Kernel-based Kullback-Leibler Divergence on Nonparametric Density Alternatives

Han Yu*

Abstract

A kernel-based Kullback-Leibler divergence is proposed. The proposed Kullback-Leibler divergence are used for tests on nonparametric density alternatives that are developed to be asymptotically distribution-free. The procedure can be viewed as a nonparametric extension of the traditional parametric likelihood ratio tests. Simulations of the proposed tests are provided for the small sample size performance.

Key Words: Goodness-of-Fit, Kullback-Leibler, kernel smoothing, nonparametric, density estimator.

1. Introduction

It is a very common statistical practice to check whether a family of well established parametric models fit a particular data set adequately to reduce the risk of misspecification since it is more and more popular to have allowed flexible and refined models. The traditional hypothesis testing is to use a large family of parametric models with the implicit assumption that the large family of parametric models specifies the form of the true underlying models correctly. Though parametric models provide great power of interpretation and ease of computation due to their parsimonious description, the parametric alternatives is not large enough to contain the nonparametric models and the erroneous conclusion of testing can still be drawn (see, e.g., the example in Fan et al. (2007), Ingster(1993, 2002) and Silverman (1986). It is the drawbacks of parametric alternatives and the increasing popularity of flexible nonparametric models that make nonparametric models as an attractive alternative hypothesis.

The question arises naturally that the traditional maximum likelihood ratio test is not applicable to the problems with nonparametric models as alternatives in general. The nonparametric maximum likelihood estimate may usually not exist in a density function space specifying the nonparametric density alternatives and hence the nonparametric maximum likelihood ratio tests are not applicable in general. Some likelihood ratio test procedures that are distribution-free under parametric alternatives may become dependent on nuisance parameters under nonparametric alternatives since infinite dimensional neighborhood is around a null hypothesis. To attenuate these difficulties, we propose a kernel-based Kullback-Leibler divergence for nonparametric density alternatives and investigate its performance in finite sample size. We replace the maximum likelihood estimate under the nonparametric density alternatives by a reasonable nonparametric estimate to construct our nonparametric likelihood ratio test statistics.

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

2. Kernel-based Kullback-Leibler Divergence

Consider the Kullback-Leibler divengence 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 the density $f(x)$. Song (2002) presents a general methodology for developing asymptotically distribution-free goodness-of-fit tests based on the Kullback-Leibler divergence using the mth-order spacings between order statistics:

$$H_{mn} := n^{-1} \sum_{i=1}^{n} \log \frac{n}{2m} (X_{(i+m)} - X_{(i-m)}). \tag{2.1}$$

Here, the window width m is a positive integer smaller than $n/2$. 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. The asymptotic theory suggests that m should be chosen adaptively according to the sample size.

However, there are some limitations to the tests: finding the optimal choice of m is clearly a difficult problem. To overcome the problem, we extend (2.1) by the kernel smoothing method to

$$\widehat{\ell_n(f)} = -\sum_{i=1}^{n} \log \left(\sum_{j \in \mathcal{J}_i} w_{h_n ij} X_{(j)} \right) \tag{2.2}$$

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

3. Simulation Study

The choice of bandwidth h is critical to implementation of a testing procedure. To test H_0 : $f_0(x, \theta) \in \mathscr{F}_0$, our asymptotic study suggests that h should be chosen adaptively according to the sample size and would ensure the distribution-free property and consistency of our test. The test statistic $T_{n,h} = 2(\widehat{\ell_n(f)} - \hat{\ell}_n(f_0))$ is based on the kernel smoothing method (2.2), where $\hat{\ell}_n(f_0) = \sum_{i=1}^{n} \log f_0(X_{(i)}, \hat{\theta}_n)$. we choose $\hat{h}_n^{opt}$ as the estimate of h_n^{opt} for

the given sample size that minimizes the $\widehat{\ell_n(f)} - \hat{\ell}_n(f_0)$ with respect to h under the null hypothesis. The following data-driven method of choosing smoothing parameter h_n^{opt} in terms of log-likelihood:

$$\hat{h}_n^{opt} := \underset{O(\frac{\log n}{n})<h<o(n^{-\frac{2}{3}}\log^{-\frac{4}{3}} n)}{\arg\min} \left\{ \widehat{\ell_n(f)} : \quad \widehat{\ell_n(f)} \geq \hat{\ell}_n(f_0) \right\}.$$

were chosen.

