Electronic Journal of Statistics
Vol. 5 (2011) 1846–1875
ISSN: 1935-7524
DOI: 10.1214/11-EJS658

Multinomial goodness-of-fit tests under inlier modification

Abhijit Mandal* and **Ayanendranath Basu**

Bayesian and Interdisciplinary Research Unit
Indian Statistical Institute, Kolkata, India
e-mail: abhijit_v@isical.ac.in; ayanbasu@isical.ac.in

Abstract: The Pearson's chi-square and the log-likelihood ratio chi-square statistics are fundamental tools in multinomial 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 real parameter. Divergences at one end of the scale are powerful against deviations of one type while being poor against deviations of the other type. The reverse property holds for divergences at the other end of the scale. Several other families of divergences available in the literature also show similar behavior. We present several inlier control techniques in the context of multinomial goodness-of-fit testing which generate procedures having reasonably high powers for both kinds of alternatives. We explain the motivation behind the construction of the inlier modified test statistics, establish the asymptotic null distribution of the inlier modified statistics and explore their performance through simulation and real data examples to substantiate the theory developed.

AMS 2000 subject classifications: Primary 62G35; secondary 62G10.
Keywords and phrases: Goodness-of-fit, disparity, inliers, power divergence, small sample studies.

Received July 2011.

Contents

*Most of the material presented in this paper is a part of the Ph.D. thesis of the first author.

1. Introduction

The Pearson's chi-square and the log likelihood ratio statistics have long been used to perform goodness-of-fit tests in multinomial settings. Some other less frequently used goodness-of-fit test statistics like the Neyman modified chi-square statistic, the Freeman-Tukey statistic and the modified log likelihood ratio statistic are also available in the literature. Several authors have compared the performance of these test statistics; an excellent discussion of this topic together with an extensive bibliography is available in Cressie and Read (1984) [14] and Read and Cressie (1988) [29].

Cressie and Read (1984) [14] developed a rich class of goodness-of-fit test statistics called the family of power divergence statistics. All the above mentioned statistics are particular members of the power divergence family. The divergences within this family are indexed by a single parameter $\lambda \in \mathbb{R}$. Cressie and Read presented an analytical discussion of the asymptotic properties of these test statistics along with a substantial amount of numerical results for the finite sample case. The power divergence measure is a subclass of the family of *f-divergences* or *disparities* (Csiszár, 1963 [15]; Lindsay, 1994 [18]).

For the illustration of the properties of their tests under multinomial models, Cressie and Read considered the equiprobable null against the class of *bump* and *dip* alternatives at a single cell. Their findings demonstrate that the power of the test statistics increases with the parameter λ under a bump alternative; unfortunately, the power of these tests decreases with λ for a dip alternative. Since the expected deviation from the null hypothesis, if any, is unlikely to be known apriori in many practical situations, the choice of an ordinary goodness-of-fit test within the power divergence family prior to the experiment may conceivably lead to poor results. This phenomenon is not limited to the power divergence family alone. Goodness-of-fit tests based on many other families of divergences behave similarly. Our intention here is to modify the *inlier* part of the divergences generating these goodness-of-fit tests to develop new *inlier modified* test statistics which might provide better overall protection against loss in power, much in the spirit of the uniformly most powerful unbiased test against the two sided alternative under the normal and some other models, or the maxi-min philosophy.

In the context of probabilistic modeling of real data, we will denote elements of the sample space having more data than predicted under the model (here, the structure specified by the null hypothesis) as *outliers*. On the other hand, *inliers* will denote elements of the sample space having less data than is predicted by the model. Inliers have so far received less attention in the literature compared to outliers. In this paper we will show that the proper controlling of inliers can significantly alter the performance of the goodness-of-fit tests, particularly in small samples.

At this point we make it clear that we follow the nomenclature considered by Lindsay (1994) in our definition of inliers and outliers as given in the previous paragraph. One could, of course, view the inliers as outliers with a negative orientation. As our set up naturally follows from that of Lindsay we stick to the "inlier" and "outlier" notations in this paper as described above.

The rest of the paper is organized as follows. In the next section we introduce the ordinary disparity test statistics for testing goodness-of-fit in multinomial models. In Section 3, we give a qualitative justification of why the controlling of the inliers is expected to moderate the performance of the test statistics and make them less extreme. Section 4 proposes several inlier control strategies. The asymptotic null distributions of the inlier modified test statistics are derived in Section 5. In Section 6 the improved performance of the inlier modified tests are numerically demonstrated. Concluding remarks are presented in Section 7.

2. Disparity test statistics

Suppose X has a k-cell multinomial distribution with parameters n and $\pi = (\pi_1, \pi_2, \ldots, \pi_k)^T$, where $\pi_i > 0$ for $i = 1, 2, \ldots, k$ and $\sum_{i=1}^k \pi_i = 1$. Let us denote the observed proportions based on the above multinomial variable by $p = (p_1, p_2, \ldots, p_k)^T$. Consider the null hypothesis of the goodness-of-fit testing problem

$$H_0 : \pi \in \Pi_0, \tag{2.1}$$

where Π_0 represents a specified set of probability vectors that are hypothesized for π. Suppose the null hypothesis is specified by s parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_s)^T$, *i.e.* $\pi_i = f_i(\theta)$, $i = 1, 2, \ldots, k$ or equivalently $\pi = (f_1(\theta), f_2(\theta), \ldots, f_k(\theta))^T = f(\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^s$, $s \leq k - 1$ and $f : \mathbb{R}^s \to \mathbb{R}^k$. We denote the true value of θ as θ^0, whereas the true value of π will be denoted by π_0.

We define the *Pearson residual* function $\delta_i(\theta)$ for the i-th cell by the relation

$$\delta_i(\theta) = \frac{p_i - f_i(\theta)}{f_i(\theta)}, \quad i = 1, 2, \ldots, k.$$

Let G be a real-valued, thrice differentiable, strictly convex function on $[-1, \infty)$ with $G(0) = 0$. The *disparity* between the probability vectors p and $f(\theta)$ based on G is denoted by $\rho_G(p, f(\theta))$ and is defined as

$$\rho_G(p, f(\theta)) = \sum_{i=1}^{k} G(\delta_i(\theta)) f_i(\theta). \quad (2.2)$$

The function G will be called the disparity generating function of the measure defined in equation (2.2). Let $\hat{\theta}_n$ be the estimator of θ that minimizes ρ_G over $\theta \in \Theta$, provided such a minimizer exists; $\hat{\theta}_n$ will be called the *minimum disparity estimator* (MDE) of θ corresponding to ρ_G. Under differentiability of the model the estimating equation for θ is of the form

$$\sum_i A_G(\delta_i(\theta)) \nabla f_i(\theta) = 0, \quad (2.3)$$

where

$$A_G(\delta) = (1+\delta)G'(\delta) - G(\delta), \quad (2.4)$$

G' is the first derivative of G with respect to its argument, and ∇ is the gradient with respect to θ. The function A_G is called the *residual adjustment function* (RAF) of the disparity generated by G; it may be redefined, without changing the estimating properties of the disparity so that it satisfies $A_G(0) = 0$ and $A'_G(0) = 1$, where A'_G is the derivative of A_G (see Lindsay, 1994 [18] and Basu *et al.*, 1997 [7]). These two conditions are automatic if, in addition to its usual properties, the associated G function satisfies

$$G'(0) = 0, \text{ and } G''(0) = 1, \quad (2.5)$$

where G' and G'' are the indicated derivatives of G. The strict convexity of G implies that A_G is an increasing function. Also it is easily checked that given a twice differentiable increasing function A_G or a non-negative differentiable function A'_G, one can reconstruct a disparity measure ρ_G by using the function

$$G(\delta) = \int_0^{\delta} \int_0^{t} A'_G(s)(1+s)^{-1} ds\ dt. \quad (2.6)$$

For our goodness-of-fit testing problem we will consider disparity test (DT) statistics of the type

$$\min_{\theta \in \Theta} 2n\rho_G(p, f(\theta)) = 2n\rho_G(p, f(\hat{\theta}_n)). \quad (2.7)$$

Equation (2.7) also serves as the defining equation of $\hat{\theta}_n$, the minimum disparity estimator of θ under the null as generated by the disparity ρ_G. More generally, one can consider test statistics of the form $2n\rho_G(p, \hat{\pi}_0)$, where $\hat{\pi}_0$ is any BAN (best asympotically normal) estimator of π under the null; see, *e.g.,* Bishop *et al.* (1975) [13] for a description of this concept.

In particular the power divergence test statistics $2nI_\lambda(p, f(\hat{\theta}_n))$ are generated by the power divergence (PD) family (Cressie and Read, 1984 [14]) given by

$$I_\lambda(p, f(\theta)) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{k} p_i \left\{ \left(\frac{p_i}{f_i(\theta)} \right)^{\lambda} - 1 \right\}, \quad \lambda \in \mathbb{R}, \quad (2.8)$$

which has the associated G function

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

For $\lambda = 0$ and $\lambda = -1$, the divergences are defined by the continuous limits of the above expressions as $\lambda \to 0$ and $\lambda \to -1$ respectively. Several standard distances such as the Pearson's chi-square (PCS), Neyman's chi-square (NCS), Hellinger distance (HD) and likelihood disparity (LD) belong to this family for different values of the tuning parameter λ. The test statistic for the case $\lambda \to 0$ is the log likelihood ratio statistic (LDT), while the Kullback-Leibler divergence (KLD) corresponds to $\lambda \to -1$. The statistic based on the Hellinger distance is also called the Freeman-Tukey statistic.

Other subfamilies of disparities include the blended weight Hellinger distance (Basu and Lindsay, 1994 [8]; Lindsay, 1994 [18]), defined by

$$\mathrm{BWHD}_\beta(p, f(\theta)) = \frac{1}{2}\sum_{i=1}^{k} \frac{(p_i - f_i(\theta))^2}{(\beta {p_i}^{1/2} + \bar{\beta} f_i^{1/2}(\theta))^2}, \ \beta \in [0,1], \ \bar{\beta} = 1-\beta. \quad (2.9)$$

The G function for this subfamily is given by $G_\beta(\delta) = \frac{1}{2}\frac{\delta^2}{[\beta(\delta+1)^{1/2}+\bar{\beta}]^2}$. For $\beta = 0, 1/2$ and 1, this family generates the PCS, HD and NCS respectively. Another such family is the blended weight chi-square divergence (Lindsay, 1994 [18]; Shin *et al.*, 1996 [31]), defined by

$$\mathrm{BWCS}_\tau(p, f(\theta)) = \frac{1}{2}\sum_{i=1}^{k} \frac{(p_i - f_i(\theta))^2}{\tau p_i + \bar{\tau} f_i(\theta)}, \ \tau \in [0,1], \ \bar{\tau} = 1-\tau. \quad (2.10)$$

The G function for this subfamily is given by $G_\tau(\delta) = \frac{1}{2}\frac{\delta^2}{[\tau(\delta+1)+\bar{\tau}]}$. This family generates the PCS and NCS for $\tau = 0$ and 1 respectively.

