SAMPLING ON LOCALLY DEFINED PRINCIPAL MANIFOLDS

Erhan Bas, Deniz Erdogmus

Cognitive Systems Laboratory, ECE Department, Northeastern University, Boston, MA 02115 - USA

ABSTRACT

We start with a locally defined principal curve definition for a given probability density function (pdf) and define a pairwise manifold score based on local derivatives of the pdf. Proposed manifold score can be used to check if data pairs lie on the same manifold. We use this score to *i*) cluster nonlinear manifolds having irregular shapes, and *ii*) (down)sample a selected principal curve with sufficient accuracy sparsely. Our goal is to provide a heuristic-free formulation for principal graph generation and curve parametrization in order to form a basis for a principled principal manifold unwrapping method.

Index Terms— Principal graphs, resampling on manifolds

1. INTRODUCTION

In general, high dimensional data is embedded in low dimensional manifolds where the intrinsic dimension of the manifold is less than the dimension of the data space. Unsupervised manifold learning and dimension reduction techniques [1, 2, 3] aim to estimate the intrinsic embedded manifold from the original data samples that are possibly subject to noise, so that mapped noise-free data samples on the manifold represents the original samples sufficiently well. A common approach to map data to a lower dimensional space is to use linear projections such as PCA that maximizes sample variance on the projected spaces. However, linear methods are incapable of representing data structures sampled from nonlinear manifolds. Nonlinear techniques are needed to map such nonlinear structures onto local piecewise-linear subspaces [3, 4, 5] namely onto principal manifolds. Principal manifolds are smooth manifolds that passes through the middle of the data space or cloud. The concept is first introduced by Hestie and Stueltz [6]. Kegl and Krzyzak [7] extend principal manifold definition to principal graphs such that self intersections and branching structures are also included in their definition. A principal graph is a set of principal manifolds that resides in the middle of the data space. After them, [5] described the principal graph (and its approximations/variations) in terms of its deviation from the ideal configuration, using rule-based complexity measure which is a function of elements in the graph and graph grammar. In addition to these advances, [8] embraces a local strategy to define principal sets in terms of data probability distribution and its first and second order statistics. With this study, following earlier work [8], we extend the concept of principal manifold projections, and define a score that represents the similarity between sample pairs on any local manifold. Proposed score can be used to approximate the principal manifolds as piecewise linear structures and to obtain principal graphs with some given compression factor. Unlike [5], our measure is not rule-based and does not require any grammar, and can be driven from local definitions without any global optimization. Moreover, we demonstrate that the same affinity measure can be used to cluster structures having irregular shapes. In this manuscript, our goal is not to come up with efficient algorithms that will speed up the process, but to establish a framework for future studies. For that reason, we tested the proposed approach mainly on 1-dimensional manifolds (principal curves) using synthetic datasets.

This work is supported by NSF under grants ECCS0929576, ECCS0934506, IIS0934509, IIS0914808, and BCS1027724. The opinions presented here are solely those of the authors and do not necessarily reflect the opinions of the funding agency.

