Journal of Multivariate Analysis 153 (2017) 64–82

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Scale and curvature effects in principal geodesic analysis

Drew Lazar [a,*], Lizhen Lin [b]

[a] Department of Mathematics, Ball State University, United States

[b] Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, United States

ARTICLE INFO

Article history:
Received 31 October 2014
Available online 17 September 2016

AMS subject classifications:
60D05
62H11
53C22

Keywords:
Curvature effects
Data scaling
Diffusion tensors
Dimension reduction
Manifold-valued statistics
Principal geodesic analysis (PGA)
Symmetric spaces

ABSTRACT

There is growing interest in using the close connection between differential geometry and statistics to model smooth manifold-valued data. In particular, much work has been done recently to generalize principal component analysis (PCA), the method of dimension reduction in linear spaces, to Riemannian manifolds. One such generalization is known as principal geodesic analysis (PGA). This paper, in a novel fashion, obtains Taylor expansions in scaling parameters introduced in the domain of objective functions in PGA. It is shown this technique not only leads to better closed-form approximations of PGA but also reveals the effects that scale, curvature and the distribution of data have on solutions to PGA and on their differences to first-order tangent space approximations. This approach should be able to be applied not only to PGA but also to other generalizations of PCA and more generally to other intrinsic statistics on Riemannian manifolds.

1. Introduction

Principal component analysis (PCA) is an important statistical method for dimension reduction and exploration of the variance structure of data in a linear space. PCA has been generalized to data in smooth manifolds in various principal geodesic procedures in which projections are done to explanatory submanifolds which serve as non-linear analogues of the linear subspaces of PCA.

Principal geodesic analysis (PGA), as introduced in [9], successively identifies orthogonal explanatory directions in the tangent space at the intrinsic mean of data and then exponentiates the span of the results to form explanatory submanifolds. In [9] first-order tangent space approximations of PGA were formulated. Subsequently methods for exact computation of PGA in specific manifolds were offered as in [17,27]. Then in [29], using the derivative of the exponential map and ODEs if necessary in gradient descent algorithms, procedures to find exact solutions in a general class of manifolds were outlined.

As pointed out in [28], however, exact computation of PGA can be computationally complex and time-intensive, and thus there is interest in determining the accuracy and effectiveness of first-order approximations to PGA. This will depend on the distribution of data and its dispersion from the tangent space, the curvature and shape of the manifold in question and the interaction of these factors.

For illustration, as in [28], consider the position of the "wrist" of a moving robotic arm while its "elbow" and "body" are fixed. In Fig. 1 the motion is restricted to a two-dimensional surface. To analyze the movement of the wrist one might collect motion capture data as represented by the red dots in the figure. Formulating the surface as a Riemannian manifold

* Corresponding author.
E-mail addresses: dmlazar@bsu.edu (D. Lazar), lizhen.lin@nd.edu (L. Lin).

http://dx.doi.org/10.1016/j.jmva.2016.09.009

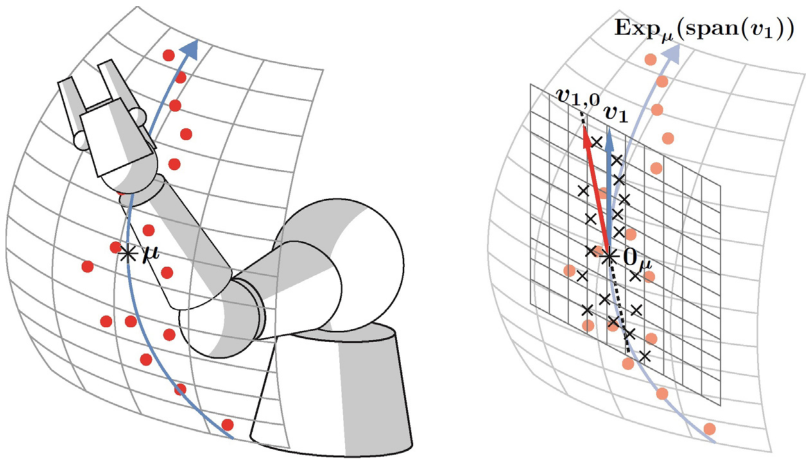

Fig. 1. First PGA motion capture data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and using intrinsic distances, an intrinsic mean of the data, μ, might be located. Then a geodesic through μ, represented by the blue curve on the surface, that best fits the data or best accounts for the data's variability might be identified.

One can find a linear direction of maximum variability, the unit vector $v_{1,0}$ in Fig. 1, of data projected by the Riemannian log map to the tangent space at the intrinsic mean. $v_{1,0}$ will be an approximation of the unit vector tangent to the geodesic v_1. Generally the greater the local curvature of the surface the less accurate this approximation will be with scale of the data or its dispersion from the intrinsic mean augmenting this effect. Conversely, projections to the tangent space will converge to the data, intrinsic distances will converge to tangent space distances and $v_{1,0}$ will converge to v_1 as the data draws in towards μ.

In this paper we quantify such effects by introducing scaling parameters on projections of data to the tangent space and by obtaining Taylor expansions of solutions to PGA procedures in these parameters. Leading terms, such as $v_{1,0}$ in Fig. 1, will originate from the Euclidean structure in the tangent space. Next-order terms will demonstrate how local curvature and scale interact to contribute to differences between first-order approximations and exact solutions. This not only allows for more accurate closed-form approximations of PGA but should also contribute to a better understanding of the parts of PGA and corresponding statistics. In this paper data in three types of symmetric spaces which have regular application are considered. Also using [17,27,29] we can compute exact solutions in these spaces which allows for comparison and testing.

1.1. Outline

Section 2 includes notations and definitions. In Section 3 a proposition which allows the expansion of PGA directions in this paper is stated and proved. In Section 4 we review the geometry of the n-spheres and obtain and test expansions using our proposition. We also carry out experiments on data sampled from an anisotropic log-normal distribution on the unit n-sphere to show improved approximations. In Section 5 we review the geometry of the space of positive definite matrices and obtain expansions using our proposition and computer algebra. In Section 6 we review the geometry of the special orthogonal group and obtain expansions of PGA in this space. Also, in Section 6.3 we take a closer look at PGA in Lie groups in [10] to show how expansions can give insight into the formulation of such intrinsic manifold statistics. In Section 7, using expansions, we obtain improvements of the *linear difference indicators* introduced in [28]. In Section 8 we discuss the results and consider their applications in similar contexts.

2. Notations and definitions

Let M be a Riemannian manifold with Riemannian metric $p \to \langle\ ,\ \rangle_p$ for $p \in M$. Given $p \in M$, T_pM is the *tangent space at* p. The unit sphere at T_pM is then $\mathrm{S}_pM = \{X \in \mathrm{T}_pM; \langle X, X\rangle_p = 1\}$. The *Riemannian exponential* and *Riemannian log maps* are denoted by $\mathrm{Exp}_p : \mathrm{T}_pM \mapsto M$ and $\mathrm{Log}_p : M \mapsto \mathrm{T}_pM$, respectively. Given smooth manifolds M_1 and M_2, $p \in M_1$ and smooth mapping $\lambda : M_1 \to M_2$ we denote the *differential of* λ *at* p by $d_p\lambda$. Then given smooth function $f : M \to \mathbb{R}$ and $p \in M$ *the gradient of* f *at* p is denoted $\nabla_p f$ so that $\langle \nabla_p f, X\rangle_p = d_pf(X)$ for all $X \in \mathrm{T}_pM$. Differential geometry texts [6,25] provide a background for and definitions of these concepts.

All the manifolds we will deal with in the paper will be of the class defined below.