A Backward Generalization of PCA for Image Analysis

Sungkyu Jung, J. S. Marron, and Stephen M. Pizer

University of North Carolina at Chapel Hill
{sungkyu,marron,pizer}@email.unc.edu

Abstract. A generalized Principal Component Analysis (PCA) for various types of image-based data is proposed. We discuss two viewpoints of classical PCA, forward and backward stepwise views, pointing out that a backward approach leads to a much more natural and accessible extension of PCA for dimension reduction on non-linear manifolds. In particular, a general framework of composite Principal Nested Spheres is proposed that generalizes PCA in a backward manner and composes one or more such non-linear analyses with Euclidean data. The method works for a variety of application areas, including point distribution models, medial representations, points and normals. In examples from a lung motion study and from a population of prostates, composite PNS is shown to give a more succinct representation than alternative methods in the literature.

Keywords: Principal Component Analysis, Principal Nested Spheres, Dimension Reduction, manifold, Principal Nested Spheres, M-reps, PDM, Points and Normals

1 Introduction

Principal component analysis (PCA) is a widely used data exploration method in a variety of fields, for many purposes including dimension reduction and visualization of important data structures. In image analysis the dimensionality of objects under investigation is usually very high, so dimension reduction through PCA is essential in some analysis; see for example, [14]. In particular, a probability distribution on object shape space can be concisely described by using PCA.

Classical PCA is based on the Euclidean properties of vector space, especially inner products and orthogonality. PCA is easily applicable for many data types with these properties, an example of which is Functional PCA [15, 16], where the data set consists of smooth curves and the goal is to understand the variation in a set of curves. By a basis expansion of curves, the Euclidean properties are still well-defined, so the Functional PCA is a complete analog of classical PCA. On the other hand, many important data types in image analysis lack the Euclidean properties, so classical PCA can not be directly applied. In particular, the feature spaces of medical imaging data form high dimensional Riemannian manifolds. Some important data types are listed, as follows:

Medial shape representations Shapes of $2D$ or $3D$ objects are represented in a parametric model, called *m-reps* in short, including directions, log sizes and points as parameters. The data space here is a manifold that is a direct product of Euclidean space and unit spheres. See [17].

Scaled Point Distribution Model Representing boundaries of $2D$ or $3D$ objects by their sampled coordinates gives the surface point distribution model (PDM). A scaled PDM (SPDM) is constructed from a PDM by moving each point towards some designated center point by some fixed factor such that the sum of squared distances from the center point is unity. Thus an SPDM is a PDM that lies on a unit hypersphere, which reflects only the shape of the object.

Points and Normals Model As an extension of PDM, at each point a direction vector that is normal to the surface is attached, which gives a richer description of the object shape. The feature space of the points and normals model is a direct product of an SPDM space, a log scale factor Euclidean space and unit spheres, which is similar to the space of m-reps.

Diffeomorphisms A common methodology for comparing shapes in image analysis is to use diffeomorphisms ([7, 6]), i.e., smooth space warping functions. A shape is considered as a distortion (i.e., diffeomorphism) of some template. Thus a set of shapes is represented as a set of diffeomorphisms and the variation in the population of diffeomorphisms can be studied to understand variation in shapes. A diffeomorphism is represented by a vector field. A useful interpretation of the vector field is to decompose the vector into the direction and length, which leads to a feature space being a direct product of Euclidean space and unit spheres.

Conventional statistical analysis, including PCA, is not directly applicable to these types of data. However, there is a growing need for PCA-like methods because the dimensionality of the data space is often very high. Generalized PCA methods for manifold data can be viewed as forward or backward stepwise approaches [11]. In the traditional forward view, PCA is constructed from lower dimension to higher dimension. In the backward point of view, PCA is constructed in reverse order from higher to lower dimensions. These two approaches are equivalent in Euclidean space but lead to different methodologies for manifold data. Previous approaches for generalized PCA to manifold data are listed and discussed in Section 2. Many commonly used methods can be viewed as the forward approach. However, the backward viewpoint is seen to provide much more natural and accessible analogues of PCA than the standard view. This is discussed further in Section 2.2.

Note that there is a common characteristic of the data types listed above. Namely, all feature spaces involve orientations either as a result of normalizing alignments or explicitly e.g. surface normals. Thus the feature space becomes a direct product of Euclidean space and unit (hyper) spheres. In this article, we propose a framework for a backward generalization of PCA that works for this type of data. This includes and generalizes Principal Nested Spheres (PNS, [8]) and composite PNS [9]. Section 3 is devoted to explaining the methodology.

