Journal of Multivariate Analysis 116 (2013) 263–277

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

Intrinsic dimension identification via graph-theoretic methods

M.R. Brito [a,c], A.J. Quiroz [b,c,*], J.E. Yukich [d]

[a] Dpto. de Matemáticas Puras y Aplicadas, Universidad Simón Bolivar, Venezuela
[b] Dpto. de Cómputo Científico y Estadística, Universidad Simón Bolivar, Venezuela
[c] Dpto. de Matemáticas, Universidad de Los Andes, Colombia
[d] Department of Mathematics, Lehigh University, USA

ARTICLE INFO

Article history:
Received 18 August 2011
Available online 11 January 2013

AMS subject classifications:
primary 62H99
60D05
secondary 62G99

Keywords:
Intrinsic dimension
Graph theoretical methods
Stabilization methods
Dimensionality reduction

ABSTRACT

Three graph theoretical statistics are considered for the problem of estimating the intrinsic dimension of a data set. The first is the "reach" statistic, $\bar{r}_{j,k}$, proposed in Brito et al. (2002) [4] for the problem of identification of Euclidean dimension. The second, M_n, is the sample average of squared degrees in the minimum spanning tree of the data, while the third statistic, U_n^k, is based on counting the number of common neighbors among the k-nearest, for each pair of sample points $\{X_i, X_j\}$, $i < j \leq n$. For the first and third of these statistics, central limit theorems are proved under general assumptions, for data living in an m-dimensional C^1 submanifold of $\mathbb{R}^d$, and in this setting, we establish the consistency of intrinsic dimension identification procedures based on $\bar{r}_{j,k}$ and U_n^k. For M_n, asymptotic results are provided whenever data live in an affine subspace of Euclidean space. The graph theoretical methods proposed are compared, via simulations, with a host of recently proposed nearest neighbor alternatives.

1. Introduction

In the statistical analysis of complex data sets, it is frequently the case that data points are represented in a high dimensional Euclidean space, $\mathbb{R}^d$, but the data actually live in an m-dimensional manifold, with $m \ll d$. This is the case, for instance, when one takes a set of digital images of the same subject. Thinking of black and white images, for simplicity, each image is coded by a very high dimensional vector, in which each coordinate represents the light intensity in a pixel, and the number of pixels may be 400 × 600 or more. Still, if we represent each image by the point in $\mathbb{R}^3$ from where it was taken, knowledge of this three dimensional representation will be enough for understanding and classification of the set of pictures. Actually, in this example, depending on how the pictures are taken, the value of m could be 1, 2 or 3. Similar situations arise in other image analysis applications, in the analysis of text and of genetic data, and in most cases, the value of m might not be as obvious as in our example. Several authors in the artificial intelligence literature have argued about the convenience of having methods to find automatically, whenever possible, a low-dimensional manifold representation of high dimensional data [2,3,7,25,27,28,30].

Methods for the representation of high dimensional data in a lower dimensional manifold are termed *manifold projection methods*. Two popular such methods are Isomap, of Tenenbaum, de Silva and Langford [30] and the Locally Linear Embedding method of Roweis and Saul [25]. In these references, it is suggested that the intrinsic dimension of the data can be "guessed" by observing the decrease in error (in the approximate representation) as the value of m varies, as was traditionally done in

[*] Corresponding author at: Dpto. de Matemáticas, Universidad de Los Andes, Carrera 1, Nro. 18A-10, edificio H, Bogotá, Colombia.
E-mail addresses: aj.quiroz1079@uniandes.edu.co, ajquiroz@gmail.com (A.J. Quiroz).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.12.007

the context of Multidimensional Scaling. A more precise initial estimation of m, of low computational cost, can be of great help in the application of manifold projection methods.

In the present article, as in [22], it will be understood that though data are given in $\mathbb{R}^d$, they actually live in an m-dimensional, C^1 submanifold $\mathcal{M}$ of $\mathbb{R}^d$, $m \leq d$, endowed with the subset topology, and that the goal is the estimation of the dimension m from an i.i.d. sample.

Various nearest-neighbor methods attempting to estimate the value of m have appeared recently in the statistical and artificial intelligence literature. Some of the ideas behind these methods can be traced to Pettis et al. [23] and Grassberger and Procaccia [11]. With the brief description of these methods we start introducing notation that will be used throughout. Let $\mathcal{X}_n := \{X_i\}_{i=1}^n$ denote the i.i.d. sample. The dimension estimator of Grassberger and Procaccia [11] is based on the U-statistic which computes the fraction of data pairs, X_i, X_j satisfying $\|X_i - X_j\| < r$, for positive r. Secondly, let $D_k(X_i) := D_k(X_i, \mathcal{X}_n)$ be the distance from the sample point X_i to its k-th nearest neighbor in the sample. Denoting $\overline{D}_k$ the sample average of these distances, for each k in a given range, the nearest neighbor dimension estimator of Pettis et al. [23] is based on plotting $\log(\overline{D}_k)$ against $\log(k)$ and estimating the slope of the plot. Thirdly, Kegl [14] suggests using sample estimators of the packing number corresponding to the dimension where the data live, an idea motivated on the fast variation of packing numbers with dimension. The difficulty with this approach is that precise estimation of the packing number is, computationally, a very hard problem, and in practice one has to work with rather crude approximations.

Neither of these three methods provides a direct estimation of intrinsic dimension m, but rather its value must be deduced indirectly, from the slope of a line, for instance. Levina and Bickel [17], on the other hand, propose a "maximum likelihood" estimator of intrinsic dimension, arguing that, asymptotically, the expected value of the statistic

$$\hat{m}_k(X_i) := \left[\frac{1}{k-2} \sum_{j=1}^{k-1} \log \frac{D_k(X_i)}{D_j(X_i)} \right]^{-1} \tag{1}$$

is the intrinsic dimension of the data. Thus, one expects that the average over the sample, namely $\overline{m}_k := n^{-1} \sum_{i=1}^n \hat{m}_k(X_i)$, is an asymptotically unbiased estimator of the intrinsic dimension. If $k \geq 4$ and if the data have a density which is bounded away from zero and infinity on a subset of a C^1 submanifold of $\mathbb{R}^d$, $m \leq d$, then Theorem 2.1 in [22] establishes that $\overline{m}_k$ converges in probability to the intrinsic dimension m, as the sample size n tends to infinity. (If $k \geq 11$ then $\overline{m}_k$ converges a.s., as $n \to \infty$, to the intrinsic dimension.) Theorem 4.3 in [3] establishes a central limit theorem for $\overline{m}_k$ as $n \to \infty$, for data in an affine subspace of $\mathbb{R}^d$, whereas for data belonging to a manifold, Theorem 2.1 in [22] establishes a central limit theorem for $\overline{m}_k$, conditioned on nearest neighbor distances being not too large. Levina and Bickel carry out comparisons of their procedure against the correlation dimension estimator of Grassberger and Procaccia [11] and the nearest-neighbor estimator of Pettis et al. [23], and conclude that $\overline{m}_k$ has a better performance.

Costa, Girotra and Hero [5] propose a method based on the total (power) length of the k-nearest-neighbor graph, namely

$$L_{\gamma,k} := L_{\gamma,k}(\mathcal{X}_n) = \sum_{i=1}^{n} \sum_{j=1}^{k} D_j(X_i)^{\gamma}, \tag{2}$$

where γ is a power weighting constant. The use of (2) is justified by the following law of large numbers (Theorem 1 in [5]). Assume that the data sample lives in an m-dimensional compact Riemannian submanifold $\mathcal{M}$, of $\mathbb{R}^d$, and is obtained from a continuous distribution on $\mathcal{M}$ with density f bounded away from zero and infinity. Let g be the Riemannian metric defined on $\mathcal{M}$ and let μ_g be the associated volume measure. If $1 \leq \gamma \leq m$, $m \geq 2$, and d' is any positive integer, then, almost surely,

$$\lim_{n\to\infty} \frac{L_{\gamma,k}}{n^{(d'-\gamma)/d'}} = \begin{cases} \infty & \text{if } d' < m \\ \beta_{m,\gamma,k} \int_{\mathcal{M}} f^{\alpha}(x) d\mu_g(x) & \text{if } d' = m \\ 0 & \text{if } d' > m \end{cases} \tag{3}$$

where $\alpha = (m-\gamma)/m$ and $\beta_{m,\gamma,k}$ is a constant not depending on f, $\mathcal{M}$ or g.

From the above law of large numbers, [5] proposes a consistent bootstrap procedure for the estimation of m: Let $p_1, \ldots, p_Q$ be integers in $(0, n)$ and N a fixed fraction of n. For each l in $1, \ldots, Q$, produce N samples of size p_l, say $\mathcal{X}_{p_l}^j$, for $j = 1, \ldots, N$, by resampling with replacement from the original sample $\mathcal{X}_n$. Fix l for the moment. For each bootstrap sample $\mathcal{X}_{p_l}^j$, compute $L_{\gamma,k}(\mathcal{X}_{p_l}^j)$. Average the N values obtained to get $\bar{L}_{p_l}$, and let $\bar{l}_{p_l} = \ln \bar{L}_{p_l}$. After this calculus has been done for each l, adjust the model

$$\bar{l}_{p_l} = a \ln p_l + b + \epsilon_l$$

by ordinary least squares. If $\hat{a}$ stands for the estimator of a in the model above, then $\hat{m} = \text{round}(\gamma/(1-\hat{a}))$ is a consistent (as n goes to infinity) estimator of m (see [5] for more details).

Farahmand, Szepesvári and Audibert [7] present a direct estimate of m, based on nearest neighbor distances. Given $r \in (0, \infty)$ and $x \in \mathbb{R}^d$, let $B_r(x)$ denote the ball in $\mathbb{R}^d$ of radius r and center x. Define $\eta(x, r)$ by

$$P[X_i \in B_r(x)] =: \eta(x, r) r^m.$$

For small r, $\eta(x, r)$ is an approximation to the data density at $x \in \mathcal{M}$. For $i \leq n$, let

$$\hat{m}(X_i) := \frac{\ln 2}{\ln(D_k(X_i)/D_{\lceil k/2 \rceil}(X_i))} \quad \text{and} \quad \hat{m} := \text{round}\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{m}(X_i) \wedge d)\right), \tag{4}$$

where $a \wedge b := \min(a, b)$. Under differentiability assumptions on the function η and regularity assumptions on $\mathcal{M}$, exponential bounds for the probability that $\hat{m}$, in (4), differs from m, for large enough n, are given in [7]. In particular, under those assumptions, this estimator is consistent in probability for the dimension m.

