Geodesic Entropic Graphs for Dimension and Entropy Estimation in Manifold Learning

Jose A. Costa, *Student Member, IEEE,* Alfred O. Hero, III, *Fellow, IEEE*

***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 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. We introduce a novel geometric approach based on entropic graph methods. Although the theory presented applies to this general class of graphs, we focus on the geodesic-minimal-spanning-tree (GMST) to obtaining asymptotically consistent estimates of the manifold dimension and the Rényi α-entropy of the sample density on the manifold. The GMST approach is striking in its simplicity and does not require reconstructing the manifold or estimating the multivariate density of the samples. The GMST method simply constructs a minimal spanning tree (MST) sequence using a geodesic edge matrix and uses the overall lengths of the MSTs to simultaneously estimate manifold dimension and entropy. We illustrate the GMST approach on standard synthetic manifolds as well as on real data sets consisting of images of faces.**

***Index Terms*— Nonlinear dimensionality reduction, minimal spanning tree, intrinsic dimension, intrinsic entropy, manifold learning, conformal embedding.**

I. 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. For support of this fact one needs only consider the success of image, video and audio compression algorithms, e.g. MP3, JPEG and MPEG, or the widespread use of efficient computational geometry methods for rendering smooth three dimensional shapes.

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. This manifold may in some cases be a linear, i.e., Euclidean, subspace but in general it is a non-linear curved surface. A problem of substantial recent interest in machine learning, computer vision, signal processing and statistics 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 [1]–[7]. This problem falls in the area of *manifold learning* which is concerned with discovering low dimensional structure in high dimensional data.

When the samples are drawn from a large population of signals one can interpret them as realizations from a multivariate distribution supported on the manifold. As this distribution is singular in the higher dimensional embedding space it has zero entropy as defined by the standard Lebesgue integral over the embedding space. However, when defined as a Lebesgue integral restricted to the lower dimensional manifold the entropy can be finite. This finite *intrinsic entropy* can be useful for exploring data compression over the manifold or, as suggested in [8], clustering of multiple sub-populations on the manifold. The question that we address in this paper is: how to simultaneously estimate the intrinsic dimension and intrinsic entropy on the manifold given a set of random sample points? We present a novel geometric probability approach to this question which is based on entropic graph methods developed by us and reported in publications [8]–[10].

Techniques for manifold learning can be classified into three categories: linear methods, local methods, and global methods. Linear methods include principal components analysis (PCA) [11] and classical multidimensional scaling (MDS) [12]. They are based on analyzing eigenstructure of empirical covariance matrices, and can be reliably applied only when the manifold is a linear subspace. Local methods include linear local imbedding (LLE) [2], locally linear projections (LLP) [13], Laplacian eigenmaps [14], and Hessian eigenmaps [3]. They are based on local approximation of the geometry of the manifold, and are computationally simple to implement. Global approaches include ISOMAP [1] and C-ISOMAP [15]. They preserve the manifold geometry at all scales, and have better stability than local methods when the number of manifold samples is limited.

With regards to estimation of the intrinsic dimension m several methods have been proposed [11], [16]. Most of these methods are based on linear projection techniques: a linear map is explicitly constructed and dimension is estimated by applying Principal Component Analysis (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 m as they don't account for non-linearities in the data. Both nonlinear PCA [4] methods and the ISOMAP circumvent this problem but they still rely on possibly unreliable and costly eigenstructure estimates. Other methods have been proposed based on local geometric techniques, e.g., estimation of local

This work was supported in part by NSF under grant CCR-032557 and by the NIH Cancer Institute under grant 1PO1 CA87634-01.

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

neighborhoods [6] or fractal dimension [17], and estimating packing numbers [5] of the manifold.

We propose a geodesic-minimal-spanning-tree (GMST) method that jointly estimates both the intrinsic dimension and intrinsic entropy on the manifold. The method is implemented as follows. First a complete geodesic graph between all pairs of data samples is constructed, as in ISOMAP or C-ISOMAP. Then a minimal spanning graph, the GMST, is obtained by pruning the complete geodesic graph down to a subgraph that still connects all points but has minimum total geodesic length. The intrinsic dimension and intrinsic α-entropy are then estimated from the GMST length functional using a simple linear least squares (LLS) and method of moments (MOM) procedure.

The GMST method falls in the category of global approaches to manifold learning but it differs significantly from the aforementioned methods. First, it has a different scope. Indeed, unlike ISOMAP and C-ISOMAP, the GMST method provides a statistically consistent estimate of the intrinsic entropy in addition to the intrinsic dimension of the manifold. To the best of our knowledge no other such technique has been proposed for learning manifold dimension. Second, unlike local methods that work on chunks of data in local neighborhoods, GMST works on resampled data distributed over the global data set. Third, the GMST method is simple and elegant: it estimates intrinsic entropy and dimension by detecting the rate of increase of a graph as a function of the number of its resampled vertices.

The aims of this paper are limited to introducing GMST as a novel method for estimating manifold dimension and entropy of the samples. As in work of others on dimension estimation [5], [17] we do not here consider the issue of reconstruction of the complete manifold. Similarly to these authors, we believe that dimension estimation and entropy estimation for non-linear data are of interest in their own right. We also do not consider the effect of additive noise or outliers on the performance of GMST. Finally, the consistency results of GMST reported here are limited to domain manifolds defined by some smooth unknown mapping. The extension of GMST methodology to general target manifolds, e.g. those defined by implicit level set embeddings [18], [19], is a worthwhile topic for future investigation.

What follows is a brief outline of the paper. We review some necessary background on the mathematics of domain manifolds in Sec. II. In Sec. III we review the asymptotic theory of entropic graphs and obtain several new results required for their extension to embedded manifolds. In Sec. IV we define the general GMST algorithm. Finally, in Sec. V we test the GMST algorithm on standard synthetic manifolds and on a real data set consisting of human faces from different subjects.

II. Geometric Background

A. A 3D Example

To illustrate ideas consider a 2D surface embedded in 3D Euclidean space, called the embedding space. Let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\} \subset U \subseteq \mathbb{R}^2$ be a set of points (samples) in a subset U of the plane. Naturally, the shortest path between any pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ of these points is given by the straight line in $\mathbb{R}^2$ connecting them, with corresponding distance given by its Euclidean (L_2) length, $|\boldsymbol{x}_i - \boldsymbol{x}_j|$. Now let U be used as a parameterization space to describe a curved surface in $\mathbb{R}^3$ via a mapping $\varphi : U \to \mathbb{R}^3$. Surfaces $\mathcal{M} = \varphi(U)$ defined in this explicit manner are called domain or parameterized manifolds and they inherit the topological dimension, equal to 2 in this case, of the parameterization space. When φ is non-linear, the shortest path on $\mathcal{M}$ between points $\boldsymbol{y}_i = \varphi(\boldsymbol{x}_i)$ and $\boldsymbol{y}_j = \varphi(\boldsymbol{x}_j)$ is a curve on the surface called the geodesic curve. In this paper we will primarily consider domain manifolds defined by conformal mappings φ. Such conformal embeddings have the property that the angles between tangent vectors to the surface are identical to angles between corresponding vectors in the parameterization space, possibly up to a smoothly varying local scale factor. This property guarantees that, regardless of how the mapping φ "deforms" U onto $\mathcal{M}$, the geodesic distances in $\mathcal{M}$ are closely related to the Euclidean distances in U. When this smooth surface representation holds there exist algorithms, e.g. ISOMAP and C-ISOMAP [1], [15], which can be used to estimate the Euclidean distances between points in U from estimates of the geodesic distances between points in $\mathcal{M}$. If a certain type of minimal spanning graph is constructed using these estimates, well established results in geometrical probability [8], [20] allow us to develop simple estimates of both entropy and dimension of the points distributed on the surface.

B. Differential Geometry Setting

In the following, we recall some facts from differential geometry needed to formalize and generalize the ideas just described. We will consider smooth manifolds embedded in $\mathbb{R}^d$. For the general theory we refer the reader to any standard book in differential geometry (for example, [21], [22], [23]). An m-dimensional *smooth manifold* $\mathcal{M} \subseteq \mathbb{R}^d$ is a set such that each of its points has a neighborhood that can be parameterized by an open set of $\mathbb{R}^m$ through a local change of coordinates. Intuitively, this means that although $\mathcal{M}$ is a (hyper) surface in $\mathbb{R}^d$, it can be locally identified with $\mathbb{R}^m$.