Suppose the disparity $\rho_G(p, \pi) = \sum_{i=1}^{k} G(\delta_i)\pi_i$ is standardized so that $G(0) = G'(0) = 0$ and $G''(0) = 1$. Moreover, G''' is bounded and continuous in the neighborhood of zero. Then under the conditions of Birch (1964) [12] (see Read and Cressie, 1988 [29]) the following theorem holds.

Theorem 2.1. *Under the notation developed in this section, the disparity statistic $2n\rho_G(p, f(\hat{\theta}_n))$ has an asymptotic $\chi^2(k-s-1)$ distribution under the null hypothesis in (2.1), where $\hat{\theta}_n$ is the minimum disparity estimator of θ under the null hypothesis.*

The theorem is proved in Basu and Sarkar (1994) [11]; also see Zografos (1990) [33]. The essential step is to expand (2.7) in a Taylor series, and show that under the null hypothesis the statistic is separated from the Pearson's chi-square by an $o_p(1)$ term only. Basu and Sarkar also argue that $2n\rho_G(p, \hat{\pi}_0)$ will have the same asymptotic distribution as in Theorem 2.1 for any BAN estimator $\hat{\pi}_0$ of π under the null hypothesis.

In the description of this section we have used the formulation of the disparities as given by Lindsay (1994) [18]. The class of disparities has an exact one to one relationship with the class of f-divergences of Csiszár (1963) [15]. The central function ϕ in Csiszár's f-divergence notation uses the argument $\delta_i(\theta) + 1 = p_i/f_i(\theta)$, while the central function G in Lindsay's notation uses the argunent $\delta_i(\theta)$. Disparities and f-divergences are simple reformulations of each other. To be specific we have used the disparity formulation in this paper which allows us to use the geometric interpretations of Lindsay (1994). However, all the conclusions of this paper could also have been arrived at by using the f-divergence formulation. The PD, BWHD and BWCS are special cases of the family of f-divergence, just as they are special cases of the family of disparities.

3. The motivation for inlier control

In this section the motivation for the application of inlier control techniques is discussed; we also qualitatively indicate why we expect this to improve the performance of the disparity test statistics. This is the key section in developing the theme of this paper. We begin with a description of the equiprobable null hypothesis for the multinomial goodness-of-fit test, examine the performance of the different methods within some prominent families of goodness-of-fit tests in this context, and use the geometry of the structure of these test statistics to explain why the observed extreme behaviors are as expected. We follow this up with a description of how the inlier modification may be expected to provide more stable results and guard against extremes.

Let us consider a multinomial distribution with parameters n and $\pi = (\pi_1, \ldots, \pi_k)^T$, where $\pi_i > 0$ for $i = 1, 2, \ldots, k$ and $\sum_{i=1}^k \pi_i = 1$. For illustration we choose the equiprobable null hypothesis given by

$$H_0 : \pi_i = \frac{1}{k},\ i = 1, 2, \ldots, k, \tag{3.1}$$

which is to be tested against the alternative

$$H_1 : \pi_i = \begin{cases} \frac{1}{k} - \frac{\eta}{k(k-1)}, & i = 1, 2, \ldots, k-1, \\ \frac{1+\eta}{k}, & i = k, \end{cases} \tag{3.2}$$

where $\eta \in [-1, k-1]$. Under H_1 the primary violator of the null hypothesis is the last cell which represents a *bump* or a *dip* depending on whether the value η is positive or negative. If $\eta > 0$, there is a bump of magnitude η/k in the probability of the k-th cell, and the probabilities of the other cells are reduced uniformly so that the total probability adds up to unity; the reverse phenomenon is observed for $\eta < 0$ which leads to a dip in the probability of the last cell. The bump and the dip alternatives have been used by several authors including Cressie and Read (1984) [14] to illustrate the performance of different goodness-of-fit test statistics. Throughout the rest of the manuscript the symbol η will be used to indicate the deviation from the equiprobable null as quantified by the alternative in equation (3.2).

Table 1

Exact power of the randomized tests for the equiprobable null against alternative (3.2) for different values of λ and η, where $n = 20$, $k = 5$ and $\alpha = 0.1$ (the cases $\lambda = -1$ and $\lambda = 0$ are evaluated using their respective forms obtained from the general power divergence through continuity)

	η			
λ	2	1	$-.6$	$-.8$
-1.0	0.4202	0.1472	0.1833	0.3970
-0.5	0.6064	0.1640	0.1832	0.3962
-0.3	0.7551	0.2170	0.1808	0.3829
0	0.8665	0.3052	0.1650	0.3178
0.5	0.9096	0.3552	0.1442	0.2448
1	0.9190	0.3725	0.1334	0.2123
2	0.9338	0.3963	0.1137	0.1607

Table 1 presents the exact power of the disparity test statistics for various values of η based on different members of the power divergence family. Here we have taken a multinomial distribution with $n = 20$ and $k = 5$. The exact powers are calculated for the appropriately randomized test of size $\alpha = 0.1$; this involves the enumeration of all possible samples, and the calculation of the randomized critical value by determining the probabilities of these samples under the null. The results show that for a bump alternative the exact powers of the tests increase with λ, while for a dip alternative the exact powers of the tests decrease with λ. Similar results have also been observed by, among others, Cressie and Read (1984) [14].

Other families of disparities display similar characteristics. For example in the BWHD_β and BWCS_τ families (see Section 2) the power increases with β and τ for negative η and decreases with β and τ for positive η. Some indications of this behavior were presented in Basu and Sarkar (1994) [11]. For brevity we refrain from further elaborating on this phenomenon.

The sensitivity of a disparity test statistic against different types of alternatives may be described by the nature of the disparity generating function G (see Basu *et al.*, 2002 [10]). Figure 1 presents the G functions of several members of the power divergence family. We will denote the positive side of the δ-axis as the outlier side and the negative side as the inlier side. Notice that in relation to the hypotheses (3.1) and (3.2), a bump in the last cell represents an outlier while a dip indicates an inlier. For large positive values of λ the G functions are fairly flat on the negative side of the δ axis, but curve away rapidly on the positive side of δ axis. Thus the corresponding test statistics are strongly sensitive to outliers (bumps), but present a dampened response to inliers (dips). So the disparity test statistics with large positive values of λ may be expected to perform well against bump alternatives while being poor discriminators for dip alternatives. On the other hand, the results are just the opposite for the statistics with large negative values of λ. No natural disparity within the power

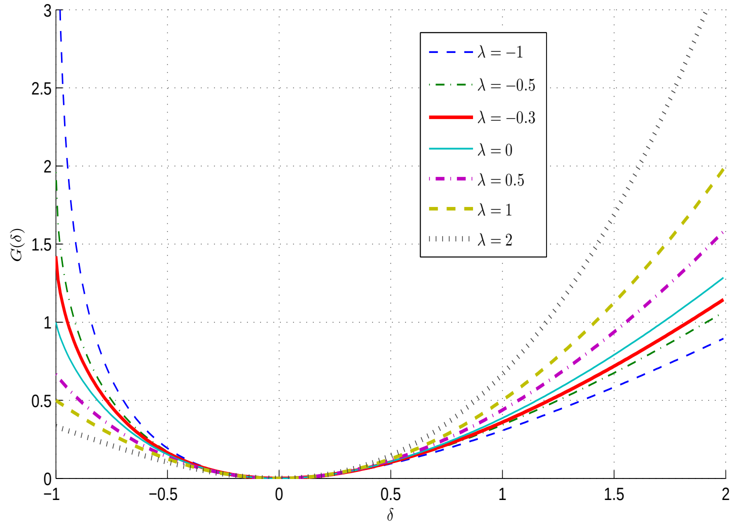

Fig 1. *G functions for different values of λ in the power divergence family.*

divergence family show high levels of sensitivity against both kinds of deviations simultaneously.

Our aim is to work with divergences which are naturally sensitive for the outlier side and appropriately modify their behavior for the inlier side so that the tests are simultaneously sensitive to both kinds of deviations. This is a delicate matter since we need to do the same without disturbing the asymptotic null distribution of the test statistics. In some of the inlier control proposals presented in the next section, the modified divergences belong to the class of disparities so that the asymptotic null distribution of the goodness-of-fit statistics based on the modified divergences follow from existing results. For the others, the asymptotic distribution of the statistics have to be freshly derived.

Our goal is to construct a G function such that it curves up rapidly on both sides of the origin so that, intuitively, the test may be expected to have good overall power at all alternatives, although it may not be best at any. For most G functions that arise naturally, sensitivity is limited to one side of the axis alone. We expect that inlier modified statistics corresponding to large positive values of λ will make the tests more sensitive on the left while preserving their sensitivity on the right. Thus inlier modified tests with large values of λ within the power divergence family (or similarly modified tests within the other families) may be expected to be effective in detecting both kind of deviations.

The inlier control techniques are also useful in the context of robust estimation based on the minimum disparity approach. However in the robust estimation problem the statistician uses the disparities to downweight the effect of large residuals by making the disparity generating function (or, equivalently, the residual adjustment function) *less* sensitive on both the inlier and outlier sides. Thus in the robust estimation case, the same inlier modification techniques are used with the opposite objective. This topic is further discussed in Mandal (2010) [19].

4. Methods for inlier control

In this section we will briefly introduce the techniques that may be used to suitably modify the inlier part of the outlier sensitive disparities. In some cases the modifications are applied directly to the disparity generating function G. In other cases the modifications are applied on the residual adjustment function A_G; subsequently we go back and recover the form of the disparity generating function using (2.6). Residual adjustment functions which dip further in the left tail produce more inlier sensitive disparities.

In this paper we will consider five different strategies of dealing with the inlier problem. The methods are based on the penalized disparities (see Harris and Basu, 1994 [16]; Basu *et al.*, 1996 [6], Basu and Basu, 1998 [5]; Yan, 2001 [32]; Park *et al.*, 2001 [25]; Basu *et al.*, 2002 [10]; Pardo and Pardo, 2003 [22]; Alin, 2007 [2], 2008 [3]; Alin and Kurt, 2008 [4]; Basu *et al.*, 2009 [9]), the combined disparities (*e.g.,* Park *et al.*, 1995 [24]; Basu *et al.*, 2002 [10]; Mandal *et al.*, 2011 [20]), the coupled disparities, the ϵ-combined disparities and the inlier shrunk disparities (Patra *et al.*, 2008 [26]).

