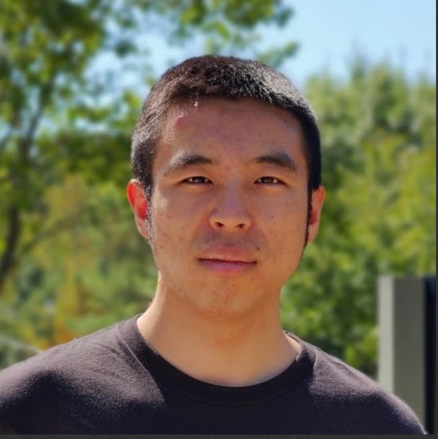
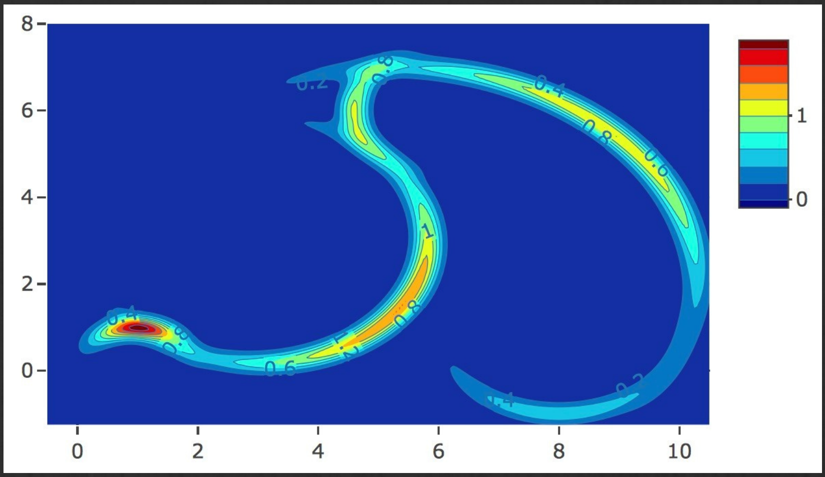


PhD Thesis

# Learning and Exploiting Low-Dimensional Structure in High-Dimensional Data

## Didong Li

Data lying in a high dimensional ambient space are commonly thought to have a much lower intrinsic dimension. In particular, the data may be concentrated near a lower-dimensional manifold. If one does not exploit the hidden geometry in the data but instead deal with the ambient high dimensional Euclidean spaces directly, both the statistical and computation efficiency are extremely low. In contrast, an accurate approximation of the unknown manifold will benefit a variety of aspects including dimension reduction, feature selection, density estimation, classification, clustering, data denoising, data visualization and so on. Most of the literature for data analysis relies on linear or locally linear methods. However, when the manifold has essential curvature, these linear methods suffer from low accuracy and efficiency. There is also an immense literature focused on non-linear methods like Variational Auto Encoders and Gaussian Process Latent Variable Model, to improve the approximation performance. However, these methods are complex black boxes lacking reproducibility, identifiability and interpretability. As a result, new non-linear tools need to be developed without introducing too much extra complexity.



My dissertation focuses on exploiting the geometry in the data through the curvature of the unknown manifold to efficiently estimate the manifold, while keeping the simple and clean close forms as in linear methods. In particular, a simple and general alternative of locally linear manifold learning method is proposed, which instead uses pieces of spheres, or spherelets, to locally approximate the unknown manifold. The spherical principal components analysis (SPCA) is developed as a non-linear alternative of PCA, to find the best sphere fitting the data. SPCA provides simple tools that can be implemented efficiency for big and complex data and allow one to learn about geometric structure in the data, without introducing much more complexity than linear methods. Inspired by spherelets, a curved kernel called the Fisher-Gaussian (FG) kernel is introduced, which outperforms multivariate Gaussians for density estimation. In particular, the Dirichlet process mixture of FG kernels model is studied for density estimation, which is proved to be posterior consistent. In addition, some applications of spherelets, including classification, geodesic distance estimation and clustering are also considered, with a variety of real data applications.

Duke MATH