BACKWARDS PRINCIPAL COMPONENT ANALYSIS AND PRINCIPAL NESTED RELATIONS

JAMES DAMON[1] AND J. S. MARRON[2]

ABSTRACT. In non-Euclidean data spaces represented by manifolds (or more generally stratified spaces), analogs of principal component analysis can be more easily developed using a backwards approach. There has been a gradual evolution in the application of this idea from using increasing geodesic subspaces of submanifolds in analogy with PCA to using a "backward sequence" of a decreasing family of subspaces. We provide a version of the backwards approach by using a "nested sequence of relations" which define the decreasing sequences of subspaces which need not be geodesic. Because these are naturally inductively added in a backward sequence, they are frequently more tractable and overcome difficulties with using geodesics.

1. INTRODUCTION

Principal component analysis is a widely applied method with many uses, including data visualization and dimension reduction (see e.g. Jolliffe [Jol05] for a good introduction and a comprehensive overview). However, an extension of this method to non-Euclidean data spaces has not been straightforward because familiar building blocks such as subspaces and orthogonality are not available. Nevertheless, the general usefulness of the method has motivated a number of different approaches to analogs of PCA in the case of data lying in a manifold. These include (see §2 for a description): principal geodesic analysis, Fletcher et al [FLPJ04], geodesic principal components, Huckemann et al [HHM10], principal arc analysis, Jung et al [JFM11], principal nested spheres, Jung et al [JDM12], and composite principal nested spheres, Pizer et al [PJG$^+$13].

In these manifold analogs of principal component analysis, the first three take a "forward approach"(using familiar terminology from stepwise regression analysis). Recall in the Euclidean case, one starts at the sample mean and finds the first eigendirection of the covariance matrix, which determines the line that best fits the data. This, together with the second eigendirection, then determines the 2-plane that best fits the data. Iteratively continuing this process results in a sequence of best fitting affine spaces, nested in order of increasing dimension.

However, principal component analysis could also be developed in the opposite "backwards direction" (still using the stepwise regression analogy). Instead, one starts with the best fitting rank r affine space, then within that finds the best fitting rank $r-1$ affine space, etc. This also results in a sequence of best fitting spaces, indexed by dimension. This approach is generally not widely considered, because

(1) Partially supported by the Simons Foundation grant 230298 and National Science Foundation grant DMS-1105470 and (2) Partially supported by National Science Foundation grant DMS-0854908.

by the Pythagorean theorem (i.e. analysis of variance type calculations) the two approaches result in the same sequence of affine spaces.

However, for non-Euclidean data spaces, these are no longer the same. Marron et al [MJD10] and Jung et al [JLMP10] observed in an empirical way that the backwards approach is more generally extendable in several cases. For example, in the case of the special class of manifolds which are spheres in $\mathbb{R}^n$, there is an alternative approach of Principal Nested Spheres introduced in [JDM12]. It finds a nested sequence of spherical submanifolds of decreasing dimension (which typically are not geodesic submanifolds). At each stage the set of data points is replaced by the set of nearest points in the next lower dimensional spherical submanifold. In turn, the next nested spherical submanifold is chosen which best approximates the new set of data points. This was extended to products of spheres and Euclidean space by [PJG$^+$13].

While these methods provided useful tools in separate contexts, an interesting question was *why* the backwards approach seemed to be so generally useful. This paper answers this question by introducing a general approach which demonstrates why backwards PCA is very natural. We begin with a collection of data points $\mathcal{S} = \{x_i\}$ on a subspace X of $\mathbb{R}^n$ where the coordinates of $\mathbb{R}^n$ have physical meaning for the data points. Now X may denote either a submanifold or more generally a Whitney stratified set (see §3). We modify the point of view from a sequence of nested submanifolds Y_i of X to a sequence of "nested relations" $f_i = c_i$. By a "relation" for a set of data points we would usually mean an equation which each of the points satisfies. Because we are considering equations of a given form, there may be no equation of this form which all of the points satisfy exactly. Thus, we seek instead an equation of the given form which the data points come closest to satisfying, as measured by a "closeness of fit criterion".

Statistical readers should consider the term *relation* a synonym for equality constraint. From traditional regression analysis, the best fitting line to a collection of data points in $\mathbb{R}^2$ can also be viewed as the equation $y = mx + b$ which the data points come closest to satisfying. This equation defines a relation between the coordinates of the data points. In the case of principal nested spheres, the spheres can be viewed as the subspaces defined by a series of linear functions which best successively describe the relations between the data points. In the general case for higher dimensions and spaces, we extend this idea to provide an inductive process for fitting the best nested sequence of relations for the data. The relations are from a (succession of) vector space(s) of functions on X, with a function chosen at the k-th stage to be a function f_k and value c_k which comes closest to defining a relation $f_k = c_k$ for the nearest points already satisfying the preceding $k-1$ relations $f_i = c_i$ for $i = 1, \ldots, k-1$.

