LEARNING INTRINSIC DIMENSION AND INTRINSIC ENTROPY OF HIGH-DIMENSIONAL DATASETS

Jose A. Costa and Alfred O. Hero III

Department of Electrical Engineering and Computer Science, University of Michigan
1301 Beal Avenue, Ann Arbor, MI 48109-2122, USA
email: jcosta@umich.edu, hero@eecs.umich.edu, web: http://www.eecs.umich.edu/˜{jcosta,hero}

ABSTRACT

Populations of measurements of objects such as faces, genes or internet data traces, lie in lower dimensional manifolds of their high dimensional embedding spaces, e.g. face images, gene microarrays, or multivariate time series records. Knowing the intrinsic dimension and relative entropy of these manifolds is important for discovering structure, classifying differences, or performing dimensionality reduction (compression). In this paper we apply a new family of entropic graph methods to the estimation of intrinsic dimension and entropy of datasets supported on synthetic manifolds and of a high dimensional dataset of handwritten digits.

1. INTRODUCTION

Several interesting classes of signals arising in diverse fields such as bioinformatics, image processing or Internet traffic analysis cannot be characterized by low dimensional statistics. This high dimensional nature of signals as images or genome sequences, among others, makes them unsuitable to the most common processing techniques and tools. However, many real life signals that have high dimensional representations, and thus appear complex, can actually be explained by only a few degrees of freedom. This is the case of signals constrained to lie on a smooth low dimensional submanifold of a higher dimensional vector space.

Understanding the aforementioned high dimensional datasets thus requires greatly reducing the dimensionality and finding intrinsic low dimensional structure. If a simple physical model generating the data is known, parametric modeling or PCA techniques can be adopted. However, when applied to general nonparametric classes of signals, these methods will result in systematic errors. Recently, more powerful methods have been proposed in the machine learning, signal processing and statistics literature. These include ISOMAP [1], LLE [2], LLP [3] or Hessian eigenmaps [4].

When characterizing high dimensional signals, two quantities are of interest to us in this paper. One is the *intrinsic dimension* of the data, which is given by the dimensionality of the manifold supporting the data. Although all of the methods mentioned above require the intrinsic dimension as an input, it is generally unknown and has to be estimated from the data. Also of interest is the *intrinsic entropy* of the data, which characterizes statistical properties of the data distribution supported on the manifold.

In this paper, we present two methods, based on entropic graphs [5], aimed at learning the intrinsic dimension and entropy of high dimensional datasets. They work by constructing a Geodesic Minimal Spanning Tree (GMST) [6] or Euclidean k-Nearest Neighbor (k-NN) graph [7] over the sample points and their total graph weight is used to estimate the quantities of interest. We compare the performance of both methods and ISOMAP on a synthetic manifold of known characteristics. We also present preliminary results on the analysis of the MNSIT database of handwritten digits using the proposed methods.

The remainder of the paper is organized as follows. In Section 2 we define the GMST and k-NN graph and discuss their asymptotic behavior. In Section 3, the asymptotic results are used to motivate an algorithm for intrinsic dimension and entropy estimation. Experimental results are reported in Section 4.

2. ENTROPIC GRAPHS ON MANIFOLDS

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, which we will also call random vertices.

By solving certain Euclidean optimization problems on the set $\mathcal{X}_n$, one can obtain special graph constructions. For example, the Euclidean minimal spanning tree over $\mathcal{X}_n$ is the acyclic graph spanning $\mathcal{X}_n$ having minimal overall length

$$L_\gamma^{\mathrm{MST}}(\mathcal{X}_n) = \min_{T \in \mathcal{T}} \sum_{e \in T} |e|^\gamma \ , \tag{1}$$

where $\mathcal{T} = \mathcal{T}(\mathcal{X}_n)$ is the set of spanning trees over $\mathcal{X}_n$, e is an edge (e.g., $e = \boldsymbol{X}_i - \boldsymbol{X}_j, i \neq j$) in the graph T, $|e|$ is the Euclidean length of e, and $\gamma \in (0, d)$ is a power-weighting constant. Another example is the Euclidean k-NN graph. Start by defining the (1-)nearest neighbor of $\boldsymbol{X}_i \in \mathcal{X}_n$ as

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

and, for general integer $k \geq 1$, define the k-nearest neighbor of a point 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:

$$L_\gamma^{k\text{-NN}}(\mathcal{X}_n) = \sum_{i=1}^{n} \sum_{\boldsymbol{X} \in \mathcal{N}_{k,i}} |\boldsymbol{X} - \boldsymbol{X}_i|^\gamma \ . \tag{2}$$