In the simulation study, we considered the problem of testing the composite hypothesis of normality when both the mean and the variance are unspecified against ten alternatives for sample size $n = 20$, $n = 50$ and $n = 100$ at the level $\alpha = 0.05$ using the triweight kernel. These alternative distributions are standard exponential denoted by Exp(1), gamma distribution with shape parameter $p = 2$, and scale parameter $\lambda = 1$ denoted by Gamma(2, 1), Uniform distribution on (0, 1) denoted by U(0, 1), Beta(2, 1), Beta(2, 6), Laplace with density function given by

$$f(x, \theta) = \frac{1}{2\phi} exp(-|x - \mu|/\phi)$$

where $\theta := (\mu, \phi) = (0, 1/4)$ denoted by Laplace(0, 0.25)), log-normal with density function given by

$$f(x, \theta) = \frac{1}{x\tau\sqrt{2\pi}} exp(-\frac{1}{2\tau^2}(\log x - \nu)^2)$$

where $\theta := (\nu, \tau) = (2, 1/4)$ denoted by Lognormal(2, 0.25). The last six alternatives were added to present various shapes of densities similar to a normal density. To determine the critical values of the test, we generalized 5000 replicate samples of size 20, 50 and 100 respectively from the standard normal distribution. For each sample, triweight kernel was used on the regular grid of h ranging in order from .05 to .45 by 0.05 and the corresponding estimated $\hat{h}_n$. The $(1-\alpha)th$ quantiles of the test statistics were then estimated. Once these critical values had been determined, the powers of the test were estimated by simulations, i.e., for each alternative and each h_n, 5000 samples of size 20, size 50 and size 100 were generated from the corresponding alternative distribution and the powers were thus estimated. These Monte Carlo power estimates are given in Table 1-2-3. Note that $h \approx n^{-1}$ (i.e., 0.05 for $n = 20$, 0.02 for $n = 50$ and 0.01 for $n = 100$), which corresponds to no smoothing at all, is included in Table 1-2-3 just for reference. It is not surprising to see poor performance with the pretty low powers for some alternatives at bandwidth h close to no-smoothing points.

Table 1: Power Estimates for Various Choices of h and Alternatives ($n = 20$, replicate= 5000, $\alpha = 0.05$)

Alternative	h=.05	h=.10	h=.15	h=.20	h=.25	h=.30	h=.35	h=.40	h=.45	$\hat{h}_n$
Exp(1)	0.0206	0.2524	0.7030	0.8142	0.8452	0.8594	0.8570	0.8576	0.8518	0.8508
Gamma(2, 1)	0.0000	0.0208	0.2484	0.4018	0.4602	0.4828	0.4790	0.4792	0.4764	0.4670
U(0, 1)	0.0000	0.0004	0.0312	0.1362	0.2576	0.3316	0.3688	0.3966	0.4170	0.3210
Beta(2, 1)	0.0000	0.0000	0.0004	0.0164	0.1322	0.2654	0.3320	0.3826	0.4124	0.2616
Beta(2, 6)	0.0000	0.0010	0.0478	0.1286	0.1702	0.1996	0.2074	0.2148	0.2180	0.1860
Laplace(0, 0.25)	0.0322	0.1532	0.2104	0.1756	0.1360	0.1072	0.0788	0.0640	0.0538	0.1052
Lognormal(2, 0.25)	0.0000	0.0000	0.0050	0.0364	0.0716	0.0948	0.1010	0.1072	0.1102	0.0778
t(3)	0.0310	0.2058	0.2634	0.2340	0.2024	0.1692	0.1442	0.1246	0.1108	0.1686
t(5)	0.0146	0.1006	0.1390	0.1192	0.0980	0.0846	0.0706	0.0654	0.0598	0.0828
Weilbull(2, 0.5)	0.0000	0.0000	0.0250	0.0718	0.1034	0.1234	0.1292	0.1322	0.1338	0.1092

Table 2: Power Estimates for Various Choices of h and Alternatives ($n = 50$, replicate=5000, $\alpha = 0.05$)

