Kolmogorov-Smirnov Test Based on Kernel Estimation

Rizky Reza Fauzi, Graduate School of Mathematics, Kyushu University
Maesono Yoshihiko, Faculty of Mathematics, Kyushu University

1 Boundary-free kernel distribution function estimators

Let $X_1, X_2, ..., X_n$ be independently and identically distributed random variables with an absolutely continuous distribution function F_X and a density f_X. The classical nonparametric estimator of F_X has been the empirical distribution function defined by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x), \quad x \in \mathbb{R}, \tag{1}$$

where $I(A)$ denotes the indicator function of a set A. It is obvious that F_n is a step function of height $\frac{1}{n}$ at each observed sample point X_i. When considered as a pointwise estimator of F_X, $F_n(x)$ is an unbiased and strongly consistent estimator of $F_X(x)$. However, given the information that F_X is absolutely continuous, it seems to be more appropriate to use a smooth and continuous estimator of F_X rather than the empirical distribution function F_n.

Parzen (1962) and Rosenblatt (1956) introduced kernel density estimator as a smooth and continuous estimator for density function. It is defined as

$$\widehat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}, \tag{2}$$

where K is a function called as kernel and $h > 0$ is called as bandwidth, which is a smoothing parameter and controls the smoothness of $\widehat{f}_X$. It is usually assumed that K is a symmetric (about 0) continuous nonnegative function with $\int_{-\infty}^{\infty} K(v)\mathrm{d}v = 1$, as well as $h \to 0$ and $nh \to \infty$ when $n \to \infty$. Since distribution function is actually an integral of density function, this kernel density estimator gave an idea to define a kernel distribution function estimator. Nadaraya (1964) defined it as

$$\widehat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}, \tag{3}$$

where $W(v) = \int_{-\infty}^{v} K(w)\mathrm{d}w$. It is easy to prove that this kernel distribution function estimator is continuous, and satisfies all the properties of a distribution function. Moreover, several authors showed that the asymptotic performance of $\widehat{F}_X(x)$ is better than that of $F_n(x)$, see Azzalini (1981), Reiss (1981), Falk (1983), Singh *et al.* (1983), Hill (1985), Swanepoel (1988), Shirahata and Chu (1992), and Abdous (1993).

Under the condition that f_X (the density) has one continuous derivative f'_X, it has been proved by the above-mentioned authors that, as $n \to \infty$,

$$Bias[\widehat{F}_X(x)] = \frac{h^2}{2} f'_X(x) \int_{-\infty}^{\infty} v^2 K(v) \mathrm{d}v + o(h^2), \tag{4}$$

$$Var[\widehat{F}_X(x)] = \frac{1}{n} F_X(x)[1 - F_X(x)] - \frac{2h}{n} r_1 f_X(x) + o\left(\frac{h}{n}\right) \tag{5}$$

where $r_1 = \int_{-\infty}^{\infty} vK(v)W(v)\mathrm{d}v$. It is easy to show that r_1 is a nonnegative number.

However, all of the previous explanations implicitly assume that the true density is supported on the entire real line. If we deal with $\mathbb{R}^+$ or unit interval for instance, the standard kernel distribution function estimator will suffer the so called boundary bias problem. This is because the estimator does not 'feel' the boundary, and puts some weights for the lack of data on the axis of zero probability.

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 $\mathbb{R}$, then doing the usual standard kernel distribution function estimation of $Y = g(X)$, instead of for the X itself. However, since the variable being analised is X, we should apply back transformation to find the estimates of $F_X(x)$. Hence, our proposed estimator is

$$\widetilde{F}_X(x) = \frac{1}{n} \sum_{i=1}^{n} W\left[\frac{g(x) - g(X_i)}{h}\right], \quad x \in A, \tag{6}$$

where $h > 0$ is a bandwidth. It is obvious that no weight will be applied outside the support A and we can assign the value of $\widetilde{F}_X$ equals to 0 for $x = \inf A$ and equals to 1 when $x = \sup A$, without abusing the properties of distribution function. No boundary bias problem involves in this setting. Another advantage of this proposed estimator is its bias and variance being still in the order of h^2 and n^{-1}, respectively, just as the standard one. They are

