Global Principal Geodesic Analysis for Latent Variable Identification

Tyrus H. Berry
Department of Mathematics
George Mason University
Fairfax, VA 22030
tyrus.berry@gmail.com

Abstract

Global Principal Geodesic Analysis (GPGA) is a new technique for latent variable analysis which is effective and computationally efficient. GPGA builds on the ideas of Principal Geodesic Analysis (PGA) and applies them in a more general setting. The motivation behind GPGA is a general model for latent variables which formally differentiates latent variable analysis from dimensionality reduction. This model shows that latent variable analysis cannot hope to preserve the topology, or even the geometry, of the observed variables. GPGA avoids these constraints, which allows for more independence in the latent variables.

1 Introduction

High-dimensional data poses a significant challenge to conventional statistics, known as the curse of dimensionality. Dimensionality reduction techniques attempt to address the problem by assuming that the data actually lie on a low-dimensional manifold. If this assumption is valid there are many successful techniques for reducing the dimensionality. The conventional wisdom is that if the data is "truly" high dimensional then the problem may simply be intractable. However, if the observed variables can be transformed into independent variables, this would also open the door to conventional statistics. Searching for independent latent variables is referred to as latent variable analysis.

Latent variable analysis has a different set of goals and priorities than dimensionality reduction. In dimensionality reduction the goal is to go from the high-dimensional observations to a low-dimensional embedding which has the same topology and local geometry. However, as we will see, we cannot hope to preserve the topology and geometry if we want to find the most independent latent variables. Moreover, we can make these changes and still maintain the probabilistic connection between the latent and observed spaces. Therefore, we consider latent variable analysis to be a tool for learning the distribution of the data, not the topological or geometric properties.

This paper introduces a general latent variable model and a new technique for latent variable analysis called *Global Principal Geodesic Analysis* (GPGA). GPGA extends the method of Principal Geodesic Analysis (PGA) to a much wider class of models. This new technique has an efficient algorithm and does not require a constrained optimization. GPGA is shown to be very effective in decomposing key examples into their latent variables.

GPGA fills a crucial gap in the current spectrum of nonlinear data analysis tools. However, the implementation uses some aspects of methods as diverse as Isomap, Locally Linear Embeddings (LLE), and even Self-Organizing Maps (SOM). Dimensionality reduction techniques such as Isomap and LLE have useful techniques for learning manifold structure, however, these techniques are constrained by topology and geometry preservation. On the other hand, the latent space of SOM allows topology and geometry to change, but goes too far by throwing out all statistical information. The key to alleviating these compromises is a unique model for latent variables which combines a latent space representation of a manifold with a probability distribution on the latent space.

2 Model for Latent Variables

2.1 Goals and their Consequences

The goal of the latent variable model is to study the statistics of the observed data, with particular focus on the dependence relationships. This goal will immediately differentiate latent variable analysis from dimensionality reduction. While dimensionality reduction techniques focus on preserving the topology or geometry of the data, the goal of latent variables is to introduce as much independence as possible into the new representation. We will see that topology and geometry preservation constrain the amount of independence in the new representation. This change in priorities is a good choice for applications such as data mining, where the main goal is to find the probability of future data lying in various subsets of the observation space. To answer these types of questions, the topology and geometry of the observed data is not relevant.

To learn the distribution of the observed data we learn a distribution f on the latent space X and a mapping $F : X \to M \subset \mathbb{R}^d$ into the observation space simultaneously. This allows us to build the latent space, distribution, and mapping in a way that introduces as much independence as possible into the latent variables, thereby avoiding the curse-of-dimensionality. Of course, the three elements of the model, (X, f, F), are constrained such that the distribution on the latent space must be the pull back of the distribution on the observed space via the mapping. By understanding this model we can easily see how various dimensionality reduction techniques restrict the amount of independence that can be introduced into the latent variables.

