Speculation on the Generality of the Backward Stepwise View of PCA

J. S. Marron
University of North Carolina
Chapel Hill, NC 27599
marron@email.unc.edu

Sungkyu Jung
University of North Carolina
Chapel Hill, NC 27599
sungkyu@email.unc.edu

Ian L. Dryden
University of South Carolina
Columbia, SC 29208
dryden@mailbox.sc.edu

ABSTRACT

A novel backwards viewpoint of Principal Component Analysis is proposed. In a wide variety of cases, that fall into the area of Object Oriented Data Analysis, this viewpoint is seen to provide much more natural and accessable analogs of PCA than the standard forward viewpoint. Examples considered here include principal curves, landmark based shape analysis, medial shape representation and trees as data.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Multivariate statistics

General Terms

Theory

1. INTRODUCTION

Principal Component Analysis (PCA) is a workhorse method in a wide variety of fields, for many purposes including data visualization and dimension reduction. The approach is so broadly applicable that it has appeared (perhaps been reinvented) in a variety of disciplines. For example, it is called Empirical Orthogonal Functions in climate and geoscience areas, Proper Orthogonal Decomposition in applied mathematics, and the Karhunen-Loeve Expansion in electrical engineering and probability. In a number of other fields PCA is called Factor Analysis, although in the social sciences where the name originated, the latter term actually refers to a related but deeper model for variation in data.

Wang and Marron (2007) proposed the term *object oriented data analysis* (OODA), to describe a wide array of modern data contexts and methodologies. The main idea is easily understood in terms of the *atom* of the data analysis. In simple statistical analyses, atoms are numbers, and the goal is to understand populations of numbers. In classical multivariate analysis, vectors are the atoms. In the relatively new field of functional data analysis, see Ramsay and Silverman (2002), (2005), the atoms are curves, and the goal is to understand the variation in a data set of curves. OODA generalizes this to more general types of data objects as atoms. Important examples include the study of populations of images, spectra, human movement traces, shape representations, and tree or graph structured objects.

While PCA continues to be broadly useful in OODA, an important limitation is that it is strongly rooted in Euclidean properties of conventional vector space data. In particular, the inner product space notions of subspace and orthogonality are fundamental. For a growing number of OODA applications, these ideas do not apply, because the data spaces are non-Euclidean. A useful distinction among non-Euclidean data types is between *mildly* (discussed further in Section 1.1) and *strongly* (discussed further in Section 1.2) non-Euclidean data.

1.1 Mildly Non-Euclidean Data

In mildly non-Euclidean contexts, the data objects are points on the surface of a curved manifold. Examples include:

- Directional data, see e.g. Fisher et al (1987), Fisher (1993), and Mardia (2000), where each data atom is an angle, represented as a point on either the unit circle or unit sphere (two simple examples of manifolds).
- Landmark based shape representations, see e.g. Dryden and Mardia (1998), where each data atom represents the shape of an object in terms of a 2- or 3-d set of corresponding (across the objects in the population) landmarks. These data are naturally viewed as lying on the surface of a non-Euclidean manifold, after location, scale and rotation have been modded out.
- Medial shape representations, see Siddiqi and Pizer (2008), where data objects are represented in terms of a parametric model involving angular parameters. The natural data space here is a manifold which is a direct product of Euclidean spaces and unit spheres.
- Diffeomorphisms for shape representation, see e.g. Joshi et al (2004), where shapes are represented as deformations of a common atlas object. The set of diffeomorphic warpings lie in a very high dimensional manifold.
- Diffusion tensor imaging, see e.g. Basser et al (1994), Pennec et al (2006) and Dryden et al (2009), a variation of magnetic resonance imaging, where 3 dimensional fluid displacement is measured in terms of tensors that are effectively represented as points on the manifold of symmetric positive definite matrices.

MIR'10, March 29–31, 2010, Philadelphia, Pennsylvania, USA.

- Manifold Learning, see e.g. Roweis and Saul (2000) and Tenenbaum (2000), where the goal is to model the variation in high dimensional Euclidean data in terms of low dimensional approximating curved manifolds.

