

# Improved Power in Multinomial Goodness-of-fit Tests

Ayanendranath Basu

*Applied Statistics Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta 700 035, India.*

Surajit Ray

*Department of Statistics, Penn State University, University Park, PA 16802, USA.*

Chanseok Park

*Dept. of Mathematical Sciences, Clemson University, Clemson, SC 29634-0975, USA.*

Srabashi Basu

*Stat-Math Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta 700 035, India.*

Submitted to JRSS(D) : 26 Mar, 2001  
Revised version submitted : 22 Jan, 2002  
Published : Sep, 2002 vol. 51, no. 3, pp. 381-393

**Summary.** The Pearson's chi-square and the log likelihood ratio chi-square statistics are fundamental tools in goodness-of-fit testing. Cressie and Read (1984) constructed a general family of divergences which includes both statistics as special cases. This family is indexed by a single parameter, and divergences at either end of the scale are more powerful against alternatives of one type while being rather poor against the opposite type. Here we present several new goodness-of-fit testing procedures which have reasonably high power at both kinds of alternatives. Graphical studies illustrate the advantages of the new methods.

*Keywords:* Disparities, empty cell penalty, goodness-of-fit, power divergence.

## 1. Introduction

The Pearson's chi-square and the log likelihood ratio statistics are long standing techniques in goodness-of-fit testing under multinomial set ups. Many authors have investigated the scope and relative performance of these tests, and have compared them with other less popular statistics such as the Neyman modified chi-square statistic, the modified log likelihood ratio statistic (based on the Kullback-Leibler divergence) and the test statistic based on the Hellinger distance. See, for example, Cochran (1952), Watson (1959), Hoeffding (1965), West and Kempthorne (1972), Moore and Spruill (1975), Chapman (1976), Larntz (1978), and Koehler and Larntz (1980). Cressie and Read (1984), hereafter referred to as C&R, and Read and Cressie (1988) presented a unified approach to goodness-of-fit testing in multinomial models through the family of power divergences denoted by  $\{I^\lambda : \lambda \in \mathbb{R}\}$ . The Pearson's chi-square, the likelihood disparity (generating the log likelihood ratio statistic), the (twice, squared) Hellinger distance, the Kullback-Leibler divergence and the Neyman modified chi-square are indexed by  $\lambda = 1, 0, -1/2, -1$  and  $-2$  respectively. Based on a comparative study, Read and Cressie (1988) recommends  $I^{2/3}$  as a compromise candidate among the different test statistics, although they noted several desirable properties of the other test statistics, including the Pearson's chi-square  $I^1$  (see, eg. Section 4.5, Section 6.7, and Appendix A11 of Read and Cressie).

Basu and Sarkar (1994) considered the disparity test statistics, a more general class of goodness-of-fit test statistics, which include the power divergence statistics and are based on the minimum disparity estimation approach of Lindsay (1994). A disparity is characterized by a function  $G(\cdot)$ , which gives geometrical insight on the behavior of the disparities in controlling the "outliers" and "inliers", representing departures from the null in opposing directions.

The present paper is motivated by the observation of C&R that there is a reverse order hierarchy in the powers of the goodness-of-fit tests within the power divergence family for the "bump" alternatives compared to the "dip" alternatives for the equiprobable null hypothesis. Bump alternatives are those where  $k-1$  cells of a multinomial with  $k$  cells have equal probability, while the remaining cell has a higher probability than the rest. Dip alternatives are similar except for the fact that the one remaining cell has lower probability than the rest. In this paper we try to explain why the above behavior of the power divergence test statistics are natural, and make a preliminary attempt to provide some new tests with reasonably high power at both kinds of alternatives.

Three new sets of goodness-of-fit test statistics are considered. The first is based upon “penalized versions” of the power divergence statistics, while the second represents a judicious combination of the members of the power divergence statistics. The third method is based on entirely new families of disparities sensitive to both kinds of departures.

In Section 2 we give a description of the disparity test statistics as well as introduce the power divergence family of C&R and the blended weight Hellinger distance family. The equiprobable null along with the dip and bump alternatives are also described in this section. In Section 3 the new test statistics are proposed and their usefulness is demonstrated through graphical studies. Section 4 provides a small comparative study where the performance of some of the new tests are compared to the present standard. The last section contains concluding remarks.

## 2. The equiprobable null hypothesis, the alternatives and the disparity test statistics

For a sequence of  $n$  observations on a multinomial distribution with probability vector  $\pi = (\pi_1, \dots, \pi_k)$ ,  $\sum_{i=1}^k \pi_i = 1$ , let  $\mathbf{X} = (x_1, \dots, x_k)$  denote the observed frequencies for  $k$  categories and  $p = (p_1, p_2, \dots, p_k) = (x_1/n, \dots, x_k/n)$  denote the observed proportions. One is often interested in a simple null hypothesis such as

$$H_0 : \pi_i = \pi_{0i}, \quad i = 1, 2, \dots, k \quad (1)$$

where  $\pi_{0i}$ ,  $i = 1, \dots, k$  are known constants. This completely specifies the null hypothesis. In particular, the equiprobable null hypothesis for this set up is obtained when one uses  $\pi_{0i} = 1/k$  for all  $i$ .

For the equiprobable null hypothesis, consider the following set of alternatives given by

$$H_1 : \pi_i = \begin{cases} \{1 - \eta/(k-1)\}/k, & i = 1, 2, \dots, (k-1), \\ (1 + \eta)/k, & i = k \end{cases} \quad (2)$$

where the value of  $\eta$  is between  $-1$  and  $k-1$ . Note that for  $\eta > 0$  a bump alternative and for  $\eta < 0$  a dip alternative is obtained. The ability of a statistic to discern the veracity of the null hypothesis depends on the sensitivity of the statistic to deviations from the null hypothesis. We will distinguish between the two kinds of deviations in this context. “Outliers” will represent those cells where  $p_i > \pi_{0i}$ ; these are the cells which have more cases than predicted. The “inliers”, on the other hand, represent those cells with fewer cases than predicted.