If we choose a latent space and a mapping, this will uniquely define the distribution on the latent space. Moreover, forcing the map to preserve topology or geometry will constrains the distributions available. To see that this restricts the amount of independence in the latent variables consider an observation space which is an annulus with uniform radial and angular distributions as in Figure 1. Topology preserving dimensionality reduction techniques cannot break the annulus. For example, graph distance based techniques found in [11] such as such as Isomap, Geodesic Nonlinear Mapping (GNLM), Curvilinear Distance Analysis (CDA), Semi-definite Embedding (SDE), LLE, Laplacian Eigenmaps (LE) and Isotop all preserve the topology of the annulus. The hole in the annulus shows up in the probability distribution as a dependence between the variables. However, this dependence can be eliminated by choosing our latent variables to be angle and radius. These latent variables are independent, but this latent space has a different topology than the annulus and therefore topology preserving techniques cannot find this representation. This shows that topology preservation restricts the amount of independence in the latent variables, and thus we will not restrict ourselves to topology preserving maps.

Two dimensionality reduction techniques which do break the topology of the annulus are SOM and GTM, but these still fail to identify the natural latent variables of angle and radius [11]. While this is partially due to the nature of the neural network algorithms, there is a deeper constraint inherent to SOM and GTM. These methods find a mapping to a uniform lattice on the latent space, and thus require the resultant distribution to be uniform. While this certainly makes the latent variables perfectly independent, it does so at too great a cost. By forcing the distribution to be uniform, the mapping to the observation space can no longer be assumed to be continuous. Thus, while the latent distribution is easy to analyze, these results cannot be extrapolated back into the observation space. At first this may seem contradictory, as we have emphasized that we want to break the topology but also want the mapping to be continuous. It turns out that we can allow for the mapping to be discontinuous on a set of measure zero, and this is enough to break the topology as required; for further information see the next subsection on the formal mathematical model.

So far we have made the case that we cannot restrict the mapping F to be topology preserving and that we cannot restrict the distribution f to be uniform. This brings us to the geometry preservation of the mapping F. This is another staple of dimensionality reduction techniques and again we claim that this is an overly restrictive property to demand of the mapping. If we consider the simple case of an ellipse in the plane, the x- and y-coordinates will not be independent. However, if we allow the major and minor axis to be rescaled separately, we can transform the ellipse into a circle which

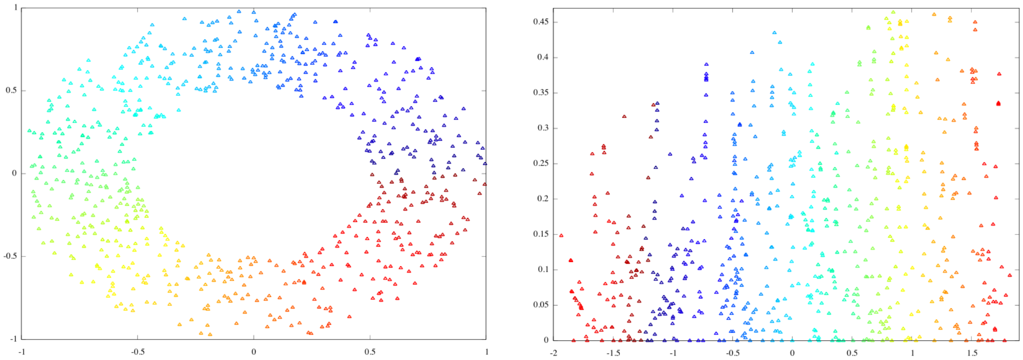

Figure 1: On the left we have an annulus with uniform radial and angular distributions. The dependence between the x-coordinates and y-coordinates cannot be removed without breaking the topology. On the right we see the new coordinates produced by the GPGA algorithm. The color shows that the new principal component (x-direction) correlates with the angle along the annulus. We see that the topology of the annulus has been broken and the new coordinates are independent.

has no linear correlation, thus making the rescaled variables more independent. This rescaling by different amounts in different directions will distort the geometry of the data. For example, if you rescale the height and width of a triangle by different amounts you can change the angles.

The GPGA technique for latent variable analysis builds a data driven latent space which is topologically trivial by analyzing the geodesic graph distances. However, GPGA does not have any inherent restrictions on the distribution of the data on the constructed latent space. While the current version of GPGA does not take advantage of the rescaling, it is described in section 3.3 how this will be accomplished. We believe that this will further improve the results shown in Section 4. The best technique to analyze the statistics on the latent space will vary depending on what aspect of the statistics is being studied. The goal of GPGA is to produce as much independence as possible in the latent space so that whatever statistical technique is chosen will have the greatest chance of success.

