SPHERICAL LAPLACIAN INFORMATION MAPS (SLIM) FOR DIMENSIONALITY REDUCTION

Kevin M. Carter[1*], *Raviv Raich*[2], *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

There have been several recently presented works on finding information-geometric embeddings using the properties of statistical manifolds. These methods have generally focused on embedding probability density functions into an open Euclidean space. In this paper we propose adding an additional constraint by embedding onto the surface of the sphere in an unsupervised manner. This additional constraint is shown to have superior performance for both manifold reconstruction and visualization when the true underlying statistical manifold is that of a low-dimensional sphere. We call the proposed method *Spherical Laplacian Information Maps* (SLIM), and we illustrate its utilization as a proof-of-concept on both real and synthetic data.

Index Terms— Information geometry, statistical manifold, dimensionality reduction

1. INTRODUCTION

Recently presented methods of manifold learning and dimensionality reduction [1, 2] focus on finding a low-dimensional representation of the data which is restricted to lie on some Riemannian submanifold of Euclidean space. These methods are designed to optimally recreate such a manifold given only a set of sample points which lie on said manifold. While each method implements this optimization differently (i.e. locally, globally, etc), all are designed to preserve some measure of the L_2 distance between sample points in a given data set.

Our recently developed algorithm, called Fisher Information Nonparametric Embedding (FINE) [3], has extended these principles towards the problem of reconstructing statistical manifolds, or manifolds of probability density functions (PDFs). With FINE, we found an embedding into an open Euclidean space in $\mathbb{R}^d$, for which the L_2-norm is an appropriate and accurate distance metric, directly related to the Fisher information distance on the original statistical manifold. This is useful when the manifold structure is unknown, as the embedding space is relatively unconstrained.

Let us now suppose, that there exists *a priori* knowledge that the statistical manifold is a portion of a hyper-sphere. An example of such a situation would involve mapping the nodes of a global network. In such situations, it may be beneficial to constrain the low-dimensional embedding to the surface of a sphere, which will enable the usage of the great-circle distance, the natural measure of a geodesic on a sphere. This additional constraint would better preserve the geometric relationships between PDFs and enable a lower-dimensional representation than would be available in a standard – open – Euclidean embedding.

In this paper we present an unsupervised method of dimensionality reduction called *Spherical Laplacian Information Maps* (SLIM), which reconstructs a statistical manifold with the constraint that all of the embedded points (PDFs) must lie on the surface of a sphere. SLIM may be interpreted as a more specific extension of FINE, and we will show it provides a more beneficial embedding for certain applications.

The remainder of the paper proceeds as follows: In Section 2 we give a brief background on measuring distances on statistical manifolds. Section 3 presents the novel algorithm of Spherical Laplacian Information Maps, while proof-of-concept simulation results are illustrated in Section 4. Conclusions and areas for future work are discussed in Section 5.

2. BACKGROUND

2.1. Fisher Information Distance

For a parametric family of probability distributions on a statistical manifold, it is possible to define a Riemannian metric using the Fisher information matrix $[\mathcal{I}(\theta)]$, which measures the amount of information a random variable contains in reference to an unknown parameter θ. The Fisher information distance between two distributions $p(x;\theta_1)$ and $p(x;\theta_2)$ is:

$$D_F(\theta_1,\theta_2) = \min_{\substack{\theta(\cdot): \\ \theta(0)=\theta_1 \\ \theta(1)=\theta_2}} \int_0^1 \sqrt{\left(\frac{d\theta}{dt}\right)^T [\mathcal{I}(\theta)] \left(\frac{d\theta}{dt}\right)}\, dt, \quad (1)$$

*Acknowledgement: This research was partially supported by the AFRL ATR Center through a Signal Innovations Group subcontract, grant number SIG FA8650-07-D-1221-TO1.

where $\theta = \theta(t)$ is the parameter path along the manifold [4, 5]. Note that the coordinate system of a statistical manifold is the same as the parameterization of the PDFs (i.e. θ). Essentially, (1) amounts to finding the length of the geodesic on $\mathcal{M}$ connecting coordinates θ_1 and θ_2.

While the Fisher information distance cannot be exactly computed without a priori knowledge about the geometry (i.e. parameterization) of the manifold, the distance between PDFs p_1 and p_2 may be approximated with the Hellinger distance,

$$D_H(p_1, p_2) = \sqrt{\int \left(\sqrt{p_1(x)} - \sqrt{p_2(x)}\right)^2 dx}, \qquad (2)$$

which converges to the Fisher information distance,

$$2D_H(p, q) \to D_F(p, q)$$

as $p_1 \to p_2$ [4]. This measure, among others, allows for the approximation of the information distance in the absence of the geometry of the statistical manifold on which the PDFs lie. For additional measures of probabilistic distance, some of which approximate the Fisher information distance, and a means of calculating them between data sets, we refer the reader to [6, 7].

