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 i.i.d. 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 thesis—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) is bounded, i.e. $\tau(\mathcal{M}) < T$*
- *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 i.i.d. samples from μ and returns an estimator $\hat{d}_n \in [D]$ such that $\hat{d}_n = d$ with high probability.

This thesis 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 thesis 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

We begin with a brief review of dimensionality reduction techniques: starting with linear methods and then easing into nonlinear methods. Throughout this discussion, let $X = [x_1, ..., x_n] \in \mathbb{R}^{D \times n}$ denote a dataset. Let $\mathbb{E}_{i \sim [n]} f(x_i) = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$ (i.e. let the expectation be taken over a uniform measure on the dataset). Let D denote the dimension of the ambient space and d denote the dimension of the manifold or the embedding output by some manifold learning algorithm. We assume throughout that $d \ll D$.

2.1 Linear Dimension Reduction

Linear dimension reduction is about finding a linear subspace that minimizes some sort of reconstruction error or distortion over data points. PCA is a canonical example of minimizing average-case distortion, while JL is a canonical example of dimension reduction posed in terms of worst-case distortion

Best-Fitting Subspace and PCA 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. We can formulate this 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 U 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:

$$\arg\min_{A \in \mathcal{A}_{D \times d}} \sum_{i=1}^{n} \|AA^T x_i - x_i\|_2^2 = \arg\min_{A \in \mathcal{A}_{D \times d}} \|AA^T X - X\|_2$$

It turns out that there are efficient algorithms for computing such an orthoprojector , which are effectively based on eigendecomposition calculations. 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 the the projected $AA^T X$. 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 $AA^T u_i = u_i \cdot \mathbf{1}(i \leq d)$. 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.* This is an important fact that will continue to show up in our discussion of manifold learning. The following correspondence will also be useful, e.g. when discussing spectral clustering.

Theorem 1. *Let $A \in \mathbb{R}^{D \times d}$ and $X \in \mathbb{R}^{D \times n}$.*

$$\operatorname*{arg\,min}_{A^T A = I} \|AA^T X - X\| = \operatorname*{arg\,max}_{A^T A = I} \ \text{tr} \left(A^T \left(XX^T \right) A \right)$$

Proof. It suffices to show that the solution to the RHS is given by the A whose columns are the top d eigenvectors of XX^T. Let a_i be the columns of A, so $\{a_1, ..., a_d\}$ is orthonormal. Note that XX^T is symmetric and PSD. Let $\{u_1, ..., u_D\}$ denote its set of eigenvectors in decreasing order of eigenvalue $\lambda_1 \geq ... \geq \lambda_D \geq 0$.

$$\text{tr} \left(A^T \left(XX^T \right) A \right) = \sum_{i=1}^{d} a_i^T (XX^T) a_i = \sum_{i=1}^{d} a_i^T \Big(\sum_{j=1}^{D} \lambda_i u_i u_i^T \Big) a_i$$

[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 $B \in \mathbb{R}^{n \times m}$.

Let $a_i = \sum_{j=1}^D c_{ij} u_j$. Then we have:

$$\begin{aligned}\sum_{i=1}^d \Big(\sum_{j=1}^D c_j u_j^T\Big)\Big(\sum_{k=1}^D \lambda_k u_k u_k^T\Big)\Big(\sum_{l=1}^D c_{il} u_l\Big) &= \sum_{i=1}^d \sum_{j=1}^D \sum_{k=1}^D \sum_{l=1}^D \lambda_k c_{ij} c_{il} (u_j^T u_k)(u_k^T u_l) \\ &= \sum_{i=1}^d \sum_{j=1}^D c_{ij}^2 \lambda_j\end{aligned}$$

Recall, since $\|a_i\| = 1$, that $\sum_{j=1}^D c_{ij}^2 = 1$. So naturally, we want to place all of this mass in λ_1, the largest eigenvalue. But we also need to have $\{a_i\}$ orthogonal. So the optimal allocation is to let $c_{11} = c_{22} = ... = c_{dd} = 1$ and all else zero. Hence, $a_i = u_i$ for $i \in [d]$. $\square$

One can consider the more general subspace approximation problem Subspace$(k,p) = \min_{U \in \mathcal{U}_k} \sum_{i=1}^n \|UU^T x_i - x_i\|_2^p$ where one adds up the pth powers of the Euclidean norm. Clearly, for $p = 2$, exact optimization can be done in polynomial time. For all $p > 2$, [DTV11] was able to show a constant factor approximation scheme which is nearly tight under the Unique Games Conjecture (i.e. it is impossible to construct a better approximation in polynomial time).

