Sample-limited L p Barycentric Subspace Analysis on Constant Curvature Spaces

Xavier Pennec

Sample-limited L_p Barycentric Subspace Analysis on Constant Curvature Spaces

Xavier Pennec[1]

Côte d'Azur University (UCA), Inria, France
xavier.pennec@inria.fr

Abstract. Generalizing Principal Component Analysis (PCA) to manifolds is pivotal for many statistical applications on geometric data. We rely in this paper on *barycentric subspaces*, implicitly defined as the locus of points which are weighted means of $k+1$ reference points [8, 9]. Barycentric subspaces can naturally be nested and allow the construction of inductive forward or backward nested subspaces approximating data points. We can also consider the whole hierarchy of embedded barycentric subspaces defined by an ordered series of points in the manifold (a flag of affine spans): optimizing the accumulated unexplained variance (AUV) over all the subspaces actually generalizes PCA to non Euclidean spaces, a procedure named Barycentric Subspaces Analysis (BSA).
In this paper, we first investigate sample-limited inference algorithms where the optimization is limited to the actual data points: this transforms a general optimization into a simple enumeration problem. Second, we propose to robustify the criterion by considering the unexplained p-variance of the residuals instead of the classical 2-variance. This construction is very natural with barycentric subspaces since the affine span is stable under the choice of the value of p. The proposed algorithms are illustrated on examples in constant curvature spaces: optimizing the (accumulated) unexplained p-variance (L_p PBS and BSA) for $0 < p \leq 1$ can identify reference points in clusters of a few points within a large number of random points in spheres and hyperbolic spaces.

1 Introduction

Principal Component Analysis (PCA) is the ubiquitous tool to obtain low dimensional representation of the data in linear spaces. To generalize PCA to Riemannian manifolds, one can analyze the covariance matrix of the data in the tangent space at the Fréchet mean (Tangent PCA). This is often sufficient when data are sufficiently centered around a central value (unimodal or Gaussian-like data), but generally fails for multimodal or distributions with a large variability with respect to the curvature. Instead of maximizing the explained variance, methods minimizing the unexplained variance were proposed: Principal Geodesic Analysis (PGA) [4] and Geodesic PCA (GPCA) [5] minimize the distance to a Geodesic Subspace (GS) spanned by the geodesics going through a point with tangent vector in a linear subspace of the tangent space.

Barycentric subspaces are a new type of subspaces in manifolds recently introduced which are implicitly defined as the locus of weighted means of k+1 reference points (with positive or negative weights) [8, 9]. Depending on the definition of the mean, we obtain the Fréchet, Karcher or Exponential Barycentric

subspaces (FBS/KBS/EBS). The Fréchet (resp. Karcher) barycentric subspace of the points $(x_0, \dots x_k) \in \mathcal{M}^{k+1}$ is the locus of weighted Fréchet (resp. Karcher) means of these points, i.e. the set of global (resp. local) minima of the weighted variance: $\sigma^2(x, \lambda) = \frac{1}{2} \sum_{i=0}^{k} \underline{\lambda}_i \operatorname{dist}^2(x, x_i)$, where $\underline{\lambda} = \lambda / (\sum_{j=0}^{k} \lambda_j)$:

$$\mathrm{FBS}(x_0, \dots x_k) = \left\{ \arg\min_{x \in \mathcal{M}} \sigma^2(x, \lambda), \lambda \in \mathcal{P}_k^* = \left\{ \lambda \in \mathbb{R}P^n / \mathbb{1}^\top \lambda \neq 0 \right\} \right\}.$$

The EBS is the locus of weighted exponential barycenters of the reference points (critical points of the weighted variance) defined outside their cut-locus by:

$$\mathrm{EBS}(x_0, \dots x_k) = \{ x \in \mathcal{M} \setminus C(x_0, \dots x_k) | \exists \lambda \in \mathcal{P}_k^* : \textstyle\sum_i \lambda_i \log_x(x_i) = 0 \}.$$

Thus, we clearly see the inclusion $FBS \subset KBS \subset EBS$. The metric completion of the the EBS is called the *affine span* $\mathrm{Aff}(x_0, \dots x_k)$. Its completeness allows ensuring that a closest point exists on the subspace, which is fundamental in practice for optimizing the subspaces by minimizing the residuals of the data to their projection. This definition works on metric spaces more general than Riemannian manifolds. In stratified metric spaces, the barycentric subspace spanned by points belonging to different strata naturally maps over several strata.

