

Title: Estimating regions of large differences in multivariate samples with a k Nearest Neighbours method

Inge Koch

School of Mathematics and Statistics
University of New South Wales, Sydney, Australia

Abstract:

We propose an extension of the k nearest neighbour estimator in density estimation to regions of density differences in a multivariate setting. Rather than estimating the density difference everywhere our method focuses on regions where the difference is large.

The density in high density regions is estimated separately for each of the two multivariate samples: the regional estimates (in regions where the density is high) combine a preprocessing step which partitions the data into disjoint clusters with a knn estimator which is applied separately to each cluster and uses different tuning parameters for each cluster. The results are combined to yield a density estimate in high density regions. Differences of the two regional density estimates will allow the determination of a region where the two samples differ mostly.

The method is applied to multivariate samples of HIV+ and HIV- data and is able to find regions where the two samples are distinct.