Manifold-Adaptive Dimension Estimation

Amir massoud Farahmand AMIR@CS.UALBERTA.CA
Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8 Canada

Csaba Szepesvári SZEPESVA@CS.UALBERTA.CA
Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8 Canada

Jean-Yves Audibert AUDIBERT@CERMICS.ENPC.FR
CERTIS - Ecole des Ponts, 19, rue Alfred Nobel - Cité Descartes, 77455 Marne-la-Vallée France

Abstract

Intuitively, learning should be easier when the data points lie on a low-dimensional submanifold of the input space. Recently there has been a growing interest in algorithms that aim to exploit such geometrical properties of the data. Oftentimes these algorithms require estimating the dimension of the manifold first. In this paper we propose an algorithm for dimension estimation and study its finite-sample behaviour. The algorithm estimates the dimension locally around the data points using nearest neighbor techniques and then combines these local estimates. We show that the rate of convergence of the resulting estimate is independent of the dimension of the input space and hence the algorithm is "manifold-adaptive". Thus, when the manifold supporting the data is low dimensional, the algorithm can be exponentially more efficient than its counterparts that are not exploiting this property. Our computer experiments confirm the obtained theoretical results.

1. Introduction

The curse of dimensionality in machine learning refers to the tendency of learning algorithms working in high-dimensional spaces to use resources (time, space, samples) that scale exponentially with the dimensionality of the space. Since most practical problems involve high-dimensional spaces, it is of uttermost importance to identify algorithms that are capable of avoiding this exponential blow-up, exploiting when additional regularity is present in the data.

One such regularity that has attracted much attention lately is when the samples lie in a low-dimensional submanifold of the possibly high-dimensional input space. Consider for example the case when the data points are images taken of a scene or object, from different angles. Although the images may contain millions of pixels, they all lie on a manifold of low dimensionality, such as 3. Another example is when the input data is enriched by adding a huge number of feature components computed from the original input components in the hope that these additional features will help some learning algorithm (generalized linear models or the "kernel trick" implement this idea).

Manifold learning research aims at finding algorithms that require less data (i.e., are more data efficient) when the data happens to be supported on a low-dimensional submanifold of the input-space. We call a learning algorithm *manifold-adaptive* when its sample-complexity depends on the intrinsic dimension of the manifold only.[1] A classical problem in pattern recognition is the estimation of the dimension of the data manifold. Dimension estimation is interesting on its own, but it is also very useful as the estimate can be fed into manifold-aware supervised learning algorithms that need to know the dimension to work efficiently (e.g., Hein 2006; Gine and Koltchinskii 2007).

In this paper we propose an algorithm for estimating the unknown dimension of a manifold from samples and prove that it is manifold-adaptive. The new algorithm belongs to the family of nearest-neighbor

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007. Copyright 2007 by the author(s)/owner(s).

[1] Of course, the sample-complexity may and will typically depend on the properties of the manifold and thus the embedding.

methods. Such methods have been considered since the late 70s. Pettis et al. (1979) suggested to average distances to k-nearest neighbors for various values of k and use the obtained values to find the dimension using an iterative method.[2] Another more recent method is due to Levina and Bickel (2005) who suggested an algorithm based on a Poisson approximation to the process obtained by counting the number of neighbors of a point in its neighborhood. In a somewhat heuristic manner they argued for the asymptotic consistency of this method. Grassberger and Procaccia (1983) suggested to estimate the dimension based on the so-called correlation dimension, while Hein and Audibert (2005) suggested a method based on the asymptotics of a smoothed version of the correlation dimension estimate. Despite the large number of works and long history, to our best knowledge no previous rigorous theoretical work has been done on the finite-sample behavior of dimension-estimation algorithms, let alone their manifold adaptivity.

2. Algorithm

The core component of our algorithm estimates the dimensionality of the manifold in a small neighborhood of a selected point. This point is then varied and results of the local estimates are combined to give the final estimate.

The local estimate is constructed as follows: Collect the observed data points into $\mathcal{D}_n = [X_1, \dots, X_n]$. We shall assume that X_i is an i.i.d. sample that comes from a distribution supported on the manifold M. Define $\eta(x, r)$ by

$$\mathbb{P}\left(X_i \in B(x,r)\right) = \eta(x,r) r^d, \tag{1}$$

or

$$\ln(\mathbb{P}\left(X_i \in B(x,r)\right)) = \ln(\eta(x,r)) + d\ln(r), \tag{2}$$

where $B(x,r) \subset \mathbb{R}^D$ is a ball around the point $x \in M$ in the Euclidean space $\mathbb{R}^D$. Our main assumption in the paper will be that in a small neighborhood of 0 the function $\eta(x,\cdot)$ is slowly varying (the assumptions on η will be made precise later). This is obviously satisfied in the commonly studied simple case when the distribution of the data on the manifold is uniform and the manifold satisfies standard regularity assumption such as those considered by Hein et al. (2006).

[2]Due to the lack of space, we cannot attempt to give a full review of existing work on dimension estimation. The interested reader may consult the papers of Kegl (2002) and Hein and Audibert (2005) which contain many further pointers.

There are two ways of using Equation (2) for estimating the dimension d. Both rely on the observation that this equation is linear in d. One approach is to fix a radius and count the number of data points within the ball $B(x,r)$, while the other approach is to calculate the radius of the smallest x-centered ball that encloses some fixed number of points. Either way, one ends up with an estimate of both $\ln(\mathbb{P}\left(X_i \in B(x,r)\right))$ and $\ln(r)$. Taking multiple measurements, we may get an estimate of d by fitting a line through these measurements, by treating η as a constant. Because η cannot be considered constant when r is large (due to the uneven sampling distribution or the curvature of the manifold), one should ideally work at small scales (small r). On the other hand, when r is too small then the measurements' variance will be high. A good estimator must thus find a good balance between the bias and the variance, making the estimation of the intrinsic dimension a non-trivial problem.