2. PRINCIPAL CURVES

Let $\mathbf{x} \in \mathbb{R}^n$ be a random vector, having a given pdf estimate of $p(\mathbf{x})$. Let $\mathbf{g}(\mathbf{x})$, $\mathbf{H}(\mathbf{x})$, and $\mathbf{C}(\mathbf{x}) = -\mathbf{H}_{log\ p(\mathbf{x})}(\mathbf{x}) = -p^{-1}(\mathbf{x})\mathbf{H}(\mathbf{x}) + p^{-2}\mathbf{g}(\mathbf{x})^T\mathbf{g}(\mathbf{x})$ be the transpose of its local gradient, the local Hessian and the inverse of the local covariance respectively. The local covariance is defined in this manner using the second order term in the Taylor series expansion of $log\ p(\mathbf{x})$ in order for principal curve projections to be consistent with PCA projections in the case of a Gaussian density. Let $\{(\lambda_1(\mathbf{x}), \mathbf{q}_1(\mathbf{x})), \ldots, (\lambda_n(\mathbf{x}), \mathbf{q}_n(\mathbf{x}))\}$ be the eigenvalue-eigenvector pairs of $\mathbf{C}(\mathbf{x})$, sorted in ascending order: $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. In general, a point, $\mathbf{x}$, is on the d-dimensional principal manifold iff the local gradient is a linear combination of eigenvectors of the local covariance inverse that span the tangent space, where gradient is also orthogonal to the remaining $n-d$ eigenvectors and all corresponding eigenvalues are strictly positive [8].[1] Similarly, minor curves satisfy the same criteria, except that eigenvalues are all negative and they pass through the valleys of the probability density function (pdf) and define a natural boundary between density modes; as well as between underlying clusters. For instance, let $S_\perp(\mathbf{x}) = span\{\mathbf{q}_{d+1}(\mathbf{x}), \mathbf{q}_{d+2}(\mathbf{x}), \ldots, \mathbf{q}_n(\mathbf{x})\}$ be the normal space spanned by the $n-d$ orthogonal eigenvectors and $S_\parallel(\mathbf{x}) = span\{\mathbf{q}_1(\mathbf{x}), \ldots, \mathbf{q}_d(\mathbf{x})\}$ be the parallel vectors that span the tangent space at $\mathbf{x}$. If a point is on the principal manifold, then $\mathbf{g}(\mathbf{x})$ is orthogonal to $S_\perp(\mathbf{x})$. For 1-dimensional manifolds this property implies that gradient is collinear with one of the eigenvectors (having smallest eigenvalue) of the local covariance inverse $\mathbf{C}(\mathbf{x})$. Since the mode of a probability density is also a member of principal manifolds, starting from a mode and following the eigenvectors of the local covariance inverse $\mathbf{C}(\mathbf{x})$, one can highlight all the principal curves of a given probability distribution. A similar strategy can also be used to project data samples to the principal curve axis. Since the local gradient is orthogonal to $S_\perp(\mathbf{x})$ on a principal curve, we use the following measure $\gamma(\mathbf{x})$ to terminate iterations and check if the point is on the principal curve.

$$\gamma(\mathbf{x}) = \frac{\mathbf{g}(\mathbf{x})^T C_\perp(\mathbf{x})\mathbf{g}(\mathbf{x})}{\|C(\mathbf{x})\mathbf{g}(\mathbf{x})\|\|\mathbf{g}(\mathbf{x})\|} \tag{1}$$

Here $C_\perp(\mathbf{x}) = Q\Lambda_\perp Q^T$, where $Q = [\mathbf{q}_{d+1}(\mathbf{x}), \ldots, \mathbf{q}_n(\mathbf{x})]$ and $\Lambda_\perp = diag(\lambda_{d+1}, \ldots, \lambda_n)$ are eigenvectors that span $S_\perp$ and

[1] Note that local covariance has the same eigenvalues as the $\mathbf{H}_{log\ p(\mathbf{x})}$ with inverted signs.

their eigenvalues respectively. $\gamma(\mathbf{x})$ has some nice properties: *i*) $\gamma(\mathbf{x})$ will attain a 0 value on the principal curve since $\mathbf{g}(\mathbf{x})$ is orthogonal to $S_\perp(\mathbf{x})$, *ii*) in a convex region around principal curve $\gamma(\mathbf{x})$ is positive since all the eigenvalues of $\Lambda_\perp$ are all positive and conversely, around minor curves it will attain negative values, and lastly *iii*) it is bounded between $-1 < \gamma(\mathbf{x}) < 1$ due to the normalization term. Due to space limitations, we skip the proofs here, but they can easily be derived from the eigendecomposition of the local covariance inverse.

It is crucial to mention that, it is not trivial to define a global ranking for the principal curves where data has intersections. For example, a T shaped Gaussian mixture with 2 components have the same local principal axis with opposite local rankings (the vertical axis is locally a principal curve but in one Gaussian it is the major direction while in the other it is the minor direction). [7] defines principal graphs to address this problem and proposed an algorithm that handles such cases with bifurcations or self intersections based on rules or grammar. Instead, we define ranking of principal curves with respect to local cluster means. Therefore, a principal curve that corresponds to the smallest eigenvalue form the first local principal curve (manifold). Similarly, principal subspace having the two smallest eigenvalues form a principal surface and the process can be extended to higher dimensions in this manner.

