High-dimensional two-sample tests based on optimal transport and related ideas

Anil Kumar Ghosh
Indian Statistical Institute, Kolkata Centre

Abstract

In recent years, the idea of optimal transport has become popular in statistics, mainly for constructing distribution-free tests for high-dimensional data. The idea is to transport the observations to a known reference distribution and construct tests based on the transported data. Under suitable regularity conditions, these resulting tests usually have large sample consistency in finite dimensions. But they often fail to yield satisfactory performance for high-dimensional data, especially when the dimension is comparable to or larger than the sample size. In this article, we first investigate the high-dimensional behaviour of some two-sample tests based on optimal transport and show that a judicious choice of reference distribution and transportation cost may lead to a better performance in high dimensions. Our theoretical investigation also leads to the construction of a class of distribution-free two-sample tests based on the idea of minimum cost derangement. Interestingly, some popular distribution-free two-sample tests belong to this class. Several simulated and benchmark data sets are analysed to study the empirical performance of our proposed test in high-dimension, low-sample-size situations.

(Joinr work with Vaibhab Sherkar, Abhradipta Ghosh and Bilol Banerjee)