Journal of Multivariate Analysis 173 (2019) 229–247

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Local angles and dimension estimation from data on manifolds

Mateo Díaz [a], Adolfo J. Quiroz [b,*], Mauricio Velasco [b]

[a] Center for Applied Mathematics, 657 Frank H.T. Rhodes Hall, Cornell University, Ithaca, NY 14853, USA
[b] Departamento de Matemáticas, Universidad de los Andes, Carrera 1 No. 18a 10, Edificio H, Primer Piso, 111711 Bogotá, Colombia

ARTICLE INFO

Article history:
Received 7 May 2018
Available online 26 February 2019

AMS 2010 subject classifications:
primary 62H12
secondary 62F12

Keywords:
Angle variance
Dimension estimation
Local U-statistics
Manifold learning

ABSTRACT

For data living in a manifold $M \subseteq \mathbb{R}^m$ and a point $p \in M$, we consider a statistic $U_{k,n}$ which estimates the variance of the angle between pairs $(X_i - p, X_j - p)$ of vectors, for data points X_i, X_j, near p, and we evaluate this statistic as a tool for estimation of the intrinsic dimension of M at p. Consistency of the local dimension estimator is established and the asymptotic distribution of $U_{k,n}$ is found under minimal regularity assumptions. Performance of the proposed methodology is compared against state-of-the-art methods on simulated data and real datasets.

1. Introduction

Understanding complex datasets often involves dimensionality reduction. This is particularly necessary in the analysis of images, and when dealing with genetic or text data. Such datasets are usually presented as collections of vectors in $\mathbb{R}^m$ and it often happens that there are non-linear dependencies among the components of these data vectors. In more geometric terms, these non-linear dependencies amount to saying that the vectors lie on a submanifold $M \subseteq \mathbb{R}^m$ whose dimension d is typically much smaller than m. The expression manifold learning has been coined in the literature for the process of finding properties of M from the data points.

Several authors in the artificial intelligence literature have argued about the convenience of having methods to find or approximate these low-dimensional manifolds [3,14,26,28,30,31]. Procedures for achieving this kind of low dimensional representation are called manifold projection methods. Two fairly successful such methods are Isomap by Tenenbaum et al. [31] and the Locally Linear Embedding method of Roweis and Saul [26]. For these and other manifold projection procedures, a key initial ingredient is a precise estimation of the integer d, ideally obtained at low computational cost.

The problem of estimating d has been the focus of much work in statistics starting from the pioneering work of Grassberger and Procaccia [15]. Many of the most recent dimension identification procedures appearing in the literature are either related to graph-theoretic ideas [7,8,22,23,34] or to nearest neighbor distances [12,20,21,24]. A key contribution of the latter group is the work of Levina and Bickel [20], who propose a "maximum likelihood" estimator of intrinsic dimension. To describe it, let $L_k(X_i)$ be the distance from the sample point X_i to its kth nearest neighbor in the sample with respect to the Euclidean distance in the ambient space $\mathbb{R}^m$. Levina and Bickel show that asymptotically, the expected value of the statistic

$$\widehat{m}_k(X_i) = \left\{ \frac{1}{k-2} \sum_{j=1}^{k-1} \ln L_k(X_i)/L_j(X_i) \right\}^{-1}$$

[*] Corresponding author.
E-mail address: aj.quiroz1079@uniandes.edu.co (A.J. Quiroz).

https://doi.org/10.1016/j.jmva.2019.02.014

coincides with the intrinsic dimension d of the data. As a result, they propose the corresponding sample average $\overline{m}_k = \{\widehat{m}_k(X_1) + \cdots + \widehat{m}_k(X_n)\}/n$ as an estimator of dimension. Asymptotic properties of this statistic have been obtained in the literature (see Theorem 2.1 in [23]) allowing for the construction of confidence intervals. Both the asymptotic expected value and the asymptotic distribution are independent of the underlying density from which the sample points are drawn and thus lead to a truly nonparametric estimation of dimension.

In addition to distances, Ceruti et al. [11] propose that angles should be incorporated in the dimension estimators. This proposal, named DANCo, combines the idea of norm concentration of nearest neighbors with the idea of angle concentration for pairs of points on the d-dimensional unit sphere. The resulting dimension identification procedure is relatively involved. The method combines two ideas. On one hand it uses the Kullback–Leibler divergence to measure the distance between the estimated probability density function (pdf) of the normalized nearest neighbor distance for the data considered and the corresponding pdf of the distance from the center of an r-dimensional unit ball to its nearest neighbor under uniform sampling. On the other hand, it uses a concentration result due to Södergren [29] for angles corresponding to independent pairs of points on a sphere.

The main contribution of this article is a new and simple dimension identification procedure based solely on angle concentration. We define a U-statistic which averages angle squared deviations over all pairs of vectors in a nearest neighbor ball of a fixed point and determine its asymptotic distribution. In the basic version of our proposed method, there is no need of calibration of distributions. Moreover, our statistic is a U-statistic among dependent pairs of data points and it is well known that these offer fast convergence to their mean and asymptotic distribution; see, in our specific context, Proposition 1.

Our method has been called ANOVA in the literature (see [6]), given that the U-statistic used, $U_{k,n}$ to be defined below, is an estimator of the variance of the angle between pairs of vectors among uniformly chosen points in the sphere S^{d-1}. Our main results are to prove the consistency of the proposed method of estimation (Proposition 5) and the description of the (suitably normalized) asymptotic distribution of the statistic considered (Theorem 5).

Statistics related to angles between uniformly distributed random points on unit spheres have been considered in the literature due to their potential applications on a number of fields, including physics, engineering, and biology. For a good account of the available results regarding the average angle distribution, maximum and minimum angle distribution and sphere packing distributions, see [9] and [35].

We describe our proposed method in Section 2, where some statistical properties are established related to angle-variance estimation. More theoretical justification for the proposed method, in the context of manifolds, is provided in Section 3. Sections 4 and 5 discuss the details of our implementation of the dimension identification procedure together with some empirical improvements. It also contains the result of performance evaluations on simulated examples and real datasets, including comparisons with current state-of-the-art methods.

2. A *U*-statistic for dimension identification

2.1. Description of the statistic

Suppose our data form a random sample $X_1, \ldots, X_n$ from a distribution P on $\mathbb{R}^m$ with support on a Riemannian C^2 manifold M of dimension $d < m$. Given a point $p \in M$, the question to be addressed is to determine the dimension d of the tangent space of M at p using only information from sample points near p. We want to allow for the value of d to depend on the point p and for M to be disconnected.

The simplest version of our dimension identification procedure is described by the following steps:

1. For an appropriate value of the constant C, to be specified below, let $k = \lceil C \ln(n) \rceil$. Assume, relabeling the sample if necessary, that $X_1, \ldots, X_k$ are the k nearest neighbors of p in the sample, according to the Euclidean distance in $\mathbb{R}^m$.
2. Define the angle-variance U-statistic, $U_{k,n}$, by the formula
$$U_{k,n} = \binom{k}{2}^{-1} \sum_{1 \le i < j \le k} \left(\arccos \left\langle \frac{X_i - p}{\|X_i - p\|}, \frac{X_j - p}{\|X_j - p\|} \right\rangle - \frac{\pi}{2} \right)^2, \tag{1}$$
where $\langle \cdot, \cdot \rangle$ denotes the dot product on $\mathbb{R}^m$.
3. Estimate the unknown dimension d as $\widehat{d}$, equal to the integer r such that β_r is closest to $U_{k,n}$, for a sufficiently large sample size n, where β_r is the quantity defined by
$$\beta_r = \begin{cases} \pi^2/4 - 2\displaystyle\sum_{j=0}^{s} (2j+1)^{-2} & \text{if } r - 2 = 2s + 1 \text{ is odd,} \\ \pi^2/12 - 2\displaystyle\sum_{j=1}^{s} (2j)^{-2} & \text{if } r - 2 = 2s \text{ is even.} \end{cases} \tag{2}$$

The key idea of our estimator goes as follows. For large n and the chosen value of k, the nearest neighbors of p in the dataset behave as uniform data on a small ball around p in the embedded tangent space of M at this point, and the corresponding unit vectors, $(X_i - p)/\|X_i - p\|$, are nearly uniform on the unit sphere of the tangent space, S^{d-1}. For uniform data on S^{d-1}, the expected angle between two random vectors is always $\pi/2$ (regardless of d), but the variance of this angle decreases rapidly with d. Formula (2) gives the value of this variance for every dimension r. Since our results below show that the U-statistic, $U_{k,n}$, will converge in probability to β_d for the actual dimension of M at p, estimation of d by choosing the r such that β_r closest to $U_{k,n}$ will be consistent. An additional fact that helps in this convergence is that the variance of $U_{k,n}$, which depends on the fourth moment of the angles, is also converging rapidly to zero.

The following section establishes useful facts about angles between random points on the unit sphere S^{d-1} of $\mathbb{R}^d$ and, in particular, about moments of the function

$$h(z, z') = \left(\arccos\left\langle z, z'\right\rangle - \pi/2\right)^2, \tag{3}$$

when computed on data uniformly distributed on S^{d-1}. Section 3, building on Section 2.2, develops the theoretical results that serve as basis for the use of $U_{k,n}$ on manifolds.

2.2. Angle-variance statistics for pairs of uniform points on S^{d-1}

Lemma 1 (*Angles Between Uniform Vectors*)**.** *Let Z_1, Z_2 be two independent vectors with the uniform distribution on the unit sphere $S^{d-1} \subseteq \mathbb{R}^d$ and let $\Theta_d = \arccos\langle Z_1, Z_2\rangle$ be the angle between them. The following statements hold.*

(i) The distribution of Θ_d is given, for all $\alpha \in [0, \pi]$, by

$$\Pr(\Theta_d \le \alpha) = \int_0^\alpha \sin^{d-2}(\phi)d\phi \Big/ \int_0^\pi \sin^{d-2}(\phi)d\phi.$$

(ii) The moment generating function of Θ_d, denoted by $\phi_{d-2}(s) = \mathrm{E}(e^{s\Theta_d})$ is given, for all $s \in \mathbb{R}$, by

$$\phi_{2k}(s) = \frac{e^{s\pi} - 1}{s\pi} \prod_{j=1}^{k} \frac{(2j)^2}{(2j)^2 + s^2}, \quad \phi_{2k+1}(s) = \frac{e^{s\pi} + 1}{2(s^2 + 1)} \prod_{j=1}^{k} \frac{(2j+1)^2}{(2j+1)^2 + s^2},$$

according to whether $d - 2$ is even or odd respectively.

