Learning Intrinsic Dimension and Entropy of High-Dimensional Shape Spaces

Jose A. Costa, Alfred O. Hero, III

Abstract

Given a finite set of random samples from a smooth Riemannian manifold embedded in $\mathbb{R}^d$ two important questions are: what is the intrinsic dimension of the manifold and what is the entropy of the underlying sampling distribution on the manifold? These questions naturally arise in the study of shape spaces generated by images or signals for the purposes of shape classification, shape compression, and shape reconstruction. This chapter is concerned with two simple estimators of dimension and entropy based on the lengths of the geodesic minimal spanning tree (GMST) and the k-nearest neighbor graph (k-NNG). We provide proofs of strong consistency of these estimators under weak assumptions of compactness of the manifold and boundedness of the Lebesgue sampling density supported on the manifold. We illustrate these estimators on the MNIST database of handwritten digits.

Index Terms

Intrinsic dimension, entropy, manifold learning, Riemannian manifold, nearest neighbor graph, minimal spanning tree, geodesics.

I. Introduction

Continuing technological advances in both sensing and media storage capabilities are enabling the development of systems that generate massive amounts of new types of data and information. However, the high dimensional nature of data sets produced by today's medical information systems or video surveillance applications, for example, poses challenging problems to the application of current signal and image processing tools. It is well known that both computational complexity and statistical performance

This research was partially supported by NSF contract CCR-0325571 and NIH contract 1P01 CA87634.

J. A. Costa and A. O. Hero, III are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122 USA (emails: {jcosta,hero}@umich.edu).

of most algorithms quickly degrades as dimension increases. This phenomenon, usually known as the *curse of dimensionality*, makes it impracticable to process such high dimensional data sets. Nevertheless, high dimensional data often contain fundamental features that are concentrated on lower dimensional subsets – curves, surfaces or, more generally, lower-dimensional manifolds – thus permitting substantial dimension reduction with little or no loss of content information.

The study of the geometry of smooth two or three dimensional objects is one of the areas where high dimensional manifolds naturally arise. The shape of a single object is characterized by mathematical shape invariants such as dimension, curvature and geodesic distance. A more challenging problem is the characterization of a smooth class of objects, e.g., imagine the set of all possible handwritten digits or all possible human faces under various poses and illumination. Having an accurate estimate of the dimension of such shape spaces is of great importance for shape modeling, compression, and classification as it provides an indication of the number of model parameters required for indexing or reconstruction of the space.

In a practical setting, the complexity of representing such manifolds or shape spaces in closed form is unmanageable and all that is available is a finite number of (possibly random) samples obtained from these shape spaces. It is thus important to be able to determine fundamental properties of shape spaces directly from this finite representation, without resorting to cumbersome algorithms that first perform shape reconstruction. In this paper we address the problem of estimating the *intrinsic dimension* of a manifold and the *intrinsic entropy* of the measured manifold random samples. These two quantities measure the geometric and statistical complexity of the underlying shape spaces and play a central role in many applications, ranging from computational biology [1] to image processing [2].

Informally, the intrinsic dimension of a manifold describes how many “degrees of freedom” are necessary to generate the observed data. The classical way to estimate such quantity is based on linear projection techniques [3]: a linear map is explicitly constructed and dimension is estimated by applying principal component analysis (PCA), factor analysis, or multidimensional scaling (MDS) to analyze the eigenstructure of the data. These methods estimate dimension by looking at the magnitude of the eigenvalues of the data covariance and determining in some *ad-hoc* fashion the numeber of such eigenvalues necessary to describe most of the data. As they do not account for non-linearities, linear methods tend to overestimate intrinsic dimension. Both nonlinear PCA [4] methods and the ISOMAP [5] try to circumvent this problem but they still rely on unreliable and costly eigenstructure estimates. Other methods have been proposed, ranging from fractal dimension [6], estimating packing numbers [7] to a maximum likelihood approach [8].

The intrinsic entropy of random samples obtained from a manifold is an information theoretic measure of the complexity of the distribution of the samples supported on the manifold. When the distribution is absolutely continuous with respect to the Lebesgue measure restricted to the lower dimensional manifold, this intrinsic entropy can be useful for exploring data compression over the manifold, registering medical images or geographical information [9] or, as suggested in [2], clustering of multiple sub-populations on the manifold.

This chapter is a follow-up to recent work by the authors on deriving intrinsic dimension and entropy estimators based on random graphs [10], [11], providing the proofs of statistical consistency of the proposed estimators for general Riemann manifolds. We note that, except for [8], our work is the only one analyzing the statistical properties of intrinsic dimension estimators. The algorithms described here are based on constructing Euclidean k-nearest neighbor (k-NN) graphs or geodesic minimal spanning trees (GMST) over all the sample points and using their growth rate to estimate the intrinsic dimension and entropy by simple linear least squares and method of moments procedure. This approach allows for the estimation of the desired quantities using algorithms with low computational complexity that avoid reconstructing the manifold or estimating multivariate distributions.

We remark that the work presented here is intimately related with recent developments on nonlinear dimensionality reduction and manifold learning [5], [12]–[16].

The remainder of this chapter is organized as follows. In Section II we introduce entropic graphs and some of the properties that make them interesting for dimension and entropy estimation. Section III describes the asymptotic behavior of such graphs in Euclidean spaces and Section IV extends these results to Riemann manifolds. The proposed algorithms are described in Section V. Experimental results are reported in Section VI. The technical proofs of the main results presented here are complied in the Appendix.

II. Entropic Graphs and their Properties

Let $\mathcal{X}_n = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ be n independent identically distributed (i.i.d.) random vectors in a compact subset of $\mathbb{R}^d$, with multivariate Lebesgue density f. $\mathcal{X}_n$ will also be called the set of random vertices.

By solving certain optimization problems on the set $\mathcal{X}_n$, one can obtain special graph constructions. One such example is the k-NN graph. Start by defi ning the (1-)nearest neighbor of $\boldsymbol{X}_i$ in $\mathcal{X}_n$ as

$$\arg \min_{\boldsymbol{X} \in \mathcal{X}_n \setminus \{\boldsymbol{X}_i\}} d(\boldsymbol{X}, \boldsymbol{X}_i) \ ,$$

where distances between points are measured in terms of some suitable distance function $d(\cdot, \cdot)$. For general integer $k \geq 1$, the k-nearest neighbor of a point is defi ned in a similar way. The k-NN graph

puts an edge between each point in $\mathcal{X}_n$ and its k-nearest neighbors. Let $\mathcal{N}_{k,i}(\mathcal{X}_n)$ be the set of k-nearest neighbors of $\boldsymbol{X}_i$ in $\mathcal{X}_n$. The total edge length of the k-NN graph is defined as:

$$L_\gamma^{k\text{-NN}}(\mathcal{X}_n) = \sum_{i=1}^{n} \sum_{\boldsymbol{X} \in \mathcal{N}_{k,i}(\mathcal{X}_n)} d^\gamma(\boldsymbol{X}, \boldsymbol{X}_i) \ , \tag{1}$$

where γ is a power weighting constant.

Another example is the MST problem, where the goal is to find a graph of minimum total edge length among the graphs $\mathcal{T}$ which span the sample $\mathcal{X}_n$. The minimum total edge length is defined as:

$$L_\gamma^{\text{MST}}(\mathcal{X}_n) = \min_{T \in \mathcal{T}} \sum_{e \in T} w^\gamma(e) \ , \tag{2}$$

where e is an edge in the graph and $w(e)$ is its weight. If edge e connects points $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ in $\mathcal{X}_n$, then its weight is $w(e) = d(\boldsymbol{X}_i, \boldsymbol{X}_j)$.

The k-NN graph or the MST are part of a larger class of graphs that were called *entropic graphs* in [2] or *continuous quasi-additive functionals* in [17]. Other graphs in this class are the minimal Euclidean matching graph, the traveling salesman tour, the Steiner tree and minimal triangulation among others. Intuitively, a graph is in this class if its total edge length functional, $L_\gamma(\mathcal{X}_n)$, can be closely approximated by the sum of the edge length functionals of the graphs constructed on a dense partition of the compact set that contains the support of $\boldsymbol{X}_i$. The following properties, which are commonly satisfied by all the graph constructions mentioned above, play a key role in formalizing (and proving) such type of results.

