MANIFOLD LEARNING USING EUCLIDEAN K-NEAREST NEIGHBOR GRAPHS

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

In the manifold learning problem one seeks to discover a smooth low dimensional surface, i.e., a manifold embedded in a higher dimensional linear vector space, based on a set of n measured sample points on the surface. In this paper we consider the closely related problem of estimating the manifold's intrinsic dimension and the intrinsic entropy of the sample points. Specifically, we view the sample points as realizations of an unknown multivariate density supported on an unknown smooth manifold. In previous work we introduced a geometric probability method called Geodesic Minimal Spanning Tree (GMST) to obtain asymptotically consistent estimates of manifold dimension and entropy. In this paper we present a simpler method based on the k-nearest neighbor (k-NN) graph that does not require estimation of geodesic distances on the manifold. The algorithm is applied to standard synthetic manifolds as well as real data sets consisting of images of faces.

1. INTRODUCTION

Consider a class of natural occurring signals, e.g., recorded speech, audio, images, or videos. Such signals typically have high extrinsic dimension, e.g., as characterized by the number of pixels in an image or the number of time samples in an audio waveform. However, most natural signals have smooth and regular structure, e.g. piecewise smoothness, that permits substantial dimension reduction with little or no loss of content information.

A useful representation of a regular signal class is to model it as a set of vectors which are constrained to a smooth low dimensional manifold embedded in a high dimensional vector space. A problem of substantial recent interest in machine learning, computer vision, signal processing and statistics [1–5] is the determination of the so-called *intrinsic dimension* of the manifold and the reconstruction of the manifold from a set of samples from the signal class. This problem falls in the area of *manifold learning* which is concerned with discovering low dimensional structure in high dimensional data. 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.

The goal of this paper is to introduce an algorithm that jointly estimates both the intrinsic dimension and intrinsic entropy given just a set of random sample points on the manifold. We construct the Euclidean 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 is similar to the GMST method introduced by us in previous work [6], in that it does not require reconstructing the manifold or estimating the multivariate density of the samples. However, the k-NN method has the main advantage of reducing complexity by one order of magnitude and is applicable to a wider class of manifolds.

The work presented here was partially supported by the National Institutes of Health through grant NIH 1PO1 CA87634-01.

2. THE EUCLIDEAN K-NN GRAPH ON A MANIFOLD

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

$$\arg \max_{\boldsymbol{X} \in \mathcal{X}_n \setminus \{\boldsymbol{X}_i\}} |\boldsymbol{X} - \boldsymbol{X}_i| \,,$$

where $|\boldsymbol{X} - \boldsymbol{X}_i|$ is the usual Euclidean (L_2) distance in $\mathbb{R}^d$ between vector $\boldsymbol{X}$ and $\boldsymbol{X}_i$. 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:

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

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

2.1. Convergence to Extrinsic α-Entropy

The k-NN edge length lies in the large class of functionals called continuous quasi-additive Euclidean functionals [7]. Other graphs in this class include the minimal spanning tree, the minimal matching graph or the traveling salesman tour among others. These functionals have remarkable asymptotic behavior as n increases:

Theorem 1 ([7, 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, with probability 1 (w.p.1)*

$$\lim_{n \to \infty} \frac{L_{\gamma,k}(\mathcal{X}_n)}{n^{\alpha}} = \beta_{d,\gamma,k} \int f^{\alpha}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \,, \tag{2}$$

where $\beta_{d,\gamma,k}$ 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 quantity that determines the limit (2) in Theorem 1 is 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.

Consider now a set of i.i.d. random vectors $\mathcal{Y}_n = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ that are constrained to lie on a compact smooth m-dimensional submanifold of $\mathbb{R}^d$ ($m < d$). In this case, the distribution of $\boldsymbol{Y}_i$ is singular with respect to Lebesgue measure, resulting in a zero limit for the right hand side of (2). However, this does not imply that the limit is zero when using a different power of n as a normalization factor. This key observation is the basis for the use of the k-NN graph for dimension and entropy estimation on manifolds.

2.2. Convergence to Intrinsic α-Entropy

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 $\boldsymbol{y}$ [8]. 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. The Riemann metric g also induces a measure μ_g on $\mathcal{M}$ via the differential volume element.