Barycentric subspaces can be characterized using the matrix field $Z(x) = [\log_x(x_0), \dots \log_x(x_k)]$ of the log of the reference points x_i. This is a smooth field outside the cut locus of the reference points. The EBS is the zero level-set of the smallest singular value of $Z(x)$. The associated right singular vector gives the weights λ that satisfy the barycentric equation $\sum_i \lambda_i \log_x(x_i) = 0$. This simple equation generates a very rich geometry: at regular points where the Hessian of the weighted distance to the reference points is not degenerate, the EBS is a stratified space of maximal dimension k. In general, the largest stratum defines locally a submanifold of dimension k.

From PCA to barycentric subspace analysis The nestedness of approximation spaces is one of the most important characteristics for generalizing PCA to more general spaces [1]. Barycentric subspaces can easily be nested by adding or removing one or several points at a time, which corresponds to put the barycentric weight of this (or these) point(s) to zero. This gives a family of embedded submanifolds called a flag because this generalizes flags of vector spaces [9].

With a forward analysis, we compute iteratively the flag of affine spans by adding one point at a time keeping the previous ones fixed. Thus, we begin by computing the optimal barycentric subspace $\mathrm{Aff}(x_0) = \{x_0\}$, which may be a Karcher mean or more generally a stationary value of the unexplained variance, i.e. a Karcher mean. Adding a second point amounts to computing the geodesic passing through the mean that best approximates the data. Adding a third point now generally differs from PGA. In practice, the forward analysis should be stopped at a fixed number or when the variance of the residuals reaches a threshold (typically 5% of the original variance). We call this method the forward barycentric subspace (FBS) decomposition. Due to the greedy nature of this forward method, the affine span of dimension k defined by the first $k+1$ points is not in general the optimal one minimizing the unexplained variance.

The backward analysis consists in iteratively removing one dimension. As the affine span of $n+1$ linearly independent points generate the full manifold, the optimization really begins with n points. Once they are fixed, the optimization boils down to test which point should be removed. In practice, we may rather optimize $k+1$ points to find the optimal k-dimensional affine span, and then reorder the points using a backward sweep to find inductively the one that least increases the unexplained variance. We call this method the k-dimensional pure barycentric subspace with backward ordering (k-PBS). With this method, the k-dimensional affine span is optimizing the unexplained variance, but there is no reason why any of the lower dimensional ones should do.

In order to obtain optimal subspaces which are embedded, it is necessary to define a criterion which depends on the whole flag of subspaces and not on each of the subspaces independently. In PCA, one often plots the unexplained variance as a function of the number of modes used to approximate the data. This curve should decreases as fast as possible from the variance of the data (for 0 modes) to 0 (for n modes). A standard way to quantify the decrease consists in summing the values at all steps. This idea gives the Accumulated Unexplained Variances (AUV) criterion [9], which is analogous to the Area-Under-the-Curve (AUC) in Receiver Operating Characteristic (ROC) curves. This leads to an interesting generalization of PCA on manifolds called Barycentric Subspaces Analysis (BSA). In practice, one can stop at a maximal dimension k like for the forward analysis in order to limit the computational complexity. This analysis limited to a flag defined by $k+1$ points is denoted k-BSA.

2 Sample-limited L_p barycentric subspace inference

This paper investigates variants of the three above barycentric subspace analysis algorithms (FBS, k-PBS and k-BSA) along two main directions. First, we limit in Section 2 the optimization of flags of barycentric subspaces to the sample points of the data: this transforms a general optimization into a very simple enumeration problem. Second, we robustify in Section 2 the optimized criteria by considering the unexplained p-variance of the residuals instead of the classical 2-variance. This construction is very natural with barycentric subspaces since the affine span is stable under the choice of the value of p.