3. SPHERICAL EMBEDDING CONSTRAINTS

Given points on the unit sphere parameterized with spherical coordinates $\theta = [\phi, \psi]^T$, $-\frac{\pi}{2} \leq \phi \leq \frac{\pi}{2}$ and $0 \leq \psi \leq 2\pi$, the distance between θ_i and θ_j is defined as

$$\begin{aligned} D_{S^2}(\theta_i, \theta_j) \quad = \quad & \arccos\left(\cos(\phi_i)\cos(\phi_j)\cos(\psi_i - \psi_j) + \right. \\ & \left. \sin(\phi_i)\sin(\phi_j)\right), \end{aligned} \qquad (3)$$

which is known as the great-circle distance.

We may utilize the Laplacian Eigenmaps (LEM) [1] framework towards an information-geometric embedding by modifying the choice of distance measure to that of the great-circle distance. Specifically, we can solve the optimization:

$$\Theta = \arg\min_{\{\theta_i\}} \sum_i \sum_j W_{ij} D_{S^2}(\theta_i, \theta_j), \qquad (4)$$

under similarly appropriate constraints and weightings, where $\Theta = [\theta_1, \ldots, \theta_N]$. While using spherical MDS [8] may be also be appropriate, optimizing (4) adds a sense of locality that better preserves the local neighborhood structure of the manifold.

Notice that under no additional constraints, the trivial solution to (4) is to collapse all samples to the same embedded point. To prevent this, we add a constraint designed to regulate the spread of the embedded points on the sphere. Specifically, let us solve (4) such that we maximize

$$\sum_i \sum_j D_{S^2}(\theta_i, \theta_j)^\gamma, \qquad (5)$$

where $0 < \gamma < 2$ is a power-weighting constant which regulates the spread on the sphere. One may view this constraint as maximizing the length of the graph formed when each embedded point represents a node and the length of the edge between nodes is the great-circle distance between points, raised to the power γ. By using (4) in conjunction with maximizing (5), we obtain the final objective function

$$\Theta = \arg\max_{\{\theta_i\}} \sum_i \sum_j D_{S^2}(\theta_i, \theta_j)^\gamma - W_{ij} D_{S^2}(\theta_i, \theta_j), \quad (6)$$

which ensures that close PDFs will be represented by close points in the embedding space, but the trivial solution is avoided.

One may also view our spread constraint (5) as having a relationship to controlling the entropy of the data. As detailed in [9], the data entropy may be estimated as a function of the length of the minimal spanning tree (MST)

$$\hat{L}_\gamma(\boldsymbol{X}) = \min_{T \in \mathcal{T}} \sum_{e \in T} D(e)^\gamma, \qquad (7)$$

where $\mathcal{T}$ is the set of spanning trees over $\boldsymbol{X}$, e is an edge between sample points, and $D(e)$ is the length of that edge (i.e. the distance between sample points). Larger values of $\hat{L}_\gamma(\boldsymbol{X})$ are related to larger entropy values of the data $\boldsymbol{X}$. We essentially measure the length of the maximal spanning tree (i.e. all nodes connected by an edge), which is indeed an element in $\mathcal{T}$. Hence, while our cost was designed to regulate the spread of embedded points to prevent trivial solutions, there is also a direct relationship to the entropy of the data. This result is also intuitive, as entropy is minimized with the trivial point solution, while maximized over a uniform distribution.

Unlike LEM, there is no closed form eigenvalue solution to this optimization, as the distance measure is highly nonlinear. Hence, we solve the optimization with gradient ascent methods. Letting our objective function be measured as

$$J = \sum_i \sum_j D_{S^2}(\theta_i, \theta_j)^\gamma - W_{ij} D_{S^2}(\theta_i, \theta_j),$$

we may iteratively determine the optimal embedding Θ through the process

$$\Theta_{l+1} = \Theta_l + \mu \frac{\partial}{\partial \Theta_l} J,$$

where μ is the step size and $\frac{\partial}{\partial \Theta} J$ is the direction of the gradient of the objective. The complete derivation of this gradient is available in [7].