(iii) In particular, $\mathrm{E}(\Theta_d) = \pi/2$ for all d and $\mathrm{var}(\Theta_d) = \beta_d$ where

$$\beta_d = \begin{cases} \pi^2/4 - 2\sum_{j=0}^{k}(2j+1)^{-2} & \text{if } d-2 = 2k+1 \text{ is odd,} \\ \pi^2/12 - 2\sum_{j=1}^{k}(2j)^{-2} & \text{if } d-2 = 2k \text{ is even.} \end{cases}$$

(iv) The variance of the centered squared angle $\sigma_d^2 = \mathrm{var}\,(\Theta_d - \pi/2)^2$ is given by

$$\sigma_d^2 = \begin{cases} -\pi^4/8 + 12\sum_{j=0}^{k}(2j+1)^{-4} + 2\left\{\pi^2/4 - 2\sum_{j=0}^{k}(2j+1)^{-2}\right\}^{-2} & \text{if } d-2 = 2k+1, \\ -\pi^4/120 + 12\sum_{j=1}^{k}(2j)^{-4} + 2\left\{\pi^2/12 - 2\sum_{j=1}^{k}(2j)^{-2}\right\}^{-2} & \text{if } d-2 = 2k. \end{cases}$$

Proof. (i) Passing to polar coordinates $r, \phi_1, \ldots, \phi_{d-1}$ with $r \le 0$, $\phi_j \in [0, \pi]$ for $j \in \{1, \ldots, d-2\}$ and $\phi_{d-1} \in [0, 2\pi]$. The probability that $\Theta_d \le \alpha$ is precisely the fraction of the surface area of the sphere defined by the inequalities $0 \le \phi_1 \le \alpha$. Since the surface element of the sphere is given by

$$dS = \sin^{d-2}(\phi_1)\sin^{d-3}(\phi_2)\cdots\sin(\phi_{d-2})d\phi_1\cdots d\phi_{d-1}$$

the probability is given by

$$\int_0^\alpha \int_0^\pi \cdots \int_0^\pi \int_0^{2\pi} dS \Big/ \int_0^\pi \int_0^\pi \cdots \int_0^\pi \int_0^{2\pi} dS = \int_0^\alpha \sin^{d-2}(\phi)d\phi \Big/ \int_0^\pi \sin^{d-2}\phi d\phi,$$

as claimed.

(ii) We begin with a claim.

Claim 1. *Let $u : \mathbb{R} \to \mathbb{R}$ be a C^2-function and define*

$$\mathrm{E}_d\{u(x)\} = \frac{1}{A_d}\int_0^{\alpha} u(x) n \sin^{d-2}(\phi) d\phi,$$

where $A_d = \int_0^{\pi} \sin^{d-2}\phi d\phi$. Then, the following recursion formula holds:

$$d\,\mathrm{E}_d\{u(x)\} = d\,\mathrm{E}_{d-2}\{u(x)\} - \mathrm{E}_d\{u''(x)\}/d.$$

Proof. Using integration by parts, one can show a recursive formula for A_d and conclude that

$$A_d = \begin{cases} (2k)!\pi/\{2(2^k\, d!)^2\} & \text{if } d-2 = 2k, \\ (2^k k!)^2/(2k+1)! & \text{if } d-2 = 2k+1. \end{cases}$$

Applying integration by parts twice and using our formula for A_d, we obtain the desired result. □

In particular if we take $u(x) = e^{sx}$, we get $\mathrm{E}_d(e^{sx}) = d^2 \mathrm{E}_{d-2}(e^{sx})/(d^2+s^2)$. As a result, we obtain the stated closed formula for the moment generating function.

(iii) All densities are, like sine, symmetric around $\pi/2$ and the first statement follows. To ease the computations, we introduce the cumulant-generating function $\psi_{d-2} = \ln\{\mathrm{E}(e^{s\Theta_d})\}$. Then it is immediate that $\mathrm{var}(\Theta_d) = \psi''_{d-2}(0)$. We consider two cases, d even and odd. First let us assume that $d = 2k+2$. Then, we write the cumulant-generating function as

$$\psi_{d-2}(s) = \underbrace{\ln\left(\frac{e^{s\pi}-1}{s\pi}\right)}_{t(s)} + \underbrace{\sum_{j=1}^{k} \ln\left\{\frac{(2j)^2}{(2j)^2+s^2}\right\}}_{r(s)}.$$

After some dry algebra, we get $t''(0) = \pi^2/12$ and $r''(0) = -2\sum_{j=1}^{k}(2j)^{-2}$, which gives the result for the even case. The odd case follows from an analogous argument.

(iv) Let μ_j be the jth moment of the random variable $(\Theta_d - \pi/2)$, i.e., $\mu_j = \mathrm{E}\,(\Theta_d - \pi/2)^j$. It is well known that $\mu_2 = \psi''(0)$ and $\mu_4 = \psi^{(4)}(0) + 3\{\psi''(0)\}^2$. Therefore,

$$\mathrm{var}\,(\Theta_d - \pi/2)^2 = \mu_4 - \mu_2^2 = \psi^{(4)}(0) + 2\{\psi''(0)\}^2. \tag{4}$$

Again, consider two cases: d even and d odd. Suppose $d = 2k-2$. Just as before, we calculate $t^{(4)}(0) = -\pi^4/120$ and $r^{(4)}(0) = 12\sum_{j=1}^{k}(2j)^{-4}$. Substituting both these into (4) yields the claim. A similar argument can be applied to the odd case. □

At first glance the formulas for β_d and σ_d in Lemma 1 might seem a little complicated. In order to derive our results, we need tangible decrease rates in terms of the dimension. The following claim gives us an easy way to interpret these quantities.

Claim 2. *The following bounds hold for β_d and σ_d^2:*

$$1/d \le \beta_d \le 1/(d-1) \quad \textit{for } d \ge 1,$$

$$1/(2d^2) \le \sigma_d^2 \le 2/(d-1)^2 \quad \textit{for } d \ge 4.$$

Moreover the upper bound for σ_d^2 holds for all $d \ge 1$.

Proof. We distinguish two cases according to whether $d \ge 1$ is even or odd. If d is even, we can define k by the equality $d-2 = 2k$ and compute

$$\beta_d = 2\sum_{j=k+1}^{\infty}(2j)^{-2}.$$

Since this series consists of monotonically decreasing terms and $d \ge 2$, we conclude that

$$1/d = 1/(2k+2) = 2\int_{k+1}^{\infty}(2x)^{-2}dx \le 2\sum_{j=k+1}^{\infty}(2j)^{-2} \le 2\int_{k+1/2}^{\infty}(2x)^{-2}dx = 1/(2k+1) = 1/(d-1)$$

as claimed. Furthermore, notice that the other term concerning the variance can be written as

$$12\sum_{j=1}^{k}(2j)^{-4} - \pi^4/120 = -12\sum_{j=k+1}^{\infty}(2j)^{-4},$$

which again can be bounded above by

$$2/d^3 = 1/\{4(k+1)^3\} = 12\int_{k+1}^{\infty}(2x)^{-4}dx \leq 12\sum_{j=k+1}^{\infty}(2j)^{-4} \leq 2\int_{k+1/2}^{\infty}(2x)^{-2}dx = 2/(2k+1)^3 = 2/(d-1)^3.$$

Then, we get

$$1/(2d^2) \leq 2/d^2 - 2/(d-1)^3 \leq \sigma_d^2 \leq 2/(d-1)^2 - 2/d^3 \leq 2/(d-1)^2,$$

where the first inequality follows since $d \geq 4$. The case when d is odd is proven similarly. □

In general, we know of no optimality results (e.g., in terms of minimal variance) for any of the dimension or local dimension estimators available in the literature for data on manifolds. We do not have such a general result, concerning dimension estimation, for $U_{k,n}$, but the next proposition gives an optimality result on $U_{k,n}$ as angle variance estimator for data on the unit ball.

Proposition 1. *Consider the nonparametric family $\mathcal{D}$ of all probability distributions on the unit sphere, S^{d-1}, that admit a density with respect to the uniform distribution. For each $P \in \mathcal{D}$, let $\eta = \eta(P) = \mathrm{E}(\langle Z_1, Z_2\rangle - \pi/2)^2$, where Z_1 and Z_2 are iid from P. The parameter η equals our β_d of the previous propositions when P is the uniform distribution. Then, for a random sample of size k, $Z_1, \ldots, Z_k$ from a P on S^{d-1},*

$$U_{k,n} = \binom{k}{2}^{-1}\sum_{i<j\leq k}(\langle Z_i, Z_j\rangle - \pi/2)^2 \tag{5}$$

is the UMVUE (Uniformly Minimum Variance Unbiased Estimator) for η over the family $\mathcal{D}$. Here, we are slightly abusing notation by using $U_{k,n}$ to name the statistic. To solve the notation conflict you can think of Z_i as the unit vector $(X_i - p)/\|X_i - p\|$ of Definition Eq. (1).

Proof. Given that $U_{k,n}$ is clearly unbiased, the result can be proved by the following classical three steps:

(i) The empirical distribution P_n is a sufficient statistic for η.
(ii) The empirical distribution is a complete statistic for η.
(iii) $U_{k,n}$ is the expected value of η, given P_n.

Step (i) of the proof is standard. For Step (ii), we follow the argument used in Chapter 13 of [33] (pp. 16–17) for another statistic. We need to show that $\mathrm{E}\{h(P_n)\} = 0$ for every $P \in \mathcal{D}$ implies $h \equiv 0$ almost surely. Let $\mathbf{x} = (x_1, \ldots, x_n) \in S^{d-1} \times \cdots \times S^{d-1}$ be a possible sample vector. A function of the sample vector $\delta(\mathbf{x})$ equals a function of the empirical distribution, $h(P_n)$, if and only if δ is symmetric on the coordinates, i.e., if $\delta(\mathbf{x}) = \delta(\tau\mathbf{x}) = \delta(x_{\tau(1)}, \ldots, x_{\tau(n)})$ for every permutation τ on $\{1, \ldots, n\}$. To see this, think of P_n as the set formed by the sample when order is ignored.

Let $f_1, \ldots, f_n$ be densities for distributions in $\mathcal{D}$ and let $\alpha_1, \ldots, \alpha_n$ be positive numbers. Consider the mixture density defined by

$$f(x) = \sum_i \alpha_i f_i(x) \Big/ \sum_j \alpha_j.$$

Clearly, f is a density for a distribution in $\mathcal{D}$. For this f, the condition $\mathrm{E}h(P_n) = 0$ translates to

$$0 = \int \cdots \int \delta(\mathbf{x}) f(x_1)\cdots f(x_n)\,\mathrm{d}(\mathbf{x}),$$

which implies

$$0 = \int \cdots \int \delta(\mathbf{x}) \prod_{j=1}^{n}\sum_{i=1}^{n}\alpha_i f_i(x_j)\,\mathrm{d}(\mathbf{x}).$$

The last expression can be written as a polynomial in the α_is, and this expression must be zero on the positive orthant. It follows that all coefficients of this polynomial are zero. In particular, the coefficient of $\alpha_1 \cdots \alpha_n$ is

$$0 = \sum_{\tau}\int \cdots \int \delta(\mathbf{x}) f(x_{\tau(1)})\cdots f(x_{\tau(n)})\,\mathrm{d}(\mathbf{x}),$$

where the summation is over all the permutations. By the symmetry of δ, all the terms being added in the last sum are equal and we conclude that

$$\int \cdots \int \delta(\mathbf{x}) f(x_1)\cdots f(x_n)\,\mathrm{d}(\mathbf{x}) = 0$$

Now, for $y_i \in S^{d-1}$ and $\rho_i \in (0,1)$, let $C(y_i,\rho_i) = \{y \in S^{d-1} : \langle y, y_i\rangle \geq \rho_i\}$. Each $C(y_i,\rho_i)$ is an angular cap neighborhood of y_i in S^{d-1}. For each $i \in \{1,\ldots,n\}$, let $f_i(x) = \mathbf{1}_{C(y_i,\rho_i)}$ and conclude, from the previous equality, that

$$\int_{C(y_1,\rho_1)} \cdots \int_{C(y_n,\rho_n)} \delta(x_1,\ldots,x_n)\,\mathrm{d}(\mathbf{x}) = 0$$

for every choice of the caps $C(y_i,\rho_i)$. Since the cartesian products of caps, $\prod_{i=1}^{n} C(y_i,\rho_i)$, form a basis for the topology of $S^{d-1} \times \cdots \times S^{d-1}$, it follows that δ is identically zero, except on a set of measure zero in the product space.