This provides flexibility in the choice of the form of the relations and avoids problems involving the use of geodesics. Already for simple subspaces which are a product of spaces, the geodesics will not generically be closed curves, and can even be dense in the space. The situation is generally worse on stratified spaces. When geodesics on a stratum meet a lower dimensional stratum, there is not a locally well-defined continuation of the geodesic onto another stratum. Such an approach overcomes the difficulties with data points being far from the mean, the ambiguity of using geodesics which are not closed, and even having to give well-defined geodesics on stratified spaces.

2. Motivating Results from Earlier Work

We begin by placing the approach we propose in the framework of recent work on PCA for manifolds. We describe in more detail the approaches we had mentioned in the introduction.

Approaches to PCA on manifolds :

i) *principal geodesic analysis* [FLPJ04]: develops standard principal component analysis in the tangent plane at the geodesic mean, and projects this back to the manifold, resulting in geodesics passing through the geodesic mean, which explain maximal amounts of variation.

ii) *geodesic principal components* of [HHM10]: finds a sequence of best fitting geodesics. This approach added important flexibility to principal geodesic analysis by eliminating the constraint that the geodesics pass through the mean.

iii) The *principal arc analysis* of [JFM11] who find the best fitting circle to data on the conventional unit sphere in 3 dimensions. This adds flexibility to geodesic principal components, by extending the set of circles that can be fit, from great circles only, to include small circles.

iv) *principal nested spheres* of [JDM12]: extends principal arc analysis to allow fitting lower dimensional spheres in arbitrary dimension. This allows natural extension to more contexts, including Kendall's shape analysis (see e.g. [DM98] for good introduction).

v) *Composite principal nested spheres*, [PJG$^+$13]: extends principal nested spheres to products of spheres and Euclidean space, which are fundamental to the medial, and skeletal approaches to object representation in image analysis (see [SP08] for a good introduction).

In the first three methods, the forward approach is taken to PCA using geodesics in the appropriate manifolds. In the fourth and fifth approaches, backwards PCA is used and a nested sequence of subspaces decreasing in dimension is constructed. These later two provide the starting points for the general approach developed here. The nested sequence of decreasing spaces is naturally constructed starting with a high dimension, and then iteratively reducing the dimension through a series of constraints which arise as the level sets of relations $f_i(x) = c_i$. This provides a dual approach to the process where the emphasis is placed on the functions defining the relations and the level sets are defined by the relations. This is the standard approach used in algebraic geometry.

We explain the general framework in Section 3 and in Section 4 we provide the inductive procedure for constructing the sequence of "principal nested relations" which best fit the data. In Section 5 we describe the abstract form the procedure takes for several examples and give a series of concrete examples.

3. Whitney Stratified Sets and Vector Spaces of Relations

We let $X \subset \mathbb{R}^n$ denote a closed Whitney stratified set. This means that it is a disjoint union of C^∞ submanifolds $\{S_i\}$, called strata, satisfying: i) the "axiom of the frontier", which means that if $Cl(S_j)$ denotes the closure of S_j and $S_i \cap Cl(S_j) \neq \emptyset$, then $S_i \subset Cl(S_j)$; and ii) the Whitney regularity conditions, which ensure that for each pair of strata (S_i, S_j) from i), S_j satisfies Whitney regularity properties along S_i (see e.g. [Mat73] or [GLDPK76]). In the special case where there is only a

single stratum, we obtain the case of a smooth submanifold of $\mathbb{R}^n$. An interesting example of a stratified set which is not a submanifold is the one introduced to model tree structures by [BHV01].

The dimenson of X, denoted $\dim(X)$, is the maximal dimension of the strata S_i. In the situations we consider, every stratum is contained in the closure of a stratum of dimension $= \dim(X)$. We also suppose that the coordinates of points on X have physical meaning, so that any equation $f = c$ for a function f on X gives a relation between physically meaningful quantities associated to points of X.

Second, we let $\dim(X) = d$. For each $k = 1, \ldots, d$, we suppose that we are given a finite dimensional inner product vector space V_k of stratawise smooth functions on X. These vector spaces need not all be distinct. One of the functions $f_k \in V_k$ will give the k-th relation $f_k = c_k$ for an appropriate $c_k \in \mathbb{R}$. The simplest example is the case where each V_k is the vector space of linear functions with the induced dual inner product from $\mathbb{R}^n$. More generally some of the V_j could denote the homogeneous polynomials of some degree m_j with appropriate inner products. It is also possible to go beyond polynomial relations, such as eigenfunctions of operators, etc.

We would like these vector spaces to satisfy a spanning condition.

Definition 3.1. Given a closed Whitney stratified set $X \subset \mathbb{R}^n$, a vector space V of stratawise smooth functions on X satisfies the *spanning condition* if for any stratum S_j of X and any $x \in S_j$, the set of derivative maps $df(x) : T_x S_j \to \mathbb{R}$ for $f \in V$ spans the dual space $T_x^* S_j$ (consisting of the linear functions on $T_x S_j$).

