Intrinsic Dimensionality Estimation of Submanifolds in $\mathbb{R}^d$

Matthias Hein MH@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Jean-Yves Audibert AUDIBERT@CERTIS.ENPC.FR
CERTIS, ENPC, Paris, France

Abstract

We present a new method to estimate the intrinsic dimensionality of a submanifold M in $\mathbb{R}^d$ from random samples. The method is based on the convergence rates of a certain U-statistic on the manifold. We solve at least partially the question of the choice of the scale of the data and can quantify the influence of the extrinsic curvature of the manifold. Moreover the proposed method is easy to implement, can handle large data sets and performs very well even for small sample sizes. We compare the proposed method to two standard estimators on several artificial as well as real data sets.

1. Introduction

The topic of intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$ has a long history. In this paper we consider the case where we have random samples from a probability distribution which has support on a submanifold in $\mathbb{R}^d$. In recent years there has been done a lot of work in estimating manifold structure from the data. However finding low-dimensional approximations of submanifolds is considerably harder than estimating their dimension and the goal of what kind of the structure of the manifold should be preserved in the approximation differs from method to method.
However the goal of estimating the dimension of a submanifold is a well-defined mathematical problem. Indeed all notions of dimensionality like e.g. topological, Hausdorff or correlation dimension agree for submanifolds in $\mathbb{R}^d$. Differences arise only if one considers more irregular sets like fractals, see (Falconer, 2003).
The methods for dimensionality estimation up to now can be roughly divided into two groups. The first one tries to determine the dimensionality by dividing the data in small subregions followed by a principal component analysis (PCA) of the points in each subregion. The number of dominant eigenvalues determines then the dimension, see (Fukunaga, 1971). This method has two drawbacks, first one has to find a suitable scale for the size of the subregions and second one has to determine what one considers as dominant eigenvalues, which is also a typical problem of standard PCA. The second type of estimators was originally designed to determine the dimension of the attractor of a chaotic dynamical system from samples of its time series. They are all based on the assumption that the volume of an m-dimensional set scales with its size s as s^m which implies that also the number of neighbors less than r apart will behave in the same way. This was the motivation for Grassberger and Procaccia (1983) to define the correlation integral as

$$C_n(s) = \frac{2}{n(n-1)} \sum_{i<j}^{n} \mathbb{1}_{\|X_i - X_j\| \leq s}$$

where $X_i = 1, \ldots, n$ are the n sample points of our manifold in $\mathbb{R}^d$. Then the correlation dimension ν is defined as

$$\nu = \lim_{r \to 0} \lim_{n \to \infty} \frac{\log C_n(s)}{\log s}$$

In practice one computes $C_n(s)$ for different s_i and then fits a line through $[\log s_i, \log C_n(s_i)]$ with least squares. Similar to the method of Fukunaga also for the correlation dimension one has the drawback that one has to chose the scales r_i. Note that this is a crucial step since the data is always 0-dimensional at a very small scale and is maybe even d-dimensional at a large scale, so that one either under-or overestimates the dimension.
The quantity we estimate is essentially the correlation integral with $\mathbb{1}$ replaced by a general kernel function. However the way we estimate now the dimension is based on the convergence rate of the modified correlation integral. The advantage is that we only have to choose once a kind of 'smallest' scale at which one examines the data, the others are then determined by the

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

convergence rate. Also we examine for the first time the influence that one uses in the correlation integral the distance in $\mathbb{R}^d$ and not the intrinsic distance of the manifold. The analysis of the limit of the modified correlation integral shows explicitly how the intrinsic and extrinsic curvature of the submanifold as well as the smoothness of the density of the probability measure influence the asymptotics of the correlation integral. Both effects lead to a scaling of $C_n(s)$ which is different from s^m.

2. Theoretical Background

We assume that the probability measure[1] P generating the data $X_i \in \mathbb{R}^d$ has support on a m-dimensional submanifold M of $\mathbb{R}^d$. That means we are not trying to separate possible noise in the data from the underlying ground truth. In fact we will argue later in an experiment that on the basis of a finite sample it is in principle impossible to judge whether one has noise in the data or a very curved manifold. Moreover we also exclude the case of probability distributions with support of fractal dimension. As in the case of noise it is in principle impossible to judge based on a finite sample whether the data has fractal dimension or just very high curvature.
The m-dimensional submanifold M is a Riemannian manifold if one considers the induced metric from $\mathbb{R}^d$. That means that the inclusion map $i : M \to \mathbb{R}^d$ is an isometry (in the sense of Riemannian manifolds). Note that we will use in the following the somehow cumbersome notation $x \in M$ and $i(x) \in \mathbb{R}^d$ in order to make it more obvious when we are working on the manifold M and when on $\mathbb{R}^d$. As any Riemannian manifold, M is also a metric space with the path-metric. A key point in the following proof will be the relation of the distance $d(x, y)$ on M and the Euclidean distance $\|i(x) - i(y)\|$ in $\mathbb{R}^d$ of two points $x, y \in M$. This relation has been derived in (Smolyanov et al., 2004):