In particular, Section 3.1 discusses a backward PCA method for data on unit hyperspheres, and then the procedure of the proposed method by composing Euclidean and manifold data is illustrated in Section 3.2.

Advantages of the proposed method are presented by some experimental results in Section 4. We show two different data types, PDMs and m-reps. We use composite PNS to describe the motion of the lung using landmark data (PDM) extracted from CT images and to fit the Gaussian distribution in the space of prostate shapes from m-rep data. We show that composite PNS captures more variation in fewer dimensions than the standard PCA.

2 Generalized PCA for manifold data

In this section, we formulate the forward and backward stepwise views for Euclidean PCA and analyze manifold extensions of PCA in terms of those.

2.1 Forward and Backward Stepwise View of PCA

In Euclidean space, or simply a vector space of dimension d, let $X_1, \ldots, X_n$ be column vectors that are inputs for Classical (Euclidean) PCA. The data matrix is formed by aggregating the data vectors: $\mathbf{X} = [X_1, , \ldots, X_n]$. Euclidean PCA can be understood as an operation of finding affine subspaces AS^i, where $i = 1, \ldots, d$ represents the dimension of AS^i.

A traditional *forward stepwise* view to Euclidean PCA is understood by increasing the dimension i of AS^i, starting from the empirical mean $\bar{X} \equiv AS^0$. In particular, given AS^i, the direction $\boldsymbol{u}_{i+1}$ of great variance is added to AS^i, resulting in AS^{i+1}. Therefore, we have

$$AS^0 \subset AS^1 \subset AS^2 \subset \cdots \subset AS^d,$$

where each AS^i is the best fit containing AS^{i-1} in the whole space AS^d. A simple example of the forward operation in depicted in Fig. 1. In 3-space, $\bar{X}$ is plotted as a black dot with the AS^1 drawn as a line segment. AS^2 is found by adding an orthogonal direction to AS^1, resulting in an affine plane AS^2 plotted in the right panel.

The viewpoint that seems more useful for generalization of PCA to manifold data is the *backward stepwise* view. In backward PCA, principal components are found in reverse order, i.e., AS^is are fitted from the largest dimension, which leads to

$$\mathbf{R}^d = AS^d \supset AS^{d-1} \supset \cdots \supset AS^1 \supset AS^0.$$

In particular, AS^{d-1} is found from AS^d by removing the direction $\boldsymbol{u}_d$ of least variance from all of the data points. Successively, AS^i is the best fit in AS^{i+1} (not in AS^d). In the toy example in Fig 1, the backward operation can be understood by viewing the plots from right to left. From $\mathbf{R}^3$, AS^2 is fitted by removing a direction $\boldsymbol{u}_3$, the direction of least variance. Then a line (AS^1) is found within AS^2 in the same fashion, and so on.

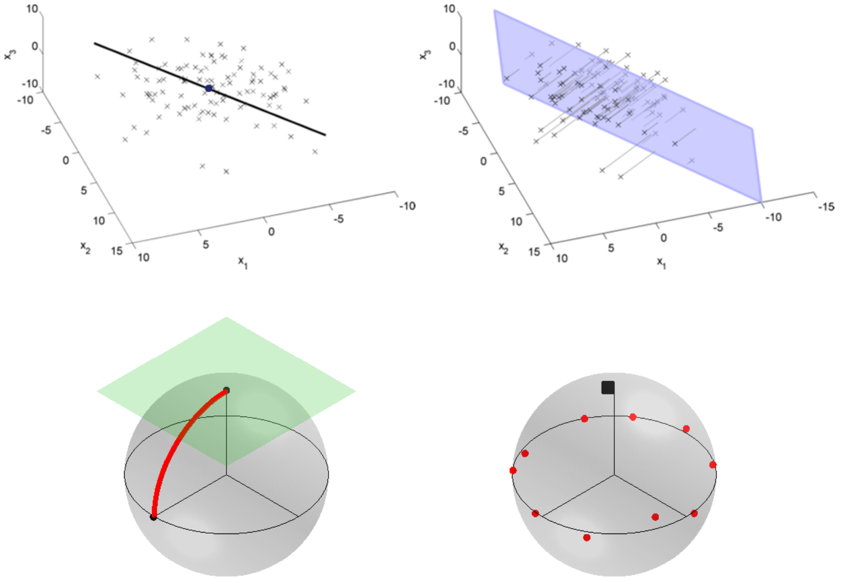

Fig. 1. (Top row) Input points in 3-space with mean and PC1 direction, and the affine subspace formed by PC1–2 directions. (Bottom left) The unit sphere S^2 with a geodesic segment (great circle segment) and the tangent plane at the north pole. (Bottom right) 10 points along the equator with random perturbation and the geodesic mean (black square) near the north pole illustrates the case where the geodesic mean on S^2 is not on the equator (S^1) and does not represent the data well.

