Tests of hypotheses in discrete models based on the penalized Hellinger distance

Ayanendranath Basu[a,*], Ian R. Harris[a], Srabashi Basu[b]

[a] *Department of Mathematics, University of Texas, Austin, TX 78712, USA*
[b] *University of Texas Health Science Center, San Antonio, TX 78284, USA*

Received February 1995; revised March 1995

Abstract

Analogues of the likelihood ratio. Rao, and Wald tests are introduced in discrete parametric models based on the family of penalized Hellinger distances. It is shown that the tests based on a particular member of this family provide attractive alternatives to the tests based on the ordinary Hellinger distance. These tests share the robustness of the Hellinger distance test, but are often closer to the likelihood-based tests at the model, especially in small samples. The convergence of ordinary Hellinger distance tests to limiting χ^2 distributions are quite slow. The proposed tests are improvements in this respect.

Keywords: Hellinger distance; Robustness; Likelihood ratio test; Rao test; Wald test

1. Introduction

The likelihood ratio test (Neyman and Pearson, 1928; Wilks, 1938) is widely used for testing of hypotheses. The Rao (score) test and the Wald test, both asymptotically equivalent to the likelihood ratio test under the null hypothesis, also utilize the likelihood in their construction. The popularity of the likelihood-based procedures is however tempered by their known lack of robustness to outliers. A major work by Beran (1977) showed that one can simultaneously obtain asymptotic efficiency and robustness properties by using the minimum Hellinger distance estimator (MHDE). Robust tests of hypotheses based on the Hellinger distance were introduced by Simpson (1989) and studied further by Lindsay (1994). These works have shown that some density-based distances can generate robust alternatives to the likelihood ratio test.

Harris and Basu (1994) have considered a general family of distances called the *penalized Hellinger distance*; this family contains the Hellinger distance and is a function of a parameter h which controls the weight of the empty cells in the distance. Harris and Basu have shown that by adjusting this penalty it may be possible to improve significantly upon the performance of the MHDE without compromising its robustness properties. In this paper we consider tests of hypotheses based on the penalized Hellinger distance. The tests considered by Simpson (1989) and Lindsay (1994) are analogues of the likelihood ratio test. Here we introduce analogues of Rao tests and Wald tests also. We compare the tests generated by a particular member of the penalized Hellinger distance family with those of the ordinary Hellinger distance and the likelihood-based methods. This distance corresponds to $h = 0.5$; Harris and Basu argued that the estimators generated by this distance share the robustness properties of the MHDE, but have smaller mean square errors than the latter in small samples. Here we discuss the asymptotic properties of the corresponding tests and investigate their small sample behavior in a modest simulation study.

2. Tests of hypotheses based on the penalized Hellinger distance

Consider a parametric family with countable support and density $m_\beta(x)$, $\beta \in \Omega$. Let $d(x)$ be the proportion of sample observations having the value x based on a sample of size n. The maximum likelihood estimator (MLE) of β minimizes the Kullback–Leibler divergence between d and m_β, defined as

$$\mathrm{KL}(d, m_\beta) = \sum_x d(x) \log\,(d(x)/m_\beta(x)). \tag{2.1}$$

The MHDE of β minimizes the distance

$$\mathrm{HD}(d, m_\beta) = 2 \sum_x (d^{1/2}(x) - m_\beta^{1/2}(x))^2 . \tag{2.2}$$

The factor of 2 in (2.2) makes the two distances asymptotically equivalent. Among others, Simpson (1987) and Lindsay (1994) have studied the MHDE in discrete models.

Let $\Omega \subset \mathscr{R}^p$, $\hat\beta_M$ be the MLE of β and suppose that the hypothesis

$$\mathrm{H}_0\colon \beta = \beta_0 \tag{2.3}$$

is of interest. It is well known that

$$2n[\mathrm{KL}(d, m_{\beta_0}) - \mathrm{KL}(d, m_{\hat\beta_M})] , \tag{2.4}$$

which equals negative of twice log likelihood ratio, has an asymptotic $\chi^2(p)$ distribution under the null; see, for example, Serfling (1980, Section 4.4). The likelihood ratio tests have several optimality properties, but are not robust against outliers. Simpson (1989) proposed the Hellinger deviance test; in discrete models, the Hellinger deviance test statistic can be obtained by replacing the Kullback–Leibler divergence by the Hellinger distance in Eq. (2.4). The Hellinger deviance test is asymptotically equivalent to the likelihood ratio test at the null and local alternatives, and has attractive breakdown robustness properties (Simpson 1989; Theorems 1 and 2). See He et al. (1990) for a nice general description of the concept of breakdown robustness of tests.