Let  $G$  be a strictly convex thrice differentiable nonnegative function on  $[-1, \infty)$  with  $G(0) = 0$ ,  $G^{(1)}(0) = 0$  and  $G^{(2)}(0) = 1$ , where  $G^{(i)}$  denotes the  $i$ -th derivative of  $G$ . Suppose that  $G^{(3)}(0)$  is finite and  $G^{(3)}$  is continuous at 0. Then the disparity  $\rho_G(p, \pi_0)$  between  $p$  and  $\pi_0$  defined in (1) is given by

$$\rho_G(p, \pi_0) = \sum_{i=1}^k G\left(\frac{p_i}{\pi_{0i}} - 1\right) \pi_{0i} = \sum_{i=1}^k G(\delta_i) \pi_{0i}, \quad \delta_i = (\pi_{0i}^{-1} p_i - 1). \quad (3)$$

For a disparity  $\rho_G(\cdot, \cdot)$  consider  $D_{\rho_G}(p, \pi_0) = D_{\rho_G} = 2n\rho_G(p, \pi_0)$  as a test statistic for the simple null hypothesis defined in (1). We will call  $\delta_i$  the “Pearson residual” at cell  $i$ . For a positive  $\delta_i$  the  $i$ -th cell is an outlier and for negative  $\delta_i$  it is an inlier.

The power divergence test statistics  $2nI^\lambda(p, \pi_0)$  are generated by the power divergence family

$$I^\lambda(p, \pi_0) = \sum_{i=1}^k \left[ \frac{p_i}{\lambda(\lambda+1)} \left\{ \left( \frac{p_i}{\pi_{0i}} \right)^\lambda - 1 \right\} + \frac{\pi_{0i} - p_i}{\lambda+1} \right], \quad \lambda \in \mathbb{R}, \quad (4)$$

corresponding to

$$G(\delta) = \frac{(\delta+1)^{(\lambda+1)} - (\delta+1)}{\lambda(\lambda+1)} - \frac{\delta}{\lambda+1}. \quad (5)$$

In particular the Pearson chi-square statistic  $D_{\text{PCS}}$  is generated by the function  $G(\delta) = \frac{1}{2}\delta^2$  (i.e.  $\lambda = 1$ ), and the Hellinger distance statistic  $D_{\text{HD}}$  corresponds to  $G(\delta) = 2[(\delta+1)^{\frac{1}{2}} - 1]^2$  (i.e.  $\lambda = -1/2$ ). The likelihood ratio chi-square statistic is generated by  $I^0$ , which corresponds to the limiting case of the form in (4) as  $\lambda \rightarrow 0$ . C&R showed that the test statistic  $2nI^\lambda$  has an asymptotic  $\chi^2(k-1)$  distribution under the simple null hypothesis  $H_0$ , for all  $\lambda \in \mathbb{R}$ . Basu and Sarkar (1994) generalised this to show that all disparity test statistics  $2n\rho_G$  have this same asymptotic distribution under the null.

The blended weight Hellinger distance family  $\{BWH D_\tau, 0 \leq \tau \leq 1\}$  defined by

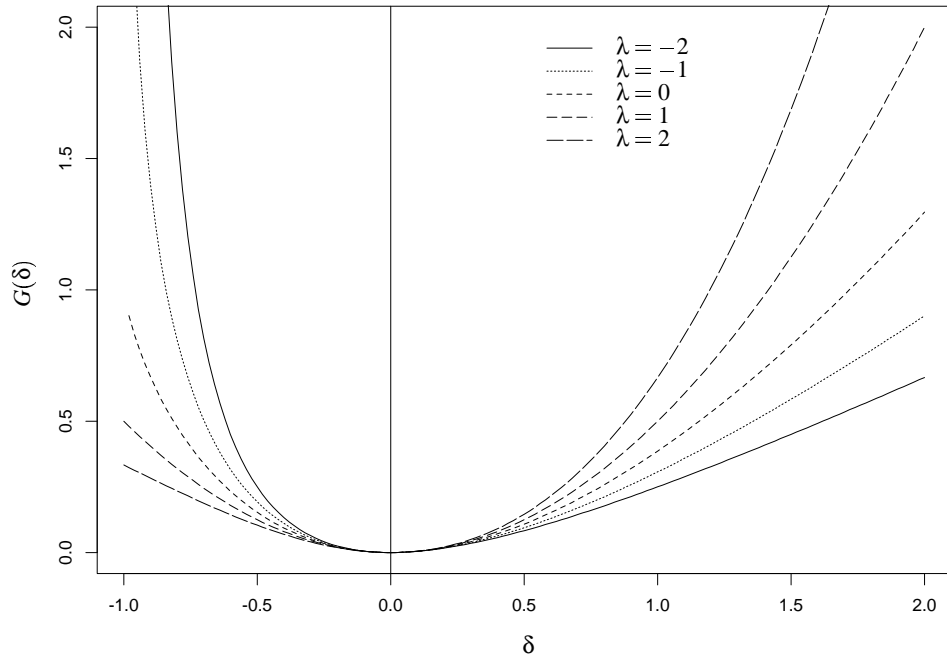
$$BWH D_{\tau}(p, \pi_0) = 2^{-1} \sum_{i=1}^k \left\{ \frac{p_i - \pi_{0i}}{\tau(p_i)^{\frac{1}{2}} + (1-\tau)(\pi_{0i})^{\frac{1}{2}}} \right\}^2 \quad (6)$$

corresponds to

$$G(\delta) = 2^{-1} \left\{ \delta / [\tau(\delta+1)^{1/2} + (1-\tau)] \right\}^2.$$

The (twice, squared) Hellinger distance is a member of  $\{BWH D_{\tau}\}$  with  $\tau = \frac{1}{2}$ .

C&R (1984, Table 2) have determined the exact power of the disparity test statistics for various values of  $\eta$  based on different members of the power divergence family, for the specific case  $n = 20$ ,  $k = 4$ , and significance level  $\alpha = 0.05$ . The exact powers are calculated for the appropriately randomized test of size  $\alpha = 0.05$  by enumerating all possible samples and calculating the appropriate critical value by determining the probabilities of these samples under the null. The power is then easily evaluated by determining the probabilities of the samples under the alternative. The results of C&R as well as Read (1984) show that for  $\eta > 0$  the exact power of the tests increases with  $\lambda$ , while for  $\eta < 0$  the exact power decreases with  $\lambda$ .



