Geometry-Aware Maximum Likelihood Estimation of Intrinsic Dimension

Marina Gomtsyan[1,2] MARINA.GOMTSYAN@SKOLTECH.RU
Nikita Mokrov[1,3] N.MOKROV@SKOLTECH.RU
Maxim Panov[1] M.PANOV@SKOLTECH.RU
Yury Yanovich[1,4] Y.YANOVICH@SKOLTECH.RU
[1] *Skolkovo Institute of Science and Technology, Moscow 121205, Russia*
[2] *Higher School of Economics, Moscow 101000, Russia*
[3] *Moscow Institute of Physics and Technology, Moscow 141701, Russia*
[4] *Institute for Information Transmission Problems, Moscow 127051, Russia*

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

The existing approaches to intrinsic dimension estimation usually are not reliable when the data are nonlinearly embedded in the high dimensional space. In this work, we show that the explicit accounting to geometric properties of unknown support leads to the polynomial correction to the standard maximum likelihood estimate of intrinsic dimension for flat manifolds. The proposed algorithm (GeoMLE) realizes the correction by regression of standard MLEs based on distances to nearest neighbors for different sizes of neighborhoods. Moreover, the proposed approach also efficiently handles the case of nonuniform sampling of the manifold. We perform a series of experiments on various synthetic and real-world datasets. The results show that our algorithm achieves state-of-the-art performance, while also being robust to noise in the data and competitive computationally.

Keywords: Intrinsic dimension estimation, Manifold learning, Maximum likelihood estimation.

1. Introduction

Dimensionality reduction is one of the critical steps of data analysis. The proper application of dimensionality reduction allows to decrease the required space for data storage and increase the speed of the data processing by machine learning algorithms. Most importantly, it often significantly improves the performance of many machine learning algorithms, which often rapidly degrades in high dimensions.

The majority of existing dimensionality reduction methods require the true dimension of the data as an input parameter. Not surprisingly, the problem of estimating the true dimension of the data known as *intrinsic dimension* estimation is a well-studied problem, and numerous specialized intrinsic dimension estimation methods exist (Bailey et al., 1979; Grassberger and Procaccia, 1983; Levina and Bickel, 2005; Hein and Audibert, 2005; Lombardi et al., 2011; Little et al., 2012; Ceruti et al., 2014; Johnsson et al., 2015; Granata and Carnevale, 2016). In addition, some dimensionality reduction methods such as principal component analysis (PCA) (Jolliffe, 1986) can be modified for estimating the intrinsic dimension, see (Fukunaga and Olsen, 1971; Bishop, 1998; Tipping and Bishop, 1999). However, the existing intrinsic dimension estimation approaches have some disadvantages: some fail on data with a non-linear structure, some require a large number of observations for efficient performance, others are computationally expensive (Campadelli et al., 2015).

In this paper, we introduce a new efficient method for intrinsic dimension estimation. We base our approach on the *Maximum likelihood estimation of intrinsic dimension* (MLE) (Levina and Bickel, 2005) which is one of the most commonly used methods due to its simplicity and computational efficiency. However, when the true dimension of the data is large, the MLE method is known to underestimate it significantly Ceruti et al. (2014). The explanation of this fact is contained in the key assumption of the method: the local neighborhood of each point is approximated by a linear subspace with a uniform density. Since real-world data often lie on or near to a nonlinear manifold with an arbitrary density, such an assumption is restrictive and leads to the bias in the procedure. To overcome the problems mentioned above we propose a data-driven approach, which explicitly introduces the correction for non-uniformity of density and nonlinearity of manifold into the likelihood and estimates unknown parameters by regression with respect to the radius of the neighborhood.

Our main contributions are the following:

- We propose a new intrinsic dimension estimation method *Geometry-aware maximum likelihood estimation of intrinsic dimension* (GeoMLE). Our approach takes into consideration the geometric properties of a manifold and corrects for a nonuniform sampling.
- GeoMLE shows the state-of-the-art results in the estimation of intrinsic dimension. In a series of experiments, GeoMLE outperforms MLE (Levina and Bickel, 2005) and other intrinsic dimension estimators. In particular, our estimator gives accurate results for datasets in high dimensions, in case of which the performance of many competitors is rather weak. Moreover, the proposed method is shown to perform well in the case of nonuniform sampling of the manifold, while being also robust to noise and competitive computationally.

In the next section, we describe the MLE approach in detail. Section 3 provides the idea of the correction for nonlinear geometry and nonuniform density and introduces the resulting GeoMLE algorithm. The experimental evaluation of the method is given in Section 4, while the review of the related literature is given in Section 5. Section 6 concludes the study.

2. Maximum Likelihood Estimator of Intrinsic Dimension

Consider data manifold of unknown dimension m:

$$\mathbb{X} = \{x = g(b) \in \mathbb{R}^p \colon b \in \mathbb{B} \subset \mathbb{R}^m\},$$

where $(\mathbb{B}, g)$ is a single coordinate chart embedded into an ambient p-dimension space $\mathbb{R}^p$, such that $m \leq p$. The mapping g is a one-to-one mapping from an open bounded set $\mathbb{B} \subset \mathbb{R}^p$ to manifold $\mathbb{X} = g(\mathbb{B})$, with a differentiable inverse map $g^{-1} \colon \mathbb{X} \to \mathbb{B}$. The manifold $\mathbb{X}$ is unknown, and a finite data set $D = \{X_1, \dots, X_n\} \subset \mathbb{X} \subset \mathbb{R}^p$ is sampled from a distribution with an unknown density $f(x)$.