$$Bias[\widetilde{F}_X(x)] = \frac{h^2}{2} c(x) \int_{-\infty}^{\infty} v^2 K(v) \mathrm{d}v + o(h^2) \tag{7}$$

$$Var[\widetilde{F}_X(x)] = \frac{1}{n} F_X(x)[1 - F_X(x)] - \frac{2h}{n} \frac{f_X(x)}{g'(x)} r_1 + o\left(\frac{h}{n}\right), \tag{8}$$

where $r_1 = \int_{-\infty}^{\infty} vK(v)W(v)\mathrm{d}v$ and

$$c(x) = \frac{f'_X(x)}{[g'(x)]^2} - \frac{f_X(x) g''(x)}{[g'(x)]^3} \tag{9}$$

For example, if the support is $(0, \infty)$, one of the simplest function that transform it to entire real line bijectively is the logarithmic function. Doing so, our proposed idea for kernel distribution function estimator is

$$\widetilde{F}_X(x) = \frac{1}{n} \sum_{i=1}^{n} W\left(\frac{\log x - \log X_i}{h}\right), \quad x \in \mathbb{R}^+ \tag{10}$$

with the bias and variance of $\widetilde{F}_X$ are

$$Bias[\widetilde{F}_X(x)] = \frac{h^2}{2}[xf_X(x) + x^2 f_X'(x)] \int_{-\infty}^{\infty} v^2 K(v) \mathrm{d}v + o(h^2), \tag{11}$$

and

$$Var[\widetilde{F}_X(x)] = \frac{1}{n} F_X(x)[1 - F_X(x)] - \frac{2h}{n} r_1 x f_X(x) + o\left(\frac{h}{n}\right), \tag{12}$$

where $r_1 = \int_{-\infty}^{\infty} vK(v)W(v)\mathrm{d}v$.

Same goes when the support of the data is the unit interval, by utilising the transformation $Y = \Phi^{-1}(X)$, where Φ is the standard normal distribution function. The estimator will be

$$\widetilde{F}_X(x) = \frac{1}{n} \sum_{i=1}^{n} W\left[\frac{\Phi^{-1}(x) - \Phi^{-1}(X_i)}{h}\right], \quad x \in [0, 1] \tag{13}$$

with the bias

$$Bias[\widetilde{F}_X(x)] = \frac{h^2}{2} f_Y'[\Phi^{-1}(x)] \int_{-\infty}^{\infty} v^2 K(v) \mathrm{d}v + o(h^2) \tag{14}$$

and the variance

$$Var[\widetilde{F}_X(x)] = \frac{1}{n} F_X(x)[1 - F_X(x)] - \frac{2h}{n} r_1 f_Y[\Phi^{-1}(x)] + o\left(\frac{h}{n}\right), \tag{15}$$

where

$$f_Y[\Phi^{-1}(x)] = \phi[\Phi^{-1}(x)] f_X(x),$$

and

$$f_Y'[\Phi^{-1}(x)] = \phi'[\Phi^{-1}(x)] f_X(x) + \phi^2[\Phi^{-1}(x)] f_X'(x),$$

with ϕ is the standard normal density function.

2 Boundary-free smoothed Kolmogorov-Smirnov type test

Continuous goodness-of-fit (GOF) is a classical hypothesis testing problem in statistics. Despite numerous suggestions, the Kolmogorov-Smirnov (KS) test is, by far, the most popular GOF test used in practice. Unfortunately, it lacks of smoothness that 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 the standard KS statistic

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - F_X(x)| \tag{16}$$

being used to test whether random variable X having F_X as its distribution, we can reformulate by smoothing it to

$$\widehat{D} = \sup_{-\infty<x<\infty} |\widehat{F}_X(x) - F_X(x)|, \tag{17}$$

where $\widehat{F}$ is 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. As usual, 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.

