Estimating Local Intrinsic Dimension with k-Nearest Neighbor Graphs

Jose A. Costa, Abhishek Girotra and Alfred O. Hero III
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI 48109
Emails: {jcosta, agirotra, hero}@umich.edu

***Abstract*— Many high-dimensional data sets of practical interest exhibit a varying complexity in different parts of the data space. This is the case, for example, of databases of images containing many samples of a few textures of different complexity. Such phenomena can be modeled by assuming that the data lies on a collection of manifolds with different intrinsic dimensionalities. In this extended abstract, we introduce a method to estimate the local dimensionality associated with each point in a data set, without any prior information about the manifolds, their quantity and their sampling distributions. The proposed method uses a global dimensionality estimator based on k-nearest neighbor (k-NN) graphs, together with an algorithm for computing neighborhoods in the data with similar topological properties.**

***Index Terms*— Manifold learning, Intrinsic dimension, Nearest neighbor graph.**

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. Today's medical information systems or video surveillance applications, for example, are producing signals that are high-dimensional in their nature and thus appear to be very complex. However, such signals 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. In the recent past, this subject has received substantial attention from researchers in machine learning, computer vision and statistics, leading to the introduction of several manifold learning algorithms (see webpage [1] for an extensive list of references).

Playing a central role in the analysis of high-dimensional data is its *intrinsic dimensionality*, given by the the dimension of the manifold supporting the data. Intuitively, this quantity describes how many "degrees of freedom" are necessary to describe the data set. When the intrinsic dimension is assumed constant over the data set, several algorithms have been proposed recently to estimate it directly from only a finite sampling of the manifold. These range from fractal dimension [2], estimating packing numbers [3], entropic graphs [4], [5] or maximum likelihood approach [6]. However, in several problems of practical interest, data will exhibit varying dimensionality across the observed data set. For example, in the protein docking problem [7], the degrees of freedom associated with the allowed movements of the reacting molecules will change during the reaction time.

In this paper, we introduce a method to estimate the local dimensionality associated with each point in a data set. If the data set is sampled from a union of disjoint manifolds, with possible different intrinsic dimensionalities, then the algorithm estimates, for each sample point, the dimension of the local manifold where it is supported. The proposed method uses a previously introduced global dimensionality estimator [5] based on k-nearest neighbor (k-NN) graphs, together with an algorithm for computing neighborhoods in the data with similar topological properties.

II. The k-Nearest Neighbor Graph and Global Dimension Estimation

Let $\mathcal{Y}_n = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ 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 $\boldsymbol{Y}_i$ in $\mathcal{Y}_n$ is given by

$$\arg \min_{\boldsymbol{Y} \in \mathcal{Y}_n \setminus \{\boldsymbol{Y}_i\}} |\boldsymbol{Y} - \boldsymbol{Y}_i| \ ,$$

where $|\boldsymbol{Y} - \boldsymbol{Y}_i|$ is the usual Euclidean (L_2) distance in $\mathbb{R}^d$ between vector $\boldsymbol{Y}$ and $\boldsymbol{Y}_i$. 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 $\mathcal{Y}_n$ and its k-nearest neighbors. Let $\mathcal{N}_{k,i} = \mathcal{N}_{k,i}(\mathcal{Y}_n)$ be the set of k-nearest neighbors of $\boldsymbol{Y}_i$ in $\mathcal{Y}_n$. The total edge length of the k-NN graph is defined as:

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

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

For many data sets of interest, the random vectors $\mathcal{Y}_n$ are constrained to lie on a m-dimensional Riemannian submanifold $\mathcal{M}$ of $\mathbb{R}^d$ ($m < d$). A Riemann manifold has an associated metric g [8], which endows $\mathcal{M}$ with both a notion of distance via geodesics and also a measure μ_g via the differential volume element. Under this framework, the asymptotic behavior of (1) is given by the following theorem [5]:

This research was partially supported by NSF contract CCR-0325571.