Levina and Bickel (2005) suggested to consider the binomial process

$$N(t, x) = \sum_{i=1}^{n} \mathbb{1}\{X_i \in S_x(t)\}, \quad 0 \le t \le R,$$

where $S_x(t)$ is a ball of radius t centered at x. They propose to approximate this process by Poisson process $N_\lambda(t, x)$ with rate $\lambda_{m,\theta}(t)$ and $\theta = \log f(x)$. Suppressing the dependence on x, the log-likelihood of the observed process $N_\lambda(t)$ is

$$L_\lambda(m, \theta) = \int_0^R \log \lambda_{m,\theta}(t) dN_\lambda(t) - \int_0^R \lambda_{m,\theta}(t) dt. \tag{1}$$

The key idea of MLE (Levina and Bickel, 2005) is to fix a point x and for an unknown smooth density f on $\mathbb{X}$ assume that $f(z) \approx \text{const}$ in a ball $z \in S_x(R) \subset \mathbb{R}^p$ of small radius R centered at x, while the intersection of $\mathbb{X}$ and $S_x(R)$ is approximated by m-dimensional ball $S_x^m(R)$. Then, the observations are treated as a Poisson process in $S_x^m(R) \subset \mathbb{R}^m$. The rate of the Poisson process for the resulting approximation is

$$\hat{\lambda}_{m,\theta}(t) = f(x) V_m m t^{m-1}, \tag{2}$$

where V_m is the volume of the unit sphere in $\mathbb{R}^m$.

Let $T_k(x)$ be the Euclidean distance from a fixed point x to its k-th nearest neighbor in the sample D. Levina and Bickel (2005) prove that the intrinsic dimension estimate for a manifold $\mathbb{X}$ at a point x obtained by maximizing the likelihood (1) with a rate (2) is equal to

$$\hat{m}_R(x) = \left(\frac{1}{N_{\hat{\lambda}}(R, x)} \sum_{j=1}^{N_{\hat{\lambda}}(R,x)} \log \frac{R}{T_j(x)} \right)^{-1}.$$

For numerical calculations it might be more convenient to fix the number of neighbors k rather than the radius of the ball R. Then the MLE reads as

$$\hat{m}_k(x) = \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(x)}{T_j(x)} \right)^{-1},$$

where k is the number of neighbors. The authors suggest to construct the resulting estimator by simple averaging over a range of small to moderate values of nearest neighbours $k \in [k_1, k_2]$.

3. Geometry-Aware MLE of Intrinsic Dimension

Levina and Bickel (2005) approximate the local neighborhood of each point by a linear subspace with a uniform density. However, usually, real-world data lie on or near to an unknown nonlinear manifold with a density far from being uniform, which leads to bias in the MLE method. In this section, we propose an improvement of the MLE by introducing a correction for non-uniformity of density and nonlinearity of manifold into the likelihood function.

3.1. Adjusted Likelihood Construction

We start from the general Poisson process-based likelihood (1) but aim to find a better approximation to the rate $\lambda_{m,\theta}(t)$. Our derivation requires several assumptions of manifold $\mathbb{X}$ and density $f(x)$.

We assume that density $f(x)$ is bounded for $x \in \mathbb{X}$ and denote $f_{\max} = \sup\limits_{x \in \mathbb{X}} f(x)$. Let us also define the bounds on maximum eigenvalues of first and second derivatives of $f(x)$:

$$C_{p,1} = \sup_{x \in \mathbb{X}, \eta \in T_x(\mathbb{X}) \colon \|\eta\|=1} \|\nabla_\eta f(x)\|, \qquad C_{p,2} = \sup_{x \in \mathbb{X}, \eta \in T_x(\mathbb{X}) \colon \|\eta\|=1} \|\nabla_\eta \nabla_\eta f(x)\|,$$

where $T_x(\mathbb{X})$ is a tangent space to the manifold $\mathbb{X}$ at the point $x \in \mathbb{X}$. Bounded values of $C_{p,1}$ and $C_{p,1}$ lead to smooth behaviour of the density $f(x)$.

We also assume that the manifold $\mathbb{X}$ is not too curved. This limitation can be expressed in terms of the second normal form $\mathbb{I}(\eta, \eta)$ and the Ricci curvature $\mathrm{Ric}(\eta, \eta)$, those are bounded for manifolds with smooth enough parametrizations according to Lemmas 3 and 4 from (Yanovich, 2016). We assume that for a given manifold $\mathbb{X}$ there exist such positive constants $C_{\mathbb{I}}$ and C_{Ric} that for all $x \in \mathbb{X}$, $\eta \in T_x(\mathbb{X})$, and $\|\eta\| = 1$

$$\mathbb{I}(\eta, \eta) \le C_{\mathbb{I}}, \quad \mathrm{Ric}(\eta, \eta) \le C_{\mathrm{Ric}}.$$

Under the considered assumptions the following result follows.

Proposition 1 *The rate of Poisson process $N_\lambda(R, x)$ on the manifold $\mathbb{X}$ can be expressed as*

$$\lambda_{m,\theta}(R) = R^{m-1} V_m \big(m f(x) + R^2 \delta(R)\big) = \hat{\lambda}_{m,\theta}(R) + R^{m+1} V_m \cdot \delta(R),$$

where the term $\delta(R)$ can be bounded as

$$\begin{aligned} |\delta(R)| \le 8 f_{\max}(m+2) \frac{m C_{\mathbb{I}}}{24} + C_{p,2}(m+2) + (m+3) R C_{p,1} C_{\mathrm{Ric}} \\ + (m+4) R^2 C_{p,2} C_{\mathrm{Ric}} + f(x) C_{\mathrm{Ric}}(m+2). \end{aligned} \tag{3}$$

The result of Proposition 1 allows us to represent the true log-likelihood (1) in the following way:

$$\begin{aligned} L_\lambda(m, \theta) = (m-1) \int_0^R \log t \; dN_\lambda(t, x) + N_\lambda(R, x) \log V_m + N_\lambda(R, x) \log m \\ + N_\lambda(R, x) \log f(x) + \int_0^R \log\big(2t^2 \delta(t)\big) dN_\lambda(t, x) - V_m R^m \left(f(x) + \frac{R^2 \delta(R)}{m+2} \right). \end{aligned}$$

The following result allows to compute the maximizer for the function $L_\lambda(m, \theta)$.

