The Korean Communications
in Statistics Vol. 6, No. 3, 1999
pp. 967-975

Minimum Hellinger Distance Based Goodness-of-fit Tests in Normal Models: Empirical Approach

Dong Bin Jeong[1)]

Abstract

In this paper we study the Hellinger distance based goodness-of-fit tests that are analogs of likelihood ratio tests. The minimum Hellinger distance estimator (MHDE) in normal models provides an excellent robust alternative to the usual maximum likelihood estimator. Our simulation results show that the Hellinger deviance test (Simpson 1989) based goodness-of-fit test is robust when data contain outliers. The proposed Hellinger deviance test (Simpson 1989) is a more direct method for obtaining robust inferences than an automated outlier screen method used before the likelihood ratio test data analysis.

1. Introduction

The likelihood ratio tests (Neyman and Pearson 1928; Wilks 1938) used widely for testing in parametric problems have certain asymptotic optimality properties but are not, in general, robust when data contain outliers. For a careful data analysis the dataset need to be screened for anomalous data points prior to an application of likelihood ratio tests. Simpson (1989) proposed a more direct procedure for robust inferences than the method of automated outlier screening and then using the likelihood based test. Simpson's Hellinger deviance tests are defined as analogs of likelihood ratio tests, and they are robust under data contamination and asymptotically equivalent to the likelihood ratio tests under local parametric alternatives. Before Simpson's (1989) work, robust versions of Wald (1943) tests and the Rao (1948) tests were studied by many authors in various settings (see Beran 1981; Hampel, Ronchetti, Rousseeuw and Stahel 1986).

The Hellinger deviance test is based on the minimum Hellinger distance estimator (MHDE) which is first-order efficient, yet has certain robustness properties (See Beran 1977; Tamura and Boos 1986; Simpson 1987). The M-estimation based robust procedures attain robustness at the cost of first order efficiency (Hampel, Ronchetti, Ronsseeuw and Stahel 1986). For the discrete models, Lindsay (1994) showed how to create a class of density based distances called disparities in order to produce estimators that are robust and the first order efficient (or even second order efficient) at the model. The class of disparities includes the Hellinger

1) Assistant Professor, Department of Statistics, Kangnung National University, Kangnung,210-702, Korea

distance (HD) as a member. The rest of this paper is organized as follows. In Section 2 we provide a review of minimum Hellinger distance estimation for continuous models. Hellinger deviance test is described in Section 3. The simulation scheme is laid out in Section 4 and the results are discussed in Section 5. Concluding remarks are provided in Section 6.

2. The Minimum Hellinger Distance Estimation (MHDE)

Suppose that we have a random sample $(X_1, X_2, \ldots, X_n)$ from a parametric class of distributions $\mathfrak{I}_{\Theta} = \{F_\theta, \theta \in \Theta\}$, where Θ is a subset of R^p. Assume that the family of distributions $\{F_\theta\}$ is dominated and f_θ represents the corresponding density for F_θ. The density based minimum disparity estimates (Lindsay 1994) can be computed by minimizing a nonnegative measure of discrepancy ρ_G, between a nonparametric density estimate $\hat{f}_n$ and the model density f_θ, defined by

$$\rho_G(\hat{f}_n, f_\theta) \equiv \int G(\delta(\hat{f}_n, \theta, x))dF_\theta(x), \tag{2.1}$$

where G is a real-valued three times differentiable, strictly convex function G on $[-1, \infty)$ with $G(0) = 0$, and

$$\delta(\hat{f}_n, \theta, x) = [\hat{f}_n(x) - f_\theta(x)]/f_\theta(x)$$

denote the "Pearson" residual at the value x, which depends on the data and the parameter θ. For simplicity of notation, we will sometimes write $\delta(\hat{f}_n, \theta, x)$ simply as $\delta(x)$. For data from a continuous distribution, one can use a nonparametric kernel density estimator defined by

$$\hat{f}_n(x) \equiv \int w(x; y, h_n)dF_n(y),$$

where w is smooth family of kernel functions like the normal densities with mean y and standard deviation h_n, F_n is the empirical distribution function. Following Beran's (1977) approach one will let the bandwidth $h_n \to 0$ as $n \to \infty$.