Theorem 1: *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, with probability 1,*

$$\lim_{n\to\infty} \frac{L_{\gamma,k}(\mathcal{Y}_n)}{n^{(d'-\gamma)/d'}} = \begin{cases} \infty, & d' < m \\ \beta_{m,\gamma,k} \int_{\mathcal{M}} f^{\alpha}(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y}), & d' = m \\ 0, & d' > m \end{cases} , \tag{2}$$

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 1 provides the basis for developing a consistent estimator of the intrinsic dimensionality m of data set $\mathcal{Y}_n$. On the one hand, the growth rate of the length functional is strongly dependent on m. In particular, the only way to obtain a nonzero finite limit in (2) is by normalizing the length functional by the right power α of n, i.e., $\alpha = (m-\gamma)/m$ when $d' = m$. On the other hand, that nonzero finite limit is determined by the *intrinsic* Rényi α-entropy of the multivariate density f on $\mathcal{M}$:

$$H_{\alpha}^{(\mathcal{M},g)}(f) = \frac{1}{1-\alpha} \log \int_{\mathcal{M}} f^{\alpha}(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y}) \ . \tag{3}$$

These observations motivate the following estimator for m. Define $l_n = \log L_{\gamma,k}(\mathcal{Y}_n)$. According to (2), l_n has the following approximation

$$l_n = a\ \log n + b + \epsilon_n \ , \tag{4}$$

where

$$\begin{aligned} a &= (m-\gamma)/m \ , \\ b &= \log \beta_{m,\gamma,k} + \gamma/m\ H_{\alpha}^{(\mathcal{M},g)}(f) \ , \end{aligned} \tag{5}$$

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

Using the additive model (4), a simple nonparametric least squares strategy based on subsampling from the population $\mathcal{Y}_n$ of points in $\mathcal{M}$ can be adopted. 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 $\bar{\boldsymbol{l}} = [\log \bar{L}_{p_1}, \ldots, \log \bar{L}_{p_1}]^T$, write down the linear vector model

$$\bar{\boldsymbol{l}} = A \begin{bmatrix} a \\ b \end{bmatrix} + \epsilon \tag{6}$$

where

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

Now, taking a method-of-moments (MOM) approach, in which (6) is used to solve for the linear least squares (LLS) estimates $\hat{a}, \hat{b}$ of a, b, $\hat{m}$ and $\hat{H}$ can be determined by inversion of the relations (5). After making a simple large n approximation, this approach yields the following estimates:

$$\begin{aligned} \hat{m} &= \text{round}\{\gamma/(1-\hat{a})\} \\ \hat{H}_{\hat{\alpha}}^{(\mathcal{M},g)} &= \frac{\hat{m}}{\gamma}\left(\hat{b} - \log \beta_{\hat{m},\gamma,k}\right) \ . \end{aligned} \tag{7}$$

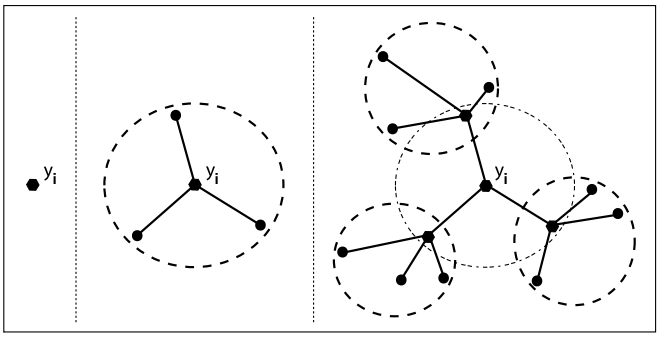


Fig. 1. Building local neighborhoods. From left to right: start with point $\boldsymbol{y}_i$; find its 3-NN points; for each of the NN points just found, compute their 3-NN points.

III. Local Intrinsic Dimension Estimation

Let $\{\mathcal{M}_1, \ldots, \mathcal{M}_P\}$ be a collection of disjoint compact Riemann submanifolds of $\mathbb{R}^d$ and define $\mathcal{M} = \cup_{j=1}^{P} \mathcal{M}_j$. Each manifold $\mathcal{M}_j$ has unknown intrinsic dimension $m_j \geq 2$, which may be different from manifold to manifold. Let f_i be the density (with respect to μ_{g_i}) of the samples on each manifold.

