Smoothing Goodness-of-Fit tests based on Kullback-Leibler Information

Han Yu[*] Kai-Sheng Song [†]

Abstract

We present asymptotically distribution-free goodness-of-fit tests based on smoothing techniques. The proposed tests is a nonparametric extension of the classical Neyman-Pearson log-likelihood ratio test. The tests are indicated to have much greater power for detecting high-frequency nonparametric alternatives than the existing classical tests such as Kolmogorov-Smirnov tests. This good performance of the proposed tests is demonstrated by Monte Carlo simulations.

Key Words: Goodness-of-Fit, smoothing, nonparametric, simulation.

1. Introduction

In the hypothesis testing setting, Song (2002) presents a general methodology for developing asymptotically distribution-free goodness-of-fit tests based on the Kullback-Leibler discrimination information. The tests are shown to be omnibus within an extremely large class of nonparametric global alternatives and to have good local power; The test procedure is a nonparametric extension of the classical Neyman-Pearson likelihood ratio test based on the mth-order spacings between order statistics cross-validated by the observed log likelihood. It can also be viewed as a procedure based on sum-log functionals of nonparametric density-quantile estimators crossed-validated by the log-likelihood. With its good power properties, the method provides an extremely simple and potentially much better alternative to the traditional empirical CDF-based test procedures.

Consider the following goodness-of-fit test problem:

$$H_0 : f(x) = f_0(x, \theta), \text{ for some } \theta \in \Theta$$

where the parameter vector θ is specified or unspecified. To test H_0, we consider the Kullback-Leibler discrimination information between two distribution functions given by

$$\begin{aligned} I(F, F_0; \theta) &= \int_{-\infty}^{\infty} f(x) \log(f(x)/f_0(x, \theta))dx \\ &= -H(F) - \int_{-\infty}^{\infty} \log f_0(x, \theta) dF(x) \end{aligned}$$

where

$$H(F) := -\int_{-\infty}^{\infty} \log f(x) dF(x)$$

is the entropy of F. The entropy estimator is given based on the mth-order spacings between order statistics:

$$H_{mn} := n^{-1} \sum_{i=1}^{n} \log \frac{n}{2m}(X_{(i+m)} - X_{(i-m)}).$$

[*] Department of Mathematics, Computer Science and Information Systems, Northwest Missouri State University

[†] Department of Mathematics, University of North Texas

Here, the window width m is a positive integer smaller than $n/2$. A test statistic of goodness-of-fit is proposed

$$I_{mn} = -H_{mn} - \frac{1}{n}\sum_{i=1}^{n} \log f_0(X_i, \hat{\theta}_n).$$

Since large values of $I(F, F_0; \theta)$ favor the alternative hypothesis to H_0 and I_{mn} is the sample estimate of $I(F, F_0; \theta)$, we reject H_0 if I_{mn} is large.

Note that the calculation of the test statistic I_{mn} involves the density function f_0 which is readily available in an explicit form in almost all commonly encountered cases. This is in contrast with procedures like empirical CDF-based tests requiring the evaluation of the cumulative distribution which may not have a closed form such as multiparameter beta and gamma distributions.

Let's consider the standardized test statistic:

$$S_{mn} := (6mn)^{1/2}(I_{mn} - log(2m) - \gamma + R_{2m-1})$$

where

$$R_m := \sum_{j=1}^{m} 1/j$$

and $\gamma := \lim_{n\to\infty}(R_n - \log n)$ is the Euler constant. Under H_0 and certain mild conditions, we have

$$S_{mn} \overset{\mathscr{D}}{\to} N(0,1), \quad as\ n \to \infty$$

The asymptotic theory suggests that m should be chosen adaptively according to the sample size. For example, any m ranging from $c(\log n)^{1+\delta}$ to $cn^{1/3}/(\log n)^{2/3+2\delta}$ for some constants $c > 0$ and $\delta > 0$ would ensure the distribution property and consistency of the test. In practice, of course, a general guide for the choice of m for a fixed and finite n would be valuable to the users since for each finite n the distribution of the test statistic is dependent on the choice of m. Data-driven method of choosing m:

$$\hat{m} : \min\left\{ m^* : m^* = \underset{m}{argmax}\{H_{mn} : H_{mn} \le -\frac{1}{n}\sum_{i=1}^{n} \log f_0(X_i, \hat{\theta}_n)\} \right\}$$

i.e., $\hat{m}$ is defined to be the smallest $\hat{m}$ that maximizes the sample entropy H_{mn} constrained by the observed log likelihood.

However, there are some limitations to this test: finding the optimal choice (say, in terms of power) of m is clearly a difficult problem; The test S_{mn} using the mth-order spacing between the order statistics can be viewed as the mth nearest neighbor method of smoothing

$$\hat{f}(x) = \frac{1}{nd_m(x)}\sum_{i=1}^{n} K\left(\frac{x - X_i}{d_m(x)}\right).$$

where for each x,

$$d_1(x) \le d_2(x) \le \cdots \le d_n(x)$$

are the distances, arranged in ascending order, from x to the points of the sample. In the tails of the distribution, the distance $d_k(x)$ will be larger than in the main part of the distribution, resulting in causing large bias due to oversmoothing in the tails.

