Statistica Sinica **36** (2026), 1-21
doi:https://doi.org/10.5705/ss.202021.0163

PRINCIPAL SUB-MANIFOLDS

Zhigang Yao*, Benjamin Eltzner and Tung Pham

National University of Singapore, University of Goettingen and University of Melbourne

Abstract: We propose a novel method of finding principal components in multivariate data sets that lie on an embedded nonlinear Riemannian manifold within a higher-dimensional space. Our aim is to extend the geometric interpretation of PCA, while being able to capture non-geodesic modes of variation in the data. We introduce the concept of a principal sub-manifold, a manifold passing through a reference point, and at any point on the manifold extending in the direction of highest variation in the space spanned by the eigenvectors of the local tangent space PCA. Compared to recent work for the case where the sub-manifold is of dimension one (Panaretos et al., 2014)—essentially a curve lying on the manifold attempting to capture one-dimensional variation—the current setting is much more general. The principal sub-manifold is therefore an extension of the principal flow, accommodating to capture higher dimensional variation in the data. We show the principal sub-manifold yields the ball spanned by the usual principal components in Euclidean space. By means of examples, we illustrate how to find, use and interpret a principal sub-manifold and we present an application in shape analysis.

Key words and phrases: Dimension reduction, manifold, principal component analysis, shape analysis, tangent space.

1. Introduction

Many quantities of interest are best described as points in a non-Euclidean space, not as vectors in a vector space. The most well-known example are directional data represented on a circle or sphere in *directional statistics*, which has been discussed as early as Fisher (1953). Higher dimensional manifold data spaces arise in the description of shapes in terms of landmarks, e.g., by Kendall (1989). In many cases, data lie close to a low dimensional sub-manifold of the data space. Approaches to restrict consideration to such a sub-manifold broadly fall into two categories. There are approaches to represent data *explicitly* on a *known sub-manifold* embedded in the data space (Kendall et al., 1999; Patrangenaru and Ellingson, 2015). Alternative approaches represent the data on an *unknown sub-manifold* in the sense that it is not embedded in the original data space, which is determined by *manifold learning* (Roweis and Sau, 2000; Donoho and Grimes, 2003; Zhang and Zha, 2004; Guhaniyogi and Dunson, 2016; Yao, Su and Yau, 2023). In this paper, we discuss a method which provides an explicit, embedded

*Corresponding author. E-mail: zhigang.yao@nus.edu.sg