Given a set of n samples $\mathcal{Y}_n \in \mathcal{M}$, the goal is to estimate the local dimension associated with each sample $\boldsymbol{Y}_i$, i.e., the dimension of manifold $\mathcal{M}_j$ where $\boldsymbol{Y}_i$ lies. Of course, this has to be accomplished without any prior knowledge on the number of different manifolds, intrinsic dimensions, sampling distribution or segmentation of the data. If the segmentation of the data set according to local manifolds was known in advance, then repeated applications of Theorem 1 to each manifold segment would yield consistent estimates for each point. However, such information is not available and local neighborhoods with similar geometric structure have to be automatically determined from the data. We propose the following general algorithm (see Figure 1):

`for` $i = 1$ `to` n `do`

1. Grow a local k-NN graph for $\boldsymbol{y}_i$:
   a) `initialize` $\mathcal{N} = \{\boldsymbol{y}_i\}$,
   b) for all $y \in \mathcal{N}$ compute the set of its k-nearest neighbors, $\mathcal{N}_{k,\boldsymbol{y}}(\mathcal{Y}_n)$.
   $\mathcal{N} \leftarrow \cup_{\boldsymbol{y}\in\mathcal{N}} \mathcal{N}_{k,\boldsymbol{y}}(\mathcal{Y}_n)$;
   c) `goto` b) `until` stopping criterion is met.
2. Apply the estimation algorithm described in Section II to the graph built in step 1, and obtain a local dimension estimate $\hat{m}(\boldsymbol{y}_i)$.

`end`.

The challenging part of the algorithm described above is the selection of a criterion that stops the growing of the local

k-NN graph. On the one hand, the graph should be small enough so that only the geometry of the local manifold where sample point $\boldsymbol{y}_i$ lies is captured by the graph. On the other hand, the graph should include enough samples so that the asymptotic regime described by Theorem 1 is valid, resulting in statistically consistent estimates. Any stopping rule should take into account this tradeoff between local geometry and asymptotic consistency. We propose an heuristic rule based on the geometric and asymptotic properties of k-NN graphs.

The k-NN graph satisfies certain geometric properties, like *subadditivity* and *superadditivity* [9], which imply that the graph can be approximately computed in a greedy fashion in the following way. First partition $\mathbb{R}^d$ into a finite number of disjoint sets. Then, build a k-NN graph on the samples that fall on each disjoint set and compute its total edge length functional. Summing all contributions from each total edge length functional provides a good approximation for the global value of the functional, as long as the number of samples falling on each individual partition set is significant. According to [10], the number of samples that minimizes upper bounds on the convergence rate of (2) to its asymptotic limit is roughly of order $O\left(n^{1/d}\right)$. According to this result, a simple stopping rule can then be to grow the local k-NN graph until it incorporates a total of $O\left(n^{1/d}\right)$ sample points.

We are currently studying other stopping rules based on *adaptive neighborhood* graphs [11] that have provable geometric properties.

IV. Related Methods

The local dimension estimation method proposed here is conceptually related to the estimation of the following functional of the density of the sample points:

$$\log \int_{B(\boldsymbol{y}_0,r)} g(f(\boldsymbol{y}))\, \mu(\mathrm{d}\boldsymbol{y}) \ , \tag{8}$$

where g is a strictly increasing function and $B(\boldsymbol{y}_0, r)$ is the ball of radius r centered at $\boldsymbol{y}_0$. Under suitable regularity conditions on f and g, using the mean value theorem results in:

$$\log \int_{B(\boldsymbol{y}_0,r)} g(f(\boldsymbol{y}))\, \mu(\mathrm{d}\boldsymbol{y}) = m_{\boldsymbol{y}_0} \log r + c + o(1) \ , \tag{9}$$

where c is a constant depending on f, g and the volume of the unit sphere and $o(1) \to 0$ when $r \to 0$. Compare equation (9) to equation (4). By choosing different functions g and radii r one can develop new estimators for the local dimensionality $m_{\boldsymbol{y}_0}$.

For example, by choosing $g(u) = 1$, then functional (8) can be estimated by the number of points falling into $B(\boldsymbol{y}_0, r)$. This is the motivation behind correlation dimension methods [3], [12]. If r is chosen adaptively according to the distance from $\boldsymbol{y}_0$ to its k-nearest neighbor, $T_k(\boldsymbol{y}_0)$, then (8) is given by k/n, the proportion of samples within a radius $T_k(\boldsymbol{y}_0)$ of $\boldsymbol{y}_0$. This is the starting point for earlier methods for estimating intrinsic dimension based on k-NN distances [13].

