective such that for each $y_i \in B_{\mathbb{R}^{d_2-d_1}}(0, w)$ and $1 \leq i \leq n$,*

  $$\mathcal{M}(y_1, ..., y_n) \cap T_i = \Phi_i([-K_I, K_I]^{d_1-1} \times [0, a] \times \{y_i\})$$

  In other words, for any choice of $x_i \in T_i$ for all $i \in [n]$, $\mathcal{M}(\{\Pi^{-1}_{d_1+1:d_2}\Phi_i^{-1}(x_i)\})$ passes through $x_1, ..., x_n$ (where $\Pi_{a:b}$ denotes projection onto the coordinates a through b).

The crucial idea here is that (1) the T_i are arranged in a zigzag fashion which makes it sort of space-filling, and (2) we can always find a manifold satisfying the regularity constraints that passes through all the samples from T_i. The next step is to show that there exists a convex combination of distributions supported on these space-filling curves whose probability density is not much different from the uniform distribution.

Lemma 5. *Let $T = \{\prod_{i=1}^n T_{\sigma(i)} : \sigma \in S_n\}$. Let Q_2 be the uniform on $[-K_I, K_I]^{d_2}$. Let $\mathcal{P}_1$ be as stated earlier. There exists $Q_1 \in \operatorname{conv}(\mathcal{P}_1)$ such that:*

$$Q_1\Big(\prod_{i=1}^n B(x_i, r)\Big) \geq 2^{-n} \cdot Q_2\Big(\prod_{i=1}^n B(x_i, r)\Big)$$

This demonstrates more generally that $q_1 \geq C q_2$ for $C > 0$ a constant. Hence by Le Cam's we find, for any estimator $\widehat{d}_n$,

$$\begin{aligned}
\sup_{P \in \mathcal{Q}} \mathbb{E}^{P^{(n)}} |\widehat{d}_n - d(P)| &\geq \frac{d_2 - d_1}{2} \int [q_1 \wedge q_2] d\lambda(x) \\
&\geq (d_2 - d_1) \int_T [q_1 \wedge q_2] d\lambda(x) \\
&\geq C(d_2 - d_1) \operatorname{vol}(T)
\end{aligned}$$

where the last step follows from the fact that q_2 is the uniform distribution.

Chapter 5

Conclusion

This thesis has explored the manifold intrinsic dimension estimation problem, a fundamental challenge in data science and machine learning. We reviewed a variety of algorithms, each with their own strengths and weaknesses. Our analysis highlighted the mathematical and statistical intricacies involved in balancing sample complexity, computational efficiency, and robustness to noise.

In the linear case, we showed how principal component analysis (PCA) provides a solid foundation for dimension estimation, with well-established theoretical guarantees. For nonlinear manifolds, local methods such as nearest-neighbor techniques [LB04] and multi-scale local PCA [Lit+09] offer practical approaches, albeit with sensitivity to neighborhood selection. Global methods, including Wasserstein metric convergence [WB19] and scaling of the traveling salesman cost [KRW16], provide robust alternatives, albeit at higher computational cost.

The complexity results discussed underscore the inherent challenges in dimension estimation, particularly in the presence of noise. The sample complexity lower bounds discussed in Section 4 highlight the need for continued research into more efficient algorithms that leverage problem-specific structure; for instance, problems where the data or noise distributions have particular independence structures.

One particularly promising direction of future work is exploring the integration of active learning techniques into ID estimation, which could significantly reduce sample requirements while maintaining accuracy. This is a completely different learning model where the established sample complexity lower bounds do not apply.

Another interesting research direction is on the interpretability side of dimension estimation. Say we have a perfect dimension estimator for single cell RNA data (i.e. gene counts for a collection of cells in the human body). What do these dimensions mean? Do they have different meanings depending on where we are on the manifold? This is an exciting opportunity for mathematical modeling to interface with domain knowledge.

Bibliography

[Aam+19] Eddie Aamari et al. “Estimating the reach of a manifold”. In: (2019).

[Apa+20] Luis Aparicio et al. “A random matrix theory approach to denoise single-cell data”. In: *Patterns* 1.3 (2020).

[Ber+00] Mira Bernstein et al. *Graph approximations to geodesics on embedded manifolds*. Tech. rep. Citeseer, 2000.

[Blo+22] Adam Block et al. “Intrinsic dimension estimation using Wasserstein distance”. In: *The Journal of Machine Learning Research* 23.1 (2022), pp. 14124–14160.

[BN01] Mikhail Belkin and Partha Niyogi. “Laplacian eigenmaps and spectral techniques for embedding and clustering”. In: *Advances in neural information processing systems* 14 (2001).

[CH03] Jose Costa and Alfred Hero. “Manifold learning with geodesic minimal spanning trees”. In: *arXiv preprint cs/0307038* (2003).

[CH06] Jose A Costa and Alfred O Hero. “Determining intrinsic dimension and entropy of high-dimensional shape spaces”. In: *Statistics and analysis of shapes* (2006), pp. 231–252.

[CS16] Francesco Camastra and Antonino Staiano. “Intrinsic dimension estimation: Advances and open problems”. In: *Information Sciences* 328 (2016), pp. 26–41.

[DF08] Sanjoy Dasgupta and Yoav Freund. “Random projection trees and low dimensional manifolds”. In: *Proceedings of the fortieth annual ACM symposium on Theory of computing*. 2008, pp. 537–546.