Let $\varphi : \Omega \mapsto \mathcal{M}$ be a mapping between two manifolds, $\Omega, \mathcal{M}$. Let γ be a curve in Ω. The *tangent map* $\mathrm{d}\varphi_{\boldsymbol{x}}$ assigns each tangent vector $\boldsymbol{v}$ to Ω at point $\boldsymbol{x}$ the tangent vector $\mathrm{d}\varphi_{\boldsymbol{x}}\boldsymbol{v}$ to $\mathcal{M}$ at point $\varphi(\boldsymbol{x})$, such that, if $\boldsymbol{v}$ is the initial velocity of γ in Ω, then $\mathrm{d}\varphi_{\boldsymbol{x}}\boldsymbol{v}$ is the initial velocity of the curve $\varphi(\gamma)$ in $\mathcal{M}$. For example, if $\boldsymbol{x} \in U \subseteq \Omega \subseteq \mathbb{R}^m$, with U an open set of $\mathbb{R}^m$, then $\mathrm{d}\varphi_{\boldsymbol{x}}\boldsymbol{v} = J_\varphi(\boldsymbol{x})\,\boldsymbol{v}$, where $J_\varphi = [\partial\varphi_i/\partial x_j]$, $i = 1, \ldots, d$, $j = 1, \ldots, m$, is the Jacobian matrix associated with φ at point $\boldsymbol{x} \in \Omega$.

The *length* of a smooth curve $\Gamma : [0,1] \to \mathcal{M}$ is defined 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:

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

We can now define the following types of embeddings.

Definition 1: $\varphi : \Omega \mapsto \mathcal{M}$ is called a conformal mapping if φ is a diffeomorphism (i.e., φ is differentiable, bijective with differentiable inverse φ^{-1}) and, at each point $\boldsymbol{x} \in \Omega$, φ preserves the angles between tangent vectors, i.e.,

$$(\mathrm{d}\varphi_{\boldsymbol{x}}\boldsymbol{v})^T (\mathrm{d}\varphi_{\boldsymbol{x}}\boldsymbol{w}) = c(\boldsymbol{x})\, \boldsymbol{v}^T \boldsymbol{w} \ , \tag{1}$$

for all vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ that are tangent to Ω at $\boldsymbol{x}$, and $c(\boldsymbol{x}) > 0$ is a scaling factor that varies smoothly with $\boldsymbol{x}$. If for all $\boldsymbol{x} \in \Omega$, $c(\boldsymbol{x}) = 1$, then φ is said to be a (global) isometry. In this case the length of tangent vectors is also preserved in addition to the angles between them.

If there is an open set $U \subseteq \Omega \subseteq \mathbb{R}^m$ [21], then the diffeomorphism φ is a conformal mapping iff $J_\varphi(\boldsymbol{x})^T J_\varphi(\boldsymbol{x}) = c(\boldsymbol{x})\, I_m$, where I_m is the $m \times m$ identity matrix. In this case, the geodesic distance in $\mathcal{M}$ can be computed as follows. Any smooth curve $\Gamma : [0,1] \mapsto \mathcal{M}$ can be represented as $\Gamma(t) = \varphi(\gamma(t))$, where $\gamma : [0,1] \mapsto \Omega$ is a smooth curve in $\mathbb{R}^m$. Then, the length $\ell(\Gamma)$ of the curve Γ is given by

$$\begin{aligned}\ell(\Gamma) &= \int_0^1 \left| \frac{\mathrm{d}}{\mathrm{d}t}\varphi(\gamma(t)) \right| \mathrm{d}t = \int_0^1 |J_\varphi(\gamma(t))\, \dot\gamma(t)|\, \mathrm{d}t \\ &= \int_0^1 \sqrt{c(\gamma(t))}\, |\dot\gamma(t)|\, \mathrm{d}t \ .\end{aligned}$$

As in $\mathbb{R}^m$ the shortest path between any two points is given by the straight line that connects them, $\gamma(t) = \boldsymbol{x}_0 + t(\boldsymbol{x}_1 - \boldsymbol{x}_0)$ minimizes $\int_0^1 |\dot\gamma(t)|\, \mathrm{d}t$, over all smooth curves with start and end points at $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$, respectively. So, if $c(\boldsymbol{x})$ is constant, i.e. $c(\boldsymbol{x}) = c$ for all $x \in \Omega$, the geodesic distance between $\boldsymbol{y}_0 = \varphi(\boldsymbol{x}_0)$ and $\boldsymbol{y}_1 = \varphi(\boldsymbol{x}_1)$ is

$$d_{\mathcal{M}}(\varphi(\boldsymbol{x}_0), \varphi(\boldsymbol{x}_1)) = \sqrt{c}\, |\boldsymbol{x}_0 - \boldsymbol{x}_1| \ . \tag{2}$$

When $c = 1$, i.e., φ is an isometry, the geodesic distance in $\mathcal{M}$ and the Euclidean distance in the parameterization space $\mathbb{R}^m$ are the same. If $c > 1$ ($c < 1$) there is a global expansion (contraction) in the distances between points.

It is evident from the above discussion that geodesic distances carry strong information about a non-linear domain manifold such as $\mathcal{M}$. However, their computation requires the knowledge of the analytical form of $\mathcal{M}$ via φ and its Jacobian. Our goal is to learn the entropy of non-linear data on a domain manifold together with its intrinsic dimension, given only the data set $\mathcal{Y}_n$ of n samples in the embedding space $\mathbb{R}^d$, and without knowledge of its embedding function φ.

III. Entropic Graph Estimators on Embedded Manifolds

Let $\mathcal{Y}_n = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ 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. Define the distance matrix D as the $n \times n$ matrix of edge weights (w.r.t. a specified metric). A spanning graph T over $\mathcal{Y}_n$ is defined as the pair $\{V, E\}$ where $V = \mathcal{Y}_n$ and E is a subset of all graph edges connecting pairs of vertices in V, with weights given by D. When D is computed from pairwise Euclidean distances, T is called a Euclidean spanning graph.

It has long been known [24] that, when suitably normalized, the sum of the edge weights of certain minimal Euclidean spanning graphs T over $\mathcal{Y}_n$ converges with probability 1 (w.p.1) to the limit $\beta_d \int_{\mathbb{R}^d} f^\alpha(\boldsymbol{y}) d\boldsymbol{y}$, where the integral is interpreted in the sense of Lebesgue, $\alpha \in (0,1)$ and $\beta_d > 0$. This a.s. limit is the integral factor $\int f^\alpha$ in what we will call the *extrinsic* Rényi α-entropy of the multivariate Lebesgue density f:

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

In the limit, when $\alpha \to 1$ we obtain the usual Shannon entropy, $-\int_{\mathbb{R}^d} f(\boldsymbol{y}) \log f(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}$. Graph constructions that converge to the integral in the limit (3) were called continuous quasi-additive (Euclidean) graphs in [20] and entropic (Euclidean) graphs in [8]. See the monographs by Steele [25] and Yukich [20] for an excellent introduction to the theory of such random Euclidean graphs. Relevant details for these results are given in the next subsection.

The α-entropy has proved to be an important quantity in signal processing, where its applications range from vector quantization [26], [27] to pattern matching [28] and image registration [8], [29]. The α-entropy parameterizes the Chernoff exponent governing the minimum probability of error [30] making it an important quantity in detection and classification problems. Like the Shannon entropy, the α-entropy also has an operational characterization in terms of source coding rates. In [31] it was shown that the α-entropy of a source determines the achievable block-code rates in the sense that the probability of block decoding error converges to zero at an exponential rate with rate constant $H_\alpha^{\mathbb{R}^d}(f)$.

