Manifold Dimension Estimation for Omics Data Analysis: Current Methods and a Novel Approach

Kerstin Johnsson

March 22, 2011

Abstract

In the field of molecular biology many data sets with thousands, tens of thousands or even more variables are produced daily, for example in genomics. Traditional statistical approaches such as hypothesis testing cannot exploit the full potential of such data sets when there are functional relations between the variables, and if the functional relations are non-linear also linear methods such as PCA do not work.

The more general approach is to look for manifolds on which the data are supported, and the first step in most manifold learning methods is to determine the dimension of the manifold. In this work we review five current methods of manifold dimension estimation: PCA, Takens' estimator, the Hill estimator, vector quantization, and k–NN. We also introduce a novel dimension estimator — the expected absolute projection (EAP) estimator, and compare its performance to the five other methods. The results do not show any significant advantage of the EAP estimator, however we do suggest improvements of the EAP estimator which might render it competitive.

∽ Tack ∼

Mitt tack går först till tre fantastiska lärare: Magnus Fontes vars engagemang och entusiasm varit inspirerande och till stor hjälp för detta arbete; Victor Ufnarovski som var den som lärde mig matematik på riktigt och så min mor — otvivelaktigt den viktigaste läraren av dem alla.

Jag vill också tacka Charlotte Soneson för den hjälp och inspiration du bistått med.

Contents

List of notation

$\lvert\cdot\rvert$	Absolute value or Euclidean length.
$a \parallel b$	a is parallell to b
$\overline{x}$	Mean value of $\{x_i\}_{i\in I}$.
$B_\delta(x)$	The open ball of radius δ centered at x.
$\overline{A}$	The closure of the set A.

Chapter 1

Introduction

In many areas of science today vast amounts of data are produced. Genomics is one area where the number of variables measured for each sample is often in the order of tens of thousands. One example is genomewide measurements of gene expression, where the amount of mRNA resulting from transcription is measured for each gene. By using hypothesis testing on such data it might be possible to immediately single out individual genes which have fundamental impact on a certain feature of the samples, but it is more often the case that groups of genes together form patterns which can be used to discriminate between groups of samples. This leads to two questions: a) What are the groups of genes and what are the relations between the genes in each group? and b) Using this information, how can we better discriminate features of samples? One way to address these questions is to use linear methods, i.e. to compute covariance or correlation between variables and use some linear model. This has been done successfully for many gene expression data sets over the last decade e.g. through principal components analysis (PCA); one of the early examples is [36]. However, the relations between genes might very well be non-linear, and then linear methods will yield sub-optimal results. This prompts for methods to find non-linear structures in data, and the natural step is to look for manifolds[1].

Consider the example in figure 1.1. The data behind the two graphs have identical variance for each variable and identical correlation, but it can easily be seen that the second data set lies on a 1-manifold (a curve) whereas the first data set does not. These two data sets are the first two in *Anscombe's quartet* from 1973 [1], which were constructed to show the importance of using graphs in statistical analyses. The problem is that when there are more than three variables it is generally impossible to plot the data since our world is three-dimensional; in principle one could utilize color and time to visualize extra dimensions, but in any case we cannot visualize more than a few dimensions. With tens of thousands of variables, as in gene expression data, we need unsupervised methods that can find manifolds in the high-dimensional data.

Manifold learning and non-linear dimension reduction is a field of much research interest, often with applications in image processing in mind. The first step in manifold learning or dimension reduction is usually to find the

[1]For the reader unfamiliar with manifolds, appendix D gives a condensed review of the background needed for this thesis.

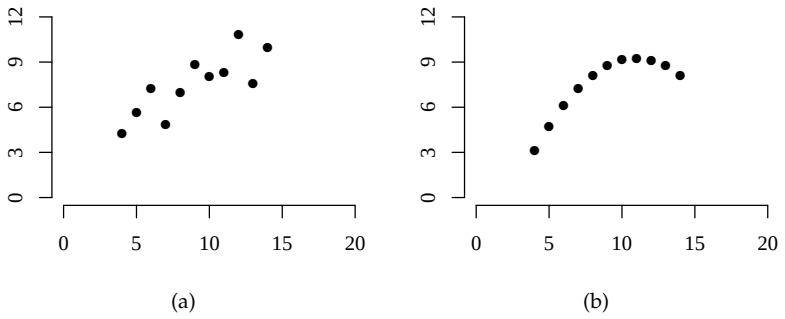


Figure 1.1: Two data sets with variance 10 and 3.75 for the first and second variable respectively, and correlation 0.898.

dimension of the manifold [28], and in the cases when the manifold learning algorithm itself can determine the dimension it is helpful to have an independent estimate of the dimension. Furthermore, as we can see from the example in figure 1.1, the dimension of the manifold on which the data lie can be a very useful piece of information. Hence it is important to find ways to estimate this dimension.

It very unlikely though that we will find data perfectly aligned on a manifold with lower dimension than the number of variables. What we might expect is a manifold with some noise added to it. This is equivalent to a model where a number of independent latent variables, that are fewer than the number of measured variables, account for a strong signal in the data, and the rest of the variation in the data is considered to be noise. The number of latent variables will be the dimension of the manifold.

When the aim is dimension reduction or manifold learning it is necessary to distinguish the dimension of the manifold and disregard the noise. However, if a dimension estimate is used to describe the degree of data dependencies it is reasonable to model the noise as due to latent variables with smaller variance than the variables defining the manifold. The intrinsic dimension of the data, which is usually defined as the number of independent latent variables, is then somewhat higher than the dimension of the manifold.

In this thesis however, intrinsic dimension estimation means to estimate the dimension of a manifold which might have noise added to it. The reason is that methods for intrinsic dimension estimation are constructed to measure dimension of manifolds and we do not try to quantify the contribution of noise in other ways than applying our dimension estimators to simulated data sets of manifolds with noise and discuss the results.

The first method to estimate intrinsic dimension was PCA, it was invented by Pearson already in 1901 [31]. It is a linear method, so it yields the dimension of the minimal linear subspace which fits the data well. The first non-linear method for dimension reduction and intrinsic dimension estimation was non-metric multidimensional scaling, introduced by Shepard in 1962 [38].

With the discovery of strange attractors in dynamical systems a need arose to estimate dimension of very complicated sets, since it was realized that the

dimension was an important characteristic of an attractor. But strange attractors do not have a dimension in the usual meaning of the word, a feature they have in common with fractals. The generalized concept of dimension used initially is usually known as box-counting dimension [12, 18] and it measures how a set fills out space. However, in 1983 a new concept of dimension was introduced independently by Grassberger and Procaccia [18] and Takens [39] — the correlation dimension. The correlation dimension is better suited for numerical estimation than the box-counting dimension and Takens developed subsequently a maximum likelihood estimator [40]. Grassberger [17] and Hentschel and Procaccia [24] generalized independently the concept of correlation dimension to an infinite series of dimensions, the Rényi dimensions (the name is due to their close relation to Rényi entropy [37]). The Rényi dimensions have subsequently become the basis for multifractal analysis [21], which is the study of systems with non-constant fractal dimension. From the Rényi dimensions the local fractal dimension can be determined by a Legendre transform under certain conditions [21]; this is an interesting approach but it has been outside the scope of this thesis.

During the last decade, a number of novel methods for intrinsic dimension estimation have been proposed [7, 23, 26, 32, 35]. Kégl presented in 2002 a novel way to estimate box-counting dimension [26]; in 2004 Costa et al published a paper in which lengths of certain random graphs on manifolds were utilized to estimate dimension, a method which later developed into what is called k–NN dimension estimation [4, 7]; Raginsky and Lazebnik presented in 2004 an estimator based on vector quantization and showed that it was a generalization of the estimator of Kégl [35]; and in 2005 Hein and Audibert published an estimator of correlation dimension that was specially adapted for data on manifolds [23]. A completely different approach was presented by Pestov in 2008 [32], he used what is called the curse of dimensionality to approach the problem of dimension estimation.

To this list we now add a dimension estimator based on the smallest possible constant in the reversed triangle inequality, which is also related to the curse of dimensionality as we will see in section 2.3. We call it the *expected absolute projection* (EAP), for reasons that will become clear in chapter 7.

In this thesis we have chosen five methods of dimension estimation to review along with our own: Local PCA [15], Takens' maximum likelihood estimator [40] and a similar estimator from [21] called the Hill estimator, the vector quantization estimator by Raginsky and Lazebnik [35] and k–NN dimension estimation [4, 5, 7, 8]. The methods have been selected for their comparatively high performance.

There is one category of methods of dimension estimation that is not touched upon in this thesis, and that is trial-and-error methods based on non-linear reduction techniques [3, 29]. These techniques tries to embed the data into a manifolds of varying dimension and returns a score of how well this can be done. The score is plotted for each dimension and the dimension estimate will be the dimension after which the score flattens out. Examples includes multidimensional scaling and neural network-methods. Trial-and-error methods are very computationally intensive, but might nevertheless be useful.

The next chapter introduces various mathematical definitions of dimension, including box-counting dimension and correlation dimension and gives some results about how these are related to each other. The not so mathematically

interested reader might read just the definitions of correlation dimension and box-counting dimension. Then there is one chapter for each dimension estimator (Takens' and the Hill estimator are together in one chapter). Finally, results from tests of the dimension estimators on a number of simulated data sets are presented and discussed.

We have assumed throughout that the reader is familiar with measure theory. For the reader who is interested in understanding the mathematics but that does not have the appropriate background, appendix E introduces a few concepts and properties.

In appendix C proofs are given of some general results that are used in this thesis.

Chapter 2

Definitions of Dimension

2.1 Topological Dimension

There are three canonical definitions of dimension in topology: Lebesgue covering dimension and the small and large inductive dimensions. These are topological invariants, i.e. constant under homeomorphisms and for separable metric spaces they are equal [13]. For a separable metric space their common value is called the *topological dimension*.

The definition of the Lebesgue covering dimension is:

Definition 2.1. *The Lebesgue covering dimension of a set S in a metric space is the smallest number n such that any finite open cover of S has a finite open refinement covering S such that any point in S belongs to at most $n + 1$ sets in the refinement. ($\{V_\alpha\}$ is a refinement of $\{U_\beta\}$ if there for each V_α exist a U_β such that $V_\alpha \subseteq U_\beta$.)*

Using that the Lebesgue covering dimension of $\mathbb{R}^d$ is d, it follows immediately that a d–manifold has locally dimension d, however it is much more difficult to prove that the entire manifold has dimension d, but it can be done by viewing the manifold as a *polyhedron* [11, 13].

It can easily be seen that any finite set has topological dimension zero, whereas the closure of a set has the same topological dimension as the set, so any countable dense set in a metric space has the same topological dimension as the space itself.

Thus topological dimension basically agrees with our intuitive idea of dimension, but the Lebesgue covering dimension definition is not possible to use as a means for dimension estimation, other than that it establishes that a d–manifold has indeed dimension d.

2.2 Fractal Dimensions

Now we will turn to generalizations of the usual concept of dimensions, where the dimension is allowed to be a non-integer.

We will present three fractal dimension definitions: the Hausdorff dimension, the box-counting dimension, and the correlation dimension. We will also give a definition of local dimension in a point. The box-counting dimension

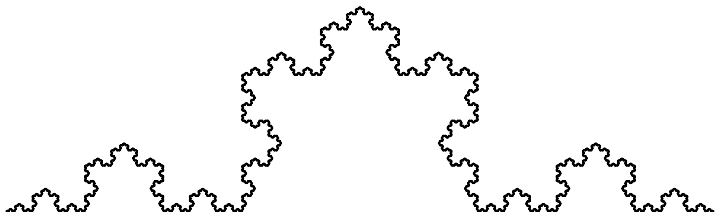


Figure 2.1: The von Koch curve.

and the correlation dimension are special cases of the Rényi dimensions, but we will not present the general form here, the interested reader can find it in [21] or [28].

For these generalized concepts of dimension to be useful, it is necessary that they agree with the topological dimension for manifolds and that they satisfy the basic inequality

$$A \subseteq B \Rightarrow \dim(A) \leq \dim(B). \tag{2.1}$$

That (2.1) holds will become immediately clear from the definitions, and the reader can either consult [12] or prove as an exercise that the Hausdorff dimension and the box-counting dimension is d for d–manifolds.

The correlation dimension differs from the other concepts of dimension presented here in that it defines dimension for a probability measure and not a set. It might seem odd, but this is in fact what makes it so suitable for empirical dimension estimation. The reason is that when we estimate dimension from a given data set we assume that the data are realizations of a random variable whose support is a manifold of a certain dimension, possibly with noise added. So if we use a dimension definition that applies to a set we need to estimate the support of the random variable, i.e. the support of the probability measure, before we can use the definition. If we have a definition of dimension that applies to a probability measure though, we can use the definition directly on the empirical distribution given by the data. This requires of course that the correlation dimension for the probability measure we study will agree with the topological dimension of its support. Towards the end of this section we will see that this is true for uniform measures on mildly regular manifolds.

The Hausdorff dimension was invented by Hausdorff in 1919 [22]; it is the oldest fractal dimension definition and it is defined for any set, as opposed to the box-counting dimension [12]. To introduce it we need another definition which also will become useful later:

Definition 2.2 (Hausdorff measure)**.** *For a subset S of a metric space (X, ρ) and a non-negative number m, we define for any $\delta > 0$*

$$H_\delta^m(S) = \inf\left\{\sum_{i\in\mathbb{N}} \operatorname{diam}(U_i)^m : \bigcup_{i\in\mathbb{N}} U_i \supseteq S;\ U_i \text{ open}, \operatorname{diam}(U_i) < \delta \ \ \forall i \in \mathbb{N}\right\}$$

where $\operatorname{diam}(U) = \sup\{\rho(x, y) : x, y \in U\}$*. The m–dimensional Hausdorff measure of S is then*

$$H^m(S) = \lim_{\delta\to 0} H_\delta^m(S).$$

Then we have

Definition 2.3 (Hausdorff dimension)**.** *The Hausdorff dimension of a set S in a metric space (X, d) is*

$$\dim_H(S) = \inf\{m : H^m(S) = 0\}.$$

The Hausdorff dimension is a valuable theoretical tool, but it is unfeasible to use for empirical dimension estimation. The box-counting dimension is closely related to it, and it allows for a straightforward method of estimation.

Box-counting dimension is defined for totally bounded sets, i.e. sets such that for any $\delta > 0$ they can be covered by a finite number of balls of radius δ. In $\mathbb{R}^n$ the totally bounded sets are the bounded sets.

Definition 2.4 (Box-counting dimension)**.** *For a totally bounded set S in a metric space, let $N_\delta(S)$ be the minimal number of balls of radius δ that cover S. The box-counting dimension is then*

$$\dim_{BC}(S) = \lim_{\delta \to 0} \frac{\log N_\delta(S)}{-\log \delta}$$

if the limit exists[1]*.*

The name box-counting dimension comes from an equivalent definition which is identical to the one above except that $N_\delta(S)$ is defined as the number of boxes in a δ–mesh that intersect S [12].

A direct result of the definitions 2.3 and 2.4 is:

Proposition 2.1. $\dim_{BC}(S) \geq \dim_H(S)$

Proof. Suppose that $\dim_{BC}(S) = d$ and $m > d$. Choose an $\epsilon > 0$ such that $\epsilon < m - d$. Then by the definition of box-counting dimension we can find a $\delta_0 > 0$, such that for all $\delta < \delta_0$

$$d + \epsilon > \frac{\log N_\delta(S)}{-\log \delta}.$$

Then

$$\delta^{-d-\epsilon} > N_\delta(S) \Rightarrow \delta^{-\epsilon} > N_\delta(S)\delta^d.$$

From the definition of $N_\delta(S)$ we know that there is a cover of S consisting of $N_\delta(S)$ open balls with radius smaller than or equal to δ. Thus

$$H^m_\delta(S) \leq N_\delta(S)\delta^m = N_\delta(S)\delta^d\delta^{m-d} < \delta^{-\epsilon}\delta^{m-d} = \delta^{(m-d)-\epsilon}.$$

But $\delta^{(m-d)-\epsilon} \to 0$ as $\delta \to 0$, since $(m - d) - \epsilon > 0$. This means that $H^m(S) = 0$ for all $m > d$, so $\dim_H(S) = \inf\{m : H^m(S) = 0\} \leq d$. □

The definition of correlation dimension as we will present it here requires that the function $x \mapsto \mu(\overline{B}_\delta(x))$ is μ–integrable for all $\delta > 0$ smaller than some δ_0, but in fact we have:

[1]Here, as well as in the following definitions of dimension, the definition is given under the condition that a certain limit exists. One could also define the upper dimension using lim sup and the lower dimension using lim inf instead of lim. The upper and lower dimensions will always exist and when they are equal the limit in the dimension definition exists. The properties given in this chapter for the different definitions of dimension also hold for the upper and lower dimensions.