Lemma 1 *Let $i : M \to \mathbb{R}^d$ be an isometric embedding of the smooth m-dimensional Riemannian manifold M into $\mathbb{R}^d$ and let $x \in M \backslash \partial M$, then $\forall y \in B_M(x, \mathrm{inj}(x))$*[2]

$$\|i(y) - i(x)\|^2_{\mathbb{R}^d} = d^2_M(x, y) - \frac{1}{12} \|\Pi(\dot\gamma, \dot\gamma)\|^2_{T_x\mathbb{R}^d} + O(d^5_M(x, y)),$$

where $\mathrm{inj}(x)$ *is the injectivity radius*[3] *at x,* Π *is the second fundamental form and $\gamma : [0, 1] \to M$, with $\gamma(0) = x$ and $\gamma(1) = y$ is the unique geodesic*[4] *from x to y.*

The second fundamental form Π is the extrinsic curvature of M, see e.g. (Lee, 1997).
Before going into further detail let us state our assumptions on M and P. We will need some regularity of the submanifold. In particular we need to bound the deviation of the extrinsic distance in $\mathbb{R}^d$ in terms of the intrinsic distance in M. For each $x \in M$ we define the *regularity radius* $r(x)$ as

$$r(x) = \sup\{r > 0 \big|\ \|i(x) - i(y)\|^2_{\mathbb{R}^d} \geq \frac{1}{2} d^2_M(x, y), \quad \forall\, y \in B_M(x, r)\}$$

Assumption 1
- *$i : M \to \mathbb{R}^d$ is a smooth, isometric embedding*[5]*,*
- *M has a bounded second fundamental form,*
- *M has bounded sectional curvature,*
- *for all $x \in M$, $r(x) > 0$, and r is continuous,*
- $\delta(x) = \inf\{\|i(x) - i(y)\|_{\mathbb{R}^d} \big| y \in M \backslash B_M(x, \frac{1}{3}\min\{\mathrm{inj}(x), r(x)\})\} > 0, \forall x \in M,$
- *Define $S_\epsilon = \{x \in M, d(x, \partial M) < \epsilon\}$, then $\forall \epsilon > 0$,* $\mathrm{inj}_\epsilon := \inf_{x \in M \backslash S_\epsilon} \mathrm{inj}(x) > 0.$
- $\forall \epsilon > 0$, $r_\epsilon := \inf_{x \in M \backslash S_\epsilon} r(x) > 0.$
- $\forall \epsilon > 0$, $\delta_\epsilon := \inf_{x \in M \backslash S_\epsilon} \delta(x) > 0.$

The first condition ensures that M is a smooth submanifold of $\mathbb{R}^d$ with the metric induced from $\mathbb{R}^d$ (this is usually meant when one speaks of a submanifold in $\mathbb{R}^d$). The next three properties guarantee that M is well behaved. The fifth condition ensures that if parts of M are far away from x in the geometry of M, they do not come too close to x in the geometry of $\mathbb{R}^d$. The last three conditions ensure that up to a small strip at the boundary we have global control over $\mathrm{inj}(x), \delta(x)$ and $r(x)$.
The reader who is not familiar with Riemannian geometry should keep in mind that locally, a submanifold of dimension m looks like $\mathbb{R}^m$.

Assumption 2
- *P has a density p with respect to the natural volume element $dvol(x) = \sqrt{\det g}\, dx$ on M,*
- *p is in $C^3(M)$,*

[1] Note that if $m < d$, P is not absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$.
[2] $B_X(x, r)$ is the ball at x of size r in (X, d).
[3] The injectivity radius of a point is the radius of the largest ball in M such that the exponential map is defined and injective.
[4] Note that γ is not parameterized by arc-length
[5] That means the Riemannian metric g_{ab} on M is induced by $\mathbb{R}^d$, $g^M_{ab} = i_* g^{\mathbb{R}^d}_{ab}$, where $g^{\mathbb{R}^d}_{ab} = \delta_{ab}$.

- $\int_M p^2(x) dvol(x) < \infty$.

