Mathematics, Statistics, and Computer Science **@ UIC**

Statistics and Data Science Seminar

A New Framework for Distance and Kernel-based Metrics in High Dimensions

Xianyang Zhang (Texas A&M University)

Abstract: We present new metrics to quantify and test for (i) the equality of distributions and (ii) the independence between two high-dimensional random vectors. We show that the energy distance based on the usual Euclidean distance cannot completely characterize the homogeneity of two high-dimensional distributions in the sense that it only detects the equality of means and the traces of covariance matrices in the high-dimensional setup. We propose a new class of metrics which inherit the desirable properties of the energy distance/distance covariance in the low-dimensional setting and is capable of detecting the homogeneity of/ completely characterizing independence between the low-dimensional marginal distributions in the high dimensional setup. We further propose t-tests based on the new metrics to perform high-dimensional two-sample testing/ independence testing and study its asymptotic behavior under both high dimension low sample size (HDLSS) and high dimension medium sample size (HDMSS) setups. The computational complexity of the t-tests only grows linearly with the dimension and thus is scalable to very high dimensional data. We demonstrate the superior power behavior of the proposed tests for homogeneity of distributions and independence via both simulated and real datasets.

Wednesday, October 23 at 4:00 PM in 636 SEO