A. Beardwood-Halton-Hammersley Theorem in $\mathbb{R}^d$

Let $\mathcal{Y}_n = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ be a set of points in $\mathbb{R}^d$. A minimal Euclidean graph spanning $\mathcal{Y}_n$ is defined as the graph spanning $\mathcal{Y}_n$ having minimal overall length

$$L_\gamma^{\mathbb{R}^d}(\mathcal{Y}_n) = \min_{T \in \mathcal{T}} \sum_{e \in T} |e|^\gamma \ . \tag{4}$$

Here the sum is over all edges e (e.g., $e = \boldsymbol{Y}_i - \boldsymbol{Y}_j, i \neq j$) in the graph T, $|e|$ is the Euclidean length of e, and $\gamma \in (0, d)$ is called the *edge exponent* or *power-weighting constant*. For example when $\mathcal{T}$ is the set of spanning trees over $\mathcal{Y}_n$ one obtains the MST. A minimal Euclidean graph is continuous quasi-additive when it satisfies several technical conditions specified in [20] (also see [9]). Continuous quasi-additive Euclidean graphs include: the minimal spanning tree (MST), the k-nearest neighbors graph (k-NNG), the minimal matching graph (MMG), the traveling salesman problem (TSP), and their power-weighted variants. While all of the results in this paper apply to this larger class of minimal graphs we specialize to the MST for concreteness.

A remarkable result in geometric probability was established by Beardwood, Halton and Hammersley [24].

Beardwood-Halton-Hammersley (BHH) Theorem [20], [25]: Let $\mathcal{Y}_n$ be an i.i.d. set of random variables taking values in a compact subset of $\mathbb{R}^d$ having common probability distribution

P. Let this distribution have the decomposition $P = F + Q$ where F is the Lebesgue continuous component and Q is the singular component. The Lebesgue continuous component has a Lebesgue density (no singular component) which is denoted $f(x)$, $x \in \mathbb{R}^d$. Let $L_\gamma^{\mathbb{R}^d}(\mathcal{Y}_n)$ be the length of the MST spanning $\mathcal{Y}_n$ and assume that $d \geq 2$ and $0 < \gamma < d$. Then, w.p.1

$$\lim_{n\to\infty} L_\gamma^{\mathbb{R}^d}(\mathcal{Y}_n)/n^\alpha = \beta_d \int_{\mathbb{R}^d} f^\alpha(\boldsymbol{y}) d\boldsymbol{y} \tag{5}$$

where $\alpha = (d-\gamma)/d$ and β_d is a constant not depending on the distribution P. Furthermore, the mean length $E[L_\gamma^{\mathbb{R}^d}(\mathcal{Y}_n)]/n^\alpha$ converges to the same limit.

The limit on the right side of (5) in the BHH theorem is zero when the distribution P has no Lebesgue continuous component, i.e., when $F \equiv 0$. On the other hand, when P has no singular component, i.e., $Q \equiv 0$, a consequence of the BHH Theorem is that

$$\hat{H}_\alpha^{\mathbb{R}^d}(\mathcal{Y}_n) \stackrel{\text{def}}{=} \frac{d}{\gamma}\left[\log \frac{L_\gamma^{\mathbb{R}^d}(\mathcal{Y}_n)}{n^{(d-\gamma)/d}} - \log \beta_d\right] \tag{6}$$

is an asymptotically unbiased and strongly consistent estimator of the extrinsic α-entropy $H_\alpha^{\mathbb{R}^d}(f)$ defined in (3). For a discussion about the role of the constant β_d in the proposed estimators see section IV.

B. Generalization of BHH Thm. to Embedded Manifolds

If the vertices $\mathcal{Y}_n$ are constrained to lie on a smooth compact m-dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$, the distribution of $\boldsymbol{Y}_i$ is singular with respect to Lebesgue measure, $F \equiv 0$, and, as previously mentioned, the limit (5) in the BHH Theorem is zero. However, as shown below, if $\mathcal{M}$ is defined by an isometric embedding from the parameterization space $\mathbb{R}^m$, if $\boldsymbol{Y}_i$ has a density f on $\mathcal{M}$, and if the geodesic estimation step of ISOMAP is used to approximate geodesic distances, then the length of an MST constructed from the geodesic edge lengths can be made to converge, after suitable normalization and transformation, to the *intrinsic* α-entropy $H_\alpha^{\mathcal{M}}(f)$ on $\mathcal{M}$ defined by

$$H_\alpha^{\mathcal{M}}(f) = \frac{m}{\gamma} \log \int_{\mathcal{M}} f^\alpha(\boldsymbol{y}) \mu_{\mathcal{M}}(d\boldsymbol{y}) \ , \tag{7}$$

where $\mu_{\mathcal{M}}(d\boldsymbol{y})$ denotes the differential volume element over $\mathcal{M}$.

More generally, assume that $\mathcal{M}$ is embedded in $\mathbb{R}^d$ through the diffeomorphism φ. As $\boldsymbol{X}_i = \varphi^{-1}(\boldsymbol{Y}_i)$ lives in $\mathbb{R}^m$, let T be the Euclidean minimal graph spanning $\mathcal{X}_n$ and having length function $L_\gamma^{\mathbb{R}^m}(\mathcal{X}_n) = L_\gamma^{\mathbb{R}^m}\left(\varphi^{-1}(\mathcal{Y}_n)\right)$ according to definition (4). We have the following extension of the BHH Theorem.

Theorem 1: Let $\mathcal{M}$ be a smooth compact m-dimensional manifold embedded in $\mathbb{R}^d$ through the diffeomorphism $\varphi : \Omega \mapsto \mathcal{M}, \Omega \subset \mathbb{R}^m$. Assume $2 \leq m \leq d$ and $0 < \gamma < m$. Suppose that $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ are i.i.d. random vectors on $\mathcal{M}$ having common density f with respect to Lebesgue measure $\mu_{\mathcal{M}}$ on $\mathcal{M}$. Then, the length functional $L_\gamma^{\mathbb{R}^m}(\varphi^{-1}(\mathcal{Y}_n))$ of the MST spanning $\varphi^{-1}(\mathcal{Y}_n)$ satisfies

$$\lim_{n\to\infty} L_\gamma^{\mathbb{R}^m}(\varphi^{-1}(\mathcal{Y}_n))/n^{(d'-\gamma)/d'} = \tag{8}$$
$$\begin{cases} \infty, & d' < m \\ \beta_m \int_{\mathcal{M}} \left[\det\left(J_\varphi^T J_\varphi\right)\right]^{\frac{\alpha-1}{2}} f^\alpha(\boldsymbol{y})\, \mu_{\mathcal{M}}(d\boldsymbol{y}), & d' = m \\ 0, & d' > m \end{cases}$$

w.p.1, where $\alpha = (m-\gamma)/m$. Furthermore, the mean $E[L_\gamma^{\mathbb{R}^m}(\varphi^{-1}(\mathcal{Y}_n))]/n^{(d'-\gamma)/d'}$ converges to the same limit.

Proof: This theorem is a simple consequence of relation (5) in the BHH Theorem and properties of integrals over manifolds. By the BHH Theorem, w.p.1,

$$L_\gamma^{\mathbb{R}^m}(\mathcal{X}_n) = n^{(m-\gamma)/m} \beta_m \int_{\mathbb{R}^m} f_X^\alpha(\boldsymbol{x})\, d\boldsymbol{x} + o(n^{(m-\gamma)/m}), \tag{9}$$

where f_X is the density of $\boldsymbol{X}_i = \varphi^{-1}(\boldsymbol{Y}_i)$. Therefore the limits claimed in (8) for $d' < m$ and $d' > m$ are obvious. For $d' = m$, relation (9) implies

$$\lim_{n\to\infty} L_\gamma^{\mathbb{R}^m}(\mathcal{X}_n)/n^{(m-\gamma)/m} = \beta_m \int_{\mathbb{R}^m} f_X^\alpha(\boldsymbol{x})\, d\boldsymbol{x}, \tag{10}$$

and it remains to show that this limit is identical to the limit asserted in (8).