The kernels used in this paper are always isotropic, that is they are functions of the norm in $\mathbb{R}^d$. Furthermore we make the following assumptions on the kernel function k:

Assumption 3
- $k : \mathbb{R}_+ \to \mathbb{R}$ *is measurable, non-negative and non-increasing,*
- $k \in C^2(\mathbb{R}_+)$, $\|k\|_\infty = K$ *and* $\frac{\partial^2 k}{\partial x^2}$ *is bounded,*
- k *has compact support on* $[0, R]$,
- $C_1 = \int_{\mathbb{R}^m} k(\|y\|^2) dy < \infty$,
  $C_2 = \int_{\mathbb{R}^m} k(\|y\|^2) y_1^2 dy < \infty$.

Furthermore we define

$$k_h(\|i(x) - i(y)\|^2) = \frac{1}{h^m} k(\|i(x) - i(y)\|^2 / h^2).$$

Now that we have stated the setting we are working in, we can introduce our estimator. We denote by X the i.i.d. sample $X_i, i = 1, \ldots, n$ of size n drawn from P. Then the U-statistic we use is defined as

$$U_{n,h}(k) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n}^{n} k_h(\|i(X_i) - i(X_j)\|^2)$$

The expectation of $U_{n,h}(k)$ is given as

$$\begin{aligned} &\mathbb{E}\, U_{n,h}(k) = \mathbb{E}\, k_h(\|i(X) - i(Y)\|^2) \\ &= \int_M \int_M k_h(\|i(x) - i(y)\|^2) p(x) p(y)\, dvol(x)\, dvol(y) \end{aligned}$$

The central point is how this U-statistic behaves as $n \to \infty$ and $h \to 0$. At first we study how the expectation behaves as $h \to 0$. A first step is the following modification of a result in (Hein et al., 2005):

Proposition 1 *Let* $S_\epsilon = \{x \in M, d(x, \partial M) < \epsilon\}$. *Under the stated assumptions on* M, P *and* k *there exists* $\forall\, \epsilon > 0$ *an* $h_0 > 0$, *such that for all* $h < h_0$

$$\begin{aligned} &\int_{M \setminus S_\epsilon} \int_M k_h \left(\|i(x) - i(y)\|^2_{\mathbb{R}^d} \right) p(x) dvol(x) p(y) dvol(y) \\ &= \int_{M \setminus S_\epsilon} \Big(C_1 p(x) + C_2 \frac{h^2}{4} p(x) \Big[-R + \frac{1}{2} \Big\| \sum_i \Pi(\partial_i, \partial_i) \Big\|^2 \Big] \\ &\quad + C_2 \frac{h^2}{2} (\Delta_M\, p)(x) + \Gamma(x) h^3 \Big) p(x) dvol(x), \end{aligned}$$

where Δ_M *is the Laplace-Beltrami operator and* R *the scalar curvature of* M *and* $\Gamma(x)$ *a function depending on* h_0, $\|f\|_{C^3}$ *and* $\|p\|_{C^3}$.

Proof: The expansion of the integral is given in Proposition 1 in (Hein et al., 2005), where $h_0(x) = \frac{1}{3} \min\{\mathrm{inj}(x), r(x)\}$. Now due to our assumptions on $\delta_\epsilon, \mathrm{inj}_\epsilon$ and r_ϵ, the expansion can be done uniformly on $M \setminus S_\epsilon$ with $h_0 = \inf_{M \setminus S_\epsilon} h_0(x) > 0$. □

This proposition shows that $U_{n,h}(k)$ has only asymptotically the expected scaling behavior. There is a second order correction with influence from the curvature of M and the possibly non-uniform probability measure P.

Proposition 2 *Under the stated assumptions on* M, P *and* k,

$$\lim_{h \to 0} \mathbb{E}\, U_{n,h}(k) = C_1 \int_M p(x)^2 dvol(x)$$

Proof: Let $f_h(x) = \int_M k_h(\|i(x) - i(y)\|^2) p(y) dvol(y)$. Then by Proposition 1 in (Hein et al., 2005), $\lim_{h \to 0} f_h(x) = C_1 p(x)$. Moreover one can show that there exists a constant C such that $f_h(x) \le C p(x)$. Namely by our assumptions on M we have for sufficiently small h

$$\begin{aligned} |f_h(x)| &\le K/h^m \int_M \mathbb{1}_{\|i(x) - i(y)\| \le hR} p(y) dvol(y) \\ &\le 2K/h^m p(x) Vol(\{y \,|\, \|i(x) - i(y)\| \le hR\}) \\ &\le C p(x) \end{aligned}$$

The proposition then follows by the dominated convergence theorem since by our assumption $\int_M p(x)^2 dvol(x) < \infty$. □