Proposition 2.2. *If μ is a Borel probability measure on $\mathbb{R}^n$, then $x \mapsto \mu(\overline{B}_\delta(x))$ is a μ–integrable function for any $\delta > 0$.*

In appendix C we have proved a lemma (lemma C.1) saying that if μ is a Borel measure on $\mathbb{R}^n$, then all Borel sets in $\mathbb{R}^{2n}$ are measurable for the product measure $\mu \times \mu$. Using this, the proposition follows from Tonelli's theorem:

Proof of proposition 2.2. We begin by noting that $\mu(\overline{B}_\delta(x)) = \int \chi_{\overline{B}_\delta(x)}\, d\mu$. Let $f\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be defined by $f((x,y)) = \chi_{\overline{B}_\delta(x)}(y)$. f is $\mu \times \mu$–measurable since

$$f((x,y)) = \begin{cases} 0 & \text{if } |x-y| > \delta \\ 1 & \text{if } |x-y| \leq \delta \end{cases}$$

and $|x-y| > \delta$ is an open set in $\mathbb{R}^{2n}$, so the inverse image of any set, in particular a measurable set, is a Borel set and by lemma C.1 $\mu \times \mu$–measurable. Then by Tonelli's theorem, $y \mapsto f(x,y)$ and $x \mapsto \int f((x,y))\, d\mu(y)$ are μ–measurable so $\int f((x,y))\, d\mu(y)$ and $\int (\int f(x,y)\, d\mu(y))\, d\mu(x)$ are well defined and since $\mu(\mathbb{R}^n) = 1$,

$$\int \left(\int f(x,y)\, d\mu(y) \right) d\mu(x) \leq \int \left(\int 1\, d\mu(y) \right) d\mu(x) \leq \int 1\, d\mu(x) \leq 1,$$

which means that $x \mapsto \int f(x,y)\, d\mu(y) = \mu(\overline{B}_\delta(x))$ is μ–integrable. □

Now we can define correlation dimension:

Definition 2.5 (Correlation dimension)**.** *For a Borel probability measure μ on $\mathbb{R}^n$ we define*

$$C_{corr}(\mu, \delta) = \int \mu(\overline{B}_\delta(x))\, d\mu(x).$$

The correlation dimension of μ is then

$$\dim_{corr}(\mu) = \lim_{\delta \to 0} \frac{\log C_{corr}(\mu, \delta)}{\log \delta}$$

if the limit exists.

To get an interpretation of $C_{corr}(\mu, \delta)$, we note (as in [21]) that if X_1 and X_2 are random variables with probability distribution μ, and we let $I(s)$ denote the indicator function (i.e. I(s) is 1 if the statement s is true and 0 otherwise), then

$$\mu(\overline{B}_\delta(x)) = \int I(|x_1 - x| \leq \delta)\, d\mu(x_1) = \Pr\left[|X_1 - x| \leq \delta\right],$$

$$\int \mu(\overline{B}_\delta(x))\, d\mu(x) = \int \Pr\left[|X_1 - x| \leq \delta\right]\, d\mu(x) = \Pr\left[|X_1 - X_2| \leq \delta\right] \tag{2.2}$$

The second equation will become useful when estimating correlation dimension.

The definitions of Hausdorff and box-counting dimension can be used to define dimension locally simply by considering subsets of the support of the probability distribution.To do the same for correlation dimension, we need a minor modification:

Definition 2.6. *For a Borel probability measure μ on $\mathbb{R}^n$ and a measurable subset $A \subseteq \mathbb{R}^n$ we define the correlation dimension of μ in A as*

$$\dim_{corr}(\mu, A) = \lim_{\delta \to 0} \frac{\log \frac{1}{\mu(A)} C_{corr}(\mu|_A, \delta)}{\log \delta},$$

where $\mu|_A(E) = \mu(E \cap A)$ for all measurable subsets $E \subseteq \mathbb{R}^n$.

We can also define the dimension at a point x, which will be referred to as just *the local dimension at x*:

Definition 2.7. *For a Borel probability measure μ on $\mathbb{R}^n$ and a point $x \in \mathbb{R}^n$ we define the local dimension of μ at x as*

$$\dim_{loc}(\mu, x) = \lim_{\delta \to 0} \frac{\log \mu(B_\delta(x))}{\log \delta}$$

if the limit exists.

The local dimensions in the points of a set bounds the Hausdorff dimension and the box-counting dimension of the set; the following inequality is proved in [41]:

Proposition 2.3. *For a Borel measure μ on $\mathbb{R}^n$ without atoms[2], and a measurable set $A \subseteq \mathbb{R}^n$ with $\mu(A) > 0$, if*

$$c \leq \dim_{loc}(\mu, x) \leq C, \ \forall x \in A$$

then $c \leq \dim_H(S), \dim_{BC}(S) \leq C$.

An immediate corollary of this proposition is that if $\dim_{loc}(\mu, x) = d$ for all x in a set S, then $\dim_H(S) = \dim_{BC}(S) = d$.

For the relation between the local dimension in the points in a set and the local correlation dimension of the set, we have at least the following:

Proposition 2.4. *If $A \subseteq \mathbb{R}^n$ is measurable and $\dim_{loc}(\mu, x) = d$ for all $x \in A$, and if there exist a C and a $\delta_0 \in \mathbb{R}$ such that*

$$\left| \frac{\log \mu(\overline{B}_\delta(x))}{\log \delta} - d \right| < C \qquad \forall \delta < \delta_0, \tag{2.3}$$

then $\dim_{corr}(\mu, A) \leq d$. Furthermore, if

$$\frac{\log \mu(\overline{B}_\delta(x))}{\log \delta} \to d$$

uniformly as $\delta \to 0$, then $\dim_{corr}(\mu, A) = d$.

Proof. Assume that $\dim_{loc}(\mu, x) = d$ for all x in a measurable set A and that (2.3) holds for some $C, \delta_0 \in \mathbb{R}$. Let

$$\epsilon_x(\delta) = \frac{\log \mu(\overline{B}_\delta(x))}{\log \delta} - d. \tag{2.4}$$

[2] A set E with $\mu(E) > 0$ is an *atom* if $\forall F \subseteq E$, $\mu(F) = \mu(E)$ or $\mu(F) = 0$.

Then $\epsilon_x(\delta) \to 0$ pointwise as $\delta \to 0$ and $|\epsilon_x(\delta)| < C$ if $\delta < \delta_0$. Rewriting (2.4) we get

$$\mu(\overline{B}_\delta(x)) = \delta^{d+\epsilon_x(\delta)}$$

which yields

$$\log C_{corr}(\mu|_A, \delta) = d \log \delta + \log \frac{1}{\mu(A)} \int_A \delta^{\epsilon_x(\delta)} \, d\mu(x). \tag{2.5}$$

Now, since $x \mapsto \log x$ is concave, Jensen's inequality gives that

$$\log \frac{1}{\mu(A)} \int_A \delta^{\epsilon_x(\delta)} \, d\mu(x) \geq \frac{1}{\mu(A)} \int_A \log \delta^{\epsilon_x(\delta)} \, d\mu(x) = \log \delta \frac{1}{\mu(A)} \int \epsilon_x(\delta) \, d\mu(x)$$

so if $\delta < 1$,

$$\frac{\log \frac{1}{\mu(A)} C_{corr}(\mu|_A, \delta)}{\log \delta} \leq d + \frac{1}{\mu(A)} \int_A \epsilon_x(\delta) \, d\mu(x).$$

Since $|\epsilon_x(\delta)| < C, \forall \delta < \delta_0$ and $\epsilon_x(\delta) \to 0$ pointwise, the Lebesgue dominated convergence theorem gives that $\int_A \epsilon_x(\delta) d\mu \to 0$ as $\delta \to 0$. Thus

$$\lim_{\delta \to 0} \frac{\log \frac{1}{\mu(A)} C_{corr}(\mu|_A, \delta)}{\log \delta} \leq d.$$

Now, if

$$\frac{\log \mu(\overline{B}_\delta(x))}{\log \delta} \to d \tag{2.6}$$

uniformly as $\delta \to 0$, then clearly (2.3) holds for some $C, \delta_0 \in \mathbb{R}$, so it is enough to prove that

$$\lim_{\delta \to 0} \frac{\log \frac{1}{\mu(A)} C_{corr}(\mu|_A, \delta)}{\log \delta} \geq d.$$

In fact, it is enough that we find a function η such that $\eta(\delta) \to 0$ as $\delta \to 0$ and

$$\frac{\log \frac{1}{\mu(A)} \int_A \delta^{\epsilon_x(\delta)} \, d\mu(x)}{\log \delta} \geq \eta(\delta) \tag{2.7}$$

since using (2.5) we have

$$\frac{\log C_{corr}(\mu|_A, \delta)}{\log \delta} \geq d + \eta(\delta) \implies \lim_{\delta \to 0} \frac{\log C_{corr}(\mu|_A, \delta)}{\log \delta} \geq d.$$

We note that (2.7) is equivalent with

$$\frac{1}{\mu(A)} \int_A \delta^{\epsilon_x(\delta)} \, d\mu(x) \leq \delta^{\eta(\delta)}.$$

Uniform convergence in (2.6) means that $\epsilon_x(\delta) \to 0$ uniformly in x as $\delta \to 0$. With $\eta(\delta) = -\sup_{x \in A}(|\epsilon_x(\delta)|)$, $\delta^{\eta(\delta)} \geq \delta^{\epsilon_x(\delta)}$ for all $x \in A$, so

$$\frac{1}{\mu(A)} \int_A \delta^{\epsilon_x(\delta)} \, d\mu(x) \leq \frac{1}{\mu(A)} \int_A \delta^{\eta(\delta)} \, d\mu(x) \leq \delta^{\eta(\delta)}$$

and since we have uniform convergence of $\epsilon_x(\delta)$, $\eta(\delta) \to 0$ as $\delta \to 0$. Thus

$$\lim_{\delta \to 0} \frac{\log \frac{1}{\mu(A)} C_{corr}(\mu|_A, \delta)}{\log \delta} \geq d.$$

□

In fact, Cutler [9] proves that if $\dim_{loc}(\mu, x)$ exists for almost every x in a set S, then $\dim_{corr}(\mu, S) \leq \dim_H(S)$. Using proposition (2.3) this immediately gives the first inequality in proposition (2.4).

Now we can prove

Proposition 2.5. *Suppose that $\mathcal{M}$ is a d–manifold such that for each $x \in \mathcal{M}$ there is a $\delta_x > 0$ and a bi-Lipschitz function $\phi_x \colon \overline{B}_{\delta_x}(x) \cap \mathcal{M} \to \overline{B}_{\epsilon_x}(\phi_x(x)) \subseteq \mathbb{R}^d$, and furthermore that there are constants C and c such that $L(\phi_x) < C$ and $L(\phi_x^{-1}) < c$, $\forall x \in \mathcal{M}$. We call such a manifold a uniform bi-Lipschitz manifold. If μ is a uniform measure on a set $A \subseteq \mathcal{M}$, then $\dim_{corr}(\mu, A) = d$.*

Proof. By proposition 2.4 it is sufficient to prove that

$$\frac{\log \mu(\overline{B}_\delta(x))}{\log \delta} \to d$$

uniformly for each $x \in A$ as $\delta \to 0$. We use the fact that since μ is a uniform measure on a d–manifold, $\mu(E) = K \cdot H^d(E)$ for all measurable $E \subseteq \mathcal{M}$ for some constant K, where H^d is the d–dimensional Hausdorff measure [19]. Similarly we have that with ν denoting Lebesgue measure on $\mathbb{R}^d$, $\nu(E) = k \cdot H^d(E)$ for all measurable $E \subseteq \mathbb{R}^d$ for a constant k.

Now if f is a Lipschitz function with Lipschitz constant $L(f) = C$, then

$$\begin{aligned}
H^d_\delta(f(U)) &= \inf\left\{\sum_i \operatorname{diam}(V_i)^d : \bigcup_i V_i \supseteq f(U);\ V_i \text{ open},\ \operatorname{diam}(V_i) < \delta\right\} \\
&\leq \inf\left\{\sum_i \operatorname{diam}(f(U_i))^d : \bigcup_i U_i \supseteq U;\ U_i \text{ open},\ \operatorname{diam}(U_i) < \frac{\delta}{C}\right\} \\
&\leq \inf\left\{\sum_i C^d \operatorname{diam}(U_i)^d : \bigcup_i U_i \supseteq U;\ U_i \text{ open},\ \operatorname{diam}(U_i) < \frac{\delta}{C}\right\} \\
&= C^d \cdot H^d_{\delta/C}(U) \\
&\Longrightarrow H^d(f(U)) \leq C^d \cdot H^d(U)
\end{aligned}$$

This gives us that

$$\begin{aligned}
\mu(\overline{B}_{\delta_x}(x)) = K \cdot H^d(\overline{B}_{\delta_x}(x)) &= K \cdot H^d(\phi_x^{-1}(\overline{B}_{\epsilon_x}(\phi_x^{-1}(x)))) \leq c^d \cdot H^d(\overline{B}_{\epsilon_x}(\phi_x^{-1}(x))) \\
&= kc^d \cdot V(d) \cdot \epsilon_x^d \leq kc^d \cdot V(d) \cdot (C \cdot \delta_x)^d = C' \cdot \delta_x^d \ ,
\end{aligned}$$

where $V(d)$ is the volume of the d–dimensional unit ball. Analogously, $\mu(\overline{B}_{\delta_x}(x)) \geq c'\delta^d$ for some constant $c' > 0$, and uniform convergence follows immediately. □

2.3 Intrinsic Dimension Defined by the Curse of Dimensionality

The curse of dimensionality is a collection of phenomena that makes it hard to do analyses of data sets with high intrinsic dimension [28]. One typical example is that the distances between points gets more and more similar, i.e. they concentrate around their mean, as the dimension increases. In normed vector spaces we have that if the d components $\{x_1, x_2, \ldots, x_d\}$ of a random vector x are independent and identically distributed with $\mathrm{E}[x_j^8] < \infty$, then

$$\mathrm{E}[\|\mathrm{x}\|] = \sqrt{\mu_2' d - \frac{\mu_4' - \mu_2'}{4\mu_2}} + O\left(\frac{1}{d}\right)$$
$$\mathrm{Var}\left[\|\mathrm{x}\|\right] = \frac{\mu_4' - \mu_2'}{4\mu_2} + O\left(\frac{1}{\sqrt{d}}\right)$$

where $\mu_k' = \mathrm{E}[x_j^k]$. Thus $\mathrm{Var}\left[\|\mathrm{x}\|\right] / \mathrm{E}[\|\mathrm{x}\|] \to 0$ as $d \to \infty$. An easy-to-follow proof of this is given in [10].

This shows that when we can consider the difference between two random points in a vector space as a vector with independent and identically distributed components with finite eigth order moment, the variance of the distance between the two points gets negligible in comparison to the mean as the dimension increases. But in fact, the concentration of distances around the mean holds in much more general circumstances, as we will see below.

Another problem in intrinsically high-dimensional spaces is that there is a sparseness of points — they seem empty. For example if we have points uniformly distributed over a ten-dimensional unit ball, and we want to cover half of the points, we need a ball of radius $0.5^{1/10} \approx 0.93$.

The good thing is that since these phenomena depend on the dimension, they can be exploited to construct definitions and/or estimators of intrinsic dimensionality. One simple example is the intrinsic dimension definition for a probability measure by Chávez et al which uses the distance D between two points randomly drawn from the probability distribution corresponding to the measure μ [6]. The definition is the following:

$$\dim_{Ch}(\mu) = \frac{\mathrm{E}[D]^2}{2 \cdot \mathrm{Var}\left[D\right]}.$$

It has been shown experimentally that with 3000 points, an estimator based on $\dim_{Ch}$ yields a good agreement with d for gaussian distributions $N(0,1) \times \ldots \times N(0,1)$ (d factors) for $1 \leq d \leq 50$ [33].

Pestov [32] uses a more mathematical approach to construct a dimension estimator based on the curse of dimensionality. He argues that the mathematical counterpart of the curse of dimensionality is the *concentration phenomenon* in certain families of metric (Borel) measurable spaces (mm-spaces). These families are called *Lévy families* and are sequences of mm-spaces $\{X_n\}$ that often have increasing dimension in the usual sense. A formal definition of the concentration phenomenon is

Definition 2.8. *The concentration phenomenon applies to a family of mm-spaces* (X_n, ρ_n, μ_n) *with* $\mu_n(X_n) = 1$*, if whenever we have a family of subsets* $\{A_n\}, A_n \subseteq X_n$

such that $\mu_n(A_n) \geq 1/2$*, then* $\mu_n((A_n)_\epsilon) \to 1$ *for every* $\epsilon > 0$*, where* $(A_n)_\epsilon = \{x \in X_n : \inf_{a \in A_n} \rho(x, a) < \epsilon\}$.

If we think of $\{\mu_n\}$ as a probability measure, this means that for any $\epsilon > 0$, the probability to draw a point from the corresponding probability distribution that is further than ϵ from A_n will approach zero as $n \to \infty$.

The definition of a Lévy family is connected to the first property of high-dimensional spaces that we stated, and it relates the *observable diameter* to the *characteristic size* of the mm-spaces in the family.

Definition 2.9. *The observable diameter of a mm-space* (X, ρ, μ), $\mathrm{ObsDiam}(X)$*, is the infimal value of* ϵ *such that for any 1–Lipschitz function* $f : X \to \mathbb{R}$

$$\Pr\left[|f(x) - f(y)| > \epsilon\right] < \kappa, \quad x, y \sim \mu$$

for some threshold value κ.