In Euclidean space the forward and backward approaches are equivalent. In practice, the basis of AS^i is formed by the eigenvectors $\boldsymbol{u}_j$, $j = 1, \dots, i$, of the sample covariance matrix $S = \frac{1}{n-1}(\mathbf{X} - \bar{X})(\mathbf{X} - \bar{X})^T$ or the left singular vectors of the centered data matrix $(\mathbf{X} - \bar{X})$.

However, in non-Euclidean spaces the choice of viewpoint affects the generalizations of PCA, discussed next.

2.2 PCA approaches for manifold data

In curved manifolds we need to generalize important notions such as the sample mean and straight lines (or directions) as they are not defined in general manifolds. A useful notion for generalization of mean is the Fréchet mean, defined as a minimizer of sum of squared distances to data points. The Fréchet mean is widely applicable, since it only requires a metric on the manifold. In Euclidean space, the sample mean is the Fréchet mean with the usual metric $\rho(x, y) = \|x - y\|$. In curved manifolds, distances are commonly measured along geodesics. A geodesic is an analog of straight lines in Euclidean space; it is roughly defined as the shortest path between two points (see Fig. 1). The geodesic distance function measures the shortest arc length between two points. With the geodesic distance as its metric, the Fréchet mean is often called geodesic mean. A detailed discussion can be found at [13].

A widely used approach to manifold PCA, called Principal Geodesic Analysis (PGA, [3]), generalizes PCA in a forward stepwise manner. The first step in PGA is to find a center point for the manifold data. Having the geodesic mean as the center point in PGA, the second step is to find a geodesic (instead of a line) that best represents the data, among all geodesics that pass through the geodesic mean. The higher order components are again geodesics that are orthogonal to the lower order geodesics. In practice, these geodesic components are computed through a projection of the data onto the tangent space at the geodesic mean. PGA and similarly defined forward approaches are developed for various types of data; see e.g. [3] for m-reps data, [2] for DTI data, and [1] for landmark shape data.

However, there has been a concern that the geodesic mean and tangent space approximation can be very poor. As a simple example, consider the usual unit sphere S^2 and the data distributed uniformly along the equator of the sphere as illustrated in Fig. 1. In this case, the equator itself is the geodesic that best represents the data. However, the geodesic mean is located near the north or the south pole, far from any data. PGA, as a forward method, finds principal geodesics through this geodesic mean, which fail to effectively describe the variation in the data.

This observation motivated [4] to propose Geodesic PCA (GPCA). In GPCA, the geodesic mean or any pre-determined mean is no longer used; instead it finds the best approximating geodesic among all possible candidates. A center point of the data is then found in the first geodesic component, and all other components must be geodesics through the center point. In the equator example above, GPCA finds the equator as the first component. GPCA can be viewed as a backward approach, particularly when applied to S^2, since the center point is found last. In higher dimensional manifolds, for example in hyperspheres S^p with $p > 2$ and Kendall's shape spaces [1], GPCA is not fully backward, since the method is built by considering lower dimensional components first, only with an exception for center point. Nevertheless, the advantage of the method indeed comes from the backward viewpoint, i.e., from reversing the order of the first two steps.

In generalizations of PCA for higher dimensional manifolds, including hyperspheres S^p and Kendall's shape spaces, the backward stepwise principle led to a fully backward generalization of PCA: Principal Nested Spheres (PNS, [8]). In taking the backward approach, it inherits the advantages of GPCA. Moreover, this allows the successive submanifolds to be non-geodesic. PNS has been shown to provide more representative description of the data (compared to other forward stepwise approaches) in a number of standard examples in [8]. In the composite PNS we propose, PNS is used as an important building block.