Sample-limited barycentric subspace inference In several domains, it has been proposed to limit the inference of the Fréchet mean to the data-points only. In neuroimaging studies, the individual image minimizing the sum of square deformation distance to other subject images is a good alternative to the mean template (a Fréchet mean in deformation and intensity space) because it conserves the original characteristics of a real subject image [7]. Beyond the Fréchet mean, [3] proposed to define the first principal component mode as the unexplained variance minimizing geodesic going through two of the data points. The method named *set statistics* was aiming to accelerate the computation of statistics on tree spaces. [11] further explored this idea under the name of *sample-limited geodesics* in the context of PCA in phylogenetic tree space. In both cases, defining higher order principal modes was seen as a challenging research topic.

With barycentric subspaces, the idea of sample-limited statistics naturally extends to any dimension by restricting the search to the (flag of) affine spans that are parametrized by points sampled form the data. The implementation boils down to an enumeration problem. With this technique, the reference points are never interpolated as they are by definition sampled from the data. This is a important advantage for interpreting the modes of variation since we may go back to other information about the samples like the medical history and disease type. The search can be done exhaustively for a small number of reference points. The main drawback is the combinatorial explosion of the computational complexity with the dimension for the optimal order-k flag of affine spans, which is involving $O(N^{k+1})$ operations, where N is the number of data points. In this paper we perform an exhaustive search, but approximate optima can be sought using a limited number of randomly sampled points [3].

Stability of barycentric subspaces by L_p norms Since barycentric subspaces minimize the weighted variance, one could think of taking a power p of the metric to define the p-variance $\sigma^p(x) = \frac{1}{p}\sum_{i=0}^{k} \operatorname{dist}^p(x, x_i)$. The global minima of this p-variance defines the Fréchet median for $p = 1$, the Fréchet mean for $p = 2$ and the barycenter of the support of the distribution for $p = \infty$. This suggests to further generalize barycentric subspaces by taking the locus of the minima of the weighted p-variance $\sigma^p(x, \lambda) = \frac{1}{p}\sum_{i=0}^{k} \underline{\lambda}_i \operatorname{dist}^p(x, x_i)$. However, it turns out that the critical points of the weighted p-variance are necessarily included in the EBS: the gradient of the p-variance at a non-reference point is

$$\nabla_x \sigma^p(x, \lambda) = \nabla_x \frac{1}{p}\sum_{i=0}^{k} \lambda_i (\operatorname{dist}^2(x, x_i))^{p/2} = -\sum_{i=0}^{k} \underline{\lambda}_i \operatorname{dist}^{p-2}(x, x_i) \log_x(x_i).$$

Thus, we see that the critical points of the p-variance satisfy the equation $\sum_{i=0}^{k} \lambda'_i \log_x(x_i) = 0$ for the new weights $\lambda'_i = \lambda_i \operatorname{dist}^{p-2}(x, x_i)$. Thus, they are also elements of the EBS and changing the power of the metric just amounts to a reparametrization of the barycentric weights. This stability of the EBS / affine span with respect to the power of the metric p shows that the affine span is really a central notion.

***L_p* barycentric subspaces fitting and analysis** While changing the power does not change the subspace definition, it has a drastic impact on its estimation: minimizing the sum of L_p distance to the subspace for non-vanishing residuals obviously changes the relative influence of points. It is well known that medians are more robust than least-squares estimators: the intuitive idea is to minimize the power of residuals with $1 \leq p \leq 2$ to minimize the influence of outliers. For $0 < p < 1$, the influence of the closest points becomes predominant, at the cost of non-convexity. In general, this is a problem for optimization. However, since we perform an exhaustive search in our sample-limited setting, this is not a problem here. At the limit of $p = 0$, all the barycentric subspaces containing $k+1$ points (i.e. all the sample-limited barycentric subspaces of dimension k that we consider) have the same L_0 sum of residuals, which is a bit less interesting.