**Fig. 1.** Plot of the  $G(\cdot)$  for various members of power divergence family.

Thus the sensitivity of a disparity test statistic depends on how the defining  $G(\cdot)$  function treats the outliers and inliers. In Figure 1 we present the  $G(\cdot)$  functions for several members of the power divergence family. For large positive values of  $\lambda$  the statistics are fairly flat on the negative side of the  $\delta$ -axis but curve away rapidly on the positive side. Thus the test statistics for large positive values of  $\lambda$  are strongly sensitive to outliers, while presenting a relatively dampened response to inliers. The opposite is true for large negative values of  $\lambda$ . As a result, large positive values of  $\lambda$  lead to high power against bump alternatives while being poor against dip alternatives. However, the findings are reversed for large negative values of  $\lambda$ .

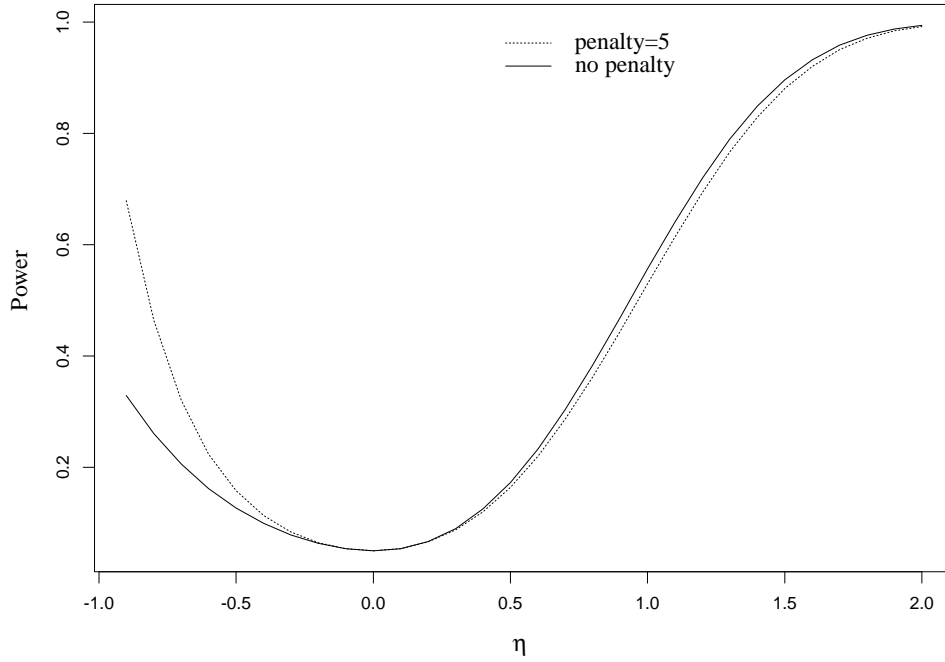


Fig. 2. Comparison of power functions for power divergence test statistic with  $\lambda = 2$  with and without penalty ( $h = 5$ ).

### 3. Proposed test statistics

#### 3.1. The penalized divergence statistics

In this section we propose a class of divergences obtained by introducing an empty cell penalty to the disparities within the class of power divergences. Recall that the disparities with large positive values of  $\lambda$  are sensitive against outliers but not against inliers. The empty cell penalty introduced here makes these disparities simultaneously sensitive to empty cells which are the extreme cases of inliers. Formally, we rewrite the power divergence family defined in (4) as

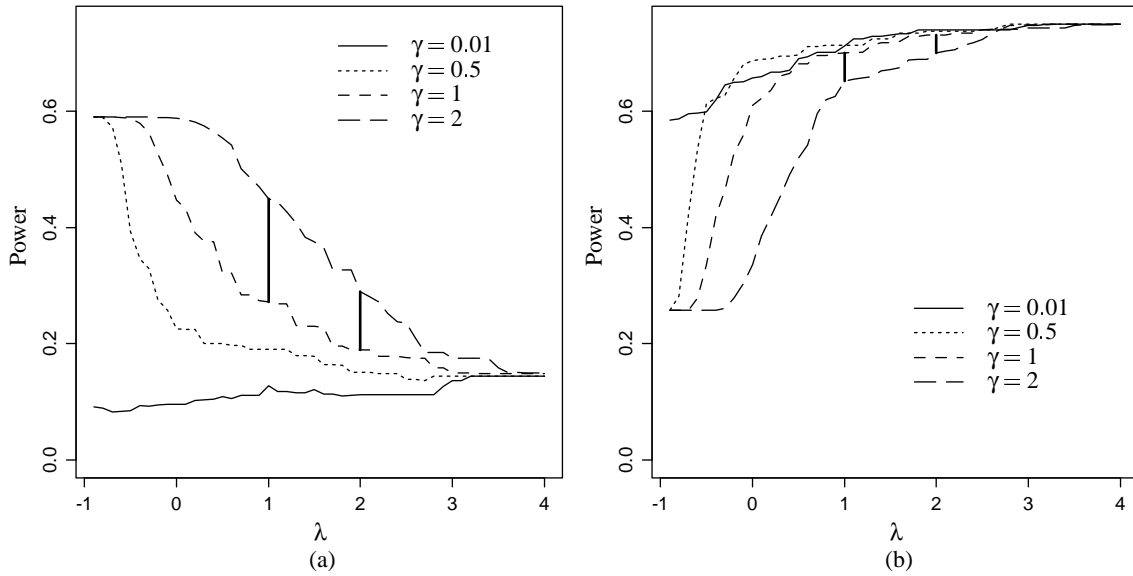
$$I^\lambda(p, \pi_0) = \sum_{p_i > 0} \left[ \frac{p_i}{\lambda(\lambda + 1)} \left\{ \left( \frac{p_i}{\pi_{0i}} \right)^\lambda - 1 \right\} + \frac{1}{\lambda + 1} (\pi_{0i} - p_i) \right] + \frac{1}{\lambda + 1} \sum_{p_i = 0} \pi_i .$$