The value of θ that minimizes (2.1) is called *the minimum disparity estimator.* The function $G(\delta) = [(\delta+1)^{1/2} - 1]^2$ produces the squared HD whereas $G(\delta) = (\delta+1)\log(\delta+1)$ generates the likelihood disparity whose minimizer is the MLE of θ.

Let ∇ represent the gradient with respect to θ. Under differentiability of the model, the minimum disparity estimating equation takes the form

$$-\nabla \rho_G = \int A(\delta(x)) \nabla F_\theta(x) = 0,$$

where

$$A(\delta) \equiv (\delta+1)[\dot{G}(\delta)] - G(\delta)$$

and $\dot{G}(\delta)$ denotes the first derivative of $G(\delta)$. The function $A(\delta)$, peculiar to the disparity, is an increasing function on $[-1, \infty)$ and can be standardized, without changing the estimates produced by the disparity, so that for the standardized $A(\delta)$ we have $A(0) = 0$ and $\dot{A}(0) = 1$, where $\dot{A}(\delta)$ denotes the first derivative of $A(\delta)$. This standardized function $A(\delta)$ is called *the residual adjustment function* (RAF) of the disparity and determines most of the theoretical properties of the estimators. For the likelihood disparity the RAF is $A(\delta) = \delta$ and is already standardized. To achieve the above mentioned standardization one multiplies the squared HD by a factor of two to get

$$A(\delta) = 2[(\delta+1)^{1/2} - 1].$$

For continuous models, Basu and Lindsay (1994) obtained asymptotically fully efficient and robust minimum disparity estimators by applying the same smoothing to the model density f_θ that is applied to the data to define

$$\hat{f}_\theta(x) \equiv \int w(x; y, h_n) dF_\theta(y)$$

with $\widehat{F}_\theta$ being the corresponding distribution function. To obtain the estimator of θ for a disparity measure ρ_G, Basu and Lindsay minimize

$$\rho_G(\hat{f}_n, \hat{f}_\theta) = \int G\left(\frac{\hat{f}_n(x) - \hat{f}_\theta(x)}{\hat{f}_\theta(x)}\right) d\widehat{F}_\theta(x)$$

with respect to θ by keeping the bandwidth h_n of the kernel function constant. In their approach $\widehat{g_n}(x)$ is an unbiased estimator of $\hat{f}_\theta(x)$, and the minimum disparity estimators are robust and in general consistent for a fixed value of the bandwidth of the kernel function. One does not need to let h_n go to zero as sample size n increases, as is usually done. In this approach there is no loss in efficiency due to the smoothing of the model, if suitable kernels called *transparent kernels*, like the normal kernel for the normal model, are used (Basu and Lindsay 1994; Basu and Sarkar 1994). If, however, a transparent kernel is not available for use, the minimum disparity estimators are asymptotically normal, but no longer enjoy full asymptotic efficiency. To overcome this problem for the HD estimation procedure, combining the ideas of Beran (1977), Tamura and Boos (1986) and Lindsay (1994), establish the asymptotic efficiency and robustness of the MHDE, irrespective of the transparency of the kernel, obtained by minimizing

$$\rho_{HD}(\widehat{f_n}, f_\theta) = \int [(\frac{\widehat{f_n}(x)}{f_\theta(x)})^{1/2} - 1]^2 dF_\theta(x),$$

which is the HD between $\widehat{f_n}$ and f_θ where $h_n > 0$ satisfies $n^{1/2}h_n^2 \to 0$ and $n^{1/2}h_n \to \infty$. Using the above MHDE we define goodness-of-fit tests in the next section.