Random Projection Perhaps the simplest conceivable method for dimension reduction is random projection, i.e. sample a random matrix $R \in \mathbb{R}^{d \times D}$ such that $\{Rx_i\}$ resembles $\{x_i\}$ with high probability. What should be our criterion for resemblance? With this introduction of randomness, it is no longer reasonable to enforce that Rx_i and x_i are close together. One can observe however, that if you provide enough dimensions for the projection, the interpoint distances do not change very much. Indeed, for $O(\log n)$ dimensions, one can guarantee the existence of a map such that the worst-case distortion factor is always within 1 ± ϵ.

Theorem 2 (Johnson-Lindenstrauss Lemma). *For any $\{x_i\}_{i \in [n]} \subset \mathbb{R}^D$ there exists $R \in \mathbb{R}^{d \times D}$ with $d \geq O(\log(n)/\epsilon)$ such that:*

$$\frac{\|Rx_i - Rx_j\|}{\|x_i - x_j\|} \in (1-\epsilon, 1+\epsilon) \qquad \forall i \neq j \in [n]$$

There are numerous methods for producing such a matrix R. Perhaps the simplest, as outlined in [DG03], is taking iid Gaussian entries in the matrix. There are many variants, including sparse JL transforms [DKS10; KN14]. There is also a rich theory of random projection on manifolds, as opposed to point cloud data.

2.2 Nonlinear Dimension Reduction

In their review of manifold learning, [MZ23] distinguishes between "one-shot" embedding algorithms as opposed to "cost-minimization" embedding algorithms. This should not, however, distract from the fact that the major "one-shot" algorithms minimize an cost functions. It just so happens that the minimization of such objectives consistently reduces to eigendecomposition.

In all of the following problems, one should imagine being given a finite metric $D \in \mathbb{R}^{n\times n}$, possibly derived from some dataset $X = [x_1, ..., x_n] \in \mathbb{R}^{D\times n}$ and wanting to output some low-dimensional collection of points $Y = [y_1, ..., y_n] \in \mathbb{R}^{d\times n}$.

Multidimensional Scaling (MDS) The classical formulation of the multidimensional scaling problem is as follows: say you are given the interpoint distances between n points in $\mathbb{R}^d$, packaged in an $n \times n$ matrix D. Given such a *Euclidean embeddable finite metric space*, how do you recover a corresponding set of points, i.e. $(x_1, ..., x_n)$ such that $D_{ij} = \|x_i - x_j\|^2$ for all $i, j \in [n]$? This turns out to a a straightforward task, based on the following fact:

Theorem 3. *Let $D \in \mathbb{R}^{n\times n}$. Then D is Euclidean embeddable if and only if the Gram matrix $B = -\frac{1}{2}HDH$ is positive semidefinite. Furthermore, if D is Euclidean embeddable, then $B = X^T X$ where $X = [x_1, ..., x_n] \in \mathbb{R}^{d\times n}$ is a point set such that $\|x_i - x_j\| = d_{ij}$.*

The key idea here is that $\|x_i - x_j\|^2 = \langle x_i, x_i\rangle + \langle x_j, x_j\rangle - 2\langle x_i, x_j\rangle$. Written more suggestively, we have:

$$-\frac{1}{2}(d_{ij}^2 - x_i^2 - x_j^2) = \langle x_i, x_j\rangle$$

The right hand side is clearly the (i, j) entry of $X^T X$. The work of the proof comes down to showing the left-hand side is the corresponding entry of the double-centered matrix $-\frac{1}{2}HDH$. Hence, if we have a Euclidean embeddable squared interpoint distance matrix, we can retrieve the embedding X as follows:

- Given D: Compute $B = -\frac{1}{2}HDH$. Compute its spectral decomposition $B = U^T \Lambda U$. Let $[\Lambda_+]_{ij} = \max\{\Lambda_{ij}, 0\}$. Let $X = U\Lambda_+$. Return $[X]_{n\times d}$ (i.e. take first d columns).

We can view the efficacy of this algorithm once again from the lens of Eckart-Young. Rewrite the optimization in matrix form:

$$\min_{x_1,...,x_n\in\mathbb{R}^d} \sum_{i,j} \left(D_{ij}^2 - \|x_i - x_j\|^2\right)^2 = \min_{X\in\mathbb{R}^{d\times n}} \left\| -\frac{1}{2}HDH - X^T X\right\|_F$$

The best rank-d approximation to $B = -\frac{1}{2}HDH$ is given by its rank-d singular value decomposition $\hat{B}_d$. So we want to find $X^T X = \hat{B}_d$. First of all, this is only possible if B and hence $\hat{B}_d$ was PSD, since $X^T X$ is PSD. But if it is, then the aforementioned algorithm recovers precisely the right matrix.