We can now state a similar result to Theorem 1 for compact Riemann manifolds with *intrinsic* dimension $m < d$ (see [9] for a proof).

Theorem 2 *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, w.p.1,*

$$\lim_{n \to \infty} \frac{L_{\gamma,k}(\mathcal{Y}_n)}{n^{(d'-\gamma)/d'}} = \begin{cases} \infty, & d' < m \\ \beta_{m,\gamma,k} \int_{\mathcal{M}} f^{\alpha}(\boldsymbol{y}) \, \mu_g(\mathrm{d}\boldsymbol{y}), & d' = m \\ 0, & d' > m \end{cases} , \tag{4}$$

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.

Theorem 2 shows 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 quantity that determines the non-zero finite limit in (4) is the *intrinsic* Rényi α-entropy of the multivariate density f on $\mathcal{M}$:

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

3. ESTIMATING INTRINSIC DIMENSION AND ENTROPY

Theorem 2 is the theoretical basis for developing a consistent estimator of both 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. In particular, the only way to obtain a non-zero finite limit in (4) is by normalizing the length functional by the right power α of n, i.e., $\alpha = (m - \gamma)/m$ when $d' = m$. We use this strong growth dependence as a motivation for a simple estimator of m. Define $l_n = \log L_{\gamma,k}(\mathcal{Y}_n)$. According to (4), l_n has the following approximation

$$l_n = a \, \log n + b + \epsilon_n \, , \tag{6}$$

where

$$\begin{aligned} a &= (m - \gamma)/m \, , \\ b &= \log \beta_{m,\gamma,k} + \gamma/m \, H_{\alpha}^{(\mathcal{M},g)}(f) \, , \end{aligned} \tag{7}$$

and ϵ_n is an error residual that goes to zero w.p.1 as $n \to \infty$.

Using the additive model (6), we propose a simple nonparametric 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 $\bar{L}_p = N^{-1} \sum_{j=1}^{N} L_{\gamma,k}(\mathcal{Y}_p^j)$. Defining $\bar{\boldsymbol{l}} = [\log \bar{L}_{p_1}, \ldots, \log \bar{L}_{p_1}]^T$ we write down the linear vector model

$$\bar{\boldsymbol{l}} = A \begin{bmatrix} a \\ b \end{bmatrix} + \epsilon \tag{8}$$

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 (8) to solve for the linear least squares (LLS) estimates $\hat{a}, \hat{b}$ of a, b followed by determination of $\hat{m}$ and $\hat{H}$ by inversion of the relations (7). After making a simple large n approximation, this approach yields the following estimates:

$$\begin{aligned} \hat{m} &= \mathrm{round}\{\gamma/(1 - \hat{a})\} \\ \hat{H}_{\hat{\alpha}}^{(\mathcal{M},g)} &= \frac{\hat{m}}{\gamma} \left(\hat{b} - \log \beta_{\hat{m},\gamma,k} \right) \, . \end{aligned} \tag{9}$$

We now discuss the role of the constants $\beta_{m,\gamma,k}$ in the above estimators. First of all, 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.

Finally, the complexity of the algorithm is dominated by determining nearest neighbors, which can be done in $O(n \log n)$ time for n sample points. This contrasts with both the GMST and ISOMAP [1] that require a costly $O(n^2 \log n)$ implementation of a geodesic pairwise distance estimation step.

4. APPLICATIONS

We applied the proposed algorithm to manifolds of known structure as well as a real data set consisting of faces images. In all the simulations we used $p_1 = n - Q, \ldots, p_Q = n - 1$. With regards to intrinsic dimension estimation, we compare our algorithm

Table 1. Number of correct dimension estimates over 30 trials as a function of the number of samples, for $k = 5$ neighbors, $N = 5$.

Sphere	n	600	800	1000	1200
S^2	$Q = 10$	30	30	30	30
S^3	$Q = 10$	27	27	28	28
S^3	$Q = 20$	29	30	30	30
S^4	$Q = 10$	23	26	26	26
S^4	$Q = 20$	28	30	30	30

(a)

(b)

Fig. 1. Samples from : (a) ISOMAP face database; (b) Yale face database B.

to ISOMAP. In ISOMAP, similarly to PCA, intrinsic dimension is usually estimated by looking at the residual errors as a function of subspace dimension.