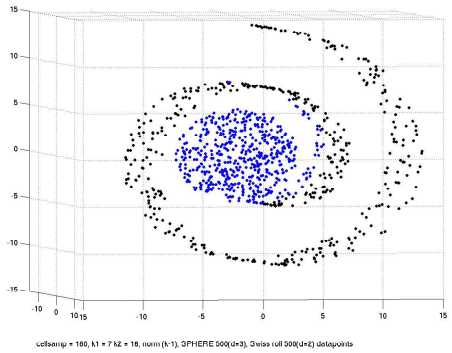


Fig. 2. Estimating the local dimension of the Swiss roll and the sphere. The estimated local dimension was 2 for the black points and 3 for the blue points.

In [6], a similar approach is followed, but the (binomial) number of points falling in $B(\boldsymbol{y}_0, T_k(\boldsymbol{y}_0))$ is approximated by a Poisson process, for samples uniformly distributed over the manifold. Then the intrinsic dimension is estimated by maximum likelihood, resulting in the following local estimate:

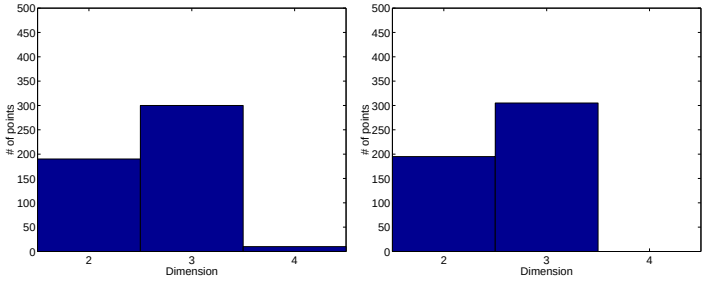
$$\hat{m}_{\boldsymbol{y}_0} = \frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(\boldsymbol{y}_0)}{T_j(\boldsymbol{y}_0)} \ .$$

V. Simulations

We now illustrate the application of the proposed method on collections of synthetic manifolds of known dimension. We compare it to the maximum likelihood (MLE) method proposed in [6] for dimension estimation.

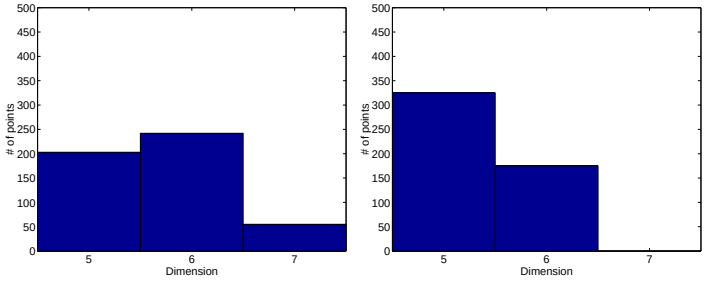
We first start with simple low-dimensional manifolds embedded in $\mathbb{R}^3$ for the purpose of visualization. Figure 2 shows the results of applying the proposed algorithm to a three-dimensional data set composed of two manifolds. This set consists of 200 hundred points sampled uniformly on the 2-dimensional "Swiss roll" and 300 points sampled uniformly on the 3-dimensional sphere. The black points have an estimated local dimension of 2, while the blue points have an estimated local dimension of 3. Figure 3 shows the histogram of the local dimension estimates. As it can be seen, almost all points were labeled with the correct dimension, except for a few that live close to the intersection of both manifolds.

The histogram of local dimension estimates obtained by the MLE method is also shown is Figure 3, where it can be observed to have a slightly better performance. This is due to the fact that that the MLE approach relies on an approximation of a binomial process by a Poisson process. This approximation converges at a rate of order $O\left(n^{-1}\right)$, as opposed to a much slower rate of order roughly $O\left(n^{-1/d}\right)$ for the graph based methods. As such, for higher dimensions, the MLE method will tend to outperform the proposed method. However, this comes at a cost, as the fast convergence rate of the MLE method is only valid for the case of sample points uniformly distributed over the manifold. When the density of the samples departs from a uniform distribution on the



(a) 4-NN graph (b) MLE method [6]

Fig. 3. Histogram of local dimension estimated for the Swiss roll + Sphere data set.