Without loss of generality, assume that the random vectors $\boldsymbol{X}_i$'s take values on $[0,1]^d$. Let F be any finite subset of $[0,1]^d$ and L_γ be a functional on F.

1) L_γ has an Euclidean structure if it satisfies:
   a) *Translation invariance*: $\forall\, \boldsymbol{y} \in \mathbb{R}^d\,,\ L_\gamma(F) = L_\gamma(F + \boldsymbol{y})$.
   b) *Homogeneity of order* γ: $\forall\, \alpha > 0\,,\ L_\gamma(\alpha\, F) = \alpha^\gamma L_\gamma(F)$.
2) L_γ is *subadditive* if, given a partition of a subset $R \in [0,1]^d$ into subsets R_1 and R_2, it satisfies

$$L_\gamma(F \cap R) \leq L_\gamma(F \cap R_1) + L_\gamma(F \cap R_2) + C_1(\text{diam}\, R)^\gamma \ ,$$

   for some constant $C_1 > 0$ independent of R, R_1 and R_2, where $\text{diam}\, R$ is the diameter of R.
3) L_γ is *superadditive* if, for the same partition defined above, it satisfies

$$L_\gamma(F \cap R) \geq L_\gamma(F \cap R_1) + L_\gamma(F \cap R_2) \ .$$

4) L_γ is *continuous* if there exists a constant $C_2 > 0$ such that for all finite subsets F and G of $[0,1]^d$,

$$|L_\gamma(F \cup G) - L_\gamma(F)| \leq C_2\, (\text{card}(G))^{(d-\gamma)/d} \ ,$$

where card(G) is the cardinality of set G. Note that continuity implies

$$|L_\gamma(F) - L_\gamma(G)| \leq 2\, C_2 \left(\text{card}(F \triangle G)\right)^{(d-\gamma)/d} ,$$

where $F \triangle G = (F \cup G) \setminus (F \cap G)$ denotes the symmetric difference of sets F and G.

III. Entropic Graphs on Euclidean Spaces

If $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^d$ and $d(\boldsymbol{X}, \boldsymbol{Y}) = |\boldsymbol{X} - \boldsymbol{Y}|$, i.e., the Euclidean distance between $\boldsymbol{X}$ and $\boldsymbol{Y}$, then both the MST graph and the k-NN graph fall under the framework of continuous quasi-additive Euclidean functionals [17]. By showing that they satisfy subadditive, superadditive and continuous properties, their almost sure (a.s.) asymptotic behavior (also convergence in the mean) follows easily from the *umbrella* theorems for such graphs:

Theorem 1 ([17], [18]): Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. random vectors with values in $[0,1]^d$ and Lebesgue density f. Let $d \geq 2$, $1 \leq \gamma < d$ and define $\alpha = (d-\gamma)/d$. Then

$$\lim_{n\to\infty} \frac{L_\gamma(\mathcal{X}_n)}{n^\alpha} = \beta_{d,L_\gamma} \int_{[0,1]^d} f^\alpha(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \qquad a.s. ,$$

where $L_\gamma(\mathcal{X}_n)$ is given by equation (1) or (2) with Euclidean distance, and, β_{d,L_γ} is a constant independent of f. Furthermore, the mean length $E\left[L_\gamma(\mathcal{X}_n)\right]/n^\alpha$ converges to the same limit.

Theorem 1 states that the limiting behavior of the graph length functional is determined by the *extrinsic* Rényi α-entropy of the multivariate Lebesgue density f:

$$H_\alpha^{\mathbb{R}^d}(f) = \frac{1}{1-\alpha} \log \int_{\mathbb{R}^d} f^\alpha(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} . \tag{3}$$

In the limit, when $\alpha \to 1$ the usual Shannon entropy, $-\int_{\mathbb{R}^d} f(\boldsymbol{x}) \log f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$, is obtained. This remarkable asymptotic behavior motivates the name *entropic graphs* given in [2].

Assume now that the random set $\mathcal{Y}_n = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ is constrained to lie on a compact smooth m-dimensional manifold $\mathcal{M}$. The distribution of $\boldsymbol{Y}_i$ becomes singular with respect to Lebesgue measure and an application of Theorem 1 results in a zero limit for the length functional of the particular graph. However, this behavior can be modified by changing the way distances between points are measured. For this purpose, we use the framework of Riemann manifolds.

IV. Entropic Graphs on Riemann Manifolds

Given a smooth manifold $\mathcal{M}$, a Riemann metric g is a mapping which associates to each point $\boldsymbol{y} \in \mathcal{M}$ an inner product $g_{\boldsymbol{y}}(\cdot,\cdot)$ between vectors tangent to $\mathcal{M}$ at $\boldsymbol{y}$ [19]. A *Riemann manifold* $(\mathcal{M}, g)$ is just a

smooth manifold $\mathcal{M}$ with a given Riemann metric g. As an example, when $\mathcal{M}$ is a submanifold of the Euclidean space $\mathbb{R}^d$, the naturally induced Riemann metric on $\mathcal{M}$ is just the usual dot product between vectors.

A Riemann metric g endows $\mathcal{M}$ with a distance $d_g(\cdot,\cdot)$ via geodesics and a measure μ_g via the volume element [19]. Given the geodesic distance, one can define nearest neighbor relations or edge weights in terms of d_g instead of the usual Euclidean distance $|\cdot|$ and, consequently, define the total edge length $L_\gamma(\mathcal{Y}_n)$ as in (1) or (2), with the correspondence $d \to d_g$.

We can now extend Theorem 1 to general compact Riemann manifolds. This extension, Theorem 2, states that the asymptotic behavior of $L_\gamma(\mathcal{Y}_n)$ is no longer determined by the density of $\boldsymbol{Y}_i$ relative to the Lebesgue measure of $\mathbb{R}^d$, but depends instead on the the density of $\boldsymbol{Y}_i$ relative to μ_g.

Theorem 2: Let $(\mathcal{M}, g)$ be a compact smooth Riemann m-dimensional manifold. Suppose $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are i.i.d. random elements of $\mathcal{M}$ with bounded density f relative to μ_g. Let L_γ be the total edge length of the MST graph or the k-NN graph with lengths computed using the geodesic distance d_g. Assume $m \geq 2$, $1 \leq \gamma < m$ and define $\alpha = (m-\gamma)/m$. Then,

$$\lim_{n\to\infty} \frac{L_\gamma(\mathcal{Y}_n)}{n^\alpha} = \beta_{m,L_\gamma} \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y}) \qquad a.s.\ , \tag{4}$$

where β_{m,L_γ} is a constant independent of f and $\mathcal{M}$. Furthermore, the mean length $E\left[L_\gamma(\mathcal{Y}_n)\right]/n^\alpha$ converges to the same limit.

Now, the limiting behavior of $L_\gamma(\mathcal{Y}_n)$ is related to the *intrinsic* Rényi α-entropy of the multivariate density f on $\mathcal{M}$:

$$H_\alpha^{(\mathcal{M},g)}(f) = \frac{1}{1-\alpha} \log \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y})\ . \tag{5}$$

An immediate consequence of Theorem 2 is that, for known m,

$$\hat{H}_\alpha^{(\mathcal{M},g)}(\mathcal{Y}_n) = \frac{m}{\gamma}\left[\log \frac{L_\gamma(\mathcal{Y}_n)}{n^{(m-\gamma)/m}} - \log \beta_{m,L_\gamma}\right] \tag{6}$$

is an asymptotically unbiased and strongly consistent estimator of the intrinsic α-entropy $H_\alpha^{(\mathcal{M},g)}(f)$.