In this paper we study an algorithm in which we fix the "scale" by fixing the number of neighbors, k: the dimension is estimated from the distance to the kth nearest neighbor. In the other approach, i.e., when a scale $h = h_n$ is selected, the typical requirement is that $h_n^d n \to \infty$, or $h_n = \Omega(n^{-1/d})$. Given that d is unknown this suggests to choose $h_n = Cn^{-1/D}$. This choice, however, is too conservative and would not lead to a dimension adaptive procedure.[3] On the other hand, for the consistency of k-nearest neighbor procedures one typically requires only $k_n/n \to 0$ and $k_n \to \infty$ (these conditions are independent of d). Therefore we prefer nearest-neighbor based techniques for this task.

In order to be more specific about the method, let $X^k(x)$ be the reordering of the data such that $||X^{(k)}(x) - x|| \le ||X^{k+1}(x) - x||$ holds for $k = 1, 2, \dots, n-1$ (ties are broken randomly). Here $\|\cdot\|$ denotes the ℓ^2-norm of $\mathbb{R}^D$. Hence, $X^{(1)}(x)$ is the nearest neighbor of x in $\mathcal{D}_n$, $X^{(2)}$ is the 2nd nearest neighbor, etc. Let $\hat{r}^{(k)}(x) = \|X^k(x) - x\|$ be the distance to the kth nearest neighbor of x. In our theoretical analysis, for the sake of proofs simplicity, we use the following simple estimation method: Take $k > 2$. Denoting $\eta(x,r) \approx \eta_0$, from (2) we have

$$\begin{aligned} \ln(k/n) &\approx \ln(\eta_0) + d\ln(\hat{r}^{(k)}(x)), \\ \ln(k/(2n)) &\approx \ln(\eta_0) + d\ln(\hat{r}^{(\lceil k/2 \rceil)}(x)), \end{aligned}$$

since if n is big, $\mathbb{P}\left(X_0 \in B(x, \hat{r}^{(k)}(x))\right)$ should be close to k/n. Reordering the above equations for d, we get

$$\hat{d}(x) = \frac{\ln 2}{\ln(\hat{r}^{(k)}(x)/\hat{r}^{(\lceil k/2 \rceil)}(x))}. \tag{3}$$

[3]Of course, other options, such as using splitting or cross-validation to select h are also possible. We leave it for future work to study such algorithms.

When a center is selected from data, this point is naturally removed when calculating the point's nearest neighbors. With a slight abuse of notation, the estimate when selecting center X_i is also denoted by $\hat{d}(X_i)$.

When used at a single random data point, the variance of the estimate will be high and due to $k \ll n$ the available data is used in a highly inefficient manner. One idea is to compute the estimate at all data points and combine the results. The simplest method is to use averaging:

$$\hat{d}_{\text{avg}} = \left[\frac{\sum_{i=1}^{n}(\hat{d}(X_i)\wedge D)}{n}\right]. \tag{4}$$

Here $a \wedge b = \min(a, b)$ and $[x]$ denotes the rounded value of x (recall that the estimated dimension is a positive integer number smaller than or equal to D). Another option is to let the estimates vote:

$$\hat{d}_{\text{vote}} = \operatorname*{arg\,max}_{d' \in \mathbb{N}^+} \sum_{i=1}^{n} \mathbb{I}_{\{\hat{d}(X_i)=d'\}}. \tag{5}$$

Here $\mathbb{N}^+$ stands for the set of positive integers.

3. Main results

The purpose of this section is to show that the procedure described in the previous section is manifold-adaptive. Let $\eta_{\min} = \inf_{x\in M} \eta(x, 0)$. In what follows we will assume that the following holds:

Assumption 1. (1) *The constant $\eta_{\min}$ is positive.* (2) *For any point $x \in M$, $\eta(x, r)$ as a function of r is continuous and differentiable at any $r > 0$ and right-differentiable at $r = 0$.* (3) *There exists a positive number B', such that for any $(x, r) \in M \times [0, r_0)$, $|\frac{\partial}{\partial r}\eta(x, r)| \leq B'\eta(x, r)$.*[4](4) *There exist $r_0 > 0$ and $B > 0$ such that η satisfies $|\eta(x, r) - \eta(x, 0)| \leq B\eta(x, 0)r$, where $(x, r) \in M \times [0, r_0)$ is arbitrary.*

Since $\eta_{\min} > 0$ the manifold has to be compact. Thus the first condition on the partial derivative of η implies the second relative Lipschitzness condition when η satisfies some additional smoothness assumptions. Assumption 1 is not very restrictive: All it says is that the sampling distribution should be well-behaving in the sense that it should not change too fast. This assumption is satisfied e.g. if η is uniform on M and if M is sufficiently regular. Define $\underline{\eta}(x, r) = \min\{\eta(x, r') \,|\, 0 \leq r' \leq r\}$. From the above assumption, it is easy to see that $\underline{\eta}(x, r) \geq \eta(x, 0)(1 - Br)$ holds for any $0 \leq r < r_0 < 1$ and $x \in M$.

The following theorem is the main result of the paper:

[4]For $r = 0$ we take the right-sided derivative of η here.

Theorem 1. *Consider the estimate $\hat{d}(X_1)$. Then under Assumption 1 provided that $n \geq ck2^d$, with probability at least $1 - \delta$*

$$|\hat{d}(X_1) - d| \leq \mathbb{E}\left[C(X_1)\right] d \left(B\left(\frac{k}{n}\right)^{\frac{1}{d}} + \sqrt{\frac{\ln(4/\delta)}{k}}\right), \tag{6}$$

where $C(x) = C'((\eta_{\min})^{-\frac{1}{d}} \wedge \eta(x, 0)^{-\frac{1}{d}} (\frac{1}{2} + 2^{\frac{1}{d}})) + C''$ and where C' and C'' are universal constants that do not depend on d, D, k, n, δ and the distribution of X_1.