Simpson, however, noted that the small sample performance of the Hellinger deviance test at some discrete models such as the Poisson is somewhat unsatisfactory, in the sense that the test requires a very large sample size for the chi-square approximation to be useful (Simpson 1989, Table 3). While it is difficult to give a complete explanation of this relatively poor behavior, the results of Lindsay (1994) suggest that this may be partially due to the large weight that the Hellinger distance puts on the *inliers*, values with less data than expected under the model. Lindsay studied the robustness and distributional properties of a subclass of minimum distance estimators including the MHDE through a characterizing function which determines how the procedure treats standardized residuals of the form $\delta(x) = d(x)/m_\beta(x) - 1$, $\delta(x) \in [-1, \infty)$. Large positive values of $\delta(x)$ represent outlying observations which are downweighted by the Hellinger distance. Negative values of $\delta(x)$ represent inliers. Lindsay noticed that while the MHDE is usually insensitive to the presence of large outliers in the data, inliers appear to cause larger biases in the MHDE compared to the MLE (Lindsay 1994, Table 3).

Empty cells, in particular, constitute part of the inlier problem of the Hellinger distance. Note that one can write the Kullback–Leibler divergence in the form

$$\mathrm{KL}(d, m_\beta) = \sum_x \left[d(x) \log (d(x)/m_\beta(x)) + (m_\beta(x) - d(x))\right]. \tag{2.5}$$

When written in the above form, the summand itself is nonnegative. Comparing (2.2) with (2.5) one can see that the contribution of a cell with $d(x) = 0$ to the distance is $2m_\beta(x)$ in (2.2), but equals just half of that in (2.5). For any given value of β, one can modify the Hellinger distance in (2.2) by considering the distance

$$2 \sum_{d(x) \neq 0} (d^{1/2}(x) - m_\beta^{1/2}(x))^2 + \sum_{d(x)=0} m_\beta(x), \tag{2.6}$$

which puts the same weight on the empty cells as (2.5). A generalization of this was considered by Harris and Basu (1994), who defined the penalized Hellinger distance (PHD) family as

$$\mathrm{PHD}_h(d, m_\beta) = 2\left[\sum_{d(x) \neq 0} (d^{1/2}(x) - m_\beta^{1/2}(x))^2 + h \sum_{d(x)=0} m_\beta(x)\right], \tag{2.7}$$

where $h = 1$ generates the ordinary Hellinger distance, and $h = 0.5$ generates the distance in (2.6).

Note that if there are empty cells, the distance in (2.6) is *strictly* closer to the Kullback–Leibler divergence than the ordinary Hellinger distance for all values of β, suggesting the possibility that compared to the MHDE the minimizer of (2.6) may be closer to the MLE. Compared to the score function (the derivative of the distance with respect to β) of the ordinary Hellinger distance, the score function of $\mathrm{PHD}_{0.5}$ is also strictly closer to that of the Kullback–Leibler divergence for all values of β. Harris and Basu have studied the effect of modifying the weight of the empty cells in parameter estimation; their results show that often the estimator obtained by minimizing (2.6) can have substantially smaller mean square errors than the ordinary MHDE in small samples. Note that since the difference between the ordinary Hellinger distance and the other members of the penalized Hellinger distance family is *only* in the empty cells, the outlier resistance properties of the MHDE are shared by all the minimum penalized Hellinger distance estimators.

Next we introduce the three classes of tests corresponding to the likelihood ratio test, the Rao test and the Wald test based on the penalized Hellinger distance. The theoretical results are based on the works of Simpson (1989) and Lindsay (1994).

The deviance test statistic for (2.3) based on the penalized Hellinger distance is

$$2n\left[\mathrm{PHD}_h(d, m_{\beta_0}) - \mathrm{PHD}_h(d, m_{\hat{\beta}_{\mathrm{PH}}})\right], \tag{2.8}$$

where $\hat{\beta}_{\mathrm{PH}}$ minimizes (2.7). When $h = 1$, this is the ordinary Hellinger deviance test statistic (Simpson, 1989), and has the same asymptotic χ^2 distribution as the likelihood ratio test statistic in (2.4) under the null. Since the other members of the penalized Hellinger deviance tests differ from the Hellinger deviance test only at the empty cells, they too have the same asymptotic distribution as the likelihood ratio test at the null. Specifically, when $h = 0.5$, our expectation is that compared to the MHDE, the minimum penalized Hellinger distance estimator will be closer to the maximum likelihood estimator, and the corresponding distance will be strictly closer to the Kullback–Leibler divergence compared to the ordinary Hellinger distance; therefore, that the penalty may cause the test statistic to be closer to the likelihood ratio test on the average, leading to more accurate type I errors.