For an integrable function F defined on a domain manifold $\mathcal{M}$ defined by the diffeomorphism $\varphi : \mathbb{R}^m \mapsto \mathcal{M}$, the integral of F over $\mathcal{M}$ satisfies the relation [22]:

$$\int_{\mathcal{M}} F(\boldsymbol{y})\, \mu_{\mathcal{M}}(d\boldsymbol{y}) = \int_{\mathbb{R}^m} F(\varphi(\boldsymbol{x}))\, g(\boldsymbol{x})\, d\boldsymbol{x} \ , \tag{11}$$

where $g(\boldsymbol{x}) = \sqrt{\det\left(J_\varphi^T J_\varphi\right)}$ is the Riemannian metric associated with $\mathcal{M}$. Specializing F to the indicator function of a small volume centered at a point $\boldsymbol{y}$, (11) implies the following relation between volume elements in $\mathcal{M}$ and $\mathbb{R}^m$: $\mu_{\mathcal{M}}(d\boldsymbol{y}) = g(\boldsymbol{x})\, d\boldsymbol{x}$. Furthermore, specializing to $F(\boldsymbol{y}) = f(\boldsymbol{y})$ it is clear from (11) that $f_X(\boldsymbol{x}) = f(\varphi(\boldsymbol{x}))g(\boldsymbol{x})$. Therefore

$$\begin{aligned} \int_{\mathbb{R}^m} f_X^\alpha(\boldsymbol{x}) d\boldsymbol{x} &= \int_{\mathbb{R}^m} (f(\varphi(\boldsymbol{x}))g(\boldsymbol{x}))^\alpha d\boldsymbol{x} \\ &= \int_{\mathbb{R}^m} f^\alpha(\varphi(\boldsymbol{x})) g^{\alpha-1}(\boldsymbol{x})\ g(\boldsymbol{x}) d\boldsymbol{x} \ , \end{aligned}$$

which, after the change of variable $\boldsymbol{x} \mapsto \varphi(\boldsymbol{x})$, is equivalent to the integral in the limit (8). ∎

C. Estimating Geodesic Distances

If φ is an isometric or conformal embedding then it has been shown that for sufficiently dense sampling over $\mathcal{M}$, i.e., for large n, the ISOMAP or the C-ISOMAP algorithm summarized in Table I will approximate the matrix of pairwise Euclidean distances between points $\mathcal{X}_n = \varphi^{-1}(\mathcal{Y}_n)$ in the domain space $\mathbb{R}^m$ without explicit knowledge of φ. This estimate is computed from an Euclidean graph G connecting all local neighborhoods of data points in $\mathcal{M}$. Specifically, in the isometric case, ISOMAP proceeds as follows. Two methods, called the ϵ-rule and the k-rule [1], are available for constructing G. The first method connects each point to

TABLE I

DISTANCE ESTIMATION STEPS OF ISOMAP/C-ISOMAP ALGORITHMS TO RECONSTRUCT EUCLIDEAN DISTANCES BETWEEN $\mathcal{X}_n$ ON THE EMBEDDING PARAMETERIZATION SPACE FROM POINTS $\mathcal{Y}_n$ OVER THE EMBEDDED MANIFOLD.

Step 1.	Determine a Euclidean neighborhood graph G of the observed data $\mathcal{Y}_n$ according to the ϵ-rule or the k-rule as defined in ISOMAP [32].
Step 2.	For isometric embeddings compute the edge matrix $\mathcal{E}$ of the ISOMAP graph [1] and for conformal imbeddings compute the edge matrix $\mathcal{E}$ of the C-ISOMAP graph [15]. The (i,j) entry of this symmetric matrix is the sum of the lengths of the edges in G along the shortest path between the pair of vertices $(\boldsymbol{Y}_i, \boldsymbol{Y}_j)$ where the edge lengths between neighboring points $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$ in G are defined as Euclidean distance $\lvert\boldsymbol{Y}_1\text{-}\boldsymbol{Y}_2\rvert$ in the case of ISOMAP or $\lvert\boldsymbol{Y}_1\text{-}\boldsymbol{Y}_2\rvert/\sqrt{M(1)M(2)}$ in the case of C-ISOMAP where $M(i)$ is the mean distance of $\boldsymbol{Y}_i$ to its immediate nearest neighbors.

all points within some fixed radius ϵ and the other connects each point to all its k-nearest neighbors. The 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. Finally, this results in a distance matrix whose (i,j) entry is the geodesic distance estimate for the (i,j)-th pair of points. The geodesic distance estimation algorithm just described is motivated by the fact that locally a smooth manifold is well "approximated" by a linear hyper-plane and, so, geodesic distances between neighboring points are close to their Euclidean distances. For faraway points, the geodesic distance is estimated by summing the sequence of such local approximations over the shortest path through the graph G.

Thus, if one uses these distances to construct an MST, its length function will approximate $L_\gamma^{\mathbb{R}^m}(\varphi^{-1}(\mathcal{Y}_n))$ and we can invoke Thm. 1 to characterize its asymptotic convergence properties. As the estimated distances will use information about the geodesic distances between pairs of points $(\boldsymbol{Y}_i, \boldsymbol{Y}_j)$ this graph will be called a *geodesic* MST (GMST).

More specifically, denote by $\hat{D}_{\mathcal{M}}$ the matrix of estimated pairwise distances between points $\varphi^{-1}(\boldsymbol{Y}_i)$ and $\varphi^{-1}(\boldsymbol{Y}_j)$ in $\varphi^{-1}(\mathcal{Y}_n)$, and by $\hat{d}(e_{ij})$ the estimated length of the corresponding edge $e_{ij} = \varphi^{-1}(\boldsymbol{Y}_i) - \varphi^{-1}(\boldsymbol{Y}_j)$. Define the GMST T as the minimal graph over $\mathcal{Y}_n$ whose length is:

$$\hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_n) = \min_{T\in\mathcal{T}} \sum_{e\in T} \hat{d}^\gamma(e) \ . \tag{12}$$

The following is the principal theoretical result of this paper and is a simple consequence of Thm. 1.

Corollary 1: Let $\mathcal{M}$ be a smooth compact m-dimensional manifold embedded in $\mathbb{R}^d$ through the diffeomorphism $\varphi : \mathbb{R}^m \mapsto \mathcal{M}$. Let $2 \le m \le d$ and $0 < \gamma < m$. Suppose that $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are i.i.d. random vectors on $\mathcal{M}$ with common density f w.r.t. Lebesgue measure $\mu_{\mathcal{M}}$ on $\mathcal{M}$. If the entries $\{\hat{d}_{ij}\}$ of matrix $\hat{D}_{\mathcal{M}}$ satisfy

$$\max_{1\le i,j\le n} \left| \frac{\hat{d}_{ij}}{|\varphi^{-1}(\boldsymbol{Y}_i) - \varphi^{-1}(\boldsymbol{Y}_j)|} - 1 \right| \to 0 \text{ as } n\to\infty \tag{13}$$

w.p.1, then the length functional of the GMST satisfies

$$\lim_{n\to\infty} \hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_n)/n^{(d'-\gamma)/d'} = \begin{cases} \infty, & d' < m \\ \beta_m \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\, g^{-\frac{\gamma}{m}}(\varphi^{-1}(\boldsymbol{y}))\, \mu_{\mathcal{M}}(d\boldsymbol{y}), & d' = m \\ 0, & d' > m \end{cases} \tag{14}$$

w.p.1, where $\alpha = (m-\gamma)/m$ and $g(\boldsymbol{x}) = \sqrt{\det\left(J_\varphi^T J_\varphi\right)}$. Furthermore, the mean $E[\hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_n)]/n^{(d'-\gamma)/d'}$ converges to the same limit.

The sufficient condition (13) of Corollary 1 simply states that the pairwise Euclidean distances between points in $\mathcal{X}_n = \varphi^{-1}(\mathcal{Y}_n)$ should be uniformly well approximated by the entries of matrix $\hat{D}_{\mathcal{M}}$. Constructions of $\hat{D}_{\mathcal{M}}$ which satisfy condition (13) will be discussed in the next subsection.

Proof of Corollary 1: Write $\hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_n)$ as

$$\hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_n) = \min_{T\in\mathcal{T}} \sum_{e\in T} \left[\frac{\hat{d}(e)}{|e|}\right]^\gamma |e|^\gamma \ .$$

The uniform convergence expressed by condition (13) implies that

$$\hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_n) = (1+o(1))^\gamma L_\gamma^{\mathbb{R}^m}(\varphi^{-1}(\mathcal{Y}_n)) \ .$$

Applying Thm. 1 and identifying $(\alpha - 1) = -\gamma/m$, $\boldsymbol{x} = \varphi^{-1}(\boldsymbol{y})$ and $\det\left(J_\varphi^T J_\varphi\right) = g(\varphi^{-1}(\boldsymbol{y}))$ provides the desired result. ∎

If $m > 2$, as the parameter d' is increased from 2 to ∞ the limit (14) in Corollary 1 transitions from infinity to a finite limit and finally to zero over three consecutive steps $d' = m-1, m, m+1$. As d' indexes the rate constant $n^{(d'-\gamma)/d'}$ of the length functional $\hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_n)$, this abrupt transition suggests that the intrinsic dimension m and the intrinsic entropy might be easily estimated by investigating the growth rate of the GMST's length functional. This observation is the basis for the estimation algorithm introduced in the next section.