The next step in the proof is to control the deviation of $U_{h,h}$ from its expectation. The following concentration inequality of Hoeffding lets us quantify the probability that $U_{n,h}$ deviates from $\mathbb{E}\, U_{n,h}$ by at most ϵ. We use the following Bernstein-type of bound, see (Hoeffding, 1963; Serfling, 1980):

Theorem 1 (Hoeffding, 1963) *Let* $\|k\|_\infty \le b$, $\mathbb{E}\, k_h(\|i(X) - i(Y)\|^2) < \infty$ *and* $\sigma^2 = \mathrm{Var}\, k_h(\|i(X) - i(Y)\|^2) < \infty$, *then*

$$P\left(|U_{n,h} - \mathbb{E}\, U_{n,h}| \ge \epsilon\right) \le 2\, e^{-\frac{[n/2]\epsilon^2}{2\sigma^2 + 2/3 |b - \mathbb{E}\, U_{n,h}| \epsilon}}$$

where $[x]$ denotes the greatest integer smaller than x. A straightforward application of this concentration inequality yields the following theorem:

Theorem 2 *Let* M, P *and* k *fulfill the stated assumptions, then*

$$P\left(|U_{n,h} - \mathbb{E}\, U_{n,h}| \ge \epsilon\right) \le 2\, e^{-\frac{[n/2] h^m \epsilon^2}{2K \mathbb{E}\, U_{n,h} + 2/3 |K - h^m \mathbb{E}\, U_{n,h}| \epsilon}}$$

Furthermore let $n \to \infty$ and $h \to 0$, then if $nh^m \to \infty$

$$\lim_{n\to\infty} U_{n,h}(k) = C_1 \int_M p(x)^2 dvol(x), \quad \text{in probability}$$

If the stronger condition $nh^m / \log n \to \infty$ holds, then

$$\lim_{n\to\infty} U_{n,h}(k) = C_1 \int_M p(x)^2 dvol(x), \quad \text{almost surely}$$

Proof: We have by assumption $\|k_h\|_\infty \leq K/h^m$ and it can be verified that $\operatorname{Var} U_{n,h} \leq K/h^m \mathbb{E}\, U_{n,h}$. Now applying Theorem 1 yields the bound. Using this concentration inequality convergence in probability of $U_{n,h}$ towards its expectation follows immediately from the condition $nh^m \to \infty$. Moreover we know from Proposition 2 the form of $\mathbb{E}\, U_{h,h}$ as $h \to 0$. Complete convergence, which implies almost sure convergence, follows from $\sum_{n=1}^{\infty} P(|U_{n,h} - \mathbb{E}\, U_{n,h}| \geq \epsilon) < \infty$. This follows if the stronger condition holds. □

The previous theorem together with the following corollary will be the cornerstones of our algorithm.

Corollary 1 *Let M, P and k fulfill the stated assumptions and define $k_h = \frac{1}{h^l} k(\|i(x) - i(y)\|^2 / h^2)$, then if $h \to 0$ and $nh^l \to \infty$*

$$\lim_{n\to\infty} U_{n,h}(k) = \infty, \qquad \text{if} \quad l > m$$
$$\lim_{n\to\infty} U_{n,h}(k) = 0, \text{ in probability} \qquad \text{if} \quad l < m$$

Proof: By Theorem 2 we have for $l = m$ convergence in probability to $C_1 \int_M p(x)^2 dvol(x)$. Now with the different power of h in front of the kernel we have convergence towards $\frac{C_1}{h^{l-m}} \int_M p(x)^2 dvol(x)$. Since the integral is finite this diverges if $l > m$ and converges to zero if $l < m$. □

Note that we get convergence to a finite number if and only if $l = m$, since $0 < \int_M p(x)^2 dvol(x) < \infty$.

3. The Algorithm

The algorithm is based on the convergence result in Theorem 2 and on Corollary 1. Using these results we know that in order to get convergence in probability the bandwidth h has to fulfill $nh^m \to \infty$. Otherwise the U-statistic either diverges or approaches zero. We will use this property by fixing a convergence rate for each dimension, that means we are fixing h as a function of the sample size n. Then we compute the U-statistic for subsamples of different sizes, where h varies according to the function we have fixed. Finally the dimension is determined by the U-statistic which has the smallest slope as a function of n.