We have validated the algorithm on standard synthetic manifolds in the literature: linear planes in several dimensions, the 2-dimensional swiss roll [1] and S-shaped surface [2] embedded in $\mathbb{R}^3$. Due to space limitations we will not present these results here.

Of greater interest is the case of the m-dimensional sphere S^m (embedded in $\mathbb{R}^{m+1}$). This is a more challenging problem, as the sphere does not satisfy any of the usual isometric or conformal embedding constraints required by ISOMAP or several other methods like C-ISOMAP [10] or Hessian eigenmap [3]. We ran 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 times that the intrinsic dimension was correctly estimated. 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 1 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 1 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 (4) to settle in and validate the estimator.

Next, we applied our method to a high dimensional synthetic image data set. For this purpose we used the ISOMAP face database [1]. This set consists of 698 images of the same face generated by varying three different parameters: vertical and horizontal pose, and lighting direction. Each image has 64×64 pixels with 256 gray levels, normalized between 0 and 1 (Figure 1.a). For processing, we embedded each image in the 4096-dimensional Euclidean space using the common lexicographic order. We applied the algorithm 30 times over the data set with results displayed in Figure 2. The first column shows the real valued estimates of the intrinsic dimension, i.e., estimates obtained before the rounding operation in (9). Any value that falls in between the dashed lines will then be rounded to the middle point. The second column of Figure 2 shows the histogram for these rounded estimates over the 30 simulations trial. The estimated intrinsic dimension oscillates between 3 and 4, which, as in [5], deviates from the "informal" intrinsic dimension of 3 estimated by ISOMAP.

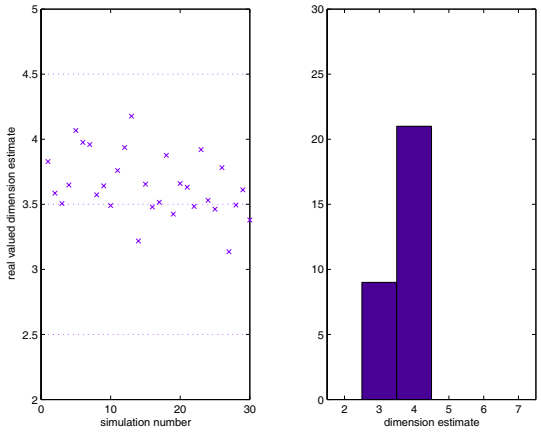

Fig. 2. Real valued intrinsic dimension estimates and histogram for the ISOMAP face database. $k = 5$, $N = 10$, $Q = 15$.

Finally, we applied the k-NN method to a real data set, and, consequently, of unknown manifold structure and intrinsic dimension. We chose the set of 256 gray levels images of several individuals taken from the Yale Face Database B [11]. This is a publicly available database[1] containing face images of 10 subjects with 585 different viewing conditions for each subject (Figure 1.b). These consist of 9 poses and 65 illumination conditions (including ambient lighting). The images were taken against a fixed background which we did not bother to segment out. We think this is justified since any fixed structures throughout the images would not change the intrinsic dimension or the intrinsic entropy of the dataset. We randomly selected 4 individuals from this data base and subsampled each person's face images down to a 64×64 pixels image. Similarly to the ISOMAP face data set, we normalized the pixel values between 0 and 1. Figures 3 and 4 display the results of running 30 simulations of the algorithm using face 1 and face 2, respectively. The intrinsic dimension estimate is between 5 and 6 for face 1 and is clearly 5 for face 2. Figure 5 shows the corresponding residual variance plots used by ISOMAP to estimate intrinsic dimension. From these plots it is not obvious how to determine the "elbow" at which the residuals cease to decrease "significantly" with added dimensions. This illustrates one of the major drawbacks of ISOMAP (and other spectral based methods like PCA) as an intrinsic dimension estimator, as it relies on a specific eigenstructure that may not exist in real data. A simple minimum angle threshold rule on ISOMAP produced estimates between 4 and 8 for face 1 and 4 and 7 for face 2. Table 2 summarizes the results of the k-NN method for the four faces, where the last column shows the results of processing two faces simultaneously. As it can