We now specialize Corollary 1 to the following cases of interest.

1) Isometric Embeddings: In the case that φ defines an isometric embedding, the geodesic estimation step of ISOMAP is asymptotically able to recover the true Euclidean distances between the points in $\mathcal{X}_n = \varphi^{-1}(\mathcal{Y}_n)$ and $\hat{D}_{\mathcal{M}}$ satisfies condition (13) [32]. Furthermore, $J_\varphi^T J_\varphi = I_m$. Thus, using ISOMAP to construct $\hat{D}_{\mathcal{M}}$, limit (14) holds with the $d' = m$ limit replaced by

$$\beta_m \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\, \mu_{\mathcal{M}}(d\boldsymbol{y}).$$

Furthermore, $m/\gamma \log\left(\hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_n)/n^{(m-\gamma)/m} - \log\beta_m\right)$ converges w.p.1 to the intrinsic entropy (7).

If φ defines an isometric embedding with contraction or expansion, the geodesic estimation step of ISOMAP algorithm is able to recover the true Euclidean distances between points in $\mathcal{X}_n$ only up to an unknown scaling constant c (c.f. (2)). As

$J_\varphi^T J_\varphi = c\, I_m$, limit (14) holds with the $d' = m$ limit replaced by

$$\beta_m c^{1-\gamma/2} \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\, \mu_{\mathcal{M}}(d\boldsymbol{y}).$$

Now, the entropy estimator defined above converges w.p.1 up to an unknown additive constant $(1-\gamma/2)\log c$ to the intrinsic entropy (7). We point out that in many signal processing applications (e.g. image registration) a constant bias on the entropy estimate does not pose a problem since an estimate of the relative magnitude of the entropy functional is all that is required.

2) Non-isometric Embeddings Defined by Conformal Mappings: In the case that φ is a general (non-isometric) conformal mapping, it was stated in [33] without proof, that the C-ISOMAP algorithm is once again able to recover the true Euclidean distances between points in $\mathcal{X}_n$. Furthermore, $J_\varphi^T J_\varphi = c(\boldsymbol{x})\, I_m$. Thus, when $\hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_n)$ is the length of the geodesic MST constructed on the distance matrix generated by the C-ISOMAP algorithm, we expect the limit (14) to hold with the $d' = m$ limit replaced by

$$\beta_m \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\; c^{-\gamma/2}(\varphi^{-1}(\boldsymbol{y}))\, \mu_{\mathcal{M}}(d\boldsymbol{y}).$$

In this case, $m/\gamma \log\left(\hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_n)/n^{(m-\gamma)/m} - \log\beta_m\right)$ would converge a.s. to the weighted intrinsic entropy

$$\frac{1}{1-\alpha}\log\int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\; c^{-\gamma/2}(\varphi^{-1}(\boldsymbol{y}))\, \mu_{\mathcal{M}}(\boldsymbol{y})\;.$$

The weighted α-entropy is a "version" of the standard unweighted α-entropy $H_\alpha^{\mathcal{M}}(f)$ which is "tilted" by the space-varying volume element of $\mathcal{M}$. This unknown weighting makes it impossible to estimate the intrinsic unweighted α-entropy. However, as can be seen from the discussion in the next section, as the growth rate exponent of the GMST length depends on m we can still perform dimension estimation in this case.

3) Non-conformal Diffeomorphic Embeddings: When φ defines a general diffeomorphic embedding, an extension of the C-ISOMAP algorithm that can provably learn the Euclidean distances between the points $\mathcal{X}_n$ in the parametrization space is needed in order to apply Corollary 1. To the best of our knowledge such an extension of C-ISOMAP does not yet exist.

IV. GMST Algorithm

Now that we have characterized the asymptotic limit (14) of the length functional of the GMST we here apply this theory to jointly estimate entropy and dimension. The key is to notice that the growth rate of the length functional is strongly dependent on the intrinsic dimension 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. Throughout we assume that the geodesic minimal graph length $\hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_n)$ is determined from a distance matrix $\hat{D}_{\mathcal{M}}$ that satisfies the assumption of Corollary 1, e.g., obtained using ISOMAP or C-ISOMAP. We also assume that $m \geq 2$. This guarantees that $\hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_n)/n^{(d'-\gamma)/d'}$ has a non-zero finite convergent limit for $d' = m$. Next define

TABLE II

GMST resampling algorithm for estimating intrinsic dimension m and intrinsic entropy $H_\alpha^{\mathcal{M}}$.

Initialize: Using entire database of signals $\mathcal{Y}_n$ construct geodesic distance matrix $\hat{D}_{\mathcal{M}}$ using ISOMAP or C-ISOMAP method.
Select parameters: $M > 0$, $N > 0$, $Q > 0$ and $p_1 < \ldots < p_Q \leq n$

$\overline{m} = 0$, $\overline{H} = 0$;
for $M' = 1, \ldots, M$
  for $p = p_1, \ldots, p_Q$
    $\overline{L} = 0$;
    for $N' = 1, \ldots, N$
      Randomly select a subset of p signals $\mathcal{Y}_p$ from $\mathcal{Y}_n$;
      Compute geodesic MST length L_p over $\mathcal{Y}_p$ and from $\hat{D}_{\mathcal{M}}$;
      $\overline{L} = \overline{L} + L_p$;
    end for
    Compute sample average geodesic MST length;
    $\hat{E}[\hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_p)] = \overline{L}/N$;
  end for
  Estimate dimension $\hat{m}_{M'}$ and α-entropy $\hat{H}_{M'}$ from $\{\hat{E}[\hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_p)]\}_{p=p_1}^{p_Q}$ via LLS/NLLS;
  $\overline{m} = \overline{m} + \hat{m}_{M'}$, $\overline{H} = \overline{H} + \hat{H}_{M'}$;
end for
$\hat{m} = \overline{m}/M$, $\hat{H} = \overline{H}/M$

$l_n = \log \hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_n)$. According to (14) l_n has the following approximation

$$l_n = a\, \log n + b + \epsilon_n\;, \tag{15}$$

where

$$\begin{aligned} a &= (m-\gamma)/m\;, \\ b &= \log\beta_m + \gamma/m\; H_\alpha^{\mathcal{M}}(f)\;, \end{aligned} \tag{16}$$

$\alpha = (m-\gamma)/m$ and ϵ_n is an error residual that goes to zero w.p.1 as $n \to \infty$.

The additive model (15) could be the basis for many different methods for estimation of m and H. For example, we could invoke a central limit theorem on the MST length functional [34] to motivate a Gaussian approximate to ϵ_n and apply maximum likelihood principles. However, in this paper we adopt a simpler non-parametric least squares strategy which is based on resampling from the population $\mathcal{Y}_n$ of available points in $\mathcal{M}$. The proposed algorithm is summarized in Table II. 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 data sets $\mathcal{Y}_p^j$, $j = 1, \ldots, N$, 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 GMST length functionals $\overline{L}_p = N^{-1}\sum_{j=1}^N \hat{L}_\gamma^{\mathcal{M}}(\mathcal{Y}_p^j)$. Defining $\overline{l} = [\log\overline{L}_{p_1}, \ldots, \log\overline{L}_{p_Q}]^T$, and motivated by (15) we write down the linear model

$$\overline{l} = A \begin{bmatrix} a \\ b \end{bmatrix} + \epsilon\;, \tag{17}$$

where

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

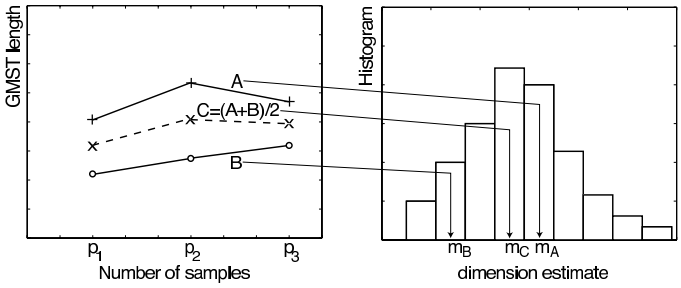


Fig. 1. Computing the dimension estimators by averaging over the length functional values, i.e., $(M, N) = (1, N)$ (dashed line), or by averaging over the dimension estimates, i.e., $(M, N) = (M, 1)$ (solid lines).