4.1. Penalized disparities

Let us assume the set up of Section 2. Suppose the disparity generating function G satisfies the conditions in (2.5) in addition to its usual properties. The disparity in (2.2) can be rewritten as

$$\rho_G(p, f(\theta)) = \sum_{i:p_i>0} G(\delta_i(\theta)) f_i(\theta) + G(-1) \sum_{i:p_i=0} f_i(\theta). \tag{4.1}$$

This shows that the natural weight for the set $\{i : p_i = 0\}$, *i.e.* the empty cells, is $G(-1)$; for the power divergence family this equals $1/(\lambda + 1)$ so that this is a small positive number for large positive values of λ. The disparity generating functions are very flat on the inlier side for such values of λ.

The penalized disparity between the densities p and f_θ for the penalty weight h is defined as

$$\rho_{G_h}(p, f(\theta)) = \sum_{i:p_i>0} G(\delta_i(\theta)) f_i(\theta) + h \sum_{i:p_i=0} f_i(\theta), \ h > 0, \tag{4.2}$$

which simply replaces the natural weight of the empty cells in (4.1) with a suitable positive constant h. It is clear that the penalized disparity in (4.2) is non-negative; also evident is the fact that if the probability mass functions p and f_θ are identically equal the penalized disparity must equal zero. Again, for $h > 0$, two probability mass functions which are not identically equal must necessarily produce a positive penalized disparity. If the support of f_θ is independent of θ, the range of h can be enhanced to include $h = 0$.

For goodness-of-fit problems with a large number of cells, replacing $G(-1)$ with a large value can lead to significant benefit particularly when the sample size is small.

Leter on, in our description of the goodness-of-fit test statistics based on the penalized disparities, we will use the minimum penalized disparity estimator of θ for testing the composite null hypothesis. In order to get the correct results in this context it is necessary that the minimum penalized disparity estimators are BAN estimators. The results of Mandal *et al.* (2010) [21] ensure that the minimum penalized disparity estimators are BAN in our context.

4.2. Combined disparities

In the combined disparity approach we combine two different disparities at the positive and negative sides of the δ axis at the origin $\delta = 0$. Suppose we have two different disparities ρ_{G_1} and ρ_{G_2}; then the combined disparity ρ_{G_c} is defined by

$$\rho_{G_c}(p, f(\theta)) = \sum_i G_c(\delta_i(\theta)) f_i(\theta),$$

where

$$G_c(\delta) = \begin{cases} G_1(\delta), & \text{if } \delta \geq 0, \\ G_2(\delta), & \text{if } \delta < 0. \end{cases} \tag{4.3}$$

Suppose A_{G_1} and A_{G_2} are the residual adjustment functions corresponding to G_1 and G_2. Then the residual adjustment function A_c of the combined disparity ρ_{G_c} is defined by

$$A_c(\delta) = \begin{cases} A_{G_1}(\delta), & \text{if } \delta \geq 0, \\ A_{G_2}(\delta), & \text{if } \delta < 0. \end{cases} \tag{4.4}$$

When using the combined disparity approach, our aim will be to combine a sharply rising RAF on the positive side of the δ axis with a RAF which magnifies the effect of inliers on the negative side of the δ axis. In general, however, the second order smoothness of the residual adjustment function at $\delta = 0$ is lost as a result of this combination, i.e. $A_c''(\delta)$ does not exist at $\delta = 0$.

As in the case of the minimum penalized disparity estimators, the minimum combined disparity estimators are also BAN, a result which follows from the Mandal *et al.* (2011) [20]. We will use this fact in deriving the asymptotic distribution of combined disparity test statistics.

4.3. Coupled disparities

Suppose we start with an initial disparity ρ_G, where G is the disparity generating function, and $A_G(\delta)$ is the corresponding residual adjustment function. In the coupled disparity approach we replace $A_G(\delta)$ for negative values of δ with a third degree polynomial such that the following conditions hold:

1. The new residual adjustment function $A_{cp}(\delta)$ is a continuous function for all $\delta \in [-1, \infty)$. So $A_{cp}(0) = A_G(0) = 0$.
2. First two derivatives of $A_{cp}(\delta)$ at $\delta = 0$ match with the original residual adjustment function $A_G(\delta)$, *i.e.* $A'_{cp}(0) = A'_G(0) = 1$ and $A''_{cp}(0) = A''_G(0)$.
3. The function $A_{cp}(\delta)$ gives a desired weight to the empty cells, *i.e.* $A_{cp}(-1) = k_0$, where $k_0 < 0$ is a suitable value. We denote k_0 as the intercept parameter of the coupled disparity.

It may be noted that the first two conditions ensure that the coupled disparity has the smoothness properties of a general disparity function (as described in Section 2) at the origin $\delta = 0$. The third condition is imposed to control the inlier part of the disparity.

A large negative value of k_0 will generate an inlier sensitive disparity. Under some conditions described below the resulting coupled disparity is a genuine disparity, and the related asymptotics can be studied within the already established properties of disparities.

Conditions 1, 2 and 3 lead to four algebraic constraints on $A_{cp}(\delta)$, *viz.* $A_{cp}(0) = 0$, $A'_{cp}(0) = 1$, $A''_{cp}(0) = A''_G(0)$ and $A_{cp}(-1) = k_0$. We therefore assume that $A_{cp}(\delta)$ is a 3-rd degree polynomial of δ for $\delta \in [-1, 0]$. Solving for the coefficients of the above polynomial under the given constraints we get

$$A_{cp}(\delta) = \delta + \frac{1}{2}A''_G(0)\delta^2 + \frac{1}{2}\{A''_G(0) - 2k_0 - 2\}\delta^3, \ \delta \in [-1, 0], \tag{4.5}$$

so that the residual adjustment function of the coupled disparity $\rho_{G_{cp}}$ has the form

$$A_{cp}(\delta) = \begin{cases} A_G(\delta), & \text{if } \delta > 0, \\ \delta + \frac{1}{2}A''_G(0)\delta^2 + \frac{1}{2}\{A''_G(0) - 2k_0 - 2\}\delta^3, & \text{if } \delta \in [-1, 0]. \end{cases} \tag{4.6}$$

Using equation (2.6), the corresponding reconstructed G function is given by

$$\begin{aligned} G_{cp}(\delta) \quad = \quad & \frac{1}{4}\left(A_2 - 2k_0 - 2\right)\delta^3 - \frac{1}{4}\left(A_2 - 6k_0 - 6\right)\delta^2 \\ & -\frac{1}{2}\left(A_2 - 6k_0 - 4\right)\left(\delta - (1+\delta)\log(1+\delta)\right), \end{aligned}$$

for $\delta \in [-1, 0]$, where $A_2 = A''_G(0)$ is the curvature parameter of the disparity ρ_G. It can be shown that G_{cp} will be strictly convex for $k_0 \leq -\frac{1}{6}\{A_2 + 4\}$ (see Mandal, 2010 [19]). In such a case the coupled disparity satisfies all the properties of a regular disparity, so that the asymptotic null distribution of the corresponding goodness-of-fit test statistic $2n\rho_{G_{cp}}(p, \hat{\pi}_0)$ follows from the existing results (*e.g.,* Basu and Sarkar, 1994 [11]), where $\hat{\pi}_0$ is any BAN estimator of π under the null hypothesis. However, also see the comments in Remark 1 later.

4.4. ϵ-combined disparities

In the ϵ-combined disparity approach we combine two residual adjustment functions A_{G_1} and A_{G_2} of two different regular disparities near the origin $\delta = 0$ by smoothing them using a seventh degree polynomial in the interval $\delta \in [-\epsilon, \epsilon]$, where ϵ is a small positive number. This smooth joining allows the combined function to retain the second order smoothness properties at $\delta = 0$; otherwise the method is similar in spirit to the combined disparity approach. The residual adjustment function of the ϵ-combined disparity is defined by

$$A_\epsilon(\delta) = \begin{cases} A_{G_1}(\delta), & \text{if } \delta > \epsilon, \\ \sum_{i=0}^{7} k_i \delta^i, & \text{if } -\epsilon \le \delta \le \epsilon, \\ A_{G_2}(\delta), & \text{if } \delta < -\epsilon. \end{cases} \tag{4.7}$$

The smoothed function A_ϵ should satisfy the following conditions:

1. $A_\epsilon(0) = 0$ and $A'_\epsilon(0) = 1$.
2. $A_\epsilon(\delta)$ is a continuous function for $-1 \le \delta < \infty$. So $A_\epsilon(\epsilon) = A_{G_1}(\epsilon)$ and $A_\epsilon(-\epsilon) = A_{G_2}(-\epsilon)$.
3. The first derivative of $A_\epsilon(\delta)$ exists for all values of δ in the interval $(-1, \infty)$. So $A'_\epsilon(\epsilon) = A'_{G_1}(\epsilon)$ and $A'_\epsilon(-\epsilon) = A'_{G_2}(-\epsilon)$.
4. The second derivative of $A_\epsilon(\delta)$ exists for all values of δ in the interval $(-1, \infty)$. So $A''_\epsilon(\epsilon) = A''_{G_1}(\epsilon)$ and $A''_\epsilon(-\epsilon) = A''_{G_2}(-\epsilon)$.

The above generates eight algebraic constraints, so a seventh degree polynomial serves the purpose. Whenever the resulting residual adjustment function A_ϵ is increasing, as it normally is, the associated disparity generating function G_ϵ is convex, so that the asymptotic distribution of the resulting goodness-of-fit test statistic again follows from existing results.

4.5. Inlier-shrunk disparities

Let G be a function satisfying the disparity conditions in equation (2.5). Define the corresponding inlier-shrunk class of disparity generating functions indexed by the inlier shrinking parameter $\gamma \in \mathbb{R}$ through the relation

$$G_\gamma(\delta) = \begin{cases} G(\delta), & \delta \ge 0, \\ \dfrac{G(\delta)}{(1+\delta^2)^\gamma}, & \delta < 0. \end{cases} \tag{4.8}$$

Notice that this strategy again keeps the G function intact on the outlier side but modifies it in the inlier side. The function is shrunk closer to zero for $\gamma > 0$ and is magnified for $\gamma < 0$. For our purpose, inlier shrunk disparities with $\gamma < 0$ are the relevant ones (inspite of the "inlier shurnk" name). It can be easily verified that $G_\gamma^{'''}(\delta)$, the third derivative of the function $G_\gamma(\delta)$, exists and is continuous at $\delta = 0$; the same is true for the corresponding second derivative of the residual adjustment function. Thus for every inlier shrinking parameter γ one would only need to verify that G_γ is a convex function to establish that the associated inlier-shrunk disparity satisfies the original disparity conditions, so that the asymptotic null distribution of the corresponding goodness-of-fit test will still be as given in Theorem 2.1.