Alternative	h=.05	h=.10	h=.15	h=.20	h=.25	h=.30	h=.35	h=.40	h=.45	h=.50	$\hat{h}_n$
Exp(1)	0.0108	0.9970	0.9986	0.9994	0.9990	0.9990	0.9990	0.9978	0.9968	0.9954	0.9992
Gamma(2, 1)	0.0000	0.7652	0.8630	0.9220	0.9174	0.9144	0.8942	0.8714	0.8330	0.7992	0.9200
U(0, 1)	0.0000	0.3626	0.8372	0.9266	0.9486	0.9638	0.9694	0.9724	0.9738	0.9746	0.9248
Beta(2, 1)	0.0000	0.0252	0.7940	0.8994	0.9370	0.9428	0.9438	0.9412	0.9386	0.9348	0.9058
Beta(2, 6)	0.0000	0.2166	0.3922	0.5594	0.5680	0.5786	0.5656	0.5486	0.5230	0.5032	0.5528
Laplace(0, 0.25)	0.0256	0.4128	0.3040	0.2278	0.1302	0.0666	0.0324	0.0190	0.0116	0.0084	0.2280
Lognormal(2, 0.25)	0.0000	0.0302	0.0978	0.2314	0.2156	0.2204	0.1946	0.1780	0.1564	0.1454	0.2242
t(3)	0.0266	0.4792	0.3868	0.3162	0.2288	0.1556	0.1042	0.0682	0.0468	0.0326	0.3156
t(5)	0.0130	0.2148	0.1534	0.1134	0.0690	0.0470	0.0312	0.0236	0.0176	0.0136	0.1136
Weilbull(2, 0.5)	0.0000	0.0916	0.1974	0.3210	0.3276	0.3394	0.3306	0.3186	0.2992	0.2822	0.3144

Table 3: Power Estimates for Various Choices of h and Alternatives ($n = 100$, replicate=5000, $\alpha = 0.05$)

Alternative	h=.05	h=.10	h=.15	h=.20	h=.25	h=.30	h=.35	h=.40	h=.45	h=.50	$\hat{h}_n$
Exp(1)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9996	1.0000
Gamma(2, 1)	0.8404	0.9972	0.9982	0.9976	0.9952	0.9900	0.9816	0.9692	0.9444	0.9150	0.9980
U(0, 1)	0.1562	0.9980	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Beta(2, 1)	0.0000	0.9932	0.9996	0.9996	0.9998	1.0000	0.9998	0.9996	0.9996	0.9994	0.9996
Beta(2, 6)	0.1050	0.8556	0.9138	0.9132	0.8974	0.8784	0.8524	0.8254	0.7988	0.7708	0.9144
Laplace(0. 0.25)	0.6974	0.6000	0.4374	0.2446	0.1004	0.0324	0.0094	0.0032	0.0024	0.0020	0.2888
Lognormal(2, 0.25)	0.0006	0.3398	0.4376	0.4026	0.3494	0.2952	0.2426	0.2078	0.1764	0.1578	0.4164
t(3)	0.7606	0.6786	0.5154	0.3248	0.1788	0.0914	0.0498	0.0270	0.0140	0.0086	0.3630
t(5)	0.3672	0.2526	0.1464	0.0708	0.0282	0.0134	0.0054	0.0034	0.0028	0.0016	0.0812
Weilbull(2, 0.5)	0.0184	0.5692	0.6724	0.6718	0.6416	0.6042	0.5658	0.5332	0.4996	0.4736	0.6770

These power simulations show that for a fixed n, there does not exist an h which is optimal uniformly for all alternatives considered. This makes quite sense in the situation of goodness of fit testing against all nonparametric density alternatives, since alternatives are vague and the choice of the bandwidth h is designed to guard against all nonparametric density alternatives, and it is natural not to expect that the chosen $\hat{h}_n$ would beat all other choices of h in terms of power. Our simulation results are very encouraging. From Tables 1-2-3, we can see that the powers for $\hat{h}_n$ are far greater than or as close as median powers for all choices of h for sample size $n = 100$ and the powers for $\hat{h}_n$ are far greater than or as close as median powers for all choices of h for sample size $n = 20$ and $n = 50$. These results suggest that the data-driven method of choosing h is a very promising procedure of overcoming the dependence problem of the power of the test on h. These results

also suggest one possible way of choosing the optimal h in the situations where one has in mind a particular alternative being tested against, i.e., if a specific alternative is of special interest then the best way of choosing h would be to choose h that yields the highest power in the direction of this alternative for the given sample size and level α. Furthermore, these results suggest another way to improve power against all nonparametric density alternatives or a particular alternative in mind is to increase the sample size by its own nature in nonparametric setting. This was shown from our simulations that all the powers increases as sample size increases.

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.

Jianqing Fan and Jiancheng Jiang (2007). Nonparametric inference with generalized likelihood ratio tests. Test, 16:409-444.