For Step (iii), observe that given P_n, the only randomness remaining is the order in which the dataset is presented and thus the pair (Z_1,Z_2) can take as value (with equal probabilities) any pair (Z_i,Z_j) from the original sample, with $i \neq j$. □

The following proposition gives a measure of the efficiency of $U_{k,n}$ as a dimension estimator in the ideal (limiting) case in which the normalized neighbors have a uniform distribution on the unit sphere.

Proposition 2 (*Asymptotic Sample Complexity*). *Assume that $Z_1,\ldots,Z_k$ form a random sample of vectors uniformly distributed on S^{d-1}, with d unknown. Estimate d by $\hat{d} = \hat{d}_{basic} = \arg\min_{j\in\mathbb{N}} |\beta_j - U_{k,n}|$, with $U_{k,n}$ as given in (5). That is, choose as dimension that one for which β_j is closest to the observed $U_{k,n}$. Fix a probability $\delta \in (0,1)$. Then $\Pr(\hat{d} \neq d) \leq \delta$, provided that $k \geq 6(d+1)/\sqrt{\delta} + 1$. Thus, a linear (in d) number of local examples suffices to get a good estimation of the dimension in the limiting case.*

Remark 1. The proof is omitted. A similar result is proved in Section 3 in the context of manifolds (Proposition 5).

3. Theoretical foundations

3.1. Statement of results

In this section, we state the theoretical results that serve as basis for the proposed methodology. Proofs are given in the following section. The setting is the following: A random sample, $X_1,\ldots,X_n$, is available from a distribution P on $\mathbb{R}^m$. Additionally, we have access to a distinguished point p, and near this point the data live on a Riemannian C^2 manifold M, of dimension $d < m$. Furthermore, at p the distribution P has a Lipschitz continuous non-vanishing density function g with respect to the volume measure on M. Without loss of generality, we assume that $p = 0$. Then, we have the following result.

Proposition 3 (*Behavior of Nearest Neighbors*). *For a positive constant C, define $k = \lceil C\ln(n)\rceil$ and let $R(n) = L_{k+1}(0)$ be the Euclidean distance in $\mathbb{R}^m$ from $p = 0$ to its $(k+1)$st nearest neighbor in the sample $X_1,\ldots,X_n$. Define $B_{R(n)}(0)$ to be the open ball of radius $R(n)$ around 0 in $\mathbb{R}^m$. Then, the following holds true.*

1. *For any sufficiently large $C > 0$, we have that, with probability 1, for large enough n ($n \geq n_0$, for some n_0 depending on the actual sample), $R(n) \leq r(n)$, where $r(n) = O[\{\ln(n)/n\}^{1/d}]$ is a deterministic function that only depends on the distribution P at p and C.*
2. *Conditionally on the value of $R(n)$, the k-nearest-neighbors of 0 in the sample $X_1,\ldots,X_n$, have the same distribution as an independent sample of size k from the distribution with density g_n, equal to the normalized restriction of g to $M \cap B_{R(n)}(0)$.*

In what follows, with a slight abuse of notation, we will write $X_1,\ldots,X_k$ to denote the k nearest neighbors of 0 in the sample and assume that these follow the distribution with density g_n of Proposition 3. Let $\pi : \mathbb{R}^m \to T_pM$ be the orthogonal projection onto the (embedded) tangent space to M at $p = 0$. For a nonzero $X \in \mathbb{R}^m$, let $W = \pi(X)$, $\widehat{X} = X/\|X\|$ and $\widehat{W} = W/\|W\|$. Thus, $\widehat{W}$ takes values in the $(d-1)$-dimensional unit sphere S^{d-1} of the tangent space of M at 0.

Our first lemma bounds the difference between the inner products $\langle \widehat{X_i}, \widehat{X_j}\rangle$ and $\langle \widehat{W_i}, \widehat{W_j}\rangle$ in terms of the length of projections. In this lemma, the random nature of the X_i is irrelevant.

Lemma 2 (*Basic Projection Distance Bounds*). *For any $X, X_1, X_2 \in M$, the following statements hold.*

(i) $\|X - \pi X\| = O(\|\pi X\|^2)$.
(ii) $\|\widehat{X} - \widehat{W}\| = O(\|\pi X\|)$.
(iii) The cosine of the angle between X_1 and X_2 is close to that between W_1 and W_2. More precisely,

$$|\langle \widehat{X_1}, \widehat{X_2}\rangle - \langle \widehat{W_1}, \widehat{W_2}\rangle| \leq Cr$$

for some $C \in \mathbb{R}$, whenever $r \geq \|\pi(X_i)\|$ for $i \in \{1,2\}$.

Using Lemma 2, we can establish the following approximation. Let $X_1, \ldots, X_k$ be the k-nearest-neighbors from the sample to $p = 0$ in $\mathbb{R}^m$. Define W_i and $\widehat{W_i}$ as above and let $V_{k,n}$ be given by the formula

$$V_{k,n} = \binom{k}{2}^{-1} \sum_{1\le i<j\le k} (\arccos\langle \widehat{W_i}, \widehat{W_j}\rangle - \pi/2)^2.$$

Proposition 4 (*Approximating the Statistic via its Tangent Analogue*)**.** *For $k = C\ln(n)$, as above, the following statements hold.*

(i) The sequence $k(U_{k,n} - V_{k,n})$ converges to 0 in probability as $n \to \infty$.
(ii) $\mathrm{E}\left(U_{k,n} - V_{k,n}\right) \to 0$ as $n \to \infty$.

When X comes from the distribution producing the sample, but is restricted to fall very close to 0, the distribution of πX will be nearly uniform in a ball centered at 0 in T_pM. This will allow us to establish a coupling between the normalized projection $\widehat{W}$ and a variable Z, uniformly distributed on the unit sphere of T_pM, an approximation that leads to the asymptotic distribution of $U_{k,n}$.

Some geometric notation must be introduced to describe these results. Since near 0, $M \subseteq \mathbb{R}^m$ is a Riemannian submanifold of dimension d, it inherits, from the Euclidean inner product in $\mathbb{R}^m$, a smoothly varying inner product $\ell_p : T_pM \times T_pM \to \mathbb{R}$, given by $\ell_p(u, v) = \langle i_*(u), i_*(v)\rangle$, where $i : M \to \mathbb{R}^m$ is the inclusion with differential i_*. This metric determines a differential d-form Ω_M which, in terms of local coordinates ∂_i for T_pM and dual coordinates dx_i of T_pM^* with $i \in \{1, \ldots d\}$, is given by

$$\Omega_M = \sqrt{\det(\langle \partial_i, \partial_j\rangle)_{1\le i,j\le d}}\, dx_1 \wedge \cdots \wedge dx_d.$$

The differential form endows M with a volume measure $\nu(U) = \int_U \Omega_M$. We say that a random variable A on M has density $g : M \to \mathbb{R}$ if the distribution μ_A of A satisfies $\mu_A(D) = \int_D g\Omega_M$ for all Borel sets D in M.

If X is a random variable taking values on M with density g and r is a positive real number, let $X(r)$ be a random variable with distribution g_r given by the normalized restriction of g to $M \cap B_r(0)$, i.e.,

$$g_r(z) = \begin{cases} g(z)/\displaystyle\int_{B_r(0)\cap M} g\Omega_M & \text{if } z \in M \cap B_r(0), \\ 0 & \text{otherwise.} \end{cases}$$

Define $W(r) = \pi\{X(r)\}$. The following geometric lemma will be used for relating the densities of $X(r)$ and $W(r)$.

Lemma 3 (*Tangent Space Approximations*)**.** *The following statements hold for all sufficiently small r and $p = 0$ in M.*

(i) The map $\pi : B_r(0) \cap M \to \pi\{B_r(0) \cap M\}$ is a diffeomorphism. Let $\Phi : B_r(0) \cap T_pM \to B_r(0) \cap M$ be its inverse.
(ii) The inclusion $\pi\{B_r(0) \cap M\} \subseteq B_r(0) \cap T_pM$ holds and moreover $|\lambda\{B_r(0) \cap T_pM\} - \lambda\{\pi(B_r(0) \cap M)\}| = O(r)$, where λ denotes the Lebesgue measure on T_pM.
(iii) The following equality holds:

$$\left|1 - \sqrt{\det\langle \partial\Phi/\partial x_i, \partial\Phi/\partial x_j\rangle_{1\le i,j\le d}}\right| = O(r).$$

Let $D(r)$ be a random variable uniformly distributed on $B_r(0) \cap T_pM$ and note that $Z = \widehat{D}(r) = D(r)/\|D(r)\|$ is uniformly distributed on the unit sphere S^{d-1}, regardless of the value of r. Our next lemma shows that under weak assumptions, there is a coupling between $W(r)$ and $D(r)$ which concentrates on the diagonal as r decreases.

Lemma 4 (*Coupling*)**.** *Let r denote a small positive number. With $D(r)$ as above and Z a random vector with the uniform distribution on the unit sphere, S^{d-1} of T_pM, if the density g of X in M, near 0, is locally Lipschitz continuous and non-vanishing at 0, then the following statements hold.*

(i) There exist a coupling $A(r) = (W(r), D(r))$ and a constant $C > 0$ such that $\Pr\{W(r) \ne D(r)\} \le Cr$ for all sufficiently small r.
(ii) There exists a coupling $A'(r) = (\widehat{W(r)}, Z)$ such that $\Pr\{\widehat{W(r)} \ne Z\} \le Cr$ for all sufficiently small r.

The previous lemma leads to the asymptotic distribution of the statistic $U_{k,n}$.

Theorem 5 (*Local Limit Theorem for Angle-variance*)**.** *Let $k = \lceil C\ln(n)\rceil$ for the constant C of the proof of Proposition 3 and assume $X_1, \ldots, X_k$ are the k nearest neighbors to $p = 0$ in the sample, with respect to the Euclidean distance in $\mathbb{R}^m$. If $\dim T_pM = d$ then the following statements hold.*

(i) The equality $\lim_{n\to\infty} \mathrm{E}(U_{k,n}) = \beta_d$ holds and
(ii) The quantity $k\left(U_{k,n} - \beta_d\right)$ converges, in distribution, to that of $\sum_{i=1}^{\infty} \lambda_i(\chi^2_{1,i} - 1)$ where the $\chi^2_{1,i}$ are mutually independent chi-squared random variables with one degree of freedom and the λ_is are the eigenvalues of the operator $\mathcal{A}$ on $L^2(S^{d-1})$

defined by

$$(\mathcal{A}u)(x) = \int_{S^{d-1}} \{h(x,z) - \beta_d\} u(z)\, d\mu(z)$$

for $u \in L^2(S^{d-1})$, where $h(v, v') = \{\arccos(v \cdot v') - \pi/2\}^2$ and μ denotes the uniform measure on S^{d-1}.

This limit theorem is obtained by the various approximation steps given in the preliminary results, together with the classical Central Limit Theorem for degenerate U-statistics, as described in Chapter 5 of [27]. Motivated by part (ii) of the previous theorem, we obtain formulas for the eigenvalues of the operator $\mathcal{A}$. We will use some basic facts about harmonic function theory on spheres and the reader is referred to [2] for details.