For a Euclidean space, minimizing the sum L_p norm of residuals under a rank k constraint is essentially the idea of the robust R1-PCA [2]. However, as noted in [6], an optimal rank k subspace is not in general a subspace of the optimal subspace of larger ranks: we loose the nestedness property. In this paper, we do not follow the PCA-L1 approach they propose, which maximizes the L_1 dispersion within the subspace. On manifolds, this would lead to a generalization of tangent-PCA maximizing the explained p-variance. In contrast, we solve this problem by minimizing the Accumulated Unexplained p-Variance (L_p AUV) over all the subspaces of the flag which is considered. Since the subspaces definition is not impacted by the power p, we can compare the subspaces' parameters (the reference points) for different powers. It also allows to simplify the algorithms: as the (positive) power of a (positive) distance is monotonic, the closest point to an affine span for the 2-distance remains the closest point for the p-distance. This give rise to three variations of our previous estimation algorithms:

- The Forward Barycentric Subspace decomposition (L_p k-FBS) iteratively adds the point that minimizes the unexplained p-variance up to $k+1$ points.
- The optimal Pure Barycentric Subspace with backward reordering (L_p k-PBS) estimates the $k+1$ points that minimize the unexplained p-variance, and then reorders the points accordingly for lower dimensions.
- The Barycentric Subspace Analysis of order k (L_p k-BSA) looks for the flag of affine spans defined by $k+1$ ordered points that optimized the L_p AUV.

3 Examples on constant curvature spaces

We consider here the exhaustive sample-limited version of the three above algorithms and we illustrate some of their properties on spheres and hyperbolic spaces. Affine spans in spheres are simply lower dimensional great subspheres [8, 9]. The projection of a point of a sphere on a subsphere is almost always unique (with respect to the spherical measure) and corresponds to the renormalization of the projection on the Euclidean subspace containing the subsphere. The same property can be established for hyperbolic spaces, which can be viewed as pseudo-spheres embedded in a Minkowski space. Affine spans are great pseudo-spheres (hyperboloids) generated by the intersection of the plane containing the reference points with the pseudo-sphere, and the closest point on the affine span is the renormalization of the unique Minkowski projection on that plane. In both cases, implementing the Riemannian norm of the residuals is very easy and the difficulty of sample-limited barycentric subspace algorithms analysis resides in the computational complexity of the exhaustive enumeration of tuples of points.

Example on real shape data For planar triangles, the shape space (quotient of the triad by similarities) boils down to the sphere of radius 1/2. The shape of three successive footprints of Mount Tom Dinosaur trackway 1 described in [10, p.181] is displayed on Fig.1 (sample of 9 shapes). In this example, the reference points of the L_2 BSA stay the same from $k = 0$ to 3 and identical to the ones

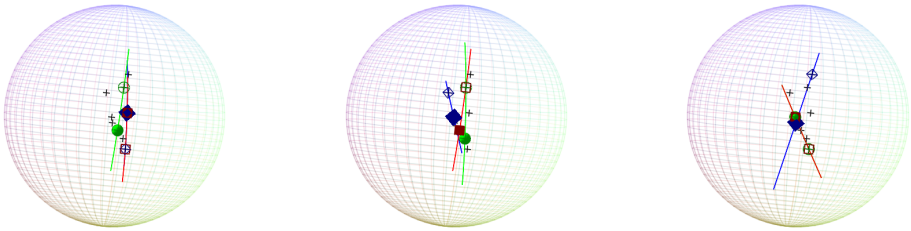

Fig. 1. Mount Tom Dinosaur trackway 1 data (symbol +), with $p = 2$ (**left**), $p = 1$ (**middle**) and $p = 0.1$ (**right**). For each method (FBS in blue, 1-PBS in green and 1-BSA in red), the first reference point has a solid symbol. The 1D mode is the geodesic joining this point to the second reference point (empty symbols).

of the L_2 FBS. This is a behavior that we have observed for simulated examples when the variance of each mode is sufficiently different. The optimal L_2 1-PBS (the best geodesic approximation) picks up different reference points. For $p = 1$, the L_1 FBS is highly influenced by the location of the Fréchet median (solid blue symbol at the center Fig.1) and we see that the optimal L_1 1-PBS and 1-BSA pick-up a different zero-th order mode (solid green and red symbols at the center). For a very low value $p = 0.1$, the optimal 1D subspace $L_{0.1}$ 1-PBS and the 1-BSA agree on points defining a geodesic excluding the 3 points located on the top right while the forward method gives something less intuitive.