One can generalize MDS in a number of ways: in particular, exploring different cost functions. For instance, consider the Kamada-Kawai MDS objective (in the matrix formulation, note that the division of matrices is done elementwise).

$$\min_{x_1,...,x_n\in\mathbb{R}^d} \sum_{i,j} \left(1 - \frac{\|x_i - x_j\|}{D_{ij}}\right)^2 = \min_{X\in\mathbb{R}^{d\times n}} \left\| \mathbf{1}\mathbf{1}^T - \frac{X^T X}{D}\right\|_F$$

The Kamada-Kawai objective is often used in force-based graph drawing. It is an active study of research in geometric optimization. [Dem+21] recently showed that this is NP-hard to minimize exactly but admits a randomized poly-time approximation scheme, which was later improved by [Bak+23].

Open Problem 1. *Does stochastic gradient descent with enough random restarts have provable guarantees for minimizing the (highly non-convex) Kamada-Kawai objective? Posed by [Dem+21]; the idea being that, if the greedy discretized optimization method has such a guarantee, so too should the greedy continuous optimization.*

Isomap Isomap, developed by [TSL00], is a portmanteau for *isometric mapping* (the reason for this name will become apparent). Isomap is effectively a clever application of MDS, where the key insight comes down to preprocessing. Suppose we have data lying on a manifold of known dimension. Due to the locally Euclidean nature of manifolds, the Euclidean distance is a good approximation for the actual geodesic distance metric on the manifold itself. This approximation falls apart for larger distances, but there is a remedy: if you construct a nearest-neighbor graph and compute the shortest path (e.g. with Djikstra's or Floyd's algorithm) along neighbors, then the sums of these small Euclidean distances is a better approximation for the geodesic distance along the manifold. Then, you could plug in these approximate geodesic distances into MDS and use the corresponding embedding.

$$Y^*_{\text{isomap}} = \min_{Y \in \mathbb{R}^{d \times n}} \left\| -\frac{1}{2} H [D^{(X)}_{\text{geodesic}}] H - Y^T Y \right\|$$

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

Later in this chapter we will discuss a consistency theorem for Isomap, i.e. a result which shows that this shortest-path graph metric approximates the manifold geodesic distance under suitable conditions. Now, in order for Isomap to be truly optimal, it should be the case that the geodesic distance is itself Euclidean (otherwise, by Theorem 2, our optimization is hopeless). So this algorithm only makes sense for a Euclidean manifold that is *isometrically* embedded, i.e. its geodesic distances are still Euclidean (hence the name IsoMap). Indeed, we find this justification in the following consistency result, which demonstrates how fast the graph shortest path metric converges to the true manifold geodesic metric.

Theorem 4 (Isomap, informal, see [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 $\epsilon \in (0,1)$. If M is geodesically convex and the graph G and the dataset $\{x_i\}$ satisfy suitable conditions (which get stricter for smaller ϵ), then:*

$$(1-\epsilon) d_M(x,y) \le d_G(x,y) \le (1+\epsilon) d_M(x,y)$$

Laplacian Eigenmaps (LE) Developed by [BN01], Laplacian Eigenmaps (also called Diffusion Maps) can be posed as solving the following very simple objective function, where W_{ij} is some similarity measure perhaps derived from the original interpoint distances, e.g. the heat kernel $W_{ij} = \exp(-D^2_{ij})$. We impose the condition on the right to avoid a trivial embedding (e.g. placing all x on the same point, so the interpoint distances are all zero).

$$\min_{x_1,\ldots,x_n} \sum_{i,j} W_{ij} \|x_i - x_j\|^2 \qquad XX^T = I$$

The key mathematical insight for the algorithm lies in the following fundamental fact about the graph Laplacian, which is a sort of second-order differential operator on a graph.

Lemma 1. *Let $L = D - W$ be the graph Laplacian of a symmetric $n \times n$ matrix W, where D is a diagonal matrix such that $D_{ii} = \sum_j W_{ij}$. Then:*

$$x^T L x = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(x_i - x_j)^2$$

Proof. $x^T L x = x^T D x - x^T W x = \sum_{i=1}^n D_{ii} x_i^2 - \sum_{i,j} x_i x_j W_{ij} = \sum_{i,j}[W_{ij} x_i^2 - 2 x_i x_j W_{ij}] = \frac{1}{2} \sum_{i,j} W_{ij}(x_i^2 + x_j^2 - 2 x_i x_j) = \frac{1}{2} \sum_{i,j} W_{ij}(x_i - x_j)^2$. □

In light of this lemma, the optimization becomes:

$$\min_{x_1,\ldots,x_n} \sum_{i=1}^{n} x_i^T L x_i = \min_{x_1,\ldots,x_n} \operatorname{tr}(X^T L X) \qquad X^T X = I$$

By Theorem 1, this is an eigenvalue problem. The optimal X is given by a matrix whose columns are, effectively, the bottom eigenvectors of L. The catch is that the bottom eigenvector of L, always given by the ones vector, has eigenvalue zero and thus would break our constraint.

