Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Minimum disparity estimation: Improved efficiency through inlier modification

Abhijit Mandal [a,*], Ayanendranath Basu [b]

[a] C.R. Rao AIMSCS, Hyderabad, India
[b] Indian Statistical Institute, Kolkata, India

ARTICLE INFO

Article history:
Received 8 May 2012
Received in revised form 13 February 2013
Accepted 23 February 2013
Available online 5 March 2013

Keywords:
Disparity
Inliers
Power divergence
Small sample studies

ABSTRACT

Inference procedures based on density based minimum distance techniques provide attractive alternatives to likelihood based methods for the statistician. The *minimum disparity estimators* are asymptotically efficient under the model; several members of this family also have strong robustness properties under model misspecification. Similarly, the *disparity difference tests* have the same asymptotic null distribution as the likelihood ratio test but are often superior than the latter in terms of robustness properties. However, many disparities put large weights on the *inliers*, cells with fewer data than expected under the model, which appears to be responsible for a somewhat poor efficiency of the corresponding methods in small samples. Here we consider several techniques which control the *inliers* without significantly affecting the robustness properties of the estimators and the corresponding tests. Extensive numerical studies involving simulated data illustrate the performance of the methods.

1. Introduction

In parametric estimation there are two fundamental ideas – efficiency at the model and robustness away from it – which are often in conflict. The maximum likelihood estimator (MLE) is asymptotically efficient at the model under standard regularity conditions, but often has poor robustness properties. On the other hand classical robust estimators – those based on M-estimation and its extensions – usually sacrifice first order efficiency at the model to achieve their robustness (*e.g.*, Hampel et al., 1986).

Some density-based minimum distance estimators, such as the minimum Hellinger distance estimator (MHDE), have been shown to attain *first order efficiency* at the model together with strong robustness properties. Beran (1977) seems to be the first to draw the attention of the statistical community to the robustness properties of the MHDE, although minimum distance estimation based on chi-square type distances has been around in the literature for some time (*e.g.*, Pearson, 1900, Rao, 1957).

Our focus in this paper will be on the chi-square type distances. In the literature this class of distances has also been referred to as the class of *disparities* (Lindsay, 1994) or as the class of *ϕ-divergences* (Csiszár, 1963; Vajda, 1989; Pardo, 2006). Here we will follow the approach of Lindsay since it provides a nice geometrical insight into the robustness of the resulting procedure.

The properties of disparities are characterized by a key function called the *residual adjustment function* (RAF) which is useful for demonstrating the adjustments between robustness and efficiency. Often, however, the particular structure

[*] Corresponding author.
E-mail address: abhijit_v@isical.ac.in (A. Mandal).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.02.030

of disparities which leads to the downweighting of *outliers*, cells with more data than expected under the model, is also responsible for an unsatisfactory treatment of *inliers*, cells with less data than expected under the model, leading to poor small sample efficiency. We provide a comprehensive description of inlier control strategies including new proposals, consolidation of scattered existing results, and a thorough comparison of these techniques. In the current literature the role of the inliers in minimum distance estimation is not sufficiently explored. We hope that this work will erase this deficiency, at least partially, and lead to a better appreciation of the inlier problem. We will restrict ourselves to the case of discrete models, since empirical evidence shows that improvements due to inlier control are much more pronounced in this case.

The rest of the paper is organized as follows. In Section 2 we introduce disparity based inference. In Section 3 we provide some numerical studies to justify the need for inlier control strategies. We define the inlier control strategies in Section 4. In Section 5 the improved performances of the inlier modified estimators and tests are numerically demonstrated. Concluding remarks are presented in Section 6.