The proof of Theorem 2 is given in Appendix A. The intuition behind it comes from the fact that a Riemann manifold $\mathcal{M}$, with associated distance and measure, looks locally like $\mathbb{R}^m$ with Euclidean distance $|\cdot|$ and Lebesgue measure λ. This implies that on small neighborhoods of the manifold, the total edge length $L_\gamma(\mathcal{Y}_n)$ behaves like a Euclidean length functional. As $\mathcal{M}$ is assumed compact, it can be covered by a finite number of such neighborhoods. This fact, together with subadditive and superadditive properties [17] of L_γ, allows for repeated applications of Theorem 1 resulting in (4).

A. Approximating Geodesic Distances on Submanifolds of $\mathbb{R}^d$

Although Theorem 2 provides a characterization of the asymptotic behavior of entropic graphs over random points supported on a manifold, one further step is missing in order to make it applicable to a wide class of practical problems. This extra step comes from the computation of the length functionals which depends on finding geodesic distances between sample points, which in turn require knowing the manifold $\mathcal{M}$. However, in the general manifold learning problem, $\mathcal{M}$ (or any representation of it) is not known in advance. Consequently, the geodesic distances between points on $\mathcal{M}$ cannot be computed exactly and have to be estimated solely from the data samples.

In [10], the geodesic minimal spanning tree (GMST) algorithm was proposed, where the pairwise geodesic distances between sample points are estimated by running Dijkstra's shortest path algorithm over a global graph of "neighborhood relations" among all sample points of the manifold. If $\hat{d}(e_{ij})$ is the estimate of the geodesic length of edge $e_{ij} = (\boldsymbol{Y}_i, \boldsymbol{Y}_j)$ obtained by this algorithm, then the GMST is defined as the minimal graph over $\mathcal{Y}_n$ whose length is:

$$\hat{L}_\gamma^{\text{GMST}}(\mathcal{Y}_n) = \min_{T \in \mathcal{T}} \sum_{e \in T} \hat{d}^\gamma(e) \ . \tag{7}$$

By using geodesic information, the GMST length functional encodes global structure about the nonlinear manifold. The geodesic distances between sample points on the manifold are uniformly well approximated by $\hat{d}$ in the following sense:

Theorem 3: Let $(\mathcal{M}, g)$ be a compact Riemann submanifold of $\mathbb{R}^d$. Suppose $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are i.i.d. random vectors of $\mathcal{M}$, with density bounded away from zero. Then, with probability 1,

$$\max_{\substack{1 \leq i,j \leq n \\ i \neq j}} \left| \frac{\hat{d}(\boldsymbol{Y}_i, \boldsymbol{Y}_j)}{d_g(\boldsymbol{Y}_i, \boldsymbol{Y}_j)} - 1 \right| \to 0 \quad \text{as } n \to \infty \ . \tag{8}$$

This theorem is proven in Appendix B. We remark that there exist alternative algorithms for computing geodesic distances that can also provide guarantees similar to theorem 3. Of particular interest for future work is the method proposed in [20] for estimating geodesics that accounts for noisy samplings of the manifold.

Unlike the MST, the k-NN graph is only influenced by local distances. For fixed k, the maximum nearest neighbor distance of all points in $\mathcal{Y}_n$ goes to zero as the number n of samples increases. For n sufficiently large, this implies that the k-NN of each point will fall in a neighborhood of the manifold where geodesic curves are well approximated by the corresponding straight lines between end points. This suggests using simple Euclidean k-NN distances ($|\boldsymbol{Y}_i - \boldsymbol{Y}_j|$) as surrogates for the corresponding

true nearest neighbor geodesic distances ($d(\boldsymbol{Y}_i, \boldsymbol{Y}_j)$). In fact, we prove in Appendix C that the geodesic k-NN distances are uniformly well approximated by the corresponding Euclidean k-NN distances in the following sense:

Theorem 4: Let $(\mathcal{M}, g)$ be a compact Riemann submanifold of $\mathbb{R}^d$. Suppose $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are i.i.d. random vectors of $\mathcal{M}$. Then, with probability 1,

$$\max_{\substack{1\le i\le n \\ \boldsymbol{Y}\in\mathcal{N}_{k,i}(\mathcal{Y}_n)}} \left| \frac{|\boldsymbol{Y}-\boldsymbol{Y}_i|}{d_g(\boldsymbol{Y},\boldsymbol{Y}_i)} - 1 \right| \to 0 \quad \text{as } n\to\infty \ . \tag{9}$$

Finally, the asymptotic behavior of the GMST or the Euclidean k-NN graph is a simple consequence of Theorem 2 and Theorems 3 and 4:

Corollary 5: Let $(\mathcal{M}, g)$ be a compact smooth Riemann m-dimensional manifold. Suppose $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are i.i.d. random elements of $\mathcal{M}$ with bounded density f relative to μ_g. Let $\hat{L}_\gamma$ be the total edge length of the GMST graph or the Euclidean k-NN graph defined over $\mathcal{Y}_n$. Then,

$$\lim_{n\to\infty} \frac{\hat{L}_\gamma(\mathcal{Y}_n)}{n^\alpha} = \beta_{m,L_\gamma} \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y}) \qquad a.s. \ , \tag{10}$$

where β_{m,L_γ} is a constant independent of f and $\mathcal{M}$. Furthermore, the mean length $E\left[L_\gamma(\mathcal{Y}_n)\right]/n^\alpha$ converges to the same limit.

Proof: For example, for the k-NN case,

$$\hat{L}_\gamma(\mathcal{Y}_n) = \sum_{i=1}^{n} \sum_{\boldsymbol{Y}\in\mathcal{N}_{k,i}} \left(\frac{|\boldsymbol{Y}-\boldsymbol{Y}_i|}{d_g(\boldsymbol{Y},\boldsymbol{Y}_i)} \right)^\gamma d_g^\gamma(\boldsymbol{Y},\boldsymbol{Y}_i) \ .$$

The uniform convergence expressed by Theorem 4 implies that

$$\hat{L}_\gamma(\mathcal{Y}_n) = (1+o(1))^\gamma L_\gamma(\mathcal{Y}_n) \ .$$

Corollary 5 now follows from an application of Theorem 2. The GMST case is similar. ■

We remark that corollary 5 differs from corollary 1 presented in [10], in that the latter discusses the asymptotic behavior of the total edge length of the MST as a function of the samples embedded on the m-dimensional Euclidean space that parameterizes the manifold (assuming a global conformal mapping), as opposed to the samples supported on the manifold itself considered here.

With regards to computational complexity, the geodesic free property of the k-NN algorithm makes it computationally inexpensive as compared with other manifold learning algorithms. In this case, complexity is dominated by determining nearest neighbors, which can be done in $O(n \log n)$ time for n sample points. This contrasts with the GMST, which, as ISOMAP, requires a costly $O(n^2 \log n)$ implementation of the geodesic pairwise distance estimation step.

V. Joint Intrinsic Dimension/Entropy Estimation

The asymptotic characterization of the GMST or k-NN length functional stated in Corollary 5 provides the framework for developing consistent estimators of both intrinsic dimension and entropy. The key observation is to notice that the growth rate of the length functional is strongly dependent on m while the constant in the convergent limit is equal to the intrinsic α-entropy. We use this strong growth dependence as a motivation for a simple estimator of m. Define $l_n = \log \hat{L}_\gamma(\mathcal{Y}_n)$. According to Corollary 5, l_n has the following approximation

$$l_n = a \ \log n + b + \epsilon_n \ , \tag{11}$$

where

$$\begin{aligned} a &= (m-\gamma)/m \ , \\ b &= \log \beta_{m,L_\gamma} + \gamma/m \ H_\alpha^{(\mathcal{M},g)}(f) \ , \end{aligned} \tag{12}$$

$\alpha = (m-\gamma)/m$ and ϵ_n is an error residual that goes to zero a.s. as $n \to \infty$.