Remark 1. The primary use of the convexity of the function $G(\delta)$ is in establishing the non-negativity of the disparity ρ_G. The condition (2.5) assures that the disparity generating function is non-negative throughout the range $\delta \in [-1, \infty)$. Any modification of the disparity generating function G which keeps the function non-negative also keeps the resulting disparity non-negative, irrespective of whether the G function remains convex or not. In such cases the asymptotic distribution of the corresponding test statistic continues to hold as long as conditions of Theorem 2.1 (Theorem 3.1 of Basu and Sarkar, 1994 [11]) are satisfied. In case of inlier shrunk disparities, for example, the modification on the inlier side either shrinks it to a fraction of its existing value or magnifies it to a larger value, and thus the modified function G_γ continues to remain non-negative. Therefore the asymptotic null distribution of the goodness-of-fit statistic $2n\rho_{G_\gamma}(p, \hat{\pi}_0)$, where $\hat{\pi}_0$ is any BAN estimator under the null hypothesis, follows from Theorem 2.1.

5. Some asymptotic results

Here we prove the asymptotic null distributions of the penalized and the combined disparity test statistics, abbreviated hereafter as PDT and CDT respectively. The proofs of the asymptotic distributions of the other three types of inlier modified test statistic – the coupled disparity statistic (CpDT), ϵ-combined disparity statistic (ϵ-CDT) and the inlier shrunk disparity statistic (ISDT) – follow directly from Theorem 2.1 as long as the modified G functions remain non-negative.

5.1. Asymptotic null distribution of the PDT

In this section we establish the asymptotic distribution of the penalized disparity test (PDT) statistic $2n\rho_{G_h}(p, \hat{\pi}_0)$, where $\hat{\pi}_0$ is any BAN estimator of π under H_0. Here ρ_{G_h} is as defined in (4.2).

Theorem 5.1. *Under the conditions of Birch (1964) [12] (see Read and Cressie, 1988 [29]) the penalized disparity statistic $2n\rho_{G_h}(p, \hat{\pi}_0)$ has an asymptotic $\chi^2(k-s-1)$ distribution under the null hypothesis given in (2.1).*

Proof. By Theorem 2.1 the disparity statistic $2n\rho_G(p, \hat{\pi}_0)$ has an asymptotic $\chi^2(k - s - 1)$ distribution under the null hypothesis. Hence it is sufficient to show that under the null

$$R_n(\theta) = 2n\rho_G(p, \hat{\pi}_0) - 2n\rho_{G_h}(p, \hat{\pi}_0) = o_p(1). \tag{5.1}$$

Now

$$R_n(\theta) = 2n(G(-1) - h) \sum_{i:p_i=0} \hat{\pi}_{0i} = 2n(G(-1) - h) \sum_{i=1}^{k} \hat{\pi}_{0i} I(p_i), \tag{5.2}$$

where $I(y) = 1$ if $y = 0$ and 0 otherwise. So

$$\begin{aligned} E\left[|R_n(\theta)|\right] &= 2n|G(-1) - h|\ E\left[\sum_{i=1}^{k} \hat{\pi}_{0i} I(p_i)\right] \\ &\leq 2n|G(-1) - h| \sum_{i=1}^{k} E\left[I(p_i)\right] \\ &\leq 2n|G(-1) - h| \sum_{i=1}^{k} (1 - \pi_{0i})^n. \end{aligned} \tag{5.3}$$

Suppose $g_n(x) = n(1 - x)^n$, where $x \in (0, 1)$. Note $g_n(x) \to 0$ for all $x \in (0, 1)$ as $n \to \infty$. As k is finite and $\pi_{0i} > 0$ for each i, we get from (5.3)

$$E\left[|R_n(\theta)|\right] \to 0 \quad \text{as } n \to \infty.$$

Hence, by Markov's inequality, the condition in equation (5.1) holds. □

In Section 4 we have listed some of the previous applications of the empty cell penalty approach in the literature. Park *et al.* (2001) [25] considered the asymptotic distribution of the penalized disparity test statistic for the special case of the blended weight Hellinger distance.

5.2. Asymptotic null distribution of the CDT

Here we will prove the asymptotic distribution of the combined disparity test (CDT) statistic $2n\rho_{G_c}(p, \hat{\pi}_0)$, where $\hat{\pi}_0$ is any BAN estimator of π under H_0. We assume that the component functions G_1 and G_2 both satisfy the conditions in (2.5); in addition G_1''' and G_2''' are assumed to be finite and conditions at zero, where G''' is the third derivative of G.

Theorem 5.2. *Under the conditions of Birch (1964) [12] (see Read and Cressie, 1988 [29]) the combined disparity statistic $2n\rho_{G_c}(p, \hat{\pi}_0)$ has an asymptotic $\chi^2(k-s-1)$ distribution under the null hypothesis given in (2.1).*

Proof. The combined disparity test statistic in given by

$$\begin{aligned} 2n \sum_{i=1}^{k} G_c\left(\frac{p_i - \hat{\pi}_{0i}}{\hat{\pi}_{0i}}\right) \hat{\pi}_{0i} = 2n \sum_{p_i \geq \hat{\pi}_{0i}} G_1\left(\frac{p_i - \hat{\pi}_{0i}}{\hat{\pi}_{0i}}\right) \hat{\pi}_{0i} \\ + 2n \sum_{p_i < \hat{\pi}_{0i}} G_2\left(\frac{p_i - \hat{\pi}_{0i}}{\hat{\pi}_{0i}}\right) \hat{\pi}_{0i}. \end{aligned}$$

When $p_i \geq \hat{\pi}_{0i}$ a Taylor series expansion about $p_i = \hat{\pi}_{0i}$ gives

$$\begin{aligned} 2nG_c\left(\frac{p_i - \hat{\pi}_{0i}}{\hat{\pi}_{0i}}\right) \hat{\pi}_{0i} = 2nG_1(0)\hat{\pi}_{0i} + 2n(p_i - \hat{\pi}_{0i})G_1'(0) + \frac{n}{\hat{\pi}_{0i}}(p_i - \hat{\pi}_{0i})^2 G_1''(0) \\ + \frac{n}{3\hat{\pi}_{0i}^2}(p_i - \hat{\pi}_{0i})^3 G_1'''\left(\frac{\xi_i - \hat{\pi}_{0i}}{\hat{\pi}_{0i}}\right), \end{aligned} \tag{5.4}$$

where ξ_i lies in the line segment joining the points p_i and $\hat{\pi}_{0i}$. Now, from assumption (2.5), $G_1(0) = G_1'(0) = 0$ and $G_1''(0) = 1$. Hence, when $p_i \geq \hat{\pi}_{0i}$, equation (5.4) reduces to

$$2nG_c\left(\frac{p_i - \hat{\pi}_{0i}}{\hat{\pi}_{0i}}\right)\hat{\pi}_{0i} = \frac{n}{\hat{\pi}_{0i}}(p_i - \hat{\pi}_{0i})^2 + \frac{n}{3\hat{\pi}_{0i}^2}(p_i - \hat{\pi}_{0i})^3 G_1'''\left(\frac{\xi_i - \hat{\pi}_{0i}}{\hat{\pi}_{0i}}\right). \quad (5.5)$$

Notice that ξ_i lies in the same side of $\hat{\pi}_0$ as p_i.

Similarly when $p_i < \hat{\pi}_{0i}$, using the properties of the function G_2 we get

$$2nG_c\left(\frac{p_i - \hat{\pi}_{0i}}{\hat{\pi}_{0i}}\right)\hat{\pi}_{0i} = \frac{n}{\hat{\pi}_{0i}}(p_i - \hat{\pi}_{0i})^2 + \frac{n}{3\hat{\pi}_{0i}^2}(p_i - \hat{\pi}_{0i})^3 G_2'''\left(\frac{\zeta_i - \hat{\pi}_{0i}}{\hat{\pi}_{0i}}\right), \quad (5.6)$$

where ζ_i lies in the line segment joining the points $\hat{\pi}_{0i}$ and p_i. Again ζ_i lies on the same side of $\hat{\pi}_{0i}$ as p_i. As G_1''' and G_2''' are bounded in the neighborhood of zero, combining equations (5.5) and (5.6) we get

$$2nG_c\left(\frac{p_i - \hat{\pi}_{0i}}{\hat{\pi}_{0i}}\right)\hat{\pi}_{0i} = \frac{n}{\hat{\pi}_{0i}}(p_i - \hat{\pi}_{0i})^2 + o_p(1), \quad i = 1, 2, \ldots, k. \quad (5.7)$$

Therefore adding the terms on both side of (5.7) for $i = 1, \ldots, k$ we get

$$2n\sum_{i=1}^{k} G_c\left(\frac{p_i - \hat{\pi}_{0i}}{\hat{\pi}_{0i}}\right)\hat{\pi}_{0i} = n\sum_{i=1}^{k}\frac{(p_i - \hat{\pi}_{0i})^2}{\hat{\pi}_{0i}} + o_p(1).$$

So the asymptotic distribution of $2n\rho_{G_c}(p, \hat{\pi}_0)$ is equivalent to the distribution of the Pearson's chi-square statistic, and the required result holds. □

In Section 4 we have listed some of the previous applications of the combined distance approach in statistical literature. Basu *et al.* (2002) [10] attempted to show that the asymptotic distribution of the combined disparity statistic differs from the Pearson's chi-square statistic by only a small order term. However, they considered only the case of the simple null hypothesis, and their proof was not completely rigorous.

6. Numerical results

In this section we will present the results of an extensive numerical study to explore the properties of the inlier modified test statistics. Our illustrations will primarily focus on the power divergence family, but will also use the blended weight Hellinger distance and the blended weight chi-square distance for illustration. To ensure that the illustration is as streamlined as possible, we denote the disparity tests according to the following convention. We denote the ordinary disparity tests as DT, with the value of the tuning parameter indicated within parentheses, and the subscripts λ, β and τ representing the power divergence, the blended weight Hellinger distance and the blended weight chi-square families respectively. Thus $\mathrm{DT}_\lambda(2)$ will represent the disparity test within the power divergence family with tuning parameter 2, $\mathrm{DT}_\beta(0.5)$ will represent the disparity test within the blended weight Hellinger distance family with tuning parameter 0.5, and so on. For the inlier modified families the prefixes P, C, Cp and IS will represent the indicated inlier modification scheme, while the second argument will represent the inlier modification parameter. For example, $\mathrm{PDT}_\lambda(2,1)$ will represent the penalized power divergence statistic with tuning parameter 2 and penalty weight $h=1$; $\mathrm{CDT}_\beta(0.2,0.5)$ will represent the combined disparity statistic within the blended weight Hellinger distance family with the outlier and inlier components represented by the statistics with tuning parameters 0.2 and 0.5 respectively; $\mathrm{CpDT}_\tau(0.2,-1)$ will represent the coupled disparity statistic within the blended weight chi-square family with tuning parameter 0.2 and intercept parameter $k_0=-1$, and so on.

In our numerical illustrations we have left out the ϵ-combined disparity statistics, as they tend to present results very similar to the corresponding combined disparity statistics for small values of ϵ.