Let $a_{n\beta} = \partial/\partial\beta[\mathrm{KL}(d, m_\beta)]$. The Rao test for the hypothesis (2.3), given by the statistic $na_{n\beta_0}^{\mathrm{T}} I_{\beta_0}^{-1} a_{n\beta_0}$, has the same limiting χ^2 distribution as the likelihood ratio test under the null, where I_β is the Fisher information about β in $m_\beta(x)$ (Serfling, 1980, Section 4.4.2). To distinguish it from similar tests based on the penalized

Hellinger distance, we will refer to this test as the *likelihood-Rao test*. Evaluation of the statistic $na_{n\beta_0}^{\mathrm{T}} I_{\beta_0}^{-1} a_{n\beta_0}$ does not require the explicit calculation of the MLE. Let $b_{n\beta} = \partial/\partial\beta[\mathrm{PHD}_h(d, m_\beta)]$. When $h = 1$, $n^{1/2} b_{n\beta_0}$ has an asymptotic $\mathrm{N}(0, I_{\beta_0})$ distribution under the null (Simpson, 1989). From Serfling (1980, Lemma A, Section 4.4.2) it follows that $nb_{n\beta_0}^{\mathrm{T}} I_{\beta_0}^{-1} a_{n\beta_0}$ has the same limiting distribution as the likelihood-Rao test statistic, and we refer to this test as the *Hellinger–Rao* test. The penalized Hellinger–Rao tests corresponding to other values of h also have the same asymptotic distribution since they differ from the Hellinger–Rao test only at the empty cells. Note also that compared to the Hellinger–Rao statistic, the penalized Hellinger–Rao statistic for $h = 0.5$ is *strictly* closer to the likelihood-Rao statistic, indicating that its type I errors could be closer to the likelihood-Rao statistic.

The *Likelihood–Wald* test for the hypothesis (2.3) is given by the statistic $n(\hat{\beta}_M - \beta_0)^{\mathrm{T}} I_{\beta_0}(\hat{\beta}_M - \beta_0)$, having an asymptotic $\chi^2(p)$ distribution under the null. As in the derivation of the penalized Hellinger deviance tests and the penalized Hellinger–Rao tests, it follows from Simpson (1989) and Lindsay (1994) that the distributions of the *penalized Hellinger–Wald* test statistics $n(\hat{\beta}_{PH} - \beta_0)^{\mathrm{T}} I_{\beta_0}(\hat{\beta}_{PH} - \beta_0)$ have the same asymptotic $\chi^2(p)$ limit under the null as that of the corresponding likelihood-based version. The results of Harris and Basu indicate that compared to the MHDE the minimum distance estimator for $h = 0.5$ can be closer to the MLE on the average; we hope that this will lead to more accurate type I errors for this penalized Hellinger–Wald test.

By inverting the three types of test statistics considered above, one can also determine robust confidence intervals for the unknown β. Given the data and any of the above test statistics, the $100(1 - \alpha)\%$ confidence interval for β is the set of values of β for which the test statistic fails to reject the null hypothesis (2.3) at level α. Also, the testing procedures described in this section extend straightforwardly to the case where the null hypothesis is composite, using the techniques of Serfling (1980, Section 4.4.4). The penalized Hellinger tests again have the same asymptotic distribution as the likelihood-based ones.

The difference between the Hellinger distance and the other members of the penalized Hellinger family asymptotically vanish at the model, and so the procedures resulting from other penalized Hellinger distances inherit the asymptotic properties of the Hellinger distance based procedures; thus the asymptotic distribution of the penalized Hellinger distance estimators, and the breakdown robustness properties of the corresponding test statistics are identical to those of the Hellinger distance procedures. We emphasize that the differences between the penalized Hellinger distance tests and the ordinary Hellinger distance tests will be significant only in small samples. While we expect that the test statistics resulting from $\mathrm{PHD}_{0.5}$ will perform better than the tests based on the Hellinger distance in small samples, the difference is expected to be minimal for large n.