Denote by $N(m, d)$ the dimension of the space of harmonic polynomials of degree m in $\mathbb{R}^d$ and by ω_d the area of the unit sphere $S^{d-1} \subseteq \mathbb{R}^d$. It is well known that

$$N(m,d) = \binom{m+d-1}{m} - \binom{m+d-3}{m-2}$$

and that $\omega_d = 2\pi^{d/2}/\Gamma(d/2)$. Among all harmonic polynomials of degree m in S^{d-1}, there is a unique polynomial L_m^d which satisfies: (i) it is invariant under all rotations which fix the "north pole" e_d and (ii) has value one at e_d. Writing a point of the sphere in polar coordinates as $\zeta_{(d)} = te_d + (1-t^2)^{1/2}\zeta_{(d-1)}$, we define the univariate Legendre polynomial $P_m^d(t) = L_m^d(\zeta_{(d)})$. These polynomials have many useful properties; we mention two:

1. The addition theorem (see Theorem 2.9 in [2]): If $Y_1, \ldots, Y_{N(m,d)}$ are an orthonormal basis for the spherical harmonics of degree m and $z, z' \in S^{d-1}$, then
$$\sum_{j=1}^{N(m,d)} Y_j(z) Y_j(z') = N(m,d)\omega_d^{-1}\, P_m^d(\langle z, z'\rangle).$$
2. The Legendre polynomials are explicitly given by the Rodrigues' representation formula (see Theorem 2.23 in [2]):
$$P_m^d(t) = R_{m,d}(-1)^m(1-t^2)^{(3-d)/2} \frac{d^m}{dt^m}\{(1-t^2)^{m+(d-3)/2}\}$$
where $R_{m,d} = \Gamma\{(d-1)/2\}/[2^m\Gamma\{m+(d-1)/2\}]$ is Rodrigues' constant.

With these preliminaries we can now describe the spectrum of the operator $\mathcal{A}$.

Lemma 6. *For an integer $m \geq 0$ let λ_m be the real number given by*

$$\lambda_m = \omega_{d-1} \int_{-1}^{1} P_m^d(t)\{\arcsin(t)^2 - \beta_d\}(1-t^2)^{(d-3)/2} dt.$$

The eigenvalues of $\mathcal{A}$ are given by the numbers λ_m with multiplicity $N(m, d)$ for $m \in \mathbb{N}$. More precisely, $\lambda_m = 0$ if m is odd while if $m \geq 1$ is even, then

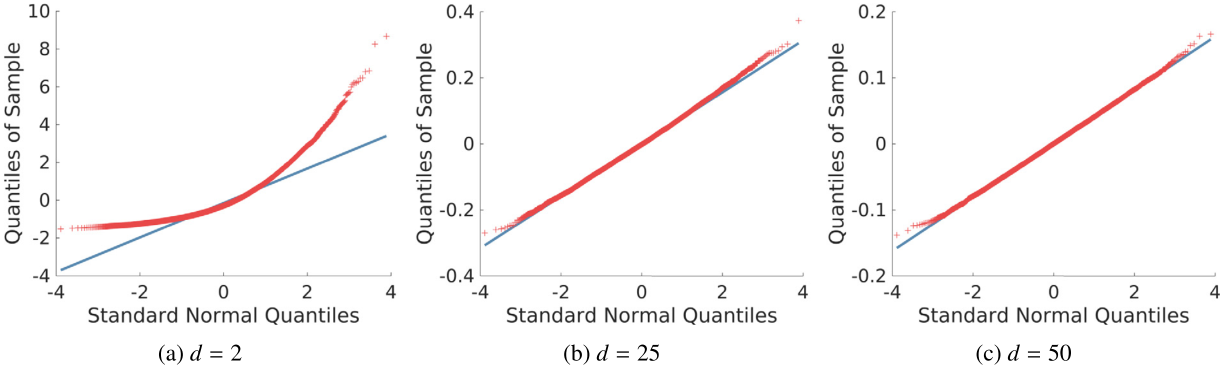
$$\lambda_m = \omega_{d-1} \sum_{k=m/2}^{\infty} \frac{2^{2k-1}}{\binom{2k}{k}k^2} R_{m,d} \frac{2k!}{(2k-m)!} B\left(k - \frac{m-1}{2}, m + \frac{d-1}{2}\right)$$

where $B(a, b)$ is the Beta function and $R_{m,d}$ is Rodrigues' constant. When $m = 0$, we have

$$\lambda_0 = \omega_{d-1}\left\{\sum_{k=1}^{\infty} \frac{2^{2k-1}}{\binom{2k}{k}k^2} B\left(k+\frac{1}{2}, \frac{d-1}{2}\right) - \beta_d \frac{(d-3)!!}{(d-2)!!} \times \kappa_d\right\},$$

where $\kappa_d = \pi$ if d is even and $\kappa = 2$ otherwise. Here, $r!!$ denotes the double or semifactorial of a number r, defined as the product of all the integers from 1 up to r that have the same parity (odd or even) as r.

Even with the formula provided by Lemma 6, it does not seem feasible at the moment to verify conditions that would guarantee that the limiting distribution in Theorem 5 is approximately Gaussian for large d. Still, we can give the following plausibility argument based on Lemma 6. The multiplicity $N(m, d)$ in Lemma 6 grows fast with m and d. For each fixed $d \geq 5$, it can be verified that $N(m, d) \geq a_d m^{d/2}$ for $m \geq 3$ and a constant a_d depending on the dimension d. To each λ_m there is an associated term, say S_m, in the sum appearing in the statement of Theorem 5, which is the sum of $N(m, d)$ mutually independent chi-squared terms with one degree of freedom, multiplied by λ_m. For large $N(m, d)$ the distribution of S_m approaches normality (at a rate that depends only on $N(m, d)$ and can be measured with the Berry–Esséen bound). This holds for every m for which $N(m, d)$ is large enough. As d grows, all the S_m will be very close to Gaussian for m larger than some small value, and the only source of non-Gaussianity in the total sum would come from the first few values of m. If these first few values do not dominate the total sum, the result should be nearly Gaussian. This reasoning leads us



(a) $d = 2$

(b) $d = 25$

(c) $d = 50$

Fig. 1. QQ-plots. As established in the proof of Theorem 5, the limiting distribution is in fact the asymptotic distribution of $k(E_n - \beta_d)$, with E_n defined in Eq. (13). For this figure, we generate 10,000 samples of the variable $k(E_n - \beta_d)$, for $k = 10\,d$ in dimensions $d \in \{5, 25, 50\}$. The plots compare the quantiles of the sample distribution against the standard normal quantiles.

to conjecture that as d increases, the limiting distribution converges to a Gaussian distribution. Numerical experiments seem to support this conjecture; see Fig. 1.

In order to get a consistency result for our basic local dimension estimator, we will use the following fact.

Corollary 1 (*Variance Convergence*)**.** *Under the conditions stated at the beginning of this subsection,*

$$\lim_{n\to\infty} \binom{k}{2} \operatorname{var} U_{k,n} = \sigma_d^2$$

for d equal to the dimension of M near 0.

Recall from Section 2 that our basic procedure estimates the dimension d as $\widehat{d}$, equal to the integer r such that β_r is closest to $U_{k,n}$. This procedure is consistent, as stated next.

Proposition 5 (*Consistency of Basic Dimension Estimator*)**.** *As in Section 2, write $\widehat{d}$ for the basic estimator described above. Let d be the true dimension of M in a neighborhood of $p = 0$. Then, in the setting of the present section, as $n \to \infty$,*

$$\Pr(\mathit{failure}) = \Pr(\widehat{d} \neq d) \to 0.$$

3.2. Proofs

Proof of Proposition 3. Let us recall a probability bound for the Binomial distribution. For N an integer-valued random variable distributed as $\mathrm{Bin}(n, p)$ and expected value $\lambda = np$, one of the Chernoff–Okamoto inequalities (see Section 1 in [17]) states that, for $t > 0$, $\Pr(N \le \lambda - t) \le \exp(-t^2/2\lambda)$. Letting $t = \lambda/2$, we get

$$\Pr(N \le \lambda/2) \le \exp(-\lambda/8). \tag{6}$$

For fixed and small enough $r > 0$, let $B_r(0)$ denote the ball of radius r around $0 \in \mathbb{R}^m$. For a random vector X, with the distribution P of our sample $X_1, \ldots, X_n$, by our assumptions on P and M near 0, we have

$$\Pr\{X \in B_r(0) \cap M\} \le \alpha \nu_d r^d,$$

where ν_d is the volume (Lebesgue measure) of the unit ball in $\mathbb{R}^d$ and α is a positive number. Let $N = N_r$ denote the number of sample points that fall in $B_r(0) \cap M$. We have that $\lambda = \mathrm{E}(N) \le \alpha \nu_d r^d n$. We choose r such that $\lambda \le C \ln n$, for a constant C to be specified in a moment. Then, by (6), we get

$$\Pr(2N \le C \ln n) \le \exp(-C \ln n/8) = (1/n)^{C/8}. \tag{7}$$

Pick any value of $C > 8$. For this choice, the bound in (7) will add to a finite value when summed over n. By the Borel–Cantelli Lemma, the inequality $N > C \ln n/2$ will hold for all n sufficiently large. It follows that if $k = C \ln n/2$, the k-nearest-neighbors of 0 in the sample, will fall in $B_r(0)$ for every n sufficiently large and the chosen value of r, namely

$$r = r(n) = \{C \ln n/(\alpha \nu_d n)\}^{1/d},$$

which is $\mathrm{O}_{\Pr}\{(\ln n/n)^{1/d}\}$. The proof of the first part of the proposition ends by renaming C.

The statement of the second part of Proposition 3 is intuitive and has been used in the literature without proof. Luckily, Kaufmann and Reiß [18] provide a formal proof of these types of results in a very general setting. In particular, (ii) of Proposition 3 holds by formula (6) of [18]. □

Proof of Lemma 2. By an orthogonal change of coordinates, we can assume that T_pM is spanned by the first d basis vectors in $\mathbb{R}^m$. The projection $\pi : M \to T_pM$ is a differentiable function whose derivative at $p = 0$ is the identity. By the Implicit Function Theorem, we can conclude that there exists an $r > 0$, such that $\pi : B_r(0) \cap M \to \pi\{B_r(0) \cap M\}$ is a diffeomorphism and that M admits, near p a chart $\Phi : B_r(0) \cap T_pM \to M$ of the form

$$\Phi(z_1, \ldots, z_d) = (z_1, \ldots, z_d, F_1(z_1, \ldots, z_d), \ldots, F_t(z_1, \ldots, z_d)), \tag{8}$$

where $m = d + t$, $\Phi(0) = p = 0$ and such that $\partial F_i(0)/\partial z_j = 0$ for all $i \in \{1, \ldots, t\}$ and $j \in \{1, \ldots, d\}$. As a result, the Euclidean distance between a point of M near 0 and the tangent space at 0 is given, in the local coordinates z, by

$$d(z_1, \ldots, z_d) = \sqrt{F_1^2(z) + \cdots + F_t^2(z)}.$$

We will prove that there exists a constant K such that, for all sufficiently small $\delta > 0$ and all z with $\|z\| \le \delta$, the inequality $d(z) \le K\|z\|^2$ holds. By Applying Taylor's Theorem at 0 to the differentiable function $d(z)$, and considering that $\Phi(0) = 0$ and $\partial F_i(0)/\partial z_j = 0$, we conclude that the constant and linear term vanish from the expansion. This proves the claim because $\|z\|^2 = \|\pi\{\Phi(z)\}\|^2$. Assume K is a constant which satisfies $\|X - \pi X\| \le K\|\pi X\|^2$. Thus,

$$\| X/\|\pi X\| - \pi X/\|\pi X\| \, \| \le K\|\pi X\|. \tag{9}$$

In addition,

$$\begin{aligned} \| X/\|X\| - X/\|\pi X\| \, \| &= \|X\| \, |1/\|X\| - 1/\|\pi X\|| \\ &= |\|X\|/-\|\pi X\|| \, /\|\pi X\| \le \|X - \pi X\|/\|\pi X\| \le K\|\pi X\|. \end{aligned} \tag{10}$$

Altogether,

$$\left\| \frac{X}{\|X\|} - \frac{\pi X}{\|\pi X\|} \right\| \le \left\| \frac{X}{\|X\|} - \frac{X}{\|\pi X\|} \right\| + \left\| \frac{X}{\|\pi X\|} - \frac{\pi X}{\|\pi X\|} \right\| \le 2K\|\pi X\|,$$

where the first inequality is just the triangle inequality and the second one follows from (9)–(10). Part (iii) of the lemma follows immediately from the triangle inequality, and part (ii) by adding and subtracting $\langle \widehat{X_i}, \widehat{W_j} \rangle$. □

Remark 2. The quadratic term of Φ is the second fundamental form of M and, therefore, the constant K can be chosen to be the largest sectional curvature of M at p.

