Entropic Graphs for Manifold Learning

Jose A. Costa and Alfred O. Hero III
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI 48109
Email: jcosta@umich.edu, hero@eecs.umich.edu

Abstract— **We propose a new algorithm that simultaneously estimates the intrinsic dimension and intrinsic entropy of random data sets lying on smooth manifolds. The method is based on asymptotic properties of entropic graph constructions. In particular, we compute the Euclidean k-nearest neighbors (k-NN) graph over the sample points and use its overall total edge length to estimate intrinsic dimension and entropy. The algorithm is validated on standard synthetic manifolds.**

I. Introduction

Several interesting classes of signals arising in fields such as bioinformatics, image processing or Internet traffic analysis live in high dimensional vector spaces. It is well known that both computational complexity and statistical performance of most algorithms quickly degrades as dimension increases. This phenomenon, usually known as *curse of dimensionality*, makes it impracticable to process such high dimensional data sets. However, many real life signals do not fill the space entirely but are constrained to lie on a smooth low dimensional non-linear manifold embedded in the high dimensional space. *Manifold learning* is concerned with the problem of discovering low dimensional structure based on a set of observed high dimensional sample points on the manifold.

In the recent past, manifold learning has received substantial attention from researchers in machine learning, computer vision, signal processing and statistics [1]–[4]. This is due to the fact that effectively solving the manifold learning problem can bring considerable improvement to the solution of such diverse problems as: feature extraction in pattern recognition; multivariate density estimation and regression in statistics; data compression and coding in information theory; visualisation of high dimensional data; or complexity reduction of algorithms. Several techniques for recovering the low dimensional structure of high dimensional data have been proposed. These range from: linear methods as principal components analysis (PCA) [5] and classical multidimensional scaling (MDS) [6]; local methods as linear local imbedding (LLE) [1], locally linear projections (LLP) [7], and Hessian eigenmaps [4]; and global methods as ISOMAP [2].

One common step to the manifold reconstruction algorithms mentioned above is that all require the explicit knowledge of the *intrinsic dimension* of the manifold. In many real life applications, this parameter cannot assumed to be known and has to be estimated from the data. A frequent way of doing this is to use linear projection techniques ([5]): a linear map is explicitly constructed and dimension is estimated by applying PCA, factor analysis, or MDS to analyze the eigenstructure of the data. These methods rely on the assumption that only a small number of the eigenvalues of the (processed) data covariance will be significant. Linear methods tend to overestimate the intrinsic dimension as they don't account for non-linearities in the data. Both nonlinear PCA [3] methods and the ISOMAP circumvent this problem but they still rely on unreliable and costly eigenstructure estimates. Other methods have been proposed based on local geometric techniques, e.g., estimation of local neighborhoods [8] or fractal dimension [9], and estimating packing numbers [10] of the manifold.

The closely related problem of estimating the manifold's *intrinsic entropy* arises if the data samples are drawn from a multivariate distribution 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 or, as suggested in [11], clustering of multiple sub-populations on the manifold.

The goal of this paper is to develop an algorithm that jointly estimates both the intrinsic dimension and intrinsic entropy on the manifold, without knowing the manifold description, given only a set of random sample points. Our approach is based on entropic graph methods; see [11] for an overview. Specifically: construct the Euclidean k-nearest neighbors (k-NN) graph over all the sample points and use its growth rate to estimate the intrinsic dimension and entropy by simple linear least squares and method of moments procedure. This method shares with the geodesic minimal spanning tree (GMST) method introduced by us in previous work [12], the simplicity of avoiding the reconstruction of the manifold or estimating the multivariate density of the samples. However, it has the main advantage of reducing runtime complexity by an order of magnitude and is applicable to a wider class of manifolds.

The remainder of the paper is organized as follows. In Section II we discuss the asymptotic behavior of the k-NN graph on a manifold and the approximation of k-NN geodesic distances by the corresponding Euclidean distances. The proposed algorithm is described in Section III. Experimental results are reported in Section IV.

The theoretical results introduced in this paper are presented without proof due to space limitations. The corresponding proofs can be found in [13].