Definition 2.10. *The characteristic size of a mm-space* (X, ρ, μ), $\mathrm{CharSize}(X)$ *is the median value of distances between points in the space, i.e.*

$$\Pr\left[|x - y| \geq \mathrm{CharSize}(X)\right] = 1/2, \quad x, y \sim \mu$$

if μ *has no atoms.*

Definition 2.11. *A family of mm-spaces* (X_n, ρ_n, μ_n) *is a Lévy family if* $\mathrm{ObsDiam}(X_n) \ll \mathrm{CharSize}(X_n)$ *as* $n \to \infty$.

The definition of a Lévy family says that not only for the norm, but for every 1-Lipschitz function from X to $\mathbb{R}$, the values of the function sharply concentrate around their mean. Examples of Lévy families are $\{\mathbb{S}^n\}$, $\{\mathbb{B}^n\}$ with the Euclidean metric and the usual measures, and the Hamming cubes $\{\{0,1\}^n\}$ with normalized Hamming distance $\rho(a_1 a_2 \dots a_n, b_1 b_2 \dots b_n) = \frac{1}{n}|\{i : a_i \neq b_i\}|$ and the counting measure [19, 32].

The novel dimension estimator we will present in chapter 7 relies on the fact that $\{(\mathbb{S}^n, \rho_n, \nu_n)\}$ is a Lévy family, with ρ_n denoting Euclidean distance, and ν_n the rotation-invariant measure. $\mathrm{CharSize}(\mathbb{S}^n) \to \sqrt{2}$ as $n \to \infty$ and $\mathrm{ObsDiam}(\mathbb{S}^n) = O(1/\sqrt{n})$ [32]. We consider the 1–Lipschitz function $p_v : X_n \to \mathbb{R}$, which is projection onto a fixed vector v, e.g. $(1, 0, 0, \dots)$. By lemma C.2 it follows that $\mathrm{E}[|p_v(x)|] \to 0$ for $x \sim \nu_n$ as $n \to \infty$ since by symmetry the median of $p_v(x)$ with $x \sim \nu_n$ is zero for any n, and $\sup_{x \in \mathbb{S}^n} |p_v(x)| = 2$. The estimator is constructed using the fact that we actually have an analytic expression for $\mathrm{E}[|p_v(x)|]$, $x \sim \nu_n$ for each n.

Pestov on the other hand used a definition of dimension more closely connected to the concentration phenomenon, based on the most common quantitative measure of it, the *concentration function*.

Definition 2.12. *For a mm-space* X *with* $\mu(X) = 1$*, the concentration function* α_X *is defined for each* $\epsilon > 0$ *by*

$$\alpha_X(\epsilon) = 1 - \inf\{\mu(A_\epsilon) : A \subseteq X, \mu(A) > 1/2\} \ .$$

His definition of intrinsic dimension was

$$\dim_{Pe}(X) = \frac{1}{\left[2 \int_0^1 \alpha_X(\epsilon)\, d\epsilon\right]^2}$$

From the next theorem it follows that $\dim_{Pe}(X_n) \to \infty$ if and only if $\{X_n\}$ was a Lévy family.

Theorem 2.1. *$\alpha_{X_n}(\epsilon) \to 0$ as $n \to \infty$ for any $\epsilon > 0$ if and only if $\{X_n\}$ is a Lévy family.*

Another advantage with Pestov's approach is that it made no difference between continuous and discrete sets, if $\rho_{Gr}(X_n, X) \to 0$ as $n \to \infty$, then $\dim_{Pe}(X_n) \to \dim_{Pe}(X_n)$, where ρ_{Gr} is the *Gromov distance* [32]. Due to high computational complexity though, $\dim_{Pe}$, is not suitable for dimension estimation.

Chapter 3

PCA

The most widely known application of intrinsic dimension estimation and dimension reduction is data compression. To compress an image, it is first divided into N regions of $a \times a$ pixels. The N regions are samples from an a^2–dimensional space. When for example a discrete cosine transform (which is used in jpeg compression) is applied to each sample, the samples are projected onto a pre-determined linear subspace of the original a^2–dimensional space. The dimension of the linear subspace can be chosen so that the rate of compression is sufficiently large, or so that the projection error is sufficiently small, but otherwise it is not adapted to the samples. When using principal components analysis, PCA, on the other hand, the subspace is chosen in an optimal way — it yields the minimum squared approximation error given the dimension.

The idea of PCA is straightforward: to project on a linear subspace such that as much as possible of the variance[1] is kept. Maximizing variance within the subspace also means minimizing variance orthogonal to the subspace, so the subspace given by PCA yields the minimal approximation error when points are approximated by their projection onto a linear subspace of lower dimension.

PCA is inherently a linear technique, and can thus only find linear submanifolds, however a local version of PCA can be used to make dimension estimates, using the idea that a smooth d–manifold is locally well approximated by its tangent space. This was first done by Fukunaga and Olsen in 1971 [15].

The principal components constitute an orthonormal basis which defines the linear subspace which is projected on for dimension reduction. Suppose we have a data set $\{\mathrm{x}_i\}_{i=1}^N$ of points in $\mathbb{R}^p$, arranged in a $N \times p$ matrix X. Using the precept that variance in the projection on the principal components should be maximal, the first principal component $\mathrm{w}_1 \in \mathbb{R}^p$ is a vector of unit length such that $\mathrm{Var}[\,\mathrm{Xw}]$ is maximal. Using the notation $\overline{\mathrm{X}} = [\overline{\mathrm{x}}\ \ \overline{\mathrm{x}}\ \ldots\ \overline{\mathrm{x}}]^T$ we have for $\mathrm{w} \in \mathbb{R}^p$

$$\mathrm{Var}[\,\mathrm{Xw}] = (\mathrm{Xw} - \overline{\mathrm{X}}\mathrm{w})^T(\mathrm{Xw} - \overline{\mathrm{X}}\mathrm{w}) = \mathrm{w}^T(\mathrm{X} - \overline{\mathrm{X}})^T(\mathrm{X} - \overline{\mathrm{X}})\mathrm{w} = \mathrm{w}^T\Sigma\mathrm{w}$$

where $\Sigma = (\mathrm{X} - \overline{\mathrm{X}})^T(\mathrm{X} - \overline{\mathrm{X}})$ is the covariance matrix[2]. Since Σ is a normal positive-semidefinite matrix, there is a unitary matrix U of eigenvectors

[1] With a slight abuse of notation, variance for a data set $\{x_i\}$ here means $\sum_i (x_i - \overline{x})^2$, keeping in mind that variance for a random variable is defined as $\mathrm{E}[(X - \mathrm{E}[X])^2]$.

[2] with the covariance between two sets of data $\{a_i\}_{i=1}^N$ and $\{b\}_{i=1}^N$ meaning $\sum_{i=1}^N (a_i - \bar{a})(b_i - \bar{b})$.

$u_1, u_2, \ldots, u_p$ to Σ, with corresponding non-negative eigenvalues $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$. We can assume that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$. With D denoting the diagonal matrix with $\sigma_1, \sigma_2, \ldots, \sigma_p$ on the diagonal, we have the factorization $\Sigma = UD^2U^T$. Thus

$$\mathrm{Var}[Xw] = w^T UDDU^T w = |w^T UD|^2.$$

Let $(w')^T = w^T U$ and let $w_1, \ldots, w_p$ denote the entries of w. With the restriction $|w| = 1$, from

$$\sum_{i=1}^{p} (w_i')^2 = |w'| = |w| = 1 \quad \text{and} \quad \mathrm{Var}[Xw] = \sum_{i=1}^{p} (w_i' \sigma_i)^2$$

it is easy to see that $\mathrm{Var}[Xw]$ is maximal when $w = u_1$, so we can choose $w_1 = u_1$.

The second principal component w_2 is defined to be a vector $w \in \mathbb{R}^p$ that maximizes $\mathrm{Var}[Xw]$ under the restrictions that w should be orthogonal to w_1 and have unit length, the third is defined the same way except that it is required to be orthogonal also to w_2, and so on. Iterating the argument above we get that $w_i = u_i$ for $i = 1, \ldots, p$ is a possible choice for the principal components.

Now, dimension reduction to d dimensions can be obtained by centering the data (i.e. subtracting the mean) and then projecting it onto the subspace spanned by the d first principal components. This projection minimizes the projection error $\|P(X - \overline{X}) - (X - \overline{X})\|_F$ among orthogonal projections P onto d–dimensional subspaces, with $\|\cdot\|_F$ standing for Frobenius norm[3]. To see why, with $\widetilde{X} = X - \overline{X}$ we have for any orthogonal projection P that

$$\|\widetilde{X}\|_F^2 = \|\widetilde{X} - P(\widetilde{X})\|_F^2 + \|P(\widetilde{X})\|_F^2.$$

It is easy to see from the definition of the principal components that if we are restricted to d dimensions in the projection, $\|P(\widetilde{X})\|_F^2$ is maximal when the projection is onto the subspace spanned by the d first principal components. Thus $\|\widetilde{X} - P(\widetilde{X})\|_F^2$ must be minimal for projection onto this subspace.

One of the main advantages of PCA is that the eigenvalues $\sigma_1^2, \sigma_2^2, \ldots \sigma_p^2$ reflect how much of the variance is kept in the projection onto each principal component, and thus also how much variance that is lost in the projection. If some of the eigenvalues are zero, the data set lies in a linear subspace of $\mathbb{R}^p$ with the dimension given by the number of non-zero eigenvalues. If the data lie close to a linear subspace, but not quite, due to noise or a slight non-linearity, the eigenvalues corresponding to principal components orthogonal to the linear subspace will be small in comparison to eigenvalues corresponding to principal components within the linear subspace as long as the variance within the linear subspace is big enough. This makes it possible to estimate the intrinsic dimension for data on linear or almost-linear manifolds by comparing eigenvalues. However, the linear nature of PCA will make this method overestimate dimension for curved manifolds. To avoid this, PCA can be done locally, using only data points within a cut-off radius from a certain point.

[3]For a matrix A with row vectors $a_1, a_2, \ldots, a_N$ the Frobenius norm is $\|A\|_F = \sum_{i=1}^{N} |a_i|^2$.

Chapter 4

Correlation Dimension Estimation

We saw in chapter 2 that the correlation dimension of a uniform measure on a bi-Lipschitz manifold equals the topological dimension of the manifold. In view of equation (2.2) this means that

$$\Pr\left[|X_1 - X_2| \leq \delta\right] = \delta^{d+\epsilon_{corr}(\delta)},$$

where $\epsilon_{corr}(\delta) \to 0$ as $\delta \to 0$. However, this does not imply that $\delta^{\epsilon_{corr}(\delta)}$ is bounded. The estimators we present in this chapter will not work if this is the case, therefore we will give some additional conditions under which we in fact have

$$\Pr\left[|X_1 - X_2| \leq \delta\right] = c \cdot \delta^d, \ \forall \delta < \delta_0; \ \ X_1, X_2 \sim \mu \tag{4.1}$$

Using a similar argument to one used in [30] we will show that (4.1) holds if the probability measure corresponds to a uniformly continuous density function and its support is a d–manifold $\mathcal{M}$ for which we have an isometry $\phi\colon \mathbb{R}^d \to \mathcal{M}$.

If this holds for a probability measure μ with probability density function f, a random variable X distributed according to μ can be written as $X = \phi(Z)$, where Z is a random variable on $\mathbb{R}^d$ with uniformly continuous probability density function $\tilde{f} = f \circ \phi$.

Now fix a point $z \in \mathbb{R}^d$ and assume that $\tilde{f}(z)$ is approximately constant in a small ball of radius δ_0 around z. If $V(d)$ denotes the volume of the unit ball in $\mathbb{R}^d$, the volume of a ball in $\mathbb{R}^d$ with radius r is $V(d) \cdot r^d$; this leads to the powerlaw

$$\Pr\left[|Z_1 - z| < \delta\right] \approx \tilde{f}(z) V(d) \delta^d, \ \ Z_1 \sim \nu$$

for $\delta < \delta_0$. If we assume that $\tilde{f}$ is approximately constant in a ball of radius δ_0 around z for any $z \in \operatorname{supp}(\nu)$ (which we can do by uniform continuity), we get

$$\begin{aligned}\Pr\left[|Z_1 - Z_2| < \delta\right] &= \int \Pr\left[|Z_1 - z| < \delta\right] d\nu(z) \approx \int \tilde{f}(z) V(d) \delta^d d\nu(z) \\ &= V(d)\delta^d \int \tilde{f}(z) d\nu(z) = V(d)\delta^d, \quad \forall \delta < \delta_0\end{aligned}$$

If $X_1 = g(Z_1)$ and $X_2 = g(Z_2)$, then using isometry we get

$$\Pr\left[|X_1 - X_2| < \delta\right] = \Pr\left[|Z_1 - Z_2| < \delta\right] = V(d)\delta^d, \quad \forall \delta < \delta_0,$$

i.e. (4.1) holds.

Using (2.2) and (4.1) we can construct an estimator of correlation dimension almost directly. For a sample $\{x_1, \ldots, x_N\}$, $C_{corr}(\mu, \delta)$ in definition 2.5 can be estimated by

$$C^*_{corr}(\mu, \epsilon) = \frac{2}{N(N-1)} \sum_{\substack{i,j=1 \\ i<j}}^{N} I(|x_i - x_j| \leq \epsilon),$$

which is an unbiased estimate of the probability that the distance between two samples drawn from μ is below or equal to δ.

However, the estimate is bad for low δ, due to the fact that there are finitely many samples. Therefore one must assume that the error

$$\epsilon^*_{corr}(\delta) = \dim_{corr}(\mu) - \frac{C^*_{corr}(\mu, \epsilon)}{\log \delta} \tag{4.2}$$

is small in a range of values of δ where the estimate $C^*_{corr}(\mu, \epsilon)$ is still reliable.

The most straightforward way to estimate the correlation dimension is then to compute $C^*_{corr}(\mu, \epsilon)$ for $\delta = \delta_1, \delta_2, \ldots, \delta_N$, and fit a line to a section of the plot of

$$\{(\log \delta_1, \log C^*_{corr}(\mu, \delta_1)), (\log \delta_2, \log C^*_{corr}(\mu, \delta_2)), \ldots, (\log \delta_N, \log C^*_{corr}(\mu, \delta_N))\}$$

where it is deemed that a linear relationship between $\log C_{corr}(\mu, \delta)$ and $\log \delta$ holds. The slope of the line will be the dimension estimate.

This was the method used by Grassberger and Procaccia when they introduced the concept of correlation dimension [18].

4.1 Takens' Estimator

Takens' estimator is the maximum likelihood estimator of d for the model (4.1) with a predetermined δ_0.

Given a sample $\{X_1, X_2, \ldots, X_N\}$ of N points with distribution μ, the estimate is constructed as follows: Determine the distances between each pair of points in the sample, $|X_i - X_j|$ with $i \neq j$. Let $R_1, \ldots, R_M$ be an enumeration of those of these distances that are smaller than δ_0. The maximum likelihood estimate of d is then

$$d^*_T = -\left(\frac{1}{M} \sum_{k=1}^{M} \ln \frac{R_k}{\delta_0} \right)^{-1} \tag{4.3}$$

This might be a non-integer, and in the case that we know the dimension to be an integer, the estimate is rounded to the closest integer. Takens determined the standard error of the estimator to be $1/\sqrt{M} \cdot 1/d$ under the assumption that $R_1, R_2, \ldots, R_M$ are independent [40]. However, the R_i's are obviously not independent by way of their construction, so the variance is really higher. The variance when this is taken into account is derived in [34]. One way to avoid this higher variance is to use only a sample of all interpoint distances; this bootstrap approach is discussed in [21].

Takens suggested making the maximum likelihood estimate for many values of δ_0 to verify the powerlaw relationship (4.1). If the powerlaw relationship doesn't hold, Takens' estimator will be biased, an expression for the bias in this case is presented in [21].

4.2 The Hill Estimator

The Hill estimator based on work by Hill in 1975 regarding inference in situations like (4.1) when the form of the probability distribution is only known (or assumed) in a certain region [21, 25]. The approach is similar to that of Takens' estimator, the difference being that instead of using a cutoff radius, the M shortest distances between points in the sample are used, for some heuristically chosen M. Let $R_{(1)}, R_{(2)}, \ldots R_{(M)}$ be the M shortest distances between pairs of points in the sample, in order of length, $R_{(1)}$ being the shortest. The mathematics involved is more complicated than for Takens' approach, since it is necessary to compute the joint probability distribution of $R_{(1)}, R_{(2)}, \ldots R_{(M-1)}$ conditioned on that $R_{(M)} = r_{(M)}$, where $r_{(M)}$ is the observed value of $R_{(M)}$. However the resulting estimator is very similar to Takens' estimator, being

$$d_H^* = -\left(\frac{1}{M-1} \sum_{k=1}^{M-1} \ln \frac{R_{(k)}}{R_{(M)}} \right)^{-1} \tag{4.4}$$

The estimator is biased,

$$\mathrm{E}[d_H^*] = \frac{M-1}{M-2} \cdot d$$

if (4.1) holds, so by dividing by $M-2$ instead of $M-1$ in (4.4) the estimator will be unbiased if the powerlaw (4.1) holds. As for Takens' estimator, other cases are discussed in [21]. The variance of the estimator is

$$\mathrm{V}[d_H^*] = \left(\frac{M-1}{M-2} \right)^2 \cdot \frac{d^2}{M-3}$$

if (4.1) holds and $R_{(1)}, R_{(2)}, \ldots R_{(M)}$ are independent. Since as before they are not independent the variance is really higher, but it can be reduced by bootstrap procedures, see [21].