3.1. First Step: Fixing $h_l(n)$

As a first step we fix $h_l(n)$ as a function of the sample size n and the dimension l. We choose the function in such a way that it is just sufficient for convergence in probability so that $h_l(n)$ approaches zero at the fastest allowed rate, that is

$$nh(n)^l = \frac{1}{c^l} \log n \;\Rightarrow\; h_l(n) = \frac{1}{c} \left(\frac{\log n}{n} \right)^{1/l},$$

where c is a constant. The crucial point of this procedure is that the scales at which we look at the data vary according to the dimension l, so that $U_{n,h}(k)$ will depend as a function of the sample size n on the chosen dimension l. We fix the constant c in the algorithm by determining a certain nearest neighbor scale. Let N be the total number of samples of our data set and define $d(X_i)$ as the distance of the sample X_i to its nearest neighbor. We set:

$$h_l(N) = \frac{1}{N} \sum_{i=1}^{N} d(X_i) \;\Rightarrow\; c = \frac{1}{h_l(N)} \left(\frac{\log N}{N} \right)^{1/l}.$$

In total we get for the function $h_l(n)$:

$$h_l(n) = h_l(N) \left(\frac{N}{n} \frac{\log n}{\log N} \right)^{1/l}$$

Note that $h_l(N)$ does not depend on the dimension, that is we examine the full data (all N sample points) for each dimension at the same scale $h_l(N)$. The crucial point however is that as we consider subsamples of size n the scale $h_l(n)$ is different for each dimension.

3.2. Second step: Computing the dimension

The choice of the kernel seems not to influence the result much. We choose a kernel with compact support to save computational time, that is

$$k(x) = (1 - x)_+$$

We consider subsamples of size $\{[N/5], [N/4], [N/3], [N/2], N\}$. For each dimension $l \in \{1, \ldots, \mathrm{l}_{\max}\}$, where we put usually $l_{\max} = \min\{d, 15\}$, we compute the empirical estimate of $U_{[N/r], h_l([N/r])}(k)$, $r = 1, \ldots, 5$.
In order to improve the estimates of the subsamples we consider not only one subsample but several ones by using the so called two-sample U-statistics which is defined as follows. Given two i.i.d. samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, one considers the following U-statistic

$$U = \frac{1}{n^2} \sum_{i,j=1}^{n} k(X_i, Y_j)$$

It was shown by Hoeffding (Hoeffding, 1963) that also this form of the U-statistic converges as in Theorem 1. In our case we will take two samples of the same distribution, so that the expectation and variance are the same as for the one-sample U-statistic.

Let us explain how we use this result for our subsamples. Consider the subsample of size $[N/2]$. Using our full set of N-data points, we can build three samples, namely $(X_{2k}, X_{2k}), (X_{2k+1}, X_{2k})$ and (X_{2k+1}, X_{2k+1}). The first and the last one lead to one-sample U-statistics and the second one to a two-sample U-statistic. For each of these subsamples we compute the estimates of $U_{[N/2],h}(k)$ and then take the mean of them. Obviously one gets for the subsample of size $[N/r]$, $r = 5, 4, 3, 2, 1$, $r(r+1)/2$ such estimates, and for each r we take the mean of them. This method looks at first quite complicated. However the implementation is straightforward and solves the problem that especially for small sample sizes N taking subsamples leads to high variances in the estimates. Using instead a set of subsamples with the described method we can in that way minimize the variance of the estimates corresponding to the subsamples.

The estimation of the U-statistics can be done for all dimensions and for all subsample sizes simultaneously. Especially for high-dimensional data, which is potentially the most interesting one, the main computational cost lies in the computation of the distances and not in the calculations of $U_{[N/r],h([N/r])}(k)$.

In order to determine the dimension we fit for each dimension l a line through the five points $\left[\log h_l([N/r]), \log U_{[N/r],h_l([N/r])}(k)\right]$, $r = 1, \ldots, 5$, with weighted least squares with weights $w(r) = 1/r$ which can be easily done in closed form. The dimension is then determined by the line with the smallest absolute value of the slope of the line. This is justified since the slope of $\log U_{n,h_l(n)}(k)$ is given by $(m-l)\log h_l(n)$ as $n \to \infty$ and $h \to 0$ from Theorem 2 and the resulting corollary 1. We use weighted least squares since for smaller subsamples we look at the data at a larger scale. Therefore if one has high curvature these estimates are less reliable.

