LEARNING ON STATISTICAL MANIFOLDS FOR CLUSTERING AND VISUALIZATION

Kevin M. Carter[1]*, *Raviv Raich*[2], *and Alfred O. Hero III*[1]

[1] Department of EECS, University of Michigan, Ann Arbor, MI 48109
[2] School of EECS, Oregon State University, Corvallis, OR 97331
{kmcarter,hero}@umich.edu, raich@eecs.oregonstate.edu

ABSTRACT

We consider the problem of analyzing data for which no straight forward and meaningful Euclidean representation is available. Specifically, we would like to perform dimensionality reduction to such data for visualization and as preprocessing for clustering. In these cases, an appropriate assumption would be that the data lies on a *statistical* manifold, or a manifold of probability density functions (PDFs). In this paper we propose using the properties of information geometry in order to define similarities between data sets. This has been done using the Fisher information distance, which requires knowledge of the parametrization of the manifold; knowledge which is usually unavailable. We will show this metric can be approximated using entirely non-parametric methods. Furthermore, by using multi-dimensional scaling (MDS) methods, we are able to embed the corresponding PDFs into a low-dimensional Euclidean space. We illustrate these methods on simulated data generated by known statistical manifolds. Rather than as an analytic or quantitative study, we present this framework as a proof of concept, demonstrating our methods which are immediately applicable to problems of practical interest.

1. INTRODUCTION

The fields of statistical learning and machine learning are used to study problems of inference, which is to say gaining knowledge through the construction of models in order to make decisions or predictions based on observed data [1]. In some problems, the observations can be represented as points in a Euclidean space with the $L2$-norm as a natural dissimilarity metric. Solutions to problems of dimensionality reduction, clustering, classification have been formulated using the Euclidean representation. Unfortunately, when no obvious natural Euclidean representation for the data is available, such inference tasks require independent solutions. A straightforward strategy is to express the data in terms of a low dimensional feature vector for which the curse of dimensionality is alleviated. This initial processing of data as real-valued feature vectors in Euclidean space, which is often carried out in an ad hoc manner, has been called the "dirty laundry" of machine learning [2]. This procedure is highly dependent on having a good model for the data and in the absence of such model may be highly suboptimal, resulting in much information loss. When a statistical model is available, the process of obtaining a feature vector can be done optimally by extracting the model parameters for a given data set and thus characterizing the data through its lower dimensional parameter vector. We are interested in extending this approach to the case in which the data follows an unknown parametric statistical model. These types of problems have been presented in document classification [3], face recognition [4], texture segmentation [5], and shape analysis [6].

In this paper, we present a framework to handle such problems in which a model for the data is unavailable. Specifically, we focus on the case where the data is high-dimensional and no lower dimensional Euclidean manifold gives a sufficient description. In many of these cases, a lower dimensional statistical manifold can be used to assess the data for various learning tasks. Our framework includes characterization of data sets in terms of a non-parametric statistical model, a geodesic distance as an information metric for evaluating distance between data sets, and a dimensionality reduction procedure to obtain a feature vector representation of a high-dimensional data set for the purposes of both clustering and visualization. While none of our methodologies are individually uncommon, the combination of them all into a common framework is, to our knowledge, a link which has not yet been presented.

This paper is organized as follows: Section 2 describes a background in information geometry and statistical manifolds. Section 3 gives the formulation for the problem we wish to solve, while Section 4 develops the algorithm for the methods we use. We illustrate the results of using our methods in Section 5. Finally, we draw conclusions and discuss the possibilities for future work in Section 6.

*Acknowledgement: This work is partially funded by the National Science Foundation, grant No. CCR-0325571.

2. BACKGROUND ON INFORMATION GEOMETRY

Information geometry is a field that has emerged from the study of geometrical structures on manifolds of probability distributions. It is largely based on the works of Shun'ichi Amari [7] and has been used for analysis in such fields as statistical inference, neural networks, and control systems. In this section, we will give a brief background on the methods of information geometry that we utilize in our framework. For a more thorough introduction to information geometry, we suggest [8] and [9].