Chapter 5

Vector Quantization

Vector quantization is a way to approximate a probability distribution μ in $\mathbb{R}^D$ with a fixed number of points, prototypes, in $\mathbb{R}^D$. The result is a *quantizer*, Q, that for each $x \in \mathbb{R}^D$ assigns a prototype. The minimal error of a quantizer depends on the number of prototypes but also on the dimension of the probability distribution. If the support of μ is compact, the *quantization dimension of order* r for $1 \leq r \leq +\infty$ can be defined, a concept introduced by Zador in 1982 [42]. The Hausdorff dimension of $\text{supp}(\mu)$ is smaller or equal to the quantization dimension of any order [16], and we will see that for $r = \infty$ the quantization dimension equals the box-counting dimension.

For $1 \leq r \leq \infty$, the error of order r of the quantizer Q is defined as $e_r(Q|\mu) = \mathrm{E}[\|X - Q(X)\|^r]^{1/r}$, where X is distributed according to μ. The error of infinite order is defined as $e_\infty(Q|\mu) = \sup\{\|x - Q(x)\| : x \in \text{supp}(\mu)\}$. Let $\mathcal{Q}_k$ denote the set of all quantizers using k prototypes. The minimal error for k prototypes is then $e_r^*(k|\mu) = \inf\{e_r(Q|\mu) : Q \in \mathcal{Q}_k\}$. Now we are ready for the definition:

Definition 5.1 (Quantization dimension)**.** *The quantization dimension of order r of the probability measure μ is*

$$\dim_{quant}^{(r)}(\mu) = -\lim_{k\to\infty} \frac{\log k}{\log e_r^*(k|\mu)}$$

if the limit exists[1]*.*

Proposition 5.1. *In the limit $r = \infty$, the quantization dimension equals the box-counting dimension.*

Proof. To simplify notation, let $N(\epsilon) := N_\epsilon(\text{supp}(\mu))$. First note that $e_\infty^*(k|\mu) \to 0$ as $k \to \infty$ since $\text{supp}(\mu)$ is compact. Thus we can define a sequence $\{n_j\}_{j\in\mathbb{N}} \subseteq \mathbb{N}$ by letting $n_1 = 1$ and then choosing n_{j+1} as the smallest number such that $n_{j+1} > n_j$ and $e_\infty^*(n_{j+1}|\mu) < e_\infty^*(n_j|\mu)$. Then $n_j = N(e_r^*(n_j))$, so

$$\dim_{quant}^{(\infty)}(\mu) = -\lim_{j\to\infty} \frac{\log n_j}{\log e_\infty^*(n_j|\mu)} = -\lim_{j\to\infty} \frac{\log N(e_\infty^*(n_j|\mu))}{\log e_\infty^*(n_j|\mu)} = \dim_{BC}(\text{supp}(\mu))$$

□

[1] see footnote 1 in chapter 2

For a certain amount of regularity of μ, the quantization dimension will also equal the Hausdorff dimension and the correlation dimension.

Definition 5.2. *A measure μ is* regular of dimension d *if it has compact support and there exists constants c and δ_0 such that*

$$\frac{1}{c}\delta^d \leq \mu(B_\delta(x)) \leq c\delta^d \quad \forall \delta < \delta_0, \forall x \in \text{supp}(\mu).$$

It has been shown [16, 42] that if μ is regular of dimension d, the minimal error for k prototypes, $e_r^*(k|\mu)$, asymptotically complies with $e_r^*(k|\mu) = \Theta(k^{-1/d})$[2] Then it follows that μ has quantization dimension d.

That a probability distribution is regular of dimension d for some d is a rather strong condition, and if μ is regular of dimension d its support has Hausdorff dimension d (see [16]) and it can easily be proven that its correlation dimension is d.

To estimate the quantization dimension from a collection of samples, an approximation of the optimal quantizer is built from a subset of the samples, the training set. Then the minimal quantization error is estimated as the quantization error when this empirically optimal quantizer is applied to the rest of the samples, the test set. As is the case for estimating correlation dimension it is not possible to go in the limit since the approximation gets worse as k increases. Thus the minimal quantization error is estimated for a range of values of k, and the resulting quantization dimension is plotted as a function of k. Based on a heuristic argument, it is stated in [35] that the first minimum in this graph is a good estimate of the dimension.

An interesting feature of the vector quantization method is that there is a theoretical bound for the errors in dimension estimates due to noise, see [35].

[2] $\exists c_1, c_2$ such that $c_1 \cdot k^{-1/d} \leq e_r^*(k|\mu) \leq c_2 \cdot k^{-1/d}$ when k is big enough.

Chapter 6

k–NN Dimension Estimation

The method presented here using the k–NN (k nearest neighbors) graph was developed by Costa and Hero in 2004 [7, 8], and has been improved and applied in subsequent papers by Carter, Raich and Hero [4, 5]. Following a brief outline of the method we will give the mathematical background and thereafter we will present an algorithm which is a somewhat simplified version of the algorithm presented in [5].

The directed k–NN graph for a set of points in $\mathbb{R}^m$ is the directed and weighted graph where the nodes are the points of the set and from each point there are edges going to the k points closest to it in $\mathbb{R}^m$. The weights of the edges are the distances between the points (according to some metric).

If we build a directed k–NN graph from n points that are distributed on a d–manifold in $\mathbb{R}^m$, under some additional conditions that we will describe below, the γ–power sum ($\gamma > 0$) of all the weights in the graph will asymptotically equal $n^{1-\frac{\gamma}{d}} \cdot c$ as $n \to \infty$, where c is a constant depending only on the probability measure. We call this γ–power sum the γ–weighted total length of the directed k–NN graph.

Through bootstrapping, subsamples from an original data set can be constructed, which will give us samples of varying sizes for which the γ–weighted total length of the respective directed k–NN graphs can be computed; then using a least squares approach d can be estimated.

6.1 Mathematical Background

Suppose that $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$ are n points in $\mathbb{R}^d$ distributed according to a probability measure that is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$ and is supported in a compact set. Let $\mathcal{N}^k_{\mathcal{X}_n}$ be the function that to each point in $\mathcal{X}_n$ assigns the set of its k nearest neighbors in $\mathcal{X}_n$, i.e. $\mathcal{N}^k_{\mathcal{X}_n}(X_i)$ is the set consisting of the k points in $\mathcal{X}_n \backslash \{X_i\}$ closest to X_i. The γ–power sum of edges in the directed k–NN graph of $\mathcal{X}_n$ is then

$$L^k_{\mathbb{R}^d}(\mathcal{X}_n) = \sum_{i=1}^{n} \sum_{j \in \mathcal{N}^k_{\mathcal{X}_n}(X_i)} d_{\mathbb{R}^d}(X_i, X_j)^\gamma = \sum_{i=1}^{n} \sum_{j \in \mathcal{N}^k_{\mathcal{X}_n}(X_i)} |X_i - X_j|^\gamma. \tag{6.1}$$

Let $\alpha = 1 - \frac{\gamma}{d}$. By the Beardwood-Halton-Hammersley theorem [2, 7],

$$L^k_{\mathbb{R}^d}(\mathcal{X}_n) = n^\alpha \cdot c + o(n^\alpha).$$

Now, suppose that we have n points $\mathcal{Y}_n = \{Y_1, Y_2, \ldots, Y_n\}$ distributed according to a probability measure μ that is supported on a compact subset of a d–manifold $\mathcal{M} \subseteq \mathbb{R}^m$, such that there is an isometry $\phi\colon \mathcal{M} \to \mathbb{R}^d$ so that μ induced on $\mathbb{R}^d$ by ϕ is absolutely continuous with respect to the Lebesgue measure. Now isometry implies that

$$d_{\mathcal{M}}(Y_i, Y_j) = |\phi(Y_i) - \phi(Y_j)|,$$

where $d_{\mathcal{M}}$ denotes geodesic distance on $\mathcal{M}$; thus $L^k_{\mathcal{M}}(\mathcal{Y}_n) = L^k_{\mathbb{R}^d}(\phi(\mathcal{Y}_n))$. Since $\phi(\mathcal{Y}_n)$ is a set of n points in $\mathbb{R}^d$ distributed according to a probability measure that is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$, we have

$$L^k_{\mathcal{M}}(\mathcal{Y}_n) = L^k_{\mathbb{R}^d}(\phi(\mathcal{Y}_n)) = n^\alpha \cdot c + o(n^\alpha).$$

However, if two points are close to each other on $\mathcal{M}$, the geodesic distance is well approximated by the Euclidean distance in $\mathbb{R}^m$. Assuming that the k nearest neighbors for any point are so close so that Euclidean distance is a good approximation, we get $L^k_{\mathbb{R}^m}(\mathcal{Y}_n) \approx L^k_{\mathcal{M}}(\mathcal{Y}_n)$ and accordingly

$$L^k_{\mathbb{R}^m}(\mathcal{Y}_n) \approx n^\alpha \cdot c + o(n^\alpha).$$

The restriction that there should exist an isometry $\phi\colon \mathcal{M} \to \mathbb{R}^d$ is rather severe, for most manifolds e.g. the sphere, this is not true. However, when the curvature of a manifold is mild there are diffeomorphisms $\phi\colon \mathcal{M} \to \mathbb{R}^d$ that are close to isometries, i.e. they don't distort distances too much. When looking at a smooth manifold at an increasingly local scale the curvature decreases, this meaning that it is possible to find local near-isometries. With dense enough sampling the k nearest neighbors will fall within a region for which there is a local near-isometry to $\mathbb{R}^d$, which makes the above theory applicable.

6.2 Algorithm

The algorithm given here which is also utilized in chapter 8 is a simplified version of the algorithm presented by Carter et al. in [5], the difference being that Carter et al. used a block bootstrapping method to get subsamples, in order to account for possible data dependencies.

We are given a sample $\{x_1, x_2, \ldots, x_N\}$ of N data points in $\mathbb{R}^m$. Choose the number of neighbors in the k–NN graph, k, and the weighting constant γ as parameters. Since the distance from a point to any of its k nearest neighbors should be a good approximation of the geodesic distances between the points we restrict k to small values ($k \leq 5$). Then choose a sequence of subsample sizes $\{n_i\}_{i=1}^P$ (where obviously $k < n_i \leq N$) and a number M giving how many subsamples that should be drawn for each subsample size.

Now for each n_i, draw M subsamples $\{\mathcal{X}^{(j)}_{n_i}\}_{j=1}^M$ from $\{x_1, x_2, \ldots, x_N\}$. For each subsample construct the k–NN graph and compute the γ–weighted total

length $L^k_{\mathbb{R}^m}(\mathcal{X}^{(j)}_{n_i})$ by equation 6.1. Using a least squares approach, the dimension estimate is defined to be the value of d that minimizes the residual

$$\sum_{i=1}^{P}\sum_{j=1}^{M}|L^k_{\mathbb{R}^m}(\mathcal{X}^{(j)}_{n_i}) - c\cdot n_i^{1-\frac{\gamma}{d}}|^2. \tag{6.2}$$

c is not given, but with d fixed it is easy to see that the value of c that minimizes 6.2 is

$$\hat{c} = \left(\sum_{i=1}^{P}\sum_{j=1}^{M} n_i^{1-\frac{\gamma}{d}} L^k_{\mathbb{R}^m}(\mathcal{X}^{(j)}_{n_i})\right) \Big/ M\sum_{i=1}^{P}(n_i^{1-\frac{\gamma}{d}})^2 \ .$$

Thus for every possible value of d, $(1 \le d \le m)$ we can determine $\hat{c}$ and compute

$$\sum_{i=1}^{P}\sum_{j=1}^{M}|L^k_{\mathbb{R}^m}(\mathcal{X}^{(j)}_{n_i}) - \hat{c}\cdot n_i^{1-\frac{\gamma}{d}}|^2 \tag{6.3}$$

The value of d which minimizes the residual 6.3 is the intrinsic dimension estimate.

Chapter 7

A Novel Approach: Expected Absolute Projection

One of the cornerstones in the theory of normed linear spaces is the triangle inequality,

$$\|x + y\| \leq \|x\| + \|y\|. \tag{7.1}$$

A remarkable fact shown in [14] is that for every finite dimensional normed linear vector space X there is a reversed version of (7.1), namely there is a constant $C \in \mathbb{R}$ such that for any finite subset $\{x_1, x_2, \ldots, x_N\} \subseteq X$

$$\sum_{i=1}^{N} \|x_i\| \leq C \cdot \max_{J \subseteq \{1,2,\ldots N\}} \|\sum_{i \in J} x_i\|. \tag{7.2}$$

Moreover, the value of the smallest possible constant C depends on the dimension of X, hence this allows for an approach to dimension estimation. We will see later that this constant can also be obtained as an expected absolute projection if $X = \mathbb{R}^n$ with Euclidean norm.

There are two major concerns that have to be addressed when adapting this approach: First, we are trying to accomplish dimension estimation for non-linear manifolds, for linear subspaces we already have an optimal approach: PCA. This means that X will not be a normed linear vector space. The remedy is the same as for PCA: to do dimension estimation locally so that the support of the probability measure will be almost linear. If the curvature is not too high it is conceivable that dimension estimation using (7.2) still works, but it is essential to investigate how curvature affects the dimension estimates.

Second, there will be only a finite number of samples from X; had we assumed X was linear this would have been much less of a problem since with centered data linear combinations of samples still would have been in X. With X being non-linear one has to be cautious when using linear combinations of samples, and this restricts severely what subsets of X we have access to. This might lead us to think that the smallest possible C in (7.2) is smaller than it actually is, since we don't find the worst-case sets.

We have addressed the above issues heuristically, with results presented in chapter 8. In this chapter we will describe how to construct estimators of intrinsic dimension based on the triangle inequality.

7.1 Background

For normed linear spaces X, we define

$$C_X = \sup_{T\subseteq X,\, |T|<+\infty} \left\{ \frac{\sum_{x\in T}\|x\|}{\max_{S\subseteq T}\|\sum_{x\in S} x\|} \right\}.$$

This smallest possible constant in the reversed triangle inequality is finite if and only if X has finite dimension [14]. In particular, when $X = \mathbb{R}^n$ and with Euclidean norm, C_X is the quotient between the area of the unit sphere in $\mathbb{R}^n$ and the volume of the unit ball in $\mathbb{R}^{n-1}$, as we shall see. An alternate way to write the denominator is given by the identity,

$$\max_{S\subseteq T}\|\sum_{x\in S} x\| = \max_{\substack{\alpha\in X\\ \|\alpha\|=1}} \sum_{x\in T}(x,\alpha)_+, \tag{7.3}$$

where $(\cdot,\cdot)$ is the scalar product and $(a,b)_+ = \max(0,(a,b))$. For a proof of this identity, see section 7.3.1. We define

$$C_X(T) = \frac{\sum_{x\in T}\|x\|}{\max_{S\subseteq T}\|\sum_{x\in S} x\|}. \tag{7.4}$$

With $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$ denoting the unit sphere in $\mathbb{R}^n$, using (7.3) and the mean value theorem we have that

$$C_X(T) \leq \frac{\sum_{x\in T}\|x\|}{\frac{1}{\mu(\mathbb{S}^{n-1})}\int_{\alpha\in\mathbb{S}^{n-1}}\sum_{x\in T}(x,\alpha)_+\, d\mu(\alpha)},$$

with μ denoting the usual surface measure on $\mathbb{S}^{n-1}$. Due to symmetry we have

$$\int_{\alpha\in\mathbb{S}^{n-1}}\sum_{x\in T}(x,\alpha)_+\, d\mu(\alpha) = \sum_{x\in T}\|x\|\int_{\alpha\in\mathbb{S}^{n-1}}(v,\alpha)_+\, d\mu(\alpha),$$

where $v \in \mathbb{R}^n$ is an arbitrary vector of length 1. Thus

$$C_X(T) \leq \left(\frac{1}{\mu(\mathbb{S}^{n-1})}\int_{\alpha\in\mathbb{S}^{n-1}}(v,\alpha)_+\, d\mu(\alpha)\right)^{-1}.$$

Now let $\{T_N\}_{N\in\mathbb{N}}$ be a family of sets of points on the unit sphere such that T_N is asymptotically evenly distributed over the unit sphere. By this we mean that $\lim_{N\to\infty}\frac{1}{N}\sum_{x\in T_N}(x,\alpha)_+$ is independent of $\alpha \in \mathbb{S}^{n-1}$. Then we have that

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\max_{\alpha\in\mathbb{S}^{n-1}}\sum_{x\in T_N}(x,\alpha)_+ &= \frac{1}{\mu(\mathbb{S}^{n-1})}\int_{\alpha\in\mathbb{S}^{n-1}}\lim_{N\to\infty}\frac{1}{N}\sum_{x\in T_N}(x,\alpha)_+\, d\mu(\alpha)\\ &= \frac{1}{\mu(\mathbb{S}^{n-1})}\int_{\alpha\in\mathbb{S}^{n-1}}(v,\alpha)_+\, d\mu(\alpha)\end{aligned}$$

Thus

$$C_X = \sup_{\substack{T\subseteq X\\ |T|<\infty}} C_X(T) = \left(\frac{1}{\mu(\mathbb{S}^{n-1})}\int_{\alpha\in\mathbb{S}^{n-1}}(v,\alpha)_+\, d\mu(\alpha)\right)^{-1}. \tag{7.5}$$

In section 7.3.2 we use (7.5) to determine the value of C_X for $X = \mathbb{R}^n$, but (7.5) can also be used to construct an estimate of C_X^{-1} for the linear (or almost linear) subspace of smallest dimension a data set lies in. To see this, suppose that A_1 is a random variable with uniform distribution over $\mathbb{S}^{n-1}$. Then $\mathrm{E}[(v, A_1)_+] = C_X^{-1}$. Furthermore, since this is not dependent on v, if A_2 is a random variable independent from A_1 with uniform distribution over $\mathbb{S}^{n-1}$, then $\mathrm{E}[(A_1, A_2)_+] = C_X^{-1}$.