II. The k-NN Graph

Let $\mathcal{X}_n = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ be n independent and identically distributed (i.i.d.) random vectors with values in a compact

0-7803-8104-1/03/$17.00 ©2003 IEEE

subset of $\mathbb{R}^d$. The (1-)nearest neighbor of $\boldsymbol{X}_i$ in $\mathcal{X}_n$ is given by

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

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 defined in a similar way. The k-NN graph puts an edge between each point in $\mathcal{X}_n$ and its k-nearest neighbors. Let $\mathcal{N}_{k,i} = \mathcal{N}_{k,i}(\mathcal{X}_n)$ 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:

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

where $\gamma > 0$ is a power weighting constant.

If $d(\boldsymbol{X}, \boldsymbol{Y}) = |\boldsymbol{X} - \boldsymbol{Y}|$, where $|\cdot|$ is the usual Euclidean (L_2) norm in $\mathbb{R}^d$, then the k-NN graph falls under the framework of continuous quasi-additive Euclidean functionals [14]. As a consequence, its almost sure (a.s.) asymptotic behavior (also convergence in the mean) follows easily from the *umbrella theorems* for such graphs:

Theorem 1 ([14, Theorem 8.3]): Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. random vectors with values in a compact subset of $\mathbb{R}^d$ and Lebesgue density f. Let $d \geq 2$, $1 \leq \gamma < d$ and define $\alpha = (d - \gamma)/d$. Then

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

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

The integral factor $\int f^{\alpha}$ in the a.s. limit is a monotonic function of the *extrinsic* Rényi α-entropy of the multivariate Lebesgue density f:

$$H_{\alpha}^{\mathbb{R}^d}(f) = \frac{1}{1-\alpha} \log \int_{\mathbb{R}^d} f^{\alpha}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} . \qquad (2)$$

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.

Assume now that $\mathcal{Y}_n = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ 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 k-NN 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.

A. Random Points in a Riemann Manifold

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 $\boldsymbol{y}$ [15]. 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.

For any tangent vector $\boldsymbol{v}$ to $\mathcal{M}$ at $\boldsymbol{y}$, we can define its norm as $|\boldsymbol{v}|_{g_{\boldsymbol{y}}} = g_{\boldsymbol{y}}(\boldsymbol{v}, \boldsymbol{v})$. Using this norm, it is natural to define the length of a piecewise smooth curve on $\mathcal{M}$, $\Gamma : [0,1] \to \mathcal{M}$, as $\ell(\Gamma) = \int_0^1 |\frac{\mathrm{d}}{\mathrm{d}t}\Gamma(t)|_{g_{\boldsymbol{y}}} dt$. The *geodesic distance* between points $\boldsymbol{y}_0, \boldsymbol{y}_1 \in \mathcal{M}$ is the length of the shortest piecewise smooth curve between the two points:

$$d_g(\boldsymbol{y}_0, \boldsymbol{y}_1) = \inf_{\Gamma}\{\ell(\Gamma) : \Gamma(0) = \boldsymbol{y}_0, \Gamma(1) = \boldsymbol{y}_1\} .$$

Given the geodesic distance, one can construct a geodesic k-NN graph on $\mathcal{Y}_n$ by computing the nearest neighbor relations between points using d_g instead of the usual Euclidean distance. Consequently, we define the total edge length of this new graph as $L_{\gamma,k}(\mathcal{Y}_n)$, where $L_{\gamma,k}(\mathcal{Y}_n)$ is given by (1) with the correspondence $d \to d_g$.

We can now extend Theorem 1 to general compact Riemann manifolds. This extension, Theorem 2 bellow, states that the asymptotic behavior of $L_{\gamma,k}(\mathcal{Y}_n)$ 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, the induced measure on $\mathcal{M}$ via the volume element [15].