The vector space of linear functions on $\mathbb{R}^n$ satisfies the spanning condition on any stratified set $X \subset \mathbb{R}^n$. In the case of principal nested spheres, these are the appropriate spaces of functions defining relations.

Also, if $\{0\}$ is an isolated stratum of X or $0 \notin X$, then the space of homogeneous polynomials of some degree $m > 1$ also satisfies the spanning condition.

One consequence of the spanning condition is the following Lemma.

Lemma 3.2. *If the space of functions V satisfies the spanning condition for the stratified space $X \subset \mathbb{R}^n$, then for almost all $(f, c) \in V \times \mathbb{R}$ (i.e. in the complement of a set of Lebesgue measure zero), the level set $f^{-1}(c) = \{x \in \mathbb{R}^n : f(x) = c\}$ is transverse to the strata of X. Hence, the intersection $X_1 = f^{-1}(c) \cap X$ is again a Whitney stratified set with strata $\{f^{-1}(c) \cap S_i\}$ for $\{S_i\}$ the strata of X. Moreover, the orthogonal subspace V_1 to f again satisfies the spanning condition on X_1.*

Proof. Define $F = (f(x), f) : \mathbb{R}^n \times V \to \mathbb{R} \times V$. Then, a straightforward argument shows that $c \in \mathbb{R}$ satisfies that $f^{-1}(c)$ is transverse to a stratum S_i of X if and only if (c, f) is not a critical value of $F|S_i \times V$. Then, by Sard's Theorem, the set of critical values of $F|S_i \times V$ has (Lebesgue) measure 0 in $\mathbb{R} \times V$. Thus, the union Z of these sets for all of the (finite number of) strata of X has measure 0. If $c \notin Z$, then $f^{-1}(c)$ is transverse to the strata of X. By a standard result about Whitney stratified sets (see e.g. [Mat73] or [GLDPK76, Chap. 1]), $f^{-1}(c) \cap X$ is again a Whitney set with strata $f^{-1}(c) \cap S_i$. Lastly, as $T_x f^{-1}(c) \cap S_i = \ker(df(x)) \cap T_x S_i$, a simple argument in linear algebra implies that the orthogonal complement to f still satisfies the spanning condition on $f^{-1}(c) \cap X$. □

This provides us the ability to inductively construct the subspaces using the relations.

Measures for "Goodness of Fit" for the Relations. Given a relation $f(x) = c$ on X, we give two ways to measure the failure of a specific point $x \in X$ to satisfy the relation using a *difference function* $\delta(x)$.

i) We let $\delta(x) = dist(x, f^{-1}(c) \cap X)$. We view X as a metric space with a metric satisfying $dist(x, x') \geq \|x - x'\|$ (the Euclidean distance between x and x'). For example, we may define $d(x, x') =$ length of the minimum path in X from x to x' if X is path-connected.

ii) We choose a *weighted penalty function* $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ which is a smooth monotonically increasing function on the half-line. Then, $\delta(x) = |\varphi(f(x)) - \varphi(c)|$.

In the second case, the weighted penalty function allows us to alter the weight assigned to the distance of $f(x)$ from c. Of these two choices i) may be more natural; however, ii) would be easier to implement.

Example 3.3. Consider $X = S^n \subset \mathbb{R}^{n+1}$ the unit sphere, with V_j the space of linear functions for all j. In case i) we obtain the geodesic distance from x to the spherical submanifold of S^n. In case ii), we use the weighted penalty functions $\varphi_j(x) = \arcsin(\frac{x}{r_j})$ for all j, where r_j is defined inductively by $r_j = r_{j-1}\sqrt{1 - (\frac{c_{j-1}}{r_{j-1}})^2}$ for $f_{j-1}(x) = c_{j-1}$ the $(j-1)$-st relation. Then, a simple calculation shows that we obtain for each j a δ_j which is a constant multiple of the geodesic distance from x to the spherical submanifold.

4. Inductive Definition of the Nested Relations

Given the stratified set $X \subset \mathbb{R}^n$ of $dim(X) = d$, the inner product vector spaces $\{V_j\}$, and difference functions δ_j, we give an inductive definition of the nested relations.

In general, consider a finite set of points $\{x_i : i = 1, \ldots, r\} \subset X$, for X a Whitney stratified set of $\dim(X) = d$, with V an inner product vector space of functions on X, and a difference function δ. We seek to minimize

$$\Phi(f, c) \;=\; \sum_{i=1}^{r} \delta(x_i)^2 \tag{4.1}$$

over $(f, c) \in S_V \times \mathbb{R}$, where S_V is the unit sphere in V about 0. For example, for a finite set of points in $X = \mathbb{R}$, with $V = \mathbb{R}^*$, then $S_V = \{\pm id\}$ and the minimizing value of c is the Fréchet mean of the points.