3 Composite Principal Nested Spheres

In this section, a method for direct product manifolds that possesses the advantage of backward generalization of PCA is discussed. In particular, the feature space is decomposed into directional parts in S^2, shape parts in S^p, and size parts in $\mathbf{R}^+$. Then the spherical parts of the manifold are analyzed by Principal Nested Spheres (PNS), and a composite space of Euclidean parts and the result of PNS is formed to take the correlation structure into account. We summarize PNS in more detail and discuss the procedure of composite PNS.

3.1 Principal Nested Spheres

PNS generalizes PCA in a non-geodesic way for hyperspheres and Kendall's shape space, which was possible by taking the backward viewpoint. The first step in PNS is to reduce the dimension d of S^d to $d-1$. Specifically, we wish to find the best approximating sub-manifold of dimension $d-1$. PNS solves this problem with a flexible class of sub-manifolds in the form of nested spheres.

A k-dimensional nested sphere A_k of S^d is nested within (i.e., sub-manifold of) higher dimensional nested spheres; and A_k itself can be thought of as a k-dimensional sphere. A_k need not be a great sphere. As an example, A_{d-1} of S^d is defined with an axis $v_1 \in S^d$ and distance $r_1 \in (0, \pi/2]$ as follows,

$$A_{d-1}(v_1, r_1) = \{x \in S^d : \rho_d(v_1, x) = r_1\},$$

where ρ_d is the geodesic distance function defined on S^d. The parameter v_1 drives the 'direction' that is not contained in A_{d-1}. In relation to the backward view of Euclidean PCA in Section 2.1, the direction coincides to $\boldsymbol{u}_d$, which is orthogonal to AS^{d-1}. The distance from v_1 to any point in A_{d-1} is r_1, which is responsible for the curvature of A_{d-1}. This flexibility of curvature in A_{d-1} allows PNS to capture a certain form of non-geodesic variation.

Lower dimensional nested spheres are defined similarly. Since A_{d-1} is essentially a sphere, A_{d-2} can be defined again with a pair (v_2, r_2) and in a way that A_{d-2} is nested within A_{d-1}. Iteratively, one can continue to build a sequence of nested spheres $S^d \supset A_{d-1} \supset \cdots \supset A_1$. Fig. 2 shows a geometric structure of nested spheres that are recursively defined and fitted.

In PNS with a data set $X_1, \ldots, X_n \in S^d$, the pair (v, r) of nested spheres is fitted to the data iteratively so that the fitted nested spheres represent the data. [8] proposed minimizing the sum of squared distances to the data, that is, the $d-1$ dimensional PNS is

$$\hat{A}_{d-1} = \operatorname{argmin} \sum_{i=1}^{n} \rho_d(A_{d-1}, X_i)^2, \tag{1}$$

where $\rho_d(A_{d-1}, X_i)$ is defined as follows. Each X_i can be projected on A_{d-1} along the minimal geodesic that joins X_i to A_{d-1}. Denote X_i^P for the projection. The length of the minimal geodesic is the distance, that is $\rho_d(A_{d-1}, X_i) = \rho_d(X_i^P, X_i)$. Note that each observation gets a signed residual $z_{d,i}$.

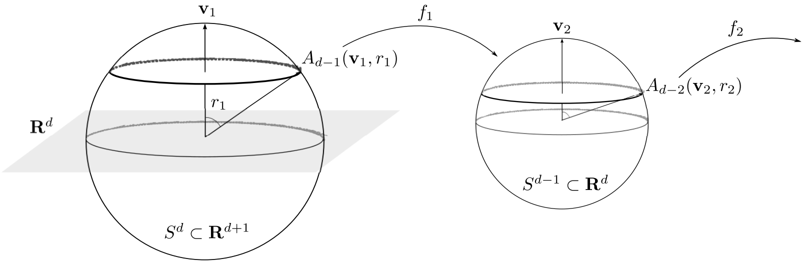
$\mathbf{v}_1$
f_1
$\mathbf{v}_2$
f_2
$A_{d-1}(\mathbf{v}_1, r_1)$
r_1
$\mathbf{R}^d$
$A_{d-2}(\mathbf{v}_2, r_2)$
$S^{d-1} \subset \mathbf{R}^d$
$S^d \subset \mathbf{R}^{d+1}$

Fig. 2. The nested sphere $A_{d-1}(v, r_1)$ in S^d and its relation to S^{d-1}, through some isomorphism f_1. Recursively, A_{d-2} is found in S^{d-1}.