(a) 4-NN graph (b) MLE method [6]

Fig. 4. Histogram of local dimension estimated for the non-uniform 6-D hyper-sphere.

manifold, the effective convergence rate may be less than order $O\left(n^{-1}\right)$, as it will be slowed down by the variations of the distribution. This phenomenon can be observed in Figure 4 that shows the histogram of dimension estimates for a 6-dimensional hyper-sphere sampled according to a Bingham distribution [14], whose density with respect to the Lebesgue measure on the hyper-shere is

$$f(\boldsymbol{y})\,\alpha\,\exp\{\boldsymbol{y}^T\,\boldsymbol{K}\boldsymbol{y}\}\ ,$$

where $\boldsymbol{K}$ is a symmetric matrix.

Figure 5 shows similar results to the ones described previously for a data set consisting of a 3-dimensional sphere and the 2-dimensional S curve in $\mathbb{R}^3$. As it can be seen, all points were labeled with the correct dimension.

A. Complexity Segmentation

We now apply the proposed method to a synthetic image database. The goal is to classify images according to their complexity, i.e., the intrinsic dimensionality of the model used to generate them. In our simplified experiment, we generated gray scale 3×3 pixel images according to the following model. For a d-dimensional database, choose d seed pixels that will be generated independently from each other. The remaining pixels are generated according to a linear or nonlinear function of the seed pixels. For example, Figure

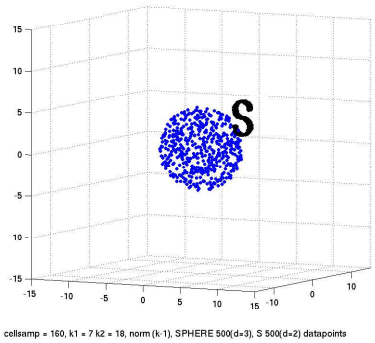


Fig. 5. Estimating the local dimension of the S curve and the sphere. The estimated local dimension was 2 for the black points and 3 for the blue points.

7(a) shows a 2-D database where the first two columns of each image are linearly dependent on the seed pixel located at the uppper rightmost corner, while the last column is a linear function of the upper leftmost corner pixel. If I_{ij} is the intensity of pixel ij, then the model is:

$$\{I_{ij}\} = \begin{bmatrix} 1 & c_{12} & 1 \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \cdot \begin{bmatrix} I_{11} & 0 & 0 \\ 0 & I_{11} & 0 \\ 0 & 0 & I_{13} \end{bmatrix} ,$$

where I_{11} and I_{13} are the independent random seeds and c_{ij} are fixed coefficients. Figure 7(b) shows a 3-D database, where each column is generated independently, according to:

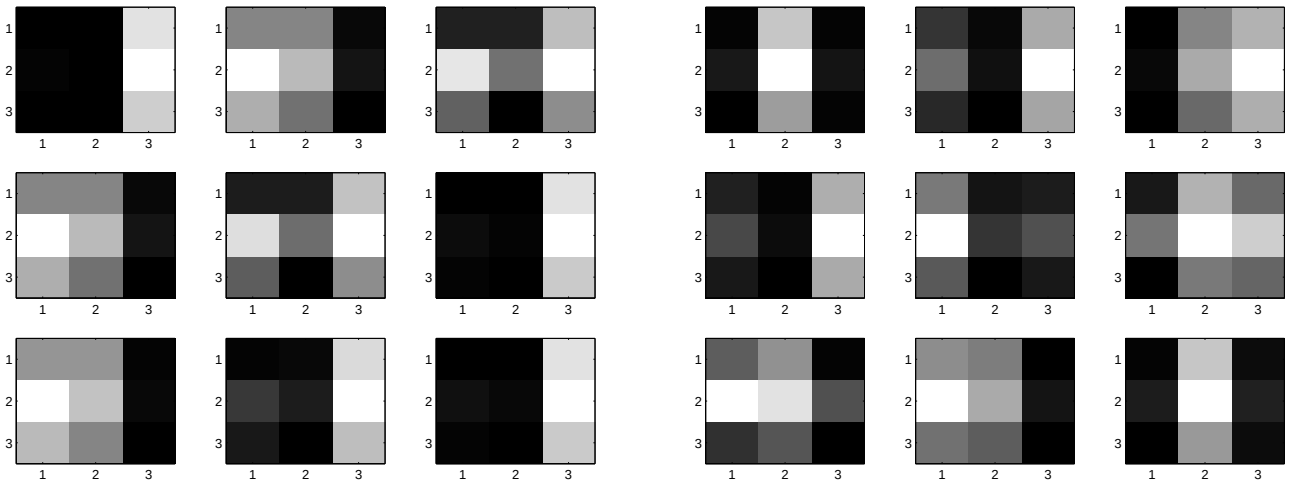
$$\{I_{ij}\} = \begin{bmatrix} 1 & 1 & 1 \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{bmatrix} \cdot \begin{bmatrix} I_{11} & 0 & 0 \\ 0 & I_{12} & 0 \\ 0 & 0 & I_{13} \end{bmatrix} ,$$

for fixed coefficients d_{ij} and independent random seeds I_{11}, I_{12} and I_{13}. The aim of these models is to simulate databases that contain images/textures with different patterns or edges, for example, which are inherently of different intrinsic dimensionality, and thus complexity.