Recently, Sricharan, Raich, and Hero [28] have proposed another direct procedure for estimation of m, using ideas related to k-nearest neighbors density estimators. Partition the sample $\mathcal{X}_n$ into two disjoint sets, say $\mathcal{Y}$ and $\mathcal{Z}$ of respective sizes N and M. For each $Y_i \in \mathcal{Y}$, let $R_k(Y_i)$ denote the distance from Y_i to its k-th nearest neighbor in $\mathcal{Z}$. For a positive γ, let

$$T_k(\mathcal{X}_n) := \frac{\gamma}{N}\sum_{i=1}^{N}\log R_k(Y_i).$$

The "improved" estimator for m proposed in [28] is

$$\hat{m} := \frac{\gamma\,(\log(k_2 - 1) - \log(k_1 - 1))}{T_{k_2}(\mathcal{X}_n) - T_{k_1}(\mathcal{X}_n)} \tag{5}$$

for $k_1 < k_2$. Although no theoretical results are given in [28] for the estimator (5), approximate formulas for the bias and variance and a central limit theorem are stated for an associated estimator and it is argued that the behavior, in terms of variance, of the improved version should be better than that of the associated statistic. Also, in [28] optimal criteria are obtained for the choice of M, N, k_1 and k_2 in (5), but these criteria are of limited practical application since the formulas obtained for the parameters depend on some unknown constants, including the dimension m to be estimated.

In the following section we present the graph theoretical methods for intrinsic dimension identification to be discussed here. Section 3 includes results of a Monte Carlo performance comparison of the graph theoretical methods with the k-nearest neighbor methods of [5,7,17,28], described above. Section 4 provides some theoretical results that support the methods introduced here, including central limit theorems for the "reach" statistic and the "mutual neighbors" statistic of Section 2, assuming that the data live in a C^1 manifold.

2. Graph theoretical methods

Since the series of papers of Friedman and Rafsky [8–10] it has been clear that graph theoretic methods offer a natural way of dealing with non-parametric statistics in the multivariate setting. The article [24] gives a brief account of the different applications of graph theoretic methods on a variety of statistical problems, including the two-sample problem, outlier identification and clustering, among others. In this reference, or in any classical book on Graph Theory (for instance, Harary, [12]) the reader can find the graph theoretic definitions required in what follows. Penrose and Yukich [19–22] as well as [31] explain some of the main mathematical ideas that validate, asymptotically, these sorts of methods.

Starting from a d-dimensional data set $\mathcal{X}_n$, or more precisely, from the corresponding interpoint distances, different graphs can be built that establish connections (edges) between nearby sample points. Examples include the k-nearest-neighbor graph, G_k, the minimum spanning tree graph and the sphere of influence graph, which are three types of "proximity graphs"; see Aldous and Shun [1]. For a description of the first two, the reader can look in the references of Friedman and Rafsky [8–10], while the sphere of influence graph, has been studied in [19]. Once a proximity graph H has been constructed from a data sample, we may abstract from the data and look at statistics which are functions only of graph theoretic properties of H, such as vertex degrees, vertex eccentricities, length of paths between vertices, diameter of the graph, and so on. Procedures based on these types of statistics are what we call graph theoretic methods. Since the estimators of intrinsic dimension of [5,7,17,28] require for their computation the actual values of the nearest neighbor distances, they do not fall in the graph theoretic category. The main advantages of using graph theoretic methods are, in our opinion:

(i) Low computational cost. Most of the graphs considered in these methods and the statistics calculated from them can be computed at a sub-quadratic cost, with respect to the sample size. See [8] and references therein.

(ii) Robustness. The construction of the graphs does not require exact knowledge of the interpoint distance. Only the ordering among these distances is required. Thus, for data coming from a continuous distribution, there is an $\epsilon > 0$ such that statistics based on these graphs are invariant if the data is perturbed by a noise component of Euclidean norm less than ϵ. Small amounts of noise will thus not affect the analysis in an important manner, unless a significant fraction of distance comparisons is altered.

(iii) Availability of theory. A number of theoretical tools exist for the study of properties of graph theoretic statistics, including methods for proving laws of large numbers, central limit theorems and consistency of the corresponding parameter estimators. See [19,22,31]. In many important problems, the graph theoretic procedures turn out to be asymptotically non-parametric.

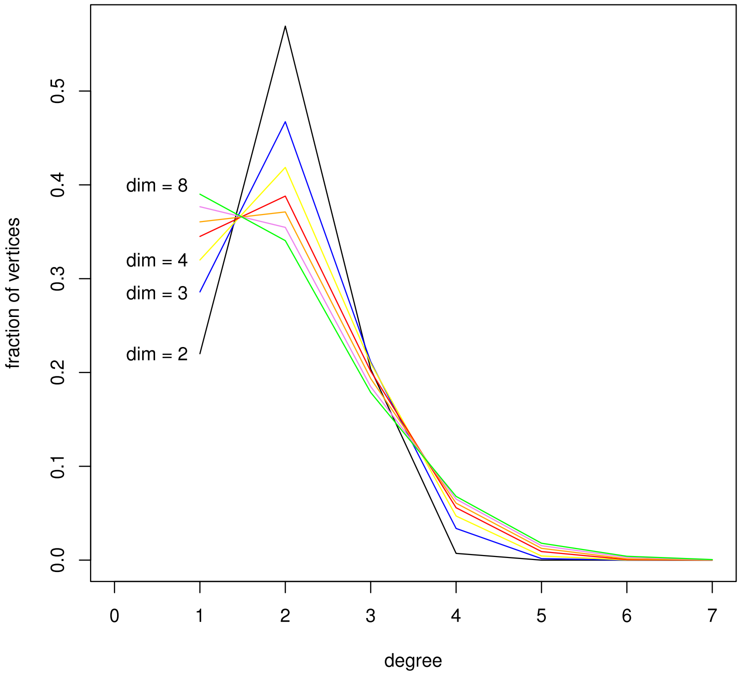

Fig. 1. Average degree frequencies in MST.

In the context of Multidimensional Scaling, Brito, Quiroz and Yukich [4] propose a graph theoretical method for identifying the Euclidean dimension of a data set. The method is based on the “reach” of data in the k-nearest neighbor graph, defined as follows. Recall that in the k-nearest-neighbor graph, $G_k := G_k(\mathcal{X}_n)$ the vertex set is the sample $\mathcal{X}_n$. For $x, y \in \mathcal{X}_n, x \neq y$ and j a fixed positive natural number, say that y can be reached in j steps from x, if there exists a path $v_0, v_1, \ldots, v_j$ in G_k with $v_0 = x$ and $v_j = y$. The reach in j steps of vertex $x \in \mathcal{X}_n$, $r_{j,k}(x, \mathcal{X}_n)$, is the total number of vertices that can be reached from x in j steps or less in G_k. The statistics considered in [4] is the average reach in j steps of points in G_k, namely

$$\bar{r}_{j,k}(\mathcal{X}_n) := \frac{1}{n} \sum_{x \in \mathcal{X}_n} r_{j,k}(x, \mathcal{X}_n).$$

The intuition for considering the statistic $\bar{r}_{j,k}(\mathcal{X}_n)$ is that, as the dimension increases, there are more directions in which a given sample point can find neighbors and, therefore, the amount of points reached in j steps in G_k, should increase with dimension. For data living in Euclidean space, a law of large numbers for $\bar{r}_{j,k}(\mathcal{X}_n)$ is established in [4]. This result is extended here to data living in a C^1 submanifold of $\mathbb{R}^d$ and, in this context, a central limit theorem for $\bar{r}_{j,k}(\mathcal{X}_n)$ is given as well (see Remark (ii) following Theorem 1 in Section 4). This leads to a consistency result for an estimator of intrinsic dimension based on $\bar{r}_{j,k}(\mathcal{X}_n)$ (Theorem 4).

To introduce a second estimator of intrinsic dimension, let $T_n := T_n(\mathcal{X}_n)$ be the minimum spanning tree (MST) associated to the sample $\mathcal{X}_n$. Let $\deg(X_i) := \deg_{T_n(\mathcal{X}_n)}(X_i)$ denote the degree of node X_i in T_n, that is, the number of X_j's, $j \neq i$, such that $\{X_i, X_j\}$ is an edge of T_n. Steele, Shepp and Eddy [29] showed that for data obtained from a continuous distribution on $\mathbb{R}^d$, the fraction of nodes with a given degree, j, in T_n, converges almost surely to a limit depending only on j and d. Fig. 1 shows the average fractions of nodes with each given degree up to 7, for MSTs in dimension 2–8 for samples of size $n = 1000$ from the Uniform distribution on the unit cube. For each dimension, the averages are computed over 50 samples. We see, in this figure, that the degree distribution in the MST is monotonically changing with dimension: As the dimension increases, the number of leaves (nodes of degree 1) is increasing, as is the fraction of nodes with large degree (degree $\geq$4). Recall that the average degree in a tree is a constant, depending only in the number of vertices. On the other hand, that the fraction of nodes with large degree increases with dimension, suggests that the average of a power greater than 1 of the node degrees in the MST might capture the difference between dimensions. We are thus motivated to consider the following estimator of intrinsic dimension:

$$M_n := M(\mathcal{X}_n) := \frac{1}{n} \sum_{i=1}^{n} (\deg(X_i))^2. \tag{6}$$

Fig. 2 shows average values of M_n for different dimensions, up to 12. For each dimension, the average is obtained from a set of 100 M_n values computed on independent samples of size $n = 1000$ from the d-dimensional Uniform distribution. It is clear that the mean of this random variable behaves monotonically with dimension in the range considered.