The second (or the $d-2$ dimensional) PNS is found with the projections X_i^P. Since X_i^P's are on $\hat{A}_{d-1}$, one can use the method (1) by treating $\hat{A}_{d-1}$ and $\{X_i^P\}$ as S^{d-1} and $\{X_i\}$, respectively. Simply put, $\hat{A}_{d-2}$ is fitted to X_i^P's by minimizing the sum of squared distances. In general, we recursively find the sequence of PNS from the (iteratively) projected data.

The lowest level principal nested sphere $\hat{A}_1$ is then a small circle, with intrinsic dimension 1. The Fréchet mean of $X_1^P, \ldots, X_n^P \in \hat{A}_1$ is used as the best 0-dimensional representation of the data in the framework of PNS. Denote the Fréchet mean as $\hat{A}_0$, and keep the signed deviations $z_{1,i}$ of X_i^P for later use.

As a result, PNS constructs the sequence of the best approximating submanifolds

$$S^d \supset \hat{A}_{d-1} \supset \cdots \supset \hat{A}_1 \supset \{\hat{A}_0\},$$

for every dimension. The backward principle is essential to PNS, since the forward stepwise generalizations of PCA are not be equivalent to PNS (see Section 2.2) and are even not clearly defined for non-geodesic variation.

Furthermore, we wish to represent the data in an Euclidean space for further analysis, especially for composite PNS, discussed later in Section 3.2. Recall that in the procedure above, we have collected the signed residuals $z_{k,i}$. The *Principal Scores matrix* of the data by PNS is obtained by combining those residuals into a $d \times n$ data matrix

$$\mathcal{Z} = \begin{pmatrix} z_{1,1} & \cdots & z_{1,n} \\ \vdots & \ddots & \vdots \\ z_{d,1} & \cdots & z_{d,n} \end{pmatrix}, \tag{2}$$

where each column is the corresponding sample's coordinates in terms of the PNS. Each entry in row k is a perfect analogue to the kth principal component score in a Euclidean space.

When using the computational algorithm proposed in [8] for the optimization task (1), the procedure is computationally fast when the dimension and number of samples are moderate. However, in the high dimension low sample size situation where for example $d > 1000$ and $n < 100$, strict application of the iterative procedure results in a very slow computation. [8] has shown that the intrinsic dimensionality of the data can be reduced to $n-1$ without losing any information and that the first $d-n$ PNS can be found trivially by an application of singular value decomposition. This fact is used when it applies, including the experiments in Section 4.

3.2 Composite of Euclidean and Manifold data

We now introduce the general framework of composite PNS. Suppose the feature space is a direct product manifold, e.g.,

$$\mathcal{M} = \mathbf{R}^{d_1} \times (S^p)^{d_2} \times (S^q)^{d_3}, \tag{3}$$

for some $d_1, d_2, d_3 \geq 1$, $p, q \geq 2$.

The various data types introduced in Section 1 fall into this structure. For example, the feature space of the Points and Normals model with k landmark points is $\mathbf{R}^{3k} \times (S^2)^k$ (PDMs × Normals). The feature space of the PDMs with k points is $S^{3k-1} \times \mathbf{R}^+$, where the spherical part represents the shape of the lung and $\mathbf{R}^+$ represents the size of the lung. The size variable is log-transformed as done in [9]. Therefore, the feature space of the the Points and Normals model becomes $S^{3k-1} \times \mathbf{R} \times (S^2)^k$. Similarly the feature space of the m-rep model with k medial atoms is $(\mathbf{R}^3 \times \mathbf{R}^+ \times S^2 \times S^2)^k = S^{3k-1} \times \mathbf{R}^{k+1} \times (S^2)^{2k}$, by a normalizing operation and log transformation.

Let $X(i) \in \mathcal{M}$ be the ith observation, $i = 1, \ldots, n$. According to (3), we decompose each $X(i)$ into a tuple of components

$$X(i) = (x(i), y_1(i), \ldots, y_{d_2}(i), z_1(i), \ldots, z_{d_3}(i)),$$

where $x(i) \in \mathbf{R}^{d_1}$, $y_j(i) \in S^p$, and $z_j(i) \in S^q$. For the jth component in S^p, collect $y_j(1), \ldots, y_j(n) \in S^p$ that are the inputs of PNS and thus we get a principal scores matrix $\mathcal{Y}_j$ of size $p \times n$ (Eq. (2)). Likewise, For each of the jth component in S^q, PNS gives a principal scores matrix $\mathcal{Z}_j$ of size $q \times n$. Euclidean components also form a $d_1 \times n$ matrix $\mathcal{X} = [\tilde{x}(1), \ldots, \tilde{x}(n)]$, where $\tilde{x}(i)$ is a centered version, i.e., $\tilde{x}(i) = x(i) - n^{-1} \sum_{i=1}^{n} x(i)$.