Before proving Proposition 4, we need a lemma on the behavior of the arccos function.

Lemma 7. *Suppose that $-1 \le c_1 \le c_2 \le 1$ and let $\delta = c_2 - c_1$ be sufficiently small (for our purposes it suffices to have $\delta \le 1/4$). Then*

$$|\arccos(c_2) - \arccos(c_1)| \le 2\sqrt{|c_2 - c_1|}.$$

Proof. Assume first that both c_1 and c_2 are positive. We have

$$|\arccos(c_2) - \arccos(c_1)| = \int_{c_1}^{c_2} (1 - x^2)^{-1/2}\, dx.$$

Using that the integrand in the last expression is increasing in [0, 1] and by the change of variables $u = 1 - x$, we get

$$|\arccos(c_2) - \arccos(c_1)| \le \int_{1-\delta}^{1} (1 - x^2)^{-1/2}\, dx = \int_0^{\delta} \{u(2 - u)\}^{-1/2} du \le \int_0^{\delta} u^{-1/2} du$$

since $2 - u \ge 1$ for $u \le \delta$. From this last bound, the result follows in this case by integration. The argument for the case in which both c_1 and c_2 are negative is identical, by symmetry. In the case $c_1 \le 0 \le c_2$, both c_1 and c_2 fall in the fixed interval $[-1/4, 1/4]$ where the derivative of arccos is bounded, and the result follows easily. □

Proof of Proposition 4. To prove part (i), putting together Proposition 3 and Lemma 2, we have

$$\max_{i \le k} \|\widehat{X_i} - \widehat{W_i}\| = \mathrm{O}_{\Pr}(r(n)) = \mathrm{O}_{\Pr}\{(\ln n/n)^{1/d}\}$$

From this, it follows easily that

$$\max_{i<j \le k} |\langle \widehat{X_i}, \widehat{X_j} \rangle - \langle \widehat{W_i}, \widehat{W_j} \rangle| = \mathrm{O}_{\Pr}\{r(n)\},$$

and, using Lemma 7, we get

$$\max_{i<j \le k} |\arccos\langle \widehat{X_i}, \widehat{X_j} \rangle - \arccos\langle \widehat{W_i}, \widehat{W_j} \rangle| = \mathrm{O}_{\Pr}\left\{\sqrt{r(n)}\right\}.$$

The bound is preserved by the application of the function $u \mapsto (u - \pi/2)^2$ (since the function is locally Lipschitz) and by taking averages over all pairs, and we get

$$U_{k,n} - V_{k,n} = \mathrm{O}_{\mathrm{Pr}}\left\{\sqrt{r(n)}\right\} = \mathrm{O}_{\mathrm{Pr}}\{(\ln n/n)^{1/(2d)}\}. \tag{11}$$

The result follows by observing that, for the value of k considered, we have

$$k\,\mathrm{O}_{\mathrm{Pr}}\{(\ln n/n)^{1/2d}\} = \mathrm{o}_{\mathrm{Pr}}(1). \tag{12}$$

To prove part (ii), notice that from part (i) it is immediate that $|U_{k,n} - V_{k,n}|$ converges to zero in probability, which implies that $\mathrm{E}\left(U_{k,n} - V_{k,n}\right) \to 0$ as $n \to \infty$, since $U_{k,n} - V_{k,n}$ is a bounded random variable. □

Remark 3. In the crucial approximation step (12), it is important that our choice of k is of order $\ln n$. This step will not hold for $k = \mathrm{O}(\sqrt{n})$, for instance, which is sometimes considered in the literature as an alternative choice for the amount of neighbors to use in classification problems.

Proof of Lemma 3. Recall, from the proof of Lemma 2, that for $r > 0$, small enough, the projection $\pi : B_r(0) \cap M \to \pi\{B_r(0) \cap M\}$ is a diffeomorphism and that M admits, near p a chart (inverse) $\Phi : B_r(0) \cap T_pM \to M$ of the form given in (8) and satisfying that $\Phi(0) = p = 0$ and such that $\partial F_i(0)/\partial z_j = 0$ for all $i \in \{1, \ldots, t\}$ and $j \in \{1, \ldots, d\}$. Also from that proof, recall that there exists a constant K such that for small enough $\delta > 0$ and all z with $\|z\| \le \delta$, we have $d(z) \le K\|z\|^2$, where $d(z)$ is the distance between a point $z \in M$ and its projection on T_pM. It follows that the image $\pi(B_r(0) \cap M)$ contains a ball of radius $r' < r$ such that $r - r' = O(r^2)$ and, therefore,

$$B_r(0) \cap T_pM \supseteq \pi\{B_r(0) \cap M\} \supseteq B_{r'}(0) \cap T_pM,$$

proving part (ii) of the lemma, since the volume of the first and last terms differ by at most $O(r^2)$. For part (iii), note that

$$\frac{\partial}{\partial z_i}\Phi = e_i + \sum_{t=1}^{m-d} \frac{\partial}{\partial z_i} F_t e_{d+t}.$$

Since $\partial F_t/\partial z_i$ are $O(r^2)$ the inner products $\langle \partial\Phi/\partial z_i, \partial\Phi/\partial z_j\rangle$ are $1 + O(r^2)$ if $i = j$ and $O(r^2)$ otherwise and we conclude that $\sqrt{\langle \partial\Phi/\partial z_i \partial\Phi/\partial z_j\rangle}$ is $1 + O(r)$ as claimed. □

Proof of Lemma 4. The total variation distance between two probability measures μ and ν, defined as $\|\mu - \nu\|_{TV} = \sup_A |\mu(A) - \nu(A)|$, satisfies

$$\|\mu - \nu\|_{TV} = \inf\{\Pr(X \neq Y) : A = (X, Y)\},$$

where the infimum runs over all couplings $A = (X, Y)$ of random variables X, Y with distributions given by μ and ν, respectively; see, e.g., p. 22 in Chapter 1 of [32]. Moreover, if μ and ν are given by densities g_1, g_2, then the following inequality holds

$$\|\mu - \nu\|_{TV} \le \|g_1 - g_2\|_{L^1}.$$

We will prove part (i) of the lemma by bounding the L^1 norm of the difference of the densities of $W(r)$ and $D(r)$. Recall the definition of $W(r)$ right before the statement of Lemma 3. The difference of the two densities is given by

$$\int_{B_r(0)\cap T_pM} |h_r - 1/\lambda(B_r)|\, d\lambda,$$

where h_r denotes the density of $W(r)$ with respect to the d-dimensional Lebesgue measure λ in T_pM and $B_r = B_r(0) \cap T_pM$. More precisely, defining Φ as in Lemma 3, the density of $W(r)$ is given by

$$h_r(u) = g_r(\Phi(u))\sqrt{\det\langle \partial\Phi/\partial x_i, \partial\Phi/\partial x_j\rangle_{1\le i,j\le d}(u)}.$$

Since g is locally Lipschitz continuous, there exists a constant K such that

$$g(0) - K_1 r \le g\{\Phi(u)\} \le g(0) + K_1 r.$$

By Lemma 3 there exist constants K_2, K_3 such that the following inequalities hold for $u \in B_r$:

$$1 - K_2 r \le \sqrt{\det\langle \partial\Phi/\partial x_i, \partial\Phi/\partial x_j\rangle_{1\le i,j\le d}(u)} \le 1 + K_2 r,$$

$$\lambda(B_r) - K_3 r \le \lambda[\pi\{B_r(0) \cap M\}] \le \lambda(B_r) + K_3 r.$$

Combining these inequalities, we conclude that there exists a constant $\tilde{K}$ such that, for all $u \in B_r$,

$$\lambda(B_r)\{g(0) - \tilde{K} r\} \le \int_{B_r(0)\cap M} g\Omega \le \lambda(B_r)\{g(0) + \tilde{K} r\}.$$

As a result, the inequality

$$\frac{1}{\lambda(B_r)}\left\{\frac{g(0)-\tilde{K}r}{g(0)+\tilde{K}r}-1\right\} \le h_r - \frac{1}{\lambda(B_r)} \le \frac{1}{\lambda(B_r)}\left\{\frac{g(0)+\tilde{K}r}{g(0)-\tilde{K}r}-1\right\}$$

holds, so using the fact that $g(0) > 0$ we conclude that there exists a constant such that

$$\int_{B_r} |h_r - 1/\lambda(B_r)| \le Cr$$

as claimed, completing the proof of part (i) of the lemma.

For part (ii), let $A'(r)$ be the random pair obtained from $A(r) = (W(r), D(r))$ by normalizing its components, that is $A'(r) = (\widehat{W(r)}, \widehat{D(r)})$ and note that $\Pr\{\widehat{W(r)} \neq \widehat{D(r)}\} = O(r)$, because this probability is bounded above by the probability that $W(r)$ and $D(r)$ differ. Note that the random vector $Z = D(r)/\|D(r)\|$ is uniform on the unit sphere, and in particular its distribution is independent of the value of r. □

For $k = \lceil C\ln(n)\rceil$, as before, let $Z_1, \ldots, Z_k$ be a random sample distributed uniformly on the unit sphere S^{d-1} of T_pM and define

$$E_n = \binom{k}{2}^{-1} \sum_{1\le i<j\le k} (\arccos\langle Z_i, Z_j\rangle - \pi/2)^2. \tag{13}$$

In view of Proposition 4, in order to prove Theorem 5, it will suffice to show that the limiting standardized distribution of the $V_{k,n}$ defined in that proposition is the same as that of E_n, and establish the asymptotics for E_n.

Proof of Theorem 5. By Proposition 3 part (i), on a set of probability 1 (on the set of infinite samples of data), for some n_0, the inequality $R_n = R(n) \le r(n)$ holds for $n \ge n_0$.

By Proposition 3 part (ii), for $i \in \{1, \ldots, k\}$ the distribution of $W_i = \pi(X_i)$, the projections on T_pM of the k nearest neighbors of 0 in the sample, is that of an independent sample $W_1(R_n), \ldots, W_k(R_n)$, with $W_1(R_n), \ldots, W_k(R_n)$ as defined before Lemma 3.