2.1. Statistical Manifolds

Let us now present the notion statistical manifolds, or a set $\mathcal{M}$ whose elements are probability distributions. A probability density function (PDF) on a set $\mathcal{X}$ is defined as a function $p : \mathcal{X} \to \mathbb{R}$ in which

$$p(x) \geq 0, \forall x \in \mathcal{X} \tag{1}$$
$$\int p(x)\, dx = 1.$$

We describe only the case for continuum on the set $\mathcal{X}$, however if $\mathcal{X}$ was discrete valued, equation (1) will still apply by switching $\int p(x)\, dx = 1$ with $\sum p(x) = 1$. If we consider $\mathcal{M}$ to be a family of PDFs on the set $\mathcal{X}$, in which each element of $\mathcal{M}$ is a PDF which can be parameterized by θ, then $\mathcal{M}$ is known as a statistical model on $\mathcal{X}$. Specifically, let

$$\mathcal{M} = \{p(x \mid \theta) \mid \theta \in \Theta \subseteq \mathbb{R}^n\}, \tag{2}$$

with $p(x \mid \theta)$ satisfying the equations in (1). Additionally, there exists a one-to-one mapping between θ and $p(x \mid \theta)$.

Given certain properties of the parametrization of $\mathcal{M}$, such as differentiability and C^∞ diffeomorphism (details of which are described in [9]), the parametrization θ is also a coordinate system of $\mathcal{M}$. In this case, $\mathcal{M}$ is known as a statistical manifold. In the rest of this paper, we will use the terms 'manifold' and "statistical manifold" interchangeably.

2.2. Fisher Information Metric

In Euclidean space, the distance between two points is defined as the length of a straight line between the points. On a manifold, however, one can measure distance by a trace of the shortest path between the points along the manifold. This path is called a geodesic, and the length of the path is the geodesic distance. In information geometry, the distance between two points on a manifold is analogous to the difference in information between them, and is defined by the Fisher information metric. This measures the amount of information a random variable X contains in reference to an unknown parameter θ. For the single parameter case it is defined as

$$\mathcal{I}(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2 |\theta\right].$$

If the condition $\int \frac{\partial^2}{\partial\theta^2} f(X;\theta)\, dX = 0$ is met, then the above equation can be written as

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]. \tag{3}$$

For the case of multiple unknown parameters $\theta = \left[\theta^1, \ldots, \theta^n\right]$, we define the Fisher information matrix $[\mathcal{I}(\theta)]$, whose elements consist of the Fisher information with respect to specified parameters, as

$$\mathcal{I}_{ij} = \int f(X;\theta)\frac{\partial \log f(X;\theta)}{\partial\theta^i}\frac{\partial \log f(X;\theta)}{\partial\theta^j}\, dX.$$

For a parametric family of PDFs, it is possible to define a Riemannian metric using the Fisher information matrix, known as the information metric. This was first presented by Cramér and Rao. The information-metric distance, or Fisher information distance, between two distributions $p(x;\theta_1)$ and $p(x;\theta_2)$ in a single parameter family is

$$D_F(\theta_1,\theta_2) = \int_{\theta_1}^{\theta_2} \mathcal{I}(\theta)^{1/2} d\theta,$$

where θ_1 and θ_2 are parameter values corresponding to the two PDFs and $\mathcal{I}(\theta)$ is the Fisher information for the parameter θ. Extending to the multi-parameter case, we obtain:

$$D_F(\theta_1,\theta_2) = \min_{\theta(\cdot):\theta(0)=\theta_1,\theta(1)=\theta_2} \int_0^1 \sqrt{\left(\frac{d\theta}{d\beta}\right)^T \mathcal{I}(\theta)\left(\frac{d\theta}{d\beta}\right)}\, d\beta. \tag{4}$$

The metric in (4) is key to our approach as it provides an information-based means of comparing PDFs on the appropriate statistical manifold. The shortest path θ^* that minimizes (4) does so by considering only routes which lie on the manifold, guaranteeing that each point along the path θ^* is a PDF governed by the $\mathcal{M}$. Other distances that do not restrict measured paths to the manifold may lead to inaccurate "short cut" distances; ie paths that consist of PDFs not governed by $\mathcal{M}$. This is clearly the case with the $L2$-distance, which only considers the straight-line path between points.