In order to incorporate the correlation between the Euclidean components $\mathcal{X}$ and spherical components $\mathcal{Y}_j$ and $\mathcal{Z}_j$, define a composite data matrix

$$\mathcal{Z}_c = [\mathcal{X}^T, \mathcal{Y}_1^T, \ldots, \mathcal{Y}_{d_2}^T, \mathcal{Z}_1^T, \ldots, \mathcal{Z}_{d_3}^T]^T,$$

by vertically stacking the scores matrices.

Let the spectral decomposition of the $d = d_1 + pd_2 + qd_3$ dimensional square matrix $\frac{1}{n-1}\mathcal{Z}_c\mathcal{Z}_c^T$ be $U\Lambda U^T$, where $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d]$ is the orthogonal matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_d$. Similar to the conventional PCA, the eigenvectors $\boldsymbol{u}_k$ represent the directions of important variation in the space of $\mathcal{Z}_c$ which lead to the *Principal Arcs* when converted back to the original space $\mathcal{M}$. Likewise, the eigenvalues λ_k represent the variation contained in each component. *Principal Arc scores* for each component are computed by $\boldsymbol{u}_k^T \mathcal{Z}_c$, which is the vector of the kth scores of all n samples. Note that we do not center the data set, as opposed to the conventional PCA approach via eigen-decomposition. This is because each variable in $\mathcal{Z}_c$ is already centered.

The analysis of composite PNS can be used in the same fashion as Euclidean PCA is used. Both provide a nested sequence of subspaces (or sub-manifolds) for dimension reduction, and PC scores (or PA scores) that are important for visualization of important data structure, and for further analysis such as PC regression.

The advantage of composite PNS comes from the flexible class of sub-manifolds instead of subspaces. As shown in Section 4, the proposed method gives a more effective decomposition of the space compared to Euclidean PCA and PGA.

4 Applications to Image-based Data

Advantages of composite PNS are illustrated through analyses of two data types. The experimental results show that composite PNS gives a more effective description of the PDMs and m-reps in lower dimension than Euclidean PCA and PGA.

4.1 Lung Motion Study by PDM

Respiratory motion analysis in the lung is important for understanding the motility of tumors in the lung of an individual patient for radiation therapy applications. In-correspondence [12] PDMs of the lung boundary across respiration are used to characterize the respiratory motion [10]. Analysis by ordinary PCA has been used; we analyze this data by composite PNS.

We consider two examples, each with 10 respiratory time points. The first data set is from 4D NURBS-based Cardiac-Torso (NCAT) phantom thorax CTs. The second data set is from Respiration-correlated CT of a real patient. Retrospectively sorted CT data sets were provided by a GE 4-slice scanner using a Varian device recording patients' chest position.

The difficulty of the problem is two-fold; the dimension is very high ($d = 31650$, which could be much higher depending on the number of points on the surface) while the sample size is small ($n = 10$), and the major variation is non-linear, as shown in Fig. 3 for the NCAT data sets.

The two examples yield similar results. Fig. 3 shows scatter plots of NCAT lung data by the usual PCA (in the left panel) and by composite PNS (in the right panel). The dimension of the data space is reduced to 3 to visualize the structure of major variation. The non-linear variation apparent in the PCA subspace is represented as a linear motion in the sub-manifold of composite PNS. Observe that the sum of variances contained in PC 1-2 is roughly the amount of variation in the first principal arc. The data set from the real patient gives a similar result, where the cumulative proportions of variances in the first three