Figure 7 shows the histograms resulting from applying the discussed methods to a database consisting of merging 400 samples of 2-D images with 400 samples of 3-D images. Unlike the MLE method, the proposed method succeeds at finding the right proportion of samples from each dimensionality. However, regarding classification rates, i.e, the number of samples whose dimensionality was correctly estimated, both methods behave similarly, with rates of correct classifications around 75%.

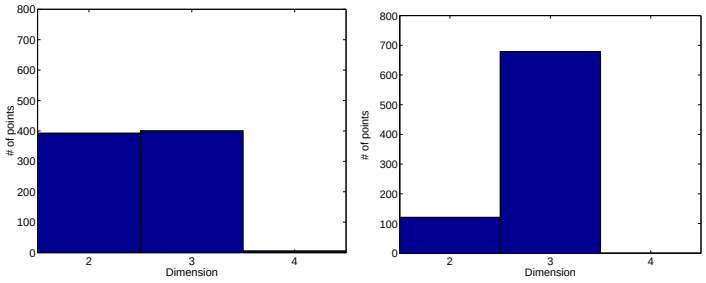
VI. Conclusions

We have introduced a new method to estimate intrinsic local dimensionality associated with each data sample. This represents the first attempt towards developing a robust non-parametric method that will be able to segment a data set into regions of different complexities. This complexities can be a product of, for example, different textures, number of edges,



(a) 2-D (b) 3-D

Fig. 6. Samples from image databases with different complexities.



(a) 4-NN graph (b) MLE method [6]

Fig. 7. Histogram of local dimension estimated for the 2-D and 3-D image databases.

etc, that impose nonlinear constraints on the data set. Several issues have to be addressed before achieving this goal.

The key block behind a local dimensionality estimator is an algorithm that finds a local adjacency graph that connects points with similar geometric properties. We are currently studying adaptive neighborhood graphs that find local neighborhoods of points that lie on the same manifold. We are also implementing a two step procedure that uses the first complexity segmentation to construct new adjacency graphs using only the points classified with the same intrinsic dimension.

Another possible improvement to the performance of the algorithm is the development of a block resampling and bootstrap procedure that will account for the dependencies among resamplings when estimating the slope in equation (4). This method might also prove useful for extending the current methodology to non i.i.d. samples. Examples of such data sets include, among others, time series obtained from Internet traffic traces.

Also of interest are applications to streaming data problems. This will require developing algorithms to compute k-NN graph neighborhoods recursively.

Finally, we are developing the asymptotic analysis necessary to guarantee the statistical consistency of the proposed method.

We remark that the problem of sampling a manifold with noise was not considered in this paper. That is a subject of future work.

REFERENCES

[1] "Manifold learning resource page," `http://www.cse.msu.edu/~lawhiu/manifold/`.

[2] 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.

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

[4] 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.

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

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

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

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

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

[10] A. Hero, J. Costa, and B. Ma, "Convergence rates of minimal graphs with random vertices," submitted to *IEEE Trans. on Inform. Theory*, 2003, `www.eecs.umich.edu/~hero/det_est.html`.

[11] J. Giesen and U. Wagner, "Shape dimension and intrinsic metric from samples of manifolds," in *Proceedings of the 19th Annual ACM Symposium on Computational Geometry*, 2003.

[12] P. Grassberger and I. Procaccia, "Measuring the stangeness of strange attractors," *Physica D*, vol. 9, pp. 189–208, 1983.

[13] K. Pettis, T. Bailey, A. Jain, and R. Dubes, "An intrinsic dimensionality estimator from near-neighbor information," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 1, no. 1, pp. 25–36, 1979.

[14] G. S. Watson, *Statistics on Spheres*, John Wiley & Sons, 1983.