Proposition 2 *Assume that $R^2\delta(R) < 1$ and $R^2\delta(R)/f(x) < 1$ for all $x \in \mathbb{X}$. Then the maximum of the function $L_\lambda(m, \theta)$ is achieved by*

$$\breve{m}_R(x) = \hat{m}_R(x)\left(1 + \varepsilon(R)\frac{R^2}{N(R, x)}\right), \tag{4}$$

where $|\varepsilon(R)| \le \mathrm{C}|\delta(R)|$ for some absolute constant C.

Unfortunately, the estimate $\breve{m}_R(x)$ cannot be computed directly as the quantities $\delta(R)$ and $\varepsilon(R)$ are unknown. We know the explicit upper bound (3) on $\delta(R)$, but it still includes a number of unknown parameters depending on manifold $\mathbb{X}$ and density $f(x)$.

However, the form of dependency in equation (4) suggests that $\breve{m}_R(x)$ can be found by computing the correction to the standard MLE $\hat{m}_R(x)$. We note that by Taylor expansion we can represent (4) in the following form

$$\breve{m}_R(x) = \hat{m}_R(x) + \sum_{j=1}^{l} \zeta_j R^j + O(R^{l+1}),$$

where vector $\zeta = (\zeta_1, \ldots, \zeta_l)$ represents coefficients of a polynomial of degree l.

The key idea is to consider the estimates $\hat{m}_R(x)$ for different values of R and try to fit polynomial approximation to them. Under the assumption that $\breve{m}_R(x)$ does not depend on R, the zero order term in the approximation will give an estimate $\breve{m}(x)$ of the intrinsic dimension. By fixing the number of neighbors k and estimating $\hat{m}_k(x)$ we obtain the following polynomial regression problem

$$\hat{m}_k(x) = \breve{m}(x) + \sum_{j=1}^{l} \zeta_j T_k^j(x) + \epsilon_k,$$

where ϵ_k represents an error due to ignoring higher order terms in polynomial approximation. The estimation of $\breve{m}(x)$ and other coefficients of polynomial ζ can be done based on estimates $\hat{m}_k(x)$ computed for different values of the number of neighbors k and corresponding distances $T_k(x)$.

3.2. Algorithmic implementation of GeoMLE

To estimate the intrinsic dimension $\breve{m}(x)$ of the manifold in the vicinity point x based on the sample $D = \{X_1, \ldots, X_n\}$ by polynomial regression, we should construct a dataset of MLEs $\hat{m}_{k_1}(x), \ldots, \hat{m}_{k_2}(x)$ for a range of values of $k = k_1 \le \cdots \le k_2$ with k_1 and k_2 being input parameters of the method. It is important to choose k_1 large enough to ensure the stability of distance estimates $T_k(x)$, while k_2 can not be very large to validate the approximations used to construct the estimates. In practice, due to the finite size of the data, the estimates $\hat{m}_k(x)$ are unstable for small and even moderate values of k. We suggest to estimate this uncertainty by special bootstrap procedure and incorporate obtained uncertainty estimates directly into regression problem. Such an approach also allows making the method less dependent on the choice of the number of nearest neighbors k.

We start by creating M bootstrapped datasets $\tilde{D}_1, \ldots, \tilde{D}_M$ of the sample $D = \{X_1, \ldots, X_n\}$. For each k we repeat the following procedure. First, we find k nearest neighbors of point x among the points in $\tilde{D}_j$ bootstrapped dataset for $j = 1, \ldots, M$. Then, for x we calculate its distance from its k-th nearest neighbor $T_k(x, \tilde{D}_j)$ in $\tilde{D}_j$ and find its dimension $\hat{m}_k(x, \tilde{D}_j)$ by MLE approach.

After that, we average the distances to neighbors and MLEs in the following way

$$\bar{T}_k(x) = \frac{1}{M} \sum_{j=1}^{M} T_k(x, \tilde{D}_j), \quad \bar{m}_k(x) = \frac{1}{M} \sum_{j=1}^{M} \hat{m}_k(x, \tilde{D}_j).$$

In addition, for each neighbor k we calculate the variances of MLE dimensions for x in the sample

$$\hat{\sigma}_k^2(x) = \frac{1}{M} \sum_{j=1}^{M} \bigl(\hat{m}_k(x, \tilde{D}_j) - \bar{m}_k(x)\bigr)^2.$$

Given estimates of the variances $\hat{\sigma}_k^2(x)$ of estimated dimension $\hat{m}_k(x)$, we can build a heteroscedastic polynomial regression model

$$\min_{\breve{m}(x), \zeta} \sum_{k=k_1}^{k_2} \frac{1}{\hat{\sigma}_k^2(x)} \Bigl(\bar{m}_k(x) - \breve{m}(x) - \sum_{j=1}^{l} \zeta_j \bar{T}_k^j(x)\Bigr)^2,$$

where we fit to the data the polynomial of degree l with constant term given by $\breve{m}(x)$ and other coefficients given by vector ζ. In order to find the resulting intrinsic dimension $\breve{m}$ we can run the procedure for each point in the sample $D = \{X_1, \ldots, X_n\}$ and average the obtained local estimates:

$$\breve{m} = \frac{1}{n} \sum_{i=1}^{n} \breve{m}(X_i).$$

Figure 1 illustrates GeoMLE approach by showing the resulting polynomial estimates for the samples from spheres of three different dimensions. We note that it is not possible to get MLE estimates for the values of R close to zero as the smallest possible values of R are determined by the distances to nearest neighbours of the central point. Thus, GeoMLE clearly improves over MLE and gives the estimate close to the true dimension. In the next section, we proceed with the evaluation of GeoMLE compared to other intrinsic dimension estimators.