3. Hellinger Deviance Tests

Let Θ_0 be a proper subset of Θ and consider the problem of testing the null hypothesis $H_0: \theta \in \Theta_0$ against the alternative hypothesis $H_a: \theta \in \Theta \backslash \Theta_0$. The log likelihood ratio statistic is given by

$$\Lambda = 2n[L_n(\widetilde{\theta}) - L_n(\widetilde{\theta}_0)],$$

where $L_n(\theta) = n^{-1} \sum_{i=1}^{n} \log(f_\theta(X_i))$ is the average log likelihood function, and $\widetilde{\theta}$ and $\widetilde{\theta}_0$ are points corresponding to the maximization of $L_n(\theta)$ over Θ and Θ_0. In general, for a disparity ρ_G one can define the disparity test statistic

$$d_G = 2n[\rho_G(\widehat{f_n}, f_{\widehat{\theta_0}}) - \rho_G(\widehat{f_n}, f_{\widehat{\theta}})],$$

where $\widehat{\theta}$ and $\widehat{\theta}_0$ correspond to the minimization of $\rho_G(\widehat{f_n}, f_\theta)$ over Θ and Θ_0. The likelihood ratio test Λ has the property that if Θ_0 is q-dimensional subset of Θ and Θ_1 is an r-dimensional subset of Θ_0, then the test of Θ_1 against $\Theta \backslash \Theta_1$ can be partitioned into a test of Θ_1 versus $\Theta_0 \backslash \Theta_1$ and a test of Θ_0 versus $\Theta \backslash \Theta_0$. This property of Λ is shared by the test d_G. The test d_{LD} corresponding to the likelihood disparity defined using $G(\delta) = (\delta+1)\log(\delta+1)$ will be a close relative of Λ. Note that when $G(\delta) = [(\delta+1)^{1/2} - 1]^2$ (corresponding to two times squared Hellinger distance) d_G gives Simpson's (1989) Hellinger deviance test $d_{HD} = 2n[\rho_{HD}(\widehat{f_n}, f_{\widehat{\theta_0}}) - \rho_{HD}(\widehat{f_n}, f_{\widehat{\theta}})]$.

4. Simulation Scheme

In our Monte Carlo study we have compared the Hellinger deviance test (d_{HD}), the likelihood disparity test (d_{LD}), and the likelihood ratio test Λ in finite samples. We have done the computations for the normal model. Simulations were run using MICROSOFT FORTRAN POWER STATION on WINDOWS 95. The subroutines DRNNOA, DRNUN and DRNSTT of the IMSL subroutine were used to generate the normal, uniform and t random numbers respectively.

For the normal model we consider the following three contaminated data generating schemes: The populations are

(i) $(1-\varepsilon)N(\mu, \sigma^2)+\varepsilon N(3,1)$;
(ii) $(1-\varepsilon)N(\mu, \sigma^2)+\varepsilon N(\mu, 25)$
(iii) $(1-\varepsilon)N(\mu, \sigma^2)+\varepsilon t(1)$

where ε is the contaminating proportion. We consider testing $H_0: \mu=0$ versus $H_a: \mu\neq 0$ (a) by treating σ^2 as known and equal to 1 and (b) by treating σ^2 to be unknown and to be estimated from data like μ. For the three test statistics d_{HD}, d_{LD} and Λ we have computed the level (and power) when data were generated with μ = 0 (and μ = 0.5 for power) under no contamination with ε = 0 as well as under contamination with ε = 0.20.

For the MHDE we computed the kernel density

$$\hat{f}_n = \frac{1}{nh_n} \sum_{i=1}^{n} w\left(\frac{x-X_i}{h_n}\right)$$

with biweight kernel $w(x)=(15/16)(1-x^2)^2$ for $|x|\leq 1$, and 0 otherwise and independent observations X_i's, $i=1,\ldots,n$. Parzen (1962) found the h_n which minimizes the integrated mean square error between a kernel density estimate and the true density f. When f is a $N(\mu, \sigma^2)$ density, for the biweight kernel, h_n is of the form

$$h_n = (35e)^{1/5}(\pi/8)^{1/10}\sigma n^{-1/5}.$$

If σ is not known as in the testing case (b) above, then in the above formula of h_n we replace σ with the robust estimate $\widehat{\sigma_n} = 1.48(median(|X_i - median(X_i)|)$. The numerical integrals were computed using Simpson's one-third rule, and the Newton-Raphson algorithm was used to solve for the roots of the estimating equations. As intial estimates of μ and σ in the iterations we used

$$\hat{\mu}^{(0)} = median(X_i), \quad \hat{\sigma}^{(0)} = 1.48 \times median(|X_i - \hat{\mu}^{(0)}|).$$

All the computations are done with 5000 replications, the same set of samples being used for the calculation of these statistics. In these calculations we have used nominal levels 0.01, 0.05 and 0.10 and sample sizes 12, 20 and 50. The empirical levels and powers of the tests under contamination and no contamination have been presented in Table 1 for case (a) ($\sigma^2 = 1$ known), and in Table 2 for case (b) (σ^2 unknown).