We used weighted variable-width kernel density estimate (KDE)[2] obtained from samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$. KDE is given as

$$p(\mathbf{x}) = \sum_{i=1}^{N} w(\mathbf{x}_i) G_{\Sigma_i}(\mathbf{x} - \mathbf{x}_i) \tag{2}$$

where $w(\mathbf{x}_i)$ is the weight and Σ_i is the variable kernel covariance[3] of the Gaussian kernel $G_{\Sigma_i}(\mathbf{x_i}) = C_{\Sigma_i} e^{-\frac{1}{2} x^T {\Sigma_i}^{-1} x}$ for the i^{th} data sample $\mathbf{x_i}$. The gradient and the Hessian of the KDE are:

$$\mathbf{g}(\mathbf{x}) = -\sum_{i=1}^{N} w(\mathbf{x}_i) G_{\Sigma_i}(\mathbf{x} - \mathbf{x}_i) {\Sigma_i}^{-1} (\mathbf{x} - \mathbf{x}_i) \tag{3}$$

$$\mathbf{H}(\mathbf{x}) = \sum_{i=1}^{N} w(\mathbf{x}_i) G_{\Sigma_i}(\mathbf{x} - \mathbf{x}_i) (\mathbf{u}_i \mathbf{u}_i^T - {\Sigma_i}^{-1}) \tag{4}$$

Here $\mathbf{u}_i = {\Sigma_i}^{-1}(\mathbf{x} - \mathbf{x}_i)$.

In the literature, there are various techniques to estimate the kernel size from given data [9]. In this paper, we used leave-one out cross validation to estimate the width of anisotropic kernels. Although density estimation using variable kernel size or anisotropic kernels is more robust to outliers and is capable of fitting the underlying density variations better, the estimation procedure requires k-nearest neighbor (k-nn) information where selection of k is crucial (we selected $k = \sqrt{N}$). This is computationally very expensive compared to isometric kernels. Fig. 1(a) displays nonlinearly separable 4-spiral data clusters (blue) and their principal curve projections (red) and (b) the estimated probability density using anisotropic kernels.

3. SIMILARITY ON MANIFOLDS

In general, two samples belong to the same manifold if their projections are connected via the same principal curve. In order to define connectivity on a principal curve, we first define a similarity score that has low values (ideally 0) between the data pairs on the same curve and high values (ideally ∞) between inter-curve samples. In Eqn. 1, we have already defined the measure of being a data sample on the principal curves in terms of local gradient and covariance inverse. Let $\wp(a, b)$ be the similarity score[4] of points a and b, given as the line integral of scalar valued function,$\gamma(.)$, from a to b evaluated on the curve $l(t)$

$$\wp(a, b) = \int_0^1 \gamma(l(t)) [\dot{l}^T(t) \dot{l}(t)]^{\frac{1}{2}} dt \tag{5}$$

here we parameterized $l(t) = \mathbf{a} + t(\mathbf{b} - \mathbf{a})$ as a line with $l(0) = \mathbf{a}$, $l(1) = \mathbf{b}$ and $\dot{l}(t) = (\mathbf{b} - \mathbf{a})$.

Since the local principal curve rankings might not coincide with the ones in the neighborhood manifolds, we used the ranking at location $\mathbf{a}$ as reference and select the ranking of eigenvalues at an intermediate step based on the pairwise inner product of reference eigenvectors with the ones at the intermediate step. Let $q_j^{l(t)}$ be the j^{th} eigenvector at $l(t)$ and $q_{\perp,k}^{\mathbf{a}}$ be the k^{th} eigenvector that spans $S_\perp(\mathbf{a})$ at reference $\mathbf{a}$, where $j = 1, \ldots, n$ and $k = d+1, \ldots, n$. Ranking of the k^{th}principal curve can be obtained as

$$\Theta_k = \arg \min_{i=1,\ldots,n} (q_i^{l(t)})^T q_{\perp,k}^{\mathbf{a}} \tag{6}$$