4. Experiments

In this section, we present the performance of GeoMLE by conducting a series of experiments on synthetic and real-world datasets that are suggested as a benchmark for evaluating intrinsic dimension estimators in (Rozza et al., 2012). Simulated datasets used in our experiments are generated from different well-known manifolds such as linear subspace with normal distribution, sphere, Swiss roll, helix, cube surface, paraboloid, and some others.

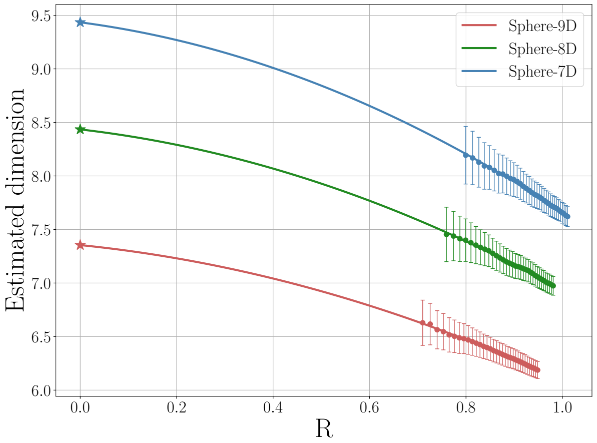

Figure 1: Illustration of GeoMLE for the samples from spheres of 3 different dimensions. Points indicate average MLEs of bootstraped datasets for different values of R with corresponding standard deviations. Curves show weighted quadratic regression fitted to the points, while stars represent the resulting estimates of the dimension.

Table 1: Parameters of each algorithms

Method	MLE	GeoMLE	MiND_{ML}	DANCo	ESS	PCA	Hein
Parameters	k1 = 20 k2 = 55	k1 = 20 k2 = 55	k = 37	k = 37	d = 1	$\alpha = 0.05$	None

Table 2: Hyperparameters of the algorithms used in the experiments.

For the experiments on synthetic data we take the size of datasets equal to 1000. Real-world datasets in our experiments include Digits (Alpaydin and Kaynak, 1998), ISOMAP face (Tenenbaum et al., 2000), and ISOLET (Fanty and Cole, 1990).

In our experiments we consider several classical baseline methods such as Local PCA (Fukunaga and Olsen, 1971), MiND_{KL} (Lombardi et al., 2011), Hein (Hein and Audibert, 2005) and MLE (Levina and Bickel, 2005), and state-of-the-art approaches DANCo (Ceruti et al., 2014) and ESS (Johnsson et al., 2015) according to the recent review (Campadelli et al., 2015). See a more detailed discussion of these methods in Section 5.

We observed that already the quadratic polynomial works well for GeoMLE and used it with 15 bootstrap samples for each local estimate in all the experiments. The other hyperparameters of all the methods were also fixed in our experiments and are summarized in the Table 2. The following link provides access to the implementation of the proposed method and all the experiments: `https://github.com/premolab/GeoMLE`.

Table 3: Estimation results on synthetic datasets averaged over 10 samples with standard deviations reported in brackets. p is the dimension of space into which the data is embedded and m is the true dimension of the data.

Dataset	p	m	MLE	GeoMLE	MiND$_{\text{ML}}$	DANCo	ESS	Hein
Affine	10	10	8.00(0.00)	**10.00**(0.00)	8.00(0.00)	9.60(0.55)	**10.00**(0.00)	8.00(0.00)
Cubic	35	30	18.00(0.00)	**29.80**(0.84)	19.20(0.45)	**29.80**(0.84)	30.40(0.55)	21.00(0.71)
Helix	3	1	**1.00**(0.00)	**1.00**(0.00)	**1.00**(0.00)	**1.00**(0.00)	3.00(0.00)	**1.00**(0.00)
Moebius	3	2	**2.00**(0.00)	**2.00**(0.00)	**2.00**(0.00)	**2.00**(0.00)	**2.00**(0.00)	**2.00**(0.00)
Nonlinear	36	6	7.00(0.00)	7.00(0.00)	7.00(0.00)	7.80(0.45)	12.00(0.00)	**6.00**(0.00)
Paraboloid	30	9	5.00(0.00)	**9.00**(0.00)	5.00(0.00)	6.00(0.00)	1.00(0.00)	2.60(0.55)
Roll	3	2	**2.00**(0.00)	**2.00**(0.00)	**2.00**(0.00)	**2.00**(0.00)	3.00(0.00)	**2.00**(0.00)
Sphere	15	10	8.80(0.45)	**10.20**(0.45)	9.00(0.00)	11.00(0.00)	11.00(0.00)	9.00(0.00)
Spiral	13	1	2.00(0.00)	1.80(0.45)	2.00(0.00)	2.00(0.00)	2.00(0.00)	**1.00**(0.00)
Uniform	55	50	25.00(0.00)	51.00(1.22)	27.00(0.00)	49.00(1.22)	**49.20**(1.30)	30.40(0.89)
MPE			0.269	**0.097**	0.248	0.164	0.385	0.188
Time			0.354	3.608	**0.050**	21.220	0.273	0.079376

4.1. Simulated and real-world data

For the evaluation we consider 45 different synthetic datasets with 10 independent samples generated for each of them. Table 3 presents the resulting estimates for selected synthetic datasets. Here p denotes the full dimension of data space and m is the true dimension of the data for synthetic datasets. The results are averaged over 10 independent samples, and the best estimates for each dataset are in bold. For more quantitative comparison of the algorithms we also calculate mean percentage error (MPE), which is MPE $= \frac{1}{n}\sum_{i=1}^{n}\frac{|m_i - \hat{m}_i|}{m_i}$, where n is the number of synthetic manifolds, m_i is the true dimension, and $\hat{m}_i$ is the estimated dimension. We also report the average computing time for all the methods. It is clearly seen that GeoMLE is the most accurate estimate in the majority of cases, while other methods give the best results only for few datasets each. Moreover, GeoMLE gives the best results in terms of MPE and is computationally significantly faster than its closest competitor in terms of MPE (DANCo).