3 clusters on a 5D sphere In this synthetic dataset, we consider three clusters of 10, 9 and 8 points around the axes e_1, e_2 and e_3 (the vertices of an equilateral triangle of side length $\pi/2$) on a 5-dimensional sphere (embedded in 6D) with an error of standard deviation $\sigma = 6°$. We add 30 points uniformly sampled on the sphere to simulate three clusters on a 2-sphere with 50% of outliers. The ideal flag of subspaces is a pure 2D subspace spanning the first three coordinates with points at the cluster centers (Fig.2).

For the L_2 metric, one first observes that at zero-th and first order, FBS, PBS and BSA estimate the same reference points which do not fall into any

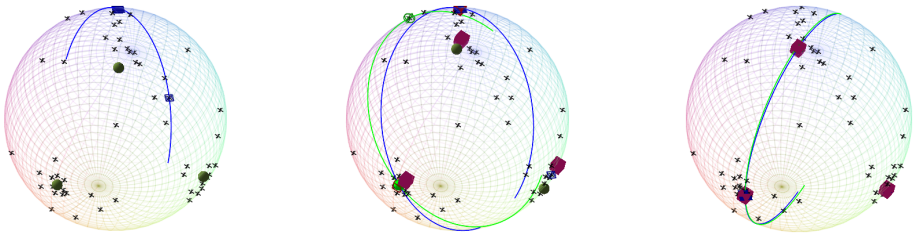

Fig. 2. Analysis of 3 clusters on a 5D sphere, projected to the expected 2-sphere, with $p = 2$ (**left**), $p = 1$ (**middle**) and $p = 0.1$ (**right**). For each method (FBS in blue, 1-PBS in green, 1-BSA in red), the 1D mode is a geodesic joining the two reference point. The three reference points of 2-PBS are represented with dark green solid circles, and the ones of 2-BSA with deep pink solid boxes.

of the three clusters (blue point and geodesic on Fig.2, left). For the second order approximation, which should cover the ideal 2-sphere, 2-BSA and 2-FBS continue to agree on the previous reference points and pick-up a third reference point within the smallest cluster (dark green circle on top of the sphere). Thus, we get at most one of the reference point in one of the clusters, except for the optimal 2-subspace (2-PBS) which makes a remarkable job by picking one point in each cluster (dark green point on Fig.2, left).

With the L_1 metric, we first observe that the FBS is fooled by the geometric median, which is not in any of the three clusters. The two other reference points successively added fall in two of the clusters. The L_1 optimal subspace (1-PBS) and 1-BSA find one of their reference points in a cluster, but the second point is still an outlier. When we come to 2D subspaces, both the 2-PBS and 2-BSA algorithms pick-up reference points in the three clusters, although they are not the same (dark green circles and deep pink solid boxes on Fig.2, center). For a lower value of the power $p = 0.1$, all three reference points of the FBS are identical and within the three clusters, demonstrating the decrease in sensibility of the method to the outliers. The first and second order PBS and BSA consistently find very similar reference points within the clusters (Fig.2, right).

3 clusters on a 5D hyperbolic space This example emulates the same example as above but on the 5D hyperbolic space: we draw 5 random points (tangent Gaussian with variance 0.015) around each vertex of an equilateral triangle of length 1.57 centered at the bottom of the 5D hyperboloid embedded in the (1,5)-Minkowski space. As outliers, we add 15 points drawn according to a tangent Gaussian of variance 1.0 truncated at a maximum distance of 1.5 around the bottom of the 5D hyperboloid. This simulates three clusters on a 2-pseudo-sphere with 50% of outliers (Fig.3). With the L_2 hyperbolic distance, the 1-FBS and 1-BSA methods select outliers for their two reference points. 1-PBS manages to get one point in a cluster. For the two dimensional approximation, the 2-FBS and the 2-PBS select only one reference points within the clusters while 2-BSA correctly finds the clusters (Fig.3 left, dark green points). With the L_1 distance,