3. Simulation results

Data were generated from the $(1 - \varepsilon)\mathrm{Poisson}(2) + \varepsilon\mathrm{Poisson}(15)$ mixture, $\varepsilon = 0, 0.1$. Our target parameter is the mean of the larger Poisson(2) component, and the second component is considered to be a contamination. Assuming a Poisson(β) model for the data, we are interested in testing H_0: $\beta = 2$. We expect that compared to the likelihood-based tests, the tests generated by the ordinary ($h = 1$) and penalized ($h = 0.5$) Hellinger distances will be less sensitive to the presence of the Poisson(15) component and provide more stable levels. For each of deviance type, Rao type and Wald type tests, we considered three sample sizes, $n = 20, 50$ and 100, and three nominal levels, 0.1, 0.05 and 0.01. (For brevity we only present the results corresponding to nominal level 0.05, but the findings are similar.) The observed level of the tests were computed as the proportion of test statistics exceeding the $\chi^2(1)$ critical value. All the results in this section are based on 5000 replications. Given a probability estimate $\hat{p}$, its estimated standard deviation may be computed as $[\hat{p}(1 - \hat{p})/5000]^{1/2}$ (assuming binomial rejection frequencies) which can be no greater than $[0.5 \times 0.5/5000]^{1/2} = 0.007$.

Table 1
Comparison of observed levels of three tests at nominal level 0.05

		Contaminating proportion					
		$\varepsilon = 0.0$			$\varepsilon = 0.1$		
		Sample size n			Sample size n		
Testing procedure	Test statistic	20	50	100	20	50	100
Deviance	Likelihood	0.053	0.050	0.048	0.724	0.940	0.994
	HD	0.107	0.084	0.077	0.122	0.092	0.079
	PHD	0.049	0.051	0.054	0.053	0.060	0.066
Rao	Likelihood	0.051	0.049	0.052	0.746	0.942	0.995
	HD	0.112	0.092	0.083	0.108	0.088	0.073
	PHD	0.030	0.043	0.047	0.024	0.042	0.048
Wald	Likelihood	0.051	0.049	0.052	0.742	0.942	0.995
	HD	0.079	0.068	0.067	0.110	0.084	0.090
	PHD	0.057	0.056	0.055	0.082	0.082	0.084

The observed levels of the tests are presented in Table 1. It is clearly seen that the penalized Hellinger deviance test is very close to the likelihood ratio test when there is no contamination, and is much better than the Hellinger deviance test in terms of the closeness of the observed levels to the nominal levels. The Hellinger deviance test, the Hellinger–Rao test, and the Hellinger–Wald test are all very anti-conservative. The penalized Hellinger–Rao test is slightly conservative, but its levels tend to the nominal levels faster than the Hellinger–Rao test. The performance of the penalized Hellinger–Wald test also appears to be quite reasonable. Under contamination, the tests based on the Hellinger distance and the penalized Hellinger distance hold their levels quite well, unlike their likelihood based counterparts.

The entries in Tables 1 can also provide an idea of the accuracy of the confidence intervals obtained by inverting the tests. The numbers in these tables represent eliminated type I errors, the difference of these numbers from 1 give estimates of the actual coverage probability of the corresponding confidence intervals. In these examples, the penalized Hellinger tests provide more accurate confidence intervals than the Hellinger tests. Both groups produce more robust confidence intervals than the likelihood based procedures.

While the results presented here at the Poisson(2) distribution, similar improvements were noticed at Poisson(5) and Poisson(10) distributions, not presented here for brevity. For example, generating data from a Poisson(10) distribution, the three observed levels (nominal levels 10%, 5%, 1%) at sample size 100 for the hypothesis H_0: $\beta = 10$ using the Hellinger deviance test were 0.136, 0.076 and 0.021. The corresponding values for the likelihood ratio test were 0.107, 0.051, 0.010; for the penalized Hellinger deviance test they were 0.097, 0.048, 0.011.

In order to study a different model, we generated data from the Geometric(0.5) distribution. Using the Geometric(p) model, the observed levels for the test H_0: $p = 0.5$ were plotted as a function of the sample size in Fig. 1, at nominal level 10%. For each value of n (at intervals of 10) between 10 and 2500 we plotted the proportion of observed rejections of the null hypothesis out of the 5000 samples. The Hellinger deviance test is far worse than the penalized version at converging to the true level.

Generating data from the mixture $(1 - \varepsilon)\text{Poisson}(2) + \varepsilon\text{Poisson}(15)$, $\varepsilon = 0, 0.1$, we also tested the hypothesis H_0: $\beta = 3$ assuming the Poisson(β) model to study and compare the behavior of the test statistics in terms of their attained power at this hypothesis. Again the observed power is computed as the proportion of test statistics exceeding the $\chi^2(1)$ critical value. The results are presented in Table 2. The power of the