From the previous lemma, conditionally on R_n, we have a coupling $A'_j(R_n) = (\widehat{W_j(R_n)}, Z_j)$ for each $j \in \{1, \ldots, k\}$. These couplings can be taken such that the $\widehat{W_j(R_n)}, j \le k$ form a random sample and the same holds for the Z_js. By the previous lemma, we have that for each j,

$$\Pr(\widehat{W_j(R_n)} \neq Z_j) \le Cr(n).$$

Then, except for a set of measure 0, we get the following event inclusion

$$\{kV_{k,n} \neq kE_n\} \subseteq \bigcup_{j=1}^{k} \{\widehat{W_i(R_n)} \neq Z_i\}.$$

By the union bound, the probability of the rightmost event is bounded by $Ck(n)r(n)$ which goes to zero, as $n \to \infty$ by the choice of $k(n)$ and the value of $r(n)$. It follows that $V_{k,n}$ and E_n have the same standardized asymptotic distributions, and being both random variables bounded, it follows that $\mathrm{E}(V_{k,n} - E_n) \to 0$ as $n \to \infty$.

For part (ii) of Theorem 5, it only remains to establish the limiting distribution of E_n. This statistic falls in the framework of classical U-statistics, that have played an important role in the theory of many nonparametric procedures. See, e.g., [25] for several applications of the theory of U-statistics. An empirical process theory is even available for U-processes; see, e.g., [1], which has found application in the study of notions of multivariate depth. Still, we will only require the classical theory, as exposed in Chapter 5 of [27] and Chapter 3 of [25].

Recall the definition of the kernel h in Eq. (3). By the symmetry of the uniform distribution for three independent vectors, Z_1, Z_2, Z_3 with uniform distribution on S^{d-1}, it can be easily verified that

$$\mathrm{E}\{h(Z_1, Z_2)h(Z_1, Z_3) - \beta_d^2\} = 0.$$

This means that the U-statistic associated to h is degenerate. It follows (see the variance calculation in [25]) that

$$\binom{k}{2}\mathrm{var}(E_n) = \mathrm{var}\{h(Z_1, Z_2)\} = \sigma_d^2 \tag{14}$$

and by the theorem for degenerate U-statistics in Section 5.5 of [27], part (ii) of our theorem follows. □

Proof of Lemma 6. Since $h(x, z) = (\arccos(\langle x, z\rangle) - \pi/2)^2$ depends only on x, z only through the inner product $\langle x, z\rangle$, the operator $\mathcal{A}$ is $O(n)$-equivariant and therefore a morphism of representations. As a result, it acts as a multiple λ_m of the identity in the irreducible subrepresentations $\mathcal{H}_m(S^{d-1}) \subseteq L^2(S^{d-1})$ which consist of homogeneous harmonic polynomials of degree m restricted to the sphere. Consequently, the spectrum of $\mathcal{A}$ is given by the numbers λ_m with multiplicity equal to $\dim\{\mathcal{H}_m(S^{d-1})\} = N(m, d)$.

From the previous paragraph, we also know that if $Y_m \in \mathcal{H}_m(S^{d-1})$ is any spherical harmonic of degree m then $\mathcal{A}(Y_m) = \lambda_m Y_m$ for some constant λ_m. Since $\arccos(t) - \pi/2 = -\arcsin(t)$, we can conclude, by applying the Funk–Hecke formula (see Theorem 2.22 in [2]), that the number λ_m is given by the claimed expression. If m is odd, then the Legendre polynomial is odd and, therefore, the integral vanishes so $\lambda_m = 0$. For $m \geq 1$, m even, we wish to compute the integral

$$\lambda_m = \omega_{d-1} \int_{-1}^{1} \arcsin(t)^2 P_m^d(t)(1-t^2)^{(d-3)/2} dt.$$

By [4], for $|t| \leq 1$, the function $\arcsin^2(t)$ has the series expansion

$$\arcsin^2(t) = \sum_{k=1}^{\infty} \frac{2^{2k-1}}{\binom{2k}{k} k^2} t^{2k}$$

and thus

$$\lambda_m = \omega_{d-1} \sum_{k=1}^{\infty} \frac{2^{2k-1}}{\binom{2k}{k} k^2} \int_{-1}^{1} t^{2k} P_m^d(t)(1-t^2)^{(d-3)/2} dt.$$

By Proposition 2.26 in [2] for every m-times continuously differentiable function f, the following identity holds:

$$\int_{-1}^{1} f(t) P_m^d(t)(1-t^2)^{(d-3)/2} dt = R_{m,d} \int_{-1}^{1} \left\{ \frac{d^m}{dt^m} f(t) \right\} (1-t^2)^{m+(d-3)/2} dt.$$

Applying this identity to $f(t) = t^{2k}$, we obtain

$$\lambda_m = \omega_{d-1} \sum_{k=m/2}^{\infty} \frac{2^{2k-1}}{\binom{2k}{k} k^2} R_{m,d} \int_{-1}^{1} \frac{(2k)!}{(2k-m)!} t^{2k-m} (1-t^2)^{m+(d-3)/2} dt.$$

Using the fact that m is even and making the substitution $u = t^2$, we can express the last integral using the Beta function, proving the claim.

Finally, let us establish the result for $m = 0$. It is immediate that $P_0^d(t) = 1$. Then

$$\begin{aligned} \lambda_0 &= \omega_{d-1} \int_{-1}^{1} \left(\arcsin(t)^2 - \beta_d \right) (1-t^2)^{(d-3)/2} dt \\ &= \omega_{d-1} \left(\sum_{k=1}^{\infty} \frac{2^{2k-1}}{\binom{2k}{k} k^2} \int_{-1}^{1} t^{2k} (1-t^2)^{(d-3)/2} dt - \int_{-1}^{1} \beta_d (1-t^2)^{(d-3)/2} dt \right) \\ &= \omega_{d-1} \sum_{k=1}^{\infty} \frac{2^{2k-1}}{\binom{2k}{k} k^2} B\left(k + \frac{1}{2}, \frac{d-1}{2} \right) - \omega_{d-1} \beta_d \frac{(d-3)!!}{(d-2)!!} \kappa_d, \end{aligned}$$

where $\kappa_d = \pi$ if d is even and $\kappa = 2$ otherwise.

Where the last identity is proven by applying integration by parts to $I_q = \int_{-1}^{1} (1-t^2)^q dt$, deriving the recursion $I_q = 2q I_{q-1}/(2q+1)$, and noting that $I_{-1/2} = \pi$ and $I_0 = 2$. □

Proof of Corollary 1. By (14), it suffices to show that $k^2(\mathrm{E}U_{k,n}^2 - \mathrm{E}E_n^2) \to 0$ as $n \to \infty$, for E_n as in the proof of Theorem 5. Applying Eq. (11) and the facts that the function h is bounded and that $k = \mathrm{O}(\ln n)$ on a set of probability 1, we have

$$k^2 \mathrm{E}(U_{k,n}^2) \leq k^2 \mathrm{E}(V_{k,n}^2) + \mathrm{O}\{\ln^2 n (\ln n/n)^{1/2d}\}.$$

The opposite inequality (interchanging the roles of $U_{k,n}$ and $V_{k,n}$) is obtained by the same reasoning, and we get

$$\lim_{n\to\infty} k^2(\mathrm{E}U_{k,n}^2 - \mathrm{E}V_{k,n}^2) = 0. \tag{15}$$

By the coupling argument of the proof of Theorem 5, we have that $V_{k,n}$ and E_n might differ at most on a set of measure $\mathrm{O}\{\ln(n) r(n)\}$. Taking expectations on the sets where they coincide and differ, we obtain

$$k^2 \, \mathrm{E}V_{k,n}^2 \leq k^2 \, \mathrm{E}E_n^2 + k^2 \, \mathrm{O}\{\ln(n) r(n)\}.$$

Observing that the second term on the right-hand side of this inequality goes to zero, as n grows, and that the reverse inequality is obtained similarly, we conclude

$$\lim_{n\to\infty} k^2(\mathrm{E}V_{k,n}^2 - \mathrm{E}E_n^2) = 0, \tag{16}$$

and the result follows by combining (15) and (16). □

Proof of Proposition 5. Let d be the true value of the dimension. β_{d+1} is the expected value closest to β_d and, by the proof of Claim 2, $(\beta_d - \beta_{d+1})/2 \geq 1/(d-1)^2$. For the basic procedure to incur an error, it is necessary that $|U_{k,n} - \beta_d| \geq |U_{k,n} - \beta_{d+1}|$, which means

$$|\mathrm{E}U_{k,n} - \beta_d| + |U_{k,n} - \mathrm{E}U_{k,n}| \geq (\beta_d - \beta_{d+1})/2 \geq 1/(d-1)^2.$$

Since $|\mathrm{E}U_{k,n} - \beta_d|$ converges to zero, the condition above requires that $2|U_{k,n} - \mathrm{E}U_{k,n}| \geq 1/(d-1)^2$, for n large enough. The probability of this last event is bounded above as follows. Let c be a supremum norm bound for $h(\widehat{X}_1, \widehat{X}_2) - \mathrm{E}h(\widehat{X}_1, \widehat{X}_2)$, for the kernel h given in (3) and $\widehat{X}_i$ as in Lemma 2. Clearly, c is bounded above by $\pi^2/4$. By Bernstein's inequality for U-statistics as reported, e.g., in Proposition 2.3(a) of [1], we get

$$\Pr\left\{|U_{k,n} - \mathrm{E}U_{k,n}| \geq \frac{1}{2(d-1)^2}\right\} \leq 2\exp\left[\frac{-k/\{8(d-1)^4\}}{2\mathrm{var}\{h(\widehat{X}_1, \widehat{X}_2)\} + 2c/\{6(d-1)^2\}}\right].$$

Now, using Claim 2 and Corollary 1, we find, after some calculations,

$$\Pr\left\{|U_{k,n} - \mathrm{E}U_{k,n}| \geq \frac{1}{2(d-1)^2}\right\} \leq 2\exp\left[\frac{-k/\{8(d-1)^4\}}{5/(d-1)^2 + \pi^2/\{12(d-1)^2\}}\right] \leq 2\exp\left\{\frac{-k}{47(d-1)^2}\right\}. \tag{17}$$

This bound goes to zero as n (and k) grow to infinity. Thus the proof is complete. □

Remark 4. Forcing the estimation error bound in (17) to be less that a given $\delta > 0$, will give a value of k of the order of $Ad^2 \ln(2/\delta)$, for some constant A, reflecting that precise estimation is more demanding, in terms of sample size, as the dimension d grows. The result obtained from the proof of Corollary 1 is more pessimistic than the one expressed by Lemma 4, which suggests that a value of k linear on d can be used to identify the dimension d with high probability. The reason for this is that, by considering the limiting case, Lemma 4 can take advantage of the degenerate nature of the limiting U-statistic, something that cannot be assumed in the proof of Corollary 1. As a guideline for the choice of C in the formula $k = C \ln n$, due to the fast convergence of U-statistics to their limiting distributions, we are inclined to trust the bound of Lemma 4 applied to a known (or guessed) upper bound $d_{\max}$ for the dimension being estimated. Using a value of δ of 0.1, for instance, this procedure gives us a value of k of around $20d_{\max}$, from where the constant C can be solved for.

4. Estimators

In this section, we present two dimension estimators based on the statistic $U_{k,n}$. First, a local estimator that gives the dimension of M around a distinguished non-singular point $p \in M$ is discussed. Then, the case in which the manifold is equidimensional is considered, by building upon our local estimator to propose a global dimension estimator. Some implementation issues are discussed and in the following section, we evaluate the performance of our estimators. The code used in these experiments is publicly available; it can be found at https://github.com/mateodd25/ANOVA_dimension_estimator.