These examples show that latent variable analysis cannot be restricted to preserve the topology or geometry of the observed data. Instead it should find a representation where the latent variables are as independent as possible, to alleviate the curse of dimensionality, while still maintaining a continuous mapping from the latent space to the observation space. With these ideas in mind we introduce the formal mathematical model in the next subsection.

2.2 Mathematical Model

Our general model for the latent variable assumption is a generalization of the latent space typified by the models of [1, 2] combined with a manifold model such as that found in [7]. As in [1, 2] we have a d-dimensional observation space $y = (y_1, ..., y_d) \in M$ and an l-dimensional latent space $x = (x_1, ..., x_l) \in X$ where $l \leq d$. We will assume that the latent space X is a compact contractible subspace of $\mathbb{R}^l$ and the observed variables result from a function $F : X \to M \subset \mathbb{R}^d$ which is one-to-one and continuous with continuous inverse except on a set of measure zero. The observed variables lie on an l dimensional manifold $M = F(X)$.

Since latent variable analysis intends to reconstruct the latent space X from observing M it is obvious that we cannot hope to preserve the topology of M. Instead we are interested in the statistics on M, and we want to learn about this by studying the statistics on the simpler latent space X. The assumption that F is one-to-one and continuous except on a set of measure zero will allow us to lift distributions on the latent space to the manifold M. While this may seem like a strong assumption, when M is a compact orientalbe Riemannian Manifold such an F always exists via the exponential map based at any point on the manifold [9].

The central statistical assumption is that the latent space splits into k independent components so that the dimension of the i-th component is l_i and $l = \sum_{i=1}^{k} l_i$. For convenience we introduce the partial sums $L(s) = \sum_{i=1}^{s} l_i$ and $L(0) = 1$. Then if $f : X \to \mathbb{R}$ is the probability density on the latent space it can be decomposed as a product over independent components:

$$f(x_1, ..., x_l) = \prod_{i=1}^{k} f_i(x_{L(i-1)}, ..., x_{L(i)})$$

Now, following [9], we assume that M is a compace orientable Riemannian Manifold and we will define a volume form dM induced by F. Define a distribution on M via f; if A is a Borel subset of M and ξ is a random variable on M we have:

$$P(\xi \in A) = \int_{F^{-1}(A)} f(x)dx = \int_A f(F^{-1}(y))|DF^{-1}(y)|dy = \int_A dM(y)$$

Note that if F^{-1} is not differentiable, we define the probability to be a limit via approximating functions. Thus we may define the volume form $dM(y) = f(F^{-1}(y)))|DF^{-1}(y)|dy$ so that: $P(\xi \in A) = \int_A dM(y)$. This volume form naturally combines information about the geometry of the manifold and the latent distribution. The GPGA algorithm learns this volume form by recording the principal vectors and values of the PGA decomposition at a network of points along the manifold.

Furthermore, if we assume that $f(x) = \prod f_i(x_{L(i-1)}, ..., x_{L(i)})$ is a continuous distribution on the latent space X, then there exists maps $H_i : [0, 1]^{l_i} \to X$ such that each $f_i(H_i(x_{L(i-1)}, ..., x_{L(i)}))$ has uniform marginals. Thus by replacing F with $F \circ H_i^{-1}$ we may assume that the probability distribution on the latent space has uniform marginal distributions on the latent variables. This is the key to understanding how GPGA can be extended to rescale the latent variables in a natural way, this is described in Section 3.3, although this has not been implemented yet. Note that in order to make this assumption we have implicitly changed the metric on the manifold.

3 Global Principal Geodesic Analysis

3.1 Overview

In this section we introduce a new technique for latent variable analysis called *Global Principal Geodesic Analysis* (GPGA). GPGA extends the ideas of another technique called Principal Geodesic Analysis (PGA) to a more general class of models. Principal Geodesic Analysis was introduced in [13] as a method for decomposing a Lie Group, and further refined in [14] which provides the theoretical basis for PGA using differential geometry. The theory shows that a manifold has a decomposition into a natural set of coordinates, however, in general finding this coordinate system may be intractable. The technique introduced in [13, 14] is able to approximate the PGA decomposition locally. Thus, GPGA connects the manifold with a network of geodesic curves, which are used to tie together all the local PGA decompositions.