Note that ordinarily the disparity puts the weight  $1/(\lambda + 1)$  on the empty cells (cells with  $p_i = 0$ ). For large positive values of  $\lambda$  this weight is fairly small. An artificial empty cell penalty can hike up the weight of the empty cells to a suitably large value so that it increases the sensitivity of the statistics to dip alternatives. Thus we consider the penalized power divergence family given by

$$I_h^\lambda(p, \pi_0) = \sum_{p_i > 0} \left[ \frac{p_i}{\lambda(\lambda + 1)} \left\{ \left( \frac{p_i}{\pi_{0i}} \right)^\lambda - 1 \right\} + \frac{1}{\lambda + 1} (\pi_{0i} - p_i) \right] + h \sum_{p_i = 0} \pi_i$$

where  $h$  represents the penalty weight. One can use  $2nI_h^\lambda(p, \pi_0)$  as the goodness-of-fit test statistic for testing the null hypothesis (1). Ideal choices would be divergences with large positive values of  $\lambda$  used in conjunction with large positive values of  $h$ . Following Park et al. (2001, Theorem 2.1), one can show that  $2nI_h^\lambda(p, \pi_0)$  has an asymptotic  $\chi^2(k - 1)$  distribution under the null hypothesis. Penalized divergences have also been used by Harris and Basu (1994), Basu and Basu (1998), and Park et al. (2001) to make the divergences *less* sensitive to empty cells, thereby increasing the robustness of their parameter estimation properties.

For illustration we computed the exact powers for the equiprobable null hypothesis with  $n = 20$ ,  $k = 4$ , and significance level  $\alpha = 0.05$ , for the power divergence statistic with  $\lambda = 2$ , as well as for its penalized version with



**Fig. 3.** The effect of the penalty on the power at two fixed alternatives for the power divergence test statistics.

penalty weight  $h = 5$ . The powers of the statistics for values of  $\eta$  between  $-1$  and  $2$  are plotted in Figure 2. Notice that the penalty clearly leads to a significant increase in power for large negative values of  $\eta$  without any appreciable loss in power for positive values of  $\eta$ .

Figure 3 gives another graphical illustration of the effect of the penalty where we plot the powers for different values of the penalty weight  $h$  for the equiprobable null hypothesis with  $n = 20$ ,  $k = 5$ ,  $\alpha = 0.05$ , and for two specific values of  $\eta$  determining two alternatives of the opposite type. However in this case, instead of choosing the penalties to be preassigned specific values, we have chosen the penalty weight  $h$  to be equal to  $\gamma/(\lambda + 1)$ , a scale factor of the natural weight to the empty cell for the disparity. We have used  $\gamma = 0.01, 0.5, 1, 2$  in our graphs. Figure 3a exhibits the powers of these penalized statistics for  $\eta = -0.9$ , which gives us some idea of the change in power due to the penalty effect for this value of  $\eta$ . For example, for  $\lambda = 1$  and  $\lambda = 2$ , the segment of the vertical lines along these values of  $\lambda$  between  $\gamma = 1$  and  $\gamma = 2$  represent the increase in power for these statistics at  $\eta = -0.9$  when the penalty weights are double their natural weights. Notice also that the increase in power between  $\gamma = 1$  and  $\gamma = 2$  is negligible for very large values of  $\lambda$ , the reason being that for such values of  $\lambda$  the ordinary weights of the empty cells are so small that even doubling them has little effect in terms of improving the power.

Figure 3b illustrates the effect of the penalty under identical set ups as in Figure 3a but now the value of  $\eta$  is  $1.5$ . In this case the line corresponding to  $\gamma = 1$  lies above the line for  $\gamma = 2$ , exhibiting that there is a loss in power due to doubling the penalty weight. Again the vertical line segments for  $\lambda = 1$  and  $\lambda = 2$  between these two values of  $\gamma$  quantify this decrease in power, but clearly the losses in these situations are substantially smaller compared to the gain for the dip alternative in Figure 3a. A comparison of Figures 3a and 3b show that in this case roughly for values of  $\lambda$  between  $1/2$  and  $2$  the application of penalty is probably meaningful and the gains outweigh the losses even by conservative estimates.

### 3.2. The combined divergence statistics

Another option for improved power is to choose a disparity test statistic where the  $G(\cdot)$  function is the combination of the functions of two different disparities so that the test is sensitive for both dip and bump alternatives. Recall that disparities corresponding to large positive values of  $\lambda$  are highly sensitive to outliers (corresponding to  $\delta > 0$ ), while those corresponding to large negative  $\lambda$  are highly sensitive to inliers (corresponding to  $\delta < 0$ ). Thus, for example, if one chooses a  $G(\cdot)$  so that it matches the  $G$  function for  $\lambda = -1$  on the negative side of the axis and the corresponding function of  $\lambda = 2$  on the positive side, the statistic is obviously going to be strongly influenced by both inliers and

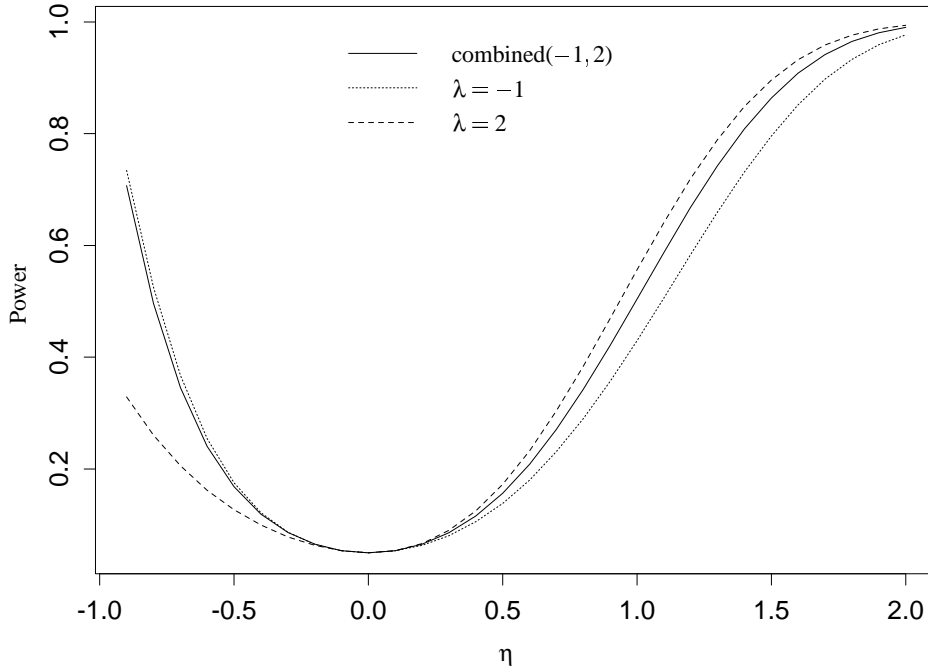


Fig. 4. Power comparison for a combined divergence statistic with simple power divergence statistics.