2.1 Asymptotics in $\mathbb{R}^d$

Both the MST and the k-NN graph are part of a large class of graphs called *entropic graphs* [5]. Other graphs in this class include the minimal Euclidean matching graph, the traveling salesman tour or minimal triangulations among others. This class of graphs exhibits remarkable asymptotic behavior of its total edge length functional. Specifically, $L_\gamma(\mathcal{X}_n)/n^\alpha$, where $\alpha = (d-\gamma)/d$, converges with probability one (w.p.1) to the limit $\beta_{d,\gamma} \int_{\mathbb{R}^d} f^\alpha(\boldsymbol{y}) d\boldsymbol{y}$, as $n \to \infty$. The value of constant $\beta_{d,\gamma} > 0$ depends on the graph considered, but is independent of the distribution of $\boldsymbol{X}_i$. The quantity that

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

determines the aforementioned limit 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.

2.2 Entropic Graphs and Geodesics

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 $\mathcal{M}$ 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 $L_\gamma(\mathcal{X}_n)/n^\alpha$. However, by changing the notion of distance and dominating measure, the length functionals can provide crucial information for dimension and entropy estimation.

When $\mathcal{M}$ is a submanifold of $\mathbb{R}^d$, the usual Euclidean distance in $\mathbb{R}^d$ induces a natural norm on the tangent space to $\mathcal{M}$, given by the usual Euclidean norm. 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 $\ell(\Gamma) = \int_0^1 |\frac{\mathrm{d}}{\mathrm{d}t}\Gamma(t)|\mathrm{d}t$. 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:

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

Geodesic distances carry strong information about the nonlinear manifold $\mathcal{M}$, but their exact computation requires the knowledge of $\mathcal{M}$, which is not known in advance. However, it is possible to accurately estimate these distances based solely on a sample of points in $\mathcal{M}$. One such geodesic estimator, used in the ISOMAP algorithm, proceeds as follows. Two methods, called the ϵ-rule and the k-rule [1], are available for constructing the estimator. The first method connects each point to all points within some fixed radius ϵ and the other connects each point to all its k-nearest neighbors. A graph G defining the connectivity of these local neighborhoods is then used to approximate the geodesic distance between any pair of points as the shortest path through G that connects them. This results in an edge matrix whose (i,j) entry is the geodesic distance estimate for the (i,j)-th pair of points.

After the geodesic distances have been estimated, they can be used to construct new (non-Euclidean) graphs. Specifically, denote by $\hat{D}_{\mathcal{M}}$ the matrix of estimated pairwise distances between points in $\mathcal{Y}_n$, and by $\hat{d}(e_{ij})$ the estimated geodesic length of the corresponding edge $e_{ij} = \boldsymbol{Y}_i - \boldsymbol{Y}_j$. Define the GMST as the minimal graph over $\mathcal{Y}_n$ whose length is:

$$L_\gamma^{\mathrm{GMST}}(\mathcal{Y}_n) = \min_{T \in \mathcal{T}} \sum_{e \in T} \hat{d}^\gamma(e) \ . \tag{4}$$

By using geodesic information, the GMST length functional encodes global structure about the nonlinear manifold.

On the other hand, the local nature of nearest neighbor distances circumvents the use of geodesic distances in the k-NN graph: as the number of sample points increases, the geodesic NN distances became small enough to be well approximated by simple Euclidean NN distances. However, the global manifold structure is still accounted for in the k-NN graph by summing over all the NN distances of all points.

2.3 Asymptotics in Submanifolds of $\mathbb{R}^d$

For the geodesic graphs defined in the previous subsection, 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 density of $\boldsymbol{Y}_i$ relative to μ_g, the induced measure on $\mathcal{M}$ via the manifold's volume element [8].