In case (a), the likelihood ratio statistic, $n\,\overline{X}^2$ with $\overline{X}$ the sample mean, is distributed as $\chi^2(1)$ distribution under standard normal sampling. For the tests d_{HD} and d_{LD} we used the $\chi^2(1)$ critical values. In case (b), the likelihood ratio statistic is given by $\Lambda = n\log[1 + T^2/(n-1)]$ where T is a t-statistic with n-1 degrees of freedom. The level $(1-\alpha)$ critical value for Λ is given by $n\log[1 + F(1-\alpha; 1, n-1)/(n-1)]$, where $F(\cdot\,; \upsilon_1, \upsilon_2)$ denotes the F distribution function with υ_1 and υ_2 degrees of freedom.

5. Discussion of the Results

The empirical levels and powers of the Hellinger deviance test (HDT), the likelihood disparity test (LDT) and the likelihood ratio test (LRT) of the hypothesis $H_0: \mu = 0$ for both pure $(\varepsilon = 0)$ and contaminated $(\varepsilon = 0.20)$ models are presented in Table 1 and 2. Now we discuss our findings based on Table 1 and Table 2.

For the pure model, the empirical sizes of the likelihood ratio test (LRT) are more or less closer to nominal level than those of Hellinger deviance test (HDT) but there are no big differences between two tests. When we increase the sample size, the empirical powers of both tests get bigger and closer with each other.

Under contamination, the levels of the LRT get perturbed while the robust Hellinger deviance test holds their levels much better than the likelihood disparity test and the likelihood ratio test. It also seems that as the sample size increase, the likelihood ratio test loses power under contamination. In the meanwhile, the results from both Table 1 and Table 2 show that the HDT based on Table 1 (with σ^2 known) has a little more relative efficiency than that based on Table 2 (with σ^2 unknown) with respect to both the LRT and the LDT.

6. Concluding Remarks

We have studied the small sample performance of goodness-of-fit tests based on HD under normal models. It is shown that the Hellinger deviance test is robust in the presence of outliers for the normal models. When the data come from the model, our study shows that compared to the likelihood ratio test and the likelihood disparity test, the empirical levels and powers of the Hellinger deviance test is satisfactory for the small sample size. Through an empirical study at the normal models it is shown that the Hellinger deviance test is good robust alternatives to the likelihood ratio test.

References

[1] Basu, A. and Lindsay, B. G. (1994), Minimum Disparity Estimation for Continuous Models: Efficiency, distributions and Robustness, *Ann. Inst. Statist. Math.*, Vol. 46, 683-705.

[2] Basu, A. and Sarkar, S. (1994). On Disparity Based Goodness-of-fit Tests for Multinomial Models, *Statistics & Probability Letters*, Vol. 19, 307-312.

[3] Beran, R. J. (1977). Minimum Hellinger Distance Estimates for Parametric Models, *The Annals of Statistics*, Vol. 5, 445-463.

[4] Beran, R. J. (1981). Efficient Robust Tests in Parametric Models, *Zeischrift fur Wahrscheeinlichkeitstheorie und Verwandte Gebiete*, Vol. 57, 73-86.

[5] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley.

[6] Lindsay,B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods, *The Annals of Statistics*, Vol. 22, 1081-1114.

[7] Neyman, J., and Pearson, E. S. (1928). On the Use and Interpretation of Certain Test Criteria for Purpose of Statistical Inference, *Biometrika*, Ser. A, Vol. 20, 175-240.