If the line integral passes through a minority curve or a region where local convexity is violated, measure defined in Eqn. 1 will attain negative values. Since we initially assumed that the data lies very close to the manifold, we only calculate the score between pairs where the connecting path also lies inside the locally defined convex region such that

$$\begin{aligned} \bar{\wp}(a, b) &= \wp(a, b) \text{ if } \forall t \in [0, 1] \lambda_{d+1,\ldots,n}(l(t)) > 0 \\ &= \infty \text{ otherwise} \end{aligned} \tag{7}$$

Note that the defined score is not symmetric. This is implied by the fact that in the presence of multiple local manifolds, the significance ranking of manifold dimensions vary as one travels through the space. Although there are two overlapping paths (actually same path defined by the local principal curve) that connects data pairs, they have different rankings. For example, in the previous T-shaped Gaussian mixture example, lets assume point $\mathbf{a}$ is at the mode of bottom mixture, whereas $\mathbf{b}$ is positioned at the upper mode. By following the first-principal axis defined at $\mathbf{a}$ (y-direction), we have access to point $\mathbf{b}$, whereas the opposite is not true since the first-principal axis of the top mixture is in x-direction.

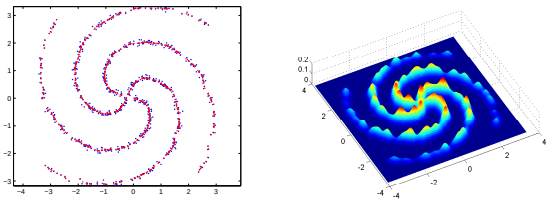

Fig. 1. (a) Spiral dataset with 4 clusters. (b) Kernel density estimate using anisotropic kernels

[2]KDE is used as an example since it encompasses parametric mixture models as a special case; the method is general for any pdf model.

[3]Assuming Gaussian kernels here for simplicity.

[4]In the rest of the paper we avoid the use of distance measure for $\wp(a, b)$, since as we will discuss it later, $\wp(a, b)$ does not have to be symmetric. However, intuitively they are similar in terms of representing dissimilarities.

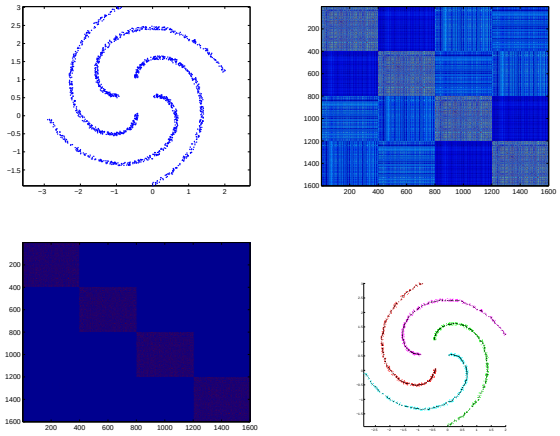

Fig. 2. (a) Mixture of 4-spirals. (b) Affinity using Euclidean distance. (c) Affinity using Manifold Score. (d) Clustering on Manifolds.