As promised, the result shows that the proposed method is manifold-adaptive: the estimate's convergence rate depends only on the intrinsic dimension of the manifold and not the dimension of the embedding space $\mathbb{R}^D$.

The first term of (6) bounds the bias of the estimate. By making k small, the bias can be held small. However, a small k makes the second term, which bounds the variance, large. The choice of k that optimizes the bound is $k = n^{2/(2+d)}$, giving rise to the rate $n^{-1/(2+d)}$. Since d is not available, we may e.g. choose $k = n^{1/2}$ giving the rate $O(n^{-\frac{1}{2d}\wedge\frac{1}{4}})$.

Since d is discrete valued when the upper bound of the left-hand-side of (6) becomes smaller than $1/2$, then by rounding the estimate, $\hat{d}(X_1)$, we get d. This gives rise to the following corollary:

Corollary 2. *Assume that the conditions of Theorem 1 are satisfied and $k/n < (1/2)^d$. Then*

$$\mathbb{P}\left([\hat{d}(X_1)] \neq d\right) \leq 4\exp\left(-k\left(\tfrac{1}{2} - \left(\tfrac{k}{n}\right)^{\frac{1}{d}}\right)^2\right). \tag{7}$$

Although we see that the probability of error decays exponentially fast, the result is only valid (just like the previous result) when the number of samples is large enough. This condition, however, seems inevitable since the algorithm uses estimates of the volume of balls of radii r and $r/2$ with small r. In d dimensions, if we have less than $(1/r)^d$ uniform random points in a unit cube centered around the origo then the expected number of points in the ball $B(0, r)$ is smaller than one. Thus we need at least $(2/r)^d > 2^d$ points if we want to have at least one point inside the ball of radius $r/2$.

Now, let us consider the global estimates $\hat{d}_{\text{avg}}$ and $\hat{d}_{\text{vote}}$. Using McDiarmid's version of the Hoeffding-Azuma inequality (McDiarmid, 1989; Hoeffding, 1963; Azuma, 1967) and a counting argument relying on the covering of the manifold by cones (essentially adopting the argument of Stone (1977) to manifolds) and un-

der very weak assumptions on the manifold[5] both estimates can be shown to enjoy exponentially fast rates. In particular, for some universal constants $c, c', c'' > 0$, we have

$$\mathbb{P}\left(\hat{d}_{\text{vote}} \neq d\right) \leq e^{-\frac{c'n}{(c^d k)^2}}, \quad (8)$$

$$\mathbb{P}\left(\hat{d}_{\text{avg}} \neq d\right) \leq e^{-\frac{c''n}{(Dc^d k)^2}}. \quad (9)$$

From these bounds we can conclude that voting should be preferred since in the case of the averaging bound the rate of convergence depends on D (though only in a very mild, polynomial way). However, our experimental results seems to suggest that the estimate for the averaging method is probably too conservative as it tends to produce better results than voting, at least for the particular dataset and choice of parameters that we considered.

Due to the lack of space the proof of this statement is deferred to the full version of this paper, but the proof of Theorem 1 which is the key to this proof as well is given in the next section.

4. Proofs

Theorem 1 is proven in a series of lemmas. First, let us remark that due to the independence of samples, it suffices to show the result for any deterministically selected point $x \in M$. Hence, in what follows we will consider this case. For the sake of brevity we shall suppress the dependence on x in the rest of this section.

Let $p = k/n$. By the triangle inequality,

$$|d - \hat{d}| \leq |d - d(p)| + |d(p) - \hat{d}|. \quad (10)$$

Here $d(p)$ is defined by

$$d(p) = \frac{\ln 2}{\ln(r_p/r_{p/2})}. \quad (11)$$

By (2), if $\eta(x, r_p) = \eta(x, r_{p/2})$ were hold true then $d(p) = d$ would hold. Hence, the source of the error $|d - d(p)|$ is the change in η in the neighborhood of x. By Assumption 1 on η, we can make this error controllable.

The following statement follows by elementary considerations (the proofs of these lemmas are given in the appendix):

Lemma 1. *$|d - d(p)| \leq CBdr_p$ provided that $r_p < (0.2/B) \wedge r_0$. Here $C \leq 8$ is a universal constant.*

[5]Bounded curvatures and that the manifold is not self-approaching are the main assumptions.

It is easy to see that $r_p \leq (\eta_{\min})^{-1/d}(k/n)^{1/d}$. When the density is non-uniform this estimate might be very conservative. We prefer a bound that depends on the properties of the density in the vicinity of x. Using the observation stated after Assumption 1, we get the following result:

Lemma 2. *Assume that $Br_p < (0.5 \wedge r_0)$. Then*

$$r_p \leq ((\eta_{\min})^{-\frac{1}{d}} \wedge \eta(x,0)^{-\frac{1}{d}}\left(\tfrac{1}{2} + 2^{\frac{1}{d}}\right)) \left(\frac{k}{n}\right)^{\frac{1}{d}}.$$

Chaining the inequalities of Lemma 1 and Lemma 2 we get that $|d - d(p)| \leq CdBr_p \leq C((\eta_{\min})^{-\frac{1}{d}} \wedge \eta(x,0)^{-\frac{1}{d}}\left(\frac{1}{2} + 2^{\frac{1}{d}}\right))Bd\left(\frac{k}{n}\right)^{\frac{1}{d}}$.