outliers (see Figure 1). Technically such combinations amount to choosing a combined divergence  $\rho_G(p, \pi)$  defined by the function  $G$  such that

$$G(\delta) = \begin{cases} G_1(\delta), & \text{if } \delta \leq 0 \\ G_2(\delta) & \text{if } \delta > 0 \end{cases}$$

where  $G_1$  and  $G_2$  are convex functions satisfying all the properties mentioned in Section 2. Notice that the combined  $G$  itself is a convex function satisfying  $G(0) = 0$ ,  $G^{(1)}(0) = 0$ , and  $G^{(2)}(0) = 1$ .

For a combined divergence  $\rho_G(\cdot, \cdot)$  consider  $D_{\rho_G} = 2n\rho_G(p, \pi_0)$  as a test statistic for the simple null hypothesis in (1). The following theorem, proved in the appendix, shows that the test statistic for combined divergences have the same asymptotic  $\chi^2$  distribution when the null hypothesis is true.

**Theorem 1** *The test statistic  $D_{\rho_G}$  corresponding to the combined divergence  $\rho_G$  has an asymptotic  $\chi^2(k-1)$  distribution under the null hypothesis in (1).*

As an illustration we present the following case. In Figure 4 we present the power for the combination where  $G_1$  corresponds to the  $G$  function of  $\lambda = -1$ , while  $G_2$  corresponds to the  $G$  function of  $\lambda = 2$ . Notice that the combined divergence test statistic is quite close to the best cases (among  $2nI^{-1}$  and  $2nI^2$ ) in terms of power for most values of  $\eta$ , while being substantially better than the worst cases for all the alternatives. In particular the power of the combined divergence tests are very close to the best case for negative values of  $\eta$ .

### 3.3. The $PBHM_\tau$ statistics

We now consider a third set of goodness-of-fit tests expected to perform well for both dip and bump alternatives. The divergences used here are mixtures of the Pearson's chi-square with members of the blended weight Hellinger distance ( $BWHD_\tau$ ) resulting in the  $PBHM_\tau$  (Pearson-blended Hellinger Mixture) family indexed by the parameter  $\tau$ . The relevant  $G$  function for this mixture disparity is

$$G_{PBHM}(\delta) = \tau \frac{\delta^2}{2} + (1 - \tau) \frac{\delta^2}{2[\tau\sqrt{\delta+1} + (1 - \tau)]^2}, \quad \tau \in (0, 1).$$

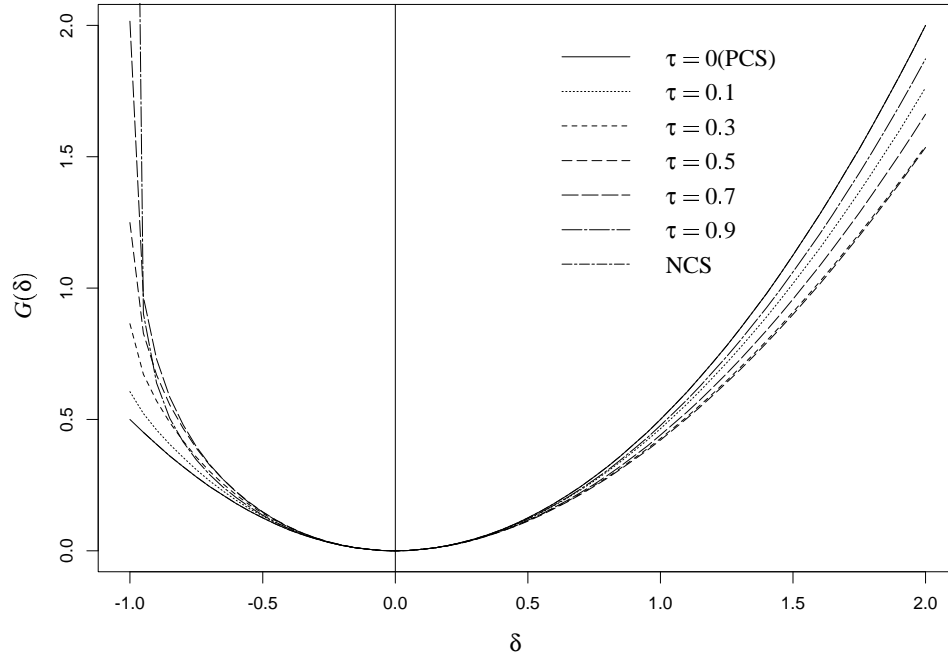


Fig. 5. Plot of the  $G(\cdot)$  function for various members of  $PBHM$  family.

Figure 5 presents the graphs of the  $G$  functions for several members of this family, together with those for the Pearson's and Neyman's modified chi-square. The graphs demonstrate the sensitivity of the new mixture disparities to both outliers and inliers.

For illustration, the exact powers of the  $2nPBHM_{\tau}(p, \pi_0)$  statistic, with  $\tau = 0.5$  are computed with  $n = 20, k = 4$  and significance level  $\alpha = 0.05$  for different values of  $\eta$  and presented in Figure 6 together with the power function of the Pearson's chi-square statistic. Once again notice that the  $2nPBHM_{\tau}$  disparity test statistic shows a comparatively larger increase in power for large negative values of  $\eta$ , together with a relatively smaller loss in power for large positive values of  $\eta$ .

Since the  $G$  functions of the  $PBHM_{\tau}$  disparities satisfy all the properties listed in Section 2, the  $2nPBHM_{\tau}$  statistics have asymptotic chi-square distributions for each  $\tau$ .

