Senior Thesis

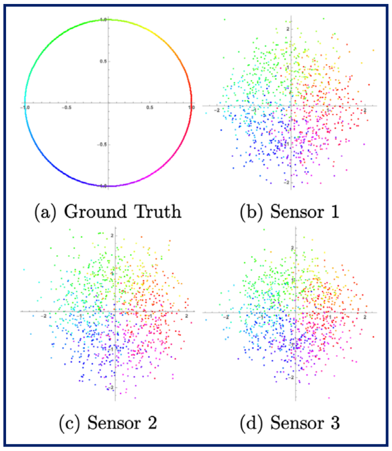
Alex Damian

advised by
Dr. Hau-Tieng Wu

Theoretical Guarantees for Signal Recovery

I am interested in understanding what type of signals various machine learning algorithms can recover from a large dataset. As a general rule, the more data points you have, the easier it is to recover a signal, and the higher the dimension of the data is, the more difficult it is to recover a signal. Although data scientists now have access to massive datasets with hundreds of thousands or even millions of data points, the data is also increasingly high dimensional. For example, the dimension of an image is equal to the number of pixels in the image which can be on the order of millions. The focus of my senior thesis was analyzing two algorithms, Principal Component Analysis (PCA) and Multi-View Diffusion Map, in both the classical setup where the number of data points is very large and the dimension is fixed and the high dimensional setup where the number of data points is comparable to the dimension of each data point.

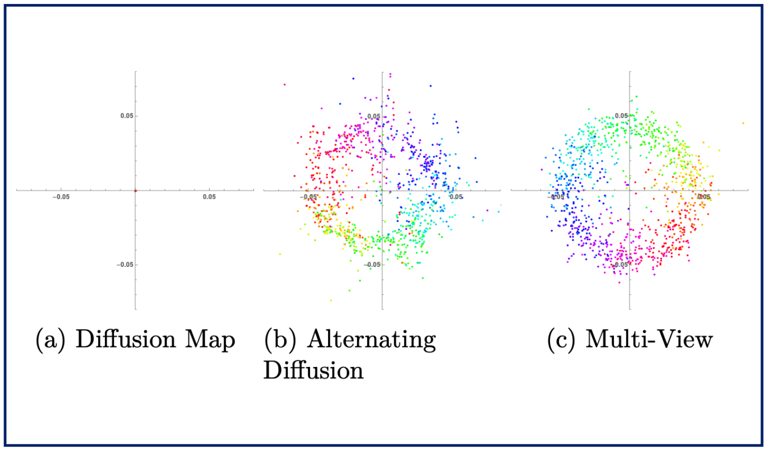
First, we studied the behavior of PCA in the "null-case" where the population covariance matrix is just the identity matrix so there is no "true" signal to recover. PCA returns a set of orthonormal vectors which can be represented as an element of the orthogonal group. We proved that when the data is Gaussian, the distribution of principal components is Haar distributed over the orthogonal group. This is not always true when the data is not Gaussian and we conjecture a necessary and sufficient fourth moment condition. Motivated by recent results showing that in the high dimensional setup, the sample principal components poorly approximate the true principal components, we proved that PCA is the optimal estimator (under the squared cosine loss) for the true principal component. This result assumes a uniform prior for the true signal. Finally, we analyzed the spectral properties of Multi-View Diffusion Map which is an algorithm for combining data from multiple sensors. We empirically demonstrate the noise reduction capabilities of Multi-View Diffusion map and provide partial theoretical justification.



(a) Ground Truth (b) Sensor 1

(c) Sensor 2 (d) Sensor 3

Input Data



(a) Diffusion Map (b) Alternating Diffusion (c) Multi-View

Reconstruction Results