The second term of (10), $|d(p) - \hat{d}|$, is bounded by relating it to the relative errors of estimating r_p by $\hat{r}^{(k)}$ (and $r_{p/2}$ by $\hat{r}^{(\lceil k/2 \rceil)}$).

Lemma 3. *If $d(p)'$ is defined by $d(p)' = \ln(2)/\ln(r'_p/r'_{p/2})$ for some positive quantities r'_p and $r'_{p/2}$ then for*

$$\alpha = \max\left(\left|\frac{r'_p}{r_p} - 1\right|, \left|\frac{r'_{p/2}}{r_{p/2}} - 1\right|\right),$$

$$|d(p) - d(p)'| \leq Cd^2\alpha \quad (12)$$

provided that $\alpha \leq c/d$ and $r_p < (0.2/B) \wedge r_0$, where c is a fixed universal constant.

Again, the proof of this lemma uses elementary analysis. By this lemma, in order to get a bound on $|d(p) - \hat{d}|$, we need to analyze the relative error of estimating r_p by $\hat{r}^{(k)}$. We get the following lemma by using Assumption 1.

Lemma 4. *Assume that $r_p < (4B')^{-1} \wedge r_0$ and $\alpha \leq 1/(4(d+1))$. Then*

$$\mathbb{P}\left(\hat{r}^{(k)} \leq r_p(1-\alpha)\right) \leq \exp(-C_1 k\alpha^2(d - \tfrac{1}{4})^2) \quad (13)$$

$$\mathbb{P}\left(\hat{r}^{(k)} \geq r_p(1+\alpha)\right) \leq \exp(-C_2 k\alpha^2(d - \tfrac{1}{4})^2) \quad (14)$$

where $C_1 = \frac{3}{8}(1 - \frac{d-2}{4(d+1)})(1 - \frac{3}{16(d+1)})$, $C_2 = \frac{3e^{-1/4}}{8}(1 - \frac{1}{8(d+1)})(1 - \frac{1}{16(d+1)(d-1/4)})^2$.

The proof of this lemma relies on Bernstein's inequality. According to these bounds, with probability at least $1 - \delta$,

$$\max\left\{\left|\frac{\hat{r}^{(k)}}{r_p} - 1\right|, \left|\frac{\hat{r}^{(\lceil k/2 \rceil)}}{r_{p/2}} - 1\right|\right\} \leq C_3 \frac{1}{d}\sqrt{\frac{\ln(4/\delta)}{k}}$$

with a suitable universal constant C_3. Hence,

$$|d - \hat{d}| \leq C(x)d\left(B\left(\frac{k}{n}\right)^{\frac{1}{d}} + \sqrt{\frac{\ln(4/\delta)}{k}}\right)$$

holds with probability at least $1-\delta$, which proves Theorem 1.

5. Experimental results

The purpose of this section is to provide some experimental evidence on the performance of our algorithm. We investigated the influence of the following factors: *(i)* number of samples (n), *(ii)* the manifold's dimension (d) *(iii)* the embedding space's dimensionality (D), *(iv)* the number of centers used when combining the local estimates (m),[6] *(v)* the number of neighbors (k) *(vi)* noise level. Due to the lack of space here we only present results for *(i)–(iii)*. The other results will be given in the longer version of this paper, here we remark only that according to our experience the algorithm's performance degrades gracefully when noise, not respecting the manifold is added to the data. Noise is the Achilles heel of manifold-aware algorithms as it changes the support of the sampling distribution. We leave it for future work to study the behaviour of manifold-aware algorithms in the presence of noise.

The default setting of the parameters are $m = n/2$ and $k = [2 \ln n]$. These parameter settings were used in all the experiments.[7] Except for the real-world dataset, we performed the measurements by repeating the calculations 100 times, for 100 different randomizations of the datasets considered. We report average errors and the percentages when a correct estimate was obtained.

The datasets used were essentially identical to those used by Hein and Audibert (2005), i.e., they include some standard datasets such as spheres of various dimensionality and some high-curvature datasets for which dimension estimation is quite challenging. In the case of *shperes* the data points are sampled uniformly from a d-dimensional sphere S^d embedded in $\mathbb{R}^{d+1}$. The *sinusoid* dataset is a one dimensional oscillating sinusoid on the circle in $\mathbb{R}^3$. The data points come from the manifold $M = \{(\sin(u), \cos(u), \frac{1}{10}\sin(10u)) \,|\, u \in [0, 2\pi)\}$, where the samples are obtained by drawing random points uniformly at random in the interval $[0, 2\pi)$. The *10-Möbius strip* is a two dimensional submanifold in $\mathbb{R}^3$, created by twisting a two dimensional rectangle 10 times. Data points are obtained by sampling points (U, V) uniformly on $[-1, 1] \times [0, 2\pi)$ and returning $x_1(U, V) = (1 + \frac{U}{2}\cos(5V))cos(V)$, $x_2(U, V) = (1 + \frac{U}{2}\cos(5V))\sin(V)$, and $x_3(U, V) = \frac{U}{2}\sin(5V)$. We used two other datasets: "Swiss roll" and the ISOMAP Face datset. The Swiss roll is a two dimensional manifold embedded in $\mathbb{R}^3$ Levina and Bickel (2005). ISOMAP Face consists of 698 64×64 images (256 gray levels) of a face sculpture Tenenbaum et al. (2000). For this dataset we obtained an estimate of four when using $\hat{d}_{\text{avg}}$, while we got an estimate of 3 when using $\hat{d}_{\text{vote}}$. Earlier results by others suggest that the intrinsic dimensionality is 3.