4. Experiments

The experiments we perform are only partially based on datasets which have been previously used for dimensionality estimation. The reason for this is that these datasets don't have high extrinsic and intrinsic curvatures. In our experiments based on artificial datasets we study the influence of high curvature as well as noise on our estimator. Later on we will evaluate the estimator on two real world datasets. The first one is the face database used in the study of ISOMAP (Tenenbaum et al., 2000) and the MNIST database. For the MNIST database we actually don't know the intrinsic dimensionality. Therefore we study first for the digit 1 an artificial dataset, where we can control the number of dimensions. This study gives then a hint how well our estimator performs. We compare the results of our method to that of the correlation dimension estimator described in the introduction and the estimator of Takens (1985) defined as

$$\nu^{-1} = < \log(\|i(X_i) - i(X_j)\| \, / h_{\text{Takens}}) >$$

where $< >$ is the mean over all distances smaller than h_{Takens}. In order to do a fair comparison we tried to optimize the scales s_i for the correlation dimension estimator as well as the 'maximal' scale h_{Takens} for the Takens estimator over all the datasets. We fixed then them to $s_i = d + 0.2\, r\, \sigma$, $r = 1, \ldots, 5$ for the correlation dimension estimator and $h_{\text{Takens}} = d + \sigma$ where d is the mean and σ the standard deviation of the nearest neighbor distances. We would like to note that also for the Takens estimator one has to determine only one scale, however since it is a kind of 'maximal scale' it is more difficult to choose then a minimal scale as for our method.

4.1. Sinusoid on the circle

In this example our one-dimensional submanifold is a strongly oscillating sinusoid on the circle in $\mathbb{R}^3$, see Figure 1.

$$s(t) : [0, 2\pi) \to \mathbb{R}^3, \quad s(t) \to (\sin t, \cos t, \frac{1}{10} \sin 150t)$$

We sample straightforward in our coordinate expression, which yields a non-uniform probability measure on this manifold where more points appear at the extreme points of the sinusoid. We compare this submanifold to a circle with uniform noise of height 0.1 in the z-direction, see Figure 2, which results in a strip of the cylinder, which is 2-dimensional. The results are shown in Table 1 for $400, 500$ and 600 sample points. Two conclusions can be drawn. The first rather obvious one is that very curved submanifolds require a large number of samples so that their dimension can be well estimated since the high curvature of the sinusoid is misinterpreted as a second dimension for small sample sizes. The second one is that for small sample sizes it is impossible to distinguish between noise and high curvature. The rather surprising fact is that already for a sample size of 600 we have an almost perfect distinction between the one-dimensional sinusoid and the two-dimensional strip of the cylinder.

Table 1. Correct estimates of dimension 1 for the sinusoid and dimension 2 for the noisy circle of 90 trials. $a/b/c$, a our method, b corr. dim. est., c Takens est.

	400	500	600
Sinuosid	15/0/12	49/57/49	86/88/90
Noisy Circle	90/90/90	90/90/90	90/90/90

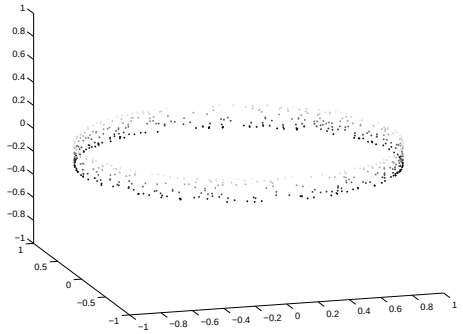

Figure 1. 600 samples of the sinusoid

4.2. The m-sphere

In this experiment we study the m-dimensional spheres S^m embedded in $\mathbb{R}^{m+1}$. The $n = 600, 800, 1000, 1200$ data points are sampled in 90 trials uniformly from the sphere S^m. The number of successful trials is given in Table 2. For S^7 and S^9 the number of samples is no longer sufficient (curse of dimensionality), most of the time the dimension is underestimated by one.

4.3. The 10-Möbius strip

The k-Möbius strip is a submanifold in $\mathbb{R}^3$ which can be created by taking a rectangle, twisting it k-times and then identifying the ends. If k is odd one gets a non-orientable manifold with surprising properties. It is obvious that this manifold has high extrinsic curvature, increasing with the number of twists k. We considered a 10-Möbius strip, see Figure 3 for an illustration with 16000 points. The coordinate representation for $u \in [-1, 1], v \in [0, 2\pi)$, is as follows:

$$x_1(u,v) = \big(1 + \frac{u}{2}\cos\big(\frac{k}{2}v\big)\big)\cos(v),$$
$$x_2(u,v) = \big(1 + \frac{u}{2}\cos\big(\frac{k}{2}v\big)\big)\sin(v),$$
$$x_3(u,v) = \frac{u}{2}\sin\big(\frac{k}{2}v\big).$$

We sampled in this coordinate representation $20, 40, 80$ and 120 points. This example is done to illustrate that even for manifolds with high extrinsic curvature the intrinsic dimension can be estimated with a relatively small sample size, see Table 3.

