A generalisation of Locally Linear Embedding to manifold-valued data.

, Elodie Maignant
, *Équipe-projet Inria EPIONE, ENS Paris Saclay*

Email : elodie.maignant@inria.fr

Mots Clés : Locally Linear Embedding, Manifold learning, Shape analysis

Biographie – I studied mathematics at the ENS Paris-Saclay before joining Xavier Pennec's ERC project G-Statistics in October 2020 at Inria for a PhD focused on manifold learning and approximations by symmetric spaces. I am also supervised by Alain Trouvé at ENS Paris Saclay.

Resumé : The study of shapes, as we would define them intuitively, involves working with data which are intrinsically non-linear. Whether we are interested in analysing conformational dynamics of a protein or we want to investigate the anatomical variability of an organ, the objects we consider carry crucial information through their shapes. Various models were introduced in the literature, depending on the application's domain, giving rise to the field known as shape analysis, at the interface of geometry and statistics. In landmark-based theories, the shape of an object is determined by a set of well-chosen coordinates points located on the object [1]. Shapes could also be described as deformation of one reference shape. In any of these representations, invariance of the shape up to some transformations leads us to work with non-Euclidean structures like differential manifolds. However, these kind of structures may have a complex geometry which make computations way more difficult, adding up to the fact that data encountered in medicine or biology are often high-dimensional. Therefore, in such domains, manifold learning and dimension reduction are key problems.

Many manifold learning methods are based on the estimation of a local linear model. This is particularly the case of LLE (Local Linear embedding [4]) or LTSA (Local tangent space approximation [6]) which determine the local tangent spaces found by local PCA (Principal Component Analysis). However, while the local linear assumption is sufficient in high data concentration, it generates many errors when the data are in low concentration or non-homogeneously distributed. It can arise for example when working with medical images of the brain which can not be acquired often, or for protein simulations which take a long time to process. For local "small data" learning, it is thus necessary to regularise the estimates. Some works have sought to generalise the global structure of the manifold by assuming spaces with constant curvature, as for example with MDS (MultiDimensional Scaling [3]). Other methods seek to directly reconstruct the data on spheres or hyperbolic spaces [5]. The current boom in of artificial intelligence also raises a new interest for approximation by non-Euclidean spaces which still remain simple enough to be explanatory and efficient.

In this talk, we introduce a new method for manifold learning, generalising the LLE algorithm [4] to manifold-valued data. The first step of the LLE method consists in computing the barycentric coordinates of each data point in the ambient space with respect to its k-nearest neighbours. If the dataset consists of real-valued vectors, it is a classical least squares regression problem. We then reconstruct the data in an embedding low-dimensional vector space by optimising the corresponding coordinates of each point so that they match the previous barycentric coordinates. Again, this is a least squares problem with explicit solution. Now suppose the data lie on some Riemannian manifold. Barycentric coordinates are still defined, but cannot be written in closed form. Therefore, the resulting minimisation problem cannot be solved explicitly, especially because the search space is non-trivial. We propose to approximate the problem by a more reasonable one, exploiting the Riemannian structure of our data space. At some point, our method requires to compute the

parallel transport of the manifold in a differentiable way. This is certainly a constraining, however crucial, condition. We implement the method for the specific case of Kendall shape spaces, for which we have computations of the parallel transport compatible with automatic differentiation [2]. At a later stage, we plan to generalise the second step of the algorithm to non-Euclidean reconstruction spaces, minimising the squared distance of each point to the barycentric subspace generated by its nearest neighbours. We thus obtain a non-linear weighted least squares criterion which can be optimised on our embedding manifold by numerical gradient descent methods. We expect our method to work better for data in low concentration, such that the local linear assumption on which the classical LLE algorithm is based does not hold anymore. We also hope for it to extend to more general manifolds apart from Kendall shape spaces.

Références

[1] Ian L Dryden and Kanti V Mardia. *Statistical shape analysis: with applications in R*, volume 995. John Wiley & Sons, 2016.

[2] Nicolas Guigui, Elodie Maignant, Alain Trouvé, and Xavier Pennec. Parallel transport on kendall shape spaces. 2021.

[3] Harold Lindman and Terry Caelli. Constant curvature riemannian scaling. *Journal of Mathematical Psychology*, 17(2):89–109, 1978.

[4] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[5] Richard C Wilson, Edwin R Hancock, Elżbieta Pekalska, and Robert PW Duin. Spherical and hyperbolic embeddings of data. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2255–2269, 2014.

[6] Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM journal on scientific computing*, 26(1):313–338, 2004.