We will see that GPGA can be used for latent variable analysis, and that the latent space is data-driven as in LLE and Laplacian Eigenmaps. However, GPGA does not fix the distribution on the latent space to be uniform like SOM and GTM. Moreover, our mapping to the latent space is not determined by a neural network or a constrained optimization. Instead GPGA implicitly exploits the natural volume form from our latent variable model.

An easily visualized application of GPGA is to an annulus in the plane, shown in Figure 1. As discussed in Section 2, latent variable analysis should find the independent variables of angle and radius as the latent space. Bialek and Chigirev in [3] dealt with the fact that the coordinates of the annulus in the plane are not an optimal representation. In the context of [3] optimality refers to an information theory metric related to minimum description length, but latent variable analysis gives a more general explanation. In the context of latent variable analysis, the planar representation is not optimal because of the removable dependence between the x-coordinates and y-coordinates caused by the hole in the data.

Bialek's method and the method of principal curves both look for a lower dimensional manifold which approximately represents the data, in this case a curve. However, these methods are not suitable for the current situation since the annulus with uniformly distributed radial and angular coordinates is truly two-dimensional. The GPGA method exploits the orientable Riemannian structure of the model to find a compatible system of local coordinates which lie along geodesic curves. These

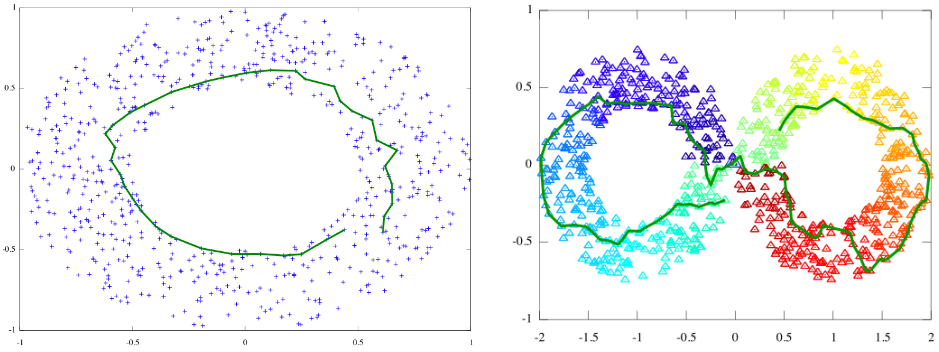

Figure 2: Above we see two examples of principal geodesics discovered by the GPGA algorithm. On the left data points are sampled from a uniform distribution on an annulus, and on the right the data is uniform on a figure-eight. The green curves are the principal geodesics discovered by the GPGA algorithm. The new coordinates of each data point are be determined by using the PGA decomposition of the nearest point on the principal geodesic. These coordinates are then adjusted according to the position along the principal geodesic.

local coordinate systems still have the same dimension as the manifold, but the final representation depends on both the local coordinates and the position relative to the geodesic curves. A more detailed explanation of the GPGA method and algorithm are provided below, but first we show the application of the method to the annulus.

Example 1: In the application to the annulus the algorithm first finds the principal geodesic, which is always following the direction of greatest variance, as shown in Figure 2. For computational efficiency the points of the principal geodesic are chosen iteratively from the actual data set. Since the angular variance of the annulus is greater than the radial variance, the principal geodesic in this example follows the angular component. Then every point on the annulus is close enough to the principal geodesic that it can be given local coordinates based at the nearest point on the principal geodesic. The final coordinates come from adding the local coordinates to the position of the nearest point along the principal geodesic. In Figure 1 we see the result, on the left is the original coordinates and on the right the GPGA coordinates.