3. PROBLEM FORMULATION

We restrict our attention to problems in which clustering and visualization of PDFs is desired. A key property of the Fisher information metric is that it is independent of the parametrization of the manifold [8]. Although the evaluation remains equivalent, calculating the FIM requires knowledge of the parametrization, which is generally not available. We instead assume that the collection of density functions lie on a manifold that can be described by some natural parametrization. Specifically, we are given $\mathcal{P} = \{p_1, \ldots, p_n\}$, where $p_i \in \mathcal{M}$ is a probability density function and $\mathcal{M}$ is a manifold embedded in $\mathbb{R}^d$. Our goal is to find an approximation for the

geodesic distance between points on $\mathcal{M}$ using only the information available in $\mathcal{P}$. Can we find an approximation function G which yields

$$\hat{D}_F(p_i, p_j) = G(p_i, p_j; \mathcal{P}), \tag{5}$$

such that $\hat{D}_F(p_i, p_j) \to D_F(p_i, p_j)$ as $n \to \infty$?

This problem is similar to the setting of classical papers [10, 11] in manifold learning and dimensionality reduction, where only a set of points on the manifold are available. As such, we are able to use these manifold learning techniques to construct a low-dimensional, information based embedding of that family. This not only allows for an effective visualization of the manifold (in 2 or 3 dimensions), but by embedding the family into a Euclidean space we can perform clustering of the PDFs lying on the manifold with existing Euclidean methods.

3.1. Approximation of Fisher Information Distance

Many metrics have been defined to approximate the Fisher information distance when the specific parameterization of the manifold is unknown. An important class of such divergences is known as the f-divergence [12], in which $f(u)$ is a convex function on $u > 0$ and

$$D_f(p\|q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right).$$

A specific and important example of the f-divergence is the α-divergence, where $D^{(\alpha)} = D_{f^{(\alpha)}}$ for a real number α. The function $f^{(\alpha)}(u)$ is defined as

$$f^{(\alpha)}(u) = \begin{cases} \frac{4}{1-\alpha^2}\left(1 - u^{(1+\alpha)/2}\right) & \alpha \neq \pm 1 \\ u \log u & \alpha = 1 \\ -\log u & \alpha = -1 \end{cases}.$$

As such, the α-divergence can be evaluated as

$$D^{(\alpha)}(p\|q) = \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{\frac{1-\alpha}{2}} q(x)^{\frac{1+\alpha}{2}} dx\right) \quad \alpha \neq 1,$$

and

$$D^{(-1)}(p\|q) = D^{(1)}(q\|p) = \int p(x) \log \frac{p(x)}{q(x)}. \tag{6}$$

The α-divergence is the basis for many important and well known divergence metrics, such as the Hellinger distance, the Kullback-Leibler divergence (6), and the Renyi-Alpha entropy [13].

3.1.1. Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence is defined as

$$KL(p\|q) = \int p(x) \log \frac{p(x)}{q(x)}, \tag{7}$$

which is equal to $D^{(-1)}$ (6). The KL-divergence is a very important metric in information theory, and is commonly referred to as the relative entropy of a probability distribution. The KL-divergence is related to the Hellinger distance, as in the limit $\sqrt{KL(p\|q)} \to D_H(p, q)$. Kass and Vos also show the relation between the Kullback-Leibler divergence and the Fisher information distance, $\sqrt{2KL(p\|q)} \to D_F(p, q)$ as $p \to q$.

It should be noted that the KL-divergence is not a distance metric, as it does not satisfy the symmetry, $KL(p\|q) \neq KL(p\|q)$, or triangle inequality properties of a distance metric. To obtain this symmetry, we will define the the KL-divergence as:

$$D_{KL}(p, q) = KL(p\|q) + KL(q\|p), \tag{8}$$

which is symmetric, but still not a distance as it does not satisfy the triangle inequality. Since the Fisher information is a symmetric measure,

$$\sqrt{2KL(q\|p)} \to D_F(q, p) = D_F(p, q). \tag{9}$$

Combining (8) and (9), we can approximate the Fisher information distance as

$$\sqrt{D_{KL}(p, q)} \to D_F(p, q), \tag{10}$$

as $p \to q$.