Effective PCA for data lying on the surface of a manifold is challenging. A simple minded approach would be treat the data as points in the Euclidean embedded space (e.g. when a data set of angles is represented as points on the unit circle, S^1, they can also be thought of as points in $\mathbb{R}^2$), but then the analysis (e.g. projections onto eigenvectors) leaves the manifold, and thus gives neither useful insights into the desired population of data objects, nor effective dimension reduction. This has motivated a number of variations of PCA that are defined within the manifold. The label *mildly non-Euclidean* is used for manifold data, because at each point of a manifold, it can be approximated by a (Euclidean) *tangent hyperplane*. This structure is exploited in *Principal Geodesic Analysis* (PGA), see Fletcher et al (2003), (2004), where the manifold surface is approximated by a tangent plane centered on the geodesic mean (defined in Section 2) of the data. Conventional PCA is performed in the tangent hyperplane, and the results are mapped back into the manifold, so that the usual lines (passing through the sample mean) that best represent the data are replaced by geodesics (through the geodesic mean). Some interesting variations on PGA are discussed in Section 3.1.

1.2 Strongly Non-Euclidean Data

Strongly non-Euclidean data is used to describe data sets where each atom is a tree or graph structured object. Important examples include phylogenetic trees, see e.g. Billera et al (2001), social networks, see e.g. Wasserman and Faust (1995), and various anatomical objects, such as blood vessel trees, see e.g. Bullitt and Aylward (2002), and lung airway trees, see e.g. Tschirren et al (2002). The label *strongly Euclidean* are used for this type of data, because tree or graph spaces do not seem to have any useful notion of approximating tangent plane. Furthermore, notions such as geodesics are much more challenging to define and work with, and have properties that are even farther from the properties of lines in Euclidean space. These issues are discussed in more detail in Section 3.4.

2. NOTATION AND MATHEMATICAL DEVELOPMENT

The input to conventional Euclidean PCA is a set of d-dimensional vectors $X_1, ..., X_n \in \mathbb{R}^d$. It is notionally convenient to aggregate the (column) vectors of data into a data matrix $\mathcal{X} = [X_1 \cdots X_n] \in \mathbb{R}^{d \times n}$. A conventional approach to PCA is given by the following steps:

1. Use the sample (Euclidean) mean, $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$. to represent the centerpoint of the data set.

2. Find the line, i.e. affine 1-d space, which best approximates the data. The Pythagorean Theorem shows that this line can be thought of as either the minimizer of the sum of squared distances from the data to the line, or equivalently as the maximizer of the variance of the data projected to the line. A different application of the Pythagorean Theorem shows that the minimizing line must go through the sample mean, $\overline{X}$. This best one dimensional approximating line (affine 1-d subspace) can be written in the form
$$AS_1^1 = \left\{ \overline{X} + cU_1 : c \in \mathbb{R} \right\},$$
where U_1 is the first eigenvector from the eigen-analysis of the sample covariance matrix, $\widehat{\widehat{\Sigma}} = n^{-1} \left(\mathcal{X} - \overline{X} \right) \left(\mathcal{X} - \overline{X} \right)^t$. Thus AS_1^1 is the line centered at $\overline{X}$ pointing in the direction U_1.

3. Next find the line, restricted to the plane through $\overline{X}$ that is orthogonal to U_1, that best approximates the data. The result can be shown to be the 1-d affine space of the form
$$AS_2^1 = \left\{ \overline{X} + cU_2 : c \in \mathbb{R} \right\},$$
where U_2 is the second eigenvector of $\widehat{\widehat{\Sigma}}$, and where $U_1 \perp U_2$. Another application of the Pythagorean Theorem shows that the 2-d affine space
$$AS^2 = AS_1^1 \cup AS_2^1 = \left\{ \overline{X} + c_1 U_1 + c_2 U_2 : c_1, c_2 \in \mathbb{R} \right\}$$
is the best two dimensional approximation of the data (again either in terms of minimum sum of squares, or maximum variation of the projections).

