Data Beyond the Euclidean Space

Jörn Schulz

Department of Electrical Engineering and Computer Science,
University of Stavanger, Stavanger, Norway

Complex data such as non-Euclidean or a mixture of Euclidean and non-Euclidean data has gained growing attention recently. However, only few methods are available to do sensitive statistical inferences on these types of data and only little is known about their asymptotic properties. In the following, we assume that the non-Euclidean data lives on a smooth manifold and in particular we will focus on directional data, i.e. data on the hypersphere $\mathbb{S}^d = \{\mathbf{x} \in \mathbb{R}^{d+1} : \mathbf{x}^T\mathbf{x} = 1\}$ and data on polyspheres $(\mathbb{S}^2)^d$. Examples of these data types are i.) shape representations including directions such as skeletal representations that live on $\mathbb{S}^{d_1} \times (\mathbb{S}^2)^{d_2}$ (Hong et al. (2016); Pizer et al. (2013); Schulz et al. (2016)), ii.) dihedral angles of protein structures on $(\mathbb{S}^1 \times \mathbb{S}^1)^d$ (Eltzner et al. (2017)) or iii.) to analyze temporal sequences of molecules on $\mathbb{S}^d$ (Dryden et al. (2019)). Especially, in examples i.) and ii.) we have usually a high dimension low sample size setting, i.e. $d \gg n$ where n is the sample size and d is the dimension.

A crucial step in the analysis in all these applications is principal nested spheres (PNS) (Jung et al. (2012)), a method for decomposition and dimension reduction of directional data on $\mathbb{S}^d$. In opposite to principal component analysis, PNS is a backward dimension reduction method. In each step, a submanifold of successively lower dimension, containing the largest total variance, is fitted to the data. A submanifold can be either a small-sphere or a great sphere, i.e. a sphere with radius $r < \pi/2$ or $r = \pi/2$. The choice of a small or a great sphere is a critical question in the PNS procedure. The fitting of a small sphere to the data might result in an overfitting, e.g. if the data is concentrated around a point at $\mathbb{S}^d$. We will discuss a new testing procedure that outperforms alternative testing methods during a simulation study and the analysis of skeletal 3D models of hippocampi. The proposed method is based on a measure of multivariate kurtosis for directional data. Given a suitable decomposition of the data, statistical inference by hypothesis testing (Schulz et al. (2016)), classification (Hong et al. (2016)) or clustering (Dryden et al. (2019)) might be performed.

In addition, we will briefly review and discuss some recent works on asymptotic results within this framework.

References

Dryden, I. L., Kim, K.-R., Laughton, C. A., and Le, H. (2019), "Principal nested shape space analysis of molecular dynamics data," *arXiv preprint arXiv:1903.09445*.

Eltzner, B., Huckemann, S., and Mardia, K. V. (2017), "Torus Principal Component Analysis with an Application to RNA Structures," *Annals of Applied Statistics*, ISSN 1932-6157 (In Press).

Hong, J., Vicory, J., Schulz, J., Styner, M., Marron, J. S., and Pizer, S. M. (2016), "Non-Euclidean classification of medically imaged objects via s-reps," *Medical image analysis*, 18, 37–45.

Jung, S., Dryden, I. L., and Marron, J. S. (2012), "Analysis of Principal Nested Spheres," *Biometrika*, 99, 551–568.

Pizer, S. M., Jung, S., Goswami, D., Zhao, X., Chaudhuri, R., Damon, J. N., Huckemann, S., and Marron, J. S. (2013), "Nested Sphere Statistics of Skeletal Models," in *Innovations for Shape Analysis: Models and Algorithms*, eds. Breus, M., Bruckstein, A., and Maragos, P., New York: Springer, pp. 93–115.

Schulz, J., Pizer, S. M., Marron, J. S., and Godtliebsen, F. (2016), "Non-linear Hypothesis Testing of Geometric Object Properties of Shapes Applied to Hippocampi," *Journal of Mathematical Imaging and Vision*, 54, 15–34, issue 1.