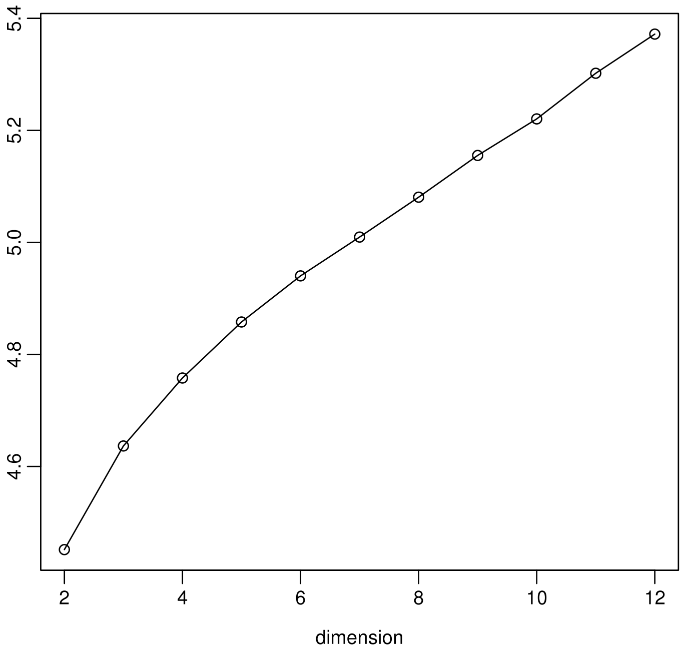

Fig. 2. Average M_n as function of dimension.

Mutual neighbor and neighbor sharing probabilities are studied by Schilling [26] in connection with test statistics for the multivariate two-sample problem. Schilling shows that these probabilities, for samples from continuous distributions, converge to limits that do not depend on the particular density producing the sample. In particular, the limits of mutual neighbor probabilities (for samples from continuous distributions in Euclidean space) depend only on the dimension of the support of the data. This fact motivates our third estimator of dimension identification, defined as follows.

The sphere of influence graph, studied, for instance, in [19], is a proximity graph defined as follows: For each sample point, X_i, consider the closed ball $B_{\rho(i)}(X_i)$ with center X_i and radius $\rho(i)$ equal to the distance between X_i and its nearest neighbor in the sample. The sphere of influence graph $S_1 := S_1(\mathcal{X}_n)$ has the sample $\mathcal{X}_n$ as its vertex set, and an edge exists between X_i and X_j iff the corresponding nearest neighbor balls intersect, that is, iff $\|X_i - X_j\| \leq \rho(i) + \rho(j)$. We shall consider a generalization of the sphere of influence graph, as follows. Let S_k, called the k-sphere of influence graph, be defined in the same way as S_1 but instead of $\rho(i)$, we shall use $\rho_k(i)$, defined as the distance between X_i and its k-th nearest neighbor in the sample. That is, an edge exists between X_i and X_j in S_k if $\|X_i - X_j\| \leq \rho_k(i) + \rho_k(j)$. This is a way of enriching the sphere of influence graph, by adjoining more edges to it. We define a third graph theoretical intrinsic dimension statistic in the context of S_k. Let $N_{i,j}$ denote the number of sample points, other than X_i and X_j, in the intersection $B_{\rho_k(i)}(X_i) \cap B_{\rho_k(j)}(X_j)$. $N_{i,j}$ is the number of neighbors, among the k-nearest, that the points X_i and X_j share. A similar intuition to the one motivating the consideration of $\bar{r}_{j,k}$, leads us to consider the statistic

$$U_n^k := U^k(\mathcal{X}_n) := \frac{1}{n}\sum_{i<j} N_{i,j}. \tag{7}$$

The intuition is that, as the dimension grows, every point is more likely to be closer to any other point in the sample, in terms of length of the connecting path in G_k, and thus, the expected number of neighbors shared by a given pair of data points, should grow with dimension. Although (7) suggests that the computation of U_n^k has quadratic complexity in the sample size, it turns out to be easy to compute, once the k-th nearest neighbor graph has been created, with computational complexity $O(nk)$, as follows. Suppose we have a $n \times k$ table that contains the identities (indices), for each $i \leq n$, of the k nearest neighbors of X_i in the sample. One pass through this table is enough to compute, for each X_i, the random variable

$$c_k(X_i, \mathcal{X}_n) := \operatorname{card}\{j \leq n : X_i \text{ is one of the } k \text{ nearest neighbors of } X_j \text{ in the sample } \mathcal{X}_n\}.$$

Since each sample point X_l, $1 \leq l \leq n$, appears in a total of $\binom{c_k(X_l, \mathcal{X}_n)}{2}$ summands $N_{i,j}$, we get that U_n^k in (7) can be computed as

$$U^k(\mathcal{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\binom{c_k(X_i, \mathcal{X}_n)}{2}. \tag{8}$$

It follows that the computational complexity of U_n^k is the same as that of constructing the k-th nearest neighbor graph, which is less than quadratic, according to Friedman and Rafsky [8].

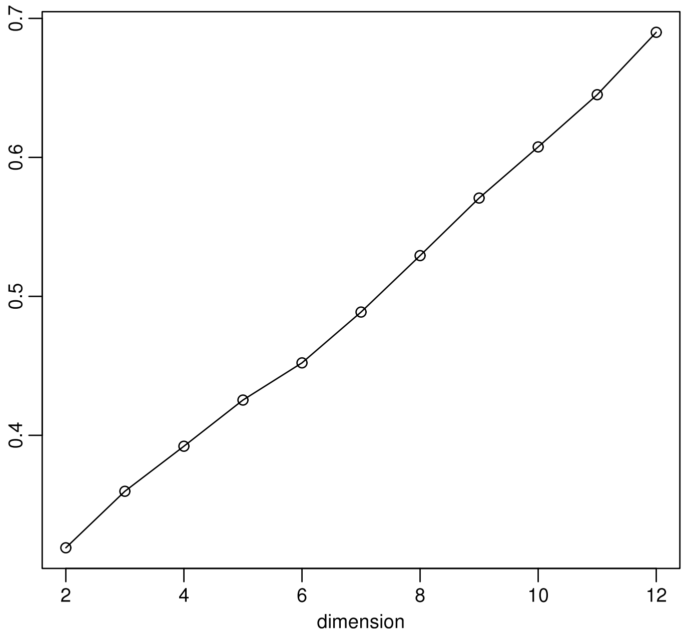

Fig. 3. Average U_n^1 as function of dimension.

Fig. 3 shows the averages of $U_n^1 := U_n^1(\mathcal{X}_n)$ values, for dimensions 2–12. As in Fig. 2, for each dimension, the U_n^1 are computed on 100 samples of size $n = 1000$, from the Uniform distribution on the d-dimensional unit cube. Again, we see a clear tendency of the statistic $U_n^1(\cdot)$ to grow nearly linearly with dimension, suggesting that it might be appropriate for the identification of intrinsic dimension. For values of k larger than 2, (experiments not included here) a similar behavior of $U_n^k(\mathcal{X}_n)$ is observed, as a function of dimension.

Figs. 2 and 3 and the results in [4] suggest that, at least for a relevant set of dimensions, a linear dependence can be established between all the graph theoretical statistics presented above and the underlying data dimension. In each case, one could find a linear formula to predict the dimension from the value of the graph theoretical statistic. We shall not pursue this line of research, since it would not take into consideration the variability that each graph theoretical statistic presents in each dimension, but only the average value of the statistics in the different dimension. To take advantage of all the information obtained from simulations, including estimated means and variances, and the information provided by theoretical results (Section 4) regarding the asymptotic Gaussian distribution of the statistics, it seems more sensible to employ a Bayesian decision theoretic procedure, described below, since it is known to converge to the smallest possible error rate based on a given statistic (see Chapter 2 in [6]). A drawback of our approach is the need to estimate parameters of the distribution of the statistic to be used, but this is a one-time cost, since simulations can be performed for the Uniform distribution on the d-dimensional unit cube and, by the asymptotic theory, the parameters estimated will be approximately valid for any data distribution satisfying the assumptions of the results in Section 4. Another possible objection to our proposed methodology is the reliance on asymptotic results, but this is essentially true of all the methods currently available for this problem. For instance, the statistic of Levina and Bickel [17] is known to be biased for small sample size, but asymptotically unbiased, and its use can be justified by the existence of a strong law of large numbers (Theorem 2.1 in [22]). Similarly the method of Costa, Girotra and Hero [5] is supported by a strong law of large numbers for the total length statistic $L_{\gamma,k}(\mathcal{X}_n)$. Perhaps the one exception to this dependence on asymptotics is the method of Farahmand, Szepesvári and Audibert [7], in which bounds for the probability of error in the estimation of m are provided for finite n, although the values of n required for the bounds to be valid depend on constants that are not known.

We shall now describe the setup for the approximate Bayesian estimators to be used with each of the three graph theoretic statistics described above (the reach statistic, $\bar{r}_{j,k}$, the average of squared degrees in the MST, M_n, and the mutual neighbors statistic, U_n^k). In the context of manifolds we prove central limit theorems only for the first and third of these statistics, though we shall assume that a Gaussian approximation is valid for all of them. In the remainder of this section, $S_n := S_n(\mathcal{X}_n)$ denotes any of these graph theoretic statistics. When the data live in an m-dimensional C^1 submanifold of $\mathbb{R}^d$, we assume that $S_n(\mathcal{X}_n)$ converges in L^2 to a limit, $\mu(m)$, that does not depend on the particular density producing the sample, but only on m. We assume further that

$$n\,\mathrm{Var}(S_n(\mathcal{X}_n)) \to \sigma^2(m), \quad \text{as } n \to \infty, \tag{9}$$

for a positive constant $\sigma^2(m)$ depending only on m. Finally, we assume that

$$\sqrt{n}(S_n(\mathcal{X}_n)) - \mathbb{E}(S_n(\mathcal{X}_n)) \xrightarrow{\mathcal{D}} Z, \tag{10}$$

Table 1
Estimated parameters for graph theoretic statistics, $n = 1000$.

Dimension	$\bar{r}_{2,4}$		M_n		U_n^1	
	$\tilde{\mu}_d$	$\tilde{\sigma}_d/\sqrt{n}$	$\tilde{\mu}_d$	$\tilde{\sigma}_d/\sqrt{n}$	$\tilde{\mu}_d$	$\tilde{\sigma}_d/\sqrt{n}$
2	12.58	0.161	4.452	0.0172	0.319	0.0176
3	15.02	0.202	4.637	0.0219	0.360	0.0235
4	17.01	0.249	4.758	0.0303	0.392	0.0233
5	18.75	0.273	4.860	0.0341	0.425	0.0251
6	20.27	0.312	4.940	0.0434	0.452	0.0275
7	21.76	0.334	5.010	0.0429	0.489	0.0310
8	23.14	0.346	5.081	0.0405	0.529	0.0303
9	24.48	0.398	5.155	0.0472	0.571	0.0342
10	25.65	0.404	5.220	0.0504	0.608	0.0419
11	26.98	0.399	5.302	0.0633	0.645	0.0452
12	28.13	0.487	5.371	0.0648	0.690	0.0459

for a Gaussian random variable Z with mean 0 and variance $\sigma^2(m)$. At least in the case of $\bar{r}_{j,k}$ and U_n^k these assumptions are supported by the results in Section 4. We consider a finite set F of candidate values for the intrinsic dimension m. For each dimension j in F, we produce, by simulation, L samples of size n (large) from the Uniform distribution on the unit j-cube. For each sample generated, the statistic S_n is computed, and from the L values produced, we obtain the natural estimators $\tilde{\mu}_j$ and $\tilde{\sigma}^2(j)$ of the parameters μ_j and $\sigma^2(j)$, for each j in F. In the Monte Carlo experiments used for parameter estimation and described in Section 3, we used $F = \{2, 3, \dots, 12\}$, $n = 1000$ and $L = 100$ for each statistic.