4. CLUSTERING ON MANIFOLDS

In the previous sections, we defined pairwise manifold scores, $\bar{\varphi}(a,b)$, that could be used to assess the similarity between the projected principal curve samples. We further bound the principal curve path that connects these pairs by a convex region. Using these pairwise scores, now we define the principal curve affinity matrix as $\mathbf{A}_{\wp}(\mathbf{a},\mathbf{b}) = e^{-max(\bar{\varphi}(\mathbf{a},\mathbf{b}),\bar{\varphi}(\mathbf{b},\mathbf{a}))}$. In order to cluster data samples, we simply cluster local manifolds based on $\mathbf{A}_{\wp}(\mathbf{a},\mathbf{b})$. Because extra connections between manifolds will merge clusters, we assign the maximum of the pairs as the score, $\wp(\mathbf{a},\mathbf{b})$. Fig. 2(a) displays 4-spiral data, each of which is radially perturbed with a uniform noise. In fact, for this particular example, iterative methods such as Medioidshift [10] employing ISOMAP distance [1] fails to extract the underlying cluster structure due to the presence of uniform perturbation along the first principal curve direction, whereas the proposed manifold score successfully captures the underlying pairwise affinity. For comparison, Fig.2(b-c) show obtained affinities using pairwise Euclidean distance and manifold score respectively. Fig. 2(d) displays the clustering result using the proposed manifold affinity. Since affinity is already in sparse block diagonal form,there is no need to threshold the affinity structure. Clustering result can be obtained by using connected component analysis directly, where each color represents a distinct cluster label in the figures. Similarly, Fig. 3(a) shows the result of the clustering algorithm with labels in color, and (b) shows the corresponding affinity matrix. In both figures, projected principal curve points are overlayed with cluster labels as black dots.

5. SAMPLING ON PRINCIPAL MANIFOLDS

Principal curves provide a nonlinear summary for the data and can be used as a compression tool that mimics the underlying sparse geometry sufficiently sparse. For that purpose, given a compression factor, we define the deviation from the original curve invariant of path length as $\bar{\varphi}^*(\mathbf{a},\mathbf{b}) = max(\bar{\varphi}(\mathbf{a},\mathbf{b}),\bar{\varphi}(\mathbf{b},\mathbf{a}))/L$, where $L = \int_0^1 [\dot{l}^T(t)\dot{l}(t)]^{\frac{1}{2}} dt$ is the arc length of the path. Let $\Gamma\wp(\mathbf{a},\mathbf{b}) = e^{-\bar{\varphi}(\mathbf{a},\mathbf{b})/L}$ be the affinity matrix obtained from this path normalized pairwise score matrix and $\mathbf{M}(\mathbf{a},\mathbf{b})$ be the neighborhood mask obtained from $\Gamma\wp(\mathbf{a},\mathbf{b})$ by thresholding with *thr*.

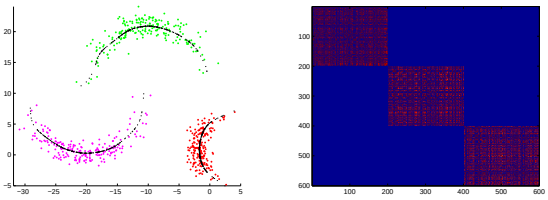

Fig. 3. (a) Mixture of 3-crescents with clustering results (in color) and principal curve projections (black). (b) Affinity matrix with 0 off-diagonals.

We propose the strategy outlined in Tab. 1 to approximate any smooth principal curve with piecewise linear lines.

Given confidence/compression threshold, *thr* and an initial reference sample $\mathbf{x}_k$, first obtain $\mathbf{M}(\mathbf{a},\mathbf{b}) = \Gamma\wp(\mathbf{a},\mathbf{b}) > thr$. Then, repeat below until termination

1. Find the set of projected samples on the principal curve,$\aleph_k = \mathbf{x}_j \in$M(k,.), $j = 1,\ldots,N$, that is accessible from $\mathbf{x}_k$
2. Find the furthest 2 samples in $\aleph_k$ having smaller $\mathbf{x}_-$ and larger $\mathbf{x}_+$ index.
3. Starting from $\mathbf{x}_-$ repeat the steps 1-2 for only - indexes until the smallest possible index is achieved.
4. Starting from $\mathbf{x}_+$ repeat the steps 1-2 for only + indexes until the largest possible index is achieved.