Algorithm 1: Local dimension estimation

```
Data: $k \in \mathbb{N}_+$ and $X_1, \ldots, X_n, p \in M \subseteq \mathbb{R}^m$
Result: Estimated dimension $\widehat{d}$ at $p \in M$
Find the $k$-nearest neighbors to $p$. Use these neighbors to compute $U_{k,n}$ as in Eq. (1). Choose $\widehat{d}$ associated with $U_{k,n}$.
```

4.1. Local estimators

The theory presented in Section 3 suggests that $k \sim \ln(n)$ should be a good choice, asymptotically speaking, for the number of neighbors to consider in the local dimension estimation procedure. However, one could potentially leverage prior knowledge of the structure of the problem to set this parameter differently. In our implementation, we set it to $k = \mathtt{round}\{10\log_{10}(n)\}$.

In our theoretical analysis presented above, it was assumed that we are given a center point p, where the local dimension is to be estimated. A natural question that arises in practice is the following: given a sample, how to select good center points. In Section 4.2 we present a simple heuristic to select "good" centers.

Both estimators presented in what follows are based on the Algorithm 1. The difference between the estimators considered lies on the last line of the algorithm, namely, on how to pick the dimension estimator, given $U_{k,n}$. Next, the two different rules to execute this step are discussed.

```
Algorithm 2: Global dimension estimation
Data: $c \in \mathbb{N}_+$, $k \in \mathbb{N}_+$ and $X_1, \ldots, X_n \in M \subseteq \mathbb{R}^m$
Result: Estimated dimension $\widehat{d}$ of $M$
Choose $c$ centers $p_1, \ldots, p_c$ from the sample. Apply Algorithm 1 to each center $p_i$, let $\widehat{d}_i$ be its output. Set $\widehat{d}$ to the median of $\widehat{d}_1, \ldots, \widehat{d}_c$.
```

4.1.1. Basic estimator

Since $U_{k,n}$ converges in probability to β_d, a natural way to estimate the dimension from $U_{k,n}$ is to set

$$\widehat{d}_{\text{basic}} = \underset{d \in [D_{\max}]}{\arg\min}\, |\beta_d - U_{k,n}|,$$

where $D_{\max}$ is the ambient dimension or some bound we know a priori on the dimension of the manifold. Interestingly, such a rule is fairly accurate, as established in Proposition 5. Another advantage of this estimator is that there is no need to train it, since all the quantities involved have been analytically computed and presented in Section 2.

Remark 5. Classical discriminant analysis results (see, e.g., Section 4.1 in [13]) would advise to incorporate available variance information on the selection of $\widehat{d}$, by choosing

$$\widehat{d}_{\text{disc}} = \arg\max\{d : U_{k,n} \geq \eta_d\},$$

where $\eta_d = \beta_d + \sigma_d(\beta_{d-1} - \beta_d)/(\sigma_d + \sigma_{d+1})$. Still, simulation evaluations (not included) show that $\widehat{d}_{\text{disc}}$ and $\widehat{d}_{\text{basic}}$ have a very similar performance in practice, as well as in theory, since both are consistent. Thus, we prefer to use the latter, being the simpler one.

4.1.2. Kernel-based estimator

For our second estimator, we start by simulating multiple instances $Y_1^{(d)}, \ldots, Y_M^{(d)}$ of the random variable $k(E_n - \beta_d)$, for a large value of M (= 5000, for instance) and with E_n as defined in Eq. (13), for each dimension $d \in [D_{\max}] = \{1, \ldots, D_{\max}\}$. From these data, the density $\hat{f}_k^{(d)}$, of $k(E_n - \beta_d)$, is estimated, for each d, as

$$\hat{f}_k^{(d)}(y) = \frac{1}{Mh} \sum_{i=1}^{M} \varphi\{(y - Y_i^{(d)})/h\},$$

where φ is the standard Gaussian density and the parameter h (the bandwidth) can be set at $h = (4/3M)^{1/5}$. This choice of bandwidth guarantees consistent density estimation; see Section 4.1 in [5]. Then, a Bayesian classification procedure with uniform prior distribution on the set $[D_{\max}]$, would select the dimension as that for which the kernel density estimator is maximized at the standardized $U_{k,n}$, namely

$$\widehat{d}_{\text{ker}} = \underset{d \in [D_{\max}]}{\arg\max}\, \hat{f}_k^{(d)}\{k(U_{k,n} - \beta_d)\}.$$

Notice that the simulations described above need to be performed only once for each dimension, since they are made on uniform data on S^{d-1} and do not depend on the particular data being studied.

4.2. Global estimators

We now turn our attention to extending the local dimension estimation algorithm to a global one. Assuming that M is equidimensional, the local method can be extended by running multiple instances of Algorithm 1 on different centers, and combining the results, as outlined in Algorithm 2.

After getting dimension estimates at each center, one could use different summary statistics to choose the global dimension, such as the mean, the mode or the median. To make a method robust against outliers, we chose to use the median. To decide about the parameter c we ran a cross-validation algorithm. Empirically, it appears that $c \sim \ln(n)$ is a good choice for this parameter.

4.2.1. Choosing centers

To pick the centers p_i in Algorithm 2, we divide the sample into c disjoint subsamples of approximately equal size. Assume, for simplicity of the exposition, that $c = 1$. Inside each subsample we pick a center by assigning each point a score of centrality and then choosing the one with highest score. Scores are assigned through the following procedure:

1. For each coordinate i, we order the sample based on the ith entry, and let τ_i be the permutation giving this ordering, i.e., the ith row in $(X_{\tau_i(1)}, \ldots, X_{\tau_i(n)})$ is nondecreasing.
2. Then, the centrality score of X_j is given by $f\{\tau_1(j)\} + \cdots + f\{\tau_m(j)\}$, where $f(x) = |1/2 - 2(x - 1)/(2n)|$.

For the ith coordinate, the weight function f gives the maximum scores to the point (or points) such that $\tau_i(j)$ is closest to $n/2$ and thus, the mechanism used chooses as center a point which for many components, appears near the center of these orderings.

4.2.2. Heuristic to discard centers

Finally, we present a simple heuristic for discarding some of the selected centers, based on the mean of the angles between its neighbors, taking as always, the point considered as origin. This is done in order to improve the performance of the statistic. Consider again the angle

$$\theta_{i,j} = \arccos\left\langle \frac{X_i - p}{\|X_i - p\|}, \frac{X_j - p}{\|X_j - p\|} \right\rangle$$

for each pair of nearest neighbors X_i, X_j of the point $p \in M$, as used in the basic definition (1). Consider the average of these angles, viz.

$$\overline{\theta}(p) = \binom{k}{2}^{-1} \sum_{1 \le i < j \le k} \theta_{i,j}.$$

If the manifold M, near p, is approximately flat, $\overline{\theta}(p)$ should be close to $\pi/2$, regardless of the value of the dimension d, since $\pi/2$ is the expected value of the angle between uniformly sampled points in every dimension and the U-statistic $\overline{\theta}(p)$ should converge rapidly to this expectation. Thus, when $\overline{\theta}(p)$ is far from $\pi/2$, it can be taken as a suggestion of strong curvature that is causing non-uniformity of the angles and, therefore, p might not be a good point to consider for dimension estimation. For these reasons, in our implementation, the user is allowed to use this heuristic and discard a fraction of the centers p_i with largest values of $|\overline{\theta}(p_i) - \pi/2|$. Experiments presented in the next section suggest that this heuristic is useful when the manifold is highly curved.

The procedures just described for inclusion/exclusion of centers aim at choosing points which are central to their regions in the manifold and where the manifold has a "flat behavior" so that the theory established for the proposed procedure has a better chance of holding with good approximation. Still, since this selection procedure is data dependent, a user could be concerned about the effect of the center selection on the statistic distribution. If that is the case, the center choice and the statistic calculation can be "decoupled", by performing the center choice procedure on a subsample (of 10% of the available sample, for instance) and then, computing the statistic on the chosen centers, using their neighbors from the remaining 90% of the data. In this way, the two stages of the process can be made stochastically independent.

5. Numerical results

We compare our methods against two powerful dimension estimators, DANCo [11] and Levina–Bickel [20], using a manifold library proposed in [16] and two real datasets. The first estimator is, arguably, the state-of-the-art for this problem, while the second one is a classical well-known estimator with great performance. To see a comparison between these and other estimators, we refer the reader to [10,11]. All the experiments were run in a 2013 MacBook Pro with 2.4 GHz Intel Core i5 Processor and 8 GB of RAM.

Manifold library

Table 1 presents a brief description of the manifolds included in the study. Additionally, Table 2 contains a list of the parameters used for each one of the estimators. We compare two error measures, namely the Mean Square Error (MSE) and the Mean Percentage Error (MPE), which are respectively defined as

$$\mathrm{MSE}(\widehat{d}) = \frac{1}{T}\sum_{i=1}^{T}(\widehat{d}_i - d_i)^2 \quad \text{and} \quad \mathrm{MPE}(\widehat{d}) = \frac{100}{T}\sum_{i=1}^{T}|\widehat{d}_i - d_i|/d_i,$$

where T is the number of trials included in the test and $\widehat{d}_i$ and d_i are the estimated dimension and the correct dimension of the ith trial, respectively.

For each one of the aforementioned manifolds, we draw $T = 50$ random samples with $n = 2500$ data points and then compute the MSE and MPE of the following six estimators: the global basic estimator (Basic); the global basic estimator combined with the centers heuristic (B+H); the global kernel-based estimator (Kernel); the global kernel-based estimator with centers heuristic (K+H); the Levina–Bickel estimator (LB); and the Fast DANCo estimator (DANCo), which is a faster version of the DANCo estimator described in Section 4.2 of [11]. Tables 3 and 4 summarize the results.

In both these tables, the last column shows the average of the performance measure over the examples. It is clear from these tables that the angle-variance methods introduced in the present article do well, in terms of average performance, against the very strong competitors considered. This is more evident when the measure of error is the MSE. Still, for many of the manifolds considered, namely M_1, M_3, M_4 (in this case tied with LB), M_9 and M_{12}, DANCo clearly displays the best performance. The Levina–Bickel estimator is the best for manifold M_6, tying for first place with DANCo in M_4, while

Table 1
Library of manifolds used for benchmark, for more details consult [16].

Manifold	d	m	Description
M_1	9	10	Sphere S^9
M_2	3	5	Affine subspace
M_3	4	6	Nonlinear manifold
M_4	4	8	Nonlinear manifold
M_5	2	3	Helix
M_6	6	36	Nonlinear manifold
M_7	2	3	Swiss roll
M_8	12	72	Highly curved manifold
M_9	20	20	Full-dimensional cube
M_{10}	9	10	9-dimensional cube
M_{11}	2	3	Ten-times twisted Mobius band
M_{12}	10	10	Multivariate Gaussian
M_{13}	1	10	Curve

Table 2
Parameters of each algorithm.

Method	Parameters
ANOVA	$k = 34, c = 16$
LB	$k_1 = 10, k_2 = 20$
DANCo	$k = 10$

Table 3
Rounded Mean Square Error for different manifolds, last column displays the average MSE over all the examples. The darker cells show the best results in each column.