Suppose now a new data sample, of size n', assumed to live in an m-dimensional submanifold of $\mathbb{R}^d$ is presented and the value of m for these data is to be estimated. Based on the simulation results, the density of $S_n(\mathcal{X}_n)$ in dimension j, evaluated at $s \in \mathbb{R}$, is approximated as $\tilde{f}_j(s)$, the Gaussian density with mean $\tilde{\mu}_j$ and variance $\tilde{\sigma}^2(j)/n'$. Then, we compute the value of the graph theoretic statistic on the new data set, $S_{n'}$. Elementary Bayesian Decision Theory tells us that, assuming equal a priori probabilities for all the dimensions in F, the a posteriori probabilities, $P[j \mid S_{n'}]$ corresponding to the dimensions considered, are

$$P[j \mid S_{n'}] = \frac{\tilde{f}_j(S_{n'})}{\sum\limits_{l \in F} \tilde{f}_l(S_{n'})} \quad \text{for } j \in F. \tag{11}$$

Choosing the value of j that maximizes (11) as the estimator of intrinsic dimension, corresponds to employing an approximation to the Bayesian Classifier, which is the best possible procedure based on $S_{n'}$ (see Chapter 2 in [6], for instance). Still, since the intrinsic dimension is a numerical (non categorical) variable, it seems sensible to combine the information in the a posteriori probabilities and to use, as an estimator, the a posteriori expected value of the intrinsic dimension, given by

$$\tilde{m} := \text{round}\left(\frac{\sum\limits_{j \in F} j \tilde{f}_j(S_{n'})}{\sum\limits_{j \in F} \tilde{f}_j(S_{n'})} \right). \tag{12}$$

Formula (12) is the one we shall use for our graph-theoretic estimation of intrinsic dimension; a proof of consistency of this method, valid when S_n is $\bar{r}_{j,k}$ or U_n^k, is given in Theorem 4.

3. Monte Carlo evaluation of procedures for intrinsic dimension identification

The simulations described in the present section were carried out on a laptop computer using the statistical language R.

We report first the result of the Monte Carlo experiments carried out for parameter estimation for the graph theoretic statistics $\bar{r}_{j,k}$, M_n, and U_n^k. For the reach statistic $\bar{r}_{j,k}$, the parameters were set to $k = 4$ and $j = 2$, while for U_n^k we worked with $k = 1$ (the statistic M_n does not require choice of parameters). For each statistic and each dimension j in $F = \{2, 3, \dots, 12\}$, $M = 100$ samples of size $n = 1000$ were generated from the Uniform distribution on the j-dimensional unit cube. For each sample, the statistic considered is computed, and from the M values available the sample mean is used as the estimator $\tilde{\mu}_j$ of the mean, while, if s^2 denotes the sample variance of the statistic values observed, $\tilde{\sigma}^2(j) = n s^2$ is used as estimator of $\sigma^2(j)$ in (9). Table 1 presents the results of this estimation for the three graph theoretical statistics. The numbers in columns $\tilde{\mu}_d$, for M_n and U_n^1, are the ones plotted in Figs. 2 and 3. We can see in the $\tilde{\mu}_d$ values, for the three statistics in this table, a nearly linear growth of the means. On the other hand, the curse of dimensionality manifests itself, in Table 1, in that the standard deviations of the statistics tend to grow with dimension, making it more difficult to distinguish between consecutive dimensions when they are larger.

To evaluate the performance of the graph theoretic methods proposed, a Monte Carlo comparison was carried out, in which we included the recently developed nearest-neighbor methods described in Section 1 and a set of manifolds with dimensions m varying from 2 to 9, most of which have appeared in similar studies in the literature:

Table 2
Estimated MSE for Costa et al. estimator.

Data set	$n = 250$		$n = 500$		$n = 1000$	
S^3	0.62	(1)	0.665	(1)	0.659	(1)
S^6	7.311	(9)	7.096	(9)	6.97	(9)
S^9	25.703	(25)	24.298	(25)	24.205	(25)
$\mathscr{P}_3$	0.62	(1)	1.001	(1)	0.85	(1)
$\mathscr{P}_6$	13.838	(16)	11.445	(9.467)	10.032	(9)
$\mathscr{P}_9$	46.98	(49)	40.186	(37.733)	35.147	(36)
Swiss Roll	0.059	(0)	0.053	(0)	0.047	(0)
Möbius	0.043	(0)	0.099	(0)	0.147	(0)

Table 3
Estimated MSE for Farahmand et al. estimator.

Data set	$n = 250$		$n = 500$		$n = 1000$	
S^3	0.298	(0.5)	0.287	(0.567)	0.286	(0.633)
S^6	0.331	(0.5)	0.35	(0.667)	0.566	(1)
S^9	0.085	(0.067)	0.098	(0.1)	0.105	(0.1)
$\mathscr{P}_3$	0.298	(0.5)	0.048	(0)	0.086	(0)
$\mathscr{P}_6$	1.904	(1.5)	0.871	(1)	0.311	(0.733)
$\mathscr{P}_9$	16.041	(16.3)	8.97	(9)	5.245	(4.167)
Swiss Roll	1.025	(1)	1.017	(1)	0.722	(1)
Möbius	0.624	(1)	0.224	(0.333)	0.122	(0)

(i) Möbius band data (as considered in [7]). This is a 2-dimensional manifold data set produced as follows: For $U \sim$Unif$(-1/2, 1/2)$ and $V \sim$ Unif$(0, 2\pi)$, let the coordinates of $X \in \mathbb{R}^3$ be given by $X_1 = (1+U\cos(5V))\cos(V)$, $X_2 = (1 + U\cos(5V))\sin(V)$ and $X_3 = U\sin(5V)$.

(ii) The "Swiss Roll" data, introduced in Tenenbaum, de Silva and Langford [30] and considered in several other articles, is another 2-dimensional manifold data set that is obtained by generating first a bivariate mixture of Gaussian sample, and then "twisting" the sample into $\mathbb{R}^3$ by applying the following transformation to each bivariate datum (Z_1, Z_2): $X_1 = Z_1\cos(Z_1)$, $X_2 = Z_2$ and $X_3 = Z_1\sin(Z_1)$.

(iii) Uniform data in the unit sphere S^m. This data was generated for $m = 3$, 6 and 9.

(iv) Random data in an m-dimensional paraboloid, $\mathscr{P}_m$. These data points fall in the manifold

$$x_{m+1} = x_1^2 + \cdots + x_m^2 \tag{13}$$

and are obtained by generating first $(X_1, \dots, X_m) \in \mathbb{R}^m$ with the Multivariate Burr distribution with parameter $\alpha = 1$ (see [13, Chapter 9]). This means that, for $i \le m$, $X_i = (1 + E_i/E_0)^{-1}$, where $E_0, E_1, \dots, E_m$ are i.i.d. exp(1) variables. After the coordinates $(X_1, \dots, X_m)$ are generated, X_{m+1} is computed according to (13). These data were generated for $m = 3$, 6 and 9. The idea in considering the paraboloid data is to have, in the larger dimensions considered, an example with non-constant curvature, in contrast to the unit sphere data.

As in [7], after generating the manifold samples just described, redundant coordinates were added to the data by adding the coordinates $\sin(X_i)$ and X_i^2, for each coordinate X_i. In this manner, although the data actually live in a manifold of dimension m, the dimension identification procedures shall be implemented with data in $\mathbb{R}^{3(m+1)}$.

The nearest neighbor methods considered in this comparison, with their parameters, are the following:

LB: The statistic of Levina and Bickel, (1), with $k = 5$.

CGH: Costa, Girotra and Hero's method, with $\gamma = 1$, $k = 5$, $Q = 3$, $p_1 = \lceil n/4 \rceil$, $p_2 = \lceil n/2 \rceil$, $p_3 = n$ and $N = 0.1n$. See the explanation of the parameters in the lines following (2).

FSA: The estimator of Farahmand, Szepesvári and Audibert, (4), with $k = 10$.

SRH: Sricharan, Raich and Hero's estimator, (5), with $k_1 = 5$, $k_2 = 10$, $N = \lfloor n/5 \rfloor$ and $M = n - N$.

These nearest neighbor methods are compared to the approximate Bayesian classifiers corresponding to the graph theoretic procedures proposed here. The Bayesian classifiers are set up with the estimated mean and variance values given in Table 1 and the choice of parameters made for that table. For comparison purposes, for each method of classification, each data type, each value of the intrinsic dimension, m, and each sample size, $n = 250$, 500 and 1000, 30 samples were generated and presented to the dimension estimator. For each estimator, the final value is obtained by rounding, to the next integer, a value produced by the estimator formula.

Tables 2–8 present the mean squared errors (MSEs) for the different estimators, before and after rounding. The value in parenthesis is the MSE for the rounded (integer) value of the estimator. In these MSE tables, we observe the following: The estimator of Costa, Girotra and Hero (CGH) presents a very good behavior in identification of dimension 2, in which case the MSE for the rounded value of the estimator is 0 for both 2-dimensional data distributions and all sample sizes, but its performance deteriorates rapidly with increasing dimension, and for $m = 9$ the estimator presents a bias of about

Table 4
Estimated MSE for Levina–Bickel estimator.

Data set	$n = 250$		$n = 500$		$n = 1000$	
S^3	0.016	(0)	0.009	(0)	0.007	(0)
S^6	0.115	(0.133)	0.09	(0.033)	0.037	(0)
S^9	1.056	(1.067)	0.612	(0.767)	0.468	(0.767)
$\mathcal{P}_3$	0.016	(0)	0.033	(0)	0.021	(0)
$\mathcal{P}_6$	2.362	(2.8)	1.634	(1)	1.101	(1)
$\mathcal{P}_9$	13.007	(13.433)	9.019	(8.833)	6.681	(7.667)
Swiss Roll	0.277	(0.533)	0.261	(0.333)	0.285	(0.367)
Möbius	0.009	(0)	0.004	(0)	0.002	(0)

Table 5
Estimated MSE for Sricharan et al. estimator.