Local Linear Embedding (LLE) A crucial feature of manifolds is the tangent space: in some sense, the best linear approximation to the manifold at a point. The idea of locally linear embedding is to learn the manifold via the tangent space. This is accomplished in a two-step process: (1) to learn the local linear structure of the high-dimensional points X and then impose that the low-dimensional points Y respect that local structure (2). Given $X \in \mathbb{R}^{D \times n}$ and a neighborhood size k, we have:

$$(1) \quad W^* = \operatorname*{arg\,min}_{W \in \mathbb{R}^{n \times n}} \sum_{i=1}^{n} \Big\| x_i - \sum_{j \in N(i)} W_{ij} x_j \Big\|^2 \qquad \sum_i W_{ij} = 1 \text{ and } \; j \notin N(i) \implies W_{ij} = 0$$

$$(2) \quad Y^* = \operatorname*{arg\,min}_{Y \in \mathbb{R}^{d \times n}} \sum_{i=1}^{n} \Big\| y_i - \sum_{j \in N(i)} W_{ij}^* y_j \Big\|^2 \qquad Y Y^T = I_d$$

The first problem can be rewritten more suggestively using the following notation: let N_i be the $D \times k$ neighbor matrix, W_i be the corresponding $k \times 1$ weight vector for the neighbors of x_i, and let e be the $k \times 1$ ones vector.

$$(1) \quad W_i^* = \operatorname*{arg\,min}_{W_i} \; W_i^T (X_i e^T - N_i)^T (X_i e^T - N_i) W_i \qquad e^T W_i = 1$$

This can be solved using Lagrange multipliers (i.e. adding in the constraint with a variable λ, differentiating with respect to W_i, setting to zero). The result is:

$$W_i^* = (\lambda/2) \Big[(X_i e^T - N_i)^T (X_i e^T - N_i) \Big]^{-1} e$$

Method	Kernel	Note
PCA	X^TX	$X^TX =$ covariance matrix
Classical MDS	$-\frac{1}{2}HD_{\text{Euclidean}}H$	
Isomap	$-\frac{1}{2}HD_{\text{geodesic}}H$	$D_{\text{geodesic}} =$ shortest path distance.
LLE	$\lambda_{max}I - (I - W^*)(I - W^*)^T$	$W_i^* = (X_ie^T - N_i)^T(X_ie^T - N_i)$
LE	$\lambda_{max}I - L$	$L =$ graph Laplacian of X

Table 2.1: Canonical manifold learning techniques and their corresponding kernels, when viewed as special cases of kernel PCA. Note that in LLE and LE, we exclude the top eigenvectors because they output trivial solutions.

where λ is chosen such that $\sum_j W_{ij}^* = 1$ for all i.

The second problem, meanwhile, reduces as follows (where we reinterpret W^* as lying in $\mathbb{R}^{n\times n}$, with neighbors matching appropriately):

$$(2) \quad Y^* = \arg\min_Y \operatorname{tr}\left(Y(I_n - W^*)(I_n - W^*)^T Y^T\right) \qquad Y^TY = I$$

The solution (in terms of the components of the Y vectors) is given by the bottom d eigenvectors of $(I - W^*)(I - W^*)^T$.

Reduction to Kernel PCA Each of the aforementioned manifold learning algorithms reduce to an eigenvalue problem which looks similar to PCA. Indeed, we can make this connection precise: the manifold learning algorithms are precisely PCA with a kernel that looks to preserve local geometry.

The general form of kernel PCA is as follows: given some kernel $K \in \mathbb{R}^{n\times n}$ capturing some notion of dot products for the original data, i.e. $K_{i,j} = K(x_i, x_j)$, solve:

$$\arg\min_{Y\in\mathbb{R}^{d\times n}} \|K - Y^TY\|$$

The generalization of each technique as kernel PCA is summarized in Table 2.1. In the spirit of unsupervised learning, it is natural to ask: how might we optimize our choice of kernel? It turns out there is a nice way to approach this using semidefinite programming. The idea of [WS06] is to learn a kernel which spreads out the data as much as possible, while preserving local distances. This is captured by maximizing the trace:

$$K^* = \arg\max_{K\succcurlyeq 0} \operatorname{tr}(K) \qquad N_{ij}(K_{ii} + K_{jj} - 2K_{ij} - \|x_i - x_j\|^2) = 0 \qquad \sum_{i,j} K_{ij} = 0$$

where $N_{ij} = 1$ if (i, j) are considered neighbors in X, and zero otherwise. Note that the $\sum_{i,j} K_{ij} = 0$ condition is there to ensure that K represents a proper covariance matrix, i.e. $K = Y^TY$ and $\mathbb{E}(Y) = 0$. This method is known as maximum variance unfolding (MVU) or semidefinite embedding.

2.3 Basics of Manifolds

One can formulate manifold learning methods without too much mathematical grounding as to what a manifold actually is. However, in order to prove consistency results about these methods, it becomes crucial to work on a more refined level of abstraction. We review some of the basic ideas and definitions below .