Table 1. Sampling on principal manifolds

Fig. 4(a) shows a spiral shaped distribution, where samples (blue) are ordered from inwards to outwards and generated uniformly along the angle which results in increasing pairwise displacement between consecutive samples with the increasing curve index. Moreover, data is perturbed with a radially increasing uniform noise. Principal curve projection of the samples are plotted in red. Fig. 4(b) displays the pairwise Euclidean distance between principal curve projections. Fig. 4(c) shows the pairwise principal affinity ($\Gamma\wp(a,b) = e^{-\bar{\varphi}(a,b)/L}$). Fig. 3(d) illustrates the masked graph that is obtained using the given compression threshold *thr* Euclidean distance graph obtained. Larger values of *thr*, will result in smaller mask area and smaller distance range for $\mathbf{x}_-$ and $\mathbf{x}_+$, implying low compression and viceversa. Lastly, Fig. 4(e) shows the algorithm result.

6. DISCUSSION

Using the locally defined principal curve definitions, we address two problems. *i*) How do you decide if two sample belongs to the same local manifold? *ii*) If you have samples from the same manifold, how do you compress (and downsample) them? In this study, we didn't deal with computational efficiency issues and tested the basic algorithm on synthetic results on 1-dimensional manifolds, namely principal curves. However, the method described here is generic and can be extended to higher dimensions with minor modifications. An interesting, yet not so-well investigated issue is the *k*-nearest neighborhood (*k*-nn) selection procedure. Almost all graph based methods

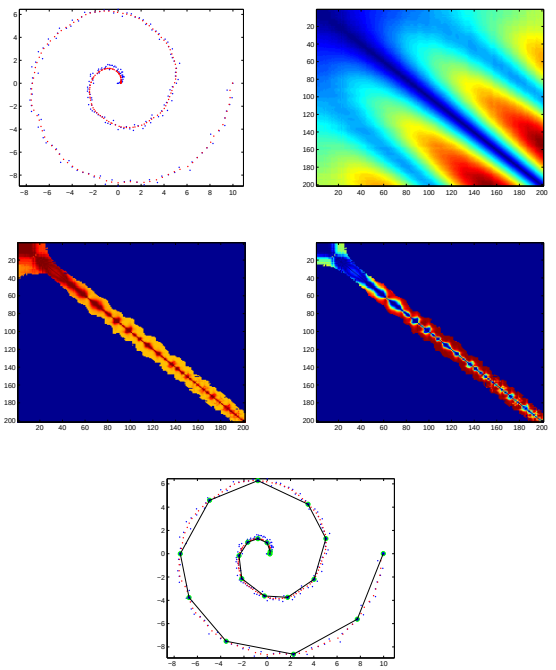

Fig. 4. (a) Single spiral. (b) Pairwise Euclidean distance. (c) Affinity using manifold score. (d) Pairwise Euclidean distance on a masked graph. (e) Resampled instances on the curve.

for dimensionality reduction and manifold learning start with some initial sparse graph, i.e. k-nn, and selection of k is arbitrary. However, although we start with a fully connected graph initially, our problem formulation yields to a varying implicit neighborhood degree around the principal curve throughout the space and further can be constrained with the threshold *thr* as seen from Fig. 3(c-d).

7. REFERENCES

[1] J. B. Tenenbaum, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, December 2000.

[2] Pierre Gaillard, Michaël Aupetit, and Gérard Govaert, "Learning topology of a labeled data set with the supervised generative gaussian graph," *Neurocomputing*, vol. 71, no. 7-9, pp. 1283–1299, 2008.

[3] Lawrence K. Saul, Sam T. Roweis, and Yoram Singer, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

[4] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," *Arxiv preprint cs/0212008*, 2002.

[5] A. Gorban and A. Zinovyev, "Elastic principal graphs and manifolds and their practical applications," *Computing*, vol. 75, no. 4, pp. 359–379, 2005.

[6] T. Hastie and W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.

[7] B. Kégl and A. Krzyzak, "Piecewise linear skeletonization using principal curves," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 59–74, 2002.

[8] D. Erdogmus and U. Ozertem, "Self-consistent locally defined principal surfaces," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 2.

[9] B.W. Silverman, *Density estimation for statistics and data analysis*, Chapman & Hall/CRC, 1998.

[10] Yaser Ajmal Sheikh, E.Khan, and Takeo Kanade, "Mode-seeking by medoidshifts," in *Eleventh IEEE International Conference on Computer Vision (ICCV 2007)*, October 2007, number 1.