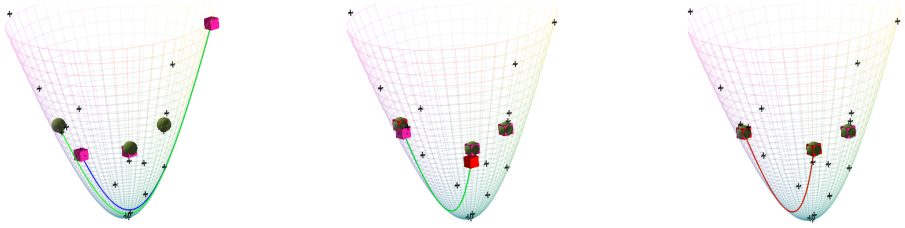

Fig. 3. Analysis of 3 clusters on a 5D hyperbolic space, projected to the expected 2-pseudo-sphere, with $p = 2$ (**left**), $p = 1$ (**middle**) and $p = 0.5$ (**right**). For each method (FBS in blue, 1-PBS in green, 1-BSA in red), the 1d mode is figured as a geodesic joining the two reference point. The three reference points of 2-PBS are represented with dark green solid circles, and the ones of 2-BSA with deep pink solid boxes.

FBS, PBS and BSA select 3 very close points within the three clusters (Fig.3 center). Lowering the power to $p = 0.5$ leads to selecting exactly the same points optimally centered within the 3 clusters for all the methods (Fig.3 right).

4 Conclusion

We have presented in this paper the extension of the barycentric subspace analysis approach to L_p norms and developed sample-limited inference algorithms which are quite naturally suited to barycentric subspaces, thanks to their definition using points rather than vectors as in more classical extensions of PCA. Experimental results on spheres and hyperbolic spaces demonstrate that the forward and optimal estimations of a k-subspace may differ from the barycentric subspace analysis optimizing the full flag of embedded subspaces together, even with the L_2 norm on residuals. This behavior differs from the one in Euclidean space where all methods are identical. Experiments also demonstrate that taking the L_p norm for $p < 2$ improves the robustness. Combined with the sample-limited estimation technique, we can even go well below $p = 1$ using exhaustive optimization. The main limitation of the optimal pure barycentric subspace (PBS) and barycentric subspace analysis (BSA) algorithms is their computational complexity which is exponential in the number of reference points. Thus, in order to increase the dimensionality, we now nee to develop efficient stochastic sampling techniques which allow to quickly pick up good reference points.

References

1. J. Damon and J. S. Marron. Backwards Principal Component Analysis and Principal Nested Relations. *J. of Math. Imaging and Vision*, 50(1-2):107–114, 2013.
2. C. Ding, D. Zhou, X. He, and H. Zha. R1-pca: Rotational invariant l1-norm principal component analysis for robust subspace factorization. In *Proc. of ICML '06*, pages 281–288, New York, NY, USA, 2006.
3. A. Feragen, M. Owen, *et al.* *Tree-Space Statistics and Approximations for Large-Scale Analysis of Anatomical Trees*, pages 74–85. In *Proc. of IPMI'13*, Asilomar, CA, USA. LNCS, Springer, 2013.
4. P. Fletcher, C. Lu, S. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE TMI*, 23(8):995–1005, Aug. 2004.
5. S. Huckemann and H. Ziezold. Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Advances in Applied Probability*, 38(2):299–319, June 2006.
6. N. Kwak. Principal component analysis based on l1-norm maximization. *IEEE TPAMI*, 30(9):1672–1680, Sept 2008.
7. N. Leporé, C. Brun, *et al.* Best individual template selection from deformation tensor minimization. In *Proc. of ISBI'08*, pages 460–463, France, 2008.
8. X. Pennec. Barycentric Subspaces and Affine Spans in Manifolds. In *Proc. of GSI'2015*, *LNCS 9389*, pages 12–21, Palaiseau, France, Oct. 2015.
9. X. Pennec. Barycentric Subspace Analysis on Manifolds. arXiv:1607.02833, July 2016. In revision.
10. C. Small. *The Statistical Theory of Shapes.* Springer series in statistics, 1996.
11. H. Zhai. *Principal component analysis in phylogenetic tree space.* PhD thesis, University of North Carolina at Chapel Hill, 2016.