Using the additive model (11), we propose a simple non-parametric least squares strategy based on resampling from the population $\mathcal{Y}_n$ of points in $\mathcal{M}$. Specifically, let $p_1, \ldots, p_Q$, $1 \leq p_1 < \ldots, < p_Q \leq n$, be Q integers and let N be an integer that satisfies $N/n = \rho$ for some fixed $\rho \in (0, 1]$. For each value of $p \in \{p_1, \ldots, p_Q\}$ randomly draw N bootstrap datasets $\mathcal{Y}_p^j$, $j = 1, \ldots, N$, with replacement, where the p data points within each $\mathcal{Y}_p^j$ are chosen from the entire data set $\mathcal{Y}_n$ independently. From these samples compute the empirical mean of the functionals $\bar{L}_p = N^{-1} \sum_{j=1}^N \hat{L}_\gamma(\mathcal{Y}_p^j)$. Defining $\bar{\boldsymbol{l}} = [\log \bar{L}_{p_1}, \ldots, \log \bar{L}_{p_1}]^T$ we write down the linear vector model

$$\bar{\boldsymbol{l}} = A \begin{bmatrix} a \\ b \end{bmatrix} + \epsilon \tag{13}$$

where

$$A = \begin{bmatrix} \log p_1 & \ldots & \log p_Q \\ 1 & \ldots & 1 \end{bmatrix}^T .$$

We now take a method-of-moments (MOM) approach in which we use (13) to solve for the linear least squares (LLS) estimates $\hat{a}, \hat{b}$ of a, b followed by inversion of the relations (12). After making a simple large n approximation, this approach yields the following estimates:

$$\begin{aligned} \hat{m} &= \text{round}\{\gamma/(1-\hat{a})\} \\ \hat{H}_\alpha^{(\mathcal{M},g)} &= \frac{\hat{m}}{\gamma} \left(\hat{b} - \log \beta_{\hat{m},L_\gamma} \right) \ . \end{aligned} \tag{14}$$

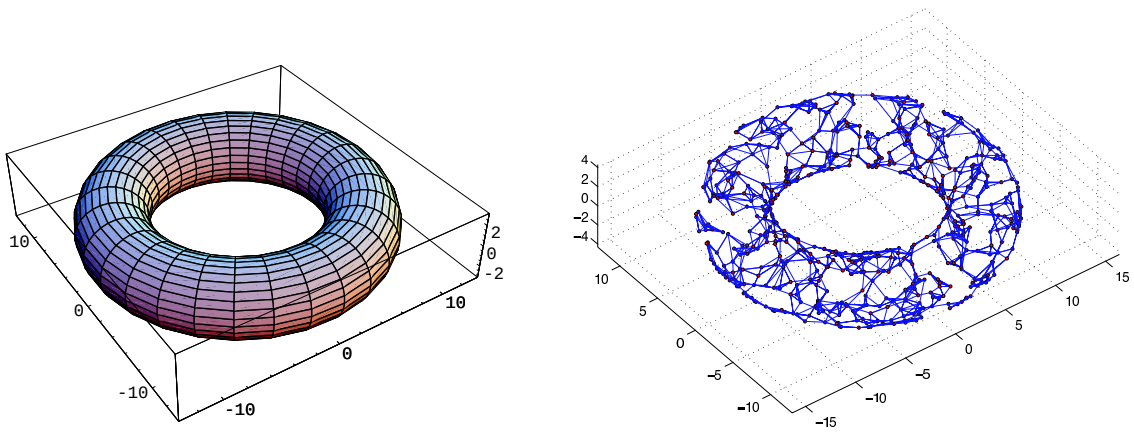

Fig. 1. The 2D-torus and the 4-NN graph on 500 points sampled uniformly from the torus.

The constants β_{m,L_γ} in the above estimators depend only on m, γ and the particular entropic graph construction algorithm, e.g., GMST or k-NN. Due to the slow growth of $\{\beta_{m,L_\gamma}\}_{m>0}$ in the large n regime for which the above estimates were derived, $\beta_{m,\gamma}$ is not required for the dimension estimator. On the other hand, the value of β_{m,L_γ} is required to obtain unbiased estimates of entropy. β_{m,L_γ} is the limit of the normalized length functional of the corresponding Euclidean entropic graph for a uniform distribution on the unit cube $[0,1]^m$. As, closed form expressions are not available, it can be determined by performing Monte Carlo simulations of the entropic graph length on the unit cube $[0,1]^m$ for uniform random samples. Another approach is to use analytical approximations and bounds for the GMST case, e.g. available in [17].

VI. Experimental Results

We illustrate the performance of the entropic graph algorithm on manifolds of known dimension as well as on a real high dimensional data set consisting of handwritten digits. In all the simulations we fi xed the parameters $\gamma = 1$ and $p_1 = n - Q, \ldots, p_Q = n - 1$. With regards to intrinsic dimension estimation, we compare our algorithms to ISOMAP. In ISOMAP, similarly to PCA, intrinsic dimension is usually estimated by detecting a knee in the residual fi tting error curve as a function of subspace dimension.

A. Torus

First, we consider the case of the 2-dimensional torus embedded in $\mathbb{R}^3$ (Figure 1). This manifold presents some challenges as it does not satisfy any of the usual isometric or conformal embedding

TABLE I

NUMBER OF CORRECT DIMENSION ESTIMATES OVER 30 TRIALS AS A FUNCTION OF THE NUMBER OF SAMPLES FOR THE TORUS ($N = 5$, Q=10).

n	200	400	600
GMST	29	30	30
5-NN	29	30	30

TABLE II

ENTROPY ESTIMATES FOR THE TORUS ($n = 600$, $N = 5$, Q=10).

	emp. mean	std. deviation
GMST	10.0	0.55
5-NN	9.6	0.93

constraints required by ISOMAP or Hessian eigenmaps [14], among others. We tested the algorithms over 30 generations of uniform random samples over the torus for different sample sizes n, and counted the number of correct dimension estimates. We note that in all the simulations ISOMAP always overestimated the intrinsic dimension as 3. The results for the GMST and k-NN are shown in Table I. Table II shows the entropy estimates obtained by both methods on uniform samples supported on the torus. The true ($\alpha = 1/2$) entropy is $H_{1/2} = \log(120\pi^2) \approx 10.21$.

B. MNIST Database of Handwritten Digits

The MNIST database[1] consists of 256 gray levels images of handwritten digits obtained by optical character recognition. This publicly available database has became one of the benchmarks for testing new digit recognition algorithms [21], containing extensive test and training sets of all digits. Each digit in the database consists of a 28×28 pixel image that was size normalized and translated so that its center of mass lies in the center of the image. For the purpose of dimensionality estimation, we chose the first 1000 samples of digits 0 to 9 (Figure 2) in the training set.

Figure 3 shows the histogram of the dimension estimates for 30 simulations of the 5-NN algorithm applied to the samples of digits 0 to 9. Figure 4 shows the boxplot of the entropy estimates for the same scenario. Although the histograms show high variability, most of the estimates are between 9 and 15. It is interesting to notice that digit 1 exhibits the lowest dimension estimate, between 9 and 10, while

[1] `http://yann.lecun.com/exdb/mnist/`

Fig. 2. Samples from digits 0 to 9 in the MNIST database.

all the other digits exhibit dimensions between 12 and 14. The lower complexity of digit 1 can also be seen from Figure 4, where its entropy estimate is much lower that all other digits. Also of interest is the bimodal behavior of the histogram of digit 7, with one mode concentrated at 10, 11 and the other at 13. After looking at the images selected in the realizations that resulted in the lower dimensional mode estimates, we realized that these images, although classified as a 7, are also very close to digit 1, thus contributing to lowering the dimension estimates. This effect can also be observed in the boxplot of entropy estimates of Figure 4, where the high variance of the entropy estimate of digit 7 and consequent overlap of confidence intervals with digit 1 suggest the presence of images with a lower complexity.