The Fisher information metric is not the only method for calculating a similarity between PDFs. Another common method is using the L_2 distance, or the integrated squared error (ISE). In this case, the distance between two densities, $p(x)$ and $q(x)$, is defined as

$$D(p, q) = \int \left(p(x) - q(x)\right)^2 dx.$$

We choose to use the KL-divergence as it is a great means of differentiating shapes of densities. Analysis of (7) shows that as $p(x)/q(x) \to \infty$, $KL(p\|q) \to \infty$. This property ensures that the KL-divergence will be amplified in regions where there is a significant difference in the probability distributions. As such, the difference in the tails of the distributions is a strong contributor to the KL-divergence.

3.2. Approximation of Distance on Statistical Manifolds

As noted earlier (10), $\sqrt{D_{KL}(p_1, p_2)} \to D_F(p_1, p_2)$ as $p_1 \to p_2$. If p_1 and p_2 do not lie closely together on the manifold, the Kullback-Leibler divergence becomes a weak approximation of the Fisher information distance. However, a good approximation can still be had if the manifold is densely sampled between the two end points by defining the path between p_1 and p_2 as a series of connected segments, and summing the length of those segments. Specifically, given the set of n probability density functions parameterized by $\mathcal{P} = \{\theta_1, \ldots, \theta_n\}$,

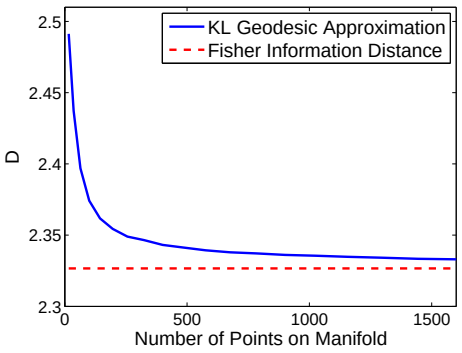


Fig. 1. Convergence of the graph approximation of the Fisher information distance using the Kullback-Leibler divergence. As the manifold is more densely sampled, the approximation approaches the true value.

the Fisher information distance between p_1 and p_2 can be approximated as:

$$D_F(p_1, p_2) \approx \min_{m, \{\theta_{(1)}, \ldots, \theta_{(m)}\}} \sum_{i=1}^{m} D_F(p(\theta_{(i)}), p(\theta_{(i+1)}))$$

where $p(\theta_{(1)}) = p_1$, $p(\theta_{(m)}) = p_2$, and $\{\theta_{(1)}, \ldots, \theta_{(m)}\} \in \mathcal{P}$. We can now form an approximation of the Fisher information distance using the Kullback-Leibler divergence for distant points on the manifold:

$$D_F(p_1, p_2) \approx \min_{m, \{p_{(1)}, \ldots, p_{(m)}\}} \sum_{i=1}^{m} \sqrt{D_{KL}(p_{(i)}, p_{(i+1)})}$$

where $p_{(1)} = p_1$ and $p_{(m)} = p_2$. Intuitively, this estimate calculates the length of the shortest path between points in a connected graph on the well sampled manifold. This is similar to the manner in which Isomap [10] approximates distances on Euclidean manifolds, which is further elaborated in Section 3.3.2.

Figure 1 illustrates this approximation by comparing the KL graph approximation to the actual Fisher information distance for the univariate gaussian case. The KL-divergence between univariate normal distributions is available in a closed-form expression:

$$KL(p_1 \| p_2) = \frac{1}{2}\left(\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{\sigma_1^2}{\sigma_2^2} + (\mu_2 - \mu_1)^2 / \sigma_2^2 - 1\right),$$

while the closed-form expression for the Fisher information distance is presented in [14]. As the manifold is more densely sampled (uniformly in mean and variance parameters for this simulation), the approximation converges to a very close approximation of the true Fisher information distance for the univariate normal case, as calculated in [14].