[8] Rao, C. R. (1948). Large Sample Tests of Statistical Hypotheses Concerning Several Parameters With Applications of Problems of Estimation, *Proceedings of the Cambridge Philosophical Society*, Vol. 44, 50-57.

[9] Simpson, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data, *Journal of the American Statistical Association*, Vol. 82, 802-807.

[10] Simpson, D. G. (1989). Hellinger Deviance Tests: Efficiency, Breakdown Points, and Examples, *Journal of the American Statistical Association*, Vol. 84, 107-113.

[11] Tamura, R. N. and Boos, D. D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance, *Journal of the American Statistical Association*, Vol. 81, 223-229.

[12] Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large, *Transactions of the American Mathematical Society*, Vol. 54, 426-482.

[13] Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypothesis, *The Annals of Mathematical Statistics,* Vol. 9, 60-62.

APPENDIX

Table 1. Level and power of the Hellinger Deviance Test (HDT), the Likelihood Ratio Test (LRT), and the Likelihood Disparity Test (LDT) for N(0, 1) vs. N(μ, 1), $\mu \neq 0$

Sampling distribution:	N(0, 1)			N(0.5, 1)		
Nominal level	0.10	0.05	0.01	0.10	0.05	0.01
n = 12						
HDT	9.58	4.34	0.90	50.92	37.88	16.56
LRT	10.06	5.02	1.02	53.96	41.48	19.64
LDT	8.86	4.68	1.06	53.44	41.10	19.38
n = 20						
HDT	10.06	4.54	0.94	70.22	57.76	33.88
LRT	10.04	5.04	1.04	72.30	60.24	37.36
LDT	10.26	5.12	0.98	72.78	61.14	37.38
n = 50						
HDT	9.78	4.80	0.92	96.62	93.18	80.78
LRT	10.02	5.04	1.02	96.80	94.14	83.74
LDT	10.28	4.86	0.88	97.18	94.14	83.32
Sampling distribution	0.8N(0, 1)+0.2N(3, 1)			0.8N(0.5, 1)+0.2N(3, 1)		
Nominal level	0.10	0.05	0.01	0.10	0.05	0.01
n = 12						
HDT	27.22	18.42	8.34	76.06	66.48	45.00
LRT	60.82	52.76	37.96	90.82	86.18	73.28
LDT	60.82	52.76	37.96	90.82	86.18	73.28
n = 20						
HDT	34.66	25.74	11.96	90.64	85.28	70.60
LRT	75.32	68.24	52.34	98.02	96.58	91.32
LDT	75.32	68.24	52.34	98.02	96.58	91.32
n = 50						
HDT	54.62	43.62	25.46	99.90	99.76	98.64
LRT	95.66	93.40	86.20	100.00	100.00	99.98
LDT	95.66	93.40	86.20	100.00	100.00	99.98
Sampling distribution	0.8N(0, 1)+0.2N(0, 25)			0.8N(0.5, 1)+0.2N(0.5, 25)		
Nominal level	0.10	0.05	0.01	0.10	0.05	0.01
n = 12						
HDT	13.52	8.04	2.36	45.86	34.94	16.50
LRT	45.26	37.92	26.86	59.98	52.70	39.00
LDT	44.48	37.52	26.42	59.66	52.18	38.44
n = 20						
HDT	13.78	7.74	2.08	61.50	50.52	29.46
LRT	45.94	38.70	26.70	65.64	58.24	45.88
LDT	45.64	38.22	26.16	65.48	58.00	45.52
n = 50						
HDT	13.28	7.02	1.92	90.16	84.68	66.92
LRT	50.22	41.84	27.90	79.76	75.26	65.90
LDT	49.60	41.26	27.26	79.68	75.08	65.62
Sampling distribution	0.8N(0, 1)+0.2t(1)			0.8N(0.5, 1)+0.2t(1)		
Nominal level	0.10	0.05	0.01	0.10	0.05	0.01
n = 12						
HDT	9.48	4.90	0.98	39.36	27.66	11.32
LRT	28.54	23.04	14.64	49.08	39.36	24.64
LDT	23.78	17.90	9.18	46.00	35.72	19.98
n = 20						
HDT	10.68	5.54	1.36	55.56	43.20	22.36
LRT	33.46	27.06	17.90	60.38	51.62	35.42
LDT	26.56	19.68	10.38	56.64	46.96	30.18
n = 50						
HDT	12.38	6.30	1.48	86.36	78.64	58.66
LRT	39.42	31.88	22.34	76.46	70.56	57.70
LDT	28.58	20.68	11.22	78.56	71.60	57.24