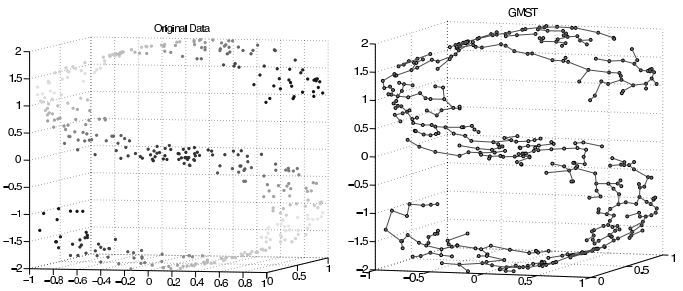


Fig. 2. The S-shaped surface manifold and corresponding GMST ($k = 7$) graph on 400 sample points.

Expressing a and b explicitly as functions of m and H_α via (16), the dimension and entropy quantities could be estimated using a combination of non-linear least squares (NLLS) and integer programming. Instead we take a simpler method-of-moments (MOM) approach in which we use (17) to solve for the linear least squares (LLS) estimates $\hat{a}, \hat{b}$ of a, b followed by inversion of the relations (16). After making a simple large n approximation, this approach yields the following estimates:

$$\begin{aligned} \hat{m} &= \text{round}\{\gamma/(1-\hat{a})\} \\ \hat{H}_\alpha^{(\mathcal{M},g)} &= \frac{\hat{m}}{\gamma}\left(\hat{b} - \log \beta_{\hat{m}}\right) . \end{aligned} \tag{18}$$

It is easily shown that the law of large numbers and Thm. 1 imply that these estimators are consistent as $N, n \to \infty$. We omit the details.

By running the algorithm M times independently over the population $\mathcal{Y}_n$, one obtains M estimates, $\{\hat{m}_i, \hat{H}_i\}_{i=1}^M$, that can be averaged to obtain final regularized dimension and entropy estimators, $\hat{m} = \sum \hat{m}_i/M$ and $\hat{H} = \sum \hat{H}_i/M$. The role of parameter M, together with parameter N, is to provide a tradeoff between the bias and variance performance of the estimators for finite n. The two cases of interest (considered in the next section) are $(M, N) = (1, N)$ and $(M, N) = (M, 1)$. In the first case, the smoothing is performed on the GMST length functional values before dimension and entropy are estimated, resulting in low variance but possibly high bias. In the second case, the smoothing is performed directly on the dimension and entropy estimates, resulting in higher variance but less bias.

Fig. 1 shows a graphical illustration of the smoothing step of the algorithm. Left panel shows $N = 2$ resampled GMST lengths, labeled "+" and "o", along with their average, labeled "x" for GMSTs built on $p_1 < p_2 < p_3$ randomly chosen vertices. For $(M, N) = (1, N)$, a linear least squares fit to the average GMST trajectory, $C = (A+B)/2$, is used to compute the dimension estimate $\hat{m}_C$. For $(M, N) = (M, 1)$, dimension estimates $\hat{m}_A$ and $\hat{m}_B$ are computed from sub-trajectories A and B, forming a histogram from which a final estimate can be computed.

A word about determination of the sequence of constants $\{\beta_m\}_m$ is in order. First of all, in the large n regime for which the above estimates were derived, β_m is not required for the dimension estimator. β_m is the limit of the normalized length functional of the Euclidean MST for a uniform distribution on the unit cube $[0, 1]^m$. Closed form expressions are not available but several approximations and bounds can be used in various regimes of m [20], [35]. For example, one could use the large m approximation of Bertsimas and van Ryzin [36]: $\log \beta_m \approx \gamma/2 \ \log(m/2\pi e)$. Another strategy, adopted in this paper, is to determine β_m by simulation of the Euclidean MST length on the m-dimensional cube for uniform random samples.

Before turning to applications we briefly discuss computational issues. For n samples, computing the MST scales as $O(n \log n)$, for which we have implemented Kruskal's algorithm [29]. On the other hand, the geodesic distances needed to compute the GMST require $O(n^2 \log n)$ operations using Dijkstra's algorithm multiple times. Thus, like ISOMAP, the GMST has overall $O(n^2 \log n)$ computational complexity.

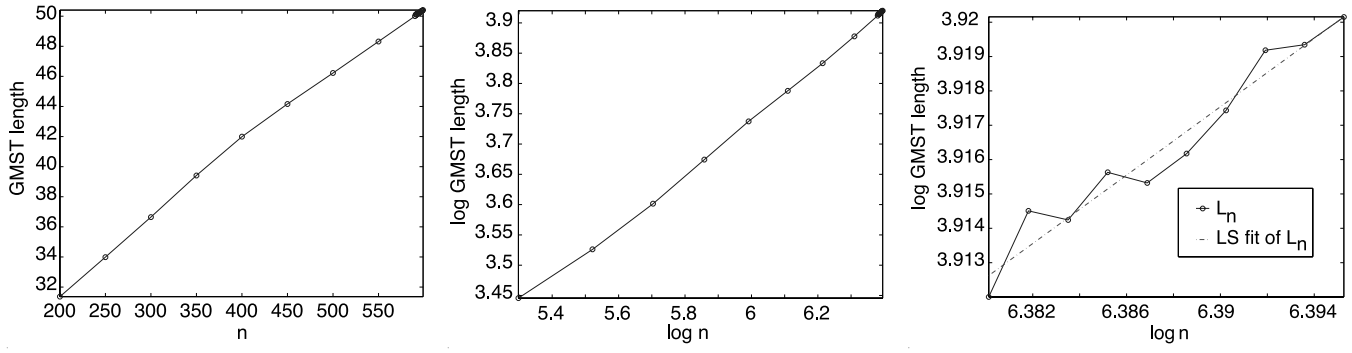
V. Applications

We illustrate the performance of the GMST algorithm on manifolds of known dimension as well as on a real data set consisting of face images. In all the simulations we fixed the parameters $\gamma = 1$ and $p_1 = n - Q, \ldots, p_Q = n - 1$. We also used the k-rule method, as described in table I, to estimate geodesic lengths. With regards to intrinsic dimension estimation, we compare our algorithm 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.

A. S-Shaped Surface

The first manifold considered is the standard 2-dimensional S-shaped surface [2] embedded in $\mathbb{R}^3$ (Fig. 2). Fig. 3 shows the evolution of the average GMST length $\bar{L}_n$ as a function of the number of samples, for a random set of i.i.d. points uniformly distributed on the surface.

To compare the dimension estimation performance of the GMST 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 follow a standard PCA order estimation procedure. Specifically, we graph the residual variance of the MDS fit as a function of the PCA dimension and try to detect the "elbow" at which residuals cease to



(a) (b) (c)

Fig. 3. Illustration of GMST dimension estimation for $(M, N) = (1, N)$: (a) plot of the average GMST length $\bar{L}_n$ for the S-shaped manifold as a function of the number of samples; (b) log-log plot of (a); (c) blowup of the last ten points in (b) and its linear least squares fit. The estimated slope is $\hat{a} = 0.4976$ which implies $\hat{m} = 2$. ($k = 7,\ M = 1,\ N = 5$).

decrease "significantly" as estimated dimension increases [1]. The elbow detector is implemented by a simple minimum angle threshold rule. Table III shows the results of this experiment. As it can be observed, the GMST algorithm outperforms ISOMAP in terms of dimension estimation error rates for small sample sizes. Fig. 4 shows the histogram of the entropy estimates for the same experiment.

B. Hyper-Planes

Next, we investigated linear 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 IV shows the results of running a Monte Carlo simulation under the same conditions as in the previous subsection. When $M = 1$ (i.e., least squares applied to the average length functional values), the GMST method showed a tendency to underestimate the correct dimension at smaller sample sizes. However, by taking $N = 1$ instead (i.e., averaging of least squares dimension estimates), this negative bias was eliminated and the GMST performed as well as the ISOMAP, which was observed to correctly predict the dimension for all sample sizes investigated.

Of course, as expected, 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 (14) to settle in and validate the limit in Corollary 1.