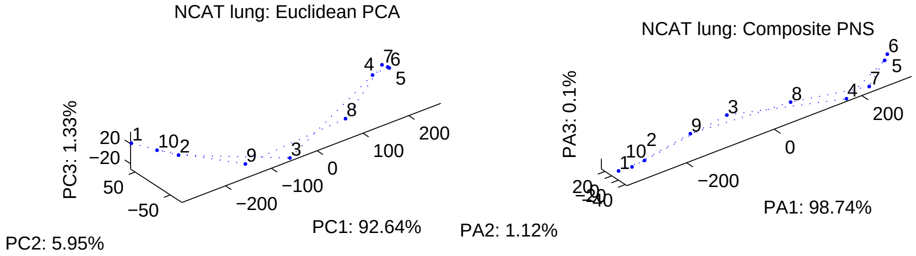

Fig. 3. (Left) Scatterplot of NCAT lung data by PC scores in the first three components of Euclidean PCA. Time points are labeled as 0-9 in the scatterplot and the proportion of variance contained in each component appears in the labels of axes. Major variation in the data is non-linear. (Right) Scatterplot of the NCAT lung data by PA scores of composite PNS. The non-linear variation is captured in the first principal arc, and thus the variation appears linear. The first component in composite PNS contains more variation (98.74% of the total variation) than 92.64% of PCA.

Table 1. Discrepancy of $1D$ approximations at each time point of the real patient lung motion. L_2 distance in real PDM scale with unit $mm/100$ is used.

time point	1	2	3	4	5	6	7	8	9	10
PCA	65.2	69.9	88.7	77.7	38.9	74.4	44.1	69.8	74.6	57.6
composite PNS	38.2	66.9	66.1	55.6	37.8	36.7	30.4	63.0	60.2	44.6

sub-manifolds (96.38%, 97.79%, and 98.63%, respectively) are higher than those of PCA (93.52%, 96.25% and 97.74%).

We also measure the discrepancy between the PDM at each time point and its $1D$ approximation by PCA or composite PNS. The discrepancy here is computed by the square root of sum of squared distances between corresponding points. In the patient lung data, the discrepancy of $1D$ approximations by composite PNS is uniformly smaller than that by PCA, as summarized in Table 1.

4.2 Shape Space of Prostate by M-reps

The m-rep models a solid human organ by few medial atoms, each of which consists of a location and two equal-length spokes to boundary. The m-rep model with k atoms lie on $\mathcal{M} = (\mathbf{R}^3 \times S^2 \times S^2 \times \mathbf{R}^+)^k$. The data set we analyze is from the generator discussed in [5]. It generates random samples of objects whose shape changes and motions are physically modeled (with some randomness) by anatomical knowledge of the bladder, prostate and rectum in the male pelvis. In the data, as in actual within-patient anatomical variation, the changes of the prostate consist of a small deformation composed with a rigid transformation. [5] has proposed and used the generator to estimate the Gaussian distribution model of shapes of human organs. Due to high dimensionality of the feature space, a dimension reduction is required.

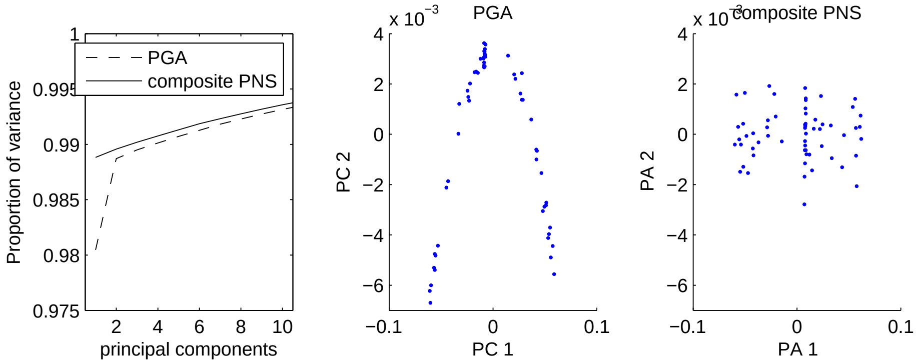

Fig. 4. Result on prostate m-reps data: (Left) Scree plot shows composite PNS captures more variation in low dimension than PGA. (Middle) Scatterplot by PG scores in the first two components. (Right) Scatterplot by composite PNS in the first two principal arcs. The curving variation left by PGA is captured in composite PNS.

The prostate m-reps consist of $k = 15$ atoms, and the dataset we analyze has 60 samples. The locations of atoms in $\mathbf{R}^{3k}$ are treated as a PDM and decomposed into SPDM (representing shapes) in S^{3k-1} and a size variable in $\mathbf{R}^+$. Performing log transformation, the feature space becomes $S^{3k-1} \times \mathbf{R}^{k+1} \times (S^2)^{2k}$. We compare the performance of dimension reduction by Principal Geodesic Analysis (PGA) as done in [5] and composite PNS.