For comparison purposes, we show in Figure 5 the eigenvalue plots for digits 2 and 3 used by ISOMAP to estimate intrinsic dimension. Even though it is not obvious how to assign a single dimension estimate from this plot - one of the main disadvantages of using spectral methods to estimate dimension - it is clear that the dataset should be at most 10-dimensional, as the residual variance ceases to decrease significantly after that value. The difference between the estimates predicted by entropic graphs and ISOMAP might be justified by the isometric assumption required by ISOMAP. The digits database contains nonlinear transformations, such as width distortions of each digit, that are not described by isometries. As consequence, ISOMAP underestimated these extra degrees of freedom, resulting in a lower dimension estimate than the entropic graphs, that are valid for a broader class of manifolds.

Finally, we present in Figure 6 the results of applying the proposed algorithm to the merged samples of digits 2 and 3. As it can be seen, the histogram of the dimension estimates shows an increase of its mode by one, being dominated by the dimensionality of the most complex digit (3). The entropy

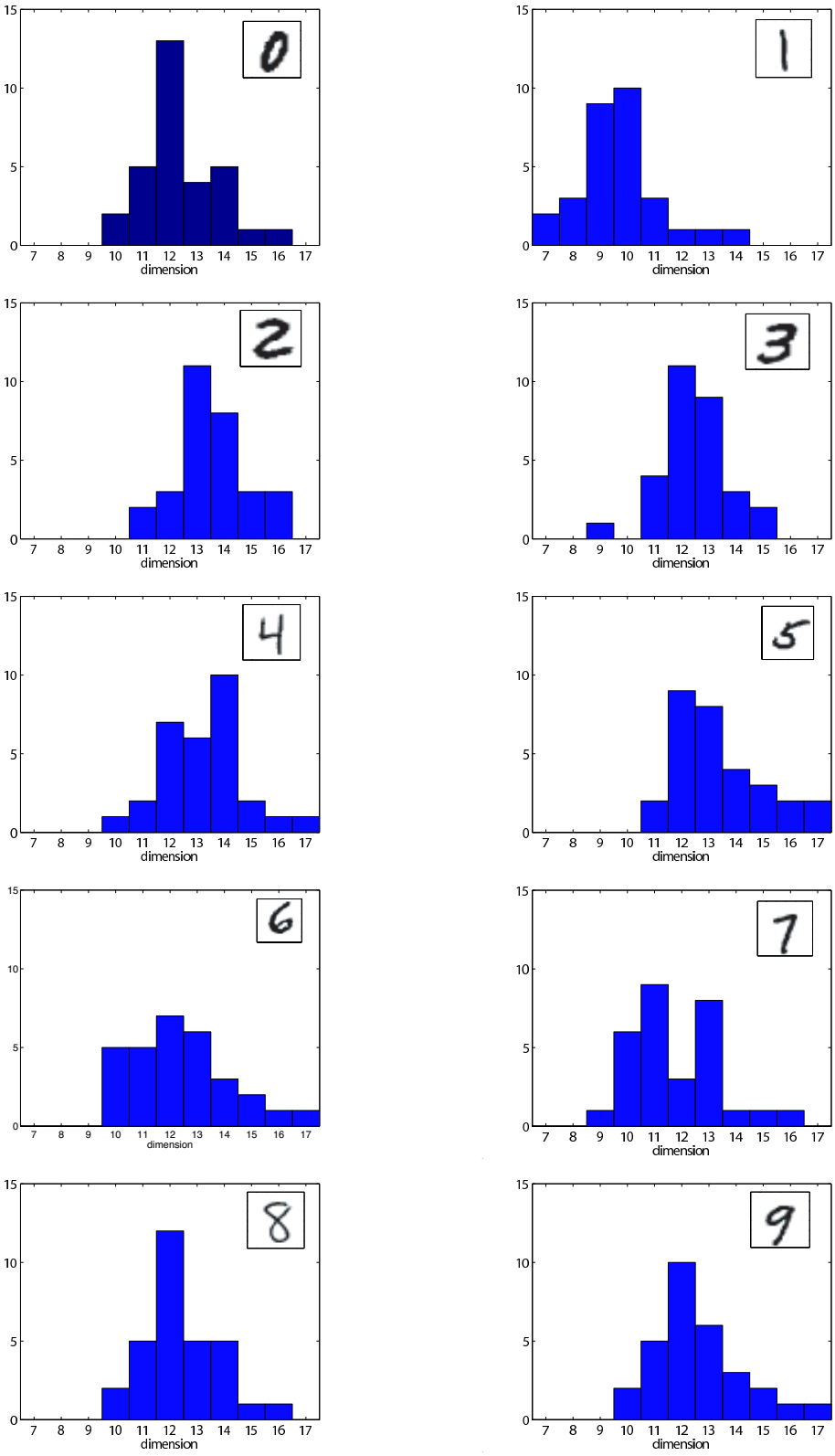

Fig. 3. Histograms of intrinsic dimensionality estimates for digits 0 to 9 in the MNIST database using a 5-NN graph ($N = 10$, $Q = 15$).

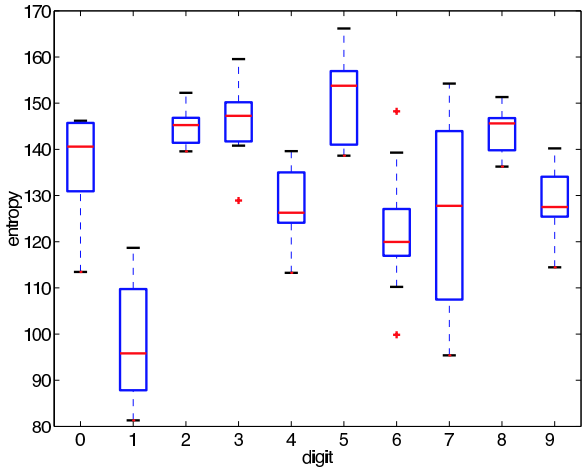

Fig. 4. Boxplot of entropy estimates for digits 0 to 9 in the MNIST database using a 5-NN graph ($N = 10$, $Q = 15$).

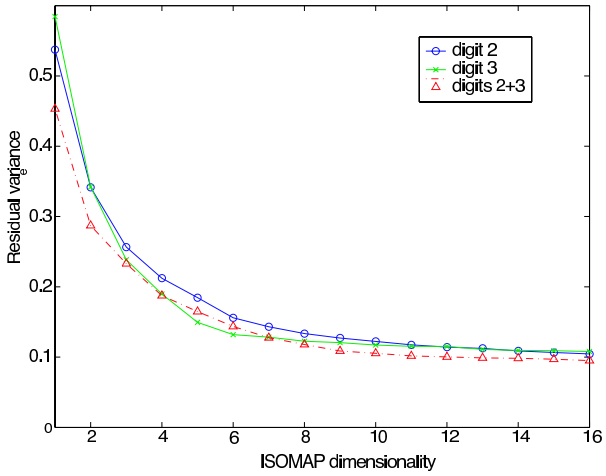

Fig. 5. ISOMAP ($k = 6$) residual variance for digits 2 and 3 in the MNIST database.

boxplot shows an increase of the median entropy estimate by roughly one bit. This should be expected, as compressing the augmented data set requires only one extra bit to identify which digit is being coded and then the individual codes for each digit can be used.