TABLE III

NUMBER OF CORRECT ISOMAP AND GMST DIMENSION ESTIMATES OVER 30 TRIALS AS A FUNCTION OF THE NUMBER OF SAMPLES FOR THE S-SHAPED MANIFOLD ($k = 7$).

n	200	400	600
ISOMAP	23	29	30
GMST ($M = 1, N = 5, Q = 10$)	29	30	30

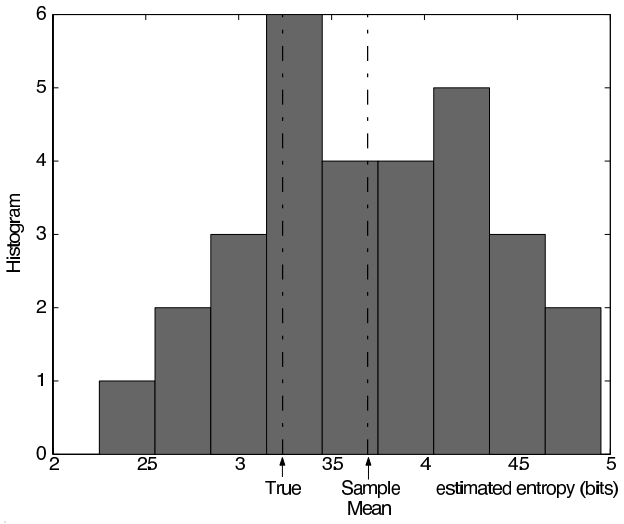


Fig. 4. Histogram of GMST entropy estimates over 30 trials of 600 samples uniformly distributed on the S-shaped manifold ($k = 7$, $M = 1$, $N = 5$, $Q = 10$). True entropy ("true") was computed analytically from the area of S curve supporting the uniform distribution of manifold samples.

TABLE IV

NUMBER OF CORRECT GMST DIMENSION ESTIMATES OVER 30 TRIALS AS A FUNCTION OF THE NUMBER OF SAMPLES FOR HYPER-PLANES ($k = 5$).

Hyper-plane dimension	Q	M	N	n = 600	n = 800	n = 1000
2	10	1	5	30	30	30
		5	1	30	30	30
3	10	1	5	24	24	27
		5	1	25	26	27
	15	1	10	30	30	30
		10	1	30	30	30
4	15	1	10	24	25	26
		10	1	27	28	28
	20	1	10	25	28	29
		10	1	29	29	30

Fig. 5. Samples from ISOMAP face database [1].

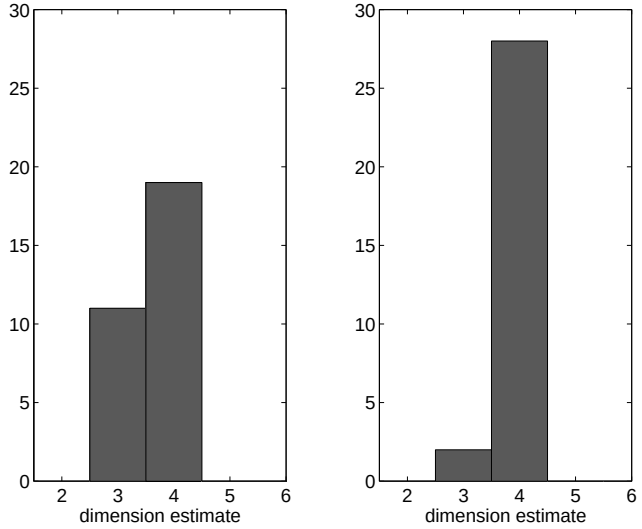


Fig. 6. GMST intrinsic dimension estimate histogram for the ISOMAP face database: left plot, $k = 6$, $M = 1$, $N = 10$, $Q = 15$; right plot, $k = 6$, $M = 10$, $N = 1$, $Q = 15$.

C. ISOMAP Face Database

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 (Fig. 5). 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 the histogram of the dimension estimates displayed in Figure 6. The estimated intrinsic dimension oscillates between 3 and 4, which, as in [5], deviates from the "informal" intrinsic dimension of 3 estimated by ISOMAP with thresholding. The estimated entropy was 21.8 bits, with a standard deviation of 0.5. Note that as $\alpha = (m-1)/m$ is close to one for the estimated values of m, the estimate of α-entropy is expected to be close to the Shannon entropy. This estimate suggests that one could, in theory, compress the ISOMAP face database, with little loss, using at most $21.8/(64 \times 64) \approx 0.005$ bits/pixel.

D. Yale Face Database B

Finally, we applied the GMST method to a real data set, and, consequently, of unknown manifold structure, intrinsic dimension and intrinsic entropy. We chose the set of 256 gray levels images of several individuals taken from the Yale Face Database B [37]. This is a publicly available database[1] containing a number of portfolios of face images under 585 different viewing conditions for each subject (Figure 7). Each portfolio consists of 9 poses and 65 illumination conditions (including ambient lighting) for each subject. The images were taken against a fixed background which we did not bother to segment out. 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.

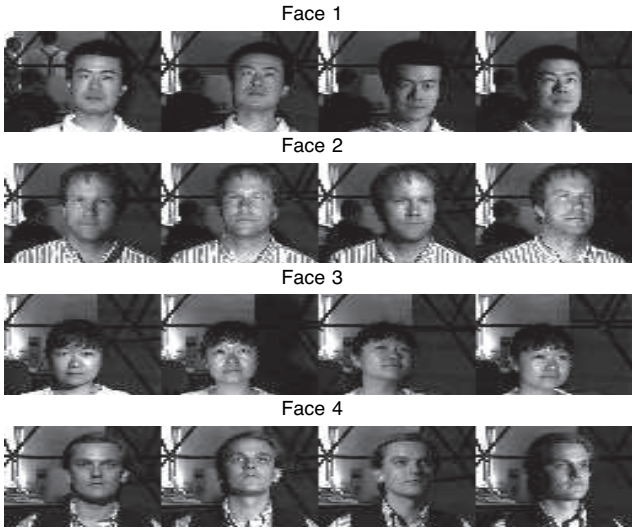


Fig. 7. Samples from Yale face database B [37].

Fig. 8 displays the results of running 30 trials of the algorithm using face 2. The first panel shows the real valued estimates of the intrinsic dimension, i.e., estimates obtained before the rounding operation in (18). Any value that falls in between the dashed lines will then be rounded to the integer at the midpoint. The second panel of Fig. 8 shows the histogram for these rounded estimates over the 30 generated trials. The intrinsic dimension estimate is between 5 and 6. Fig. 9 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. The simple minimum angle threshold rule on ISOMAP produced estimates between 3 and 6. Table V summarizes the results of the GMST method for the four faces. The intrinsic entropy estimates expressed in log base 2 were between 24.9 and 28 bits. Similarly to the ISOMAP face database, as α is close to one, these values suggest that the portfolio of a person's face image could be accurately

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

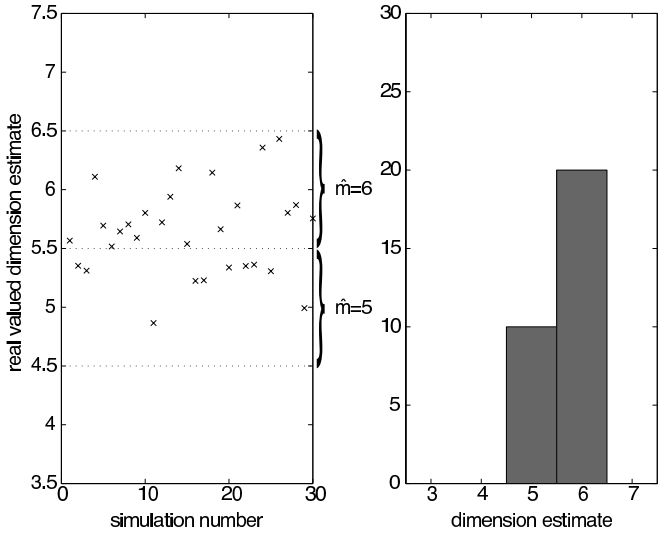


Fig. 8. GMST real valued intrinsic dimension estimates and histogram for face 2 in the Yale face database B ($k = 7$, $M = 1$ $N = 10$, $Q = 20$).