In Figure 2 we summarize the results for synthetic datasets by plotting Dolan-More curves (Dolan and More, 2002) which are a benchmarking tool for comparison of the performance of different methods. Each curve $p_a(\tau)$ defines the fraction of problems in which the a-th algorithm has the error not more than τ times bigger than the best competitor. Thus, the higher curve, the better performance of the algorithm, and $p_a(1)$ is equal to the fraction of problems for which algorithm a gives the best result over all the algorithm. We see that GeoMLE shows the best result in more than 80% of the problems. The closest competitor to GeoMLE is DANCo, while other methods perform significantly worse.

In Table 4 we report the results for real-world datasets. True dimensions m for real-world datasets are not known, so we provide only the expert opinion on them, see (Rozza et al., 2012). We see that in 2 out of 3 cases GeoMLE clearly outperforms the competitors.

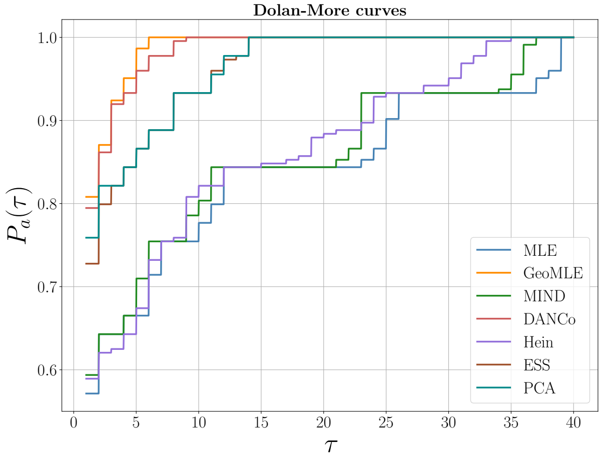

Figure 2: Dolan-More curves for all synthetic datasets to compare the estimates of MLE, GeoMLE, MiND$_{\mathrm{KL}}$, DANCo, Hein, ESS, and PCA. $p_a(\tau)$ shows the ratio of problems on which the performance of the a-th method is the best.

Table 4: Estimation results achieved on real-world datasets. p is the dimension of space into which the data is embedded and m is the estimate by expert on the true dimension of the data.

Dataset	p	m	MLE	GeoMLE	MiND$_{\mathrm{ML}}$	DANCo	ESS	Hein
Isomap	4096	3	4	**3.3**	4.0	6.0	7.4	**3.0**
Digits	64	9-11	7.7	**11.0**	8.0	**9.0**	13.2	7.0
ISOLET	617	16-22	**16.9**	25.0	15.0	14.0	12.4	14.0

4.2. Robustness to noise

We also evaluate the robustness of GeoMLE and other methods with respect to noise. We add zero mean Gaussian noise to samples for synthetic datasets. Standard deviations of noise are taken to be from 0 to 0.05 with step size equal to 0.01. The results are averaged over all synthetic datasets and 5 independent realizations of noise. We see in Figure 3 that PCA and ESS are almost not affected by noise, while GeoMLE still shows the best quality of intrinsic dimension estimation for considered levels of noise.

4.3. Effect of nonuniform sampling

Finally, we want to explicitly test whether GeoMLE allows to correct for nonuniform density, as in all the previous synthetic experiments density was always uniform. In Table 5 we compare the performance of GeoMLE and MLE on 5-dimensional spheres with uniform and nonuniform densities embedded into 7-dimensional space. Non-uniformity was achieved by generating points with uniform density in 5 dimensional space and then projecting them

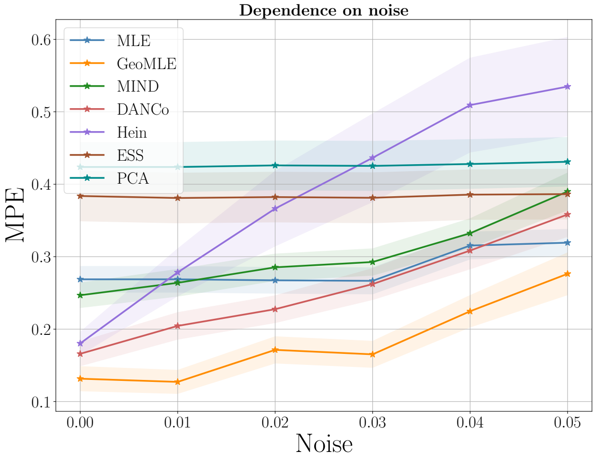

Figure 3: Dependence of estimates of MLE, GeoMLE, $\mathrm{MiND}_{\mathrm{KL}}$, DANCo, ESS, and PCA on noisy 4-dimensional sphere data.

Table 5: Dimension estimates of GeoMLE and MLE of 5-dimensional sphere in 7-dimensional space with uniform and nonuniform densities. The results are averaged over 10 samples of 1000 points each.

Method	Uniform	Nonuniform
GeoMLE	5.13 (0.08)	5.04 (0.10)
MLE	4.87 (0.05)	4.64 (0.05)

on the sphere. The presented estimates are averaged over 10 samples of 1000 points each. Despite there are no major differences between the methods for spheres with uniform densities, in case of nonuniform densities MLE underestimates the dimension while GeoMLE gives much more accurate result.

5. Related Work

This section reviews most recent and efficient intrinsic dimension estimators, which can be classified into 4 big groups: projective, fractal, nearest neighbor based, and simplex based.

Projective intrinsic dimension estimation methods are based on Multidimensional Scaling (MDS) (Romney et al., 1972) that try to maintain as much as possible pairwise distances in the data, and Principal Component Analysis (PCA) (Jolliffe, 1986), that find the best projection subspace. One of the most efficient methods in this group is local PCA (Fukunaga and Olsen, 1971).