VII. Conclusions

We have discussed the use of computational geometry graph constructions and geometric probability tools for the estimation of intrinsic dimension and entropy of shape spaces based solely on a finite random sampling of the underlying shapes. In particular, we have shown the strong statistical consistency of estimators based on k-nearest neighbor graphs or minimal spanning trees under the very general

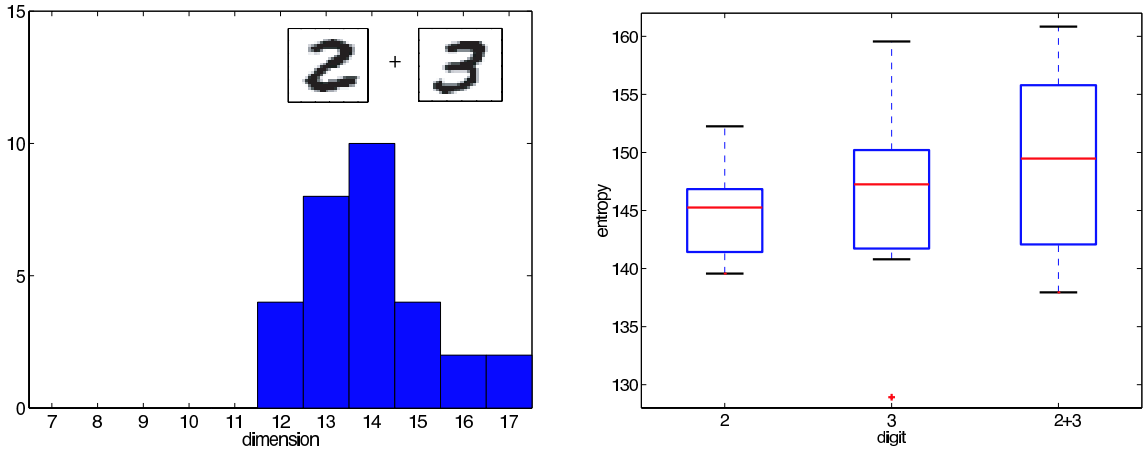

Fig. 6. Histogram of intrinsic dimensionality estimates and boxplot of entropy estimates for digits $2+3$ in the MNIST database using a 5-NN graph ($N = 10$, $Q = 15$).

assumption of high dimensional data supported on a compact Riemann manifold. These results provide a departure from usually strong assumptions of linear, isometric or conformal embeddings expressed in the previous literature on the subject.

We are currently working on extending the proposed methods to data sets that exhibit a varying complexity across the data, characterized by a changing intrinsic dimension. This will allow the analysis of interesting datasets, like images composed of textures of different complexity or computational biology models of protein interaction [1]. Future work also includes developing bias correction mechanisms to improve the bootstrapping resampling step of the algorithm and account for dependencies in the sampling process.

APPENDIX

A. *Proof of Theorem 2*

In this appendix, Theorem 2 is proven. We first introduce two auxiliary lemmas and take a small detour to discuss Euclidean boundary functionals, which are a key tool in proving asymptotic results for continuous quasi-additive Euclidean functionals [17].

The first lemma formalizes the intuition that a a Riemann manifold $\mathcal{M}$, with associated distance d_g and measure μ_g, looks locally like $\mathbb{R}^m$ with Euclidean distance $|\cdot|$ and Lebesgue measure λ:

Lemma 1 ([22, Lemma 5.1]): Let $(\mathcal{M}, g)$ be a smooth Riemann m-dimensional manifold. For any $\boldsymbol{x} \in \mathcal{M}$ and $\varepsilon > 0$, there exists a chart (U, ϕ) for $\mathcal{M}$, with $\boldsymbol{x} \in U$, such that

$$(1+\varepsilon)^{-1} \ |\phi(\boldsymbol{y}) - \phi(\boldsymbol{z})| \leq d_g(\boldsymbol{y}, \boldsymbol{z}) \leq (1+\varepsilon) \ |\phi(\boldsymbol{y}) - \phi(\boldsymbol{z})| \qquad \forall \ \boldsymbol{y}, \boldsymbol{z} \in U \tag{15}$$

and for any measurable subset $B \subset U$

$$(1-\varepsilon) \ \lambda(\phi(B)) < \mu_g(B) < (1+\varepsilon) \ \lambda(\phi(B)) \ . \tag{16}$$

Recall that a chart (U, ϕ) consists of a neighborhood U such that $\phi : \mathcal{M} \cap U \to \mathbb{R}^m$ determines a parametric representation of $\mathcal{M} \cap U$ in the Euclidean m-dimensional space, i.e., for $\boldsymbol{y} \in \mathcal{M} \cap U$, $\phi(\boldsymbol{y})$ represents $\boldsymbol{y}$ in an Euclidean m-dimensional coordinate system.

1) Boundary Functionals on Jordan Measurable Sets: We now informally introduce the notions of boundary functional. For formal defi nitions and details, we refer the reader to [17].

By appropriate canonical modifi cations of an Euclidean subadditive functional $L(F)$, it is possible to construct an associated *boundary functional* $L_B(F, R)$ on any subset R of $[0,1]^d$ [17]. Informally, in a boundary functional all the edges connecting point on the boundary of R (∂R) have zero length, so that ∂R can be seen as single point: all edges joined to the boundary are joined to one another, or, in other words, joining edges using ∂R adds no additional cost to the functional.

The importance of boundary functionals resides in the fact that they are superadditve, a property that many of the standard total edge functionals lack. If R is partitioned into sets R_1 and R_2 then L_B is superadditive if

$$L_B(F, R) \geq L_B(F \cap R_1, R_1) + L_B(F \cap R_2, R_2) \ .$$

When R, R_1, R_2 are rectangles, translation invariance and homegeneity properties of any Euclidean functional, endow $L_B(\cdot, R)$ with a self similarity property, in a way that, for a uniform sample, the value of the functional on a set of the partition is statistically similar to its value on any other partition set. However, when R is an arbitrary set, this self similarity property is lost. We now show that if R is Jordan measurable a superadditive functional has the same type of asymptotic behavior as when R is a rectangle.

Lemma 2: Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. random vectors with values in $R \subset [0,1]^d$ and bounded Lebesgue density f. Assume R is Jordan measurable. Let $L_B(\cdot, R)$ be a continuous superadditive Euclidean boundary functional of order γ on $\mathbb{R}^d$. Then

$$\liminf_{n \to \infty} \frac{L_B(\mathcal{X}_n, R)}{n^\alpha} \geq \beta_{d,L} \int_R f^\alpha(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \qquad a.s. \tag{17}$$

Furthermore, the same result holds for the mean length $E\left[L_B(\mathcal{X}_n, R)\right]/n^\alpha$.

Proof: The proof of this result relies on the fact that a Jordan measurable set is "well approximated" from below by an union of disjoint cubes. We then use the known results about the behavior of Euclidean functionals over cubes.

Let $\varepsilon > 0$. As R is Jordan measurable, there exists a finite number of disjoint cubes $\{Q_i\}$ (with faces parallel to the axis) such that $Q = \cup_i Q_i \subset R$ and $\lambda(R \setminus Q) < \varepsilon$. By superadditivity,

$$L_B(\mathcal{X}_n, R) \geq \sum_i L_B(\mathcal{X}_n \cap Q_i, Q_i) \ . \tag{18}$$

Let $p_i = \int_{Q_i} f \, \mathrm{d}\lambda$. By the strong law of large numbers, $\mathcal{X}_n \cap Q_i$ consists of $n(p_i + o(1))$ i.i.d. points in Q_i distributed with density $p_i^{-1} f$. By the usual umbrella theorem,

$$\frac{L_B(\mathcal{X}_n \cap Q_i, Q_i)}{(p_i n)^\alpha} \to \beta_{d,L} \int_{Q_i} (p_i^{-1} f)^\alpha \, \mathrm{d}\lambda \qquad a.s. \tag{19}$$

We also have

$$\left| \int_R f \, \mathrm{d}\lambda - \int_Q f \, \mathrm{d}\lambda \right| \leq \|f\|_\infty \, \lambda(R \setminus Q) < \varepsilon \, \|f\|_\infty \ , \tag{20}$$

where $\|f\|_\infty = \sup\{f(\boldsymbol{x}) : \boldsymbol{x} \in R\}$ is finite by assumption. Combining (18), (19) and (20) results in

$$\liminf_{n\to\infty} \frac{L_B(\mathcal{X}_n, R)}{n^\alpha} \geq \beta_{d,L} \sum_i \int_{Q_i} f^\alpha \, \mathrm{d}\lambda \geq \beta_{d,L} \left(\int_R f^\alpha \, \mathrm{d}\lambda - \varepsilon \, \|f\|_\infty \right) \ .$$

Letting $\varepsilon \to 0$ produces the desired result. ■

Remark 1: If L_B is close in mean [17, c.f. Definition 3.9] to the underlying smooth subadditive Euclidean functional, then $\liminf$ and the inequality in equation (17) can be replaced, respectively, by $\lim$ and an equality.