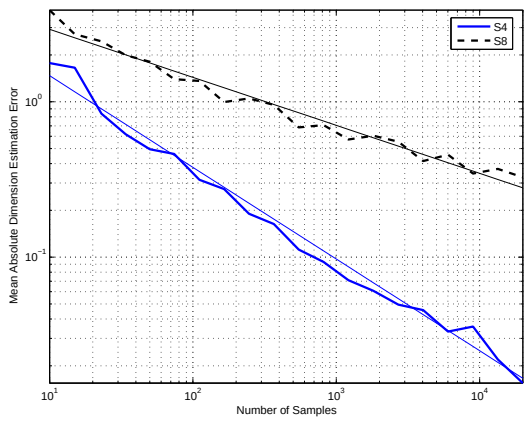

Figure 1. Average absolute error of the dimension estimate for different sample sizes for S^4 and S^8. Note the logarithmic scales. The straight lines show lines fitted to the measured curves.

Results for the different artificial datasets when the number of data points (n) is varied are given Table 1. As expected, the number of samples required for an accurate estimate increases with the intrinsic dimension of the manifold. We can conclude that (at least for the parameter settings considered) the averaging method performs better than the voting method. In particular, voting seems to have troubles when the number of datapoints is small or the intrinsic dimension is higher. Therefore in what follows we consider only the averaging method. Overall the performance seems comparable to those reported by Hein and Audibert (2005).

Figure 1 shows the average absolute error measured as the number of samples for S^4 and S^8. It turns out that the error behave roughly as $O(n^{-c/d})$ with $c = 2.4$.

One crucial property of our bounds is that they do not depend (explicitly) on the dimension of the embedding space $\mathbb{R}^D$. In order to test this we picked the 10-Möbius dataset and added additional dimensions

[6] In the theoretical analysis we assumed that $m = n$ (see Equations (4),(5)). However, one can also select datapoints participating in the computation in a random fashion (by sampling data points uniformly with replacement). The hope is that an equivalently good estimate can be obtained by less work.

[7] Note that according to the theory developed this choice of k is inferior to e.g. $k = n^{1/2}$. However, $k = O(\ln n)$ yields much less computation and was therefore preferred in the experiments.

Table 1. Percentage of correct dimension estimates for different sample sizes. The first values in a cell (not in parentheses) is for the averaging method, while those in parentheses are for the voting method.

DATA SET	N=50	N=100	N=500	N=1000	N=5000
S^1	98 (99)	100 (100)	100 (100)	100 (100)	100 (100)
S^3	75 (19)	95 (20)	100 (15)	100 (19)	100 (62)
S^5	33 (5)	50 (10)	100 (9)	98 (2)	100 (0)
S^7	18 (2)	17 (3)	57 (1)	54 (1)	100 (0)
SINUSOID	92 (98)	100 (100)	100 (100)	100 (100)	100 (100)
10-MÖBIUS	69 (47)	13 (74)	100 (98)	100 (99)	100 (100)
SWISS ROLL	62 (71)	49 (91)	88 (96)	100 (100	100 (100)

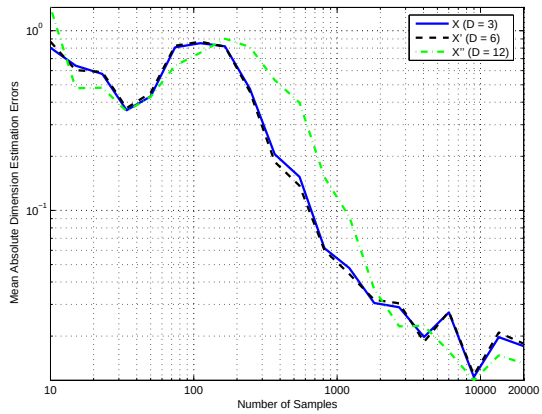

Figure 2. The effect of extrinsic dimension on the mean absolute dimension estimation error for 10-Möbius problem. For more information see the text.

("features") to it by using two functions, ϕ_1 and ϕ_2.[8] The results for the original manifold and M_1, M_2 are shown in Figure 2. We see that the behavior of the error-curves is almost identical in all the three cases.[10]

This figure reinforces us in that, as predicted by the theory, the embedding dimension has essentially no effect on the quality of estimates.

6. Conclusions

In this paper, we introduced an algorithm for estimating the intrinsic dimension of a manifold, and analyzed its finite-sample convergence properties. We showed that the method is manifold-adaptive: the convergence behavior of the method is determined by the dimension of the manifold, and not the dimension of the embedding space. In addition to the theoretical analysis, we examined our method on several test problems. It was shown that the performance of the method is comparable to other works. As for future work, it would be interesting to prove manifold-adaptivity results for other learning problems, such as regression or classification. Another interesting open question is if manifold learning can succeed at all in the presence of noise.

[8] In particular, we let $\phi_1 : \mathbb{R}^3 \to \mathbb{R}^6$ defined by $\phi_1(x) = (x, \sin(x))$, and $\phi_2 : \mathbb{R}^3 \to \mathbb{R}^{12}$ defined by $\phi_2(x) = (x, \sin(x), x^2, x^3)$.[9] Clearly, $M_j = \{\phi_j(x) \,|\, x \in M\}$ has the same dimensionality as M $(j = 1, 2)$, but the extrinsic dimensionality of the data points, $X_i' = \phi_1(X_i)$, $X_i'' = \phi_2(X_i)$ is increased.

[10] The differences in the case of the points (X_i'') can probably be explained by the additional curvature introduced by the non-linear functions.

Acknowledgments

Csaba Szepesvári greatly acknowledges the support received through the Alberta Ingenuity Center for Machine Learning (AICML) and the Computer and Automation Research Institute of the Hungarian Academy of Sciences.

A. Proof of Lemma 1

We will need the following result that we state without a proof :

Proposition 5. *If* $x < 1$ *then* $2x \le \ln((1+x)/(1-x)) \le \frac{2x}{(1-x)(1+x)}$.