### 3.4. Composite null hypothesis

In this paper we have mostly focused on the simple null hypothesis (more specifically on the equiprobable null hypothesis) so that the concepts are easily explained. However it is not difficult to show that the results extend to the case of the composite hypothesis, there the cell probabilities are a function of a parameter  $\theta$  of dimension  $s < k - 1$ , under the regularity conditions of Birch (1964), and provided a BAN (best asymptotically normal) estimator of  $\theta$  is used. In particular, the conclusions of Theorem 4.3 of Basu and Sarkar (1994) remain valid under the above conditions for all the three classes of new disparities developed in Sections 3.1-3.3. The proofs are straightforward and are not included here so as to retain the applied focus of the paper; however, we will present a data example with a composite null hypothesis in the following section.

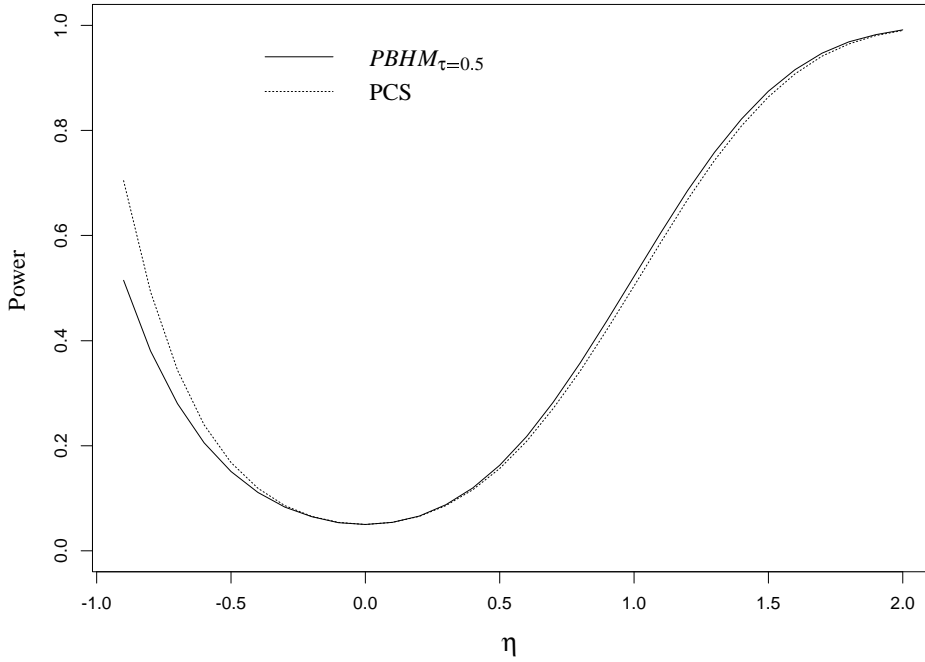


Fig. 6. Power comparison of  $PBHM_{\tau=0.5}$  with Pearson's  $\chi^2$ .

#### 4. A comparison of some of the test statistics

In this paper we have suggested three classes of new goodness-of-fit statistics. In terms of individual test statistics, this amounts to millions of choices. Precise recommendations about their use will require extensive future research including large scale comparisons involving several scenarios; however, for the benefit of the applied statistician whose final interest would be in the choice of the particular statistic to be used, we present some limited comparisons which could provide some initial indication of the possible test statistics that could serve as reasonable alternatives to the currently available tests. To do this we provide one set of exact power comparisons with several test statistics, and also consider a couple of data examples to contrast the different methods. As C&R and Read and Cressie (1988) recommended the  $I^{\frac{2}{3}}$  divergence to be used as a compromise candidate, we use this statistic for the basis of comparison.

**Exact power comparisons:** For the exact power comparisons we retain the set up considered in most of our numerical studies, i.e. we use the equiprobable null hypothesis, use the dip and bump alternatives as functions of  $\eta$ , and use sample size  $n = 20$ , and number of groups  $k = 4$ . The statistics we use in this comparison correspond to (a)  $\lambda = 2/3$ , (b) penalized statistic with  $\lambda = 2$  and penalty weight 1, (c) penalized statistic with  $\lambda = 1$  and penalty weight 5, (d) combined statistic for  $\lambda = -1$  and  $\lambda = 2$ , (e) combined statistic for  $\lambda = -2$  and  $\lambda = 1$ , and (f) the  $PBHM_{\tau}$  statistic with  $\tau = 0.5$ . The calculated exact powers at level  $\alpha = 0.05$  are graphically presented in Figure 7. It appears that except for the combined ( $\lambda = -2, \lambda = 1$ ) statistic, all the other five statistics are very close in terms of exact attained power in this case. While this does not show either of the other statistics to be actually “better” than the  $\lambda = 2/3$  case, it does show that there are many other compromise candidates with practically identical performances in this case. Next we present a couple of data examples. The examples were not restricted to equiprobable null types, but are of more complex nature representing the complexities of real life. Since the basic premise of our construction is to make the test statistics more sensitive to deviations from the null irrespective of the nature of the null, we expect the new statistics to exhibit their high sensitivity with these real data as well.

**Data Example 1:** This dataset is taken from Agresti (1990; Table 3.10, page 72). A total of 182 psychiatric patients on drugs were classified according to their diagnosis. The frequency distribution of the diagnosis is given in Table 1.

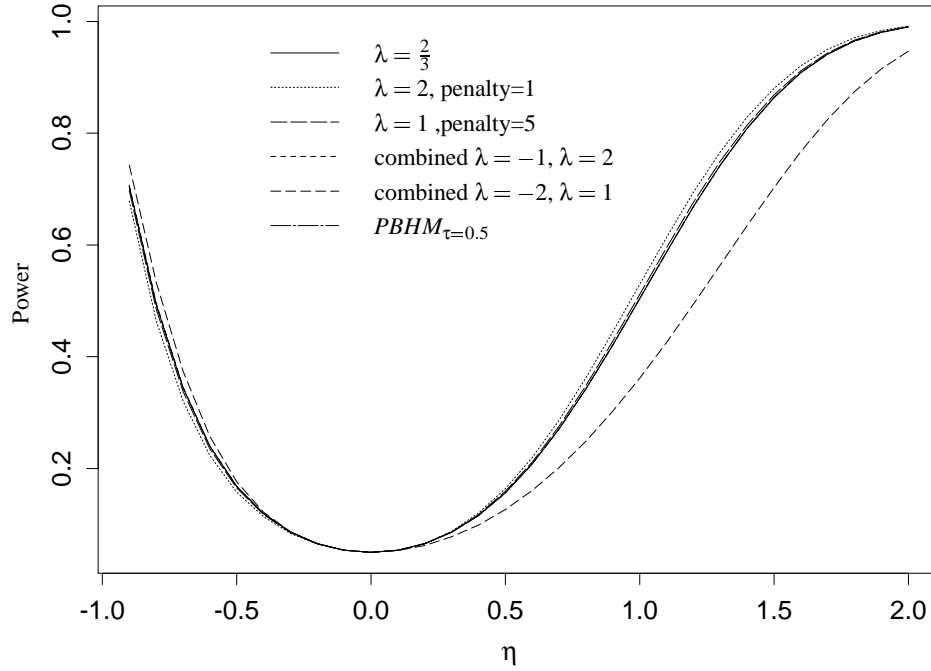


Fig. 7. Power comparison of some of the new statistics with  $2nI^{\frac{2}{3}}$ .