6.1. Comparison of the exact power

In this section we present some evidence of the "improved" performance of the inlier modified test statistics. For the first set of illustrations, we choose the statistics within the power divergence family. Our aim is to start with a test statistic with a large positive value of λ (so that the statistic is sensitive to outliers) and suitably modify the inlier side. We consider the same equiprobable testing problem as discussed in Section 3 for the particular values $n=20$ and $k=5$. Choosing the disparity test corresponding to $\lambda=2$ within the power divergence family, we consider bump and dip alternatives over the range of different values of η. In Figures 2 and 3 we present the exact powers (as discussed in Section 3) of the penalized and coupled test statistics, respectively, for the parameters indicated in the figures, together with the ordinary test statistic. In either case the figure shows that the inlier modified disparity test significantly improves the power of the test statistics on the inlier side, and is associated with a small average drop in power on the outlier side. In Figure 4 the power of the combined disparity test corresponding to the tuning parameters $\lambda=2$ in the outlier side and $\lambda=-0.5$ in the inlier side is presented together with the two ordinary disparity tests; in the sense of protecting against the worst case scenario, the combined statistic is clearly much more balanced in this illustration compared to the two ordinary statistics.

6.2. Comparison of the corrected power

To calculate the exact power as reported in Table 1 and Figures 2, 3 and 4, we need to determine the values of the test statistics for all $\binom{n+k-1}{n}$ possible samples. So the exact calculation is not feasible when n and k are large. However, we can still get a reasonable comparison of the powers of the test statistics by using simulated finite sample critical values (instead of the asymptotic chi-square critical value). We now present some simulation results where the parameters

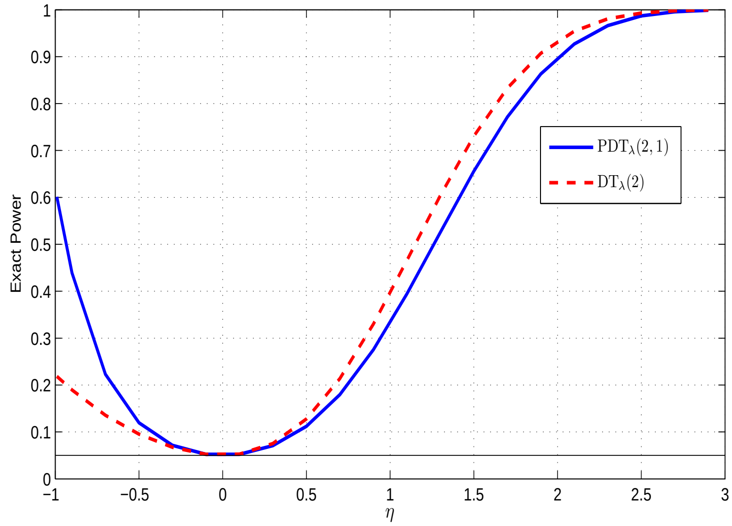

Fig 2. *Comparison of the exact powers of the ordinary and the penalized disparity test statistics considered in Section 6.1.*

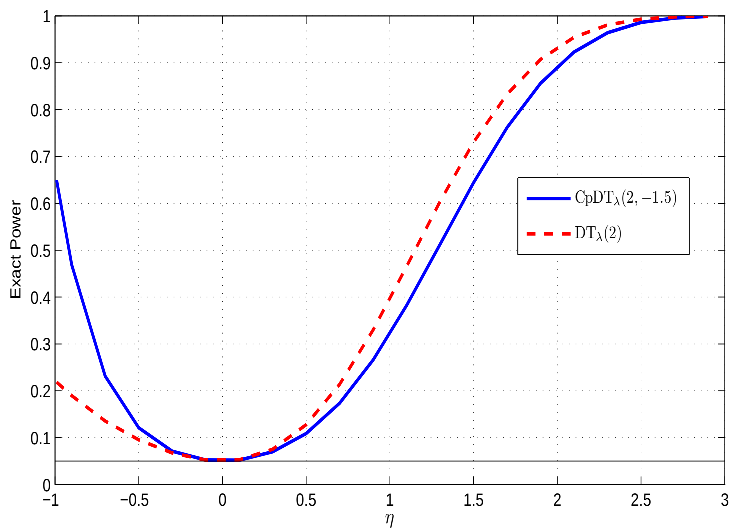

Fig 3. *Comparison of the exact powers of the ordinary and the coupled disparity test statistics considered in Section 6.1.*

for the multinomial model are $n = 50$ and $k = 10$. The null and the alternative hypotheses are given in equations (3.1) and (3.2) respectively, and the value of η ranges from -1 to 4. The test statistics we have considered are (a) the log likelihood ratio statistic (LDT), (b) the $DT_\lambda(-0.5)$ statistic; (c) the $DT_\lambda(3)$

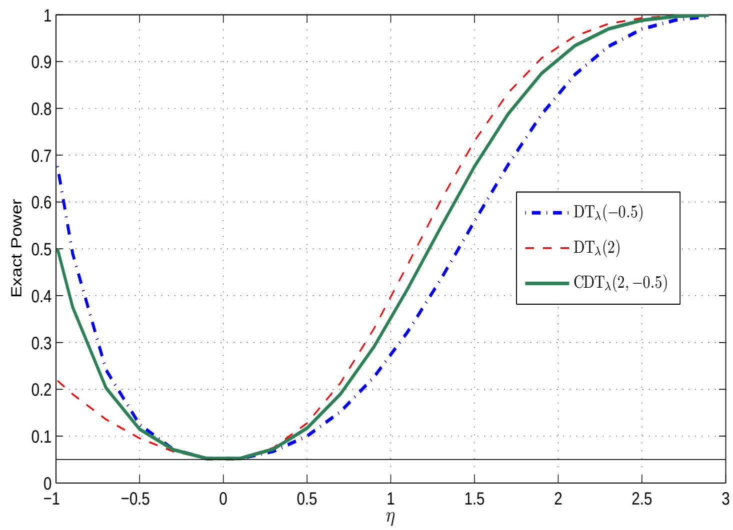

FIG 4. *Comparison of the exact powers of the ordinary and the combined disparity test statistics considered in Section 6.1.*

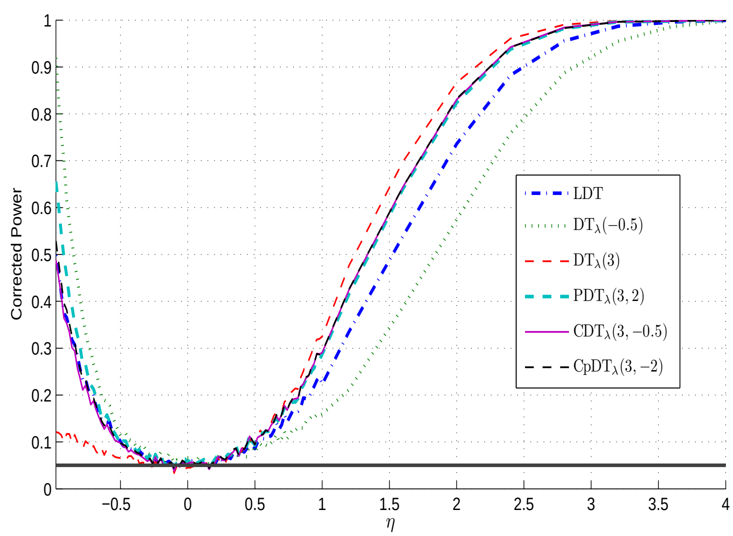

FIG 5. *Comparison of the corrected powers of the disparity test statistics considered in Section 6.2.*

statistic; (d) the $\mathrm{PDT}_\lambda(3,2)$ statistic; (e) the $\mathrm{CDT}_\lambda(3,-0.5)$ statistic and (f) the $\mathrm{CpDT}_\lambda(3,-2)$ statistic. Figure 5 shows the corrected powers where we have used the simulated critical points. Here one can observe that the corrected power for $\mathrm{DT}_\lambda(3)$ is best for the bump alternatives, whereas it is worst for the dip alternatives. Exactly the opposite situation is seen for the $\mathrm{DT}_\lambda(-0.5)$. Our inlier modified test statistics show overall good performance for all alternatives. It may be noticed that the LDT also provides a reasonable balance in all the situations, but all three inlier modified tests appear to be either competitive or better than the LDT everywhere over the range of η.

6.3. Real data examples

Example 6.1. A total of 182 psychiatric patients on drugs were classified according to their diagnosis. This data set has been analyzed previously by Agresti (1990 [1], p. 72, Table 3.10) and Basu *et al.*, (2002) [10]. The breakdown of the frequencies in the different classes is as shown in Table 2. The sample size is fairly large, and we expect that the chi-square approximation will be reasonable for all our ordinary and modified statistics.

Consider testing a simple null hypothesis where the probability vector is given by $\boldsymbol{\pi}_0 = (0.56, 0.06, 0.09, 0.25, 0.04)^T$. The deviations of the observed relative frequencies from the null probabilities appear to be small; the research question is whether these small deviations are enough to indicate a significant difference from the null given the large sample size. The χ^2 critical value at 4 degrees of freedom and 5% level of significance is given by 9.488. The ordinary power divergence statistics for $\lambda = 2, 1$ and $2/3$ are 5.273, 7.690 and 9.142 respectively, and these tests fail to reject the null hypothesis. However, for the associated penalized tests (note that the data set contains one empty cell) corresponding to $\lambda = 2,\ 1$ and $2/3$ with penalty weight $h = 2$ the test statistics are 29.540, 29.530, and 29.526, so that these statistics reject the null hypothesis comfortably. A bunch of other test statistics within the domain of our discussion are presented in Table 3. Most of the ordinary test statistics either fail to reject H_0 or provide a marginal rejection; however the inlier modified statistics clearly improve the discrimination.