We need to prove the following: Let $p > 0$ and assume that $r_p \le 0.2/B$. Then for $d(p) = \ln(2)/\ln(r_p/r_{p/2})$, $|d - d(p)| \le CdBr_p$, where $C \le 8$.

Proof. Let $r_1 = r_p$, $r_2 = r_{p/2}$ and $\eta_1 = \eta(x, r_p)$, $\eta_2 = \eta(x, r_{p/2})$. Note that $d = (\ln(2) + \ln(\eta_1/\eta_2))/\ln(r_1/r_2)$. Hence $|d - d(p)| = \ln(\eta_1/\eta_2)/\ln(r_1/r_2)$ and thus we plan to upper bound the numerator and lower bound the denominator.

Let $\eta_0 = \eta(x, 0)$. By Assumption 1, $\eta_1/\eta_2 \ge \eta_0(1 - Br_1)/(\eta_0(1 + Br_2)) \ge (1 - Br_1)/(1 + Br_1)$. Similarly, $\eta_1/\eta_2 \le (1 + Br_1)/(1 - Br_1)$. Since by assumption

$Br_1 \leq 0.2$, taking logarithms and using the upper bound in Proposition 5 we get

$$|\ln(\eta_1/\eta_2)| \leq \frac{2Br_1}{(1-Br_1)(1+Br_1)}. \tag{15}$$

Now, using the identities $p = \eta_1 r_1^d$, $p/2 = \eta_2 r_2^d$ we get $(r_1/r_2)^d = 2\,\eta_2/\eta_1 \geq 2(1-Br_1)/(1+Br_1)$. Taking logarithms and using the lower bound in Proposition 5 we get $d\ln(r_1/r_2) \geq \ln 2 - 2Br_1/((1-Br_1)(1+Br_1))$. Combining the inequalities obtained gives $|d - d(p)| \leq 2dBr_1/((1-Br_1)(1+Br_1)\ln 2 - 2Br_1)$. Using the assumption $r_1 = r_p \leq 0.2/B$ allows us to lower bound the denominator here by a constant, yielding the final result. □

B. Proof of Lemma 2

Proof. Since $p = k/n = \eta(x, r_p) r_p^d$, $r_p = (\eta(x, r_p))^{-\frac{1}{d}} (k/n)^{\frac{1}{d}} \leq (\eta(x, r_p))^{-\frac{1}{d}} (k/n)^{\frac{1}{d}} \leq (\eta(x, 0)(1 - Br_p))^{-\frac{1}{d}} (k/n)^{\frac{1}{d}}$, where we have used that $Br_p < 1$. Using the elementary inequality $(1-x)^{-1/d} \leq 1 + 2^{1+\frac{1}{d}} x$ which holds for $0 \leq x \leq 0.5$ and assuming that $Br_p \leq 0.5$ we get $r_p \leq (k/n)^{\frac{1}{d}} \eta(x, 0)^{-\frac{1}{d}} (1 + 2^{1+\frac{1}{d}}) Br_p \leq (k/n)^{\frac{1}{d}} \eta(x, 0)^{-\frac{1}{d}} (1 + 2^{1+\frac{1}{d}})/2$. Also, from $r_p = (\eta(x, r_p))^{-\frac{1}{d}} (k/n)^{\frac{1}{d}}$ we get $r_p \leq (\eta_{\min})^{-\frac{1}{d}} (k/n)^{\frac{1}{d}}$. Combining this with the previous inequality for r_p gives the result. □

C. Proof of Lemma 3

Let $\alpha = \max\left(\left|\frac{r_p'}{r_p} - 1\right|, \left|\frac{r_{p/2}'}{r_{p/2}} - 1\right|\right)$. The lemma states that $|d(p) - d(p)'| \leq Cd^2\alpha$ provided that $r_p \leq 0.2/B$ and $\alpha < 0.5$.

Proof. Let $e(p) = 1/d(p)$, $e(p)' = 1/d(p)'$. Let $\epsilon = |d(p) - d(p)'|$, $\gamma = |e(p) - e(p)'|$. Then $\epsilon = |d(p) - d(p)'| = \frac{\gamma}{e(p)\,e(p)'} = d(p)d(p)'\gamma \leq d(p)(d(p) + \epsilon)\gamma$. Ordering this for ϵ, provided that $\gamma d(p) < 1$ we get that $\epsilon \leq \gamma d(p)^2/(1 - \gamma d(p))$. Since by Lemma 1, $d(p) \leq d + CdBr_p$, assuming that $d\gamma(1 + Br_p) < 1$ we may further bound ϵ by

$$\epsilon \leq \gamma d^2 \frac{(1 + CBr_p)^2}{1 - \gamma d(1 + CBr_p)}. \tag{16}$$

Hence it suffices to show that $\gamma \leq C\alpha$ since then for α sufficiently small the denominator can be bounded from below with a positive constant and the whole expression will be bounded by $O(d^2\alpha)$ as promised.