Theorem 2: Let $(\mathcal{M}, g)$ be a compact 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 $L_{\gamma,k}$ be 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,

$$\lim_{n \to \infty} \frac{L_{\gamma,k}(\mathcal{Y}_n)}{n^{\alpha}} = \beta_{m,\gamma,k} \int_{\mathcal{M}} f^{\alpha}(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y}) \qquad a.s. , \qquad (3)$$

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

Now, the integral factor in the a.s. limit of (3) is a monotonic function of the intrinsic Rényi α-entropy of the multivariate density f on $\mathcal{M}$:

$$H_{\alpha}^{(\mathcal{M},g)}(f) = \frac{1}{1-\alpha} \log \int_{\mathcal{M}} f^{\alpha}(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y}) . \qquad (4)$$

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

$$\hat{H}_{\alpha}^{(\mathcal{M},g)}(\mathcal{Y}_n) = \frac{m}{\gamma}\left[\log \frac{L_{\gamma,k}(\mathcal{Y}_n)}{n^{(m-\gamma)/m}} - \log \beta_{m,\gamma,k}\right] \qquad (5)$$

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

The intuition behind the proof of Theorem 2 comes from the fact that a Riemann manifold $\mathcal{M}$, with associated distance and measure, looks locally like $\mathbb{R}^m$ with Euclidean distance and Lebesgue measure. This implies that on small neighborhoods of the manifold the total edge length $L_{\gamma,k}(\mathcal{Y}_n)$ 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 [14] of $L_{\gamma,k}$, allows for repeated applications of Theorem 1 resulting in (3).

B. Approximating Geodesic k-NN Distances

Assume now that $\mathcal{M} \subset \mathbb{R}^d$. In the 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 the GMST algorithm [12] (or the ISOMAP [2]), this is done by running a costly optimization algorithm over a global graph of "neighborhood relations" among all points.

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 as surrogates for the corresponding true geodesic distances. In fact, we prove that the geodesic k-NN distances are uniformly well approximated by the corresponding Euclidean k-NN distances 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 $\mathcal{M}$. Then, with probability 1,

$$\max_{\substack{1 \leq i \leq n \\ \boldsymbol{Y} \in \mathcal{N}_{k,i}(\mathcal{Y}_n)}} \left| \frac{|\boldsymbol{Y} - \boldsymbol{Y}_i|}{d_g(\boldsymbol{Y}, \boldsymbol{Y}_i)} - 1 \right| \to 0 \quad \text{as } n \to \infty \,. \tag{6}$$

III. Joint Intrinsic Dimension/Entropy Estimation

Let $\hat{L}_{\gamma,k}(\mathcal{Y}_n)$ be the total edge length of the Euclidean k-NN graph over $\mathcal{Y}_n$. Its asymptotic behavior is a simple consequence of Theorems 2 and 3:

Corollary 4: Let $(\mathcal{M}, g)$ be a compact Riemann m-dimensional submanifold of $\mathbb{R}^d$. Suppose $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are i.i.d. random vectors of $\mathcal{M}$ with bounded density f relative to μ_g. Assume $m \geq 2$, $1 \leq \gamma < m$ and define $\alpha = (m - \gamma)/m$. Then,

$$\lim_{n \to \infty} \frac{\hat{L}_{\gamma,k}(\mathcal{Y}_n)}{n^\alpha} = \beta_{m,\gamma,k} \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y}) \qquad a.s.\,, \tag{7}$$

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

We are now ready to apply this result to jointly estimate intrinsic dimension and entropy. The key 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 $l_n = \log \hat{L}_{\gamma,k}(\mathcal{Y}_n)$. According to Corollary 4, l_n has the following approximation

$$l_n = a\ \log n + b + \epsilon_n\,, \tag{8}$$

where

$$\begin{aligned} a &= (m - \gamma)/m\,, \\ b &= \log \beta_{m,\gamma,k} + \gamma/m\ H_\alpha^{(\mathcal{M},g)}(f)\,, \end{aligned} \tag{9}$$

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

Using the additive model (8), 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 k-NN length functionals $\bar{L}_p = N^{-1} \sum_{j=1}^N \hat{L}_{\gamma,k}(\mathcal{Y}_p^j)$. Defining $\bar{\boldsymbol{l}} = [\log \bar{L}_{p_1}, \ldots, \log \bar{L}_{p_1}]^T$ we write down the linear vector model

$$\bar{\boldsymbol{l}} = A \left[\begin{array}{c} a \\ b \end{array} \right] + \epsilon \tag{10}$$

where

$$A = \left[\begin{array}{ccc} \log p_1 & \ldots & \log p_Q \\ 1 & \ldots & 1 \end{array} \right]^T .$$