3.3. Dimensionality Reduction

Given a matrix of dissimilarities between entities, many algorithms have been developed to find a low dimensional embedding of the original data $\psi : \mathcal{M} \rightarrow \mathbb{R}^n$. These techniques have been classified as a group of methods referred to as Multi-Dimensional Scaling (MDS). There are supervised methods, which are generally used for classification purposes, and unsupervised methods, which are often used for clustering and manifold learning. For our purposes, we will use unsupervised methods, which will reveal any natural separation or clustering of the data sets. Using these MDS methods allows us to find a single low-dimensional coordinate representation of each high-dimensional, large sample, data set.

3.3.1. Classical Multi-Dimensional Scaling

Classical MDS takes a matrix of dissimilarities and converts them into Cartesian coordinates. This is performed by first centering the dissimilarities about the origin, then calculating the singular value decomposition (SVD) of the centered matrix. This permits the calculation of the low-dimensional embedding coordinates.

Define D as a dissimilarity matrix of Euclidean distances (may also approximate Euclidean distances). Let B be the "double centered" matrix which is calculated by taking the matrix D, subtracting its row and column means, then adding back the grand mean and multiplying by $-\frac{1}{2}$. As a result, B is a version of D centered about the origin. Mathematically, this process is solved by

$$B = -\frac{1}{2} H D^2 H,$$

where $H = I - (1/N)11^T$, D^2 is the matrix of squared distances, I is the N-dimensional identity matrix, and 1 is an N-element vector of ones.

The embedding coordinates, $\boldsymbol{Y} \in \mathbb{R}^{d \times n}$, can then be determined by taking the eigenvalue decomposition of B,

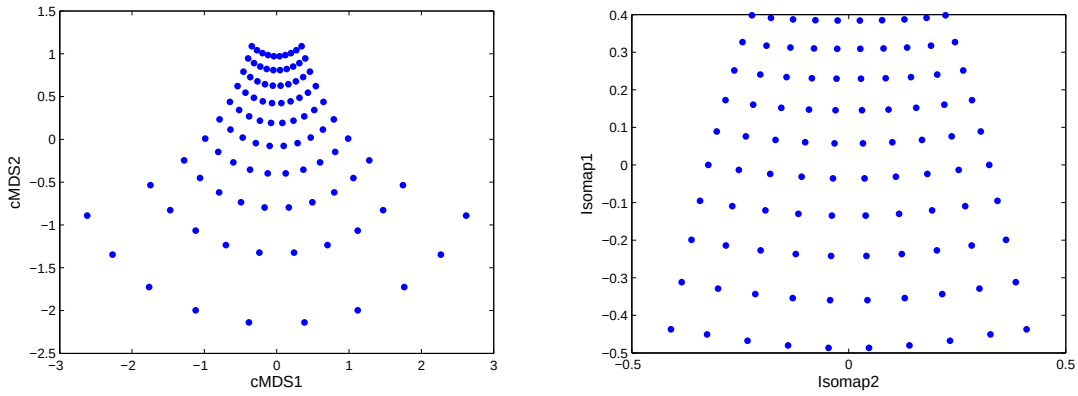
$$B = [V_1 V_2] \text{diag}\,(\lambda_1, ..., \lambda_N)\,[V_1 V_2]^T,$$

and calculating

$$\boldsymbol{Y} = \text{diag}\left(\lambda_1^{1/2}, ..., \lambda_d^{1/2}\right) V_1^T.$$

The matrix V_1 consists of the eigenvectors corresponding to the d largest eigenvalues $\lambda_1, \ldots, \lambda_d$ while the remaining $N - d$ eigenvectors are represented as V_2. $\text{diag}(\lambda_1, \ldots, \lambda_N)$ refers to an $N \times N$ diagonal matrix with λ_i as its i^{th} diagonal element.

For visualization, let us define a set of probability densities $\mathcal{P} = \{p_i(x)\}$ on a grid, such that $p_i = p_{k,l}$ is parameterized by $(\mu_i, \sigma_i) = (\alpha k, 1 + \beta l)$, $k, l = 1 \ldots n$ and $\alpha, \beta \in \mathbb{R}$. Additionally, let D be the matrix of Fisher information distances defined in [14] for the set of univariate normal densities $\mathcal{P}$, where $D(i, j) = D_F(p_i, p_j)$. Figure 2(a) displays the



(a) (b)

Fig. 2. a) Classical MDS and b) Isomap applied to the matrix of Fisher information distances on a grid of univariate normal densities, parameterized by (μ, σ)

results of applying cMDS to D. It is clear that while the densities defining the set $\mathcal{P}$ are parameterized on a rectangular grid, the manifold on which $\mathcal{P}$ lives is not rectangular itself.