2) Proof of Theorem 2: Before proving Theorem 2, we note that both the MST and the k-NN functional and respective boundary functionals defined on a Riemann manifold satisfy strong forms of subadditivity and superadditivity. Namely, if $R_1, R_2 \in \mathcal{M}$ are arbitrary sets that partition $\mathcal{M}$, then

$$L_B(F \cap R_1, R_1) + L_B(F \cap R_2, R_2) \leq L_B(F, \mathcal{M}) = L(F) \leq L(F \cap R_1) + L(F \cap R_2) + C \ , \tag{21}$$

where C is an error term independent of R_1 and R_2 (C is zero for the k-NN case). Note that the usual subadditivity and superadditivity conditions needed to prove umbrella theorems for Euclidean functionals only require that these conditions hold for partitions made of rectangles.

Proof: [Proof of Theorem 2] Let $\varepsilon > 0$. For $\boldsymbol{x} \in \mathcal{M}$ let $(U_{\boldsymbol{x}}, \phi_{\boldsymbol{x}})$ be the chart specified by Lemma 1. Without loss of generality, $U_{\boldsymbol{x}}$ may be chosen such that $\phi_{\boldsymbol{x}}(U_{\boldsymbol{x}})$ is an open ball in $\mathbb{R}^m$ (this can be

achieved by possibly shrinking the set $U_{\boldsymbol{x}}$ whose existence is guaranteed by Lemma 1). By compactness of $\mathcal{M}$, there exists a finite collection of such sets, say $\{U_i\}$, that cover $\mathcal{M}$. Define the set sequence $\{V_j\}$ by $V_1 = U_1$ and $V_j = U_j \setminus \cup_{1\le i\le j-1} V_i$, for $j \ge 2$. The sets V_j are disjoint, form a partition of $\mathcal{M}$, and $V_j \subset U_j$, for all j.

Let $p_j = \int_{V_j} f \,\mathrm{d}\mu_g$ and $\mathcal{X}_{n,j} = \phi_j(\mathcal{Y}_n \cap V_j)$. By the strong law of large numbers, $\mathcal{X}_{n,j}$ consists of $n(p_j + o(1))$ i.i.d. points in $\phi_j(V_j)$ distributed with density

$$g_j(\boldsymbol{u}) = p_j^{-1} h_j(\phi_j^{-1}(\boldsymbol{u})) f(\phi_j^{-1}(\boldsymbol{u})) \ , \quad \boldsymbol{u} \in \phi_j(V_j) \ ,$$

where h_j is the function defined in the proof of Lemma 1 in [22] (c.f. Lemma 5.1). h_j accounts for the differential changes in volume between V_j and $\phi_j(V_j)$, i.e., $\mu_g(B) = \int_{\phi(B)} h_j(\phi_j^{-1}(u)) \,\mathrm{d}\boldsymbol{u}$, for $B \subset U_j$. Recall from [22] that $1 - \varepsilon < h_j(\boldsymbol{x}) < 1 + \varepsilon$ for $\boldsymbol{x} \in V_j$.

We are now ready to apply sub and superadditivity to the partition $\{V_j\}$. By (21)

$$\sum_j L_B(\mathcal{Y}_n \cap V_j, V_j) \le L_B(\mathcal{Y}_n, \mathcal{M}) = L(\mathcal{Y}_n) \le \sum_j L(\mathcal{Y}_n \cap V_j) + C' \ . \tag{22}$$

As the sets V_j were chosen such that the geodesic lengths and Euclidean lengths are ε-close, we have by (15)

$$L(\mathcal{Y}_n \cap V_j) \le (1+\varepsilon)\, L(\mathcal{X}_{n,j}) \ . \tag{23}$$

As $L(\mathcal{X}_{n,j})$ satisfies the usual quasi-additive continuous Euclidean properties, it follows from the usual umbrella theorem that

$$\frac{L(\mathcal{X}_{n,j})}{(p_j\, n)^\alpha} \to \beta_{d,L} \int_{\phi_j(V_j)} g_j^\alpha(\boldsymbol{u}) \,\mathrm{d}\boldsymbol{u} \qquad a.s. \tag{24}$$

Changing the integration back to μ_g and using the fact that h_j is $(1 \pm \varepsilon)$-valued,

$$\begin{aligned} p_j^\alpha \int_{\phi_j(V_j)} g_j^\alpha(\boldsymbol{u}) \,\mathrm{d}\boldsymbol{u} &= \int_{\phi_j(V_j)} f^\alpha(\phi_j^{-1}(\boldsymbol{u}))\, h_j^{\alpha-1}(\phi_j^{-1}(\boldsymbol{u}))\, h_j(\phi_j^{-1}(\boldsymbol{u})) \,\mathrm{d}\boldsymbol{u} \\ &= \int_{V_j} f^\alpha(\boldsymbol{y})\, h_j^{\alpha-1}(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y}) \le (1-\varepsilon)^{\alpha-1} \int_{V_j} f^\alpha(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y}) \end{aligned} \tag{25}$$

Combining the upper bound in (22) with (23)-(25), we get:

$$\limsup_{n\to\infty} \frac{L(\mathcal{Y}_n)}{n^\alpha} \le (1+\varepsilon)(1-\varepsilon)^{\alpha-1} \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y}) \ . \tag{26}$$

The lower bound implicit in equation (4) follows in a similar way. Start by noticing that, due to (15),

$$L_B(\mathcal{Y}_n \cap V_j, V_j) \ge (1+\varepsilon)^{-1} L_B(\mathcal{X}_{n,j}, \phi_j(V_j)) \ .$$

Recall that V_j is a finite intersection of sets U_i with smooth boundary (U_i was constructed to be the inverse image of a ball trough the smooth transformation ϕ_j). So, the set $\phi_j(V_j)$ will have smooth piecewise boundary and, consequently, will be Jordan measurable. Lemma 2 can now be applied to conclude that:

$$\liminf_{n\to\infty} \frac{L_B(\mathcal{X}_{n,j}, \phi_j(V_j))}{(p_j\, n)^\alpha} \geq \beta_{d,L} \int_{\phi_j(V_j)} g_j^\alpha(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \qquad a.s.$$

Repeating the same arguments used above, we have

$$\liminf_{n\to\infty} \frac{L(\mathcal{Y}_n)}{n^\alpha} \geq (1+\epsilon)^{-1}(1+\varepsilon)^{\alpha-1} \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y})\ . \tag{27}$$

Finally, combining equations (26) and (27) and letting $\epsilon \to 0$ establishes Theorem 2. ■

B. Proof of Theorem 3

Here, we prove Theorem 3 for the case when geodesic distances are estimated using the "ϵ-rule" [23]. This rule estimates geodesic distances by running Dijkstra's shortest path algorithm over the graph constructed by putting an edge between each point and all points within a fixed radius ϵ. Of course, for the algorithm to be consistent as the number of samples n grows, ϵ has to decrease to 0 as $n \to \infty$. In particular, our proof shows that $\epsilon_n = o\left(n^{-\xi/m}\right)$, for some $0 < \xi < 1$, is sufficient to guarantee consistency.

Proof: [Proof of Theorem 3] According to [23], proving the consistency result expressed by equation (8) reduces to showing that the "δ-sampling" condition holds with probability one. This condition states that for all $\boldsymbol{x} \in \mathcal{M}$ there is a sample $\boldsymbol{x}_i$ such that $d_g(\boldsymbol{x}, \boldsymbol{x}_i) \leq \delta$.

In the following, we use the same notation as defined in the *Sampling Lemma* of [23]. In particular, $B_i(\delta)$ is the metric ball in $\mathcal{M}$ of radius δ, centered at some point $\boldsymbol{p}_i$; $V_{min}(\delta)$ is the volume of the smallest metric ball in $\mathcal{M}$ of radius δ. For Riemann submanifolds of $\mathbb{R}^d$ without boundary, $V_{min}(\delta) \asymp \delta^m$; V is the volume of $\mathcal{M}$; $f_{min} = \inf_{\boldsymbol{y}\in\mathcal{M}} f(\boldsymbol{y}) > 0$.