Based on the test statistic proposed by Song(2002), we propose a new scheme to improve it. Due to the fact that the mth nearest neighbor method in effect can be viewed as a variable triangle kernel, its derivative of the triangular kernel will be symmetric boxes. Geometrically, we naturally extend the symmetric boxes to symmetric smoothing curve, which is the derivative of the smoothing kernel. Then we propose the general kernel smoothing method to be our smoothing strategy.

$$H_{mn} := n^{-1} \sum_{i=1}^{n} \log \left(\sum_{\underline{m}_i < j \leq \bar{m}_i} \omega_{ijn} X_{(j)} \right).$$

where $\omega_{ijn} {:=} \frac{1}{h^2} \int_{\frac{j-1}{n}}^{\frac{j}{n}} k(\frac{\frac{1}{n}-y}{h}) dy$, $\underline{m}_i {:=} \lfloor i - nh \rfloor$, $\bar{m}_i {:=} \lceil i + nh \rceil$. The kernel smoothing strategy will provide more flexibility and overcome the drawbacks of the mth nearest neighbor method. With the kernel smoothing methodology, the selection of the smoothing parameter h can be made much easier than that of the smoothing parameter m in the nearest neighbor method.

2. Simulation

In this section we explore practical performance of our testing procedures via Monte Carlo simulation studies. We focus mainly on investigating the error level and power of the proposed tests, obtained via calculating the number of times the null hypothesis was rejected among the number of simulations carried out. If the null hypothesis were true, this proportion should be small, and if the null hypothesis were false this proportion should be close to one.

We study the results of Monte Carlo simulations based on samples from a standard uniform distribution for sample sizes from $n = 800$ to 1700 by 50 with repetitions of 50000. We choose $h_n \asymp n^{-2/3} \log^{-4/3} n \log\log^{-2} n$ suggested by our asymptotic results. Tests of nominal level 0.05 are considered. The results of the level study are given in Table 1. The table show the proposed nonparametric test held its level reasonably well.

Table 1: Table 1: Level Values for The Test under H_0

sample size	n=800	n=850	n=900	n=950	n=1000
level	0.08	0.0642	0.0792	0.0508	0.0562
sample size	n=1050	n=1100	n=1150	n=1200	n=1250
level	0.049	0.0546	0.0692	0.0574	0.0588
sample size	n=1300	n=1350	n=1400	n=1450	n=1500
level	0.0582	0.0482	0.0474	0.0414	0.0392
sample size	n=1550	n=1600	n=1650	n=1700	
level	0.042	0.0322	0.0422	0.0476	

To investigate the power of the proposed tests, power comparison of the Kolmogorov-Smirnov tests and the proposed nonparametric tests are made for the alternative (1) in the

Table 2. The procedure was conducted as follows: generate random samples of different sizes from $n = 800$ to 2000 by 50 from the collection of probability densities of the form

$$f_k(t) = f_0(t) + \rho_n sin2k\pi t \tag{1}$$

where $\rho_n \asymp n^{-\frac{1}{4}}$ and $k \asymp n^{\frac{1}{4}}$ by rejection method. We choose $h_n \asymp n^{-2/3} \log^{-4/3} n \log\log^{-2} n$ for the proposed tests suggested by our asymptotic results. The significant level α equals 0.05. All Monte Carlo experiments were replicated 50000 times.

The simulations show that the powers of the Kolmogorov-Smirnov tests do not exceed 10% even when the sample size is large enough while our proposed nonparametric tests have power around 90%. As would be expected, the proposed nonparametric tests perform more powerful in comparison to the Kolmogorov-Smirnov tests especially for alternatives containing the high frequency data components.

Table 2: Table 2: Power Values for The Test under H_1

sample size	n=800	n=850	n=900	n=950	n=1000
power	0.916	0.884	0.952	0.884	0.888
KS power	0.096	0.116	0.072	0.1	0.088
sample size	n=1050	n=1100	n=1150	n=1200	n=1250
power	0.912	0.9	0.912	0.928	0.892
KS power	0.072	0.068	0.08	0.096	0.1
sample size	n=1300	n=1350	n=1400	n=1450	n=1500
power	0.908	0.884	0.912	0.856	0.856
KS power	0.08	0.084	0.104	0.096	0.084
sample size	n=1550	n=1600	n=1650	n=1700	n=1750
power	0.876	0.856	0.864	0.884	0.892
KS power	0.084	0.088	0.112	0.108	0.1
sample size	n=1800	n=1850	n=1900	n=1950	n=2000
power	0.912	0.896	0.916	0.876	0.896
KS power	0.068	0.072	0.092	0.088	0.08

In summary, the proposed nonparametric tests did a reasonable job of holding their levels. In terms of power, based on the sample size, the selected bandwidth suggested by our asymptotic results is the best choice to distinguish the nonparametric alternatives from the null.

REFERENCES

Ingster, Yu. I. (1993), "Asymptotically minimax hypothesis testing for nonparametric alterantives I, II, III", Math. Methods Statist., 2:85-114, 171-189, 249-268.

Ingster, Yu. I. and Suslina Irina A. (2002), *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, Springer.

Silverman, Bernard. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC.

Song, Kai-Sheng (2002), Goodness-of-Fit tests based on kullback-leiber discrimination information. IEEE Transactions on Information Theory, 48(5):357-361.