For some illustrations, we provide the results of a numerical simulation of naive kernel distribution function estimator, and compare them with the theoretical distribution function. We generated 20 observations from two distributions, $exp(2)$ and $U(0,1)$. As we can see at

Figure 1: naive kernel DF estimator(Fh) vs $exp(2)$ distribution function(F)

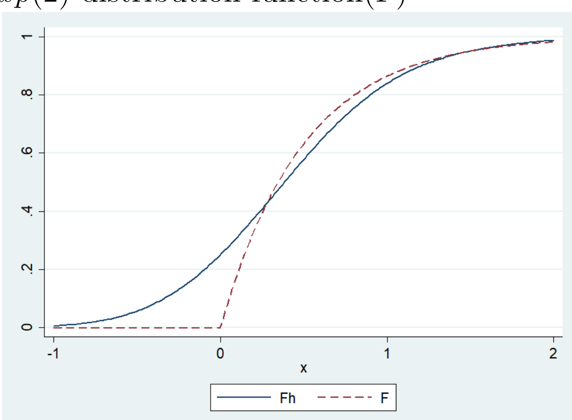
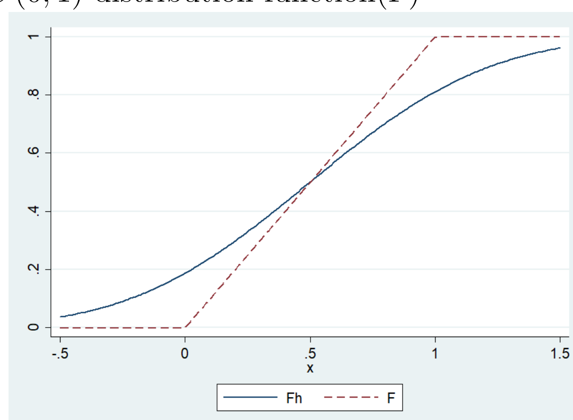


Figure 2: naive kernel DF estimator(Fh) vs $U(0,1)$ distribution function(F)



both figures, the gap between $\widehat{F}_X$ and F_X is going larger near the boundary, and this can lead to wrong conclusion of the test. Even, the estimated graphs fairly resemble normal distribution, which means if the null hypothesis is the data being normally distributed, H_0 may not be rejected. This situation can enlarge the probability of type 2 error.

To overcome this problem, we propose to use our estimator in section 1 to substitute empirical distribution function in standard KS statistic. Therefore, we define the boundary-free smoothed KS type test as

$$\widetilde{D} = \sup_{-\infty<x<\infty} |\widetilde{F}_X(x) - F_X(x)|, \tag{18}$$

where $\widetilde{F}_X$ is our boundary-free kernel distribution function estimator. Better result can be seen when the same data sets for Figure 1 and 2 are used in our proposed formula $\widetilde{F}_X$.

We also did a second numerical study by calculating simulated power of our proposed test with $n = 50$, and then we compared it with the result of the standard KS test.

Figure 3: proposed kernel DF estimator(Ft) vs $exp(2)$ distribution function(F)

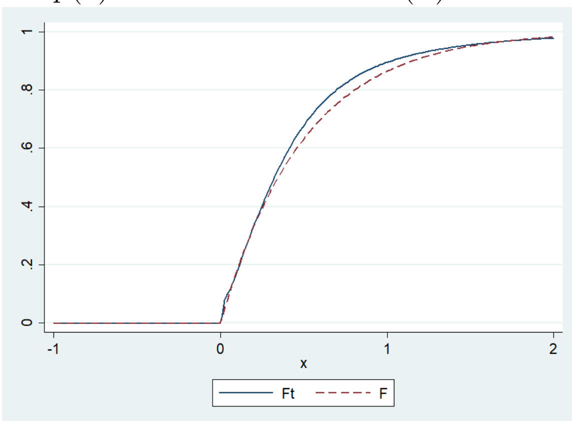
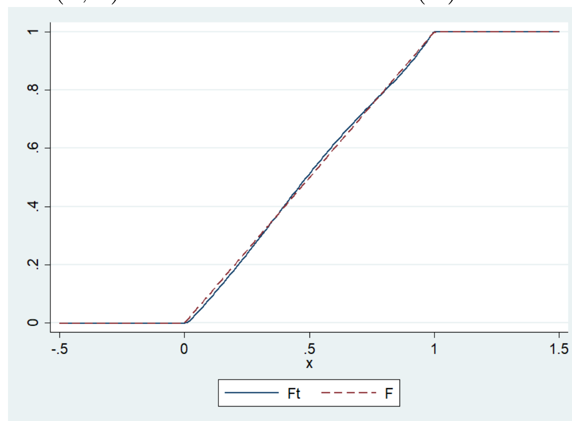