3.2 Principal Geodesic Analysis

Principal Geodesic Analysis (PGA) is a generalization of Principal Component Analysis (PCA) to a non-linear setting. Where PCA projects data onto linear subspaces, PGA projects onto geodesic submanifolds of a manifold. PGA was successfully applied to Lie Groups in [14], but the PGA method developed there is valid locally for general compact Riemannian manifolds. The key to generalizing the PGA algorithm is connecting the local decompositions by a geodesic curves called principal geodesics. In some ways a principal geodesic is similar to a principal curve [5, 6], however they optimize different quantities.

For a detailed description of PGA see [14], here we present a condensed version. Let M be a compact orientable Riemannian n-manifold. Given a point $\mu \in M$ there is a map $\text{Log}_\mu : M \to T_\mu M$, which maps the manifold into the tangent space at μ, and a map $\text{Exp}_\mu : T_\mu M \to M$ back to the manifold. For a small neighborhood $U \subset T_\mu M$ of zero, the image $\text{Exp}_\mu(U)$ can be decomposed into a sequence of submanifolds of codimension one:

$$V_1 \subset V_2 \subset \cdots \subset V_n = \text{Exp}_\mu(U)$$

such that every geodesic of V_i is still a geodesic in V_j for all $j \geq i$. These are called geodesic submanifolds, and if we choose the sequence which maximizes the variance of each V_i in turn then the sequence is unique.

In [14], Fletcher and Thomas show that locally we can approximate this sequence by simply performing PCA in the tangent space. Thus if we let $u_1, ..., u_n \in T_\mu M$ be the principal vectors ordered by variance, then

$$V_i = \text{Exp}_\mu(\text{span}(\{u_1, ..., u_i\}) \cap V)$$

is the approximate PGA decomposition. Furthermore, the PCA vectors and the singular values give the volume form locally. The PGA decomposition is also only valid locally, so we need a way to extend the decomposition to the entire manifold. This leads us to GPGA, which connects the local decompositions by a network of geodesics.

3.3 Global Principal Geodesic Analysis

First, note that a PGA decomposition exists uniquely at each point $\mu \in M$. Since M is a Riemannian manifold, standard results (see, for example, [12]) tell us that the principal vectors will vary continuously as we move the point μ on M. Thus we can define a principal geodesic based at $\mu \in M$ as a geodesic $\gamma_\mu(t)$ passing through μ such that $\gamma'_\mu(t) \in T_{\gamma_\mu(t)}M$ is always equal to the k-th largest principal vector at $\gamma_\mu(t)$. So intuitively a principal geodesic is the geodesic which always moves in the direction of k-th largest variance, as measured in the tangent space. The existence of principal geodesics is mentioned in [13, 14]; the novel aspect of GPGA is the use of these curves to tie together local decompositions.

While both principal geodesics and principal curves are embeddings of curves into the data set, it is important to differentiate them. Principal curves are restricted by a smoothness constraint, and optimize the linear projection onto the curve [5]. Principal geodesics are defined by the variance in the direction of travel, and they optimize projection in the intrinsic metric of the manifold [13, 14]. Principal geodesics are more natural for curved spaces such as Riemannian manifolds. Furthermore, they have the added benefit of being efficient to find since they do not require a solving a constrained optimization problem like principal curves.

Now we can define the first GPGA component of a point $x \in M$ by:

$$\pi_1(x) = \text{arg min}_t\{d(x, \gamma_\mu(t))\}$$

which is the coordinate of the point on the principal geodesic which is closest to x.

Note that π_1 is only defined up to the parameterization of the principal geodesic γ_μ, so we have a choice. If we are interested in the individual distributions of the latent variables we should parameterize the principal geodesic according to its geodesic length in the observation space. This will preserve aspects such as variance so that latent variables can be compared directly. On the other hand, if we were only interested in the dependence relationships of the latent variables could now rescale as described in Section 2. While this feature has not been implemented yet, it is a promising course for further research. In this case we would choose $\gamma_\mu : [0, 1] \to M$ such that the number of points in $\pi_1^{-1}([a, b])$ is proportional to $b - a$, which would insure that the distribution of points along the principal geodesic is uniform. Thus the parameterization of the principal geodesics can determine which aspects of the distribution we are emphasizing in the latent space.