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 is *locally* similar to Euclidean spac.

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) $\phi \in \tau$, (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 ($\tau = \{\phi, X\}$) 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.

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 $\mathbb{R}^n \to \mathbb{R}^n$):*

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

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 $\phi^{-1} \circ F \circ \psi : \mathbb{R}^m \mapsto \mathbb{R}^n$ is a smooth map.*

The graph of a function, say the curve $\{(x, e^x)\}_{x\in\mathbb{R}}$ 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 $g \in C^\infty_{f(p)}(N)$, 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 $\mathsf{Length}(\gamma) = \int_0^1 \mathsf{g}_{\gamma(\mathsf{t})}(\gamma'(\mathsf{t}), \gamma'(\mathsf{t}))\mathsf{dt}$*, 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).

2.3.1 Notions of Regularity

Working with general manifolds can be extremely difficult. In particular, they can have high curvature, nearly self-intersecting themselves. This can really mess with intrinsic dimension estimators, and make them severely overestimate the dimension of high-dimensional data. 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.*

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

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.*

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

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:

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

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* $\mathrm{vol}(M) = \int_M \omega$.

2.4 Big Results in Manifold Learning

Now that we have a richer sense of some of the mathematical tools at play in the study of manifolds, we are in a better position to formulate and discuss the major algorithmic results and goals of manifold learning. We focus on two results: one, an algorithmic realization of the Nash embedding theorem, and the other, an algorithm for testing whether data actually satisfies the manifold hypothesis.

2.4.1 Euclidean Embedding Algorithm

Though we often imagine manifolds as floating in an ambient Euclidean space, this need not be the case. A priori, they are topological spaces that are only *locally Euclidean.* Nonetheless, there are structure-preserving maps (i.e. smooth embeddings) one can construct to place a manifold in Euclidean space. We make this defintion precise below, before stating the two main theorems to this effect.

Definition 14. *A smooth map $f : N \to M$ is an embedding if*

- *(1) it is a one-to-one immersion, and*
- *(2) the image $f(N)$ with the subspace topology is homeomorphic to N under f.*

Definition 15. *A smooth map between Riemannian manifolds $F : (M, g) \to (N, g')$ is called **metric-preserving** if for all $p \in N$ and $u, v \in T_pN$, $\langle u, v\rangle_p = \langle F_*u, F_*v\rangle_{F(p)}$. An **isometry** is a metric-preserving diffeomorphism (i.e. F is smooth and so is F^{-1}).*

There are two famous theorems ensuring that we can always find embeddings or isometric embeddings into Euclidean space given enough structure on our manifold.

Theorem 5 (Whitney Embedding Theorem)**.** *Any smooth real m-dimensional manifold can be smoothly embedded in $\mathbb{R}^{2m}$.*

Theorem 6 (Nash Embedding Theorem)**.** *A compact m-dimensional Riemannian manifold (M, g) can be isometrically C^1 embedded in Euclidean space of dimension $2n + 1$ and C^∞ embedded in dimension $O(n^2)$.*

These theorems are in some sense the models for all manifold learning algorithms. However, most of the aforementioned methods do not have such guarantees. One approach to achieving reasonable guarantees has been to emulate some of the proof techniques involved in the theorems.

Theorem 7 (Approximate Nash Embedding Algorithm, informal, see [Ver12])**.** *Given a sufficiently tight finite sample X from a C_M-regular n-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:*

- *$\mathcal{A}$ is a $(1 \pm \epsilon)$-isometric embedding of M into $\mathbb{R}^d$.*
- *$d = \Omega(n \log(C_M/\tau))$ (assuming $d \leq D$).*

In particular, X must be a α-bounded (ρ, d) cover (see original paper, Definition 3).

2.4.2 Testing the Manifold Hypothesis

How can we tell, a priori, that the manifold hypothesis is true? [NM10] treat this as a real hypothesis-testing problem, and formulate as follows.

Let $\mathcal{G}(d, V, \tau)$ be the family of d-dimensional $\mathcal{C}^2$-submanifolds in the unit ball of $\mathbb{R}^D$ with volume $\leq V$ and reach $\geq \tau$. Given i.i.d. samples from $\mathcal{P}$, the estimator determines with probability $\geq 1-\delta$, under the promise that one of the following must be true, whether:

- There exists $\mathcal{M} \in \mathcal{G}(d, CV, \tau/C)$ such that $\mathcal{L}(\mathcal{M}, \mathcal{P}) \leq C\epsilon$.
- There exists no $\mathcal{M} \in \mathcal{G}(d, V/C, C\tau)$ such that $\mathcal{L}(\mathcal{M}, \mathcal{P}) \leq \epsilon/C$.

where C is a universal constant, and our measure of fitting the manifold is:

$$\mathcal{L}(\mathcal{M}, \mathcal{P}) = \int d(x, \mathcal{M})^2 d\mathcal{P}(x)$$

In the run of the algorithm, this loss measure is approximated by the empirical loss, $L_{\text{emp}}(\mathcal{M}) = \frac{1}{s}\sum_{i=1}^{s} d(x_i, \mathcal{M})^2$, which for enough samples is close enough (by standard empirical process methods).