From this we also see that $C_X^{-1} = \frac{1}{2}\mathrm{E}[|(A_1, A_2)|]$, hence the name expected absolute projection.

If we assume that data are distributed uniformly over a ball of dimension d (this can be data cut out from a data set with uniform distribution of dimension d) we can model the data as observations of a number of independent identically distributed random vectors $\{x_0 + V_i\}_{i=1}^N$, where x_0 is the center of the ball and $V_i = R_i \cdot A_i$ with R_i being a scalar uniformly distributed over $[0, 1]$ and A_i having uniform distribution over the unit sphere and R_i is independent from A_i. Then

$$\begin{aligned} \mathrm{E}[(V_i, V_j)_+] &= \mathrm{E}[R_i R_j]\,\mathrm{E}[(A_i, A_j)_+] = \mathrm{E}[R_i R_j]\,C_X^{-1} \\ &= \mathrm{E}[|V_i||V_j|]C_X^{-1} \\ \Rightarrow C_X^{-1} &= \frac{\mathrm{E}[(V_i, V_j)_+]}{\mathrm{E}[|V_i||V_j|]} \end{aligned} \tag{7.6}$$

Thus an estimate of C_X can be done by bootstrapping pairs from a data set and compute the empirical version of (7.6). However, this results in high variance for moderately sized data sets, and therefore this is not the approach we have adopted. Instead we try to construct subsets $T \in X$ such that $C_X(T)$ is close to C_X and thus can be used to estimate dimension.

7.2 Methods

As proved in section 7.3.2, for $X = \mathbb{R}^n$, C_X equals the quotient between the area unit sphere in $\mathbb{R}^n$ and the area of its projection onto $\mathbb{R}^{n-1}$. An explicit formula for C_X is given by

$$C_X = n\sqrt{\pi}\,\frac{\Gamma(1/2 + n/2)}{\Gamma(1 + n/2)}, \tag{7.7}$$

and the values of C_X for n up to 15 is given in table 7.1.

n	C_X	n	C_X	n	C_X
1	2	6	≈ 5.9	11	≈ 8.1
2	π	7	6.4	12	≈ 8.5
3	4	8	≈ 6.8	13	≈ 8.9
4	$\frac{3\pi}{2} \approx 4.7$	9	≈ 7.3	14	≈ 9.2
5	$\frac{16}{3} \approx 5.3$	10	≈ 7.7	15	≈ 9.5

Table 7.1: Values of C_X for $X = \mathbb{R}^n$.

We have constructed an estimator of dimension that tries to compute C_X for the linear or almost linear subspace of smallest dimension that the data lie in.

Then using the formula (7.7) the dimension estimate is found. The dimension estimate n is the smallest integer such that $C_X(T) \leq C_{\mathbb{R}^n}$, where T is a finite set of vectors constructed from our data. There are many ways to choose T, if a large number of various linear combinations of elements in the data are used it seems reasonable that we will get a $C_X(T)$ corresponding to the extrinsic dimension of the data. To get an estimate of the *intrinsic* dimension we can only use few linear combinations. We have considered four options:

1. T consists of all vectors that go from one point in the data set to another.
2. T consists of a subset of the vectors in 1.
3. T consists of the vectors going from the mass center of the data set to the points of the data set, as well as the vectors going in the opposite direction.
4. T consists of the same vectors as in 3., but the midpoints for each pair of points in the data set are added to the data set. There is an option of giving the added points lower weight simply by multiplying the vectors corresponding to these by a scale factor.

To compute $C_X(T)$ we need to find $\max_{S\subseteq T}\|\sum_{x\in S} x\|$. It is not feasible to try all $2^{|T|}$ possibilities for S if T is not very small, so we use the identity (7.3) and search for the maximal projection on vectors of unit length. The sum of the projections of vectors in T onto another vector is differentiable function with respect to the coordinates of the vector, so we can use standard optimization procedures. It should be noted though that the maximal sum of projections gets harder to find as dimension increases.

If PCA is used on T, the first principal component u_1 will have the property that $\sum_{x\in T}|(x,u_1)|^2 = \max_{\alpha\in X,\|\alpha\|=1}\sum_{x\in T}|(x,\alpha)|^2$, but this does not imply that $\sum_{x\in T}|(x,u_1)| = \max_{\alpha\in X,\|\alpha\|=1}\sum_{x\in T}|(x,\alpha)|$. However, Jensen's inequality says that

$$\sum_{x\in T}|(x,\alpha)| \leq \sqrt{|T|}\sqrt{\sum_{x\in T}|(x,\alpha)|^2}.$$

This could in principle be used to estimate C_X when $T = -T$, since

$$\begin{aligned}\sum_{x\in T}\|x\| &\leq C_X \max_{S\subseteq T}\|\sum_{x\in S} x\| = \frac{C_X}{2}\max_{\alpha\in X,\|\alpha\|=1}\sum_{x\in T}|(x,\alpha)| \\ &\leq \frac{C_X\sqrt{|T|}}{2}\max_{\alpha\in X,\|\alpha\|=1}\sqrt{\sum_{x\in T}|(x,\alpha)|^2} \\ &= \frac{C_X\sqrt{|T|}}{2}\sqrt{\sum_{x\in T}|(x,u_1)|^2}\end{aligned}$$

but the inequality turns out to be too weak.

7.3 Proofs

7.3.1 Proof of the identity (7.3).

Proposition. $\max_{S\subseteq T}\|\sum_{x\in S} x\| = \max_{\substack{\alpha\in X\\ \|\alpha\|=1}}\sum_{x\in T}(x,\alpha)_+$.

Proof. $\boxed{\leq}$: For $S \subseteq T$ define $\alpha_S = \sum_{x\in S} x / \|\sum_{x\in S} x\|$. Then

$$\|\sum_{x\in S} x\| = \left(\sum_{x\in S} x, \alpha_S\right) = \sum_{x\in S}(x, \alpha_S) \leq \sum_{x\in T}(x, \alpha_S)_+.$$

$\boxed{\geq}$: For $\alpha \in X$ with $\|\alpha\| = 1$ and a given T, define $S_\alpha = \{x \in T : (x, \alpha) \geq 0\}$. Then

$$\sum_{x\in T}(x, \alpha)_+ = \sum_{x\in S_\alpha}(x, \alpha) = \left(\sum_{x\in S_\alpha} x, \alpha\right) \leq \|\sum_{x\in S_\alpha} x\|.$$

□

7.3.2 Determining C_X for $X = \mathbb{R}^n$ with Euclidean norm

We consider $X = \mathbb{R}^n$ with Euclidean norm. From (7.5) we have that

$$C_X = \left(\frac{1}{\mu(\mathbb{S}^{n-1})} \int_{\alpha\in\mathbb{S}^{n-1}} (v, \alpha)_+ \, d\mu(\alpha)\right)^{-1},$$

where v is a vector of unit length and $\mathbb{S}^{n-1}$ is the $(n-1)$–sphere. Let π_v denote hyper plane of dimension $n-1$ with v as its normal. Let $\mathbb{S}_+^{n-1}$ denote the part of the unit sphere which is on the same side of π_v as v. Then

$$C_X = \left(\frac{1}{\mu(\mathbb{S}^{n-1})} \int_{\alpha\in\mathbb{S}_+^{n-1}} (v, \alpha) \, d\mu(\alpha)\right)^{-1}.$$

Now fix $\alpha \in \mathbb{S}_+^{n-1}, \alpha \neq v$. Let π_α denote the hyper plane of dimension $n-1$ which has α as its normal. Note that (v, α) is the length of the projection of α onto v. We will show that (v, α) also equals the area scale factor for projection of an area element in π_α onto π_v.

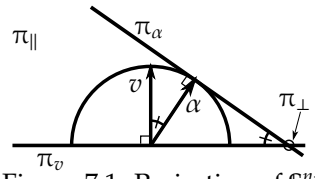


Figure 7.1: Projection of $\mathbb{S}_+^{n-1}$ onto $\pi_\parallel$.

Denote the plane spanned by v and α by $\pi_\parallel$. The intersection between π_v and π_α is orthogonal to $\pi_\parallel$, denote this $(n-2)$–dimensional hyper plane by $\pi_\perp$. We construct an orthonormal basis for π_α, $\{a_1, a_2, \ldots a_{n-1}\}$, such that a_1 lies in $\pi_\parallel$ and $a_2, \ldots, a_{n-1}$ are parallel to $\pi_\perp$. What happens when an area element $da_1 da_2 \ldots da_{n-1}$ in π_α is projected onto π_v? Before projection the area is $|da_1 da_2 \ldots da_{n-1}| = |da_1||da_2| \ldots |da_{n-1}|$. Denote the projection of a_i onto π_v by $\hat{a}_i$. Since $a_2, \ldots, a_{n-1}$ are parallel to $\pi_\perp$, which is a subset of π_v, $a_i = \hat{a}_i$ for $i = 2, \ldots, n-1$. $\hat{a}_1$ will be orthogonal to $\hat{a}_2, \ldots, \hat{a}_{n-1}$, since $\hat{a}_1$ is in $\pi_\parallel$. Moreover, $|d\hat{a}_1| = (v, \alpha)|da_1|$ by similarity, see figure 7.3.2. Thus

$$|d\hat{a}_1 d\hat{a}_2 \ldots d\hat{a}_{n-1}| = |d\hat{a}_1||d\hat{a}_2| \ldots |d\hat{a}_{n-1}| = (v, \alpha)|da_1 da_2 \ldots da_{n-1}|,$$

i.e. (v, α) is the area scale factor when projecting π_α on π_v. Thus

$$C_X^{-1} = \frac{1}{\mu(\mathbb{S}^{n-1})} \int_{\alpha\in\mathbb{S}_+^{n-1}} (v, \alpha) \, d\mu(\alpha) = \frac{1}{\mu(\mathbb{S}^{n-1})} \mu(\mathbb{B}^{n-1})$$

where $\mathbb{B}^{n-1}$ is the unit ball in $\mathbb{R}^{n-1}$, which can be embedded in $\mathbb{R}^n$ as the projection of $\mathbb{S}_+^{n-1}$ on π_v.

Chapter 8

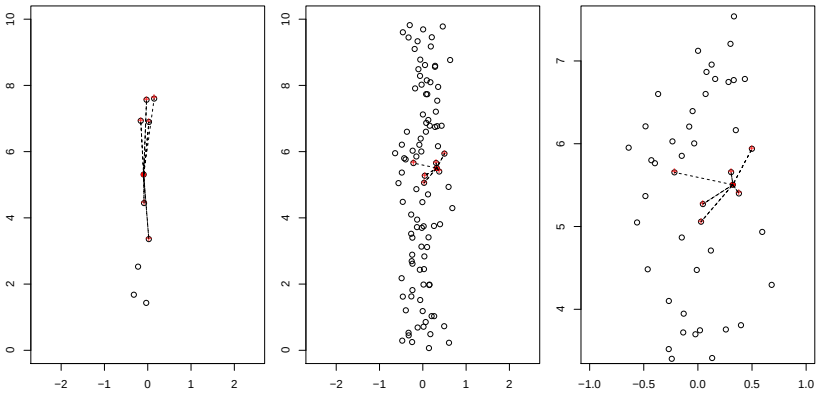
Method Comparisons: Local Dimension Estimation

When we find data lying on a manifold with lower dimension than the extrinsic dimension, there are continuous functional relationships between the variables. If the functions describing these relations are non-linear, the manifolds will be non-linear. However, for example in gene expression data, it is likely that we have continuous functional relations between variables that are not globally valid; there might be groups of samples with different functional relations. This means that the support of the data is a union of manifolds, possibly with different dimension. Another type of data where this is the case is images. Regions with high complexity have high intrinsic dimension, and regions with low complexity have low intrinsic dimension; this has been exploited to segment out regions of different complexity [5].

Motivated by this we will apply the dimension estimators presented in earlier chapters to local dimension estimation problems. Local PCA and the EAP estimator are naturally estimators of local dimension, but the other methods estimate more naturally global dimension. Most global dimension estimators can perform local dimension estimation by considering subsets of data, the single requirement is that the dimension estimator must be applicable to data sets with relatively few points. This does not mean that the estimator should be able to estimate dimension of manifolds which are very sparsely sampled in relation to their curvature, this is of course impossible, but it means that when the curvature is low it should suffice with a small number of points to do dimension estimation. This rules out vector quantization as a method for local dimension estimation, since already for a few hundreds of points sampled from a flat manifold its performance is poor, see appendix A.

On the other hand, local dimension estimators can always be used to estimate global dimension by averaging over the local estimates.

A caveat to local dimension estimation of manifolds with high-dimensional noise is that if we use a subset with little expansion we get the dimension of the noise. This is especially a problem if we fix the number of points we use for local dimension estimation, as can be seen in figure 8.1. Therefore when using a fixed number of points for the subsets used for local dimension estimation, one has to estimate the average diameter of these subsets and make sure that it



(a) 10 samples (b) 100 samples (c) Enlargement

Figure 8.1: Samples from the distribution $N(0, 0.3) \times U(0, 10)$, with a point and its six nearest neighbors marked.

is large enough in comparison to the standard deviation of the noise.

We have considered manifolds which have very low intrinsic dimension (≤ 15) compared to the extrinsic dimension (i.e. the number of variables) of typical genomics data (≥ 10000). However, the intrinsic dimension must be smaller than the number of data points, which are considerably fewer than the number of variables. Furthermore, principal components analyses often show that most of the variance in gene expression data can be explained by a handful principal components (see e.g. [36]), indicating that a model where the data lie close to a manifold with a dimension at least less than one hundred is reasonable.

The extrinsic dimensions of the manifolds we have considered here are also low, since with higher extrinsic dimension the computational cost increases. The investigation should be considered as an attempt to characterize the dimension estimators as to guide further research.

8.1 Simulated data sets

We use six categories of synthetic data sets to test the dimension estimators. Each test set is designed so that it is the restriction of samples uniformly distributed over a manifold (with noise added in one case) to a neighborhood of a point on the manifold; in some cases this point is one of the samples. The manifolds behind each of the data set categories are: 1) $\mathbb{R}^n$ with n ranging from 3 to 10. 2) The 5–sphere. 3) The n–sphere with $n = 2, 4, 6, 8$. 4) The faces of the $n + 1$–dimensional hypercube with $n = 2, 4, 6, 8$. 5) "Edges" of dimension 2 to 9 of the 10–dimensional hypercube. 6) $\mathbb{R}^5$. For the data sets in category 6), normal noise of varying dimension is added.

In category 1), which is only used to test the EAP estimator, we vary the number of samples in the neighborhood between 25 and 70. Otherwise we use always about 50 samples, in fact exactly 50 samples except in category 2). The

reason that the number of samples in the neighborhood in category 2) is not fixed, is that we fix the number of samples over the whole 5–sphere and we use a cut-off radius r to determine the neighborhood (i.e. all points closer than r to the top of the sphere are inside the neighborhood). The number of samples over the 5–sphere is determined so that approximately 50 points fall within the cut-off radius. In category 3) and 4) we also fix the total number of samples over the manifold, but the neighborhood is chosen as a randomly selected point and its 49 nearest neighbors.

In each category of data sets we vary one or two features, e.g. dimension, and for each value we have simulated 100 data sets for which we have carried out dimension estimation.

The details and the results are presented in sections 8.3–8.7. We will first give some details on the implementation of the dimension estimators.

8.2 Parameter and Design Choices

There are no clear-cut ways to define which parameter values are optimal. In general we will have a tradeoff between bias and variance, one evident example of this is shown in appendix B, where we have used Takens' estimator with different values for the cut-off radius to estimate the dimension of the data sets in category 2), which is also estimated in figure 8.3. We have strived to balance bias and variance in our choice of parameters, but we do not argue that our results are optimal.

Local PCA

When we use PCA for dimension estimation, the estimate is taken to be the number of eigenvalues of the covariance matrix that are at least 5% of the biggest eigenvalue; this is a common tolerance level [3, 15]. We can apply PCA directly on the test data sets to get local PCA by way of construction of the data sets, however for the problems we consider here we can produce better results with PCA if fewer points are used for the local estimates. Thus we have made local PCA dimension estimates for each point in the data set based on its $k = 10$ or 20 nearest neighbors and then taken the median of the dimension estimates to be the dimension estimate for the data set. In figure 8.3 we show results both for PCA applied the whole data set and for the median of the local PCA estimates in each point (PCA_{loc}).