If every point is close enough to the principal geodesic that it lies in the PGA decomposition of a point on the principal geodesic then we are done. Each point will have new coordinates given by $(\pi_1(x), 0, ..., 0) + \text{PGA}(x)$, where $\text{PGA}(x)$ gives the coordinates of the PGA decomposition in the neighborhood containing x. Otherwise, if not every point is covered, we need to iteratively repeat the process by constructing more principal geodesics which branch off perpendicular to the first principal geodesic to reach the remaining points. This process may need to be iterated up to n times until all points are reached by some branch of a geodesic. In this way we construct further maps $\pi_2, ..., \pi_n$ which project onto the needed number of branches, and the final coordinates of each point are $(\pi_1(x), ..., \pi_n(x)) + PGA(x)$.

3.4 Algorithm and Complexity

Below we outline the GPGA algorithm. The main procedure is a single-source shortest path procedure on the neighbor graph of the observed data. Since the graph is sparse, a Fibonacci Heap performs Dijkstra's algorithm in $O(N \log N)$ time, where N is the number of observed points. This procedure must be performed multiple times, at most N/k for a k-nearest neighbor graph. Thus the maximum complexity is $O(N^2 \log N)$ which is the same as that of Isomap.

Note that the algorithm saves the original coordinates of each point on the subgraph of principal geodesic points. Along with the principal vectors at each point, this gives a representation of the mapping from the latent space to the observation space. Furthermore, we represent the volume form on the observed manifold with the principal vectors and eigenvalues of the PCA decomposition at each point on the principal geodesics. In most applications one needs to extrapolate statistical results on the latent space to give results on the observation space. The mathematical model shows that the volume form relates the distributions on the latent space and the observation space, so these outputs are useful.

GPGA Algorithm

```
Inputs: Collection of $N$ points $Y = \{y_i\}_{i=1}^N$ with $d$ components.
        Either $k$ for k-nearest neighbor or $\delta$ for fixed neighborhoods or neither for adaptive.
Outputs: New coordinates $X = \{x_i\}_{i=1}^N$ with $l \leq d$ components.
         Subgraph $G$ of neighbor graph on initial dataset $Y$.
         Collection of principal vectors and values, for each Vertex of $G$.
begin:
    Choose an initial point $\hat{x} = x_0$ from the observed data (many methods).
    while, not all observed points have been classified:
        Add $\hat{x}$ to the subgraph $G$ for output.
        Compute single-source shortest path procedure based at $\hat{x}$.
        Estimate Log map based at $\hat{x}$ to produce points $\hat{Y}$ in the tangent space at $\hat{x}$.
        Compute the PCA decomposition of the $\hat{Y}$ points.
        Store these vectors and values for output.
        Choose the next point, $\hat{x}$, in the direction of the largest principal vector.
    end while
```

There are several candidates for the initial point x_0, the most promising is the geodesic mean, which can be approximated efficiently as shown in [13].

4 Performance and Conclusions

Testing the independence of latent variables requires going beyond linear correlation. Some common metrics are mutual information, correlation ratio, and higher-order moments of the joint distribution. Since the mutual information of random variables is zero if and only if the random variables are independent, this metric has the most solid theoretical foundation. However, since we would like to compare various data sets we use a standardized version of mutual information sometimes called *redundancy*. Thus we can compute the redundancy between the sequence of x-coordinates and the sequence of y-coordinates for the original data, and compare this to the redundancy in the new latent variables. If the new redundancy is smaller then we conclude that the algorithm was successful in making the variables more independent.

In order to see the robustness of the GPGA algorithm we introduce a distorted annulus, see Figure 4, where the radius is now correlated with the angle. The algorithm was applied to five different versions of this distorted annulus with varying strengths of this correlation. For each strength level we applied the algorithm to five separate random samplings from the distorted annulus. The reduction in redundancy was very consistent, with an overall average decrease of 42%. However, we can see in Figure 4 that the GPGA decomposition still has a residual correlation. We expect a further reduction in redundancy when we implement the normalization as described in Section 3.3.

The GPGA algorithm is capable of decomposing even more complex topologies, as seen in 3. In this example there are two circles which need to be cut, and the cuts are effectively taking place at the ends of the principal geodesic. The gradient coloring shows how the shape is smoothly transformed into independent components. In this case the redundancy in the observed x- and y-coordinates was 0.147 and the GPGA coordinates had a redundancy of 0.011, a reduction of 92.5%, which is a significant improvement in independence.