4. This process can be iterated, for $k = 3, 4, \cdots, d$, to obtain U_k and AS^k, which results in the k-d affine space
$$AS^k = \bigcup_{j=1}^{k} AS^j = \left\{ \overline{X} + \sum_{j=1}^{k} c_j U_j : c_1, \cdots, c_k \in \mathbb{R} \right\}$$
as the best k-dimensional affine approximation of the data (again in terms of either minimum residuals of maximum variation).

One hurdle to the analysis of either data on a manifold, or a population of tree-graph structured objects, is to obtain an appropriate generalization of the sample mean for non-Euclidean spaces. A common approach to this is the Fréchet mean, defined as:
$$\arg\min_{x \in \mathbb{R}^d} \sum_{i=1}^{n} \delta \left(x, X_i \right)^2, \tag{1}$$
where δ is some metric. When δ is Euclidean distance, a simple calculation shows that the Fréchet mean is the same as the sample mean, $\overline{X}$. The Fréchet representation, (1), is far more general because it is well defined in any metric space. As noted in Koenker (2007) ideas such as the Fréchet notion of centerpoint have been central to the development of a large number of robust statistical methods, such as M estimation, where an important goal is finding definitions of population center that are not strongly affected by outliers (e.g. replacing the power 2 in (1) by 1 gives a version of median). For manifold data, it is natural to take the distance δ to be geodesic distance (essentially arc length along the surface of the manifold). This results in the *geodesic mean*, which provides the centerpoint for PGA as discussed in Section 1. The geodesic mean is not always unique for manifolds of positive curvature, although there are sufficient conditions on the support of the data to check for uniqueness, e.g. see Le (1995). Interesting work combining notions of robustness and manifold data can be found in Fletcher et al (2009).

More challenging is the generalization of PCA to non-Euclidean data contexts. As noted in Section 3, it is not

always clear how to even define the first principal component. The notion of second principal component can be even more challenging from the conventional viewpoint, as there is usually no notion of orthogonal subspace available. This makes it generally hard to usefully formulate an appropriate analog of the approximating sequence AS_1^1, AS_2^1, $\cdots$, AS_d^1, which then generates the nested sequence of (affinely shifted) subspaces

$$\{\overline{X}\} \subseteq AS_1^1 \subseteq AS^2 \subseteq \cdots \subseteq AS^d. \quad (2)$$

Borrowing terminology from step-wise multiple linear regression, this approach to PCA is called the *forward view of PCA*. An alternate viewpoint, which comes from rewriting (2) as

$$AS^d \supseteq AS^{d-1} \supseteq \cdots \supseteq AS_1^1 \supseteq \{\overline{X}\},$$

is called the *backward view of PCA*. The idea is that instead of building the sequence from lower dimensions into higher dimensions, as in Steps 1-4 above, instead the process goes in the opposite direction.

For Euclidean data forward and backward PCA are equivalent. However, for non-Euclidean data, this is no longer the case. While the forward approach is typically thought of as the most intuitive, especially for teaching PCA, the perhaps surprising premise of this paper is that, as shown in the next section, the backward approach seems to yield the easiest generalization to non-Euclidean spaces.

3. APPLICATIONS AND SPECULATION

Some specific applications, and speculation about the future, of the backwards viewpoint of PCA are discussed in the following subsections.

3.1 Principal Arc Analysis

There are a number of variations of the notion of PCA for data that lie on the surface of S^2, the usual unit sphere in $\mathbb{R}^3$. As noted in Section 1, Fletcher proposed PGA, where the data are approximated by geodesics (great circles) that are constrained to pass through the geodesic mean. A limitation of this approach is that performance will clearly be best when the data points are closest to the geodesic mean,. resulting in effective tangent hyperplane approximation. However, this approximation can be very poor. An extreme case is data distributed roughly uniformly along the equator. In this case the equator itself will be the geodesic that best fits the data. However, the (non-unique) geodesic mean is either the north or the south pole, so the constraint that the first PG pass through the geodesic mean is a strong one. This will result in two components being required to describe the variation in this essentially one dimensional data.