Fractal methods rely on the assumption that data points are drawn thought a smooth probability density function from the manifold on which they lie. Some of the widely used fractal methods are Correlation dimension (Grassberger and Procaccia, 1983), the method by Camastra and Vinciarelli (2002), and the method by Hein and Audibert (2005). The later is among the best performing algorithms according to the recent review by (Campadelli et al., 2015). Another fractal dimension estimation method is proposed by Hino et al. (2017), which also considers de-biasing. Nevertheless, their results clearly show that there is still significant bias in the estimation.

The main assumption of nearest neighbor based approaches is that close points are uniformly drawn from m-dimensional balls with sufficiently small radii, where m is the true dimension of the data. Some of the most successful nearest neighbor based methods are MLE (Levina and Bickel, 2005), $\mathrm{MiND}_{\mathrm{KL}}$ (Lombardi et al., 2011), and DANCo (Ceruti et al., 2014). $\mathrm{MiND}_{\mathrm{KL}}$ (Lombardi et al., 2011) calculates the empirical probability density function of nearest neighbor distances. Then, it finds the distribution of the nearest neighbor distances of points uniformly sampled from synthetic hyperspheres of known dimension. The idea of $\mathrm{MiND}_{\mathrm{KL}}$ is to minimize the Kullback-Leibler divergence between these two distributions to obtain the dimension estimate. DANCo (Ceruti et al., 2014) is an extension of $\mathrm{MiND}_{\mathrm{KL}}$ and reduces the underestimation, which is the main downside of $\mathrm{MiND}_{\mathrm{KL}}$. In addition to the probability density function of the distribution of neighborhood distances, DANCo includes a second probability density function representing the distribution of pairwise angles.

Finally, simplex based methods evaluate simplex volumes and then analyze their geometric properties. One of the best performing methods in this category is Expected Simplex Skewness (ESS) (Johnsson et al., 2015).

5.1. Improvements of MLE

Maximum Likelihood technique has several refined approaches. MacKay and Ghahramani (2005) replace the arithmetic mean in the formulas of $\hat{m}_k$ and $\hat{m}$ by the harmonic mean. The modified estimator shows improved performance for some problems.

Another extension of the MLE approach is presented in (Gupta and Thomas, 2010), where authors apply regularized maximum likelihood to the neighborhood distances. They compute the Kullback-Leibler divergence between the rate parameters of the Poisson process in order to do regularization. The purpose is the reduction of bias in case if the number of nearest neighbors is small.

In (Karbauskaite et al., 2011) the MLE method is performed with geodesic distances instead of Euclidean distances, resulting in considerable improvement of the estimates for some datasets. Additionally, Karbauskaite and Dzemyda (2015) suggest to use the formula where MLE is calculated with radius instead of the number of neighbors k, since the bias of MLE is high for both large and small values of k. However, this approach is based on the assumption that the neighbors around each point are independent.

6. Conclusions

In this paper we have introduced a state-of-the-art intrinsic dimension estimator GeoMLE. It was inspired by one of the most widely used intrinsic dimension estimation approaches suggested by Levina and Bickel (2005). We extended the method by taking into consideration geometric properties of unknown support and possible non-uniformity of the data sampling. In the result, we propose a data-driven correction which allows to overcome the main drawbacks, which are underestimation of the true dimension in high dimensions and sensitivity to nonuniform sampling.

We compare the performance of GeoMLE to other intrinsic dimension estimators in the variety of synthetic and real-world problems. The comparison shows that GeoMLE achieves state-of-the-art performance with DANCo (Ceruti et al., 2014) being its closest competitor. Moreover, our approach is computationally faster than DANCo, while also being more robust to noise.

Acknowledgments

The authors thank Oleg Bulanov for his participation in the early stages of this research.

References

E. Alpaydin and C. Kaynak. Cascading classifiers. *Kybernetika*, 34:369–374, 1998.

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

C. M. Bishop. Bayesian pca. In *Proceedings of the 11th International Conference on Neural Information Processing Systems*, NIPS'98, pages 382–388, 1998.

Francesco Camastra and Alessandro Vinciarelli. Vinciarelli, a.: Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 05 2002.

P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering*, 2015:21 pages, 2015.

C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi, and P. Campadelli. Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration. *Pattern Recognition*, 08 2014.

E. D. Dolan and J. J. More. Benchmarking optimization software with performance profiles. *Math. Program.*, 91:201–213, 2002.

M. A. Fanty and R. Cole. Spoken letter recognition. In *NIPS*, page 220, 1990.

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

D. Granata and V. Carnevale. Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets. *Scientific Reports*, 08 2016.

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

M.D. Gupta and H. S. Thomas. Regularized maximum likelihood for intrinsic dimension estimation. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, pages 220–227. AUAI Press, 2010.

M. Hein and J.Y. Audibert. Intrinsic dimensionality estimation of submanifolds in rd. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 289–296, New York, NY, USA, 2005. ACM.

H. Hino, J. Fujiki, S. Akaho, and N. Murata. Local intrinsic dimension estimation by generalized linear modeling. *Neural Comput.*, 29(7):1838–1878, July 2017.

K. Johnsson, C. Soneson, and M. Fontes. Low bias local intrinsic dimension estimation from expected simplex skewness. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37:196–202, 09 2015.

I.T. Jolliffe. *Principal Component Analysis.* Springer Verlag, 1986.

R. Karbauskaite and G. Dzemyda. Optimization of the maximum likelihood estimator for determining the intrinsic dimensionality of high–dimensional data. *International Journal of Applied Mathematics and Computer Science*, 25(4):895–913, 2015.

R. Karbauskaite, G. Dzemyda, and E. Mazetis. Geodesic distances in the maximum likelihood estimator of intrinsic dimensionality. *Nonlinear Analysis: Modelling and Control*, 16:387–402, 2011.

E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 777–784. MIT Press, 2005.

A. Little, M. Maggioni, and L. Rosasco. Multiscale geometric methods for data sets i: Multiscale svd, noise and curvature. *Applied and Computational Harmonic Analysis*, 09 2012.