**Table 1.** Frequency distribution of diagnosis of psychiatric patients

Diagnosis	Frequency
Schizophrenia	105
Affective Disorder	12
Neurosis	18
Personality Disorder	47
Special Symptoms	0

Consider testing the null hypothesis where the probability vector is given by  $\pi = (0.56, 0.06, 0.09, 0.25, 0.04)$ . The chi-square critical value at 4 degrees of freedom, and at level of significance 0.05 is given by 9.488. The ordinary statistics for  $\lambda = 2, 1$ , and  $2/3$  cannot reject the null hypothesis with this critical value, with the corresponding test statistics being 5.273, 7.689, and 9.142 respectively. However, for the penalized statistics (notice that the data set contains one empty cell) corresponding to  $\lambda = 1$  and  $\lambda = 2$  with penalty weight  $h = 1$  the test statistics are 14.969 and 14.980 respectively, which rejects the null hypothesis comfortably.

Similarly, the combined statistics for  $(\lambda = -1/2, \lambda = 1)$  and  $(\lambda = -1/2, \lambda = 2)$  combinations are, respectively, 29.529 and 29.540. Once again they comfortably reject the null hypothesis unlike the ordinary statistics corresponding to  $\lambda = 2$  and  $\lambda = 1$ . Notice that we have not computed the  $(\lambda = -1, \lambda = 2)$  or  $(\lambda = -2, \lambda = 1)$  combinations. This is because if there are empty cells, any combination component for the inliers with  $\lambda = -1$  or less will make the statistic infinite. Of course technically we reject for such statistics, but it is not very informative.

The  $PBHM_{0.5}$  statistic for this data set is 18.602. For this example, therefore, the new statistics looked at lead to rejection, whereas the  $\lambda = 2/3$  statistic as well as some of the other ordinary ones fail to reject. With an expected frequency of over 7 in the last cell being matched against an observed frequency equal to zero, the null hypothesis is certainly in doubt in our opinion.

**Data Example 2:** Next we consider the time passage example data (Read and Cressie 1988; pp 12-16) which studies the relationships between life stresses and illnesses in Oakland, California. The data is in the form of an 18 cell multinomial, with the frequencies representing the total number of respondents for each month who indicated one stressful event between 1 and 18 months prior to interview (see Read and Cressie for more details). The null hypothesis  $H_0 : \pi_i = 1/18, i = 1, \dots, 18$  is rejected for each of the ordinary test statistics. However if one considers a loglinear time trend model  $H_0 : \log(\pi_i) = \vartheta + \beta i, i = 1, \dots, 18$ , the model fit appears to be much better. Notice that the null does not completely specify the probability structure any more. Expected frequencies on the basis of estimates of  $\vartheta$  and  $\beta$  obtained using maximum likelihood are given in Read and Cressie (1988; Table 2.2). The test statistics are now compared with the critical value of a chi-square with 16 degrees of freedom (rather than 17), and the critical value at level of significance 0.05 is 26.296.

The  $\lambda = 2/3$  statistic equals 23.076, and fails to reject the null (in fact so does all the statistics with  $\lambda$  between 0 and 3. However the combined statistics for  $(\lambda = -1, \lambda = 2)$ , and  $(\lambda = -2, \lambda = 1)$  are 35.271 and 44.840 respectively, with the null being rejected in both cases. The  $PBHM_{0.5}$  statistic equals 24.629, and although larger than the  $\lambda = 2/3$  statistic, this also fails to reject the null. On the whole this analysis shows that the time trend model is on the borderline of being significant, but some of the newer statistics are more likely to reject the null than the compromise suggested by C&R.

Notice that the penalized statistics are not meaningful in this case because this data has no empty cell.

## 5. Concluding Remarks

In this paper we have presented some new candidates for goodness-of-fit testing in multinomial models. A final and definitive recommendation will require much deeper research, but we feel that the initial indications are promising enough for some of the proposed statistics to be further pursued. It does appear that some of these tests are competitive in comparison to the compromise suggested by C&R, and some of them are more sensitive than  $I^{\frac{2}{3}}$  in certain cases.

What is the cost incurred in making the test statistics more sensitive to both kinds of deviations? While all the test statistics considered by us have asymptotic chi-square distributions, we believe that the cost of making the test statistics more sensitive is paid through a slower convergence to the asymptotic chi-square limit. We do not perceive this to be a problem for large sample sizes, but for small to moderate sample sizes the level of our tests may be a little off from the nominal levels if one uses the chi-square critical values. At present we are investigating better small sample approximations to the null distributions in the spirit of Read (1984). However, exact critical values in small samples and simulated critical values in moderate samples can be easily calculated for the simple null. The authors will be happy to provide codes for determination of the above so that one can perform accurate level  $\alpha$  goodness-of-fit tests when applying the proposed methods (which may be too inaccurate with  $\chi^2$  critical values in small samples). The Splus codes are readily available from the site <http://www.stat.psu.edu/~surajit/goodness/>.

In conclusion, we emphasize again that this is a small preliminary study involving only a limited number of scenarios. To determine the general scope of the proposed statistics more extensive studies are necessary which the authors hope to undertake in future.