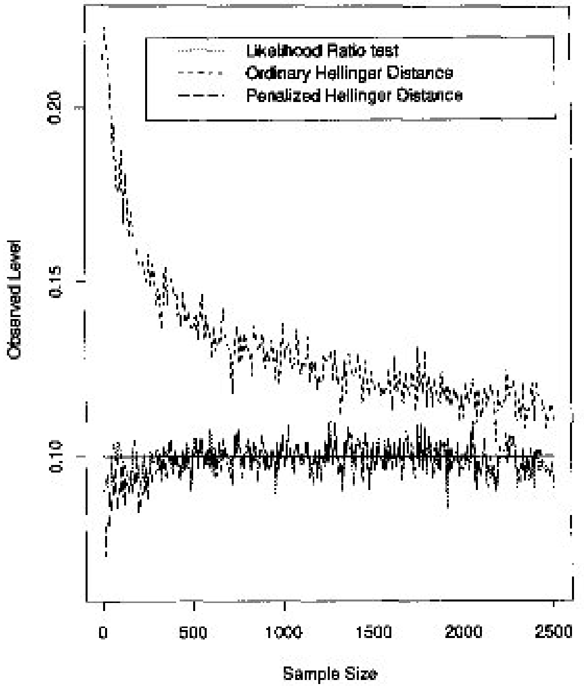

Fig. 1. Observed levels for the geometric examples.

Table 2
Comparison of observed powers of three tests at nominal level 0.05

		Contaminating proportion					
		$\varepsilon = 0.0$			$\varepsilon = 0.1$		
		Sample size n			Sample size n		
Testing procedure	Test statistic	20	50	100	20	50	100
Deviance	Likelihood	0.810	0.993	1.000	0.435	0.451	0.496
	HD	0.916	0.998	1.000	0.874	0.993	1.000
	PHD	0.798	0.994	1.000	0.728	0.983	1.000
Rao	Likelihood	0.766	0.992	1.000	0.409	0.445	0.504
	HD	0.913	0.998	1.000	0.847	0.988	0.999
	PHD	0.697	0.993	1.000	0.542	0.961	0.999
Wald	Likelihood	0.768	0.992	1.000	0.422	0.445	0.504
	HD	0.880	0.997	1.000	0.854	0.993	1.000
	PHD	0.806	0.994	1.000	0.770	0.987	1.000

penalized Hellinger distance tests were found to be much closer to the likelihood-based tests under the model compared to the power of the Hellinger distance tests. Under contamination there is a significant loss in power for the likelihood-based procedures; the robust tests did better at maintaining their powers. Note that the observed powers of the ordinary Hellinger distance tests are much higher than the other two types of tests, since the $\chi^2(1)$ critical values represent quite inaccurate approximation for the true critical values of three tests in small samples.

5. Discussion

Parametric testing of hypothesis using likelihood methods fare poorly in the presence of outliers and under model misspecification. Simpson (1989) and Lindsay (1994) have shown that tests based on robust distances like the Hellinger distance may do much better in such situations. In small samples, however, these tests may behave quite differently than the likelihood based tests when the assumed models are true, suggesting their slow convergence to the limiting chi-square distribution. In this paper we have provided alternative tests based on a modification of the Hellinger distance; these tests often do much better than those generated by the ordinary Hellinger distance in small samples. In addition, these tests share the robustness of the ordinary Hellinger distance tests. Thus, it seems that the penalized Hellinger tests can be good alternatives to the tests of hypothesis based on likelihood methods and the ordinary Hellinger distance.

Acknowledgements

The research of the first and second authors were partially supported by a Mathematics Department Research Award, University of Texas at Austin. The research of the third author was partially supported by a grant from MERECE at the University of Texas Health Science Center at San Antonio. The authors thank a referee whose suggestions led to an improved version of the paper. Discussions with Professor Bruce G. Lindsay have also been helpful.

References

Beran, R.J. (1977), Minimum Hellinger distance estimates for parametric models, *Ann. Statist.* **5**, 445–463.

Harris, I.R. and A. Basu (1994), Hellinger distance as a penalized log likelihood, *Comm. Statist. Simul. Compta.* **23**, 1097–1113.

He, X., D.G. Simpson and S.L. Portnoy (1990), Breakdown robustness of tests. *J. Amer. Statist. Assoc.* **85**, 446–452.

Lindsay, B.G. (1994), Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Statist.* **22**, 1081–1114.

Neyman, J. and E.S. Pearson (1928), On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika Ser A* **20**, 175–240.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics* (Wiley, New York).

Simpson, D.G. (1987), Minimum Hellinger distance estimation for the analysis of count data, *J. Amer. Statist. Assoc.* **82**, 802–807.

Simpson, D.G. (1989), Hellinger deviance test: efficiency, breakdown points, and examples. *J. Amer. Statist. Assoc.* **84**, 107–113.

Wilks, S.S. (1938), The large sample distribution of the likelihood ratio test for testing composite hypothesis. *Ann. Math. Statist.* **9**, 60–62.