Takens' and the Hill Estimator

For the Hill estimator we need to decide how many interpoint distances k to use (M in (4.4)), and for Takens' estimator we need to decide the cut-off radius r (δ_0 in (4.3)). The fewer interpoint distances we use, the better the Euclidean distances approximate the true geodesic distances, but when we use fewer distances we get higher variance. We have chosen $k = 50$ for the Hill estimator so that on average one distance per point is used (the distance to the nearest neighbor). For Takens' estimator we have chosen $r = \mu_1 + \sigma_1$ where μ_1 is the average distance to the nearest neighbor for the points and σ_1 is the standard deviation of this distance.

We have used the unbiased version of (4.4) for the Hill estimator.

There is also the possibility of using intrinsically local versions of Takens' and the Hill estimator, that is maximum likelihood estimates of the local dimension $\dim_{loc}$ as defined in definition 2.7.This approach is utilized in [30] and [5], but in our tests this approach proved to be suboptimal, even if we take the median of dimension estimates of subsets of the data set as for local PCA.

k–NN estimator

For the k–NN estimator we have four parameters to set: the number of neighbors in the k–NN graph, k; the weighting constant γ; the sample sizes n_i; and the number of bootstrapped samples N. We have throughout all simulations chosen $N = 10$. As for choosing k we have the same considerations of how well the nearest neighbor distances approximate the geodesic distances as for the Hill and Takens' estimator, but here the effect of choosing large k will be more severe since we will consider the k nearest neighbors in a subsample of the set for which we make the dimension estimation. This also means that the sample sizes have to be large enough so that the k:th nearest neighbor is not too far away. For the dimension estimations presented here we have chosen $k = 2$, and $n_1 = \lceil K/2 \rceil$, $n_2 = \lceil K/2 \rceil + 3$, $\ldots$, $K - 4$, $K - 1$, where K is the number of points in the subset where we make the local dimension estimation (usually $K = 50$). γ is chosen to be 1, since we have not detected much difference of the dimension estimates when γ is varied.

EAP Dimension Estimator

For the EAP dimension estimator we need to decide upon a method to construct the set T. To get the results presented here we have let T be the set of all vectors between points in the data set, the reason for this is discussed in the next section. Apart from this there are no parameters to set, disregarding that there are many ways to do the maximization

$$\max_{\substack{\alpha \in X \\ \|\alpha\|=1}} \sum_{x \in T} (x, \alpha)_+ \quad .$$

How we have chosen to do this is discussed in appendix B.

8.3 EAP Dimension Estimator: Initial Tests

We begin with investigating the fundamental properties of the EAP dimension estimator: how well does it estimate dimension of data sets from "ideal" probability measures? The "ideal" probability measures are the uniform ones, since we saw in section 7.1 that data evenly distributed over a sphere yields the maximal value for $C_X(T)$. It follows that this is also the case for data evenly distributed over a ball, and local dimension estimation means estimating the dimension of the probability measure restricted to a ball.

We have sampled data sets of varying size from the uniform distribution over the unit ball in $\mathbb{R}^n$ with n ranging from 3 to 10, and applied various versions of the EAP dimension estimator. The version with the best results (with the

least bias), which still is reasonably fast, is the one where we let T consist of all vectors that go from one point in the data set to another. The results for this version are displayed in figure 8.2. We could achieve somewhat better results by instead letting an algorithm choose a subset of these vectors in order to maximize $C_X(T)$ but this approach is very computationally intensive and has therefore not been tested extensively.

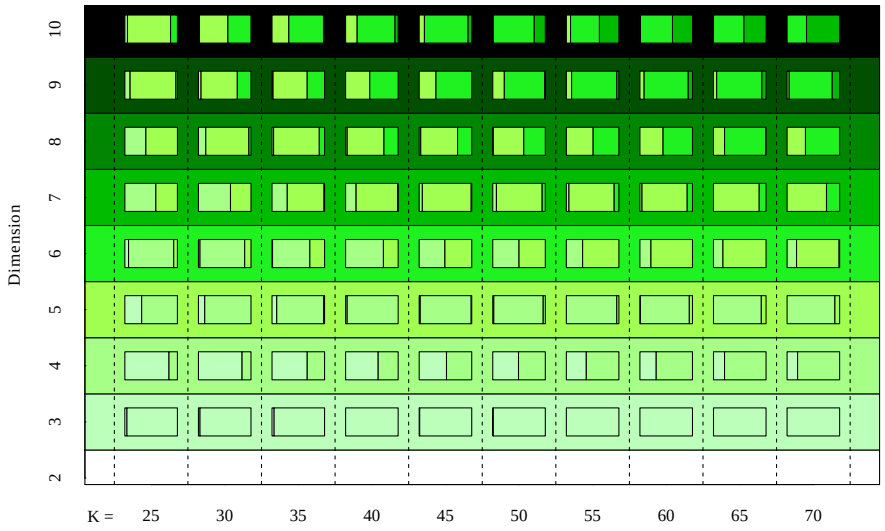


Figure 8.2: Results of the EAP dimension estimator using all vectors between points in the data set, applied to uniformly sampled balls with K samples and varying dimension. The bars in each cell show the proportion of the data sets for which a certain dimension estimate was yielded.

We see that the EAP dimension estimator systematically underestimates dimension. This is not surprising since with a finite number of samples we cannot expect to get the worst constant in the reversed triangle inequality. It is more surprising that adding convex combinations of data points does not increase the values of the dimension estimates. In appendix B the corresponding plot to figure 8.2 is presented when we use the vectors to the center of the ball from the data points and the midpoints between each pair of data points, with the midpoints given lower weight so that the influence of all the midpoint vectors equal that of the original data point vectors. The dimension estimates are in general lower, this is probably due to the fact that we introduce more data dependencies when adding midpoints.

8.4 Curvature

Now we will compare the EAP dimension estimator with the four other methods of local dimension estimation: Local PCA, Takens' estimator, the Hill estimator and the k–NN estimator. First we investigate the effect of curvature on the dimension estimators; in view of the theory presented in earlier chapters this an essential test, since with curved d–manifolds we do not have isometric embeddings to $\mathbb{R}^d$.

Strictly, it is not the curvature itself that has an adverse effect on dimension estimators, since the data can be scaled so that the curvature changes without changing the dimension estimates. What affects the dimension estimates is the total amount of bending that appears in a data set, i.e. the largest angle between tangent spaces (or normals to tangent spaces) for points in the data set.

We have applied the dimension estimators to uniform probability distributions over top segments of 5–spheres. Specifically, we have a number of data sets each of which consists of approximately 50 points, these points are the points within a certain cut-off radius r from the top of the sphere. The total number of samples on the sphere, M, varies so that the data sets cover different proportions of the sphere. To characterize the total amount of bending in the data set we define θ to be the angle between the tangent plane on the top of the sphere and the tangent plane at distance r from the top. θ can be computed from

$$\theta = \arccos(1 - \frac{r^2}{2R^2}),$$

where R is the radius of the sphere. The results are presented in figure 8.3, where we also see how the dimension estimators perform when there is zero curvature (the manifold is a unit ball of dimension 5).

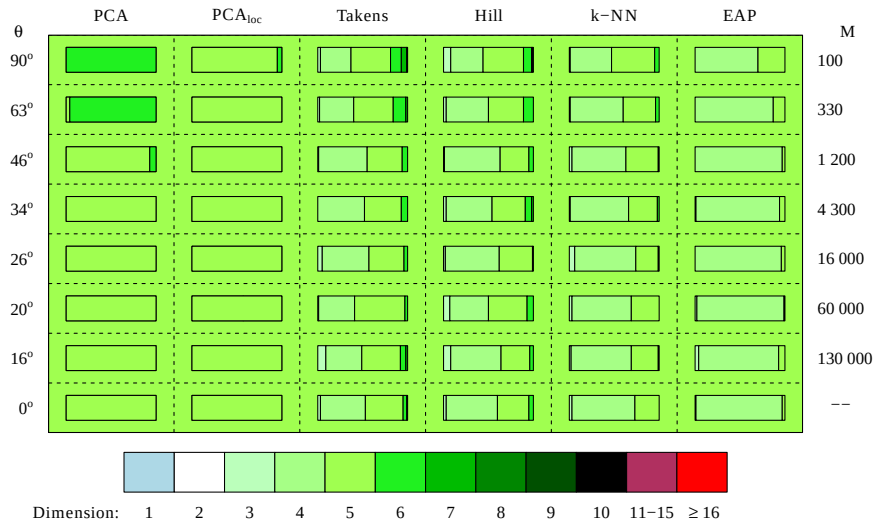


Figure 8.3: Dimension estimation results for sphere segments with $K \approx 50$ points. Parameters: Local PCA: $k = 10$; Takens' estimator: $r = \mu_1 + \sigma_1$; Hill: $k = 50$; k–NN: $k = 2$, $\gamma = 1$, 10 bootstrapped samples per sample size, with sample sizes $\lceil K/2 \rceil, \lceil K/2 \rceil + 3, \ldots, K - 4, K - 1$.

With a tolerance of 5% we see that local PCA outperforms all the other estimators, however we can see on the PCA estimates in the first column (PCA applied on the whole set, see section 8.2) that there is a sharp dependence on the curvature as expected. PCA manages to give the correct dimension when $\theta \lesssim 46°$, but fails when $\theta = 63°$. Considering that if we project a top segment of a sphere with a θ as defined above onto the the tangent plane at the top, the radius of the projection will be $R \sin\theta$, whereas if we project on the normal of the tangent plane the length of the projection will be $R(1 - \cos\theta)$, this is

reasonable since

$$\left(\frac{1-\cos 45^\circ}{\sin 45^\circ}\right)^2 \approx 0.043 \quad \text{and} \quad \left(\frac{1-\cos 63^\circ}{\sin 63^\circ}\right)^2 \approx 0.38 \ .$$

(Remember that we use the eigenvalues of the covariance matrix.)

For the other estimators, dependence on the curvature is not so clear, but they tend to underestimate the dimension, especially the EAP dimension estimator.

8.5 n-Spheres and Hyper Cube Faces

We compare dimension estimates of uniformly sampled hyper cube faces (the border of a hyper cube) with dimension estimates of uniformly sampled n-spheres with $n = 2, 4, 6, 8$, to see how well the estimators perform on manifolds of different dimension, and if the sharp edges between the hyper cube faces affect the dimension estimates. The results are presented in figure 8.4, and we can conclude that in general the sharp edges make no difference. It seems that the two estimators that are most affected though are local PCA and the EAP estimator, which is reasonable since they depend on tangent planes.

For the n–spheres local PCA gives the best results, for the others none is significantly better than the others, but the EAP dimension estimator gives the worst results.

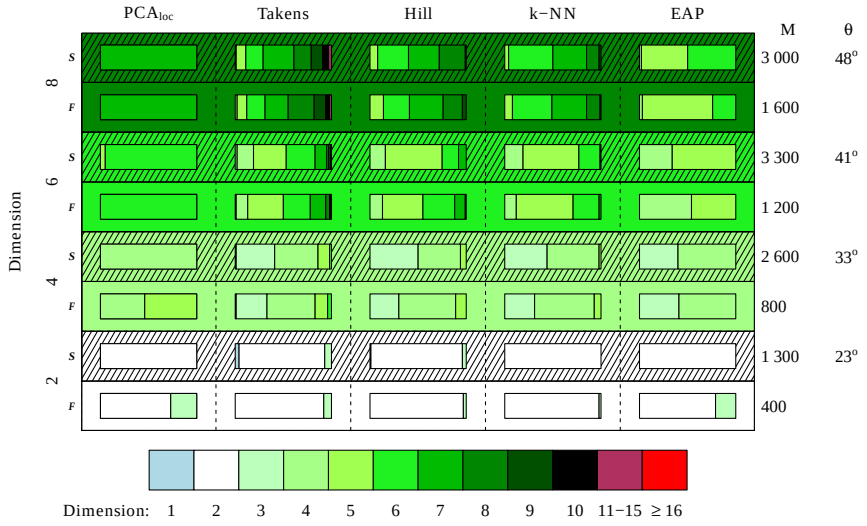


Figure 8.4: Dimension estimation results for n-spheres and hyper cube faces of varying dimension. Parameters: Local PCA: $k = 10$; Takens' estimator: $r = \mu_1 + \sigma_1$; Hill: $k = 50$; k–NN: $k = 2$, $\gamma = 1$, 10 bootstrapped samples per sample size, with sample sizes $25, 28, \ldots, 46, 49$.

8.6 Hyper Cube Edges

The manifolds considered so far have had an extrinsic dimension of one more than the intrinsic dimension. To really verify that the dimension estimators estimate intrinsic dimension and not just underestimate extrinsic dimension we need to test them on manifolds with high extrinsic dimension. Therefore we have constructed a manifold which we call hyper cube edges, since it is a generalization of the edges of a 3-dimensional cube. The hyper cube edges with codimension one (extrinsic dimension one more than the intrinsic dimension) are the faces of the hyper cube. The hyper cube edges of codimension 2 are the union of the borders of each face. The hyper cube edges of codimension 3 are the borders of each linear piece in the hyper cube edges of codimension 2 and so on.

We have generated hyper cube edges with extrinsic dimension 10 and intrinsic dimension varying between 2 and 9. The results are presented in figure 8.5.

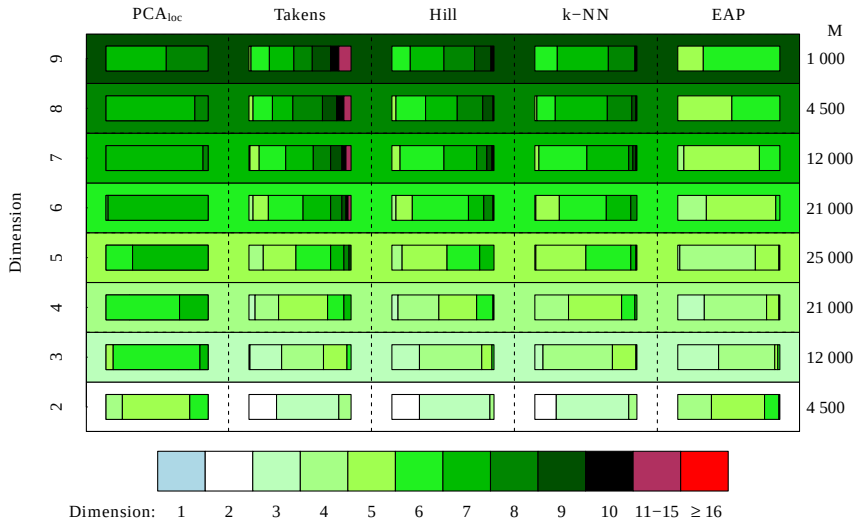


Figure 8.5: Dimension estimation results for hyper cube edges of varying intrinsic dimension. The density of the points is kept constant at 100 data points per area unit. Parameters: Local PCA: $k = 10$; Takens' estimator: $r = \mu_1 + \sigma_1$; Hill: $k = 50$; k–NN: $k = 2$, $\gamma = 1$, 10 bootstrapped samples per sample size, with sample sizes $25, 28, \ldots, 46, 49$.

We see that now there is a tendency to overestimate the dimension instead, at least for the two lowest dimensions. This goes against the general notion that dimension estimation algorithms always give a negative bias, expressed in [5] as

> To our knowledge, a phenomenon common to all algorithms of intrinsic dimension estimation is a negative bias in the dimension estimate. It is believed that this is an effect of undersampling the high-dimensional manifold.

We see here that the bias in the dimension estimate depends on the extrinsic dimension of the manifold. And in fact we have constructed an estimator of dimension from equation (7.6) which has positive bias in general, but due its

high variance we have not considered it further. It is an interesting problem to investigate where the bias really comes from for the Hill, Takens and k–NN estimator.

But it is clear that the dimension estimators measure intrinsic dimension and not the extrinsic dimension, since the dimension estimates in general increases when the dimension increases. There is one exception to this and that is the EAP dimension estimators estimates in dimension 2. We have not investigated why this is the case, but one speculation is that it is caused by that in each corner or the manifold it branches out in very many directions and the lower the intrinsic dimension, the more branches.

Another interesting thing to notice is how bad the discrimination between the manifolds of higher dimension is (the dimension estimates are very similar for $n = 8$ and 9). By looking at the diagram it seems that the EAP dimension estimator actually is slightly better than the others at doing this, but it needs further investigation.

Local PCA gives here the worst results, for $n = 3, \ldots, 9$, almost all dimension estimates are either 6 or 7. It should be noted though that since $k = 10$, the maximal possible dimension estimate is 10.

8.7 High-Dimensional Noise

Finally we will consider noisy manifolds; we will consider the simplest case — a flat uniformly sampled manifold with normal noise added to it, a generalization to higher dimensions of the manifold depicted in figure 8.1. We have used a five-dimensional manifold and added noise of varying dimension d. As before, local dimension estimation is done by cutting out a ball. Thinking of figure 8.1, we see that if the radius of the ball is small enough in comparison to the standard deviation of the noise, the local dimension estimate should be the dimension of the manifold plus the dimension of the noise.

We have used normal noise orthogonal to the manifold with a standard deviation σ in each direction, with varying σ and dimension. The radius of the ball defining the local data is kept constant. To get an idea of what the distribution of the data looks like at a local level for varying σ, we have applied PCA to it. The eigenvalues of the principal components, i.e. the variance we get if we project the data onto each principal component is shown in figure 8.6.

