Randomized incomplete U-statistics in high dimensions

Xiaohui Chen

University of Illinois at Urbana
E-mail: xiaohui@mit.edu

Abstract: This paper studies inference for the mean vector of a high-dimensional U-statistic. In the era of Big Data, the dimension d of the U-statistic and the sample size n of the observations tend to be both large, and the computation of the U-statistic is prohibitively demanding. Data-dependent inferential procedures such as the empirical bootstrap for U-statistics is even more computationally expensive. To overcome such computational bottleneck, incomplete U-statistics obtained by sampling fewer terms of the U-statistic are attractive alternatives. In this paper, we introduce randomized incomplete U-statistics with sparse weights whose computational cost can be made independent of the order of the U-statistic. We derive non-asymptotic Gaussian approximation error bounds for the randomized incomplete U-statistics in high dimensions, namely in cases where the dimension d is possibly much larger than the sample size n, for both non-degenerate and degenerate kernels. In addition, we propose novel and generic bootstrap methods for the incomplete U-statistics that are computationally much less-demanding than existing bootstrap methods, and establish finite sample validity of the proposed bootstrap methods. The proposed bootstrap methods are illustrated on the application to nonparametric testing for the pairwise independence of a high-dimensional random vector under weaker assumptions than those appearing in the literature.