## Appendix

**Proof of Theorem 1.** Consider the combined divergence  $\rho_G(p, \pi)$  defined by

$$G(\delta) = \begin{cases} G_1(\delta), & \text{if } \delta \leq 0 \\ G_2(\delta) & \text{if } \delta > 0 \end{cases}$$

where  $G_1$  and  $G_2$  are convex functions satisfying the properties listed in Section 2. For a combined divergence  $\rho_G(p, \pi)$  let  $D_{\rho_G} = 2n\rho_G(p, \pi_0)$  be the test statistic for the hypothesis defined in (1). By a first order Taylor series expansion of the test statistic (as a function of  $p_i$  around  $\pi_{0i}$ ) we get

$$\begin{aligned} \sum_{i=1}^k G((p_i - \pi_{0i})/\pi_{0i})\pi_{0i} &= \sum_{i=1}^k G(0)\pi_{0i} + \sum_{i=1}^k (p_i - \pi_{0i})G^{(1)}(0) + \sum_{i=1}^k \frac{1}{2}(p_i - \pi_{0i})^2 G^{(2)}(0)\pi_{0i}^{-1} + \\ &\quad \sum_{i=1}^k \frac{1}{6}(p_i - \pi_{0i})^3 G^{(3)}(\pi_{0i}^{-1}\xi_i - 1)\pi_{0i}^{-2} \\ &= S_1 + S_2 + S_3 + S_4, \quad \text{say} \end{aligned}$$

where  $\xi_i$  lies on the line segment joining  $p_i$  and  $\pi_{0i}$ . Note that (a)  $G_1(0) = G_2(0) = 0$  and hence  $G(0) = 0$ , (b)  $G^{(2)}$  exists everywhere and  $G_1^{(2)}(0) = G_2^{(2)}(0) = G^{(2)}(0) = 1$ , and (c) both  $p_i$  and  $\pi_{0i}$  are nonnegative terms that sum to 1 over  $i$ . The first two terms  $S_1$  and  $S_2$  are, therefore, equal to 0. Next note that

$$\begin{aligned} 6nS_4 &= \sum_{i=1}^k n(p_i - \pi_{0i})^3 [G^{(3)}(\pi_{0i}^{-1}\xi_i - 1)\pi_{0i}^{-2}] \\ &\leq \left\{ \sum_{i=1}^k n(p_i - \pi_{0i})^2 \right\} \left\{ \sup_i |p_i - \pi_{0i}| \right\} \\ &\quad \times \left\{ \sup_i \pi_{0i}^{-2} \right\} \left\{ \sup_i G^{(3)}(\pi_{0i}^{-1}\xi_i - 1) \right\}. \end{aligned}$$

$\sup_i [\pi_{0i}]^{-2}$  is bounded,  $\sup_i |p_i - \pi_{0i}| = o_p(1)$ ,  $\sum_{i=1}^k n(p_i - \pi_{0i})^2 = O_p(1)$ ,  $(\xi_i - \pi_{0i}) = o_p(1)$  for every  $i$ . Notice also that  $\delta_i = (\pi_{0i}^{-1}p_i - 1)$  and  $(\pi_{0i}^{-1}\xi_i - 1)$  have the same sign and by the assumptions  $G_1^{(3)}(0)$  and  $G_2^{(3)}(0)$  are bounded and  $G_1^{(3)}$  and  $G_2^{(3)}$  are continuous at 0. Hence  $G^{(3)}(\pi_{0i}^{-1}\xi_i - 1) = O_p(1)$  for all  $i$  and therefore,  $6nS_4 = o_p(1)$ . Then the result follows by noting that

$$2nS_3 = n \sum_{i=1}^k \pi_{0i}^{-1} (p_i - \pi_{0i})^2$$

is the Pearson chi-square statistic whose asymptotic  $\chi^2(k-1)$  distribution under the simple null hypothesis is well known.

## References

- [1] Agresti, A. (1990). *Categorical Data Analysis*. John Wiley & Sons, New York.
- [2] Basu, A. and Basu, S. (1998). Penalized minimum disparity methods for multinomial models. *Statistica Sinica*, **8**, 841–860.
- [3] Basu, A. and Sarkar, S. (1994). On disparity based goodness-of-fit tests for multinomial models. *Statist. Probab. Lett.*, **19**, 1994, 307–312.
- [4] Birch, M.W. (1964). A new proof of the Pearson–Fisher Theorem. *Ann. Math. Statist.*, **35**, 817–824.
- [5] Chapman, J.W. (1976). A comparison of the  $\chi^2$ ,  $-2 \log R$ , and the multinomial probability criteria for significance testing when expected frequencies are small. *J. Amer. Statist. Assoc.*, **71**, 854–863.
- [6] Cochran, W.G. (1952). The  $\chi^2$  test of goodness-of-fit. *Ann. Math. Statist.*, **23**, 315–345.
- [7] Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. B*, **46**, 440–464.
- [8] Harris, I. R. and Basu, A. (1994). Hellinger distance as a penalized log likelihood. *Commun. Statist. Comput. Simul.*, **23**, 1097–1113.
- [9] Hoeffding, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.*, **36**, 369–408.
- [10] Koehler, K.J. and Larntz, K. (1980), An empirical investigation of goodness-of-fit statistics for sparse multinomials, *J. Amer. Statist. Assoc.*, **75**, 336–344.
- [11] Larntz, K. (1978). Small sample comparisons of exact levels of chi-squared goodness-of-fit statistics. *J. Amer. Statist. Assoc.*, **73**, 253–263.
- [12] Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Statist.*, **22**, 1081–1114.
- [13] Moore, D.S. and Spruill, M. C. (1975). Unified large–sample theory of general chi-squared statistics for tests of fit. *Ann. Statist.*, **3**, 599–616.
- [14] Park, C., Basu, A. and Harris, I. R. (2001). Tests of hypothesis in multiple samples based on penalized disparities. *J. Korean. Statist. Soc.*, **30**, 347–366.
- [15] Read, T. R. C. (1984). Small sample comparisons for power divergence goodness-of-fit statistics. *J. Amer. Statist. Assoc.*, **79**, 929–935.
- [16] Read, T. R. C. and Cressie, N. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York.
- [17] Watson, G.S. (1959). Some recent results in chi-square goodness-of-fit tests. *Biometrics*, **15**, 440–468.
- [18] West, E.N. and Kempthorne, O. (1972). A comparison of  $\chi^2$  and likelihood ratio tests for composite alternatives. *J. Statist. Comput. Simul.*, **1**, 1–33.