Data set	$n = 250$		$n = 500$		$n = 1000$	
S^3	0.112	(0.167)	0.096	(0.1)	0.071	(0)
S^6	0.154	(0.233)	0.19	(0.333)	0.15	(0.167)
S^9	0.401	(0.567)	0.286	(0.367)	0.164	(0.133)
$\mathcal{P}_3$	0.112	(0.167)	0.037	(0.033)	0.012	(0)
$\mathcal{P}_6$	5.317	(4.967)	2.987	(3.3)	1.592	(1.3)
$\mathcal{P}_9$	29.097	(28.733)	18.592	(18.167)	11.464	(10.867)
Swiss Roll	0.184	(0.3)	0.144	(0.167)	0.123	(0.033)
Möbius	0.297	(0.5)	0.095	(0.067)	0.039	(0)

Table 6
Estimated MSE for reach estimator, $\bar{r}_{2,4}$.

Data set	$n = 250$		$n = 500$		$n = 1000$	
S^3	0.003	(0)	0.009	(0.033)	0	(0)
S^6	0.672	(0.867)	0.454	(0.633)	0.036	(0.033)
S^9	6.558	(6.667)	4.049	(4)	3.722	(4)
$\mathcal{P}_3$	0.003	(0)	0	(0)	0	(0)
$\mathcal{P}_6$	3.734	(3.767)	1.075	(1.1)	0.122	(0.2)
$\mathcal{P}_9$	26.792	(26.467)	13.906	(14.367)	4.324	(4.333)
Swiss Roll	0.31	(0.333)	0.166	(0.2)	0.064	(0.067)
Möbius	0.098	(0.133)	0	(0)	0	(0)

Table 7
Estimated MSE for estimator M_n.

Data set	$n = 250$		$n = 500$		$n = 1000$	
S^3	0.186	(0.267)	0.015	(0.033)	0.001	(0)
S^6	2.284	(2.333)	1.323	(1.4)	0.801	(0.967)
S^9	12.571	(12.767)	9.827	(10.2)	7.577	(8.367)
$\mathcal{P}_3$	0.186	(0.267)	0.014	(0)	0	(0)
$\mathcal{P}_6$	1.107	(1.167)	0.435	(0.6)	0.388	(0.467)
$\mathcal{P}_9$	5.393	(5.567)	1.2	(1.467)	0.542	(0.567)
Swiss Roll	0.078	(0.1)	0.042	(0.067)	0	(0)
Möbius	0.186	(0.2)	0	(0)	0	(0)

Table 8
Estimated MSE for mutual neighbors estimator, U_n^1.

Data set	$n = 250$		$n = 500$		$n = 1000$	
S^3	0.709	(0.667)	0.128	(0.167)	0.164	(0.2)
S^6	2.388	(2.4)	2.231	(2.033)	1.728	(1.8)
S^9	10.005	(9.967)	11.877	(12.067)	10.906	(10.8)
$\mathcal{P}_3$	0.709	(0.667)	0.371	(0.367)	0.235	(0.333)
$\mathcal{P}_6$	1.09	(1.067)	0.579	(0.633)	0.365	(0.567)
$\mathcal{P}_9$	4.293	(4.733)	3.185	(3.4)	2.717	(2.433)
Swiss Roll	1.104	(1.233)	0.461	(0.5)	0.146	(0.167)
Möbius	1.398	(1.5)	0.241	(0.333)	0.058	(0.1)

5 units for the sphere data and even more bias for the paraboloid data. This, together with the fact that this estimator is, computationally, by far the most expensive of all considered here, due to the intensive resampling required, makes it very difficult to recommend its use. The FSA estimator has a very good performance for the unit sphere data, even for the larger

value of m, while its performance deteriorates with dimension for the $\mathcal{P}_m$ data, reaching a standard error of about 4 units for the $\mathcal{P}_9$ data for $n = 250$, although we observe, as well, that this performance tends to improve with sample size. The FSA estimator is not as good on the 2-dimensional distributions considered, presenting an error of about one unit on the Swiss Roll data for all sample sizes. For the distributions considered, the LB estimator has, in general, a similar performance to FSA, being superior for the S^3 and Möbius band data, for which LB is practically perfect, and being inferior to FSA for the S^9 data. The estimator of Sricharan et al. presents a behavior similar to those of the FSA and LB statistics, with a tendency to perform slightly below (with a larger MSE) the LB statistic across the table. The reach statistic, for the 2-dimensional manifolds, performs fairly well, better than FSA and SRH, and similar to LB. With respect to S^m and $\mathcal{P}_m$, $\bar{r}_{j,k}$ does well for $m = 3$, but its performance deteriorates with growing m, reaching a standard error of about 2.5 units for S^9 and of about 5 units for $\mathcal{P}_9$ for sample size $n = 250$, but we observe that these errors tend to improve rapidly with sample size in both cases (S^9 and $\mathcal{P}_9$). For the 2-dimensional manifolds, the M_n estimator displays a good behavior, being better than FSA and SRH and similar in mean squared error to LB and $\bar{r}_{j,k}$. For the unit sphere data, M_n performance is inferior to that of all the nearest neighbor methods (except CGH), while for the $\mathcal{P}_m$ data, M_n displays consistently, the smallest errors of all the methods considered, suggesting that it might be a very good method in the case of non-constant curvature manifolds. The Mutual Neighbor method, for the 2-dimensional manifolds, displays a better behavior than the FSA statistic, but is inferior to the other nearest neighbor methods, although U_n^k's performance improves rapidly with sample size in this case. For S^m and $\mathcal{P}_m$, U_n^k's offers a behavior similar to that of M_n, being inferior to the nearest neighbor methods (except CGH) for the unit sphere data, while being clearly better than those methods for the paraboloid data.

A conclusion that can be extracted from this comparison is that no clear winner emerges. It would appear that the graph theoretic methods could have a certain edge in the case of manifolds of non-constant curvature.

4. Limit theory for dimension estimators

In this section, we establish weak laws of large numbers, variance asymptotics, and central limit theorems for the statistics M_n and U_n^k, $k \in \mathbb{N}$, as $n \to \infty$. The corresponding theory for $\bar{r}_{j,k}$ follows the pattern of that for U_n^k, as explained in the remarks after Theorem 1. To obtain the limit theory for U_n^k, we shall slightly extend some of the general results of Penrose and Yukich [22].

4.1. Statement of results

For all $x \in \mathbb{R}^d$, $k = 1, 2, \ldots$ and all locally finite point sets $\mathcal{X} \subset \mathbb{R}^d$, we put as in Section 1

$$c_k(x, \mathcal{X}) := \operatorname{card}\{y \in \mathcal{X} : x \text{ is one of the } k \text{ nearest neighbors of } y \text{ in } \mathcal{X}\}.$$

Let

$$\zeta_k(x, \mathcal{X}) := \binom{c_k(x, \mathcal{X})}{2}. \tag{14}$$

Recalling that $\mathcal{X}_n := \{X_i\}_{i=1}^n$, we have $U_n^k = n^{-1}\sum_{i=1}^n \zeta_k(X_i, \mathcal{X}_n)$.

As in [22], let $\mathbb{M} := \mathbb{M}(m, d)$ be the class of all m-dimensional C^1 submanifolds of $\mathbb{R}^d$ which are also closed subsets of $\mathbb{R}^d$. Given $\mathcal{M} \in \mathbb{M}$, let $\mathbb{P}_c(\mathcal{M})$ denote the class of probability density functions κ on $\mathcal{M}$ whose support $\mathcal{K}(\kappa)$ is a compact C^1 submanifold-with-boundary of $\mathcal{M}$, and which are bounded away from zero and infinity on their support.

Let $\mathcal{H}$ denote a homogeneous rate one Poisson point process on $\mathbb{R}^m$. Say that a functional ξ defined on pairs $(x, \mathcal{X})$, with $x \in \mathbb{R}^m$ and $\mathcal{X}$ locally finite in $\mathbb{R}^m$, is translation invariant if $\xi(x, \mathcal{X}) = \xi(x + y, \mathcal{X} + y)$ for all $y \in \mathbb{R}^m$. Letting $\mathbf{0}$ denote a point at the origin of $\mathbb{R}^m$, we put, for all $m \in \mathbb{N}$,

$$V^\xi(m) := \mathbb{E}[\xi(\mathbf{0}, \mathcal{H})^2] + \int_{\mathbb{R}^m} \{\mathbb{E}\xi(\mathbf{0}, \mathcal{H} \cup \{z\})\xi(z, \mathcal{H} \cup \{\mathbf{0}\}) - (\mathbb{E}\xi(\mathbf{0}, \mathcal{H}))^2\} dz \tag{15}$$

and

$$\Delta^\xi(m) := \mathbb{E}\xi(\mathbf{0}, \mathcal{H}) + \int_{\mathbb{R}^m} \mathbb{E}[\xi(\mathbf{0}, \mathcal{H} \cup \{z\}) - \xi(\mathbf{0}, \mathcal{H})] dz. \tag{16}$$

For all $\kappa \in \mathbb{P}_c(\mathcal{M})$ and ξ we define

$$\sigma^2(\xi, \kappa) := \int_{\mathcal{M}} V^\xi(\kappa(x))\kappa(x) dx - \left(\int_{\mathcal{M}} \Delta^\xi(\kappa(x))\kappa(x) dx\right)^2, \tag{17}$$

provided that both integrals in (17) exist and are finite. Put

$$\sigma^2(\xi) := V^\xi(m) - (\Delta^\xi(m))^2.$$

The scalar $V^{\xi}(m)$ may be interpreted as the mean pair correlation function for the functional ξ on homogeneous Poisson points $\mathcal{H}$ whereas we may view $\Delta^{\xi}(m)$ as an expected "add-one cost".

We have the following limit theory for the dimension estimator U_n^k; $N(0, \sigma^2)$ denotes a mean zero normal random variable with variance σ^2.