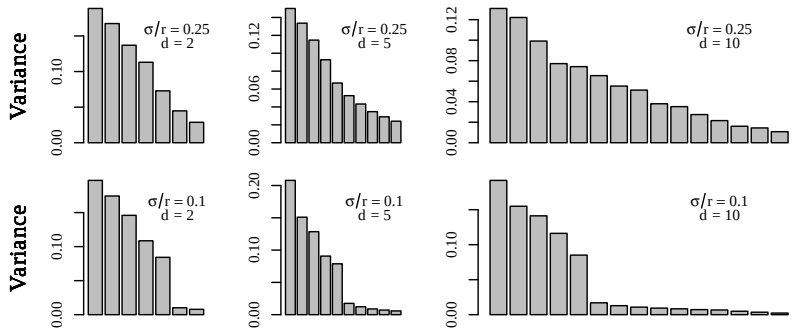


Figure 8.6: The eigenvalues of the principal components for 50 samples inside the unit ball, drawn from the uniform distribution on the 5-dimensional hyper plane with orthogonal normal noise of dimension d and variance σ^2 added.

For data sampled from an isotropic distribution the eigenvalues will decrease continuously, much like in the top diagrams in figure 8.6. However, if the data lie close to a linear subspace of dimension d, i.e. most of the variance in the data is kept when we project onto this subspace, the d first eigenvalues will be substantially higher than the rest, as in the bottom diagrams in figure 8.6.

This means that an accurate dimension estimator should yield 5 as the dimension estimate if $\sigma/r \leq 0.1$ and $5 + d$ if $\sigma/r = 0.25$. The results from the dimension estimators that we have tested is shown in figure 8.7; we have used 50 samples for each manifold.

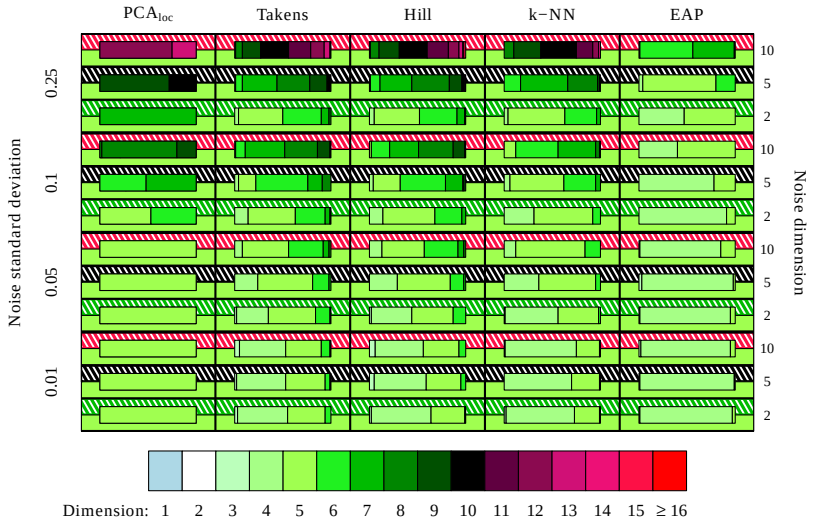


Figure 8.7: Local dimension estimation with 50 samples and cut-off radius $r = 1$ for the five-dimensional hyper plane with orthogonal normal noise of dimension d and standard deviation σ. Parameters: Local PCA: $k = 20$; Takens' estimator: $r = \mu_1 + \sigma_1$; Hill: $k = 50$; k–NN: $k = 2$, $\gamma = 1$, 10 bootstrapped samples per sample size, with sample sizes $25, 28, \ldots, 46, 49$.

Taking the ubiquitous bias into account, the EAP dimension estimator performs remarkably well. It is especially interesting to see that with a standard deviation of 0.1 and a noise dimension of 10, all the other estimators give a much higher dimension estimate than with a standard deviation of 0.25 and noise dimension 2, which is incorrect. However, comparing with figure 8.2 we see that the dimension estimates at $\sigma = 0.25$, $d = 10$ are much more biased than the estimates of the uniformly sampled unit ball with 50 samples. This is probably due to the fact that the distribution is non-uniform, the EAP dimension estimator is optimal for uniform distributions.

Otherwise local PCA gives the best estimates in general, with little variance and highest estimates when $\sigma = 0.25$. However, it is more sensitive to the moderate noise, i.e. $\sigma = 0.1$, than the others, especially the k–NN estimator. The more biased estimators, the k–NN estimator and the EAP estimator, are less sensitive to moderate noise.

Chapter 9

Conclusions

We have introduced a novel intrinsic dimension estimator, the EAP (Expected Absolute Projection) estimator for data sets supported on smooth manifolds. The estimator is based on a constant C_X defined for each $X = \mathbb{R}^n$ with a given norm, which is monotonically increasing in n. The constant C_X can be interpreted as the minimal constant in the reversed triangle inequality, but we also have that if $X = \mathbb{R}^n$ with Euclidean norm, $(C_X/2)^{-1}$ is the expected absolute projection of a random vector from the unit sphere onto any given vector on the unit sphere if the distribution is uniform, hence the name.

The EAP estimator carries out local dimension estimation, i.e. we restrict the attention to a local data set defined as the data points within a cut-off radius from a certain point. If the manifold is sufficiently smooth so that the probability measure from which the data are sampled is well approximated by a tangent space $\mathbb{R}^d$ within the cut-off radius, and the probability measure is sufficiently uniform, then as long as we have sufficiently many points we can from the data points construct a set T for which the empirical constant $C_X(T)$ defined by (7.4) is close enough to $C_{\mathbb{R}^d}$. The dimension estimate is given by the minimal d such that $C_X(T) \leq C_{\mathbb{R}^d}$.

By construction of the estimator, it will have a negative bias as long as the manifold on which the probability measure is supported is not very curved. This was also seen clearly in the experiments.

We compared the EAP estimator on local dimension estimation problems with four dimension estimators from the literature: local PCA, Takens' estimator, the Hill estimator, and the k–NN estimator. In general the EAP estimator showed much more bias than the other estimators, but it had lower variance. On n–spheres, local PCA was the best estimator, but it was also the most sensitive to sharp edges and its performance on the manifolds with high extrinsic dimension was poor. It was hard to distinguish Takens' estimator, the Hill estimator and the k–NN estimator from each other. For the parameters we used, Takens' estimator had the most variance and the least bias and the k–NN estimator had the least variance and most bias of the three. However this will be different for with other parameters, even though the estimates of the k–NN estimator were not so sensitive to changes in parameters.

When we applied the dimension estimators to a flat uniformly sampled manifold with normal noise we saw that both the dimension and the standard deviation of the noise affected the dimension estimates, especially for moderate

noise. But all the estimators had good results when the standard deviation of the noise was at most 5% of the radius of the sphere delimiting the neighborhood used for dimension estimation. The EAP dimension estimator was the least sensitive to noise, partly because of its strong bias, but probably also because its bias is bigger for non-uniform distributions.

From our results we conclude that so far the EAP estimator is not very competitive; but there are multiple ways in which one might improve it and alleviate the bias.

One disadvantage with the EAP estimator is that it requires relatively many points to give a reasonable estimate. If we can find a better way to choose T from the data points, we might get a smaller bias and also reduce the number of points needed. As we noted in chapter 7, if we use more linear combinations of the points the dimension estimate should increase. We can see in figure B.2 that adding the midpoints does not lead to this, but there might be other linear combinations for which we can achieve it. Another possible way to improve the EAP estimator is to use a different norm than the Euclidean norm.

Finally we want to emphasize the importance of estimating the impact of noise on estimators of intrinsic dimension. In all real-world data sets there will be noise, probably both from experimental noise and from variance in variables not important enough to include in the model. The vector quantization approach is very appealing in that there is a theoretical bound for the impact of noise, but the fact that it requires very many data points is prohibitive. There are also methods to reduce the impact of noise which one might apply first. Vector quantization is actually one such method, deconvolution is another [27].

Appendix A

Quantization Dimension Estimation

We have attempted to estimate the quantization dimension of data sets consisting of 200 samples drawn from the uniform probability distribution on the unit cubes of dimension 2, 3, 4 and 5.

To estimate the quantization dimension we first partition our data into test sets and training sets. Following [35] we divide our samples into four quarters and use each quarter in turn as training data and the rest as test data. For each $k < 50$ we use the k-means algorithm to construct an optimal quantizer for each set of training data, $Q_k^{(i)}$, $i = 1, \ldots 4$. The estimate of the second order error for the optimal quantizer with k points is then taken to be

$$\hat{e_2^*}(k) = \frac{1}{4} \sum e_2(Q_i^{(i)} | \mu_i),$$

where μ_i is the empirical measure of test data set i. Recall that the quantization dimension of order 2 of the probability measure μ is defined as

$$\dim_{quant}^{(2)}(\mu) = -\lim_{k \to \infty} \frac{\log k}{\log e_2^*(k|\mu)} \quad .$$

Based on a heuristic argument, it is stated in [35] that the first minimum in the graph of

$$-\frac{\log k}{\log \hat{e_2^*}(k)}$$

is a good estimate of the dimension in the manifold. As can be seen in figure A.1, this is not the case when we use only 200 points. We conclude that vector quantization as a method for dimension estimation is only applicable if we have considerably more data points.

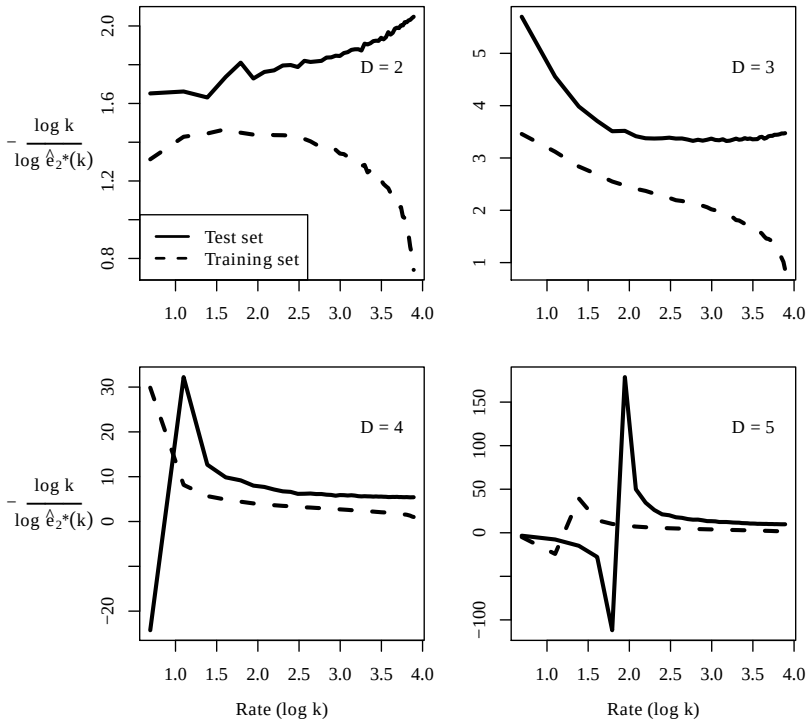


Figure A.1: Test and training curves of quantization error e_r for four data sets consisting of 200 points uniformly sampled on unit cubes of varying dimension.

Appendix B

Additional Details About the Design of the Estimators

B.1 Maximal projection

To compute

$$\max_{\substack{\alpha \in X \\ \|\alpha\|=1}} \sum_{x \in T} (x, \alpha)_+ \ ,$$

we use the `R` function `optim` with method `'BFGS'`. This means that a quasi-Newton method is used for the optimization, for details of the algorithm see the `R` documentation. To test the performance of the algorithm we applied it one hundred times with random initial values to a number of sets consisting of all the vectors between 20 points in $\mathbb{R}^4$ uniformly sampled in $[-1,1] \times [-1,1] \times \{0\} \times \{0\}$. We also computed $\sum_{x \in T} (x, \alpha)_+$ for $\alpha = (\cos\theta, \sin\theta, 0, 0)$, with $\theta \in [0, \pi]$. In most of the cases there was a single maximum which was found in all 100 cases, but sometimes we got a result as shown in figure B.1. Based on the observations from this we decided to apply the `optim` function 10 times with randomly selected initial values and take the maximal value when doing the dimension estimation by EAP.

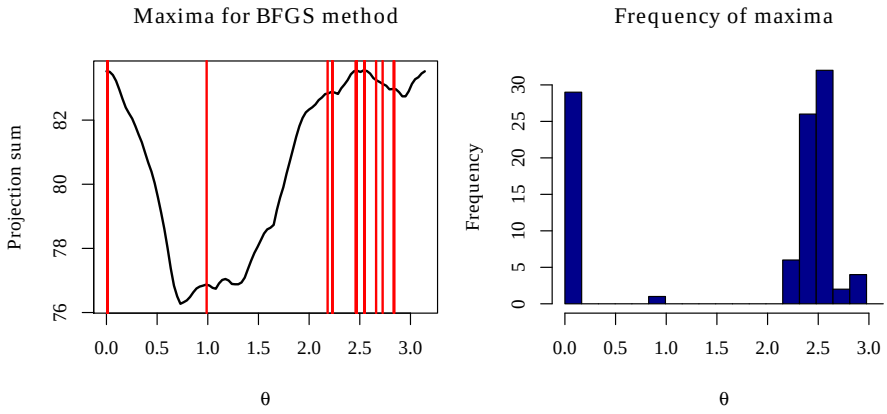


Figure B.1: Results of optimization by BFGS method.

B.2 Choice of T

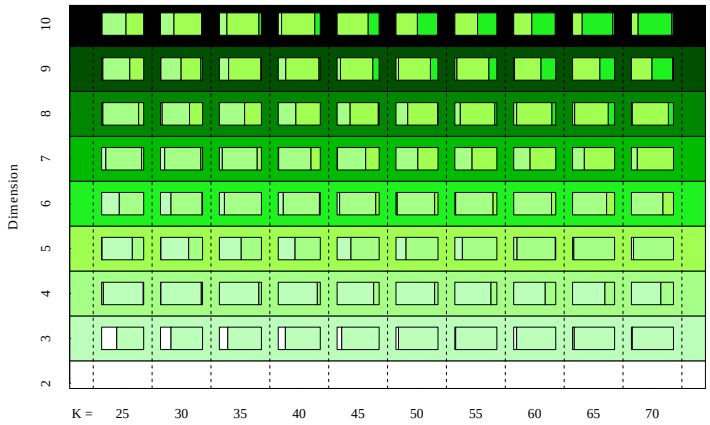


Figure B.2: Results of novel dimension estimator using the vectors to the center of the data set from the data points and the midpoints between data points, with midpoints given lower weight, applied to uniformly sampled balls with K samples and varying dimension.

B.3 Varying cut-off radius r for Takens' estimator

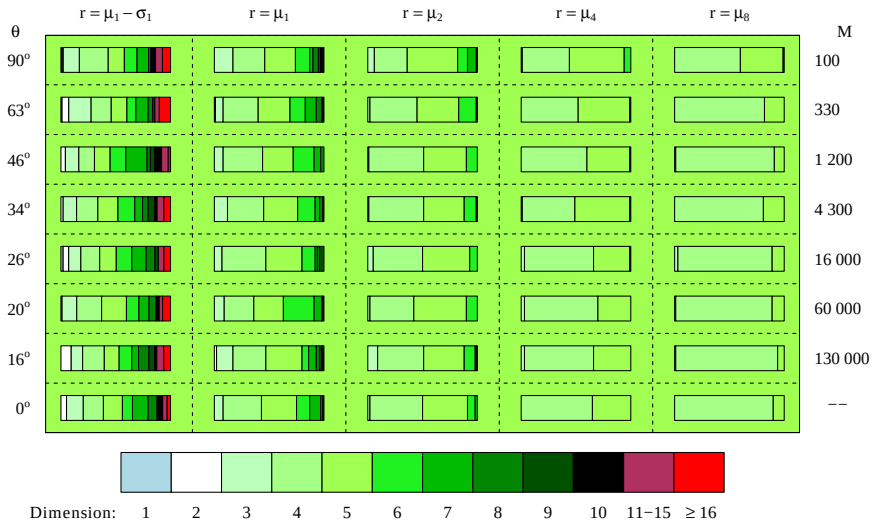


Figure B.3: Results from dimension estimation of top segments of 5–spheres by Takens estimator with varying cut-off radius (cf. figure 8.3).