Figure 4: proposed kernel DF estimator(Ft) vs $U(0,1)$ distribution function(F)



Probability rejecting H_0, proposed

Real \ H_0	$exp(1/2)$	$Gamma(3,2)$	$abs.N(0,1)$	$log.N(0,1)$
$exp(1/2)$	**0.050**	0.934	0.957	0.976
$Gamma(3,2)$	0.834	**0.051**	0.872	0.836
$abs.N(0,1)$	0.951	0.936	**0.050**	0.981
$log.N(0,1)$	0.871	0.829	0.895	**0.050**

Probability rejecting H_0, KS test

Real \ H_0	$exp(1/2)$	$Gamma(3,2)$	$abs.N(0,1)$	$log.N(0,1)$
$exp(1/2)$	**0.051**	0.746	0.855	0.724
$Gamma(3,2)$	0.887	**0.050**	0.851	0.834
$abs.N(0,1)$	0.784	0.748	**0.051**	0.878
$log.N(0,1)$	0.862	0.830	0.891	**0.052**

The asymptotic behaviours of our proposed test statistic are stated in the following theorems.

Theorem 1. Let X be a random variable with distribution function F_X supported on a set A. If $\widetilde{F}_X$ is the proposed boundary-free kernel distribution function estimator, then

$$\sup_{-\infty<x<\infty} |\widetilde{F}_X(x) - F_X(x)| = o_p\left(\frac{1}{\sqrt{n}}\right) \tag{19}$$

Theorem 2. Let X be a random variable with distribution function F_X supported on a set A. If $\widetilde{D}$ is the proposed boundary-free smoothed KS-type statistic, then

$$\lim_{n\to\infty} \Pr(\sqrt{n}\widetilde{D} \leq x) = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty} \exp\left[\frac{-(2i-1)^2\pi^2}{8x^2}\right]. \tag{20}$$

References

[1] B. Abdous, Note on the minimum mean integrated squared error of kernel estimates of a distribution function and its derivatives. *Comm. Statist. Theory Methods* Vol. 22 (1993) 603-609.

[2] N. Altman and C. Léger, Bandwidth selection for kernel distribution function estimation. *J. Statist. Plann. Inf.* Vol. 46 (1995) 195-214.

[3] A. Azzalini, A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika* Vol. 68 (1981) 326-328.

[4] M. Falk, Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. *Statist. Neerl.* Vol. 37 (1983) 73-83.

[5] P. D. Hill, Kernel estimation of a distribution function. *Comm. Statist. Theory Methods* Vol. 14 (1985) 605-620.

[6] A. Kolmogorov, Sulla determinazione empirica di una legge di distribuzione. *Inst. Ital. Attuari, Giorn.* Vol. 4 (1933) 1-11.

[7] E. A. Nadaraya, Some new estimates for distribution functions. *Theory Probability and Applications* Vol. 15 (1964) 497-500.

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

[9] E. Parzen, On estimation of a probability density function and mode. *The Annals of Mathematical Statistics* Vol. 32 (1962) 1065-1076.

[10] R. D. Reiss, Nonparametric estimation of smooth distribution functions. *Scand. J. Statist.* Vol. 8 (1981) 116-119.

[11] M. Rosenblatt, Remarks on some non-prametric estimates of a density function. *The Annals of Mathematical Statistics* Vol. 27 (1956) 832-837.

[12] S. Shirahata and I. S. Chu, Integrated squared error of kernel type estimator of distribution function. *Ann. Inst. Statist. Math.* Vol. 44 (1992) 579-591.

[13] R. S. Singh, T. Gasser, and B. Prasad, Nonparametric estimates of distribution functions. *Comm. Statist. Theory Methods* Vol. 12 (1983) 2095-2108.

[14] J. W. H. Swanepoel, Mean integrated squared error properties and optimal kernels when estimating a distribution function. *Comm. Statist. Theory Methods* Vol. 17 (1988) 3785-3799.

[15] B. B. Winter, Convergence rate of perturbed empirical distribution functions. *J. Appl. Probab.* Vol. 16 (1979) 163-173.

[16] H. Yamato, Uniform convergence of an estimator of a distribution function. *Bulletin of Mathematical Statistics* Vol. 15 (1973) 69-78.