Theorem 1. *Let $\mathcal{M} \in \mathbb{M}$ and let $\kappa \in \mathbb{P}_c(\mathcal{M})$. If X_i, $i \geq 1$, are i.i.d. with density κ, then, for all $k \in \mathbb{N}$, we have in L^2 and almost surely,*

$$\lim_{n\to\infty} U_n^k = \mathbb{E}\zeta_k(\mathbf{0}, \mathcal{H}). \tag{18}$$

If, additionally, κ is a.e. continuous, then

$$\lim_{n\to\infty} n \operatorname{Var}[U_n^k] := \sigma^2(\zeta_k) := V^{\zeta_k}(m) - (\Delta^{\zeta_k}(m))^2, \tag{19}$$

and as $n \to \infty$,

$$n^{1/2}(U_n^k - \mathbb{E}U_n^k) \xrightarrow{\mathcal{D}} N(0, \sigma^2(\zeta_k)). \tag{20}$$

Remarks. (i) The limits in (18) and (19) are independent of the density κ of the underlying point set. (ii) The proof of Theorem 1 shows that the statistic $\bar{r}_{j,k}$ also satisfies the limit theory of Theorem 1, with ζ_k replaced by the functional

$$r_{j,k}(x, \mathcal{X}) := \operatorname{card}\{y \in \mathcal{X} : \ y \neq x, \ y \ \text{ is reached in } l \text{ steps from } x;\ l \leq j\}.$$

(iii) Positivity of $\sigma^2(\zeta_k)$ and $\sigma^2(r_{j,k})$ follows as in the proof of the last part of Theorem 2.1 of [19].

For all $x \in \mathbb{R}^d$ and all locally finite point sets $\mathcal{X} \subset \mathbb{R}^d$, define the functional

$$\varphi(x, \mathcal{X}) := \left(\deg_{\mathrm{MST}(\mathcal{X})}(x)\right)^2, \tag{21}$$

where we recall that $\deg_{\mathrm{MST}(\mathcal{X})}(x)$ denotes the degree of the node x in the graph of the minimal spanning tree on $\mathcal{X}$. We have $M_n := M(\mathcal{X}_n) := n^{-1}\sum_{i=1}^n \varphi(X_i, \mathcal{X}_n)$.

Theorem 2. *Assume that X_i, $i \geq 1$, have density κ on $\mathbb{R}^m$. Then in L^2*

$$\lim_{n\to\infty} M_n = \mathbb{E}\varphi(\mathbf{0}, \mathcal{H}). \tag{22}$$

If the density κ is uniform on $[-1, 1]^m$ then there is a $\sigma^2(m)$, such that

$$\lim_{n\to\infty} n \operatorname{Var}[M_n] = \sigma^2(m) \tag{23}$$

and, as $n \to \infty$,

$$n^{1/2}(M_n - \mathbb{E}M_n) \xrightarrow{\mathcal{D}} N(0, \sigma^2(m)). \tag{24}$$

Remark. If $\mathcal{M} \in \mathbb{M}$ and $\kappa \in \mathbb{P}_c(\mathcal{M})$, then no limiting theory is currently available for M_n, but we conjecture that the limit theory of Theorem 1 applies to M_n, with ζ_k replaced by φ. The method of proof shows that Theorem 2 also holds if the data $\mathcal{X}_n$ belongs to any affine subspace of $\mathbb{R}^d$.

4.2. A general result

We first slightly extend some general results of [22]. This goes as follows. For all $r \in (0, \infty)$ and $x \in \mathbb{R}^d$, let $B_r(x)$ be the Euclidean ball of radius r centered at x. Let $F \subset \mathbb{R}^l$, $l \in \mathbb{N}$, be a locally finite set. For all $x \in \mathbb{R}^l$, $j \in \mathbb{N}$, we let $\mathcal{N}_j(x, F)$ be the set of points in F having x as one of their j nearest neighbors in $\{x\} \cup F$. Let $n_j(x, F)$ be the maximum of the distances between x and the elements of $\mathcal{N}_j(x, F)$, i.e.,

$$n_j(x, F) := \max\{|x - w| : w \in \mathcal{N}_j(x, F)\}.$$

If $\mathcal{N}_j(x, F) = \emptyset$, then put $n_j(x, F) := \operatorname{diam}(F)$.

Given $k \in \mathbb{N}$, let $\Xi(k)$ be the class of translation and rotation invariant functionals ξ such that (i) for all $x, \mathcal{X}$ with $\operatorname{card}(\mathcal{X} \setminus \{x\}) \geq k$, we have

$$\xi(x, \mathcal{X}) = \xi(x, \mathcal{X} \cap B_{n_k(x,\mathcal{X})}(x))$$

and (ii) for all n, Lebesgue-almost every $(x_1, \ldots, x_n) \in (\mathbb{R}^m)^n$ (with $\mathbb{R}^m$ embedded in $\mathbb{R}^d$) is at a continuity point of the mapping from $(\mathbb{R}^d)^n \to \mathbb{R}$, given by

$$(x_1, \ldots, x_n) \mapsto \xi(\mathbf{0}, \{x_1, \ldots, x_n\}).$$

The functional ζ_k, defined at (14), belongs to the class $\Xi(k)$. Indeed, the continuity criteria (ii) hold when the points $(x_1, \ldots, x_n)$ have distinct interpoint distances.

Next, for $x \in \mathbb{R}^l$ and $j \in \{0, 1, 2, \ldots\}$, let $N_j(x, F)$ be the Euclidean distance between x and its jth nearest neighbor in $F \setminus \{x\}$, i.e.

$$N_j(x, F) := \inf\{r \geq 0 : \mathrm{card}(F \cap B_r(x) \setminus \{x\}) \geq j\} \tag{25}$$

with the infimum of the empty set taken to be $+\infty$. In particular, $N_0(x, F) = 0$. Given $k \in \mathbb{Z}$, let $\Xi_0(k)$ be the class of translation and rotation invariant functionals, ξ, such that (i) for all x, $\mathcal{X}$ with $\mathrm{card}(\mathcal{X} \setminus \{x\}) \geq k$, we have

$$\xi(x, \mathcal{X}) = \xi(x, \mathcal{X} \cap B_{N_k(x,\mathcal{X})}(x))$$

and (ii) for all n, Lebesgue-almost every $(x_1, \ldots, x_n) \in (\mathbb{R}^m)^n$ (with $\mathbb{R}^m$ embedded in $\mathbb{R}^d$) is at a continuity point of the mapping from $(\mathbb{R}^d)^n \to \mathbb{R}$, given by

$$(x_1, \ldots, x_n) \mapsto \xi(\mathbf{0}, \{x_1, \ldots, x_n\}).$$

When $\xi \in \Xi_0(k)$, $\mathcal{M} \in \mathbb{M}$, and $\kappa \in \mathbb{P}_c(\mathcal{M})$, then under moment conditions on ξ, [22] establishes weak laws of large numbers, variance asymptotics, and central limit theorems for

$$H_n^{\xi}(\mathcal{X}_n) := \sum_{i=1}^{n} \xi_n(X_i, \mathcal{X}_n),$$

where $\xi_n(x, \mathcal{X}) := \xi(n^{1/d}x, n^{1/d}\mathcal{X})$.

Here we shall show that if $\xi \in \Xi(k)$, then a similar limit theory holds for $H_n^{\xi}(\mathcal{X}_n)$. Given $i = 1, 2, 3$, recall that $\mathcal{S}_i$ is the collection of all subsets of $\mathcal{K}(\kappa)$ of cardinality at most i, including the empty set. Consider the following moment conditions on ξ:

$$\sup_n \mathbb{E}|\xi_n(X_1, \mathcal{X}_n)|^p < \infty, \tag{26}$$

$$\sup_{n\geq 1, x\in\mathcal{K}(\kappa), \mathcal{A}\in\mathcal{S}_3} \ \sup_{(n/2)\leq \ell \leq (3n/2)} \mathbb{E}|\xi_n(x, \mathcal{X}_\ell \cup \mathcal{A})|^p < \infty \tag{27}$$

and

$$\sup_{\lambda\geq 1, x\in\mathcal{K}(\kappa), \mathcal{A}\in\mathcal{S}_1} \mathbb{E}|\xi_\lambda(x, \mathcal{P}_\lambda \cup \mathcal{A})|^p < \infty. \tag{28}$$

Theorem 3 (*Limit Theory for $H_n^{\xi}(\mathcal{X}_n)$, $\xi \in \Xi(k)$*)**.** *Let $\mathcal{M} \in \mathbb{M}$, $\kappa \in \mathbb{P}_c(\mathcal{M})$, $k \in \mathbb{N}$, and put $q = 1$ or $q = 2$. Let $\xi \in \Xi(k)$ and suppose there exists $p > q$ such that (26) holds. Then as $n \to \infty$ we have L^q convergence*

$$n^{-1}H_n^{\xi}(\mathcal{X}_n) \to \int_{\mathcal{M}} \mathbb{E}[\xi(\mathbf{0}, \mathcal{H}_{\kappa(x)})]\kappa(x)dx. \tag{29}$$

If also (27) holds for some $p > 5$, then (29) holds a.s. If, additionally, κ is a.e. continuous, and if ξ satisfies (27) and (28) for some $p > 2$, then $\sigma^2(\xi, \kappa) < \infty$ and

$$\lim_{n\to\infty} n^{-1}\,\mathrm{Var}[H^{\xi}(\mathcal{X}_n)] = \sigma^2(\xi, \kappa)$$

and as $n \to \infty$,

$$n^{-1/2}(H_n^{\xi}(\mathcal{X}_n) - \mathbb{E}H_n^{\xi}(\mathcal{X}_n)) \xrightarrow{\mathcal{D}} N(0, \sigma^2(\xi, \kappa)).$$

Remark. We say that ξ is homogeneous of order zero if, for all scalars $a > 0$, we have $\xi(x, \mathcal{X}) = \xi(ax, a\mathcal{X})$. In this case the constants in Theorem 3 simplify, since

$$\int_{\mathcal{M}} \mathbb{E}[\xi(\mathbf{0}, \mathcal{H}_{\kappa(x)})]\kappa(x)dx = \mathbb{E}[\xi(\mathbf{0}, \mathcal{H}_1)] \tag{30}$$

and $\sigma^2(\xi, \kappa) := \sigma^2(\xi) := V^{\xi}(m) - (\Delta^{\xi}(m))^2$. The functionals ζ_k and φ are both homogeneous of order zero.