By the definition of $e(p)$ and $e(p)'$, $e(p)' - e(p) = 1/\ln(2)\,\ln((\frac{r_p'}{r_p})/(\frac{r_{p/2}'}{r_{p/2}}))$. By the definition of α, $\frac{r_p'}{r_p} \leq 1 + \alpha$ and $\frac{r_{p/2}'}{r_{p/2}} \geq 1 - \alpha$. Hence, since by assumption $\alpha < 1$, $e(p)' - e(p) = 1/\ln(2)\,\ln((1+\alpha)/(1-\alpha))$. Using Proposition 5, we thus get $e(p)' - e(p) = 2/\ln(2)\,\alpha/((1+\alpha)(1-\alpha)$. Since $(1+\alpha)(1-\alpha)$ is decreasing in α, we may upper bound the right-hand side by $C\alpha$ with an appropriate positive constant C, thus finishing the proof. □

D. Proof of Lemma 4

Introduce $\lambda(x, r) = \mathbb{P}(X_1 \in B(x, r))$. Since x is fixed, in what follows for the sake of brevity we will drop x from the arguments of λ. Similarly, we drop x from $\eta(x, r)$. We need the following properties of λ:

Proposition 6. *Let $r > 0$, $0 \leq \epsilon < r$, $0 \leq \alpha < 1$. The following inequalities hold for λ:*

$$\lambda(r) - \lambda(r-\epsilon) \geq \eta(r)(1 - B'\epsilon)(r-\epsilon)^{d-1}(d - B'r)\epsilon, \tag{17}$$

$$\lambda(r+\epsilon) - \lambda(r) \geq \eta(r)(1 - B'\epsilon)r^{d-1}(d - B'(r+\epsilon))\epsilon, \tag{18}$$

$$\lambda(r(1-\alpha)) \geq \lambda(r)(1 - \alpha(1 + B'\alpha r)(d + B'r)), \tag{19}$$

$$\lambda(r-\epsilon) \leq \eta(r)(1 + B'\epsilon)(r-\epsilon)^d, \tag{20}$$

$$\lambda(r+\epsilon) \leq \eta(r)(1 + B'\epsilon)(r+\epsilon)^d. \tag{21}$$

Proof. Note that (18) follows immediately from (17). Inequalities (20),(21) follow directly from $\lambda(r) = \eta(r)r^d$ and Assumption 1. Hence, it remains to prove (17) and (19). Let us start with (17).

Since η is differentiable, $\lambda(r) = \eta(r)r^d$ is differentiable, too. Further, $\lambda'(r) = \eta'(r)r^d + \eta(r)dr^{d-1}$ and hence using Assumption 1, $\eta(r)r^{d-1}(d - B'r) \leq \lambda'(r) \leq \eta(r)r^{d-1}(d + B'r)$.

Let $0 < a, b$, $v = a \wedge b$, $u = a \vee b$. Since by assumption λ is differentiable in (v, u) and continuous on $[v, u]$, by the Mean-Value Theorem, $\lambda(a) - \lambda(b) = \lambda'(\xi)(a - b)$ where ξ is some number in (v, u). Using the bound derived for λ', with the choice $a = r$, $b = r - \epsilon$ we get $\lambda(r) - \lambda(r-\epsilon) = \lambda'(\xi)\epsilon \geq \eta(\xi)\xi^{d-1}(d - B'\xi)$ for some $\xi \in (r-\epsilon, r)$. Using $\eta(\xi) \geq \eta(r)(1 - B'\epsilon)$ which holds by Assumption 1, we get $\lambda(r) - \lambda(r-\epsilon) \geq \eta(r)(1 - B'\epsilon)(r-\epsilon)^{d-1}(d - B'r)$, which proves (17).

Let us show (19). Let $\epsilon = \alpha r$. By Taylor's theorem there exists $\xi \in (r-\epsilon, r)$, such that $\lambda(r-\epsilon) = \lambda(r) - \lambda'(\xi)\epsilon$. Hence, $\lambda(r-\epsilon) \geq \lambda(r) - \eta(\xi)\xi^{d-1}(d + B'\xi)\epsilon \geq \lambda(r) - \eta(r)(1 + B'\epsilon)r^{d-1}(d + B'r)\alpha r = \lambda(r)(1 - \alpha(1 + B'\alpha r)(d + B'r))$, where we have used the bound on λ' and Assumption 1. □

Now, let us prove Lemma 4.

Proof. Let $r'_p < r_p$ be some positive number. Then

$$\mathbb{P}\left(\hat{r}^{(k)} \leq r'_p\right) = \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{I}_{\{X_i \in B(x,r'_p)\}} \geq k\right)$$
$$= \mathbb{P}\left(\tfrac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_i \in B(x,r'_p)\}} - \lambda(r'_p) \geq k/n - \lambda(r'_p)\right)$$
$$\leq \exp\left(-\tfrac{n}{2} F(\lambda(r_p), \lambda(r'_p))\right), \qquad (22)$$

where $F(\lambda_1, \lambda_2) = (\lambda_1 - \lambda_2)^2/(\lambda_2(1-\lambda_2) + \frac{1}{3}(\lambda_1 - \lambda_2))$. The last inequality in (22) follows from Bernstein's inequality thanks to $k/n - \lambda(r'_p) = p - \lambda(r'_p) = \lambda(r_p) - \lambda(r'_p) > 0$ and $\mathrm{Var}\left[\mathbb{I}_{\{X_1 \in B(x,r)\}}\right] = \lambda(r)(1 - \lambda(r))$. Similarly, if $r'_p > r_p$ then

$$\mathbb{P}\left(\hat{r}^{(k)} \geq r'_p\right) \leq \exp\left(-\frac{n}{2} F(\lambda(r'_p), \lambda(r_p))\right). \qquad (23)$$