3.3.2. ISOMAP

Isometric feature mapping (Isomap) is a technique developed by Tenenbaum et. al. and first presented in [10]. Isomap performs non-linear dimensionality reduction by utilizing classical MDS on an adjacency matrix which approximates the geodesic distance between data points. As such, this algorithm is able to discern low-dimensional structure in high dimensional spaces that were previously indiscernible with methods such as principal components analysis (PCA) and classical MDS. The algorithm contains three steps and works as follows:

1. Construct Neighborhood Graph
   Given dissimilarity matrix D_X between data points in in the set $\boldsymbol{X}$, define the graph G over all data points by adding an edge between points i and j if $\boldsymbol{X}_i$ is one of the k-nearest neighbors of $\boldsymbol{X}_j$.

2. Compute Shortest Paths
   Initialize $d_G(i,j) = d_X(i,j)$ if there is an edge between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, $d_G(i,j) = \inf$ otherwise. Complete $D_G = \{d_G(i,j)\}$ by computing $d_G(i,j)$ to be equal to the shortest path distance between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$.

3. Construct low-dimensional embedding
   Apply classical MDS to the dissimilarity matrix D_G to obtain $\boldsymbol{Y}$, a low-dimensional embedding of the set $\boldsymbol{X}$.

Figure 2(b) shows the results of Isomap (neighborhood size $k = 6$) on the dissimilarity matrix formed by the Fisher information distance of the univariate normal family of distributions. Notice how the paths become nearly linear between all points. This is a result of Isomap obtaining a low dimensional embedding based on the approximated geodesic distance between distant points, rather than the strict distance defined by the metric, which is only accurate as $p_i(x) \to p_j(x)$.

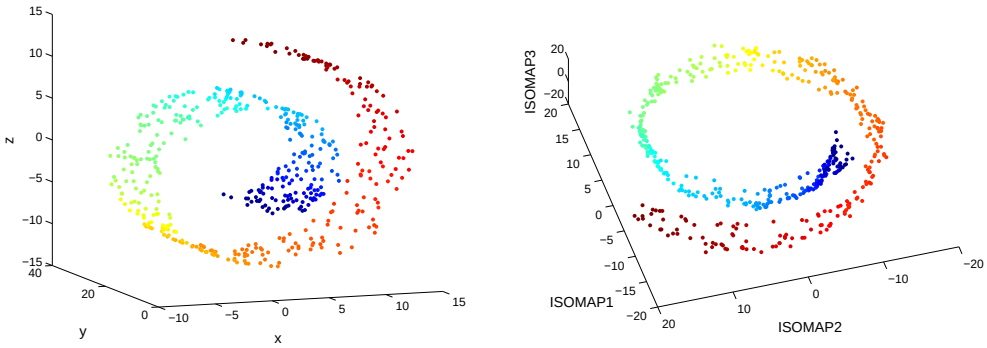
3.3.3. Additional MDS Methods

While we choose to only detail the cMDS and Isomap algorithms, there are many other methods for performing dimensionality reduction in a linear fashion (PCA) and non-linearly (Laplacian Eigenmaps [11] and Local Linear Embedding [15]) for unsupervised learning, all of which can be applied to our framework.

4. OUR TECHNIQUES

We have presented a series of methods for manifold learning developed in the field of information geometry. By performing dimensionality reduction on a family of data sets, we are able to both better visualize and cluster the data. In order to obtain a lower dimensional embedding, we calculate a dissimilarity metric between data sets within the family by approximating the Fisher information distance between their corresponding probability densities. This has been illustrated with the family of univariate normal PDFs.