G. Lombardi, A. Rozza, C. Ceruti, E. Casiraghi, and P. Campadelli. Minimum neighbor distance estimators of intrinsic dimension. In *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, pages 374–389, 2011.

D.J. MacKay and Z. Ghahramani. Comments on "maximum likelihood estimation of intrinsic dimension" by e. levina and p. bickel. Comments on "Maximum likelihood estimation of intrinsic dimension" by E. Levina and P. Bickel, 2005.

A. K. Romney, R. N. Shepard, and S. B. Nerlove. Multidimensional scaling: Theory and applications in the behavioral sciences. *Oxford, England: Seminar Press*, I, 1972.

A. Rozza, G. Lombardi, C. Ceruti, E. Casiraghi, and P. Campadelli. Novel high intrinsic dimensionality estimators. *Machine Learning*, 89(1):37–65, 10 2012.

J.B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319, 2000.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999.

Y. Yanovich. Asymptotic properties of local sampling on manifolds. *Journal of Mathematical Statistics*, 12(3):157–175, 2016.

Y. Yanovich. Asymptotic Properties of Nonparametric Estimation on Manifold. *JMLR Workshop and Conference Proceedings*, 60:18–38, 2017.

Appendix A. Proof of Proposition 1

The manifold $\mathbb{X}$ is generally nonlinear and density $f(x)$ is non-constant. Let us estimate $\frac{\partial P\left(x \in S_x(R)\right)}{\partial R}|_{R=t}$ by considering the results obtained in (Yanovich, 2016, 2017). Firstly, we replace the domain of integration with sphere $\tilde{S}_{\tilde{X}}(R)$ in tangent space $T_X(\mathbb{X})$ and calculate the error of this replacement. From Lemma 8 (Yanovich, 2016) we obtain

$$\left|P\big(x \in S_{\tilde{X}}(R+\Delta R)\big) - P\big(x \in \tilde{S}_{\tilde{X}}(R+\Delta R)\big) - P\big(x \in S_{\tilde{X}}(R)\big) + P\big(x \in \tilde{S}_{\tilde{X}}(R)\big)\right|$$

$$\leq 8V_m f_{max}\left((R+\Delta R)^{m+2} - R^{m+2}\right)\frac{mC_{\mathbb{I}}}{24} = 8V_m f_{max}\Delta R\frac{mC_{\mathbb{I}}}{24}\sum_{i=0}^{m+1}(R+\Delta R)^i R^{m+1-i}$$

$$\leq 8V_m f_{max}(m+2)\Delta R(R+\Delta R)^{m+1}\frac{mC_{\mathbb{I}}}{24}.$$

We replace the density $f(\tilde{X})$ with the density at a point $f(x)$ and calculate the error of this replacement

$$\left|\int_{\tilde{S}_{\tilde{X}}(x)} f(R)dV(\tilde{X}) - \int_{\tilde{S}_X(R)} f(x)dV(\tilde{X})\right|$$

$$= \left\{f(x) = p(x) + t\nabla_\eta p(x) + t^2/2\nabla_{\tilde{\eta}}\nabla_{\tilde{\eta}}p(\breve{X}), \breve{X} \in \tilde{S}_X(R)\right\}$$

$$= \int_{S^{q-1}}\int_0^r (t\nabla_\eta p(x) + t^2/2\nabla_{\tilde{\eta}}\nabla_{\tilde{\eta}}p(\breve{X}))(t^{q-1} + t^{q+1}Ric_{\breve{X}}(\tilde{\eta},\tilde{\eta}))dtd\eta$$

$$\leq \int_{A^{m-1}}\int_0^R t^m\nabla_\eta p(x)dtd\eta + \int_{A^{m-1}}\int_0^R t^{m+1}/2\nabla_{\tilde{\eta}}\nabla_{\tilde{\eta}}p(\breve{X})dtd\eta$$

$$+ \int_{A^{m-1}}\int_0^R t^{m+2}\nabla_\eta p(x)Ric_{\breve{X}}(\tilde{\eta},\tilde{\eta})dtd\eta + \int_{A^{m-1}}\int_0^R t^{m+3}/2\nabla_{\tilde{\eta}}\nabla_{\tilde{\eta}}p(\breve{X})Ric_{\breve{X}}(\tilde{\eta},\tilde{\eta})dtd\eta$$

$$\leq R^{m+2}V_m(C_{p,2} + RC_{p,1}C_{\text{Ric}} + R^2C_{p,2}C_{\text{Ric}}).$$

Similarly,

$$\left|\int_{\tilde{S}_X(R+\Delta R)} f(\tilde{X})dV(\tilde{X}) - \int_{\tilde{S}_X(R+\Delta R)} f(x)dV(\tilde{X}) - \int_{\tilde{S}_X(R)} f(\tilde{X})dV(\tilde{X}) + \int_{\tilde{S}_X(R)} f(x)dV(\tilde{X})\right|$$

$$\leq V_m C_{p,2}((R+\Delta R)^{m+2} - R^{m+2}) + V_m C_{p,1} C_{\mathrm{Ric}}((R+\Delta R)^{m+3} - R^{m+3})$$

$$+ V_m C_{p,2} C_{\mathrm{Ric}}((R+\Delta R)^{m+4} - R^{m+4})$$

$$\leq V_m \Delta R \left(R+\Delta R)^{m+1}(C_{p,2}(m+2) + (m+3)(R+\Delta R)C_{p,1}C_{\mathrm{Ric}} + (m+4)(R+\Delta R)^2 C_{p,2}C_{\mathrm{Ric}}\right).$$