The key insight of the method used here—which will not find a full treatment in this thesis—is translating the optimization of $\mathcal{L}(\mathcal{M}, \mathcal{P})$ over a family of manifolds to an optimization over sections of a disc bundle. The benefit of the latter space is that it is parameterized and can be approached via a convex program. We give a brief background on the algorithm.

Definition 16. *Given $x_1, ..., x_n$ sampled from $\mathcal{P}$, we say $\mathcal{M} \in \mathcal{G}(d, V, \tau)$ is an ϵ-optimal interpolant if for some constant C (depending only on dimension),*

$$L_{emp}(\mathcal{M}) \leq \epsilon + \inf_{\mathcal{M}' \in \mathcal{G}(d, V/C, C\tau)} L_{emp}(\mathcal{M}')$$

Definition 17. *For $\mathcal{D}$ an open set of $\mathbb{R}^D$ and $\mathcal{M}$ an embedded submanifold of $\mathcal{D}$ of dimension d, let $\pi : \mathcal{D} \mapsto \mathcal{M}$ be a C^k map such that for all $z \in \mathcal{M}$, $\pi(z) = z$ and $\pi^{-1}(z)$ is isometric to a Euclidean disc of dimension $n - d$. We call π a **disc bundle**. We call $s : \mathcal{M} \mapsto \mathcal{D}$ a **section** of $\mathcal{D}$ if for all $z \in \mathcal{M}$, $s(z) \in \pi^{-1}(z)$ and for some $\hat{\tau}, \hat{V}$, $s(\mathcal{M}) \in \mathcal{G}(d, \hat{\tau}, \hat{V})$.*

With this, we can specify the algorithm.

- Construct a set of disc bundles $\overline{\mathcal{D}}^{\text{norm}}$ over manifolds in $\mathcal{G}(d, CV, \tau/C)$ rich enough that every ϵ-interpolant is a section of some member of $\overline{\mathcal{D}}^{\text{norm}}$.
- Given $D^{\text{norm}} \in \overline{\mathcal{D}}^{\text{norm}}$, use convex optimization to find a minimal $\hat{\epsilon}$ such that D^{norm} has a section which is a $\hat{\epsilon}$-optimal interpolant.

  This is a achieved by finding good local sections of D^{norm} and then patching these up using a partition of unity supported on the base manifold of D^{norm}.

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. Choosing this cut-off is somewhat subjective, though; not to mention the fact that some of these manifold optimization procedures, among them Kamada-Kawai MDS, are NP-hard to optimize in the first place.

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:

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

with $\mu \in \mathbb{R}^D$ and $W \in \mathbb{R}^{D\times d}$ defining the underlying subspace, and $d \ll D$ presumably. We assume isotropic noise. Note that in the noiseless regime, the problem 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, i.e. $p(y) = \int_x p(y|x)p(x)dx$, then indeed y is still a multivariate normal. If $C = WW^T + \sigma^2 I_D$, then $y \sim \mathcal{N}(\mu, C)$. 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:

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

It is shown in [TB99] that, for fixed dimension (i.e. the interior minimization) this objective recovers the usual PCA method. 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

```
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 $C = \frac{1}{N}\sum_{i=1}^{n}(y_i - \mu)(y_i - \mu)^T$.
  Return $\hat{d}$ as the number of eigenvalues of $C$ 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.

Theorem 8. *Let $\lambda_{min}(WW^T) > \sigma^2 + 2\epsilon$ for some $\epsilon > 0$. Then with cutoff parameter $\eta = \sigma^2 + \epsilon$ and sample size $n \geq O(\frac{D\log(1/\delta)^2}{\epsilon^2})$, Algorithm 2 outputs a correct estimate of the dimension with probability $\geq 1 - \delta$.*

In order to prove this, we invoke the following theorem regarding the convergence of the empirical covariance matrix to its true value.

Theorem 9 (see [Ver10], Corollary 5.50)**.** *Let X_i denote independent samples from a subgaussian distribution in $\mathbb{R}^D$ with covariance matrix Σ, and let $\epsilon \in (0, 1)$, $t \geq 1$. Then with probability $\geq 1 - 2\exp(-t^2 D)$,*

$$n \geq C(t/\epsilon)^2 D \implies \|\Sigma_n - \Sigma\| \leq \epsilon$$

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

Proof of Theorem 8. Let $t \geq \sqrt{\log(1/2\delta)/D}$. Then with probability $1 - \delta$, for $n \geq O(\frac{D\log(1/\delta)^2}{\epsilon^2})$, we have$\|\Sigma_n - \Sigma\| \leq \epsilon$ and indeed, by basic properties of the operator norm, every eigenvalue of Σ_n is within ϵ of its corresponding eigenvalue in Σ. In this event, out cutoff will indeed only select for the non-noisy eigenvalues and hence output the correct intrinsic dimension. □

Of course, the difficulty in practice is getting the right cutoff between signal and noise. There are various 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).

Open Problem 2. *Establish matching upper and lower bounds on ID estimation in the linear case. Furthermore, develop an algorithm which can adaptively learn the noise from the signal, given the promise that there is a separation.*

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

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

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

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

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 19. *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:

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

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 20 (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****.*

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

Definition 21 (Assouad 1983)**.** *The* ***doubling dimension*** *of a set $S \subset \mathbb{R}^D$ is:*

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

Definition 22. *Let (X, d) be a metric space. A set $V \subset X$ is r-**separated** if $d(x, y) \geq r$ for all distinct $x, y \in V$. The r-**packing number** $M_\epsilon(S)$ 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:*

$$\mathcal{N}_\epsilon(S) \leq M_\epsilon(S) \leq \mathcal{N}_{\epsilon/2}(S)$$

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

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

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 23. *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 $M_\epsilon(S)$ 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 $n^{1-\epsilon}$ for all $\epsilon > 0$ [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.

With the estimate for $M_\epsilon(S)$ 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.

Algorithm 3 Kegl's algorithm to estimate $M_\epsilon(S)$

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

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 24 (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:*

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

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

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

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 25. *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:*

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

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 $l\epsilon^d \to \infty$. 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 $\epsilon = \epsilon_d(l)$ as a function of l and d.
- Break data into subsamples of varying sizes $N_1, ..., N_k = [n]$.

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

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

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

3.3.3 Local Covariance Structure

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 26 (see [DF08], Definition 2)**.** *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 $\sum_{i\in[d]} \sigma_i^2 \geq (1-\epsilon)\sum_{i\in[D]} \sigma_i^2$.*

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 *kd*-trees) is able to adapt to this notion of intrinsic dimension of data.

Ultimately, like other ID estimation techniques, local PCA of the nature suggested in this definition suffers from the multi-scale problem: what is an appropriate neighborhood size to consider the data to be approximately linear? This motivates the multi-scale approach by [Lit+09]. 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 $r > 0$. The singular values corresponding to noise are the ones which do not scale with r. There is also a key difference between singular values corresponding to tangent directions or curvatures, depending on whether they scale linearly or quadratically with r. This analysis of the local behavior helps determine an appropriate neighborhood size for looking at singular values (i.e. principal components).

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."

$$\hat{d}_{\text{avg}} = \frac{1}{n} \sum_{i=1}^{n} \hat{d}(x_i) \wedge D$$

$$\hat{d}_{\text{vote}} = \arg\max_{d' \in \mathbb{N}^+} \sum_{i=1}^{n} \mathbf{1}[\hat{d}(x_i) = d']$$

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

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

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

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

3.4.1 Minimal Subgraphs

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.

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

They formulate the ID estimation problem as a binary decision problem. Note that τ_g is the global reach of the manifold.

- If $\mathrm{TSP}(X_{1:n}; d_1) \leq O(\max\{1, \tau_g^{d_1 - D}\})$, return $\hat{d} = d_1$. Otherwise, return $\hat{d} = d_2$.

With some heroic effort, this makes way for a minimax upper bound, described further in a later section of this thesis.

Minimum Spanning Tree Following [CH03], we consider the use of the *minimum spanning tree* of an undirected weighted graph as a proxy for intrinsic dimension. In this case they use the scaling of the weight of the MST with respect to the size of a sample of the data: larger samples should yield weight that grows exponentially with the dimension of the dataset.