Remark 4.1. A potential L^1 variation of the squared distances used in (4.1) would use absolute distances instead. This is expected to result in enhanced robustness properties. For example, in the case of a finite set of points, this would result in the Fréchet median.

This minimum is achieved by the following Lemma.

Lemma 4.2. *The function Φ attains a minimum ≥ 0 on $S_V \times \mathbb{R}$.*

Proof. This is proven by finding a $C > 0$ such that if $(f', c') \in S_V \times \mathbb{R}$, with $|c'| > C$, then there is an (f, c) with $|c| \leq C$ such that $\Phi(f', c') \geq \Phi(f, c)$. Hence, $\min(\Phi(f, c))$ occurs on the compact set $S_V \times [-C, C]$.

To see this, we choose an $R > 0$ so that all $x_i \in D_R$, for the disk D_R about 0 of radius $R > 0$, and such that for some $(f, c) \in S_V \times \mathbb{R}$, $\Phi(f, c) \leq rR^2$. Then,

$F : S_V \times D_{2R} \to \mathbb{R}$ defined by $F(f, x) = f(x)$ has compact image $Y \subset [-C, C]$ for some $C > 0$. Then, for (g, c') with $|c'| > C$, if $x \in g^{-1}(c')$, then $x \notin D_{2R}$. Thus, if $\delta(x_i)$ is achieved as the distance in X from x_i to x, then $\delta(x_i) \geq \|x_i - x\| \geq R$. Thus, $\Phi(g, c') \geq rR^2 \geq \Phi(f, c)$ as claimed. □

Generic Properties needed for PCA.

In order to carry out PCA, whether forwards or backwards, the data set must be assumed to be generic in an appropriate sense. This already occurs even in the simplest case of usual PCA on Euclidean space, where the data must have a covariance matrix whose eigenvalues are distinct. For the case of data on manifolds, the properties of the manifolds and the data relative to subspaces becomes important. In the forward cases these involve the geodesic subspaces, and constraining the data so the method can be applied. For backward PCA, there is the requirement that data points have unique minimum geodesics to the subspaces. For the case of nested relations we identify the following properties that generic data sets should possess. These are described in terms of the *cut locus* for a subspace $Y \subset X$ (the definition may vary but we use the following).

$$C_X(Y) = \{x \in X : dist(x, Y) \text{ is achieved at more than one point of } Y \text{ or at a degenerate minimum point of } Y\}$$

Generic properties for data sets:

a) The set of $(x_1, \ldots, x_r) \in X^r$ such that Φ does not have a unique minimum is a set of rd-dimensional measure zero.
b) For almost all $(f, c) \in S_V \times \mathbb{R}$, $C_X(f^{-1}c))$ has d–dimensional measure 0.

It would then follow that for a generic set of data which lies outside $C_X(f^{-1}(c))$ for the minimum (f, c), there will be a unique nearest point to a data point on $f^{-1}(c)$.

Remark 4.3. These properties can be established for certain symmetric spaces such as spheres and Lie groups, and they are partially established for manifolds with generic metrics by Buchner [Buc78]. In general considerable work is needed to establish the general validity of these properties.

Example 4.4. As a simple example, if $V = \mathbb{R}^{n\,*}$, then the mapping $x \mapsto f(x)$ defines a linear mapping $\mathbb{R}^n \to V^*$. Hence, if the number of data points $r < \ell = \dim V$, then generically the image has dimension r in V^*. The appropriate c will correspond to an element of this subspace. Also, the image vanishes on a subspace $V' \subset V$ of dimension $\ell - r$. Any $f \in V'$ will vanish on all of the data points and c so f could be modified by adding any element of V'. Thus, we would not obtain a unique (f, c). Hence we would need at least n data points to obtain a unique (f, c).

If a generic set of data $\{x_i\}$ satisfies both a) and b) for the unique minimum (f, c), then there is a minimum of the distance from x_i to $X' = f^{-1}(c) \cap X$. We consider the unique $x_i' \in X'$ such that $d(x_i, x_i')$ is minimum. Suppose moreover that for the generic relation $f(x) = c$, $f^{-1}(c)$ is transverse to X. Then, X' is a Whitney stratified set of one lower dimension, and we may replace the data set $\{x_i\}$ by the corresponding closest set of points $\{x_i'\}$ which lie in X'. These now exactly satisfy the relation $f(x) = c$. We may then proceed inductively.

We now give the inductive procedure for constructing the sequence of nested relations for a generic set of data points $\{x_i\}$ based on a sequence of vector spaces V_j of functions and difference functions δ_j.

Inductive Construction of Nested Relations:

This procedure will yield in the generic case: i) a sequence of vector subspaces $W_j \subset V_j$, ii) a sequence of relations $(f_j, c_j) \in S_{W_j} \times \mathbb{R}$, iii) a nested sequence of closed Whitney stratified sets $X = X_0 \supset X_1 \supset \cdots \supset X_d$, with $\dim(X_j) = d - j$ where $X_j = f_{j-1}^{-1}(c_{j-1}) \cap X_{j-1}$, and iv) a sequence of sets of data points $\{x_i^{(j)}\} \subset X_j$.