While much has been accomplished in the area of dimensionality reduction, published techniques are not suitable for high-dimensional data with many independent components. In these cases, only latent variable analysis provides relief from the curse of dimensionality. However, to identify the most independent latent variables we must reject the dimensionality reduction concepts of topology and geometry preservation. Instead we only ask that the model allow us to transfer probabilistic

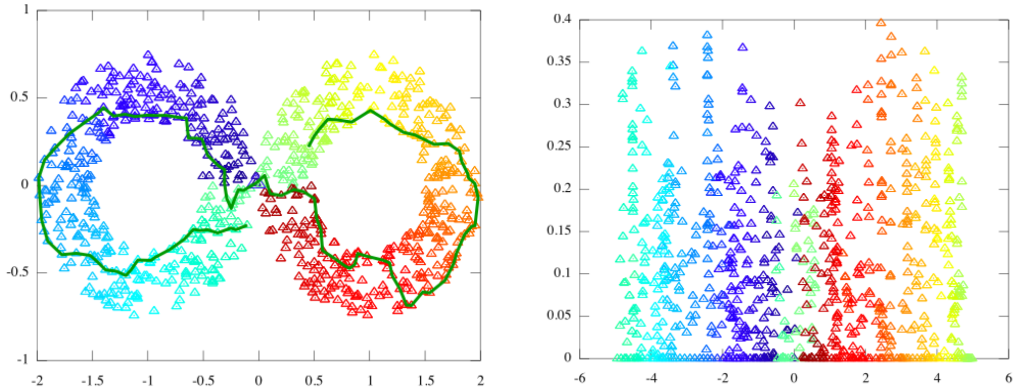

Figure 3: An example of a more complex topology, on the left are the observed variables with the principal geodesic shown as a green curve. On the right is the GPGA decomposition. In this case the redundancy was reduced from 0.147 in the observed coordinates to 0.011 in the latent variables discovered by GPGA.