Begin by covering $\mathcal{M}$ with a finite family of metric balls of radius $\delta/2$, choosing the centers $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots$ such that

$$\boldsymbol{p}_{i+1} \notin \cup_{j=1}^{i} B_i(\delta/2)$$

and stopping when this is no longer possible. As no two centers $\boldsymbol{p}_i$ are within distance $\delta/2$ of each other, the balls $B_i(\delta/4)$ are disjoint and, consequently, at most $V/V_{min}(\delta/4)$ points can be chosen before the process terminates.

The δ-sampling condition will be satisfied if each ball B_i contains at least one sample, as the diameter of B_i is δ and every $\boldsymbol{x} \in \mathcal{M}$ belongs to a ball B_i. The probability of this event is:

$$P(\delta\text{-sampling condition holds}) \geq P(\text{no ball } B_i \text{ is empty}) \geq 1 - \sum_i P(B_i \text{ is empty}) \ . \tag{28}$$

Under the i.i.d. assumption on the samples, the probability $P(B_i \text{ is empty})$ can be computed as:

$$\begin{aligned} P(B_i \text{ is empty}) = \left(1 - \int_{B_i} f \mathrm{d}\mu_g\right)^n &\leq (1 - V_{min}(\delta/2)\, f_{min})^n \\ &\leq \exp\left\{-n\, V_{min}(\delta/2)\, f_{min}\right\} \ , \end{aligned} \tag{29}$$

where the last inequality follows from the inequality $\log(1-x) \leq -x$. Substituting equation (29) in (28) and using the asymptotic value for $V_{min}(\delta/2)$ results in:

$$\begin{aligned} P(\delta\text{-sampling condition holds}) &\geq 1 - \frac{V}{V_{min}(\delta/4)} \exp\left\{-n\, V_{min}(\delta/2)\, f_{min}\right\} \\ &= 1 - C_1\, V \delta^{-m} \exp\left\{-C_2\, f_{min}\, n\, \delta^m\right\} \ , \end{aligned} \tag{30}$$

where C_1 and C_2 are constants.

Now, choose $\delta = \delta_n$ as a function of the number of samples such that $\delta_n \to 0$ and $n\,\delta_n^m \to \infty$ as $n \to \infty$. For example, $\delta_n = O\left(n^{-\xi/m}\right)$, for some $0 < \xi < 1$, will satisfy these conditions. Then choose a sequence ϵ_n such that $\epsilon_n \to 0$ and $\epsilon_n/\delta_n \to 0$ as $n \to \infty$. For example, $\epsilon_n = o\left(n^{-\xi/m}\right)$. Given $\lambda > 0$, there exists an integer n_0 such that for all $n > n_0$, ϵ_n is small enough to satisfy conditions 5, 6 and 7 of *Main Theorem A* of [23]. This theorem, together with equation (30), implies that

$$P\left(\max_{\substack{1 \leq i,j \leq n \\ i \neq j}} \left| \frac{\hat{d}(\boldsymbol{Y}_i, \boldsymbol{Y}_j)}{d_g(\boldsymbol{Y}_i, \boldsymbol{Y}_j)} - 1 \right| \geq \lambda \right) \leq C_1\, V \delta_n^{-m} \exp\left\{-C_2\, f_{min}\, n\, \delta_n^m\right\} \ ,$$

for $n > n_0$. As the choice of δ_n implies that $\sum_{n>n_0} \delta_n^{-m} \exp\left\{-C_2\, f_{min}\, n\, \delta_n^m\right\} < \infty$, the desired result follows by the Borel-Cantelli Lemma. ■

C. Proof of Theorem 4

Without loss of generality, assume that $\mathcal{M} \in [0,1]^d$. We first prove that $M_{n,k} = M_{n,k}(\mathcal{Y}_n)$, the length of the longest k-NN link, converges to zero with probability 1.

Given $\varepsilon > 0$, partition $[0,1]^d$ into a finite number of cubes, $\{Q_j\}$, with edge length at most ε. Let $p_j = \int_{Q_j \cap \mathcal{M}} f(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y})$. By the strong law of large numbers, there will be $n\,(p_j + o(1))$ points in Q_j with probability 1. This implies, for $p_j > 0$, that there exists an integer N_j such that for all $n > N_j$, $n\,(p_j + o(1)) \geq k$. Let $N = \max_j N_j$. Ignoring the cubes with $p_j = 0$ (with probability 1 they will have no points), each cube has at least k points for $n > N$. This implies that for all $n > N$, $M_{n,k} < O(\varepsilon)$, i.e, $M_{n,k} \to 0$ as $n \to \infty$. With this result in hand, Theorem 4 follows directly by an application of Corollary 4 from [23].

REFERENCES

[1] H. Edelsbrummer, M. Facello, and J. Liang, "On the definition and the construction of pockets on macromolecules," *Discrete Applied Math.*, vol. 88, pp. 83–102, 1998.

[2] A.O. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Processing Magazine*, vol. 19, no. 5, pp. 85–95, October 2002.

[3] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*, Prentice Hall, Englewood Cliffs, NJ, 1988.

[4] M. Kirby, *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns*, Wiley-Interscience, 2001.

[5] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[6] F. Camastra and A. Vinciarelli, "Estimating the intrinsic dimension of data with a fractal-based method," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 10, pp. 1404–1407, October 2002.

[7] B. Kégl, "Intrinsic dimension estimation using packing numbers," in *Neural Information Processing Systems: NIPS*, Vancouver, CA, Dec. 2002.

[8] E. Levina and P. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Neural Information Processing Systems: NIPS*, Vancouver, CA, Dec. 2004.

[9] H. Neemuchwala, A. O. Hero, and P. Carson, "Image registration using entropy measures and entropic graphs," *European Journal of Signal Processing, Special Issue on Content-based Visual Information Retrieval*, vol. 85, no. 2, pp. 277–296, 2005.

[10] J. A. Costa and A. O. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *IEEE Trans. on Signal Processing*, vol. 52, no. 8, pp. 2210–2221, August 2004.

[11] J. A. Costa and A. O. Hero, "Entropic graphs for manifold learning," in *Proc. of IEEE Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA, November 2003.

[12] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear imbedding," *Science*, vol. 290, no. 1, pp. 2323–2326, 2000.

[13] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, June 2003.

[14] D. Donoho and C. Grimes, "Hessian eigenmaps: locally linear embedding techniques for high dimensional data," *Proc. Nat. Acad. of Sci.*, vol. 100, no. 10, pp. 5591–5596, 2003.

[15] Z. Zang and H. Zha, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," *SIAM Journal of Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2004.

[16] K. Weinberger and L. Saul, "Unsupervised learning of image manifolds by semidefinite programming," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington D.C., 2004.

[17] J. E. Yukich, *Probability theory of classical Euclidean optimization problems*, vol. 1675 of *Lecture Notes in Mathematics*, Springer-Verlag, Berlin, 1998.

[18] M. Penrose and J. Yukich, "Weak laws of large numbers in geometric probability," *Annals of Applied Probability*, vol. 13, no. 1, pp. 277–303, 2003.

[19] M. Carmo, *Riemannian geometry*, Birkhäuser, Boston, 1992.

[20] F. Mémoli and G. Sapiro, "Distance functions and geodesic disntances on point clouds," *to appear in SIAM Journal of Applied Math.*, 2005, (Tech. Rep. 1902, IMA, University of Minnesota, Mineapolis).

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.

[22] M. Penrose, "A strong law for the largest nearest-neighbour link between random points," *J. London Math. Soc*, vol. 60, no. 2, pp. 951–960, 1999.

[23] M. Bernstein, V. de Silva, J. C. Langford, and J. B. Tenenbaum, "Graph approximations to geodesics on embedded manifolds," Tech. Rep., Department of Psychology, Stanford University, 2000.