Step 1: Let $W_1 = V_1$ and minimize Φ for a unique minimum relation $(f_1, c_1) \in S_{W_1} \times \mathbb{R}$. Generically $f_1^{-1}(c_1)$ is transverse to X so we let $X_1 = f_1^{-1}(c_1) \cap X$.

Step 2: Generically each x_i has a unique closest point $x_i^{(1)} \in X_1$.

Step 3: Repeat steps 1 and 2 for the closed Whitney stratified set X_1, and the set of data points $\{x_i^{(1)}\}$ in X_1, except if $V_2 = V_1$. then we use instead of V_2, $W_2 = \langle f_1 \rangle^{\perp}$. By the independence condition and the transversality of $f^{-1}(c)$ to X, V_2' satisfies the independence condition. We obtain a relation $(f_2, c_2) \in S_{W_2} \times \mathbb{R}$ which minimizes Φ for closed Whitney stratified set X_1, and the set of data points $\{x_i^{(1)}\}$ in X_1. Also, $f_2^{-1}(c_2)$ is transverse to X_1 and we let $X_2 = f_2^{-1}(c_2) \cap X_1$. Again generically the $\{x_i^{(1)}\}$ has unique closest points $\{x_i^{(2)}\}$ in X_2.

Inductive step

j+1: Given X_j and data points $\{x_i^{(j)}\}$ in X_j, with $j < d$, we repeat the argument in Step 2 where now if $V_j = V_{j_i}$ with $j_i < j$ for $i = 1, \ldots, m$, then we let $W_j = \langle f_{j_1}, \ldots, f_{j_m} \rangle^{\perp}$ in V_j.

Conclusion : After d steps, we obtain a nested sequence of subspaces W_j, relations $f_j(x) = c_j \in S_{W_j} \times \mathbb{R}$, $j = 1, \ldots, d$, nested sequence of closed Whitney stratified sets $\{X_j\}$ of $\dim(X_j) = d - j$ and corresponding data points $\{x_i^{(j)}\} \subset X_j$.

Note that $X_d = X \cap_{j=1}^{d} f_j^{-1}(c_j)$ is a finite set of points. This set is what is called the *backwards mean.*

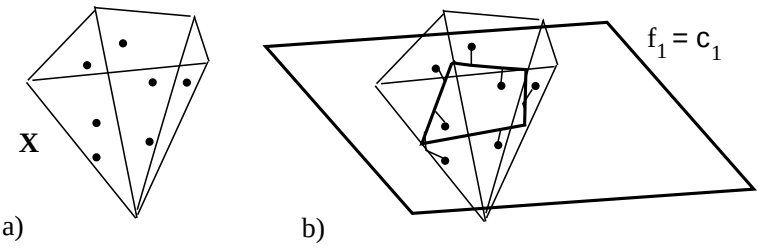

FIGURE 1. a) Stratified space X which is a quadrilateral cone, with data points. b) The best fit linear relation $f_1 = c_1$, measured with δ denoting the distance (in X) of data points to the dark quadrilateral $X_1 = X \cap P_1$, where P_1 is the plane defined by the relation.

Example 4.5. We illustrate the algorithm for a stratified space X, which consists of four planar regions (for different planes) each defined in their plane by linear inequality constraints. This yields X as a quadrilateral cone in part a) of Figure 1. In b) of the same figure we show the best fitting linear relation between data points in X. Here for a data point $x \in X$, two possible choices for $\delta(x)$ are either the

distance in X from x to the intersection $X_1 = X \cap P_1$, with the plane P_1 defined by the linear relation $f_1 = c_1$; or alternatively, the minimum value of $|f_1(x) - f_1(x')|$ for $x' \in X \cap P_1$. On a face F_i of X, these different choices for δ differ by $\cos(\theta)$, where θ is the angle between F_i and the normal to P_1.

Then, the data points are projected to their nearest points in the dark quadrilateral X_1 as shown in a) of Figure 2. Then, a second linear relation $f_2 = c_2$ is chosen, with f_2 orthogonal to f_1, to be the best fit for the projected data points in X_1. The subset satisfying both relations is the intersection of X_1 with the plane P_2 defined by the second relation $f_2 = c_2$. It is the intersection of the line $\ell = P_1 \cap P_2$ with X_1.

Part b) of figure 2 shows the two points of intersection. Here for the first choice of $\delta(x)$, all distances are measured in the quadrilateral, so distances of points to the intersection are measured along the shortest paths illustrated by arrows. The pair of intersection points represents the nested mean, and the points are divided into two groups based on which mean point they are closest to.

