Study of U-statistics

Shunan Yao

U-statistic is popular tool in statistics when one seeks a permutation invariant estimation. For example, let $X_1, X_2, \dots, X_n$ be iid random variables with finite second moment. One way to write the variance of X_1 is

$$\operatorname{var}(X_1) = \frac{1}{2}\mathbb{E}\left(X_i - X_j\right)^2 \ ,$$

for $i \neq j$. Therefore, following this construction, the estimation of variance can be formulated as

$$\frac{1}{\lfloor n/2 \rfloor}\sum_{j=1}^{\lfloor n/2 \rfloor}(X_{2(j-1)+1} - X_{2j})^2 = \frac{1}{\lfloor n/2 \rfloor}\left((X_1 - X_2)^2 + (X_3 - X_4)^2 + \ldots + (X_K - X_{K+1})^2\right) \ , \tag{1}$$

where K is the largest odd number such that $K + 1 \leq n$. It is easy to see that the exact result in (1) can be affected by how the random variables are indexed. For example, if we interchange the index of X_1 and X_3, equation (1) may yield a different value. This permutation variant phenomenon is often undesirable since two users may come to different conclusions when analyzing the same data just because of random shuffling of data. In view of this, U-statistic is proposed to "stablize" the estimation result. Besides, U-statistic enjoys many nice theoretical properties beyond permutation invariance. We are going to explore the details of U-statistic in this project.