We refer to this framework as *Spherical Laplacian Information Maps* (SLIM), as we find an information-geometric embedding of a statistical manifold, constrained to the surface of an intrinsically 2-dimensional sphere. The weights W_{ij} are calculated in a similar way to LEM, using the Fisher information distance (or approximation thereof) rather than Euclidean distance,

$$W_{ij} = \exp(-G(p_i, p_j; \mathcal{P})/t),$$

if nodes i and j are connected, with t being some constant.

Algorithm 1 Spherical Laplacian Information Maps

Input: Collection of data sets $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$; power-weighting constant γ; step size μ

for $i = 1$ to N **do**

  Calculate $\hat{p}_i(\boldsymbol{x})$, the density estimate of $\boldsymbol{X}_i$

end for

Calculate the pairwise weight matrix $[W]_{ij}$

$l = 1$

while $|J_l - J_{l-1}| > \epsilon$ **do**

  Calculate $\frac{\partial}{\partial \Theta_l} J$

  $\Theta_{l+1} = \Theta_l + \mu \frac{\partial}{\partial \Theta_l} J$

  $J = \sum_i \sum_j D_{S^2}(\theta_i, \theta_j)^\gamma - W_{ij} D_{S^2}(\theta_i, \theta_j)$

  $l = l + 1$

end while

Output: Embedding of $\mathcal{X}$, constrained to the sphere $\Theta = [\theta_1, \ldots, \theta_N]$

3.1. SLIM Algorithm

We now present the full algorithm for SLIM, which embeds PDFs onto a 2-dimensional spherical subspace. The resultant embedding is parameterized through spherical coordinates $\theta = [\phi, \psi]^T$, which maps to a 3-dimensional Euclidean subspace, constrained to lie on the surface of a sphere. The user-defined constant γ determines how large a portion of the sphere the embedding should occupy.

The full description of the SLIM algorithm is available in Algorithm 1. Empirical testing suggests that a value of $0.1 < \gamma < 1$ yields desirable results, although we would suggest users empirically determine an appropriate γ for the data of interest. We note that although we restrict our SLIM embedding to the 2-dimensional sphere in $\mathbb{R}^3$, it may be formulated for embedding onto an arbitrary d-dimensional hypersphere, although the implementation details are more difficult.

4. SIMULATIONS

4.1. Dimensionality Reduction

Let $\alpha^{(i)} = [\alpha_1^{(i)}, \ldots, \alpha_5^{(i)}]^T$ be uniformly distributed as a 5-dimensional vector satisfying the properties of a multinomial distribution: $\alpha_j^{(i)} \geq 0$ and $\sum_j \alpha_j^{(i)} = 1$. For each $\alpha^{(i)}$, we draw an i.i.d. realization $\boldsymbol{X}_i$ from a Dirichlet distribution

$$f\left(x_1, \ldots, x_4; \alpha_1^{(i)}, \ldots, \alpha_5^{(i)}\right) = \frac{1}{B(\alpha^{(i)})} \prod_{j=1}^{5} x_j^{\alpha_j^{(i)} - 1},$$

where $x_5 = 1 - \sum_j^4 x_j$ and

$$B(\alpha^{(i)}) = \frac{\prod_j \Gamma(\alpha_j^{(i)})}{\Gamma\left(\sum_j \alpha_j^{(i)}\right)}$$

Method	Classification Rate (%)	
	Mean	STD
SLIM 2-D	80.3	6.3
FINE 2-D	76.9	6.9
FINE 3-D	80.4	5.5

Table 1. Classification rates for dimensionality reduction on Dirichlet distributions parameterized by multinomials.

is the multinomial beta function, expressed in terms of the gamma function. Hence, we create a collection of data sets $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$ from a statistical manifold paramaterized by the simplex. Given that the simplex can be mapped to a portion of the sphere by the square root, this may be a good scenario for SLIM.

Let us further add a classification aspect to the problem, by defining class labels such that those data sets generated with parameters $\alpha_1^{(i)} + \alpha_2^{(i)} > 0.4$ belong to class 1, while all other sets belong to class 2. Essentially, this measures whether or not more than 40% of the probability mass was covered in the first 40% (2 out of 5) of the variates of the parameterization.

Using $N = 100$ data sets, we perform leave-one-out cross validation over 20 classification trials, i.i.d. in $\{\alpha^{(i)}\}$. We compare the k-NN classification performance of SLIM to that of FINE, embedded in both 2 and 3 dimensions, and illustrate the best performance results (optimized over k) in Table 4.1. We believe that SLIM shows superior performance to FINE in 2-D, and comparable to FINE in 3-D, due to the fact that the original PDFs could be easily parameterized by the non-negative portion of the hyper-sphere. When using SLIM for dimensionality reduction, we maintain the spherical constraint while the mapping allows for negativity, essentially yielding an additional degree of freedom. This explains the similar results to the 3-dimensional embedding with FINE.