Table 2. Level and Power of Testing H_0: $\mu = 0$ vs. H_a: $\mu \neq 0$ with σ^2 Unspecified

Sampling distribution:	N(0, 1)			N(0.5, 1)		
Nominal level	0.10	0.05	0.01	0.10	0.05	0.01
n = 12						
HDT	9.32	5.46	1.14	47.06	35.14	17.16
LRT	10.10	5.04	0.98	48.32	35.00	14.22
LDT	6.72	3.38	0.86	42.94	30.64	14.18
n = 20						
HDT	9.42	5.10	1.14	66.90	53.64	31.14
LRT	10.00	5.04	1.02	68.62	55.96	29.10
LDT	7.30	3.74	0.78	64.00	49.98	27.22
n = 50						
HDT	9.34	4.74	0.92	96.22	92.16	78.14
LRT	10.06	5.02	1.02	96.48	93.24	79.22
LDT	8.02	3.80	0.74	95.70	91.42	75.80
Sampling distribution:	0.8N(0, 1)+ 0.2N(3,1)			0.8N(0.5, 1)+0.2N(3,1)		
Nominal level	0.10	0.05	0.01	0.10	0.05	0.01
n = 12						
HDT	31.78	21.54	7.50	78.68	62.96	36.58
LRT	32.12	27.52	9.18	78.64	63.10	37.34
LDT	32.56	26.36	8.70	78.12	62.26	37.10
n = 20						
HDT	48.62	34.20	14.44	93.80	88.52	70.02
LRT	50.62	35.06	15.02	94.26	88.38	65.20
LDT	49.82	34.90	14.94	93.40	86.92	64.66
n = 50						
HDT	81.82	73.06	49.26	100.00	99.96	99.42
LRT	87.20	78.06	50.18	100.00	99.98	99.66
LDT	84.70	74.54	50.10	100.00	99.92	99.40
Sampling distribution:	0.8N(0, 1)+ 0.2N(0,25)			0.8N(0.5, 1)+ 0.2N(0.5,25)		
Nominal level	0.10	0.05	0.01	0.10	0.05	0.01
n = 12						
HDT	13.64	8.12	2.50	44.58	34.78	18.96
LRT	4.14	2.36	0.40	25.38	14.98	3.74
LDT	4.58	2.40	0.46	25.48	15.24	5.44
n = 20						
HDT	13.88	7.88	2.16	56.34	45.46	27.94
LRT	4.96	2.38	0.32	30.56	20.48	7.24
LDT	5.06	2.42	0.38	31.22	21.46	7.30
n = 50						
HDT	12.52	7.20	1.98	85.30	77.38	58.86
LRT	5.60	2.58	0.34	46.76	35.20	16.40
LDT	5.64	2.62	0.40	47.22	35.52	16.76
Sampling distribution:	0.8N(0, 1)+ 0.2t(1)			0.8N(0.5, 1)+ 0.2t(1)		
Nominal level	0.10	0.05	0.01	0.10	0.05	0.01
n = 12						
HDT	12.12	4.28	1.78	41.48	30.70	15.34
LRT	5.88	3.12	0.48	27.26	17.52	5.32
LDT	6.62	4.14	0.70	30.24	20.84	8.38
n = 20						
HDT	11.02	5.90	1.28	53.36	43.04	23.20
LRT	6.36	3.28	0.58	34.54	24.36	10.24
LDT	7.52	3.46	0.62	38.38	28.42	12.86
n = 50						
HDT	10.72	5.58	1.26	84.14	75.40	54.62
LRT	6.78	3.44	0.40	45.54	35.86	20.10
LDT	7.42	3.56	0.46	58.44	47.14	27.56