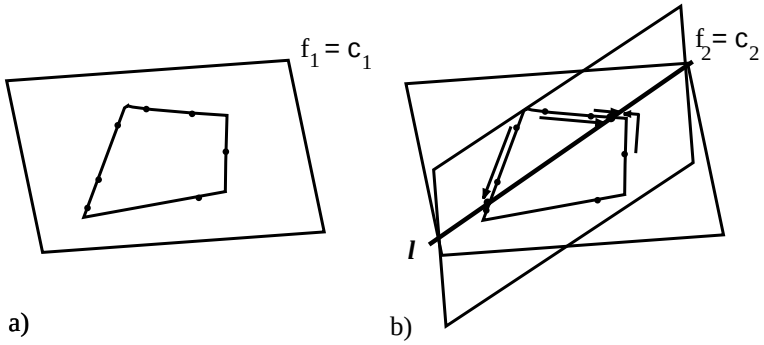

FIGURE 2. a) The nearest points in the dark quadrilateral X_1 to the original data points from Figure 1. Then, a second linear relation $f_2 = c_2$, with f_2 orthogonal to f_1, is the best fit for the projected data points in X_1. The subset satisfying both relations is the intersection with X_1 of the plane P_2 defined by the second relation $f_2 = c_2$. Part b) shows the line $\ell = P_1 \cap P_2$ which intersects X_1 in the pair of points. Here distance is measured in the quadrilateral, so the points' distances are measured along the paths illustrated by arrows.

5. Examples from Statistics

In this section we consider several examples from statistics which can be viewed as a form of backwards PCA using principal nested relations. For these examples, V_j are all the same vector space $V = \mathbb{R}^{n*}$, the space of linear functions on $\mathbb{R}^n$. Then, as we inductively define the relations $f_j = c_j$, $W_j = \langle f_1, \ldots, f_{j-1} \rangle^{\perp}$ for all j. Then, each X_j is generically the transverse intersection of X with the affine subspace defined by $F_{j-1}^{-1}(C_{j-1})$ where $F_{j-1} = (f_1, \ldots, f_{j-1}) : \mathbb{R}^n \to \mathbb{R}^{j-1}$ and $C_{j-1} = (c_1, \ldots, c_{j-1})$.

In the special case where $X = \mathbb{R}^n$, $V = \mathbb{R}^{n*}$, and we use the difference measure of type i) (or equivalently case ii) with $\varphi = Id$). Then, we obtain a backwards

version of PCA and the resulting subspaces are exactly those that would result from usual PCA, except in reverse order. We explain this in several frameworks.

5.1. **Singular value decomposition.** Given a $d \times n$ data matrix A, of n column vectors, the singular value decomposition can be written as $A = US$, where $U = (u_1 \cdots u_r)$ is a $d \times r$ matrix of orthonormal eigenvectors, $S = \begin{pmatrix} s_1 \\ \vdots \\ s_r \end{pmatrix}$ is an $r \times n$ matrix of scores, for r the rank of A.

This gives a simple form of dimensionality reduction as follows. Given $r' < r$, the best rank r' approximation (in the least squares sense) of A is $\sum_{k=1}^{r'} u_k s_k$. An often useful visualization of the relationship between the data vectors (the columns of A) is the scores scatterplot matrix, where the i, jth entry is a scatterplot of the elements of s_i vs. the elements of s_j. The $1, 2$ scatterplot is the projection of the data onto Y_2, the subspace generated by the first two eigenvectors, u_1 and u_2. For $k = 1, \cdots, r$, let Y_k denote the subspace generated by $u_1, \cdots, u_k$. These subspaces are nested in the sense that $\{0\} = Y_0 \subseteq Y_1 \subseteq \cdots \subseteq Y_r$. The backwards approach to singular value decomposition starts with Y_r as the subspace generated by the columns of A. Each Y_k can be written as a constrained version of Y_{k+1},

$$Y_k \;=\; \left\{x \in Y_{k+1} : x^t u_{k+1} = 0\right\}.$$

where u_{k+1} is chosen from among unit vectors orthogonal to $\{u_j : j = k+2, \ldots, r\}$ to minimize $\sum_{i=1}^{n}((A_i^{(k+1)})^t u_{k+1}^2$. Here each $A_i^{(k+1)}$ denotes the orthogonal projection of A_i onto Y_{k+1}. Thus, in our earlier notation, $X = Y_r$, and $X_j = Y_{r-j} = \mathbb{R}^r$, and the relations are defined by $f_k(x) = x^t u_{r-k} = 0$. Because the u_{r-k} are unit vectors, these are linear functions on $\mathbb{R}^r$ which have length 1 for the inner product on $\mathbb{R}^{r\,*}$. Here we have restricted to relations $f(x) = c$ for which $c = 0$.