Example 6.2. Next we consider the time passage example data (Read and Cressie, 1988 [29], pp. 12-16) which studied the relationships between stresses in life and illnesses in Oakland, California based on a sample of size $n = 147$, fairly large value. These data have also been analyzed by Basu *et al.* (2002) [10]. The data are in the form of an 18-cell multinomial, with the frequencies

TABLE 2

The null probabilities and the observed relative frequencies for 182 psychiatric patients used in Example 6.1

Diagnosis	Frequency	Relative Frequency	Null Probability
Schizophrenia	105	0.577	0.56
Affective disorder	12	0.066	0.06
Neurosis	18	0.099	0.09
Personality disorder	47	0.258	0.25
Special symptoms	0	0.000	0.04

TABLE 3
The test statistics for the schizophrenia data used in Example 6.1

$\mathrm{DT}_\lambda(2/3)$	$\mathrm{DT}_\lambda(1)$	$\mathrm{DT}_\lambda(2)$
9.142	7.690	5.273
$\mathrm{PDT}_\lambda(2/3{,}2)$	$\mathrm{PDT}_\lambda(1{,}2)$	$\mathrm{PDT}_\lambda(2{,}2)$
29.526	29.530	29.540
$\mathrm{CDT}_\lambda(2/3,-0.5)$	$\mathrm{CDT}_\lambda(1,-0.5)$	$\mathrm{CDT}_\lambda(2,-0.5)$
29.526	29.530	29.540
$\mathrm{CpDT}_\lambda(2/3,-2)$	$\mathrm{CpDT}_\lambda(1,-2)$	$\mathrm{CpDT}_\lambda(2,-2)$
29.526	29.530	29.540
$\mathrm{ISDT}_\lambda(2/3,-1)$	$\mathrm{ISDT}_\lambda(1,-1)$	$\mathrm{ISDT}_\lambda(2,-1)$
17.878	14.970	10.127
$\mathrm{DT}_\beta(0.1)$	$\mathrm{DT}_\beta(0.2)$	$\mathrm{DT}_\beta(0.3)$
9.394	11.779	15.258
$\mathrm{CpDT}_\beta(0.1,-2)$	$\mathrm{CpDT}_\beta(0.2,-2)$	$\mathrm{CpDT}_\beta(0.3,-2)$
29.527	29.524	29.521
$\mathrm{DT}_\tau(0.1)$	$\mathrm{DT}_\tau(0.2)$	$\mathrm{DT}_\tau(0.3)$
8.495	9.503	10.801
$\mathrm{ISDT}_\tau(0.1,-1)$	$\mathrm{ISDT}_\tau(0.2,-1)$	$\mathrm{ISDT}_\tau(0.3,-1)$
16.584	18.603	21.201

TABLE 4
The test statistics for the time passage data used in Example 6.2

$\mathrm{DT}_\lambda(2/3)$	$\mathrm{DT}_\lambda(1)$	$\mathrm{DT}_\lambda(2)$
22.896	22.349	21.706
$\mathrm{PDT}_\lambda(2/3{,}2)$	$\mathrm{PDT}_\lambda(1{,}2)$	$\mathrm{PDT}_\lambda(2{,}2)$
22.896	22.349	21.706
$\mathrm{CDT}_\lambda(2/3,-0.5)$	$\mathrm{CDT}_\lambda(1,-0.5)$	$\mathrm{CDT}_\lambda(2,-0.5)$
28.429	29.119	31.511
$\mathrm{CpDT}_\lambda(2/3,-2)$	$\mathrm{CpDT}_\lambda(1,-2)$	$\mathrm{CpDT}_\lambda(2,-2)$
28.584	28.753	29.530
$\mathrm{ISDT}_\lambda(2/3,-1)$	$\mathrm{ISDT}_\lambda(1,-1)$	$\mathrm{ISDT}_\lambda(2,-1)$
28.014	27.177	25.761
$\mathrm{DT}_\beta(0.1)$	$\mathrm{DT}_\beta(0.2)$	$\mathrm{DT}_\beta(0.3)$
22.855	23.503	24.301
$\mathrm{CpDT}_\beta(0.1,-2)$	$\mathrm{CpDT}_\beta(0.2,-2)$	$\mathrm{CpDT}_\beta(0.3,-2)$
28.595	28.472	28.384
$\mathrm{DT}_\tau(0.1)$	$\mathrm{DT}_\tau(0.2)$	$\mathrm{DT}_\tau(0.3)$
22.567	22.942	23.486
$\mathrm{ISDT}_\tau(0.1,-1)$	$\mathrm{ISDT}_\tau(0.2,-1)$	$\mathrm{ISDT}_\tau(0.3,-1)$
27.577	28.152	28.914

representing the total number of respondents for each month who indicated one stressful event between 1 and 18 months before the interview. The null hypothesis $H_0 : \pi_{0i} = 1/18$, $i = 1, 2, \ldots, 18$, of equiprobability is clearly an untenable one and is soundly rejected by practically all the ordinary disparity tests. However, the model fit appears to improve if we consider a log-linear time trend model $H_0 : \log(\pi_{0i}) = \vartheta + \beta i$, $i = 1, 2, \ldots, 18$. The values of the test statistics are given in Table 4. Expected frequencies on the basis of estimates of ϑ and β were obtained by using the same minimum disparity estimators as used in the corresponding test statistics, while Read and Cressie (1988 [29], Table 2.2)

Table 5
Questionnaire data from Haut et al. *(1987) used in Example 6.3*

Husband's Rating	Wife's Rating			
	Never or Occasionally	Fairly Often	Very Often	Almost Always
Never or Occasionally	7	7	2	3
Fairly Often	2	8	3	7
Very Often	1	5	4	9 (8)
Almost Always	2	8	9	14

used the maximum likelihood estimators in all cases. The test statistics are now compared with the critical value of a χ^2-statistic with 16 degrees of freedom (rather than 17), and the critical value at 5% level of significance is 26.296. All the ordinary power divergence test statistics with $\lambda \in [0, 2]$ fail to reject the time trend null hypothesis; for example, the $\text{DT}_\lambda(2)$ test generates a statistic equal to 21.706. However the combined statistics corresponding to $\text{CDT}_\lambda(2/3,-0.5)$ and $\text{CDT}_\lambda(2,-0.5)$ are equal to 28.429 and 31.511 respectively, and in either case the null is soundly rejected. Thus, unlike the ordinary disparity tests, the time trend model fails to pass the goodness-of-fit standards of these inlier modified disparity tests. Notice however that the penalized statistics provide no improvement as there are no empty cells.

Remark 2. In Examples 6.1 and 6.2, the failure of the ordinary statistics considered in Tables 3 and 4 to reject the null hypothesis is primarily due to their inability to properly address the effect of the inliers. Some of the ordinary inlier sensitive statistics would have also rejected the null hypothesis in these cases; for example, the $\text{DT}_\lambda(-0.5)$ test statistic is equal to 29.515 and 26.377 for Examples 6.1 and 6.2 respectively, and this would lead to rejections in either case at 5% level of significance. Similarly $\text{DT}_\beta(0.7)$ statistic has values 81.278 and 29.129 in Example 6.1 and 6.2 respectively which would have led to rejection in either case at the 5% level. The point this paper tries to make is that our inlier modified statistics will continue to provide sharp discrimination even when the ordinary inlier sensitive statistics fail, as we will see in Example 6.3.

Example 6.3. This example is from Haut *et al.* (1987) [17] which summarizes the agreement between husband and wife based on the answers to some questions of relevance. The data are given in Table 5. The survey was conducted on 91 married couples from the Tucson metropolitan area, again a reasonable large sample size. For better illustration we have intentionally changed the (3,4)-th element of the contingency table from 9 to 8, so the sample size is changed to 90. We want to test the null hypothesis that the response between husband and wife is independent. The values of different goodness-of-fit test statistics are presented in Table 6. The asymptotic null distribution of the test statistics is χ^2 with $3 \times 3 = 9$ degrees of freedom, and the critical point at 5% level of significance is 16.919. Here the ordinary power divergence statistics fail to reject H_0 for all $\lambda \in [-1, 1.198]$. But most of the inlier modified test statistics soundly reject the null hypothesis. The results are provided in Table 6.

Table 6
The test statistics for the questionnaire data used in Example 6.3

$DT_\lambda(2/3)$	$DT_\lambda(1)$	$DT_\lambda(2)$
15.703	16.388	20.254
$PDT_\lambda(2/3,2)$	$PDT_\lambda(1,2)$	$PDT_\lambda(2,2)$
15.703	16.388	20.254
$CDT_\lambda(2/3,-0.5)$	$CDT_\lambda(1,-0.5)$	$CDT_\lambda(2,-0.5)$
17.585	18.680	23.538
$CpDT_\lambda(2/3,-3)$	$CpDT_\lambda(1,-3)$	$CpDT_\lambda(2,-3)$
18.513	19.342	23.405
$ISDT_\lambda(2/3,-2)$	$ISDT_\lambda(1,-2)$	$ISDT_\lambda(2,-2)$
19.536	19.954	23.179
$DT_\beta(0.1)$	$DT_\beta(0.4)$	$DT_\beta(0.7)$
15.747	14.929	15.421
$CpDT_\beta(0.1,-4)$	$CpDT_\beta(0.4,-4)$	$CpDT_\beta(0.7,-4)$
19.751	18.287	17.602
$DT_\tau(0.1)$	$DT_\tau(0.4)$	$DT_\tau(0.7)$
15.455	14.393	14.948
$ISDT_\tau(0.1, -2)$	$ISDT_\tau(0.4, -2)$	$ISDT_\tau(0.7, -2)$
19.224	18.943	20.697

To systematically summarize some of our findings in the above examples, we note the following points.

1. When the primary violation is through a large inlier, as in the data in Table 2, positive values of λ have little impact on the value of the inlier modified statistics. Thus most of the inlier modified statistics in Table 3 are of the order of the ordinary statistics $DT_\lambda(-0.5)$ (reported in Remark 2), irrespective of the value of λ in the outlier component.
2. Instead of the frequencies reported in Table 2, if the observed frequency vector in Example 6.1 was (98, 21, 15, 42, 6), then, for testing the same null hypothesis the primary violator would be a (relatively moderate) outlier in cell 2. In this case the ordinary statistic $DT_\lambda(-0.5)$ is merely 7.336, nowhere near the rejection region. Here the ordinary $DT_\lambda(1)$ and $DT_\lambda(2)$ statistics are 10.066 and 12.904, respectively, leading to comfortable although not overwhelming rejection. The inlier modified statistics $CDT_\lambda(1, -0.5)$ and $CDT_\lambda(2, -0.5)$ are 10.107 and 12.970, respectively, and are of the same order of the ordinary statistics. Thus in this case the positive values of λ (and not the inlier component) primarily drive the statistics.
3. Example 6.3 presents a combination of outliers and inliers. While the ordinary power divergence statistics fail to reject the null for all $\lambda \in [-1, 1.198]$, inlier modified statistics such as $CDT_\lambda(1, -0.5)$ comfortably reject the null by taking into account the cumulative effect of both inliers and outliers which any ordinary statistics fail to do.
4. The power divergence statistic $DT_\lambda(2/3)$ is one of the statistics highly recommended by Cressie and Read (1984) [14]. This statistic provides a moderate level of sensitivity for both inliers and outliers, and thus stands

out among the class of ordinary statistics. However, our purpose is to make the statistic strongly sensitive in both directions, and therefore $\mathrm{DT}_\lambda(2/3)$ does not appear to be an appropriate starting point for inlier modification. In Example 6.1 the value of $\mathrm{DT}_\lambda(2/3)$ is 9.142 which, although borderline, does not lead to rejection. If the observed frequency vector was (98, 21, 15, 42, 6) instead of the frequencies reported in Table 2, the value of $\mathrm{DT}_\lambda(2/3)$ statistic would have been 9.329, again on the borderline but not strong enough to recommend rejection. For positive values of λ there are many choices (such as $\lambda = 1$ or 2) for which outlier sensitivity is stronger than that of $\lambda = 2/3$. While $\mathrm{DT}_\lambda(2/3)$ clearly has a leading role in the theory of the ordinary power divergence statistics we expect it will have a more little impact on the theory of the inlier modified statistics.