Proof of Theorem 3. We first recall a definition (see Definition 5.1 in [22]). Recall that if ξ is *continuous* if for any linear $F : \mathbb{R}^m \to \mathbb{R}^d$ of full rank, for almost all $z \in \mathbb{R}^m$ both $F(\mathcal{H})$ and $F(\mathcal{H} \cup \{z\})$ lie a.s. at continuity points of $\xi(\mathbf{0}, \cdot)$ with respect to the topology $\mathcal{T}_d$ in [22]. We first give some definitions, following closely [22]. Assume $\mathcal{M} \in \mathbb{M}$ and $\kappa \in \mathbb{P}(\mathcal{M})$ are given, and recall $\mathcal{K} := \mathcal{K}(\kappa)$. Suppose $k \in \mathbb{N}$ is given, along with the density κ. For $x \in \mathcal{K}$ and locally finite $\mathcal{X} \subset \mathcal{K}$ define

$$R_\lambda(x, \mathcal{X}) := \begin{cases} n_k(\lambda^{1/m}x, \lambda^{1/m}\mathcal{X}) & \text{if } \mathrm{card}(\mathcal{X} \setminus \{x\}) \geq k \\ \lambda^{1/m}\,\mathrm{diam}(\mathcal{K}) & \text{otherwise.} \end{cases}$$

Then $R := R_\lambda(x, \mathcal{X})$ is a *radius of stabilization* for any $\xi \in \Xi(k)$, in the following sense: for all finite $\mathcal{A} \subset (\mathcal{K} \setminus B_{\lambda^{-1/m}R}(x))$, we have

$$\xi_\lambda\left(x, (\mathcal{X} \cap B_{\lambda^{-1/m}R}(x)) \cup \mathcal{A}\right) = \xi_\lambda\left(x, \mathcal{X} \cap B_{\lambda^{-1/m}R}(x)\right). \tag{31}$$

Recall that $\mathcal{X}_n := \{X_i\}_{i=1}^n$ and for $\lambda \in [1, \infty)$, let $\mathcal{P}_\lambda$ denote the Poisson point process on $\mathcal{M}$ having intensity density $\lambda\kappa(\cdot)$, that is $\mathbb{E}\mathcal{P}_\lambda(dx) = \lambda\kappa(x)dx$. Given $\epsilon > 0$ and $t > 0$, we define the tail probabilities for R_λ denoted by $\tau(t)$ and $\tau_\epsilon(t)$, for Poisson input $\mathcal{P}_\lambda$ and binomial input $\mathcal{X}_n$, respectively, as follows:

$$\tau(t) := \sup_{\lambda \geq 1} \operatorname*{ess\,sup}_{x \in \mathcal{K}} P[R_\lambda(x, \mathcal{P}_\lambda) > t]$$

$$\tau_\epsilon(t) := \sup_{\lambda \geq 1, n \in \mathbb{N} \cap ((1-\epsilon)\lambda, (1+\epsilon)\lambda),\ \mathcal{A} \in \mathcal{S}_2} \operatorname*{ess\,sup}_{x \in \mathcal{K}} P[R_\lambda(x, \mathcal{X}_n \cup \mathcal{A}) > t]$$

where the ess sup denotes essential supremum with respect to the measure $\kappa(x)dx$.

Definition 1. Given k, we say that all $\xi \in \Xi(k)$ are *exponentially stabilizing* for κ if $\limsup_{t\to\infty} t^{-1} \log \tau(t) < 0$. We say that all $\xi \in \Xi(k)$ are *binomially exponentially stabilizing* for κ if there exists $\epsilon > 0$ such that $\limsup_{t\to\infty} t^{-1} \log \tau_\epsilon(t) < 0$.

To prove Theorem 3, by the remark at the end of Section 6 of [22], it suffices to show that $\xi \in \Xi(k)$ is continuous, exponentially stabilizing, and binomially exponentially stabilizing. We prove continuity by following verbatim the proof of Lemma 6.1 in [22], replacing $N_k(\mathbf{0}, F(\mathcal{H}))$ in that proof by $n_k(\mathbf{0}, F(\mathcal{H}))$.

We now show that ξ is exponentially stabilizing and binomially exponentially stabilizing in the sense of Definition 1. Let $C_1^0, \ldots, C_J^0$ be a finite collection of open cones in $\mathbb{R}^d$, each with vertex at the origin and angular radius $\pi/6$, so that $\mathbb{R}^d \setminus \{\mathbf{0}\} = \cup_{j=1}^J C_j^0$. For $1 \leq j \leq J$, let C_j be the translate of C_j^0 with its vertex at x. Put $K_j := C_j \cap \mathcal{M}$.

Elementary geometry shows that if $\mathrm{card}(K_j \cap B_t(x) \cap \mathcal{P}_\lambda) \geq k + 1$ for all $1 \leq j \leq J$, then points outside $B_t(x)$ will not affect the value of $\xi(x, \mathcal{P}_\lambda)$. Let $T_\lambda(x, \mathcal{P}_\lambda)$ be the minimum ρ such that each $K_j \cap B_{\lambda^{-1/m}\rho}(x)$ contains at least $k + 1$ points from $\mathcal{P}_\lambda$, that is

$$T_\lambda(x, \mathcal{P}_\lambda) := \begin{cases} \inf\{\rho > 0 :\ \mathrm{card}(K_j \cap B_{\lambda^{-1/m}\rho}(x) \cap \mathcal{P}_\lambda) \geq k + 1\} & \text{if } \mathrm{card}(K_j \cap \mathcal{P}_\lambda) > k \\ \lambda^{1/m}\,\mathrm{diam}(\mathcal{K}) & \text{otherwise.} \end{cases}$$

Then $R_\lambda(x, \mathcal{P}_\lambda) \leq T_\lambda(x, \mathcal{P}_\lambda)$ and so $T_\lambda(x, \mathcal{P}_\lambda)$ is a radius of stabilization for ξ. Also, $T_\lambda(x, \mathcal{P}_\lambda)$ exceeds t only when there is a j, $1 \leq j \leq J$, such that

$$\mathrm{card}(K_j \cap B_{\lambda^{-1/m}t}(x) \cap \mathcal{P}_\lambda) \leq k.$$

The number of points in $K_j \cap B_{\lambda^{-1/m}t}(x) \cap \mathcal{P}_\lambda$ is Poisson distributed with parameter $b(t) := b(t, x, \lambda)$ equal to the $\lambda\kappa$ measure of $B_{\lambda^{-1/m}t}(x) \cap \mathcal{M}$. By Lemma 4.3 in [22], there is a constant $C_1 \in (0, \infty)$ such that, uniformly in $\lambda \in [1, \infty)$, $x \in \mathcal{K}$, and $t \in (0, \lambda^{1/m}\,\mathrm{diam}(\mathcal{K}))$, we have $b(t) \geq C_1^{-1}t^m$.

By bounds for the Poisson distribution (e.g. Lemma 1.2 of [18]), there is a constant $C_2 \in (0, \infty)$ such that for $t \in (0, \lambda^{1/m}\,\mathrm{diam}(\mathcal{K}))$ we have uniformly in j that

$$P[\mathrm{card}(K_j \cap B_{\lambda^{-1/m}t}(x) \cap \mathcal{P}_\lambda) \leq k] = P[\mathrm{Pois}(b(t)) \leq k] \leq kC_2 \exp(-C_2^{-1}t^m).$$

Thus for $t \in (0, \lambda^{1/m}\,\mathrm{diam}(\mathcal{K}))$ we have

$$P[T_\lambda(x, \mathcal{P}_\lambda) > t] \leq C_3(k) \exp(-C_2^{-1}t^m).$$

For $t \in (\lambda^{1/m}\,\mathrm{diam}(\mathcal{K}), \infty)$ this also holds since $P[T_\lambda(x, \mathcal{P}_\lambda) > t] = 0$ in this case. This shows that ξ stabilizes exponentially fast on Poisson input. Modifications of these arguments give that ξ stabilizes exponentially fast on binomial input. This completes the proof of Theorem 3. □

4.3. *Proof of Theorem 1*

We shall deduce Theorem 1 from Theorem 3. Since $U_n^k = n^{-1} \sum_{i=1}^n \zeta_k(X_i, \{X_i\}_{i=1}^n)$, it suffices to show that ζ_k satisfies the conditions of Theorem 3.

Note that $\zeta_k \in \Xi(k)$ and so it only remains to show that ζ_k satisfies the moment conditions of Theorem 3. However this follows since ζ_k is deterministically bounded. Indeed, a given point in $\mathbb{R}^d$ is the nearest neighbor of at most a finite number of other points in $\mathbb{R}^d$ (see Lemma 8.4 of [31]). For all $k \in \mathbb{N}$ there is thus a constant C_k such that, for all $x, \mathcal{X}$ we have $c_k(x, \mathcal{X}) \le C_k$, that is $\zeta_k(\cdot, \cdot) \le C_k^2/2$. Applying Theorem 3 and recalling that ζ_k is homogeneous of order zero, we obtain Theorem 1 as desired. □

4.4. Proof of Theorem 2

The limit (22) is an immediate consequence of the boundedness of φ (as noted on pp. 811–812 of Steele, Shepp and Eddy [29]), the fact that φ is stabilizing as shown in Lemma 2.1 of [20], the general weak law of large numbers for sums of stabilizing functionals, as given by Theorem 2.1 of [20], and the fact that $\varphi(x, \mathcal{X})$ is homogeneous of order zero, and thus (30) applies.

To prove (23) and (24), we proceed as follows. The proof of (24) follows from Lee [16] or [19]. We provide the details as follows. Consider the functional

$$H^{\varphi}(\mathcal{X}) := \sum_{x \in \mathcal{X}} \varphi(x, \mathcal{X}).$$

It will suffice to show that H^{φ} satisfies the conditions of Corollary 2.1 of [19]. Since φ is bounded it follows that $H^{\varphi}(\mathcal{X}) \le C\,\mathrm{card}(\mathcal{X})$ and, therefore, H^{φ} is polynomially bounded, that is to say there exists a constant $\beta \in (0, \infty)$ such that, for all finite sets $\mathcal{X} \subset \mathbb{R}^d$, we have

$$H^{\varphi}(\mathcal{X}) \le \beta(\mathrm{diam}(\mathcal{X}) + \mathrm{card}(\mathcal{X}))^{\beta}.$$

Put $\Delta(\mathcal{X}) := H^{\varphi}(\mathcal{X} \cup \{\mathbf{0}\}) - H^{\varphi}(\mathcal{X})$. If $\mathcal{H}_{\lambda}$ denotes a homogeneous Poisson point process of intensity λ on $\mathbb{R}^d$, then there exist a.s. finite random variables S and $\Delta(\infty)$ such that, with probability one,

$$\Delta(\mathcal{H}_{\lambda} \cap B_S(\mathbf{0}) \cup \mathcal{A}) = \Delta(\infty)$$

for all finite $\mathcal{A} \subset \mathbb{R}^d \setminus B_S(\mathbf{0})$. This condition, known as *strong stabilization* of H^{φ} (cf. Definition 2.1 of [19]), follows from Kesten and Lee [15] and Lee [16]. Actually Kesten and Lee [15] show that the above condition holds if $H_{\mathrm{MST}}(\mathcal{X})$ is the total edge length of the minimal spanning tree on $\mathcal{X}$, but the random variable S which works for H_{MST} will also work for H^{φ}.