In problems of practical interest, however, the parameterization of the PDFs are usually unknown. We instead are given a family of data sets $\mathcal{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$, in which we may assume that each data set $\boldsymbol{X}_i$ is a realization of some underlying probability distribution to which we do not have knowledge of the parameters. As such, we rely on non-parametric techniques to estimate both the probability density and the approximation of the Fisher information distance. We utilize kernel density estimation (KDE) methods for deriving our probability density function estimates, although mixture



(a) Swiss Roll (b) Isomap embedding

Fig. 3. Given a collection of data sets with a Gaussian distribution having means equal to points a sampled 'swiss roll' manifold, our methods are able to reconstruct the original statistical manifold from which each data set is derived.

Algorithm 1 Calculate d-dimensional manifold embedding

Input: Collection of data sets $\mathcal{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$ and the desired embedding dimension d

1: **for** $i = 1$ to N **do**
2: Calculate $\hat{p}_i(\boldsymbol{x})$, the density estimate of $\boldsymbol{X}_i$
3: **end for**
4: Calculate D_{KL}, where $D_{KL}(i,j) = KL(\hat{p}_i||\hat{p}_j) + KL(\hat{p}_j||\hat{p}_i)$
5: $\boldsymbol{Y} = \text{embed}(\sqrt{D_{KL}}, d)$

Output: d-dimensional embedding of $\mathcal{X}$, into Cartesian coordinates $\boldsymbol{Y} \in \mathbb{R}^{d \times N}$

models as well as other density estimation techniques will suffice as well. Following these approximations, we are able to perform the same multi-dimensional scaling operations as previously described.

4.1. Algorithm

Algorithm 1 combines all of the methods we have presented in order to find a low-dimensional embedding of a collection of data sets. If we assume each data set is a realization of an underlying probability density, and each of those densities lie on a manifold with some natural parameterization, then this embedding can be viewed as an embedding of the actual manifold into Cartesian coordinates. Note that in line 5, 'embed$(\sqrt{D_{KL}}, d)$' refers to using any multi-dimensional scaling method (such as Isomap, cMDS, Laplacian Eigenmaps, etc) to embed the dissimilarity matrix $\sqrt{D_{KL}}$ into Cartesian coordinates with dimension d.

5. APPLICATIONS

We now present simulations to illustrate our methods on sample problems. Our examples are not intended to be considered the desired usages of our methods. Rather, we use simple examples with known manifolds to demonstrate how are methods may be immediately applicable to problems of practical interest. We would like to stress that in the following examples, with even a minimum amount of a priori knowledge of the data, there are simple Euclidean methods for analysis. Our methods, however, are entirely non-parametric, make no assumptions of the data, and require no a priori knowledge.

5.1. Manifold Visualization

To demonstrate the ability of our methods to reconstruct the statistical manifold, we create a known manifold of densities. Let $\boldsymbol{Y} = \{y_1, \ldots, y_n\}$, where each y_i is uniformly sampled on the 'swiss roll' manifold (see Fig. 3(a)). Let $\mathcal{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ where each $\boldsymbol{X}_i$ is generated from a normal distribution $\mathcal{N}(y_i, \Sigma)$, where Σ is held constant for each density. As such, we have developed a statistical manifold of known parameterization, which is sampled by known PDFs. Utilizing our methods in an unsupervised manner, we are able to recreate the original manifold $\boldsymbol{Y}$ strictly from the collection of data sets $\mathcal{X}$. This is shown in Fig. 3(b) where each set is embedded into 3 Isomap dimensions, and the 'swiss roll' is reconstructed. While this embedding could easily be constructed using the mean of each set $\boldsymbol{X}_i$ as a Euclidean location, it illustrates that the Kullback-Leilber divergence along with an Isomap embedding can be used for visualizing the statistical manifold as well.

5.2. Clustering

We now illustrate the ability to cluster using our methods. We create data sets comprised of either the swiss roll or S-curve manifolds in Euclidean space. Specifically, let $\mathcal{X} = \{\mathcal{Y}_1, \mathcal{Y}_2\}$ where $\mathcal{Y}_1$ is a family of data sets uniformly sampled on the swiss roll, while the family $\mathcal{Y}_2$ contains data sets uniformly