Choose $r'_p = r_p(1-\alpha)$. Then (22) gives an upper bound on $\mathbb{P}\left(\hat{r}^{(k)} \leq r_p(1-\alpha)\right)$. We further bound this by lower bounding the numerator of $F(\lambda(r_p), \lambda(r'_p))$ and upper bounding its denominator. For the lower bound we use (17) to get $\lambda(r_p) - \lambda(r'_p) \geq \eta(r_p)(1 - B'\alpha r_p) r_p^{d-1}(1-\alpha)^{d-1}(d - B'r_p) r_p \alpha = \alpha(1-\alpha)^{d-1} r p^d \eta(r_p)(1 - B'\alpha r_p)(d - B'r_p)$. The first term in the denominator is upper bounded by $\lambda(r'_p)$ (since $1 - \lambda(r'_p) \leq 1$). We now show that the second term can be bounded from above by $\lambda(r'_p)$ thanks to the assumptions $\alpha \leq 1/(4(d+1))$ and $B'r_p \leq 1$. Indeed, by (19) of Proposition 6, $\lambda(r_p(1-\alpha)) \geq \lambda(r_p)(1 - \alpha(1 + B'\alpha r_p)(d + B'r_p)) \geq 1/2\lambda(r_p)$, where the last inequality follows since $(1 + B'\alpha r_p)(d + B'r_p) \leq 2(d+1)$. Hence, the denominator can be upper bounded by $4/3\lambda(r'_p)$. Now using (20), this can be further upper bounded by $4/3\eta(r_p)(1 + B'r_p\alpha)(r_p(1-\alpha))^d$. Combining these bounds gives

$$\frac{n}{2} F(\lambda(r_p), \lambda(r'_p)) \geq$$
$$\frac{3n}{8} \frac{\alpha^2(1-\alpha)^{2d-2}(r_p)^{2d}\eta^2(r_p)(1-B'\alpha r_p)^2(d-B'r_p)^2}{\eta(r_p)(1+B'r_p\alpha)(r_p(1-\alpha))^d}$$
$$= \frac{3n}{8} \frac{\alpha^2(1-\alpha)^{d-2}(r_p)^{d}\eta(r_p)(1-B'\alpha r_p)^2(d-B'r_p)^2}{(1+B'r_p\alpha)}$$
$$\geq \frac{3k}{8}\alpha^2(1-\alpha)^{d-2}(1 - 3B'\alpha r_p)(d - B'r_p)^2$$
$$\geq \frac{3k}{8}\alpha^2(1-(d-2)\alpha)(1 - 3B'\alpha r_p)(d - B'r_p)^2$$
$$\geq \frac{3k}{8}\alpha^2(1-(d-2)\alpha)(1 - \tfrac{3}{16(d+1)})(d - \tfrac{1}{4})^2$$

where to get the first inequality we used $(1-x)/(1+x) \geq 1 - 2x$ which holds for $x > 0$, $(1-x)(1-2x) \geq (1-3x)$, and $\eta(r_p)r_p^d = p = k/n$, which holds thanks to the definition of r_p. In the last inequality we used the assumption $B'r_p < 1/4$ and $\alpha < 1/(4(d+1))$. This finishes the proof of the bound of (13).

For bounding $\mathbb{P}\left(\hat{r}^{(k)} \geq r_p(1+\alpha)\right)$ we start with (23). Again, we lower bound the numerator. This time, we use (18) to get $\lambda(r_p(1+\alpha)) - \lambda(r_p) \geq \eta(r_p)(1 - B'r_p\alpha)r_p^d(d - B'(r_p(1+\alpha)))\alpha$. The denominator of (23) is upper bounded by $4/3\lambda(r_p(1+\alpha)) \leq 4/3\eta(r_p)(1 + B'r_p\alpha)(r_p(1+\alpha))^d \leq 4/3\eta(r_p)(1+B'r_p\alpha)r_p^d e^{1/4}$, which follows by (21) and since $(1+\alpha)^d \leq (1 + 1/(4(d+1)))^{4(d+1)\times\frac{d}{4(d+1)}} \leq e^{1/(4+1/d)} \leq e^{1/4}$. Hence, the exponent of (23), $n/2F(\lambda(r'_p), \lambda(r_p))$, is bounded from below by

$$\frac{3n}{8} \frac{(\eta(r_p)(1 - B'r_p\alpha)r_p^d(d - B'(r_p(1+\alpha)))\alpha)^2}{\eta(r_p)(1 + B'r_p\alpha)r_p^d}$$
$$\geq \frac{3ne^{-1/4}}{8} \frac{\alpha^2\eta(r_p)r_p^d(1-B'r_p\alpha)(d-B'(r_p(1+\alpha)))^2}{(1+B'r_p\alpha)}$$
$$\geq \frac{3ke^{-1/4}}{8}\alpha^2(1 - \tfrac{1}{8(d+1)})(d - \tfrac{1}{4} - \tfrac{1}{16(d+1)}))^2,$$

where we again used $(1-x)/(1+x) \geq 1 - 2x$, $k/n = \eta(r_p)r_p^d$, $B'r_p \leq 1/4$, $\alpha \leq 1/(4(d+1))$. This finishes the proof of (14). □

References

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Math. J.*, 19(2):357–367.

Gine, E. and Koltchinskii, V. (2007). Empirical graph Laplacian approximation of Laplace-Beltrami operators: Large sample results. In *Proc. of the* 4th *Int. Conf. on High Dimensional Probability.* to appear.

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

Hein, M. (2006). Uniform convergence of adaptive graph-based regularization. In *COLT-2006*, pages 50–64.

Hein, M. and Audibert, J.-Y. (2005). Intrinsic dimensionality estimation of submanifolds in Euclidean space. In *ICML-2005*, pages 289–296.

Hein, M., Audibert, J.-Y., and von Luxburg, U. (2006). From graphs to manifolds - weak and strong pointwise consistency of graph Laplacians. submitted.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

Kegl, B. (2002). Intrinsic dimension estimation using packing numbers. In *NIPS-15*, pages 681–688.

Levina, E. and Bickel, P. J. (2005). Maximum likelihood estimation of intrinsic dimension. In *NIPS-17*, pages 777–784.

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in Combinatorics*, pages 148–188.

Pettis, K. W., Bailey, T. A., Jain, A. K., and Dubes, R. C. (1979). An intrinsic dimensionality estimator from near-neighbor information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:25–37.

Stone, C. (1977). Consistent nonparametric regression. *Annals of Statistics*, 5(4):595–620.

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