5.2. **Principal component analysis.** Then, we can view principal component analysis as just a mean centered version of the singular value decomposition. In particular, let $\bar{a}$ denote the sample mean of the column vectors of the data matrix A. Principal component analysis is the application of singular value decomposition to $\tilde{A} = A - \bar{a}$. This results in a rank k approximation of the data by affine spaces $\tilde{Y}_k = \bar{a} + Y_k$. The nested sequence of affine spaces $\tilde{Y}_1, \cdots, \tilde{Y}_r$ can be derived in terms of relations by starting with $\tilde{Y}_r$ as the r dimensional hyperplane generated by the columns of A, and iteratively calculating

$$\tilde{Y}_k \;=\; \left\{x \in \tilde{Y}_{k+1} : (x - \bar{a})^t\, u_{k+1} = 0\right\}.$$

The smallest affine space is the 0 dimensional space $\{\bar{a}\}$.

Now $X = \tilde{Y}_r = \mathbb{R}^r$, and $X_j = \tilde{Y}_{r-j}$, and we are using the same vector space of linear functions as in the preceding case. However, now we allow relations $f_j(x) = c_j$ with $c_j \neq 0$. In fact, the relations defining $\tilde{Y}_k$ are $f_{r-(k+1)}(x)(= x^t u_{k+1}) = \bar{a}^t u_{k+1}$ so $c_{r-k} = \bar{a}^t u_{k+1}$.

5.3. **Principal nested spheres.** The first non-Euclidean method generated using an explicitly backwards approach is the method of principal nested spheres in [JDM12]. For this approach the data points lie on the unit sphere S^d about the origin in $\mathbb{R}^{d+1}$. The method sequentially derives a nested sequence of subspheres of decreasing dimensions which need not be geodesic spheres. At each step the data

points are projected along geodesics onto the next lower dimensional sphere. Each sphere is the intersection of the previous sphere with a hyperplane chosen to best fit the data points on the preceding sphere.

This approach can be viewed as backward PCA for $X = S^d$ for δ given in Example 3.3 with the vector space $V = \mathbb{R}^{d+1\,*}$ of linear functions. This yields the same decreasing sequence of spherical subspaces as would be obtained using principal nested spheres. It gives the successive intersection spheres X_j as a constraint on the preceding sphere X_{j-1} defined by a linear relation $f_j(x) \,(\stackrel{def}{=} x^t \cdot u_j) = c_j$. At the last step, the dimension 0 point is just the geodesic mean of the data projected to the 1 dimensional circle, which is an intuitively appealing notion of centerpoint.

We remark that if we had weighted the coordinates of the data points, then in place of a sphere S^n, we obtain an ellipsoid X defined by $\sum_{i=1}^{n+1} \frac{1}{a_i^2} x_i^2 = 1$. If for the space of functions we again use $V = \mathbb{R}^{n\,*}$, then for either choice of δ_i, we will obtain a sequence of nested ellipsoids.

As noted in Section 2, Principal Nested Spheres was seen in [JDM12] to give substantial improvements in landmark based shape analysis. It has also proven to be very successful in the context of medial (see[SP08]) and skeletal (see [PJG$^+$13]) representations of shape in medical imaging. In particular, [PJG$^+$13] showed that summarizing population variation using an enhanced version called Composite Principal Nested Spheres gave major improvements in automatic segmentation.

5.4. **Principal curves and surfaces.** Principal curves, introduced in [HS89], is the foundation of the field called manifold learning. The idea was to find a flexible one dimensional spline that best approximates higher dimensional data in an appropriately regularized fashion. This was extended to higher dimensional representations by [LT94]. Rank k approximating splines were developed for each k, and a cross validated choice among these was proposed. These ideas spawned the field now called "manifold learning", where the central goal is to find low dimensional manifolds that explain large amounts of structure in high dimensional data sets. The papers [RS00] and [TDSL00] were the cornerstones of this development. See [GTW09] for an elegant mathematical framework for manifold learning.

However, all of these approaches are challenging to interpret in a multi-scale way, i.e. to analyze the data using insights gained simultaneously from several values of k (scales), because the solutions are not nested. It is not at all obvious how to build a nested sequence of manifolds in a forward manner. However, the backwards approach offers a strategy to construct a nested sequence of principal surfaces in a straightforward way. In particular, start with a high dimensional manifold representation, and then successively add constraints to find a nested sequence of lower dimensional manifolds.

5.5. **Tree structured data objects.** There have been several approaches to this data analytic challenge. One of these is the Dyck Path approach of [SSB$^+$12], where a backwards approach is expected to be useful, because the tree representations are restricted to lie in the non-negative cone. Another approach uses phylogenetic tree representations, based on the ideas of [BHV01]. A method for principle component analysis in the space of phylogenetic trees has been proposed by [Nye11]. This gave a reasonable notion of first component, but did not consider higher dimensional approximations, likely because it is not at all clear how this can be done in a forwards manner. Hence, we suggest a backwards approach to this, via a series of

constraints in tree space. A major challenge will be finding an appropriate starting space, in this combinatorially very large space.