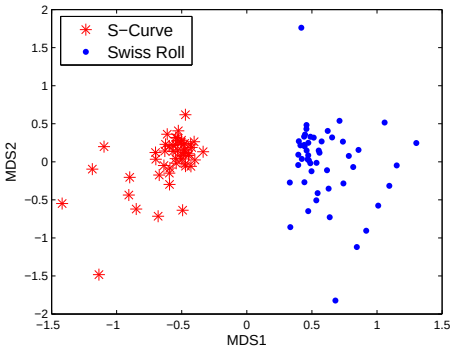


Fig. 4. When used for clustering, we can see the natural seperation arising from a set containing families of distributions drawn from uniform sampling on the swiss roll and S-curve manifolds.

sampled on the S-curve. Both the swiss roll and s-curve contain the same support and are centered about the same point. Utilizing the KL-divergence and cMDS, we are able to find a low-dimensional embedding of $\mathcal{X}$ (Fig. 4). It is clear that the two families $\mathcal{Y}_1$ and $\mathcal{Y}_2$ form distinct clusters, which is brought upon by the natural dissimilarity between the probability distributions each family is drawn from. While the procedure was performed entirely unsupervised, we assign each family a different marker when plotting to illustrate the distinct clusters.

6. CONCLUSIONS

Many problems of practical interest involve data sets which are not naturally represented in Euclidean space. Due to the *curse of dimensionality* it is difficult to both visualize and find a natural separation within the data for clustering purposes. We have presented a framework which may be used to solve both of these problems. By using methods from information geometry, we are able to learn the manifold from which the probability distributions governing the data lie. We have shown the ability to find a low-dimensional embedding of the manifold, which allows us to not only find the natural separation and clustering of the data, but to also reconstruct the original manifold and visualize it in a low-dimensional space.

While the methods we have presented here express the use of the Kullback-Leibler divergence as our dissimilarity measure, we want to stress that the framework is not tied to it. Many other methods of determining a 'distance' between probability distributions will easily fit into our framework. For example, when dealing with high-dimensional, sparse data sets (such as term-frequencies in document classification), the KL-divergence is not an appropriate measure, due to divide-by-zero issues. In this case, the Hellinger distance may be more representative.

In future work we plan to apply our framework to real data sets coming from unknown underlying probability distributions. This will include document classification, internet anomaly detection, as well as biological problems. We intend to show that our methods can be used for a variety of different problems as long as they can be formatted into the following setting: large sample size data sets derived from an underlying probability distribution in which the parameterization is unknown.

7. REFERENCES

[1] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to statistical learning theory," *Advanced Lectures on Machine Learning*, pp. 169–207, 2004.

[2] T. Dietterich, "Ai seminar," Carnegie Mellon, 2002.

[3] G. Lebanon, "Information geometry, the embedding principle, and document classification," in *Proceedings of the 2nd International Symposium on Information Geometry and its Applications*, 2005.

[4] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," in *Proceedings IEEE Conf. On Computer Vision and Pattern Recognition*, June 2005, pp. 581–588.

[5] S. Lee, A. Abbott, N. Clark, and P. Araman, "Active contours on statistical manifolds and texture segmentation," in *International Conference on Image Processing 2005*, 2005, vol. 3, pp. 828–831.

[6] J. Kim, *Nonparametric statistical methods for image segmentation and shape analysis*, Ph.D. thesis, Massachusetts Institute of Technology, February 2005.

[7] S. Amari and H. Nagaoka, *Differential-geometrical methods in statistics*, Springer, 1990.

[8] R. Kass and P. Vos, *Geometrical Foundations of Asymptotic Inference*, Wiley Series in Probability and Statistics. John Wiley and Sons, NY, USA, 1997.

[9] S. Amari and H. Nagaoka, *Methods of Information Geometry*, vol. 191, American Mathematical Society and Oxford University Press, 2000, Translations of mathematical monographs.

[10] 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.

[11] 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.

[12] I. Csiszár, "Information type measures of differences of probability distribution and indirect observations," *Studia Sci. Math. Hungarica 2*, pp. 299–318, 1967.

[13] A. Renyi, "On measures of information and entropy," in *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, 1961, pp. 547–561.

[14] S.I.R. Costa, S. Santos, and J. Strapasson, "Fisher information matrix and hyperbolic geometry," in *Proceedings of IEEE ITSOC Information Theory Workshop on Coding and Complexity*, August 2005.

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