6.4. 3D plots for the inlier modified statistics

Now we present 3D plots of the corrected power surfaces of some of our inlier modified test statistics applied to the hypotheses presented in (3.1) and (3.2); the calculation of the corrected power is as described in Section 6.2. For a specific value of the alternative, this allows us to look at the inlier modified statistics over the totality of combinations of the tuning parameters and the inlier modification parameters for fixed values of η. The results in this subsection are based on 2000 replications, and correspond to a sample size of $n = 50$, number of cells $k = 10$, and a nominal level of $\alpha = 0.05$.

6.4.1. Penalized disparity statistics

The penalized disparity statistics within the power divergence family are formed by taking different members of the family and combining them with different empty cell penalty weights. The values of η considered are $\eta = 1.5$ and $\eta = -0.8$, and the resulting corrected 3D power surface plots are presented in Figure 6. The tuning parameter λ and the penalty weight h are varied along the two axes of the base. The black dots on the plot indicate the optimum observed value of h (in terms of maximized power) when the tuning parameter λ is fixed. The black lines on the surface and the base of the figures indicate the points where $h = 1/(1+\lambda)$, *i.e.* where the value of h generates the ordinary disparity tests. The figure shows that there is a marginal loss in power for the bump alternative when the statistics corresponding to $\lambda = 2$ or 3 are penalized with a large penalty weight, but there is a substantial gain in power for the same set of parameters in case of the dip alternative.

6.4.2. Coupled disparity statistics

The 3D plots of the corrected power surfaces of the CpDTs within the power divergence family for different combinations of the inlier intercept parameter k_0

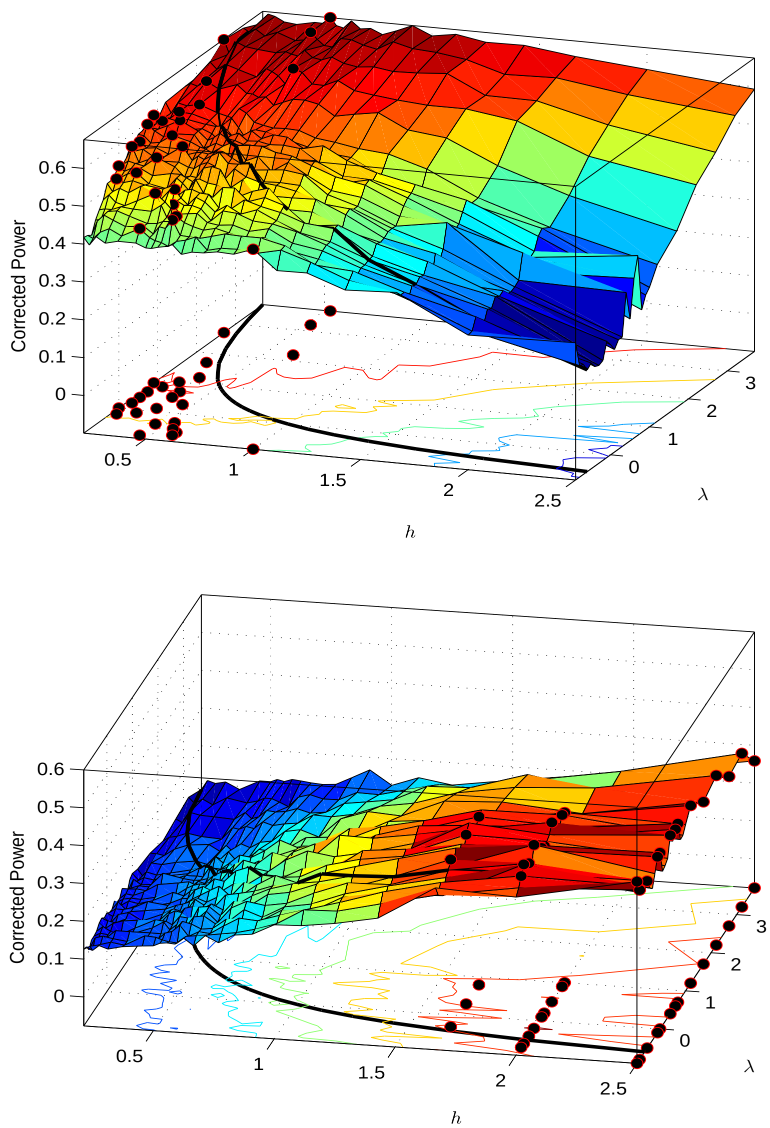

FIG 6. *(a) Corrected powers of the* PDT_λ *statistic for different values of the tuning parameter and the penalty parameter* h*, where* $\eta = 1.5$*, and (b) Corrected powers of the* PDT_λ *statistic for different values of the tuning parameter and the penalty parameter* h*, where* $\eta = -0.8$*.*

and the tuning parameter λ are presented in Figure 7 corresponding to $\eta = 1.5$ and $\eta = -0.8$. Once again there are marginal drops in the power of these tests for large positive values of the tuning parameter λ when coupled with large negative values of k_0 in case of the bump alternative; however in case of

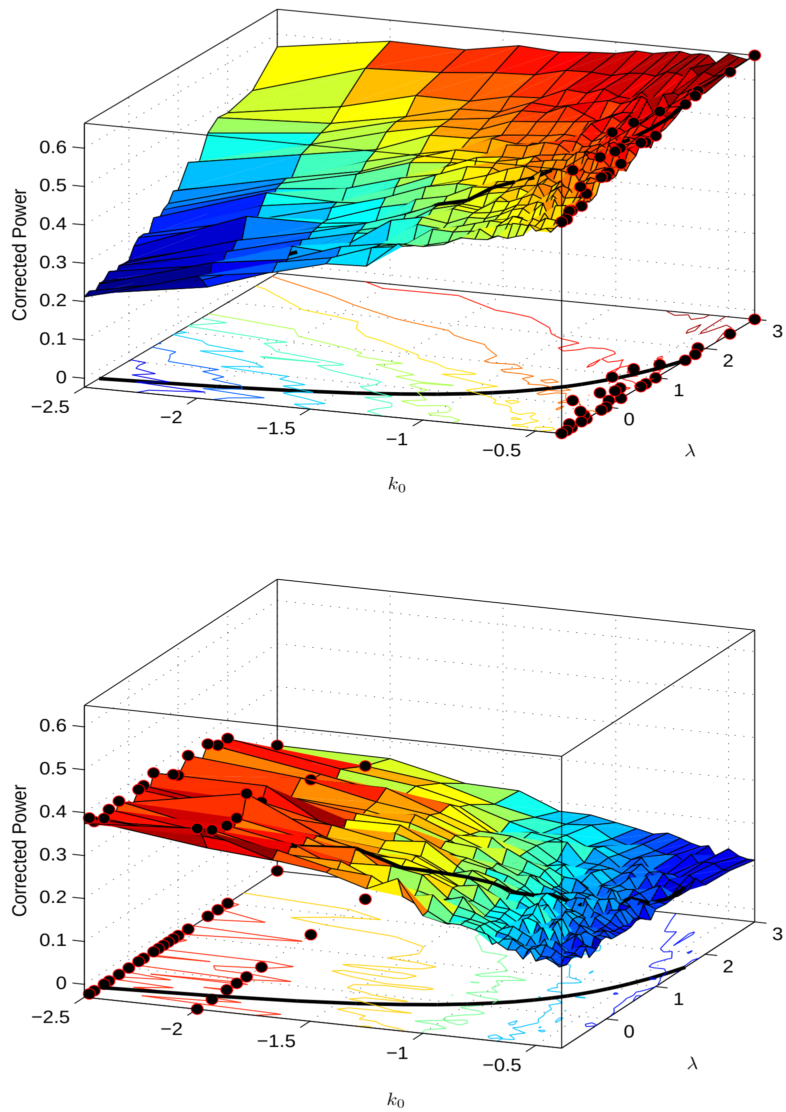

FIG 7. *(a) Corrected powers of the* CpDT_λ *statistic for different values of the tuning parameter and the intercept parameter* k_0*, where* $\eta = 1.5$*, and (b) Corrected powers of the* CpDT_λ *statistic for different values of the tuning parameter and the intercept parameter* k_0*, where* $\eta = -0.8$.

the dip alternative these same set of modifications lead to substantial gains in power.

Figure 8 shows the rising pattern of the G function of a typical coupled disparity on the inlier side as the value of k_0 decreases, and gives an indication of the increased inlier sensitivity of the corresponding statistic.

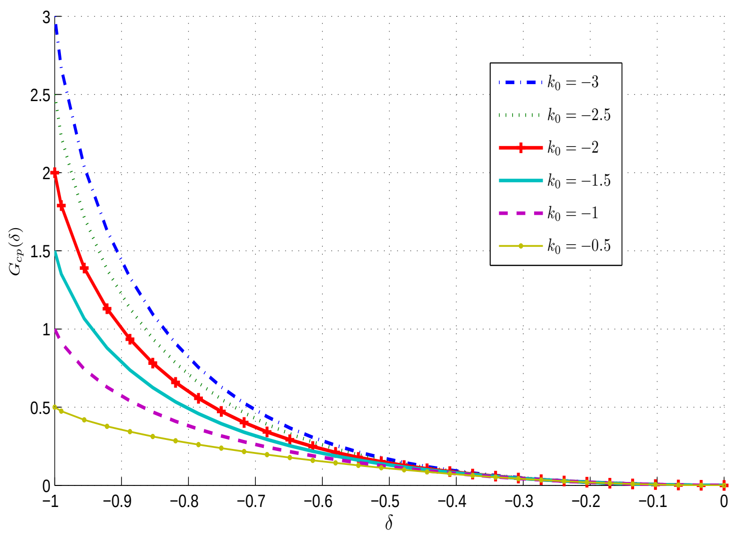

FIG 8. *Inlier part of the coupled disparities for* $\mathrm{CpDT}_\lambda(2, k_0)$ *for different values of* k_0.

6.5. Plot over the sample sizes

We next considered the performance of the inlier modified test statistics for several sets of values of the tuning parameters to demonstrate the performance over increasing sample sizes. Here we have taken the same set of disparity test statistics as presented in Figure 5; the null and the alternative hypothesis are the ones presented in equations (3.1) and (3.2) respectively, where $k = 10$. All the tests are of level $\alpha = 0.05$. For brevity we only present the corrected powers at sample sizes varying from 25 to 300 for the case $\eta = 1.5$. The results clearly show that the power of the inlier modified tests are competitive with those of the ordinary disparity test $\mathrm{DT}_\lambda(3)$, which is expected to do well in this situation. Inlier modified tests corresponding to other values of η lead to similar conclusions in general.

