Reconstruction of the multivariate density from a k-nearest neighbor graph

Johannes Schmidt-Hieber

schmidthieberaj@math.leidenuniv.nl

Internet users are often asked to provide quantitative information about an item, for instance to rate a movie. The answers are then used to estimate underlying parameters of this item and to decide whether it is recommended to another user. Such quantitative user data suffer, however, from many biases and often are of poor quality. Another approach is to ask a user for his ten favourite items (for instance movies or universities in an university ranking). Such data are more reliable, as they only require the user to order his preferences but never ask for quantitative data. The statistical challenge is to reconstruct the true underlying quantities from these data.

We will study the following (slightly) simplified problem. Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables from an unknown probability density function p on $\mathbb{R}^d$. What we observe is the k-nearest neighbor graph, that is a directed graph with vertices $\{1, \ldots, n\}$ and a directed edge from vertex i to vertex j if and only if X_j is among the k-nearest neighbors of X_i. The statistical problem is to reconstruct the density p only observing the k-nearest neighbor graph. What makes this problem particularly challenging is that it is a non-local problem. Indeed, in order to estimate the density at one point it might be necessary to look at the reconstruction at points that are far away.

In a first part of this project, we will study the article [1], which provides some fundamental ideas describing the geometry of the problem. The focus of the project will then be to determine the estimation accuracy that can be achieved for reconstruction of the density p depending on (n, k) and smoothness properties of p. Due to the non-local behavior of the problem, we need to think about defining appropriate function classes which capture the difficulty of estimation and are natural in a global sense. The project will include programming and running simulations (in R or Matlab).

References

[1] U. von Luxburg and M. Alamgir (2013). Density estimation from unweighted k-nearest neighbor graphs: a roadmap *NIPS*.