Kernel-based Goodness-of-fit Test for Data with Boundaries

Rizky Reza Fauzi, Graduate School of Mathematics Kyushu University
Maesono Yoshihiko, Faculty of Science and Engineering Chuo University

Continuous goodness-of-fit (GOF) is a classical hypothesis testing problem in Statistics. Despite of numerous suggestions, the Kolmogorov-Smirnov (KS) test is, by far, the most popular GOF test used in practice. Unfortunately, the lacks of smoothness can lead to smaller power at the tails, which is important in many practical applications. It is natural if one uses the naive kernel distribution function estimator in place of the empirical distribution function. Thus, instead of using the standard KS statistic

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F_X(x)|, \tag{1}$$

where $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$, to test whether a random variable X having F_X as its distribution, we can do a reformulation by smoothing it to

$$\widehat{D} = \sup_{-\infty<x<\infty} |\widehat{F}_X(x) - F_X(x)|, \tag{2}$$

where $\widehat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x-X_i}{h}\right)$ and $W(x) = \int_{-\infty}^{x} K(y)\mathrm{d}y$, the naive kernel distribution function estimator.

However, a new problem is raising when the support of the random variable we are dealing with is not the entire real line, i.e. boundary problem. Since the naive kernel distribution function estimator puts some weight outside the support, the value $|\widehat{F}_X(x) - F_X(x)|$ is larger than it is supposed to be when x is in the boundary region. This situation can lead to a rejection of the null hypothesis and lowering the power of the test near the boundary.

To solve this problem, we propose a new kernel based estimator for distribution function by transforming the data. The idea is by utilising a function g which bijectively transform the support A of the random variable under consideration into $\mathbb{R}$, and then doing the usual standard kernel distribution function estimation of $Y = g(X)$, instead of for the X itself. Hence, our proposed estimator is

$$\widetilde{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} W\left[\frac{g(x) - g(X_i)}{h}\right], \quad x \in A, \tag{3}$$

where $h > 0$ is a bandwidth. Therefore, we define the boundary-free smoothed KS type test as

$$\widetilde{D} = \sup_{-\infty<x<\infty} |\widetilde{F}_X(x) - F_X(x)|. \tag{4}$$

References

[1] M. Omelka, I. Gijbels, and N. Veraverbeke, Improved kernel estimation of copulas: weak confergence and goodness-of-it testing. *Annals of Statistics* Vol. 37 (2009) 3023-3058.