[DG03] Sanjoy Dasgupta and Anupam Gupta. “An elementary proof of a theorem of Johnson and Lindenstrauss”. In: *Random Structures & Algorithms* 22.1 (2003), pp. 60–65.

[Dud69] Richard Mansfield Dudley. “The speed of mean Glivenko-Cantelli convergence”. In: *The Annals of Mathematical Statistics* 40.1 (1969), pp. 40–50.

[EJS05] Thomas Erlebach, Klaus Jansen, and Eike Seidel. “Polynomial-time approximation schemes for geometric intersection graphs”. In: *SIAM Journal on Computing* 34.6 (2005), pp. 1302–1323.

[FMN16] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. “Testing the manifold hypothesis”. In: *Journal of the American Mathematical Society* 29.4 (2016), pp. 983–1049.

[FSA07] Amir Massoud Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. "Manifold-adaptive dimension estimation". In: *Proceedings of the 24th international conference on Machine learning*. 2007, pp. 265–272.

[GP83] Peter Grassberger and Itamar Procaccia. "Measuring the strangeness of strange attractors". In: *Physica D: nonlinear phenomena* 9.1-2 (1983), pp. 189–208.

[HA05] Matthias Hein and Jean-Yves Audibert. "Intrinsic dimensionality estimation of submanifolds in Rd". In: *Proceedings of the 22nd international conference on Machine learning*. 2005, pp. 289–296.

[Hsu10] Daniel Joseph Hsu. "Algorithms for active learning". PhD thesis. UC San Diego, 2010.

[Kég02] Balázs Kégl. "Intrinsic dimension estimation using packing numbers". In: *Advances in neural information processing systems* 15 (2002).

[Kol00] Vladimir I Koltchinskii. "Empirical geometry of multivariate data: a deconvolution approach". In: *Annals of statistics* (2000), pp. 591–629.

[KRW16] Jisu Kim, Alessandro Rinaldo, and Larry Wasserman. "Minimax rates for estimating the dimension of a manifold". In: *arXiv preprint arXiv:1605.01011* (2016).

[LB04] Elizaveta Levina and Peter Bickel. "Maximum likelihood estimation of intrinsic dimension". In: *Advances in neural information processing systems* 17 (2004).

[LCY00] Lucien Marie Le Cam and Grace Lo Yang. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2000.

[Lee12] John M Lee. *Smooth manifolds*. Springer, 2012.

[Lit+09] Anna V Little et al. "Estimation of intrinsic dimensionality of samples from noisy low-dimensional manifolds in high dimensions with multiscale SVD". In: *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*. IEEE. 2009, pp. 85–88.

[MNW24] Tudor Manole and Jonathan Niles-Weed. "Sharp convergence rates for empirical optimal transport with smooth costs". In: *The Annals of Applied Probability* 34.1B (2024), pp. 1108–1135.

[Nas56] John Nash. "The imbedding problem for Riemannian manifolds". In: *Annals of mathematics* 63.1 (1956), pp. 20–63.

[SR00] Lawrence K Saul and Sam T Roweis. "An introduction to locally linear embedding". In: *unpublished. Available at: http://www. cs. toronto. edu/˜ roweis/lle/publications. html* (2000).

[TB99] Michael E Tipping and Christopher M Bishop. "Probabilistic principal component analysis". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61.3 (1999), pp. 611–622.

[TSL00] Joshua B Tenenbaum, Vin de Silva, and John C Langford. "A global geometric framework for nonlinear dimensionality reduction". In: *science* 290.5500 (2000), pp. 2319–2323.

[Tu11] Loring W Tu. “Manifolds”. In: *An Introduction to Manifolds.* Springer, 2011, pp. 47–83.

[Val84] Leslie G Valiant. “A theory of the learnable”. In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142.

[Ver10] Roman Vershynin. “Introduction to the non-asymptotic analysis of random matrices”. In: *arXiv preprint arXiv:1011.3027* (2010).

[Ver11] Nakul Verma. “A note on random projections for preserving paths on a manifold”. In: (2011).

[Ver12] Nakul Verma. “Distance preserving embeddings for general n-dimensional manifolds”. In: *Conference on Learning Theory.* JMLR Workshop and Conference Proceedings. 2012, pp. 32–1.

[Vil+09] Cédric Villani et al. *Optimal transport: old and new.* Vol. 338. Springer, 2009.

[Wah22] Martin Wahl. “Lower bounds for invariant statistical models with applications to principal component analysis”. In: *Annales de l’Institut Henri Poincare (B) Probabilites et statistiques.* Vol. 58. 3. Institut Henri Poincaré. 2022, pp. 1565–1589.

[WB19] Jonathan Weed and Francis Bach. “Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance”. In: (2019).

[Weg21] Sven-Ake Wegner. “Lecture notes on high-dimensional data”. In: *arXiv preprint arXiv:2101.05841* (2021).

[Wei14] Shmuel Weinberger. “The complexity of some topological inference problems”. In: *Foundations of Computational Mathematics* 14 (2014), pp. 1277–1285.

[WS06] Kilian Q Weinberger and Lawrence K Saul. “An introduction to nonlinear dimensionality reduction by maximum variance unfolding”. In: *AAAI.* Vol. 6. 2006, pp. 1683–1686.

[YA97] Bin Yu and Fano Assouad. “Le Cam”. In: *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics, Springer-Verlag, New York* (1997), pp. 423–435.

[Zam06] Leonardo Zambito. “The traveling salesman problem: a comprehensive survey”. In: *Project for CSE* 4080 (2006).