This consideration motivated Huckemann et al (2010) to improve PGA to *Geodesic PCA*. The main improvement is that in Geodesic PCA the best fitting geodesic is drawn from all the set of all geodesics, instead of being constrained to go through the geodesic mean. In the simple equatorial data example described above, this gives a simple one dimensional mode of variation of the data. A consequence worth noting is the geodesic mean no longer appears in the sequence (2). Huckemann et al (2010) address this issue by redefining the center of the data to be an appropriately chosen intersection of the first two Geodesic PCs. Thus this falls in the framework of a backwards approach to PCA.

A related example, that has motivated the idea of Principal Arc Analysis (PAA), proposed by Jung et al (2010), is data distributed along the Tropic of Capricorn. In this case, both PGA and Geodesic PCA will use two components to adequately describe this intrinsically one dimensional data set. PAA addresses this challenge by seeking the small (i.e. not necessarily great) circle which best fits the data. For the Tropic of Capricorn example, the tropic becomes the best fit small circle, which gives an efficient one dimensional description of the data. PAA was shown to provide just this type of efficient data description for m-rep data in Jung et al (2010). PAA also falls in the domain of a backward approach to PCA, because one starts with the full data sphere, reduces to the best fitting small circle, then takes an appropriate mean of the projected data.

3.2 Principal Nested Spheres

A more serious example of the concept of backwards PCA, and in fact its motivation, comes from some unpublished work by Jung, Dryden and Marron, who are developing Principal Nested Sphere (PNS) Analysis, for analysis of landmark based shape data objects. That work was motivated by the problem of extending PAA to landmark based shape data.

Classical approaches to PCA for landmark based data, see e.g. Section 5.5 of Dryden and Mardia (1998), is a tangent plane approach. For a number of interesting data sets, this has resulted in an analysis with an apparent one dimensional mode of variation again curving through more than one component (apparently an analog of the problems encountered by PGA in the above examples).

PNS addresses this by starting with the full data space (essentially a high dimensional sphere), and directly focusing on backwards PCA, by iteratively reducing the dimension of the fit sphere. This has produced more representative modes of variation in a number of standard examples.

3.3 Principal Curves

Hastie and Stuetzle (1989), proposed *principal curves*, as an extension of PCA. The idea is to replace the linear affine approximating line, AS_1^1, with a possibly curved version, that is appropriately regularized (resulting in a spline fit of the data) to avoid overfitting. While the basic idea is very appealing, no truly convincing analog of AS_2^1 has yet been proposed. The *principal surface* approach of Leblanc and Tibshirani (1994) contains some interesting preliminary ideas and methodologies, but it completely ignores nesting issues, and thus cannot be viewed as an extension of forward PCA. We believe that, despite the large literature in this area (there are hundreds of references to Hastie and Stuetzle (1989)), the lack of forward PCA proposals is because, as in Section 3.2, the forward approach to PCA unnecessarily obscures the key issues, which are made very clear by taking a backward approach. In particular, instead of trying to find an analog of AS_2^1, which is perpendicular in some sense to AS_1^1, it seems far more natural to fit a two dimensional spline, the natural analog of AS^2, to the data. Similarly, the k dimensional analog of AS^k, is a spline of dimension k.

3.4 Trees as Data

It is interesting to study the tree line analysis of Aydin et al (2009), from the perspective of backwards PCA. Tree lines are an attempt at a one dimensional representation of

tree data, defined in terms of successive tree grow growth. A notion of projection of a data point onto a tree line leads to the development of a best fitting tree line, which is thus taken to be a tree based analog of the first principal component. Originally thinking from a forward perspective on PCA, Aydin et al (2009) went on to develop a notion of second component by finding another direction which minimizes the projection onto the union of the first line, and the new one. In retrospect, this appears more like a backwards approach because instead of finding the second component by optimization, and then constructing the two dimensional entity, the two dimensional representation is the foundation of the analysis.