We now take a method-of-moments (MOM) approach in which we use (10) to solve for the linear least squares (LLS) estimates $\hat{a}, \hat{b}$ of a, b followed by inversion of the relations (9). After making a simple large n approximation, this approach yields the following estimates:

$$\begin{aligned} \hat{m} &= \mathrm{round}\{\gamma/(1 - \hat{a})\} \\ \hat{H}_\alpha^{(\mathcal{M},g)} &= \frac{\hat{m}}{\gamma} \left(\hat{b} - \log \beta_{\hat{m},\gamma,k} \right)\,. \end{aligned} \tag{11}$$

The importance of constants $\beta_{m,\gamma,k}$ is different whether dimension or entropy estimation is considered. On one hand, due to the slow growth of $\{\beta_{m,\gamma,k}\}_{m>0}$ in the large n regime for which the above estimates were derived, $\beta_{m,\gamma,k}$ is not required for the dimension estimator. On the other hand, the value of $\beta_{m,\gamma,k}$ is required for the entropy estimator to be unbiased. From the proof of Theorem 2, it comes out that $\beta_{m,\gamma,k}$ is the limit of the normalized length functional of the Euclidean k-NN graph for a uniform distribution on the unit cube $[0,1]^m$. As closed form expressions are not available, this constant must be determined by Monte Carlo simulations of the k-NN length on the corresponding unit cube for uniform random samples. We note, however, that in many applications all that is required is the knowledge of the entropy up to a constant. For example, when maximum or minimum entropy is used as a discriminant on several data sets [11], only the relative ordering of the entropies is important.

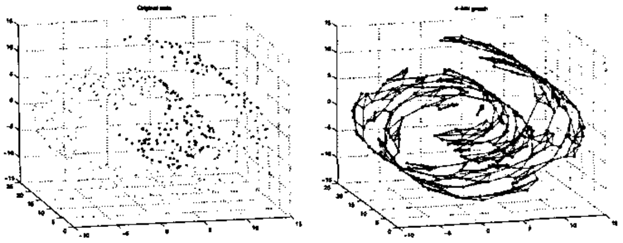

Fig. 1. The Swiss roll manifold and corresponding 4-NN graph on 400 sample points.

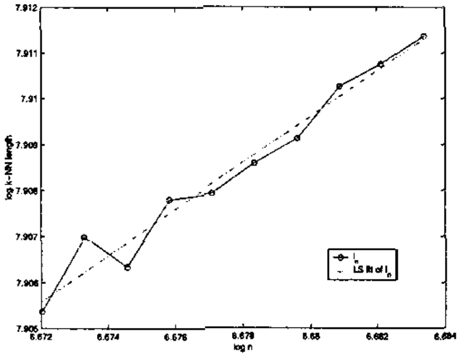

Fig. 2. Log-log plot of the average k-NN length $\bar{L}_n$ for the Swiss roll manifold and its least squares linear fit, for $n = 800$ sample points, $k = 3$ and $N = 5$. The estimated slope is $\hat{a} = 0.504$ which implies $\hat{m} = 2$.

Finally, the complexity of the algorithm is dominated by the search of nearest neighbors in the Euclidean metric. Using efficient constructions such as K-D trees, this task can be performed in $O(n \log n)$ time for n sample points. This contrasts with both the GMST and ISOMAP that require a costly $O(n^2 \log n)$ implementation of a geodesic pairwise distance estimation step.

IV. Experimental Results

We illustrate the performance of the proposed k-NN algorithm on manifolds of known dimension. In all the simulations we used $\gamma = 1$ and $p_1 = n - Q, \ldots, p_Q = n - 1$. With regards to intrinsic dimension estimation, we compare our algorithm to ISOMAP. In ISOMAP, similarly to PCA, intrinsic dimension is usually estimated by looking at the residual errors as a function of subspace dimension.