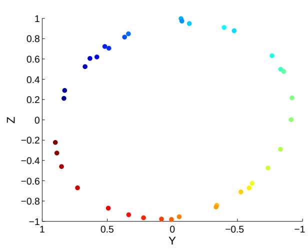
Note that we are not implying that SLIM is in general a superior algorithm to FINE. In fact the spherical constraint forces significant limitations on SLIM's usage. However, when *a priori* knowledge states that the manifold is indeed a sphere, or portion thereof, the constraint is appropriate and yields potentially significant gains for the final embedding.

4.2. Object Orientation Angle Identification

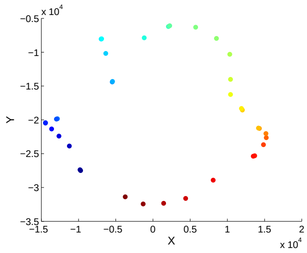
Object recognition from single images is an area that has seen much research. One of the difficulties in recognizing objects is that the orientation angle of the object in the frame must be fairly consistent in order for algorithms to properly work. Hence, the identification of this orientation angle could be of great assistance to recognition algorithms. Given that orientation is constrained to changes in pitch and yaw, one can model the manifold of angles as that of a 2-dimensional sphere.

The data for this analysis was collected at Tech-edge building, in the Air Force Research Laboratory. The experiment

Fig. 1. Sample images of rotated LCD monitor.



(a) SLIM embedding



(b) PCA embedding

Fig. 2. Embeddings of rotated images with SLIM and PCA.

was performed by positioning an LCD monitor on a swiveling desk, with a stationary camera (Canon VB-50iR) located above and to the left side of the object. The desk was then spun and the camera captured still frames of the object at 15 fps with a 640×480 resolution, for roughly 10 seconds. An illustration of the retrieved data set may be found in Fig. 1.

We proceed by sampling a portion of the image trajectory, corresponding to one complete rotation in yaw (36 images total). Each rasterized image I_i is characterized as a multinomial distribution over the entire pixel space, such that

$$p_i(I) = \left[\frac{I_i(1)}{\sum_j I_i(j)}, \ldots, \frac{I_i(m)}{\sum_j I_i(m)} \right]^T,$$

where m is the length of I_i ($m = 307200$ in this case). Given these multinomial PDFs, we calculate the pairwise Hellinger distances and implement SLIM with $\gamma = 0.75$. Results are illustrated in Fig. 2(a), where we see the clear trajectory which governs the images. Colors are applied sequentially to the points so one can view the order for which the path takes (starting at blue and ending at red).

Note that when we perform the embedding using principal component analysis (PCA) on the set of images, we see that a trajectory is still formed. While PCA discerns the order of the change in angle, it does not properly identify the shape of the trajectory (i.e. circular). This would become a crucial flaw if the camera were to change in pitch as well as yaw angle. SLIM is able to identify the constant pitch and simply define the trajectory by change in yaw.

5. CONCLUSIONS

In this paper we have presented a novel approach to dimensionality reduction, referred to as *Spherical Laplacian Information Maps*. This information-geometric method of dimensionality reduction embeds probability density functions into a common low-dimensional space, constrained to lie on the surface of the sphere. While this additional constraint restricts the usages of SLIM as compared to FINE, we have shown that if the underlying manifold is indeed that of a low-dimensional sphere, the additional constraint yields superior performance.

In future work we plan to apply SLIM towards orientation angle identification with capture changes in both pitch and yaw, and compare performance on actual recognition. We also intend to use SLIM towards mapping global networks, in which we know the sensor locations lie on an intrinsically 2-dimensional sphere.

6. REFERENCES

[1] 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. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002.

[2] 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.

[3] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero, "Fine: Fisher information non-parametric embedding," *IEEE Transactions on Pattern Analyis and Machine Intelligence*, 2009, to appear. `http:\\tbayes.eecs.umich.edu\kmcarter\papers\tpami_fine.pdf`.

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

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

[6] S.K. Zhou and R. Chellappa, "From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 917–929, June 2006.

[7] K. M. Carter, *Dimensionality Reduction on Statistical Manifolds*, Ph.D. thesis, University of Michigan, Jan. 2009.

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

[9] J. A. Costa and A. O. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *IEEE Trans. on Signal Processing*, vol. 52, no. 8, pp. 2210–2221, August 2004.