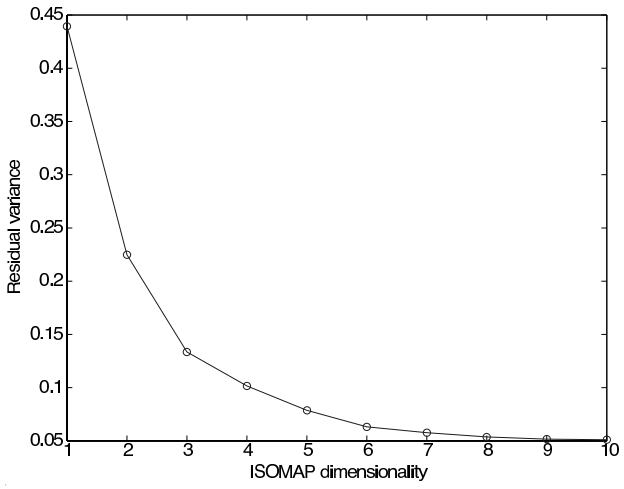


Fig. 9. ISOMAP ($k = 7$) residual variance for face 2 in the Yale face database B.

compressed using at most $28/(64 \times 64) \approx 0.007$ bits/pixel.

VI. Conclusion

We have introduced a novel method for intrinsic dimension and entropy estimation based on the growth rate of the geodesic total edge length functional of entropic graphs. The proposed method has two main advantages. First, it is global in the sense that the MST is constructed over the entire set and thus avoids local linearizations. Second, it does not require reconstructing the manifold or estimating the multivariate density of the samples. We validated the new method by testing it on synthetic manifolds of known dimension and on high dimensional real data sets.

One drawback of GMST, or any other dimension estimator based on ISOMAP geodesic fitting to data, is the restriction to isometric embeddings. We are currently working on extending Thm. 1 and Corollary 1 to general (non-isometric) Riemann manifolds, thus avoiding any assumptions about global embeddings and eliminating the effect of the Jacobian on the intrinsic entropy. We are also studying the use of entropic graphs that bypass the complex step of geodesic estimation. In particular, in [38], we consider k-nearest neighbor graphs due to their low complexity and local properties. Future work includes the characterization of the statistics in the linear model (15), optimization of the bias/variance tradeoff parameters of the GMST algorithm and the study of the effect of additive noise on the manifold samples.

TABLE V
GMST dimension estimates $\hat{m}$ and entropy estimates $\hat{H}$ for four faces in the Yale Face Database B.

	Face1	Face2	Face3	Face 4
$\hat{m}$	6	6	7	7
$\hat{H}$ (bits)	24.9	26.4	25.8	28.0

Acknowledgment

The authors acknowledge and thank the creators of Yale Face Database B for making their face image data publicly available. The authors would also like to thank Huzefa Neemuchwala and Arpit Almal for their help in acquiring and processing these face images.

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

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

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

[6] 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.

[7] V. I. Koltchinskii, "Empirical geometry of multivariate data: A deconvolution approach," *Annals of Statistics*, vol. 28, no. 2, pp. 591–629, 2000.

[8] 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.

[9] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. on Inform. Theory*, vol. 45, no. 6, pp. 921–1939, September 1999.

[10] A. Hero, J. Costa, and B. Ma, "Convergence rates of minimal graphs with random vertices," submitted to *IEEE Trans. on Inform. Theory*, 2002, `www.eecs.umich.edu/~hero/det_est.html`.

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

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

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

[14] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems, Volume 14*, T. G. Diettrich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002.

[15] 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.

[16] 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.

[17] 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.

[18] F. Memolia, G. Sapiro, and S. Osher, "Solving variational problems and partial differentyial equations mapping into general target manifolds," Tech. Rep. 1827, IMA, January 2003.

[19] F. Memoli and G. Sapiro, "Fast computation of weighted distance functions and geodesics on implicit hyper-surfaces," *Journal of Computationals Physics*, vol. 73, pp. 730–764, 2001.

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

[21] M. Carmo, *Differential geometry of curves and surfaces*, Prentice-Hall, Englewood Cliffs, N.J., 1976.

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

[23] W. Boothby, *An introduction to differentiable manifolds and Riemannian geometry*, Academic, San Diego, Calif., rev. 2nd edition, 2003.

[24] J. Beardwood, J. H. Halton, and J. M. Hammersley, "The shortest path through many points," *Proc. Cambridge Philosophical Society*, vol. 55, pp. 299–327, 1959.

[25] J. M. Steele, *Probability theory and combinatorial optimization*, vol. 69 of *CBMF-NSF Regional Conferences in Applied Mathematics*, Society for Industrial and Applied Mathematics (SIAM), 1997.

[26] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. on Inform. Theory*, vol. 28, pp. 373–380, 1979.

[27] D. N. Neuhoff, "On the asymptotic distribution of the errors in vector quantization," *IEEE Trans. on Inform. Theory*, vol. 42, pp. 461–468, March 1996.

[28] A.O. Hero and O. Michel, "Estimation of Rényi information divergence via pruned minimal spanning trees," in *IEEE Workshop on Higher Order Statistics*, Caesaria, Israel, Jun. 1999.

[29] H. Neemuchwala, A. O. Hero, and P. Carson, "Image registration using entropy measures and entropic graphs," *to appear in European Journal of Signal Processing, Special Issue on Content-based Visual Information Retrieval*, 2003.

[30] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[31] I. Csiszar, "Generalized cutoff rates and Rényi's information measures," *IEEE Trans. on Inform. Theory*, vol. 41, no. 1, pp. 26–34, January 1995.

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

[33] V. de Silva and J. B. Tenenbaum, "Unsupervised learning of curved manifolds," in *Nonlinear estimation and classification*, D.D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu, Eds. Springer-Verlag, New York, 2002.

[34] K. S. Alexander, "The RSW theorem for continuum percolation and the CLT for Euclidean minimal spanning trees," *Ann. Applied Probab.*, vol. 6, pp. 466–494, 1996.

[35] F. Avram and D. Bertsimas, "The minimum spanning tree constant in geometrical probability and under the independent model: a unified approach," *Ann. Applied Probab.*, vol. 9, pp. 223–231, 1990.

[36] D. Bertsimas and G. van Ryzin, "An aysmptotic determination of the minimum spanning tree and minimum matching constants in geometrical probability," *Oper. Research Letters*, vol. 2, pp. 113–130, 1992.

[37] 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.

[38] 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.

Jose A. Costa (S'99) was born in Lisbon, Portugal. He received the Licenciatura degree in Aerospace Engineering (1998) from Instituto Superior Técnico, Technical University of Lisbon, Portugal. He received the M.S. degree in Electrical Engineering (2002) and the M.A. degree in Statistics (2003) from the University of Michigan, Ann Arbor, where he is currently pursuing the Ph.D. degree in Electrical Engineering.

He performed research at the Communication Theory and Pattern Recognition group, Institute of Telecommunications, Lisbon, between 1998 and 2000 in statistical algorithms for positioning and navigation. His research interests include statistical signal and image processing, nonparametric estimation, pattern recognition and machine learning.

Mr. Costa was awarded a Ph.D. fellowship in 2001 by Fundação para a Ciência e Tecnologia (Portugal) and is currently a Rackham Predoctoral Fellow at the University of Michigan.

Alfred O. Hero, III (S'79-M'80-SM'96-F'98) received the B.S. in Electrical Engineering (summa cum laude) from Boston University (1980) and the Ph.D from Princeton University (1984), both in Electrical Engineering. Since 1984 he has been a Professor with the University of Michigan, Ann Arbor, where he has appointments in the Department of Electrical Engineering and Computer Science, the Department of Biomedical Engineering and the Department of Statistics.

He has held visiting positions at I3S University of Nice, Sophia-Antipolis, France (2001), Ecole Normale Supérieure de Lyon (1999), Ecole Nationale Supérieure des Télécommunications, Paris (1999), Scientific Research Labs of the Ford Motor Company, Dearborn, Michigan (1993), Ecole Nationale Superieure des Techniques Avancees (ENSTA), Ecole Superieure d'Electricite, Paris (1990), and M.I.T. Lincoln Laboratory (1987 - 1989). His research interests are in areas including: estimation and detection, statistical communications, bioinformatics, signal processing and image processing.

He is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE), a member of Tau Beta Pi, the American Statistical Association (ASA), the Society for Industrial and Applied Mathematics (SIAM), and the US National Commission (Commission C) of the International Union of Radio Science (URSI). He has received the 1998 IEEE Signal Processing Society Meritorious Service Award, the 1998 IEEE Signal Processing Society Best Paper Award, and the IEEE Third Millenium Medal. In 2002 he was appointed IEEE Signal Processing Society Distinguished Lecturer.