Theorem 1 *Let $\mathcal{M}$ be a compact m-dimensional submanifold of $\mathbb{R}^d$, with metric structure induced by the usual Euclidean metric 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, for $L_\gamma(\mathcal{Y}_n)$ given by equation (2) or (4), w.p.1,*

$$\lim_{n\to\infty} \frac{L_\gamma(\mathcal{Y}_n)}{n^{(d'-\gamma)/d'}} = \tag{5}$$

$$\begin{cases} \infty, & d' < m \\ \beta_{m,\gamma} \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\,\mu_g(\mathrm{d}\boldsymbol{y}), & d' = m \\ 0, & d' > m \end{cases} ,$$

where $\beta_{m,\gamma}$ 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 quantity that determines the non-zero finite limit in (5) is the *intrinsic* Rényi α-entropy of the multivariate density f on $\mathcal{M}$:

$$H_\alpha^{\mathcal{M}}(f) = \frac{1}{1-\alpha} \log \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\,\mu_g(\mathrm{d}\boldsymbol{y}) \ . \tag{6}$$

3. LEARNING INTRINSIC DIMENSION AND ENTROPY

The asymptotic characterization of the GMST or k-NN length functional stated in Thm. 1 provides the framework for developing consistent estimators of both intrinsic dimension and entropy. The key observation is to note that parameter d' in expression (5) indexes the different growth rates in n that the length functional can have. In particular, only when d' is equal to the true intrinsic dimension, m, will $L_\gamma(\mathcal{Y}_n)/n^{(d'-\gamma)/d'}$ have a non-zero finite limit, determined by the intrinsic entropy of the random vectors involved.

Let $l_n = \log L_{\gamma,k}(\mathcal{Y}_n)$. According to (5), l_n has the following approximation

$$l_n = a \ \log n + b + \epsilon_n \ , \tag{7}$$

where

$$\begin{aligned} a &= (m-\gamma)/m \ , \\ b &= \log \beta_{m,\gamma} + H_\alpha^{(\mathcal{M},g)}(f)\,\gamma/m \ , \end{aligned} \tag{8}$$

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

Using the additive model (7), 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 graph length functionals $\bar{L}_p = N^{-1} \sum_{j=1}^N L_{\gamma,k}(\mathcal{Y}_p^j)$. Now, considering the vector of observations $\bar{\boldsymbol{l}} = [\log \bar{L}_{p_1}, \ldots, \log \bar{L}_{p_Q}]^T$ and using the linear model (7), one can 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 (8). After making a simple large n approximation, this approach yields the following estimates:

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

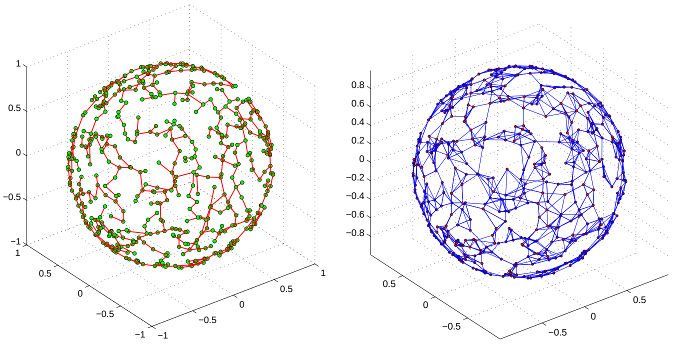


Figure 1: The GMST ($k = 7$) and 4-NN graph on 500 sample points on the 2D-sphere.

The constants $\beta_{m,\gamma}$ 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 $\{\beta_{m,\gamma}\}_{m>0}$ 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 $\beta_{m,\gamma 2}$ is required to obtain unbiased estimates of entropy. $\beta_{m,\gamma}$ 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 Monte Carlo simulations of the entropic graph length on the unit cube $[0,1]^m$ for uniform random samples. Another approach, in the GMST case, is to use known approximations and bounds [9].

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.

4. 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 fixed the parameters $\gamma = 1$ and $p_1 = n - Q, \dots, p_Q = n - 1$. With regards to intrinsic dimension estimation, we compare our algorithms to ISOMAP. In ISOMAP, similarly to PCA, intrinsic dimension is usually estimated by detecting changes in the residual fitting errors as a function of subspace dimension.

4.1 Hyper-spheres

We first consider the case of the m-dimensional sphere S^m embedded in $\mathbb{R}^{m+1}$ (Figure 1). This manifold presents some challenges as it does not satisfy any of the usual isometric or conformal embedding constraints required by ISOMAP or Hessian eigenmap [4] among others. We tested the algorithms 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 the GMST and k-NN are shown in Table 1 for different values of the parameter Q. As it can be seen, both methods succeed in finding the correct intrinsic dimension. It can also be noticed that, as intrinsic dimension increases, the GMST has a slightly better performance than the k-NN graph. As the GMST uses geodesic information to build its estimates, it is able to learn better the global geometry of the manifold, as opposed to the more local nature of the k-NN graph.

Table 1: Number of correct dimension estimates over 30 trials as a function of the number of samples for hyper-spheres ($N = 5$).