7. Concluding remarks

In this paper our main goal has been to develop inlier modified goodness-of-fit tests; our purpose in doing so is to generate procedures which provide balanced performances irrespective of the alternative rather than variable performances fluctuating over extremes. Intuitively, disparity generating functions which are sharply increasing on both sides of the δ axis are likely to generate the kinds of tests we are looking for.

We have seen that the disparity test statistics with large positive values of λ perform well against bump alternatives while being poor in case of dip alter-

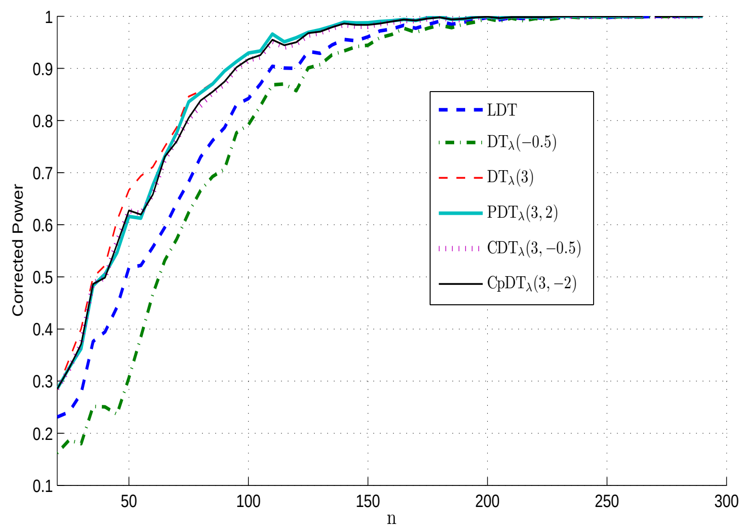

FIG 9. *Corrected powers of the disparity test statistics for different values of* n*, where* $\eta = 1.5$.

natives. The results are just the opposite for the statistics with large negative values of λ. In real life it is often the case that the direction of deviation from the null, if any, is unknown. In this respect a test which performs reasonably well in all situation may be preferred to other more extreme tests. Our inlier modified tests serve this objective.

In this paper we have focused on improving the small sample performance of the procedures. Although we do handle the necessary asymptotics, our emphasis is on small sample applications. In many cases we can actually derive the exact small sample distributions of the statistics as we have described earlier in the paper. In this connection it is worth mentioning that several approximations to the exact power functions of disparity based goodness-of-fit tests in small samples are available in the literature. See, for example, Cressie and Read (1984) [14], Read (1984a [27] and 1984b [28]), Shin *et al.* (1995 [30] and 1996 [31]) and Pardo (1998) [23]. While we do not explore this issue further, this technique may be a useful tool in our analysis since inlier modifications can make the chi-square approximation worse in small samples.

Our illustrations in this paper have been provided mostly in terms of the equiprobable null hypothesis. While this provides a very convenient platform for illustration, they are not, by any means the only type of models where these modifications are useful. The main issue is the nature of discrepancy between the observed and expected frequencies. In Example 6.2 the null hypothesis is of a time trend type (and not the equiprobable type) and the ordinary test statistics with large positive values of λ fail to detect the presence of a couple of large inliers. However, the situation changes dramatically when inlier correction is added.

Read and Cressie (1988, Section 7.2) [29] have considered the application of goodness-of-fit tests for continuous models based on the power divergence statistics. In our case, all the inlier modified statistics (except for the penalized statistics) will be meaningful and useful for testing goodness-of-fit hypotheses in continuous models. For obvious reasons, penalized disparities cannot be constructed in such cases.

Another important issue is the asymptotic distribution of the test statistics under contiguous alternatives. Cressie and Read (1984, Section 3) [14] have discussed the asymptotic distribution of the power divergence statistics in this case. We plan to study these distributions for the inlier modified statistics in a sequel paper.

Finally, we conclude with some brief reflections on whether it is meaningful to perform the exact opposite modifications to those considered in this paper. For example, one could start with a highly inlier sensitive disparity on the left and use modifications to make them outlier sensitive on the right. In some cases like combined and ϵ-combined distances the concepts coincide. In other cases such as those of the penalized disparities and coupled disparities, no natural analogues of the inlier modified tests are available, since the range of the argument δ of the G function is unbounded on the right. However the inlier shrunk disparity can be adapted in this case by modifying the G function on the right to get a test with the appropriate properties.

Acknowledgments

The authors gratefully acknowledge the comments of the Referee, the Editor and the Associate Editor which led to an improved version of the paper.

References

[1] Agresti, A. *Categorical data analysis.* John Wiley & Sons Inc., New York, 1990. MR1044993

[2] Alin, A. A note on penalized power-divergence test statistics. *Int. J. Math. Sci. (WASET) 1*, 3 (2007), 209–215 (electronic). MR2365638

[3] Alin, A. Comparison of Ordinary and Penalized Power-Divergence Test Statistics for Small and Moderate Samples in Three-Way Contingency Tables via Simulation. *Comm. Statist. Simulation Comput. 37*, 8 (2008), 1593–1602. MR2542419

[4] Alin, A., and Kurt, S. Ordinary and penalized minimum power-divergence estimators in two-way contingency tables. *Comput. Statist. 23*, 3 (2008), 455–468. MR2425172

[5] Basu, A., and Basu, S. Penalized minimum disparity methods for multinomial models. *Statist. Sinica 8*, 3 (1998), 841–860. MR1651512

[6] Basu, A., Harris, I., and Basu, S. Tests of hypotheses in discrete models based on the penalized Hellinger distance. *Statist. Probab. Lett. 27*, 4 (1996), 367–373.

[7] Basu, A., Harris, I. R., and Basu, S. Minimum distance estimation: the approach using density-based distances. In *Robust inference*, vol. 15 of *Handbook of Statist.* North-Holland, Amsterdam, 1997, pp. 21–48. MR1492707

[8] Basu, A., and Lindsay, B. G. Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Ann. Inst. Statist. Math. 46*, 4 (1994), 683–705. MR1325990

[9] Basu, A., Mandal, A., and Pardo, L. Hypothesis testing for two discrete populations based on the Hellinger distance. *Statist. Probab. Lett. 80* (2010), 206–214. MR2575447

[10] Basu, A., Ray, S., Park, C., and Basu, S. Improved power in multinomial goodness-of-fit tests. *The Statistician 51*, 3 (2002), 381–393. MR1920765

[11] Basu, A., and Sarkar, S. On disparity based goodness-of-fit tests for multinomial models. *Statist. Probab. Lett. 19*, 4 (1994), 307–312. MR1278665

[12] Birch, M. W. A new proof of the Pearson-Fisher theorem. *Ann. Math. Statist 35* (1964), 817–824. MR0169324

[13] Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. *Discrete multivariate analysis: theory and practice.* The MIT Press, Cambridge, Massachusetts, 1975. MR0381130

[14] Cressie, N., and Read, T. R. C. Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B 46*, 3 (1984), 440–464. MR0790631

[15] Csiszár, I. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magyar Tud. Akad. Mat. Kutató Int. Közl. 8* (1963), 85–108. MR0164374

[16] Harris, I. R., and Basu, A. Hellinger distance as a penalized log likelihood. *Comm. Statist. Simulation Comput. 23*, 4 (1994), 1097–1113. MR1309951

[17] Hout, M., Duncan, O., and Sobel, M. Association and heterogeneity: Structural models of similarities and differences. *Sociological Methodology 17* (1987), 145–184.

[18] Lindsay, B. G. Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Statist. 22*, 2 (1994), 1081–1114. MR1292557

[19] Mandal, A. *Minimum Disparity Inference: Strategies for Improvement in Efficiency.* PhD thesis, Indian Statistical Institute, Kolkata, India, 2010.

[20] Mandal, A., Bhandari, S. K., and Basu, A. Minimum disparity estimation based on combined disparities: Asymptotic results. *J. Statist. Plann. Inference 141*, 2 (2011), 701–710. MR2732941

[21] Mandal, A., Pardo, L., and Basu, A. Minimum disparity inference and the empty cell penalty: Asymptotic results. *Sankhyā Ser. A 72*, 2 (2010), 376–406. MR2746118

[22] Pardo, L., and Pardo, M. C. Minimum power-divergence estimator in three-way contingency tables. *J. Stat. Comput. Simul. 73*, 11 (2003), 819–831. MR2037848

[23] Pardo, M. C. Improving the accuracy of goodness-of-fit tests based on Rao's divergence with small sample size. *Comput. Statist. Data Anal. 28*, 4 (1998), 339–351. MR1659211

[24] Park, C., Basu, A., and Basu, S. Robust minimum distance inference based on combined distances. *Comm. Statist. Simulation Comput. 24*, 3 (1995), 653–673.

[25] Park, C., Basu, A., and Harris, I. R. Tests of hypotheses in multiple samples based on penalized disparities. *J. Korean Statist. Soc. 30*, 3 (2001), 347–366. MR1895981

[26] Patra, R. K., Mandal, A., and Basu, A. Minimum Hellinger Distance Estimation with Inlier Modification. *Sankhyā Ser. B 70*, 2 (2008), 310–322. MR2563992

[27] Read, T. R. C. Closer asymptotic approximations for the distributions of the power divergence goodness-of-fit statistics. *Ann. Inst. Statist. Math. 36*, 1 (1984a), 59–69. MR0752006

[28] Read, T. R. C. Small-sample comparisons for the power divergence goodness-of-fit statistics. *J. Amer. Statist. Assoc. 79*, 388 (1984b), 929–935. MR0770289

[29] Read, T. R. C., and Cressie, N. A. C. *Goodness-of-fit statistics for discrete multivariate data*. Springer Series in Statistics. Springer-Verlag, New York, 1988. MR0955054

[30] Shin, D. W., Basu, A., and Sarkar, S. Comparisons of the blended weight Hellinger distance based goodness-of-fit test statistics. *Sankhyā Ser. B 57*, 3 (1995), 365–376. MR1394883

[31] Shin, D. W., Basu, A., and Sarkar, S. Small sample comparisons for the blended weight chi-square goodness-of-fit test statistics. *Comm. Statist. Theory Methods 25*, 1 (1996), 211–226. MR1378956

[32] Yan, H. *Generalized minimum penalized Hellinger distance estimation and generalized penalized Hellinger deviance testing for discrete generalized linear models.* PhD thesis, Utah State University, Utah, USA, 2001.

[33] Zografos, K., Ferentinos, K., and Papaioannou, T. Divergence statistics: sampling properties and multinomial goodness of fit and divergence tests. *Communications in Statistics-Theory and Methods 19*, 5 (1990), 1785–1802. MR1075502