The notion of backwards PCA can also generate new approaches to tree line PCA. In particular, following the backwards PCA principal in full suggests first optimizing over a number of lines together, and then iteratively reducing the number of lines.

4. REFERENCES

[1] B. Aydin, G. Pataki, H. Wang, E. Bullitt, and J. S. Marron. A principal component analysis for trees. *to appear in Annals of Applied Statistics*, 2009.

[2] P. J. Basser, J. Mattiello, and D. Le Bihan. MR diffusion tensor spectroscopy and imaging. *Biophysics Journal*, 66:259–267, 1994.

[3] L. J. Billera, S. P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.*, 27:733–767, 2001.

[4] E. Bullitt and S. Aylward. Volume rendering of segmented image objects. *IEEE Trans. Medical Imaging*, 21:998–1002, 2002.

[5] I. L. Dryden, A. Koloydenko, and D. Zhou. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Annals of Applied Statistics*, 3:1102–1123, 2009.

[6] I. L. Dryden and K. V. Mardia. *Statistical shape analysis*. John Wiley & Sons Ltd., Chichester, UK., 1998.

[7] N. I. Fisher. *Statistical analysis of circular data*. Cambridge University Press, Cambridge, UK, 1993.

[8] N. I. Fisher, T. Lewis, and B. J. J. Embleton. *Statistical analysis of spherical data*. Cambridge University Press, Cambridge,UK, 1993.

[9] P. T. Fletcher, C. Lu, and S. Joshi. Statistics of shape via principal geodesic analysis on lie groups. In *In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 95–101, Los Alamitos, CA, 2003.

[10] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. Medical Imaging*, 23:995–1005, 2004.

[11] P. T. Fletcher, S. Venkatasubramanian, and J. S. The geometric median on Riemannian manifolds with application to robust atlas estimation. *Neuroimage*, 45 Suppl 1:S143–152, 2009.

[12] T. Hastie and W. Stuetzle. Principal curves. *J. Amer. Statist. Assoc.*, 84(406):502–516, 1989.

[13] S. Huckemann, T. Hotz, and A. Munk. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric lie group actions. *to appear in Statistica Sinica*, 20(1), 2010.

[14] S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *Neuroimage*, 23 Suppl 1:S151–160, 2004.

[15] S. Jung, I. L. Dryden, and J. S. Marron. Principal Nested Spheres with application to planar shape data. *manuscript in preparation*, 2010.

[16] R. Koenker. The Median Is the Message : Toward the Fréchet Median. *Journal de la Société Francaise de Statistique*, 147:61–64, 2007.

[17] H.-L. Le. Mean size-and-shapes and mean shapes: a geometric point of view. *Advances in Applied Probability*, 27:44–55, 1995.

[18] M. LeBlanc and R. Tibshirani. Adaptive principal surfaces. *Journal of the American Statistical Association*, 89:53–64, 1994.

[19] K. V. Mardia and P. E. Jupp. *Directional statistics*. John Wiley & Sons Ltd., 2000.

[20] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. Springer, New York, 2002.

[21] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2nd ed. edition, 2005.

[22] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[23] K. Siddiqi and S. Pizer. *Medial Representations: Mathematics, Algorithms and Applications*. Springer, 2008.

[24] J. B. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2322, 2000.

[25] J. Tschirren, K. Palágyi, J. M. Reinhardt, E. A. Hoffman, and M. Sonka. Segmentation, skeletonization and branchpoint matching—a fully automated quantitative evaluation of human intrathoracic airway trees. *Proc. Fifth International Conterence on Medical Image Computing and Computer-Assisted Intervention, Part II. Lecture Notes in Comput. Sci.*, 2489:12–19, 2002.

[26] H. Wang and J. S. Marron. Object oriented data analysis: sets of trees. *Annals of Statistics*, 35:1849–1873, 2007.

[27] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge Univeristy Press, UK, 1995.

[28] P. Xavier, F. Pierre, and N. Ayache. A Riemannian Framework for Tensor Computing. *International Journal of Computational Vision*, 66:41–66, 2006.