Finally, if $\mathcal{X}_m$ is the point process consisting of m i.i.d. uniform random variables on $[-1, 1]^d$, then straightforward modifications of Kesten and Lee [15] as well as Lee [16] show that

$$\sup_{n} \sup_{m \in [\lambda/2, 3\lambda/2]} \mathbb{E}[\Delta(\mathcal{X}_m)^4] < \infty,$$

and so H^{φ} satisfies the uniform bounded moment condition of [19]. Thus, H^{φ} satisfies all of the conditions of Corollary 2.1 in [19] and so (23) and (24) follow. □

4.5. Consistency of a dimension identification procedure

To prove that the procedures for intrinsic dimension identification based on $\bar{r}_{j,k}$ and U_n^k are consistent, we assume that the limiting constants in the corresponding LLNs are different, for different dimensions in the range considered. Actually, our simulations support the hypothesis that these constants are increasing with dimension. As in Section 2, let $S_n := S_n(\mathcal{X}_n)$ represent one of our statistics, $\bar{r}_{j,k}$ or U_n^k, or another graph theoretic statistic satisfying the same conditions. Let $\mathcal{X}_n$ denote an i.i.d. data set obtained from a density κ with support on an m-dimensional compact submanifold-with-boundary of $\mathbb{R}^d$ and assume that κ is bounded away from zero and infinity on its support. Assume that, for this kind of data S_n satisfies a SLLN: $S_n \to \mu(m)$ in L^2, and almost surely, where the limit depends only on m, and the $\mu(m)$ values are different for the different $m \in F$, F being a finite set. Assume, as well, that conditions (9) and (10) of Section 2 hold. Suppose that the parameter estimation for the approximate Bayesian dimension identification procedure based on S_n, is made as described in Section 2, using L samples of size n from the Uniform distribution on the unit cube for each $j \in F$ and that the value of the graph theoretic statistic, $S_{n'}$ is computed on a new data set $\mathcal{X}_{n'}$ that satisfies the conditions for manifold data given above. Then, we have the following.

Theorem 4. *Let m^* and $\tilde{m}$ denote, respectively, the actual intrinsic dimension of $\mathcal{X}_{n'}$ and its estimator, as given by* (12)*. Under the conditions stated above, $\tilde{m} \longrightarrow m^*$, almost surely, as $L, n, n' \to \infty$.*

Proof of Theorem 4. In view of formula (12), it suffices to show that, for each $j \in F, j \neq m^*$,

$$\frac{\tilde{f}_j(S_{n'})}{\tilde{f}_{m^*}(S_{n'})} \to 0, \quad \text{almost surely, as } n, n',\ L \to \infty.$$

Now, for each dimension $j \in F$, the estimator, $\tilde{\mu}(j)$, is converging a.s. to $\mathbb{E}(S_n)$ for j-dimensional unit cube data, as $L \to \infty$. Then, using the L^2 convergence of S_n to $\mu(j)$, we have that $\tilde{\mu}(j) \to \mu(j)$, a.s., as $L, n \to \infty$. On the other hand, both $\tilde{\sigma}^2(j)$ and

$\tilde{\sigma}^2(m^*)$ are a.s. consistent for their estimated parameters, as $n, L \to \infty$ and then, the ratio $\tilde{\sigma}^2(j)/\tilde{\sigma}^2(m^*)$ will almost surely, be bounded away from zero and ∞ for large L and n. Applying the definition of $\tilde{f}_j$, we have

$$\frac{\tilde{f}_j(S_{n'})}{\tilde{f}_{m^*}(S_{n'})} = \frac{\tilde{\sigma}(m^*)}{\tilde{\sigma}(j)} \exp\left(\frac{n'}{2}\left[\frac{(S_{n'} - \tilde{\mu}(m^*))^2}{\tilde{\sigma}^2(m^*)} - \frac{(S_{n'} - \tilde{\mu}(j))^2}{\tilde{\sigma}^2(j)}\right]\right). \tag{32}$$

Using both the L^2 and almost sure convergence to the mean, we have that $S_{n'} - \tilde{\mu}(m^*)$ goes to zero, almost surely, as n, n', L go to ∞, while $S_{n'} - \tilde{\mu}(j)$ converges, almost surely, to a non-zero limit. It follows that the exponent in (32) diverges to $-\infty$, almost surely, establishing the result. □

Remarks. The hypotheses of Theorem 4 are met by both $\bar{r}_{j,k}$ and U_n^k, and the result is valid for the procedures based on these statistics. The fact that n and n' do not need to be equal for the consistency of the procedure, partially answers the comment of Levina and Bickel [17] regarding the need of a different calibration for every sample size. It is clear, from the proof just given, that the rounding in formula (12) does not affect the consistency of the estimator. That is, Theorem 4 holds whether or not we use rounding in (12).

Acknowledgment

The research of the third author was supported in part by NSF grant DMS-1106619.

References

[1] D. Aldous, J. Shun, Connected spatial networks over random points and a route-length statistic, Statistical Science 25 (2010) 275–288.
[2] M. Belkin, P. Niyogi, Semi-supervised learning on Riemannian manifolds, Machine Learning. Special Issue on Clustering 56 (2004) 209–239.
[3] P.J. Bickel, D. Yan, Sparsity and the possibility of inference, Sankhya A 70 (1) (2008) 1–24.
[4] M.R. Brito, A.J. Quiroz, J.E. Yukich, Graph theoretic procedures for dimension identification, Journal of Multivariate Analysis 81 (2002) 67–84.
[5] J.A. Costa, A. Girotra, A.O. Hero, Estimating local intrinsic dimension with k-nearest neighbor graphs, in: IEEE/SP 13th Workshop on Statistical Signal Processing, IEEE Conference Publication, 2005, pp. 417–422.
[6] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., John Wiley and Sons, New York, 2001.
[7] A. Farahmand, C. Szepesvári, J.-Y. Audibert, Manifold-adaptive dimension estimation, in: Z. Ghahramani (Ed.), Proceedings of the 24th International Conference on Machine Learning, ACM, New York, 2007, pp. 265–272.
[8] J.H. Friedman, L.C. Rafsky, Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests, Annals of Statistics 7 (4) (1979) 697–717.
[9] J.H. Friedman, L.C. Rafsky, Graphics for the multivariate two-sample problem, Journal of the American Statistical Association 76 (374) (1981) 277–295.
[10] J.H. Friedman, L.C. Rafsky, Graph theoretic measures of multivariate association and prediction, Annals of Statistics 11 (2) (1983) 377–391.
[11] P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors, Physica D 9 (1983) 189–208.
[12] F. Harary, Graph Theory, Addison-Wesley, Reading, Mass, 1969.
[13] M.E. Johnson, Multivariate Statistical Simulation, John Wiley and Sons, New York, 1987.
[14] B. Kegl, Intrinsic dimension estimation using packing numbers, in: S. Becker, S. Thrun, K. Obermayer. (Eds.), Advances in Neural Information Processing Systems, Volume 15, MIT Press, Cambridge, Massachusetts, 2003.
[15] H. Kesten, S. Lee, The central limit theorem for weighted minimal spanning trees on random points, The Annals of Applied Probability 6 (1996) 495–527.
[16] S. Lee, The central limit theorem for Euclidean minimal spanning trees I, The Annals of Applied Probability 7 (1997) 996–1020.
[17] E. Levina, P.J. Bickel, Maximum likelihood estimation of intrinsic dimension, in: L.K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems, Volume 17, 2005.
[18] M.D. Penrose, Random Geometric Graphs, Clarendon Press, Oxford, 2003.
[19] M.D. Penrose, J.E. Yukich, Central limit theorems for some graphs in computational geometry, The Annals of Applied Probability 11 (2001) 1005–1041.
[20] M.D. Penrose, J.E. Yukich, Weak laws of large numbers in geometric probability, The Annals of Applied Probability 13 (2003) 277–303.
[21] M.D. Penrose, J.E. Yukich, Normal approximation in geometric probability, in: A.D. Barbour, L.H.Y. Chen (Eds.), Stein Methods and Applications, in: Lecture Notes Series, vol. 5, Institute for Mathematical Sciences, National University of Singapore, 2005, pp. 37–58.
[22] M.D. Penrose, J.E. Yukich, Limit theory for point processes in manifolds, The Annals of Applied Probability (2012). arXiv:1104.0914 (in press).
[23] K.W. Pettis, T.A. Bailey, A.K. Jain, R.C. Dubes, An intrinsic dimensionality estimator from near-neighbor information, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (1979) 25–37.
[24] A.J. Quiroz, Graph-theoretical methods, in: S. Kotz, C.B. Read, N. Balakrishnan, B. Vidakovic (Eds.), Encyclopedia of Statistical Sciences, Vol. 5, second ed., Wiley and Sons, New York, 2006, pp. 2910–2916.
[25] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[26] M.F. Schilling, Mutual and shared neighbor probabilities: finite and infinite dimensional results, Advances in Applied Probability 18 (1986) 388–405.
[27] V. Sindhwani, M. Belkin, P. Nigoyi, The geometric basis of semi-supervised learning, in: O. Chapelle, B. Schölkopf, A. Zien (Eds.), Semi-supervised Learning, MIT Press, Cambridge, Massachusetts, 2006.
[28] K. Sricharan, R. Raich, A.O. Hero, Optimized intrinsic dimension estimation using nearest neighbor graphs, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE Conference Publication, 2010, pp. 5418–5421.
[29] J.M. Steele, L.A. Shepp, W.F. Eddy, On the number of leaves of a Euclidean minimal spanning tree, Journal of Applied Probability 24 (1987) 809–826.
[30] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[31] J.E. Yukich, Probability Theory of Classical Euclidean Optimization Problems, in: Lecture Notes in Mathematics, vol. 1675, Springer, New York, 1998.