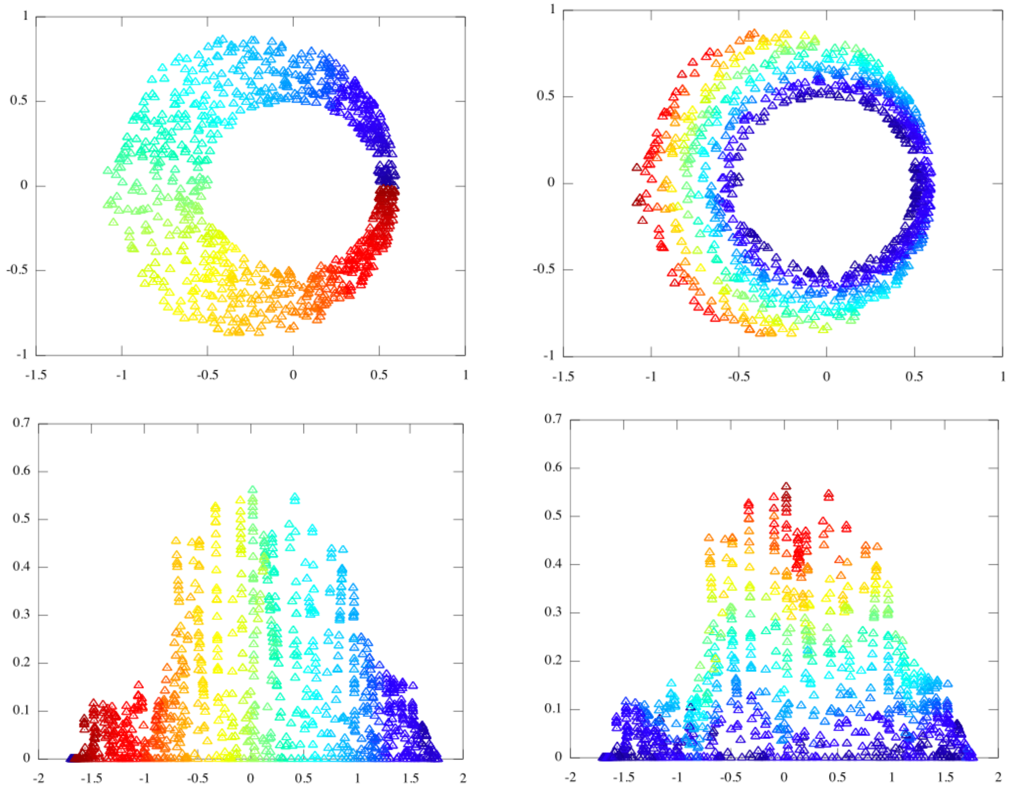

Figure 4: In the top row we have the observed data on a distorted annulus where the angles are uniformly distributed but the radius is correlated with the angle. Below we have the GPGA decomposition of the data. The left two pictures are colored by the observed angle and the right two are colored by the observed radius. Thus we can clearly see that GPGA effectively recovers the latent variables of angle and radius.

questions between the latent space and the observation space. Global Principal Geodesic Analysis (GPGA) is not constrained by the topology or geometry of the observed data, and thus is able to introduce greater independence in the latent variables than dimensionality reduction techniques. Moreover, the GPGA mapping has enough continuity to connect the latent and observed distributions. It has an efficient algorithm, comparable with the best dimensionality reduction techniques, and good performance in an objective metric. As we have seen, the proficiency of GPGA results from its ability to conveniently decouple the competing goals of dimension reduction and latent variable analysis.

References

[1] Marrs, A. D. and Webb, A. R. 1999. *Exploratory data analysis using radial basis function latent variable models*. In Advances in Neural information Processing Systems 11. MIT Press, Cambridge, MA, pp. 529-535.

[2] Bishop, C. M., Svensen, M., and Williams, C. 1996. *EM Optimization of Latent-Variable Density Models*. In Advances in Neural information Processing Systems 8. MIT Press, Cambridge, MA, pp. 465-471.

[3] Chigirev, D. V. and Bialek, W. 2004. *Optimal Manifold Representation of Data: An Information Theoretic Approach.*, In Advances in Neural information Processing Systems 16. MIT Press, Cambridge, MA, pp. 161-168.

[4] Murray, I. and Salakhutdinov, R. 2009. *Evaluating probabilities under high-dimensional latent variable models*. In Advances in Neural information Processing Systems 21, pp. 1137-1144.

[5] Hastie, T. J. *Principal Curves and Surfaces.* Ph.D. Thesis. Stanford University, 1984.

[6] Einbeck, J., Tutz, G., and Evers, L. 2005. *Local Principal Curves.* Statistics and Computing. Springer Netherlands, Volume 15, Number 4, pp. 301-313.

[7] Zha, H. and Zhang, Z. 2009. *Spectral Properties of the Alignment Matrices in Manifold Learning*. SIAM Review, Volume 51, Issue 3, pp. 545-566.

[8] Ozakin, A., and Gray, A. 2009. *Submanifold density estimation*. In Advances in Neural information Processing Systems 22, pp. 1137-1144.

[9] Pennec, X. 2004. *Probabilities and Statistics on Riemannian Manifolds*. In IEEE Workshop on Nonlinear Signal and Image Processing, pp. 194-198.

[10] Wang, J. M., Fleet, D. J., and Hertzmann, A. 2006. *Gaussian Process Dynamical Models.* In Advances in Neural information Processing Systems 18. MIT Press, Cambridge, MA, pp. 1441-1448.

[11] Lee, J. A. and Verleysen, M. Nonlinear Dimensionality Reduction. Springer. New York, NY. 2007.

[12] Carmo, M. P. Riemannian Geometry. Birkhauser. Boston, MA. 1992.

[13] Fletcher, P. T., Lu, C., and Joshi, S. 2003. *Statistics of Shape via Principal Geodesic Analysis on Lie Groups.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Volume 1, pp. 95.

[14] Fletcher, P. T., Lu, C., Pizer, S. M., and Joshi, S. 2004. *Principal Geodesic Analysis for the Study of Nonlinear Statistics of Shape.* IEEE Transactions on Medical Imaging, Volume 23, pp. 995-1005.