Now, we find the error of the replacement of density with a constant in a small neighborhood of x

$$\left|P\big(x \in \tilde{S}_X(R+\Delta R)\big) - P\big(x \in \tilde{S}_X(R)\big) - V_m m R^{m-1} f(x)\right|$$

$$= \left|\frac{\partial}{\partial R}\left(\int_{\tilde{S}_X(R)} f(x)dV(\tilde{X}) - \int_{A^{m-1}}\int_0^R t^{m-1} f(x)dtd\eta\right)\right|$$

$$= \left|\int_{A^{m-1}}\int_R^{R+\Delta R} t^{m-1} f(x)dtd\eta\right| \leq f(x)V_m C_{\mathrm{Ric}}\big((R+\Delta R)^{m+2} - R^{m+2}\big)$$

$$\leq f(x)V_m C_{\mathrm{Ric}}\Delta R(R+\Delta R)^{m+1}(m+2).$$

By substituting all the obtained errors we find the estimator $\lambda(R) = \frac{\partial P\left(x \in S_x(R)\right)}{\partial R}|_{R=t}$:

$$\lim_{\Delta R \to 0} \frac{P\big(x \in S_{\tilde{X}}(R+\Delta R)\big) - P\big(x \in S_{\tilde{X}}(R)\big)}{\Delta R}$$

$$= V_m m R^{m-1} f(x) + 8V_m f_{max}(m+2)R^{m+1}\frac{mC_{\mathrm{I\!I}}}{24} + V_m R^{m+1}\big(C_{p,2}(m+2)$$

$$+ (m+3)RC_{p,1}C_{\mathrm{Ric}}(m+4)R^2 C_{p,2}C_{\mathrm{Ric}}\big) + f(x)V_m C_{\mathrm{Ric}} R^{m+1}(m+2)$$

$$= R^{m-1}V_m\big(mf(x) + R^2\delta(R)\big),$$

where

$$|\delta(R)| \leq 8f_{max}(m+2)\frac{mC_{\mathrm{I\!I}}}{24} + C_{p,2}(m+2) + (m+3)RC_{p,1}C_{\mathrm{Ric}}$$

$$+ (m+4)R^2 C_{p,2}C_{\mathrm{Ric}} + f(x)C_{\mathrm{Ric}}(m+2).$$

A.1. Proof of Proposition 2

In order to consider geometric properties of a manifold, we plugin the obtained estimate

$$\lambda_{m,\theta}(R) = R^{m-1}V_m\big(mf(x) + R^2\delta(R)\big)$$

in the log-likelihood function. We note that below we ignore the dependence of the term $\delta(R)$ on parameters m and θ.

$$L = L_\lambda(m,\theta) = \int_0^R \log \lambda_{m,\theta}(t) dN(t) - \int_0^R \lambda_{m,\theta}(t)dt$$
$$= (m-1)\int_0^R \log t\ dN(t) + \log V_m \int_0^R dN(t) + \int_0^R \log\bigl(mf(x) + t^2\delta(t)\bigr)dN(t)$$
$$- V_m m f(x)\int_0^R t^{m-1}dt - V_m\delta(R)\int_0^R t^{m+1}dt = (m-1)\int_0^R \log t dN(t) + N(R)\log V_m$$
$$+ \int_0^R \log\bigl(mf(x) + t^2\delta(t)\bigr)dN(t) - V_m R^m \left(f(x) + \frac{R^2\delta(R)}{m+2}\right),$$

which equals for small $\frac{R^2\delta(R)}{f(x)}$ and $m \geq 1$

$$L = (m-1)\int_0^R \log t dN(t) + N(R)\log V_m + \frac{\delta(R)}{mf(x)}\int_0^R t^2 dN(t) - V_m R^m \left(f(x) + \frac{R^2\delta(R)}{m+2}\right)$$
$$= (m-1)\int_0^R \log t dN(t) + N(R)\log V_m + \frac{\Theta(\delta(R))}{f(x)}R^3 N(R) - V_m R^m \left(f(x) + \frac{R^2\delta(R)}{m+2}\right),$$

where $\Theta(\delta(R))$ means both above and below asymptotically bounded function by $\delta(R)$.

We maximize the likelihood with respect to $\theta = \log f(x)$ and m:

$$\frac{\partial L}{\partial \theta} = N(R) - V_m R^m e^\theta - R^3 N(R)\frac{\Theta(\delta(R))}{me^\theta} \Rightarrow e^\theta = \frac{N(R)}{V_m R^m}\cdot\bigl(1 - R^3\Theta(\delta(R))\bigr);$$
$$\frac{\partial L}{\partial m} = \int_0^R \log t dN(t) + \frac{V_m'}{V_m}N(R) + \frac{N(R)}{m} - V_m' R^m \frac{N(R)}{V_m R^m}$$
$$- V_m R^m \frac{N(R)}{V_m R^m}\log R - \delta(R)\frac{R^{m+2}V_m}{(m+2)^2}\left(\frac{V_m'(m+2)}{V_m} + m + 2 - 1\right) - \frac{\delta(R)}{m^2 f(x)}R^3 N(R)$$
$$= \int_0^R \log t dN(t) + \frac{N(R)}{m} - N(R)\log R - \delta(R)\frac{R^{m+2}V_m}{m+2}\left(\frac{V_m'}{V_m} + 1 - \frac{1}{m+2}\right) - \Theta(\delta(R))R^3 N(R).$$

$$\breve{m}_R(x) = \left(\frac{1}{N(R,x)}\sum_{j=1}^{N(R,x)} \log\frac{R}{T_j(x)}\right)^{-1}$$
$$\cdot\left(1 + \delta(R)\left(2\frac{R^{m+2}V_m m}{N(R,x)(m+2)}\left(\frac{V_m'}{V_m} + 1 - \frac{1}{m+2}\right) - \Theta(1)R^3\right)\left(\frac{1}{N(R,x)}\sum_{j=1}^{N(R,x)}\log\frac{R}{T_j(x)}\right)\right).$$

Finally, we obtain for small R: $\breve{m}_R(x) = \left(1 + \Theta(\delta(R))\frac{R^2}{N(R,x)}\right)\hat{m}_R(x)$.