Appendix C

General Results

Lemma C.1. *If μ is a Borel measure on $\mathbb{R}^n$, all Borel sets in $\mathbb{R}^{2n}$ are measurable for the product measure $\mu \times \mu$.*

Proof. It is sufficient to show that every open set in $\mathbb{R}^{2n}$ belongs to the σ–algebra generated by the semi-ring $\mathcal{P} = \{E \times F : E \text{ and } F \text{ open}, \ E, F \in \mathbb{R}^n\}$, i.e. that each open set in $\mathbb{R}^{2n}$ can obtained by countable unions and finite differences of elements in $\mathcal{P}$. Let $Q_\epsilon(x) \subseteq \mathbb{R}^{2n}$ denote the open hyper cube with side 2ϵ centered at x. Clearly $Q_\epsilon(x) \in \mathcal{P}$ for each $x \in \mathbb{R}^{2n}$. Furthermore, $|x - y| < \epsilon\sqrt{2n}$ for any $y \in Q_\epsilon(x)$. Now for a given open set $U \in \mathbb{R}^{2n}$, define for each $x \in U$

$$\epsilon_x = \inf\{|x - y| : y \in U^c\} \,/\, \sqrt{2n} \ .$$

Clearly $Q_{\epsilon_x}(x) \subseteq U$ for each $x \in U$. But in fact, since $\mathbb{Q} \cap U$ is dense in U, $U = \bigcup_{x \in \mathbb{Q} \cap U} Q_{\epsilon_x}(x)$ and we are done. □

Lemma C.2. *If $\Pr[|f(x) - f(y)| > \epsilon] < \kappa$ with $x, y \sim \mu$ for some $\epsilon, \kappa > 0$, then with m denoting the median, $\mathrm{E}[|f(x) - m|] \leq \sup|f(x) - m| \cdot 2\kappa + \epsilon$.*

Proof. We have that with $x, y \sim \mu$ and m being the median of $f(x)$,

$$\begin{aligned}
\Pr[|f(x) - f(y)| > \epsilon] &\geq \frac{1}{2}\Pr\left[|f(x) - f(y)| > \epsilon \;\middle|\; \begin{array}{l} \mathrm{sign}(f(x) - m) \\ = \ -\mathrm{sign}(f(y) - m) \end{array} \right] \\
&= \frac{1}{2}\Pr\left[|f(x) - m| + |f(y) - m| > \epsilon \;\middle|\; \begin{array}{l} \mathrm{sign}(f(x) - m) \\ = \ -\mathrm{sign}(f(y) - m) \end{array} \right] \\
&\geq \frac{1}{2}\Pr\left[|f(x) - m| > \epsilon \;\middle|\; \begin{array}{l} \mathrm{sign}(f(x) - m) \\ = \ -\mathrm{sign}(f(y) - m) \end{array} \right] \\
&= \frac{1}{2}\Pr[|f(x) - m| > \epsilon]
\end{aligned}$$

Thus if $\Pr[|f(x) - f(y)| > \epsilon] < \kappa$,

$$\mathrm{E}[|f(x) - m|] \leq \Pr[|f(x) - m| > \epsilon] \cdot \sup|f(x) - m| + \epsilon \leq 2\kappa \cdot \sup|f(x) - m| + \epsilon$$

□

Appendix D

Manifolds

We will here give a short introduction to the subject of manifolds, as a reference for the reader.

Manifolds are spaces that locally look like some Euclidean space, i.e. $\mathbb{R}^d$. The manifolds we encounter in this thesis are subsets of a Euclidean space of higher dimension $\mathbb{R}^p$ (for general manifolds this is not necessarily the case) so we can think of them as a generalization of the concept of curves and surfaces into higher dimensions.

To give a formal definition, we need the concept of *homeomorphisms*:

Definition D.1. *A bijective continuous function with continuous inverse is called a* ***homeomorphism****. Two topological spaces X and Y are* ***homeomorphic*** *if there is a homeomorphism from X to Y.*

If two topological spaces are homeomorphic there is a one-to-one correspondence of open sets given by the homeomorphism. Therefore topological properties such as compactness, connectedness and even topological dimension are stable under homeomorphisms.

Now we can define

Definition D.2. *A topological space $\mathcal{M}$ is a d–**manifold** if $\mathcal{M}$ is connected, Hausdorff (i.e. points can be separated by open sets) and for each $x \in M$ there is a neighborhood $U_x \ni x$ such that U_x is homeomorphic to $\mathbb{R}^d$.*

Note 1: $B_r(x) \subseteq \mathbb{R}^d$ is homeomorphic to $\mathbb{R}^d$ for any $r > 0$, $x \in \mathbb{R}^d$. Thus any point with a neighborhood homeomorphic to an open subset of $\mathbb{R}^d$ contains a neighborhood homeomorphic to $\mathbb{R}^d$. Therefore we could have used the condition that U_x should be homeomorphic to an open subset of $\mathbb{R}^d$ instead.

Note 2: The Hausdorff condition is trivially fulfilled for any subset of a Euclidean space $\mathbb{R}^p$, since $\mathbb{R}^p$ is Hausdorff.

Even though it might seem that being locally homeomorphic to $\mathbb{R}^d$ would entail a d–manifold to behave nicely, we often need more regularity. Therefore we will often work with *smooth manifolds*. Smooth manifolds are defined the same way as manifolds, but with with the stronger condition of being locally *diffeomorphic* to a Euclidean space.

Definition D.3. *A **diffeomorphism** is a smooth bijective function with smooth inverse. If there exists a diffeomorphism $\phi\colon X \to Y$, then X and Y are said to be **diffeomorphic**.*

If $\mathcal{M}$ is a smooth d–manifold, then for each point $x \in \mathcal{M}$ there is a neighborhood $U_x \subseteq \mathcal{M}$ and a diffeomorphism $\phi\colon \mathbb{R}^d \to U_x$. Using directional derivatives of ϕ, it is possible to obtain a best linear approximation of $\mathcal{M}$ at x. Thus we can generalize the concept of tangents and tangent planes:

Definition D.4. *The d–dimensional hyper plane which is the best linear approximation of a d–manifold $\mathcal{M}$ at a point $x \in \mathcal{M}$ is the called the **tangent space** of $\mathcal{M}$ at x.*

For a rigorous treatment, see e.g. [20].

For a smooth manifold $\mathcal{M}$ embedded into $\mathbb{R}^p$ we can define a scalar product on the tangent planes of $\mathcal{M}$ by using the usual scalar product on $\mathbb{R}^p$. This scalar product is called a *Riemannian metric* on the manifold (note the difference to the usual meaning of the term metric). Riemannian metrics can be defined also for smooth manifolds not embedded into Euclidean space, and the same manifold can be given different Riemannian metrics.

A smooth manifold with a Riemannian metric is called a *Riemannian manifold*. Using the Riemannian metric we can define angles between vectors and length of vectors in tangent spaces the same way as they are defined from the scalar product in $\mathbb{R}^n$. This makes it possible to define length of curves and volume of open subsets. Using length of curves one can define distance between two points in a manifold:

Definition D.5. *The distance between two points on a Riemannian manifold is the smallest possible length of a curve within the manifold going between the two points.*

Note: When a manifold is embedded in $\mathbb{R}^p$ the length of a curve $\gamma\colon [a,b] \to \mathbb{R}^p$ can be computed the usual way:

$$L(\gamma) = \int_a^b |\dot{\gamma}(t)|\, dt.$$

This can be done also for non-Riemannian manifolds, so we can define distance in a general manifold embedded in $\mathbb{R}^p$ the same way as distance is defined in Riemannian manifolds.

When talking about distance between two points in a manifold we will often use the term *geodesic*. The definition of a geodesic is as follows:

Definition D.6. *A curve in a Riemannian manifold, $\gamma\colon [a,b] \to \mathcal{M}$, whose acceleration $\ddot{\gamma}(t)$ is orthogonal to the tangent space at $\gamma(t)$ for all $t \in [a,b]$ is called a **geodesic**.*

The reason that geodesics are mentioned when computing distance between two points is the following theorem:

Theorem D.1. *A curve with minimal length between two points in a manifold is a geodesic.*

For a proof, see [29].

Some d–manifolds can be isometrically embedded in $\mathbb{R}^d$, i.e. there is a distance-preserving bijective function going from the manifold to a subset of $\mathbb{R}^d$. For this to be possible it needs to have zero *curvature*.

The surface of a cylinder has zero curvature since it is flat in one dimension. It is easy to see that if we cut it, it can be isometrically embedded into $\mathbb{R}^2$ by rolling it up. A sphere on the other hand has non-zero curvature, if it has radius R its curvature is $1/R^2$, and it is well known that a sphere cannot be mapped to a surface isometrically (hence there are no maps of the earth which do not distort distances).

Appendix E

Measure theory

A *measure* is a countably additive function defined on a σ–ring. What this is will become clear from the following three definitions:

Definition E.1. *A σ–**ring** $\mathcal{S}$ on a set X is a collection of subsets of X such that*

- $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{S} \Rightarrow \bigcup_{i\in\mathbb{N}} A_i \in \mathcal{S}$, *and*
- $A, B \in \mathcal{S} \Rightarrow A\backslash B \in \mathcal{S}$.

In other words $\mathcal{S}$ is closed under countable unions and differences.

Definition E.2. *If $\mathcal{S}$ is a σ–ring on X and $X \in \mathcal{S}$, then $\mathcal{S}$ is called a σ–**algebra**.*

Note: If $\mathcal{S}$ is a σ–algebra, then it is closed under finite intersections, since $A \cap B = X\backslash((X\backslash A) \cup (X\backslash B))$.

Definition E.3. *A function f defined on a σ–ring $\mathcal{S}$ is **countably additive** if, when $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{S}$ is a countable collection of disjoint sets*

$$f(\bigsqcup_{i\in\mathbb{N}} A_i) = \sum_{i\in\mathbb{N}} f(A_i),$$

where $\sqcup$ denotes union of disjoint sets.

If $\mathcal{S}$ is a σ–ring on X and μ is a countably additive function defined on $\mathcal{S}$, we say that μ is a measure on X, and call $(X, \mathcal{S}, \mu)$ a *measure space*. The sets in $\mathcal{S}$ are called the *measurable sets*. From the definitions of a σ–ring and countably additive functions, we have the natural rules

- If $A \subseteq X$ and $B \subseteq X$ are disjoint measurable sets, then $A \sqcup B$ is measurable, and $\mu(A \sqcup B) = \mu(A) + \mu(B)$; this generalizes to countably many sets.
- If $A \subseteq X$ and $B \subseteq X$ are measurable sets, then $A\backslash B$ is measurable, and $\mu(A\backslash B) = \mu(A) - \mu(B) + \mu(B\backslash A)$.

Definition E.4. *A **probability measure** on X is a measure taking values in $[0, 1]$ for which X is measurable and $\mu(X) = 1$.*

Definition E.5. *A function f which can be written as*

$$f = \sum_{j=1}^{n} b_j \chi_{E_j} \ ,$$

where $b_1, \ldots, b_n$ are values in a Banach space and $E_1, \ldots, E_n$ are measurable sets for a measure μ is called a ***measurable simple function****. If $\mu(E_j) < \infty$ for $j = 1, \ldots, n$, it is also an* ***integrable simple function****.*

For the function f defined above the integral is defined by

$$\int f \, d\mu = \sum_{j=1}^{n} b_j \mu(E_j) \ .$$

Furthermore, the L^1–norm, $\|\cdot\|_1$ for the function f is defined as

$$\|f\|_1 = \sum_{j=1}^{n} \|b_j\| \mu(E_j) \ .$$

Definition E.6. *A function f is* ***measurable*** *if there exist a sequence of measurable simple functions converging pointwise to f except on a set of zero measure.*

Theorem E.1. *f is measurable if and only if the inverse images of measurable sets under f are measurable.*

Definition E.7. *A function f is μ–**integrable** if there exist a sequence of integrable simple functions $\{f_n\}_{n \in \mathbb{N}}$ that is a Cauchy-sequence for the L^1–norm and that converges pointwise to f except on a set of zero measure. Then the integral of f is defined by*

$$\int f \, d\mu = \lim_{n \to \infty} \int f_n \, d\mu \ .$$

Definition E.8. *A measure μ is said to be* ***absolutely continuous*** *with respect to another measure ν if they are defined on the same σ–ring and*

$$\nu(E) = 0 \Rightarrow \mu(E) = 0.$$

Theorem E.2 (Radon-Nikodym)**.** *If μ is absolutely continuous with respect to ν, then $\exists f \in L^1$ such that $\mu = \int f \, d\nu$.*

Theorem E.3. *If $h \in L^\infty$ and $\mu = \int f \, d\nu$, then $\int h \, d\mu = \int fh \, d\nu$.*

Bibliography

[1] F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, 1973.

[2] J. Beardwood, J. H. Halton, and J. M. Hammersley. The shortest path though many points. *Mathematical Proceedings of the Cambridge Philosophical Society*, 55(4):299–327, 1959.

[3] F. Camastra. Data dimensionality estimation methods: a survey. *Pattern Recognition*, 36:2945–2954, 2003.

[4] K. M. Carter, A. O. Hero III, and R. Raich. De-biasing for intrinsic dimension estimation. In *IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 601–605, 2007.

[5] K. M. Carter, R. Raich, and A. O. Hero III. On local intrinsic dimension estimation and its applications. *IEEE Transactions on Signal Processing*, 58(2):650–663, 2010.

[6] E. Chávez et al. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, 2001.

[7] J. A. Costa and A. O. Hero III. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, 52(8):2210–2221, 2004.

[8] J. A. Costa and A. O. Hero III. Determining intrinsic dimension and entropy of high-dimensional shape spaces. In *Statistics and Analysis of Shapes*. Birkhäuser, 2006.

[9] C. D. Cutler. Some results on the behavior and estimation of the fractal dimensions of distributions on attractors. *Journal of Statistical Physics*, 62(3–4):651–708, 1991.

[10] P. Demartines. *Analyse de données par réseaux de neurones auto-organisés*. PhD thesis, Institut National Polytechnique de Grenoble, 1994. Available at `http://demartines.com/publi.html`.

[11] Encyclopaedia Mathematica. Dimension theory. SpringerLink, `http://eom.springer.de/d/d032500.htm`. [Online; accessed 25-November-2010].

[12] K. Falconer. *Fractal Geometry*. John Wiley & Sons, 1990.

[13] V. V. Fedorchuk. The fundamentals of dimension theory. In *General Topology I*, volume 17 of *Encyclopaedia of Mathematical Sciences*. Springer-Verlag, 1990.

[14] M. Fontes. Some remarks concerning a geometric constant and its connection with the Banach-Mazur distance between finite dimensional normed linear spaces. Lund Institute of Technology, Department of Mathematics, Lund, 1993.

[15] K. Fukunaga and D. R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, c-20(2):176–183, 1971.

[16] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, 2000. Chapter III.

[17] P. Grassberger. Generalized dimensions of strange attractors. *Physics Letters*, 97A(6):227–230, 1983.

[18] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1–2):189–208, 1983.

[19] M. Gromov. *Metric structures for Riemannian and non-Riemannian spaces*, volume 152 of *Progress in mathematics*. Birkhäuser Verlag, 1999. Chapter 3 1/2.

[20] V. Guillemin and A. Pollack. *Differential topology*. Prentice-Hall, 1974.

[21] D. Harte. *Multifractals — Theory and Applications*. Chapman and Hall/CRC 2001, 2001. eBook available at `http://www.crcnetbase.com/isbn/9781584881544`.

[22] F. Hausdorff. Dimension und äußeres maß. *Mathematishe Annalen*, 79:157–179, 1919.

[23] M. Hein and J.-Y. Audibert. Intrinsic dimensionality estimation of submanifolds in R^d. In *Proc. 22nd Int. Conf. Machine Learn.*, pages 289–296, 2005.

[24] H. G. E. Hentschel and I. Procaccia. The infinite number of generalized dimensions of fractals and strange attractors. *Physica*, 8D:435–444, 1983.

[25] B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

[26] B. Kégl. Intrinsic dimension estimation using packing numbers. In *Advances in Neural Information Processing Systems 15*, pages 681–688. MIT Press, 2003.

[27] V. I. Koltchinskii. Empirical geometry of multivariate data: A deconvolution approach. *The Annals of Statistics*, 28(2):591–629, 2000.

[28] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer New York, 2007. eBook available at `http://www.springerlink.com/content/v88v83`.

[29] J. M. Lee. *Riemannian Manifolds: An Introduction to Curvature*. Springer-Verlag New York, 1997.

[30] E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17*, pages 777–784. MIT Press, 2005.

[31] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[32] V. Pestov. An axiomatic approach to intrinsic dimension of a dataset. *Neural Networks*, 21:204–213, 2008.

[33] V. Pestov. Intrinsic dimensionality. *SIGSPATIAL Special*, 2(2):8–11, 2010.

[34] D. V. Pisarenko and V. F. Pisarenko. Statistical estimation of the correlation dimension. *Physics Letters A*, 197:31–39, 1995.

[35] M. Raginsky and S. Lazebnik. Estimation of intrinsic dimensionality using high-rate vector quantization. In *Advances in Neural Information Processing Systems 18*, pages 1105–1112. MIT Press, 2006.

[36] S. Raychaudhuri, J. M. Stuart, and R. B. Altman. Principal components analysis to summarize microarray experiments: Application to sporulation time series. In *Pacific Symposium on Biocomputing*, volume 5, pages 452–463, 2000.

[37] A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561, 1961.

[38] R. N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function, part I and II. *Psychometrika*, 27(2, 3):125–140, 219–245, 1962.

[39] F. Takens. Invariants related to dimension and entropy. In *Atas do* 13^o *Colóqkio Brasileiro de Matemática*, Rio de Janeiro, 1983.

[40] F. Takens. On the numerical determination of the dimension of an attractor. In *Dynamical Systems and Bifurcations, Proceedings Groningen 1984*, volume 1125 of *Lecture Notes in Mathematics*, pages 99–106. Springer-Verlag, 1984.

[41] L.-S. Young. Dimension, entropy and Lyapunov exponents. *Ergodic Theory and Dynamic Systems*, 2:109–124, 1982.

[42] P. L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Trans. on Inf. Th.*, 28(2):139–149, 1982.