5.6. **An Example for Products of Spheres.** Let $X = S^{m_1} \times \cdots \times S^{m_k}$, where each S^{m_i} is a unit sphere in $\mathbb{R}^{m_i+1}$. Then we can embed X in S^m where $m = \sum_{i=1}^k m_i + (k-1)$, by $\psi(v_1, \ldots, v_k) = \frac{1}{\sqrt{k}}(v_1, \ldots, v_k)$. Viewed as a map $\psi : \prod_{i=1}^k \mathbb{R}^{m_i+1} \to \mathbb{R}^{m+1}$, this is a vector space isomorphism. Hence, if f is a homogeneous polynomial of degree r, then so is $f \circ \psi$. Thus the vector space of homogeneous polynomials on $\mathbb{R}^{m+1}$ yield the vector space of homogeneous polynomials on $\prod_{i=1}^k \mathbb{R}^{m_i+1}$, and hence restricted to X. The difference between using the space X and viewing it as a submanifold of S^m is the definition of δ_i at each step. For S^m it is that given by principal nested spheres, while for X itself, it requires either the computation of geodesic distance in X or choosing an appropriate weighted penalty function, which will generally be different.

References

[BHV01] Louis J. Billera, Susan P. Holmes, and Karen Vogtmann, *Geometry of the space of phylogenetic trees*, Advances in Applied Mathematics **27** (2001), no. 4, 733–767.

[Buc78] M. Buchner, *The structure of the cut locus in dimensions less than or equal to six*, Compositio Math. (1978), 103–119.

[DM98] I. L. Dryden and K.V. Mardia, *Statistical analysis of shape*, Wiley, 1998.

[FLPJ04] P.T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi, *Principal geodesic analysis for the study of nonlinear statistics of shape*, Medical Imaging, IEEE Transactions on **23** (2004), no. 8, 995–1005.

[GLDPK76] C. G. Gibson, E. J. N. Looijenga, A. Du Plessis, and Wirthmüller K., *Topological stability of smooth mappings*, vol. 552, Springer Lecture Notes, 1976.

[GTW09] Samuel Gerber, Tolga Tasdizen, and Ross Whitaker, *Dimensionality reduction and principal surfaces via kernel map manifolds*, Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 529–536.

[HHM10] S. Huckemann, T. Hotz, and A. Munk, *Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions*, Statistica Sinica **5** (2010), 1–58.

[HS89] T. Hastie and W. Stuetzle, *Principal curves*, Journal of the American Statistical Association **84** (1989), no. 406, 502–516.

[JDM12] S. Jung, I. L. Dryden, and J. S. Marron, *Analysis of principal nested spheres*, Biometrika **99** (2012), no. 3, 551–568.

[JFM11] S. Jung, M. Foskey, and J. S. Marron, *Principal arc analysis on direct product manifolds*, The Annals of Applied Statistics **5** (2011), no. 1, 578–603.

[JLMP10] S. Jung, X. Liu, J. S. Marron, and S. Pizer, *Generalized PCA via the backward stepwise approach in image analysis*, Brain, Body and Machine (2010), 111–123.

[Jol05] I. Jolliffe, *Principal component analysis*, Wiley Online Library, 2005.

[LT94] M. LeBlanc and R. Tibshirani, *Adaptive principal surfaces*, Journal of the American Statistical Association **89** (1994), no. 425, 53–64.

[Mat73] J. Mather, *Stratifications and mappings*, Dynamical Systems, M. Peixoto, Editor, Academic Press, 1973.

[MJD10] J. S. Marron, S. Jung, and I. L. Dryden, *Speculation on the generality of the backward stepwise view of PCA*, Proceedings of the international conference on Multimedia information retrieval, ACM, 2010, pp. 227–230.

[Nye11] T.M.W. Nye, *Principal components analysis in the space of phylogenetic trees*, The Annals of Statistics **39** (2011), no. 5, 2716–2739.

[PJG+13] S.M. Pizer, S. Jung, D. Goswami, X. Zhao, R. Chaudhuri, J. N. Damon, S. Huckemann, and J. S. Marron, *Nested sphere statistics of skeletal models*, Innovations for Shape Analysis (M. Breuss, A. Bruckstein, and P. Maragos, eds.), Mathematics and Visualization series, Springer-Verlag, 2013, pp. 93–115.

[RS00] Sam T. Roweis and Lawrence K. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science **290** (2000), no. 5500, 2323–2326.

[SP08] K. Siddiqi and S. Pizer (eds.), *Medial representations: mathematics, algorithms and applications*, vol. 37, Springer, 2008.

[SSB+12] D. Shen, H. Shen, S. Bhamidi, Y. Muñoz Maldonado, Y. Kim, and J.S. Marron, *Functional data analysis of tree data objects*, submitted to Journal of Computational and Graphical Statistics (2012), 2716–2739.

[TDSL00] Joshua B. Tenenbaum, Vin De Silva, and John C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science **290** (2000), no. 5500, 2319–2323.

James Damon: Department of Mathematics, University of North Carolina, Chapel Hill, NC 27599-3250, USA

J. S. Marron: Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-3260, USA