In the left panel of Fig. 4, the proportion of the cumulative variances, as a function of number of components, shows that composite PNS captures the variation more succinctly than PGA. The scatterplots in Fig. 4 illustrate the advantage of taking backward and composite ideas. In particular, the obvious center point of the data cloud is not in the origin (geodesic mean) of PGA scatterplot (middle panel). Moreover, the curving variation in the the first two components will lead to a poor fitting of a Gaussian distiribution. On the other hand, the quadratic form of variation that requires two PGA components is captured by a single composite PNS component, and the the data spread more elliptically, which in turn leads to a better Gaussian fit.

5 Conclusion

We propose a general framework for doing backward PCA for various image data types. In particular, composite PNS works for a wide variety of applications and leads to a more succinct description of the data, as shown in the example of size and SPDM shape changes with application to the lung motion and in the example of the population of prostates. The advantage of composite PNS comes from *1)* decomposition of the feature space into spheres where the backward PCA can be applied and *2)* composition of Euclidean and manifold data to take the correlation into account.

The general framework of composite PNS we propose can be used to a variety of applications over both computer vision and medical imaging.

References

1. Dryden, I.L., Mardia, K.V.: Statistical shape analysis. Wiley Series in Probability and Statistics, John Wiley & Sons Ltd., Chichester (1998)
2. Fletcher, P.T., Joshi, S.: Riemannian geometry for the statistical analysis of diffusion tensor data. Signal Processing 87(2), 250–262 (February 2007)
3. Fletcher, P.T., Lu, C., Pizer, S.M., Joshi, S.: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans. Medical Imaging 23, 995–1005 (2004)
4. Huckemann, S., Hotz, T., Munk, A.: Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric lie group actions. Statistica Sinica 20(1), 1–58 (2010)
5. Jeong, J.Y., Stough, J.V., Marron, J.S., Pizer, S.M.: Conditional-mean initialization using neighboring objects in deformable model segmentation. In: SPIE Medical Imaging (February 2008)
6. Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased diffeomorphic atlas construction for computational anatomy. Neuroimage 23, S151–160 (2004)
7. Joshi, S.C., Miller, M.I.: Landmark matching via large deformation diffeomorphisms. IEEE Trans. Image Process. 9(8), 1357–1370 (2000)
8. Jung, S., Dryden, I.L., Marron, J.S.: Analysis of Principal Nested Spheres. Submitted in Biometrika (2010)
9. Jung, S., Liu, X., Marron, J.S., Pizer, S.M.: Generalized pca via the backward stepwise approach in image analysis. In: et al., J.A. (ed.) Brain, Body and Machine,. vol. 83, pp. 111–123. Springer (2010)
10. Liu, X., Oguz, I., Pizer, S.M., Mageras, G.S.: Shape-correlated deformation statistics for respiratory motion prediction in 4D lung. SPIE Medical Imaging 7625 (2010)
11. Marron, J.S., Jung, S., Dryden, I.L.: Speculation on the generality of the backward stepwise view of pca. In: Proceedings of MIR 2010: 11th ACM SIGMM International Conference on Multimedia Information Retrieval, Association for Computing Machinery, Inc., Danvers, MA, 227-230. (2010)
12. Oguz, I., Cates, J., Fletcher, T., Whitaker, R., Cool, D., Aylward, S., Styner, M.: Cortical correspondence using entropy-based particle systems and local features. In: Biomedical Imaging, ISBI 2008. 5th IEEE International Symposium, 1637-1640 (2008)
13. Pennec, X.: Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. J. Math. Imaging Vis. 25, 127–154 (July 2006), `http://portal.acm.org/citation.cfm?id=1166859.1166868`
14. Rajamani, K.T., Styner, M.A., Talib, H., Zheng, G., Nolte, L.P., Ballester, M.A.G.: Statistical deformable bone models for robust 3d surface extrapolation from sparse data. Medical Image Analysis 11, 99–109 (2007)
15. Ramsay, J.O., Silverman, B.W.: Applied Functional Data Analysis: Methods and Case Studies. Springer, New York (2002)
16. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer, New York, 2nd ed. edn. (2005)
17. Siddiqi, K., Pizer, S.M.: Medial Representations: Mathematics, Algorithms and Applications. Springer (2008)