A. Swiss Roll

The first manifold considered is the standard 2-dimensional Swiss roll surface [2] embedded in $\mathbb{R}^3$ (Fig. 1). Fig. 2 shows a log-log plot of the average k-NN length $\bar{L}_n$ as a function of the number of samples. The good agreement between $\bar{L}_n$ and its least squares linear fit confirms the large sample behavior predicted by Corollary 4 and shows evidence in favor of linear model (8).

To compare the dimension estimation performance of the k-NN method to ISOMAP we ran a Monte Carlo simulation. For each of several sample sizes, 30 independent sets of i.i.d. random vectors uniformly distributed on the surface were generated. We then counted the number of times that the intrinsic dimension was correctly estimated. To automatically estimate dimension with ISOMAP, we look at its eigenvalue residual variance plot and try to detect the "elbow" at which residuals cease to decrease "significantly" as estimated dimension increases [2]. This is implemented by a simple minimum angle threshold rule. Table I shows the results of this experiment. As it can be observed, the k-NN algorithm outperforms ISOMAP for small sample sizes.

B. Hyper-spheres

A more challenging problem is the case of the m-dimensional sphere S^m (embedded in $\mathbb{R}^{m+1}$). This manifold does not satisfy any of the usual isometric or conformal embedding constraints required by ISOMAP or other methods like C-ISOMAP [16] and Hessian eigenmap [4]. Once again, we tested the algorithm over 30 generations of uniform random samples over S^m, for $m = 2, 3, 4$ and 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 $m + 1$. The results for k-NN are shown in Table II for different values of the parameter Q. As it can be seen, the k-NN method succeeds in finding the correct intrinsic dimension. However, Table II also shows that the number of samples required to achieve the same level of accuracy increases with the manifold dimension. This is the usual curse of dimensionality phenomenon: as the dimension increases, more samples are needed for the asymptotic regime in (7) to settle in and validate the limit in Corollary 4.

C. Hyper-planes

We also investigate m-dimensional hyper-planes in $\mathbb{R}^{m+1}$ for which PCA methods are designed. We consider hyper-planes of the form $x_1 + \ldots + x_{m+1} = 0$. Table III shows the results of running a Monte Carlo simulation under the same conditions as in the previous subsections. Unlike the ISOMAP,

TABLE I
NUMBER OF CORRECT DIMENSION ESTIMATES OVER 30 TRIALS AS A FUNCTION OF THE NUMBER OF SAMPLES FOR THE SWISS ROLL MANIFOLD.

n	200	400	600
ISOMAP ($k = 7$)	18	29	30
3-NN ($N = 5, Q = 9$)	29	30	30

TABLE II
NUMBER OF CORRECT DIMENSION ESTIMATES OVER 30 TRIALS AS A FUNCTION OF THE NUMBER OF SAMPLES FOR HYPER-SPHERES, $k = 5$ NEIGHBORS, $N = 5$.

Sphere	n	600	800	1000	1200
S^2	$Q = 9$	30	30	30	30
S^3	$Q = 9$	27	27	28	28
S^3	$Q = 19$	29	30	30	30
S^4	$Q = 9$	23	26	26	26
S^4	$Q = 19$	28	30	30	30

TABLE III

NUMBER OF CORRECT DIMENSION ESTIMATES OVER 30 TRIALS AS A FUNCTION OF THE NUMBER OF SAMPLES FOR HYPER-PLANES, $k = 7$ NEIGHBORS.

Hyper-plane dimension	n	600	800	1000	1200
2	$N = 5, Q = 9$	30	30	30	30
3	$N = 5, Q = 9$	27	27	28	28
3	$N = 10, Q = 14$	30	30	30	30
4	$N = 10, Q = 14$	22	23	26	26
4	$N = 10, Q = 19$	24	26	28	28