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}	M_{13}	Mean
Basic	0.95	0.00	0.58	0.03	0.00	0.67	0.00	1.72	10.39	0.00	0.00	0.12	0.00	1.11
B+H	1.09	0.00	0.66	0.28	0.00	1.30	0.01	2.14	4.64	0.03	0.00	0.10	0.00	0.79
Kernel	0.95	0.00	0.68	0.08	0.00	0.69	0.29	1.27	10.20	0.00	0.00	0.15	0.00	1.10
K+H	0.99	0.00	0.76	0.45	0.00	1.31	0.31	2.48	4.06	0.01	0.00	0.04	0.00	0.80
LB	0.49	0.02	0.05	0.00	0.00	0.10	0.00	2.27	29.33	2.23	0.00	0.54	0.00	2.69
DANCo	0.16	0.00	0.00	0.00	0.00	1.00	0.00	25.22	0.96	0.10	0.00	0.00	0.00	2.11

the procedures proposed in this article show the best performance in the cases of manifolds M_8 and M_{10}, having very good performance also in cases M_3, M_5, M_7 and M_{12}. It is interesting that our methods do particularly well in case M_8, a high curvature manifold of a relatively high dimension in a large dimension ambient space. It does not appear to exist a significant difference in performance between the Basic procedure and the procedure that uses Kernel Density Estimation. In contrast, the introduction of the heuristics discussed in Section 4 turns out to be beneficial for our estimators in some of the relatively high-dimensional cases, namely M_9 and M_{12}, while these heuristics degrade somehow the performance in the intermediate dimension cases, M_4 and M_6. In all other cases, the use of the heuristics for center selection and center elimination does not appear to have a strong effect.

The value of T (number of repetitions) chosen for the experiments just described is enough to get a clear idea of the methods performance. A significant increase of T would change very little the results in Tables 3 and 4, but would mean a large increment in computational cost for the experiments. By contrast, if the sample size n is increased, the performance of all the methods considered will improve. In particular, $U_{k,n}$ will improve since the neighbors of each center will be located in smaller regions, making the behavior of the local statistic closer to the limiting case. Still, in our limited experiments with larger sample sizes (not reported) the relative performance of the methods considered did not change.

Real datasets

In order to test the computational complexity of our methods, we ran experiments on two real datasets. Namely, the ISOMAP face dataset [31] and MNIST dataset [19]. The former consists of 698 gray-scale images of size 64 × 64, thus $m = 4096$. Each image displays three-dimensional model of a face with a different rotation and lighting parameters, see Fig. 2. The later dataset consists of gray-scale images displaying hand-written digits of size 28 × 28, i.e., $m = 784$. For our experiment we only use the images labeled as ones; see Fig. 3. The exact dimensions of these datasets are unknown, but they are believed to be low-dimensional [10,16].

For these datasets, we compare the global basic estimator, the same estimator combined with the center heuristic, the Levina–Bickel estimator, the DANCo estimator and the Fast DANCo. We used the same parameters displayed in Table 2.

Table 4
Rounded Mean Percentage Error for different manifolds, last column displays the average MPE over all the examples.

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}	M_{13}	Mean
Basic	10.56	0.00	15.50	1.00	0.00	10.83	0.00	9.00	15.50	0.00	0.00	1.50	0.00	4.93
B+H	11.33	0.00	17.75	8.25	0.00	17.83	0.50	9.83	10.00	0.67	0.00	1.20	0.00	5.95
Kernel	10.56	0.00	17.75	2.50	0.00	11.50	16.50	7.75	15.55	0.00	0.00	1.80	0.00	6.45
K+H	11.00	0.00	20.00	12.25	0.00	18.00	17.50	10.75	9.60	0.11	0.00	0.60	0.00	7.67
LB	7.77	4.84	5.77	1.49	1.84	5.13	2.63	12.55	27.07	16.57	1.64	7.34	0.50	7.31
DANCo	1.77	0.00	0.00	0.00	0.00	16.66	0.00	41.83	4.80	1.11	0.00	0.00	0.00	5.09

Fig. 2. Images from the face dataset.

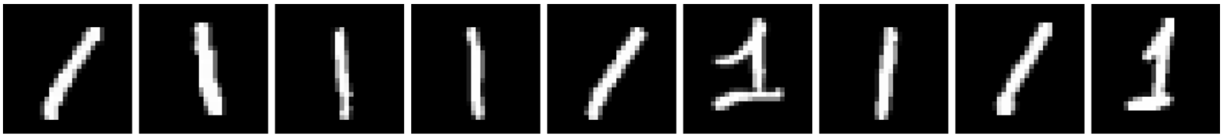


Fig. 3. Images from the MNIST dataset.

Table 5
Dimension estimates and rounded times (in seconds) of each estimator when applied to real datasets.

	Basic		B+H		LB		DANCo		Fast DANCo	
	d	Time	*d*	Time	*d*	Time	*d*	Time	*d*	Time
MNIST1	6	0.85	8.5	0.85	11.17	227.64	13	10 111.85	14	23.89
ISOMAP faces	5	0.89	5	0.89	3.86	13.62	4	8 017.10	4	3.39
Total time (s)		1.73		1.73		241.26		18 127.96		27.29

Table 5 shows their estimates and running times (in seconds). We report the same time for both the basic estimator and its enhanced version with the heuristic. This is because we compute the two statistics through a single procedure. Notice that the only difference between them is that we discard some data when using the heuristic.

Considering both the theoretical developments of the previous sections and the performance evaluation of the present one, we would like to point out the advantages that our procedure offers over the two excellent competitors included in our comparison. With respect to DANCo, our method uses a much simpler statistic, both conceptually and computationally, and our method enjoys available limiting theory, which is valid in the manifold context. In comparison with the statistic of Levina and Bickel, for which there exists available asymptotic theory developed in [23], our method offers a better performance in most of the examples considered.

Acknowledgments

We wish to thank Paul Breiding and Bernd Sturmfels for useful conversations during the completion of this work. We also wish to thank the Editor-in-Chief, Christian Genest, and two anonymous referees for their careful reading of the manuscript. M. Velasco was partially supported by research funds from Universidad de los Andes. M. Díaz wishes to thank Damek Davis for research funding during the completion of this work. A.J. Quiroz was partially supported by the STAI program from Universidad de los Andes.

References

[1] M.A. Arcones, E. Gine, Limit theorems for *U*-processes, Ann. Probab. 21 (1993) 1494–1542.
[2] K. Atkinson, W. Han, Spherical Harmonics and Approximations on the Unit Sphere: An Introduction, Springer, Heidelberg, 2012.
[3] M. Belkin, P. Niyogi, Semi-supervised learning on Riemannian manifolds, Mach. Learn. 56 (2004) 209–239.
[4] J.M. Borwein, M. Chamberland, Integer powers of arcsin, Int. J. Math. Math. Sci. (2007) 19381, 10.
[5] A.W. Bowman, P.J. Foster, Adaptive smoothing and density-based tests of multivariate normality, J. Amer. Statist. Assoc. 88 (1993) 529–537.
[6] P. Breiding, S. Kališnik, B. Sturmfels, M. Weinstein, Learning algebraic varieties from samples, Rev. Mat. Complut. 31 (2018) 545–593.

[7] M.R. Brito, A.J. Quiroz, J.E. Yukich, Graph-theoretic procedures for dimension identification, J. Multivariate Anal. 81 (2002) 67–84.
[8] M.R. Brito, A.J. Quiroz, J.E. Yukich, Intrinsic dimension identification via graph-theoretic methods, J. Multivariate Anal. 116 (2013) 263–277.
[9] T. Cai, J. Fan, T. Jiang, Distributions of angles in random packing on spheres, J. Mach. Learn. Res. 14 (2013) 1837–1864.
[10] P. Campadelli, E. Casiraghi, C. Ceruti, A. Rozza, Intrinsic dimension estimation: relevant techniques and a benchmark framework, Math. Probl. Eng. 2015 (2015).
[11] C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi, P. Campadelli, DANCo: an intrinsic dimensionality estimator exploiting angle and norm concentration, Pattern Recognit. 47 (2014) 2569–2581.
[12] J.A. Costa, A. Girotra, A. Hero, Estimating local intrinsic dimension with k-nearest neighbor graphs, in: Statistical Signal Processing, 2005 IEEE/SP 13th Workshop on, IEEE, 2005, pp. 417–422.
[13] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, 1996.
[14] A.m. Farahmand, C. Szepesvári, J.-Y. Audibert, Manifold-adaptive dimension estimation, in: Proceedings of the 24th International Conference on Machine Learning, in: ICML '07, ACM, New York, NY, USA, 2007, pp. 265–272.
[15] P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors, Physica D 9 (1983) 189–208.
[16] M. Hein, J.-Y. Audibert, Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$, in: Proceedings of the 22nd International Conference on Machine Learning, in: ICML '05, ACM, New York, NY, USA, 2005, pp. 289–296.
[17] S. Janson, On concentration of probability, in: Contemporary Combinatorics, János Bolyai Math. Soc., Budapest, 2002, pp. 289–301.
[18] E. Kaufmann, R.-D. Reiß, On conditional distributions of nearest neighbors, J. Multivariate Anal. 42 (1992) 67–76.
[19] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2324.
[20] E. Levina, P.J. Bickel, Maximum likelihood estimation of intrinsic dimension, in: Advances in neural information processing systems, 2005, pp. 777–784.
[21] G. Lombardi, A. Rozza, C. Ceruti, E. Casiraghi, P. Campadelli, Minimum neighbor distance estimators of intrinsic dimension, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2011, pp. 374–389.
[22] M.D. Penrose, J.E. Yukich, Central limit theorems for some graphs in computational geometry, Ann. Appl. Probab. 11 (2001) 1005–1041.
[23] M.D. Penrose, J.E. Yukich, Limit theory for point processes in manifolds, Ann. Appl. Probab. 23 (2013) 2161–2211.
[24] K.W. Pettis, T.A. Bailey, A.K. Jain, R.C. Dubes, An intrinsic dimensionality estimator from near-neighbor information, IEEE Trans. Pattern Anal. Mach. Intell. (1979) 25–37.
[25] R.H. Randles, D.A. Wolfe, Introduction to the Theory of Nonparametric Statistics, John Wiley & Sons, New York, 1979.
[26] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[27] R.J. Serfling, Approximation Theorems of Mathematical Statistics, Wiley, New York, 1980.
[28] V. Sindhwani, M. Belkin, P. Niyogi, The geometric basis of semi-supervised learning, in: O. Chapelle, B. Schlkopf, A. Zien (Eds.), Semi-Supervised Learning, The MIT Press, 2006, pp. 35–54.
[29] A. Södergren, On the distribution of angles between the N shortest vectors in a random lattice, J. Lond. Math. Soc. (2) 84 (2011) 749–764.
[30] K. Sricharan, R. Raich, A.O. Hero, Optimized intrinsic dimension estimator using nearest neighbor graphs, in: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, IEEE, 2010, pp. 5418–5421.
[31] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[32] C. Villani, Optimal Transport: Old and New, Springer, Berlin, 2009.
[33] J.A. Wellner, Lecture notes for course in advanced theory of statistical inference, https://www.stat.washington.edu/jaw/COURSES/580s/581/lectnotes.18html, 2006 Accessed: January 29, 2019.
[34] J.E. Yukich, Probability Theory of Classical Euclidean Optimization Problems, Springer-Verlag, Berlin, 1998.
[35] K. Zhang, Spherical cap packing asymptotics and rank-extreme detection, IEEE Trans. Inform. Theory 63 (2017) 4572–4584.