Victor de Fontnouvelle
4/2/2019

Thesis Introduction

My thesis will explore various methods of analyzing high-dimensional data. I'll investigate techniques that make inferences about the underlying structure of the data, cluster the data, or reduce the dimension of the data. I'll apply these techniques to real-world datasets to gain insights that haven't been produced before. Homology analysis and the Helmholtz decomposition provide insight on the underlying structure of the data. Homology analysis can detect features such as holes and clusters. Helmholtz decomposition takes as input ranked data and outputs ordinal data that best accounts for these rankings. Mapper clusters the data, by mapping clusters of points onto intervals on the real line using a filter function, and connecting overlapping clusters. Principal component analysis reduces the dimension of the data.

Calculating the eigenvectors of the laplacian is useful both in detecting features of the data, and in reducing its dimension. The laplacian is a symmetric matrix encoding the pairwise distances between points, normalized by row. This matrix can be thought of as a linear operator encoding heat flows. Given an input of initial temperatures, it outputs the changes in temperatures after one time step. Eigenvectors corresponding to low eigenvalues thus correspond to stable temperature configurations. The eigenvectors are perpendicular, and thus eigenvectors corresponding to slightly higher eigenvalues often capture geometric structure existing in the data. Additionally, nearby points will have similar values in the eigenvectors, and thus the eigenvectors are also a useful tool for reducing the dimension of the data.