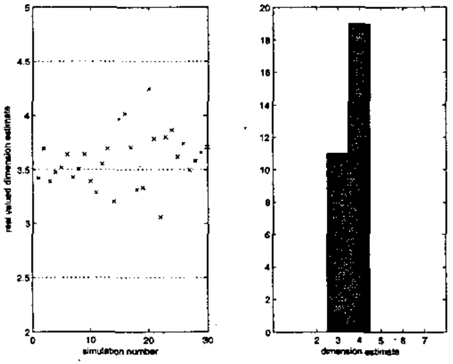

Fig. 3. Real valued intrinsic dimension estimates and histogram for the 4-D hyper-plane. for $n = 400$ sample points, $k = 5$, $N = 10$ and $Q = 14$.

which was observed to correctly predict the dimension for all sample sizes investigated, the k-NN method has a tendency to underestimate the correct dimension at smaller sample sizes. This fact can be observed in Fig. 3. The first column shows the real valued estimates of the intrinsic dimension, i.e., estimates obtained before the rounding operation in (11). Any value that falls in between the dashed lines will then be rounded to the middle point. The second column of Fig. 3 shows the histogram for these rounded estimates over the 30 simulations trial. We believe that the resampling strategy of the algorithm may be responsible for this underestimation. Several methods for improving the performance of the k-NN algorithm are currently under investigation.

D. Full Dimensional Uniform Samples on the Unit Cube

Finally, we consider uniformly distributed samples on the full dimensional unit cube $[0, 1]^d \subset \mathbb{R}^d$. The results summarized by Table IV are similar to those for hyper-planes in the previous subsection. ISOMAP correctly estimated the dimensionality of the data for all sample sizes.

V. CONCLUSION

We have introduced a novel method for intrinsic dimension and entropy estimation based on the growth rate of the Euclidean k-NN graph length functional. The proposed algorithm is applicable to a wider class of manifolds than previous methods and has reduced computational complexity. We have validated the new method by testing it on synthetic manifolds of known dimension.

TABLE IV

NUMBER OF CORRECT DIMENSION ESTIMATES OVER 30 TRIALS AS A FUNCTION OF THE NUMBER OF SAMPLES FOR FULL DIMENSIONAL UNIFORM DISTRIBUTION, $k = 7$ NEIGHBORS

Unit cube	n	600	800	1000	1200
$[0, 1]^3$	$N = 5, Q = 9$	26	27	27	27
$[0, 1]^3$	$N = 10, Q = 14$	30	30	30	30
$[0, 1]^4$	$N = 10, Q = 14$	24	25	26	26
$[0, 1]^4$	$N = 10, Q = 19$	27	28	29	29

In order to improve the performance of the derived estimators, a better understanding of the statistics of the error term in the linear model (8) would be important. Also of great interest is the study of the effect of additive noise on the manifold samples. With regards to applications, we plan to test the proposed algorithm on databases of faces, handwritten digits and genetic data.

ACKNOWLEDGMENT

The work presented here was partially funded by NIH 1PO1 CA87634-01 and NSF CCR-0325571.

REFERENCES

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

[2] 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.

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

[4] D. Donoho and C. Grimes, 'Hessian eigenmaps: new locally linear embedding techniques for high dimensional data," Tech. Rep. TR2003-08, Dept. of Statistics, Stanford University, 2003.

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

[6] T. Cox and M. Cox, *Multidimensional Scaling*, Chapman & Hall, London, 1994.

[7] X. Huo and J. Chen, 'Local linear projection (LLP)," in *Proc. of First Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, 2002.

[8] P. Verveer and R. Duin, 'An evaluation of intrinsic dimensionality estimators," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, no. 1, pp. 81–86, January 1995.

[9] 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.

[10] B. K´egl, 'Intrinsic dimension estimation using packing numbers," in *Neural Information Processing Systems: NIPS*, Vancouver, CA, Dec. 2002.

[11] 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.

[12] J. A. Costa and A. O. Hero, 'Geodesic minimal spanning trees for dimension and entropy estimation in manifold learning," *IEEE Trans. on Signal Processing*, 2003, under revision.

[13] J. A. Costa and A. O. Hero, 'Manifold learning using Euclidean k-neartest neighbor graphs," in preparation, 2003.

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

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

[16] V. de Silva and J. B. Tenenbaum, 'Global versus local methods in nonlinear dimensionality reduction," in *Neural Information Processing Systems 15 (NIPS)*, Vancouver, Canada, Dec. 2002.