[1] `http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html`

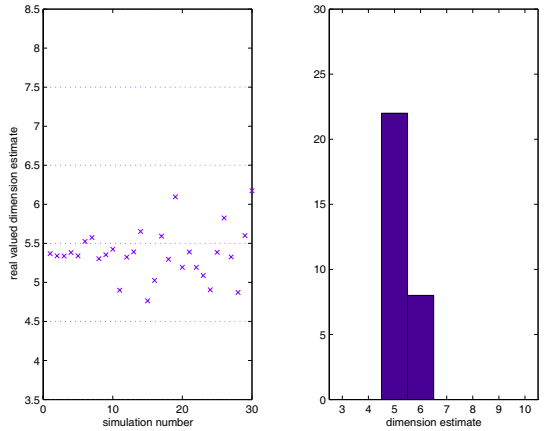

Fig. 3. Real valued intrinsic dimension estimates and histogram for face 1 in the Yale face database B. $k = 3$, $N = 10$, $Q = 20$.

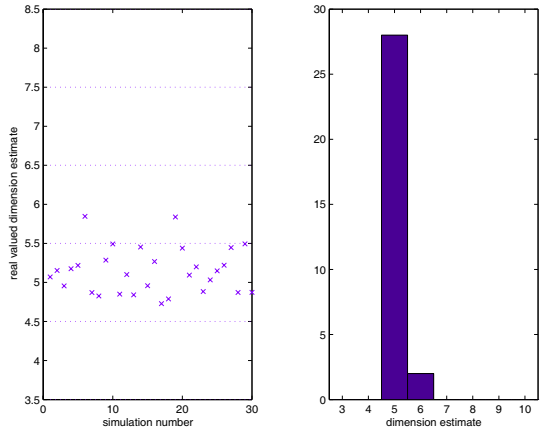

Fig. 4. Real valued intrinsic dimension estimates and histogram for face 2 in the Yale face database B. $k = 3$, $N = 10$, $Q = 15$.

be seen, the joint dimensionality of the two faces is determined by the dimension of the most complex one, while the entropy grows roughly by one bit. This should be expected, as compressing the augmented dataset requires only one extra bit to identify which face is being coded.

5. CONCLUSION

We have presented a new method for intrinsic dimension and entropy estimation on Riemann compact manifolds. Its key features are its applicability to a wide class of manifolds, ability to produce consistent estimates for both synthetic and real data, and reduced complexity. Future work includes implementing bootstrap confidence intervals for the estimators and study of the effect of additive noise on the manifold samples.

6. REFERENCES

[1] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 1, pp. 2319–2323, 2000.

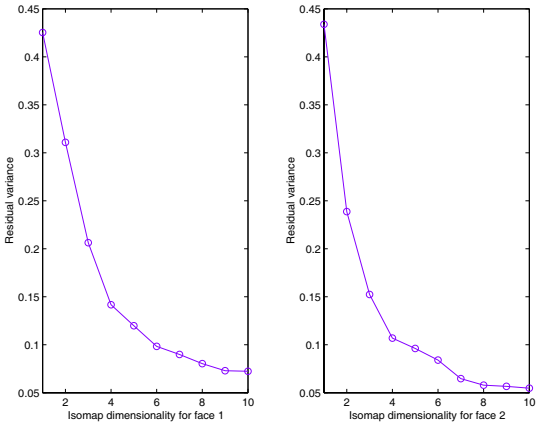

Fig. 5. ISOMAP ($k = 6$) residual variance for face 1 and face 2 in the Yale face database B.

Table 2. Dimension estimates $\hat{m}$ and entropy estimates $\hat{H}$ for four faces in the Yale Face Database B.

	Face 1	Face 2	Face 3	Face 4	Face 1 + 3
$\hat{m}$	5	5	6	6	6
$\hat{H}$ (bits)	20.8	22.9	20.3	24.0	21.8

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

[3] 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.

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

[5] B. Kégl, "Intrinsic dimension estimation using packing numbers," in *Neural Information Processing Systems 15 (NIPS)*, Vancouver, Canada, Dec. 2002.

[6] J. A. Costa and A. O. Hero III, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *IEEE Trans. on Signal Processing*, to appear, 2004.

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

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

[9] J. A. Costa and A. O. Hero III, "Manifold learning using Euclidean k-nearest neighbor graphs," in preparation, 2003.

[10] 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.

[11] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.