Sphere	Q	n	600	800	1000	1200
S^2	10	GMST	29	30	30	30
		5-NN	30	30	30	30
S^3	10	GMST	27	28	28	28
		5-NN	27	27	28	28
	20	GMST	29	30	30	30
		5-NN	29	29	30	30
S^4	10	GMST	23	27	29	29
		5-NN	23	26	26	26
	20	GMST	28	30	30	30
		5-NN	28	30	30	30

Table 2: Entropy estimates $\hat{H}$ for S^2 ($n = 600, N = 5, Q = 10$).

	emp. mean	std. deviation
GMST	5.4266	0.3514
5-NN	4.7424	0.9737

Table 2 shows the entropy estimates obtained by both methods on uniform samples supported in S^2. The true entropy is $H_{1/2} = 2\log(2\pi) \approx 5.3$.

4.2 MNSIT 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 [10], containing extensive test and training sets of all digits. Each digit in the database consists of a 28×28 pixels 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 2 and 3 (Figure 4) in the training set. As a real life high dimensional dataset, its manifold structure and intrinsic dimension are unknown.

Figure 3 shows the histogram of the dimension estimates for 30 simulations of both the GMST and k-NN algorithm applied to the samples of digit 2. Although the histograms show high variability, the GMST predicts an intrinsic dimension between 13 and 14, while the k-NN graph predicts the dimension to be in between 12 and 13. Figure 4 shows the corresponding residual plot 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

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

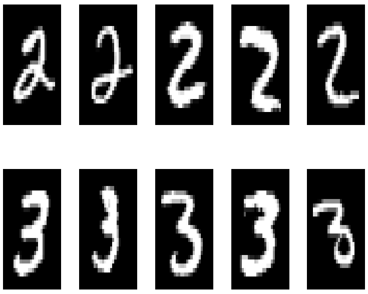
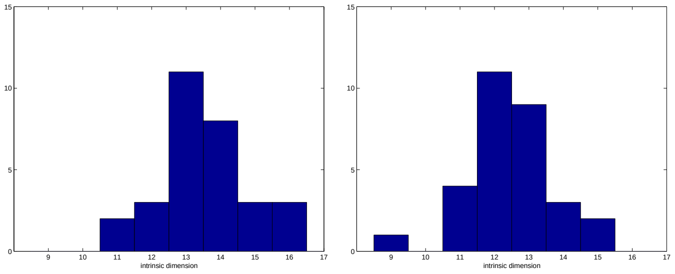


Figure 2: Samples from digits 2 and 3 of MNIST database.



(a)

(b)

Figure 3: Histogram of intrinsic dimensionality estimates for digit 2 in the MNIST database: (a) GMST; (b) 5-NN ($N = 10$, Q=15).

class of manifolds.

Table 3 displays the results of applying the entropic graphs to both digits, where the last column shows the results of processing both digits simultaneously.

5. CONCLUSION

We have applied entropic graph methods to the problem of estimating intrinsic dimension and entropy of high dimensional datasets constrained to lie on a manifold. In particular, our first results on a database of handwritten digits reveals interesting consequences: the low dimensionality of this dataset, as predicted by the proposed methods, contrasts with the complex state of the art classifiers developed for digit recognition [10], that require learning hundreds or even thousands of parameters/degrees of freedom.

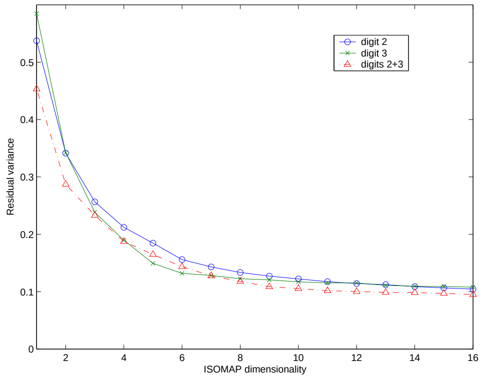


Figure 4: ISOMAP ($k = 6$) residual variance for digits 2 and 3 in the MNIST database.

Table 3: Dimension estimates $\hat{m}$ for digits 2 and 3 in the MNIST database.

	digit 2	digit 3	digit $2+3$
GMST	13	12	13
5-NN	12	11	12

Future work includes the development of bias correction mechanisms to improve the bootstrapping resampling step of the algorithm.

REFERENCES

[1] 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.

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

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

[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. 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.

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

[7] 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.

[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] 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.