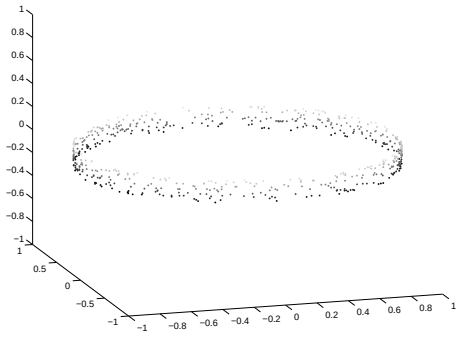

Figure 2. 600 samples of the circle with uniform noise of height 0.1 in the z-direction

Table 2. Number of correct estimates of 90 trials. $a/b/c$, a our method, b corr. dim. est., c Takens est.

	600	800	1000	1200
S^3	90/89/90	90/90/90	90/90/90	90/90/90
S^5	83/80/88	87/81/90	89/86/90	90/89/90
S^7	68/57/65	73/66/79	78/66/78	79/72/84
S^9	30/36/32	47/30/43	50/33/47	58/45/50

4.4. A 12-dimensional manifold in $\mathbb{R}^{72}$

As the last artificial dataset we present a high-dimensional dataset, a 12-dimensional manifold in 72-

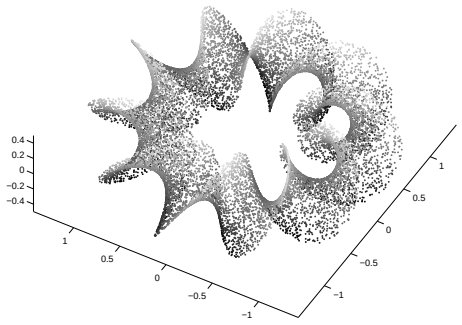

Figure 3. The 10-Möbius strip with 16000 points.

Table 3. Correct estimates of 90 trials of the 10-Möbius strip. $a/b/c$, a our method, b corr. dim. est., c Takens est.

20	40	80	120
49/34/44	71/68/73	83/78/86	88/82/90

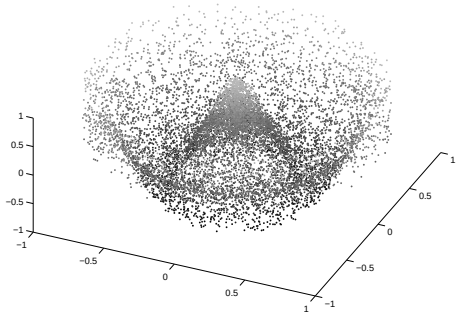

Figure 4. A 3D projection of the 12-dim manifold of Subsection

dimensions. The submanifold is generated by

$$\begin{aligned} x(\alpha) :&[0,1]^{12} \to \mathbb{R}^{72}, \\ x_{2i-1}(\alpha) &= \alpha_{i+1}\cos(2\pi\alpha_i), \quad i = 1,\dots,11, \\ x_{2i}(\alpha) &= \alpha_{i+1}\sin(2\pi\alpha_i), \quad i = 1,\dots,11, \\ x_{23}(\alpha) &= \alpha_1\cos(2\pi\alpha_{12}), \quad x_{24}(\alpha) = \alpha_1\sin(2\pi\alpha_{12}) \\ x_{j+24} &= x_{j+48} = x_j,\ j = 1,\dots,24 \end{aligned}$$

By this construction the 12-dimensional manifold lies effectively in a 24-dimensional subspace. In order to give some impression how it looks like we show the projection on the first three coordinates in Figure 4. We sample directly in these coordinates which yields a non-uniform probability measure on the manifold, which is concentrated around the origin. This leads to an interesting phenomenon, when we try to estimate the dimension. The results shown in Table 4 illustrate the connection between high curvature and non-trivial probability measure effects on the manifold. We believe that they somehow cancel out in this case. Namely a highly curved manifold leads to an overestimation of the dimension whereas a concentrated probability measure leads to an underestimation. For a relatively small sample size of 800 we already get a quite good estimate of the dimension, which is probably due to the high concentration around the origin.

4.5. The ISOMAP face database

The ISOMAP face database consists of 698 images (256 gray levels) of size 64×64 of the face of a sculpture. This dataset has three parameters: the vertical and horizontal pose and the lighting direction (one-dimensional). All estimators get for this dataset in $\mathbb{R}^{4096}$ the correct intrinsic dimension of 3.

Table 4. Correct est. of 90 trials on the 12-dim manifold. $a/b/c$, a our method, b corr. dim. est., c Takens est.

200	400	800	1600
46/42/43	60/51/61	64/70/68	84/85/85

4.6. The MNIST dataset

The MNIST dataset consists of 70000 images (256 gray levels) of size 28×28 of handwritten digits. In the generation of the MNIST dataset for all images the center of mass was computed and then the image translated such that the center of mass lies at the center of the image. However note that this does not mean that there are no translational degrees of freedom in this dataset since e.g. the digit 1 can be written with a line below or not and therefore the center of mass will vary.

