Non-Parametric Estimation of Manifolds from Noisy Data

Barak Sober
Department of Statistics and Data Science, Digital Humanities
The Hebrew University of Jerusalem
Barak.Sober@mail.huji.ac.il

A common observation in data-driven applications is that high dimensional data has a low intrinsic dimension, at least locally. In this work, we consider the problem of estimating a d dimensional sub-manifold of $\mathbb{R}^D$ from a finite set of noisy samples. Assuming that the data was sampled uniformly from a tubular neighborhood of $\mathcal{M} \in \mathcal{C}^k$, a compact manifold without boundary, we present an algorithm that takes a point r from the tubular neighborhood and outputs $\hat{p}_n \in \mathbb{R}^D$, and $\widehat{T_{\hat{p}_n}\mathcal{M}}$ an element in the Grassmanian $Gr(d, D)$. We prove that as the number of samples $n \to \infty$ the point $\hat{p}_n$ converges to $p \in \mathcal{M}$ and $\widehat{T_{\hat{p}_n}\mathcal{M}}$ converges to $T_p\mathcal{M}$ (the tangent space at that point) with high probability. Furthermore, we show that the estimation yields asymptotic rates of convergence of $n^{-\frac{k}{2k+d}}$ for the point estimation and $n^{-\frac{k-1}{2k+d}}$ for the estimation of the tangent space. These rates are known to be optimal for the case of function estimation.

Joint work with: Yariv Aizenbud.