Open Problem 3. *Develop guarantees for minimum spanning tree based estimators of the intrinsic dimension, in a similar spirit to the minimax upper bound given by [KRW16] using TSP. If* ST *is the set of spanning trees of a graph consisting of points* $X_{1:n}$ *in Euclidean space then the following quantity might be of interest:*

$$\mathrm{MST}(X_{1:n}; d_1) = \min_{T \in \mathsf{ST}} \left\{ \sum_{(i,j) \in T} \|X_i - X_j\|_{\mathbb{R}^d}^{d_1} \right\}$$

The benefit, of course, is that MST is poly-time computable (via Prim's or Kruskal's algorithm) while TSP is NP-hard.

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 11 (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 27. *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 12 (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 [MN24] 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:

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

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 most interested in lower bounds, which characterize how well *any* estimator could do. More specifically, we are interested in bounding the minimax rate R_n: the performance of the best 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 rate is *exponential*, i.e. $R_n = \Theta(q^n)$ for some $q \in (0, 1)$.
- In the noiseless model of [KRW16], the minimax rate is *superexponential*, 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** $\hat{\theta}_n : \mathbb{R}^n \to \Theta$ takes the observations (i.e. $X_{1:n}$) and outputs a prediction for the statistic.

Definition 28. *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:*

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

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. But first, some preliminaries:

To upper bound the minimax risk, it suffices to isolate an estimator $\hat{\theta}_n$ 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 ineqaulities like that of Le Cam, Fano, and Assouad. See [YA97]. We recall Le Cam's lemma because it will be crucial for the lower bound set up by [KRW16].

Theorem 13 (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 be any distribution in the convex hull of P_i.*

$$Q_i \in \operatorname{conv}(\mathcal{P}_i) = \Big\{ \sum_j \alpha_j P_j : \alpha_j \geq 0, \sum_j \alpha_j = 1, P_j \in \mathcal{P}_i \Big\} \subset \mathcal{P} \qquad i \in \{1, 2\}$$

Let q_i be the density of Q_i with respect to a measure μ. Then, for all estimators $\hat{\theta}$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}^P \Big[d(\hat{\theta}, \theta(P)) \Big] \geq \frac{d(S_1, S_2)}{2} \int [q_1(x) \wedge q_2(x)] d\nu(x)$$

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

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

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

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

Substituting for M gives the desired formula. □

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

4.2 Noise Deconvolution Model

We imagine, at first, a very straightforward generative model. Suppose you observe samples $Y_j \in \mathbb{R}^D$ generated as follows:

$$Y_j = X_j + \eta_j$$

where $\eta_j \sim \mu$ and $X_j \sim P$, each sampled i.i.d. 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 $\hat{Q}$. The true distribution is $Q = P \star \mu$, where $\star$ denotes the convolution operator.

Think even more generally than dimension estimation for a moment. If $\mathcal{P}$ is 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.

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,*

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

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*

- $\operatorname{supp}(P) \subset B(0,1)$
- *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 $d = \dim(P)$, $|\{B \in \mathcal{N}(\epsilon) : \operatorname{dist}(B^+, \operatorname{supp}(P)) \leq \epsilon\}| \leq \Theta\epsilon^{-d}$, where B^+ denotes the same ball with radius doubled, and $\mathcal{N}(\epsilon)$ is any ϵ-cover of* $\operatorname{supp}(P)$.
- *For $\epsilon > 0$ and for any ball of radius ϵ, $P(B) \leq C\epsilon^d$*

Key to their analysis is the idea of a **deconvolving empirical measure** $\hat{P}_n$. 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 $\mathbb{R}^m$.

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

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. $\mathbb{E}\hat{P}_{n,\Psi}(A) = \mathbb{P}_\Psi(A)$. They use this measure to define first a sort of empirical covering number,

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

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

$$\hat{d}_n = \left\lfloor \frac{\log \hat{N}_n}{\log(1/\epsilon)} + \frac{1}{2} \right\rfloor \tag{4.1}$$

Theorem 14. *Suppose $\epsilon < (\Theta^{-1} \wedge (2C)^{-1})^{2/\delta(\mathcal{D})}$ and $\gamma < (1/2) \wedge (\epsilon^D/(12\Theta))$. Suppose $\Psi(\{x : |x| \geq \epsilon\}) \leq \gamma$. Then there exists $\Lambda > 0$ and $q \in (0, 1)$ such that:*

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

4.3 Noiseless Model

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

4.3.1 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 29 (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.

4.3.2 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 2. *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 3. *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 $\hat{d}_n$,

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

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

Bibliography

[Bak+23] Ainesh Bakshi et al. "A quasi-polynomial time algorithm for Multi-Dimensional Scaling via LP hierarchies". In: *arXiv preprint arXiv:2311.17840* (2023).

[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.

[Dem+21] Erik Demaine et al. "Multidimensional scaling: Approximation and complexity". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 2568–2578.

[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.

[DKS10] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. "A sparse johnson: Lindenstrauss transform". In: *Proceedings of the forty-second ACM symposium on Theory of computing*. 2010, pp. 341–350.

[DTV11] Amit Deshpande, Madhur Tulsiani, and Nisheeth K Vishnoi. "Algorithms and hardness for subspace approximation". In: *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*. SIAM. 2011, pp. 482–496.

[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.

[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.

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

[KN14] Daniel M Kane and Jelani Nelson. "Sparser johnson-lindenstrauss transforms". In: *Journal of the ACM (JACM)* 61.1 (2014), pp. 1–23.

[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).

[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.

[MN24] 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.

[MZ23] Marina Meilă and Hanyu Zhang. "Manifold learning: what, how, and why". In: *Annual Review of Statistics and Its Application* 11 (2023).

[NM10] Hariharan Narayanan and Sanjoy Mitter. "Sample complexity of testing the manifold hypothesis". In: *Advances in neural information processing systems* 23 (2010).

[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.

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

[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.

[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.