4.6.1. The artificial 1-digit dataset

The intrinsic dimension of each digit is in principle unknown. In order to validate our experiment we constructed an artificial dataset of the digit 1 where we can control the dimensionality. Namely we have 5 degrees of freedom: two for translations (T), one for rotation (R), one for line thickness (L) and one for having a small line at the bottom (V). The images are constructed by having an abstract 1 as a function on $[0,1]^2$ where the different transformations are applied and then this function on $[0,1]^2$ is discretized to an image of size 28×28. We constructed 5 datasets each of size 10000, the letter combination shows which transformations have been applied, see Figure 5 for samples of the TRLV dataset. The results of the estimators on this four datasets are shown in Table 5. In three cases we are able to estimate the correct intrinsic dimension, whereas in one case we overestimate the dimension. Regarding these results on this artificial dataset we have some confidence in the results on the real MNIST dataset.

Figure 5. Samples of the artificial 1-dataset T+R+L+V.

Table 5. Estimated Dimension of the artificial 1-data sets.

Art. Digit 1	T	TR	TRL	TRLV
int. dim.	2	3	4	5
est. int. dim.	2/1/2	3/4/4	5/4/4	5/5/5

Table 6. Number of samples and estimated intrinsic dimensionality of the digits in MNIST.

1	2	3	4	5
7877	6990	7141	6824	6903
8/7/7	13/12/13	14/13/13	13/12/12	12/12/12
6	7	8	9	0
6876	7293	6825	6958	6903
11/11/11	10/10/10	14/13/13	12/11/11	12/11/11

4.6.2. Intrinsic Dimensionality of the Digits in MNIST

The estimated intrinsic dimensions are reported for each digit in Table 6 together with the number of samples of the digit in the MNIST database. Considering our result of the artificial dataset for the digit 1 we think that an estimated dimension 8 seems quite reasonable. Additional degrees of freedoms could be the length of the main line, the angle between the main line and the upper line and the length of the upper line. The intrinsic dimensions of digit 2 and 3 were estimated in (Costa & Hero, 2004) for a subsample of size 1000 as 13 and 12 respectively 12 and 11 depending on the way they build their neighborhood graph. We estimate an intrinsic dimension of 14 for digit 2 and 13 for digit 3. In comparison the results roughly agree. The difference could arise since we consider the whole dataset we look at the data at a smaller scale and therefore estimate a higher dimension.

5. Discussion

We have presented an algorithm for intrinsic dimensionality estimation of a submanifold in $\mathbb{R}^d$ from random samples. The assumptions we impose on the submanifold and the probability measure on this submanifold are not restrictive. A more careful analysis might even reveal that some of these assumptions are redundant. Our theoretical analysis clarifies the influence of the curvature of the submanifold and smoothness of the density on the asymptotic behavior of our estimated quantity. Opposite to the standard correlation dimension estimator we only have to choose once a scale at which we examine the data, the scales at which we examine subsamples are then fixed, so that we have only one free parameter in our algorithm. Even more we fixed this parameter by choosing the somehow smallest scale at which it makes sense to look at the data. In that sense we have presented an algorithm without parameters which estimates the dimension for all kinds of submanifolds irrespectively of their intrinsic and extrinsic curvature and works well also for real world datasets.

References

Costa, J. A., & Hero, A. O. (2004). Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets. *Proc. of European Sig. Proc. Conference (EUSIPCO)*.

Falconer, K. (2003). *Fractal geometry*. Wiley. 2nd edition.

Fukunaga, K. (1971). An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, *20*, 176–183.

Grassberger, P., & Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D*, *9*, 189–208.

Hein, M., Audibert, J.-Y., & von Luxburg, U. (2005). From graphs to manifolds - weak and strong pointwise consistency of graph Laplacians. Proceedings of the 18th Conference on Learning Theory (COLT).

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, *58*, 13–30.

Lee, J. M. (1997). *Riemannian manifolds*. Springer.

Serfling, R. (1980). *Approximation theorem in mathematical statistics*. Wiley.

Smolyanov, O. G., von Weizsäcker, H., & Wittich, O. (2004). Chernoff's theorem and discrete time approximations of Brownian motion on manifolds. Preprint, available at http://lanl.arxiv.org/abs/math.PR/0409155.

Takens, F. (1985). On the numerical determination of the dimension of an attractor. *Dynamical systems and bifurcations* (pp. 99–106).

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 2319–2323.