2. Minimum disparity inference

Let $X_1, X_2, \dots, X_n$ be n independent and identically distributed observations from a discrete distribution F having probability mass function f with respect to the appropriate dominating measure. The true probability mass function f will be modelled by the family of probability mass functions $\{m_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$. Let $\mathcal{X}$ be the sample space and let $d_n(x)$ be the proportion of sample observations at $x \in \mathcal{X}$. We define the *Pearson residual* function $\delta_{n\theta}(x)$ by the relation $\delta_{n\theta}(x) = \frac{d_n(x) - m_\theta(x)}{m_\theta(x)}$, $x \in \mathcal{X}$. Let G be a real-valued, thrice differentiable, strictly convex function on $[-1, \infty)$ with $G(0) = 0$. The *disparity* $\rho_G(d_n, m_\theta)$ between the probability vectors d_n and m_θ based on G is defined as

$$\rho_G(d_n, m_\theta) = \sum_{x \in \mathcal{X}} G(\delta_{n\theta}(x)) m_\theta(x). \tag{2.1}$$

The function G is called the disparity generating function of the above measure. Let $\hat{\theta}_n$ minimize ρ_G over $\theta \in \Theta$, provided such a minimizer exists; $\hat{\theta}_n$ is 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

$$\sum_{x \in \mathcal{X}} A_G(\delta_{n\theta}(x)) \nabla m_\theta(x) = 0, \tag{2.2}$$

where $A_G(\delta) = (1 + \delta)G'(\delta) - G(\delta)$, 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; 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; Basu et al., 1997). These two conditions are automatic if the associated G function satisfies

$$G'(0) = 0, \quad \text{and} \quad G''(0) = 1, \tag{2.3}$$

where G' and G'' are the indicated derivatives of G. The strict convexity of G implies that A_G is increasing. 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

$$G(\delta) = \int_0^\delta \int_0^t A'_G(s)(1+s)^{-1} ds\, dt. \tag{2.4}$$

There are several important subfamilies of the class of disparities which include the power divergence family (Cressie and Read, 1984) given by

$$\mathrm{PD}_\lambda(d_n, m_\theta) = \frac{1}{\lambda(\lambda+1)} \sum_{x \in \mathcal{X}} d_n(x) \left\{ \left(\frac{d_n(x)}{m_\theta(x)} \right)^\lambda - 1 \right\}, \quad \lambda \in \mathbb{R}, \tag{2.5}$$

with associated G function $G_\lambda(\delta) = \frac{(\delta+1)^{\lambda+1} - (\delta+1)}{\lambda(\lambda+1)} - \frac{\delta}{\lambda+1}$. For $\lambda = 0$ and $\lambda = -1$ the divergences are the limits of the above expressions as $\lambda \to 0$ and $\lambda \to -1$ respectively. The values $\lambda = 1, 0, -1/2$ and -2 generate the Pearson's chi-square (PCS), the likelihood disparity (LD), the (twice, squared) Hellinger distance (HD) and the Neyman's chi-square (NCS) respectively. The LD and HD have the form

$$\mathrm{LD}(d_n, m_\theta) = \sum_{x \in \mathcal{X}} \left[d_n(x) \log \left(\frac{d_n(x)}{m_\theta(x)} \right) + (m_\theta(x) - d_n(x)) \right] = \sum_{x \in \mathcal{X}} d_n(x) \log \left(\frac{d_n(x)}{m_\theta(x)} \right), \tag{2.6}$$

$$\mathrm{HD}(d_n, m_\theta) = 2 \sum_{x \in \mathcal{X}} \left(d_n^{1/2}(x) - m_\theta^{1/2}(x) \right)^2, \tag{2.7}$$

with associated G functions $G_{\mathrm{LD}}(\delta) = (\delta + 1)\log(\delta + 1) - \delta$ and $G_{\mathrm{HD}}(\delta) = 2[(\delta + 1)^{1/2} - 1]^2$. The MLE is the MDE corresponding to the LD. Notice that the RAF of the LD is linear, satisfying $A_{\mathrm{LD}}(\delta) = \delta$.

Other subfamilies of disparities include the blended weight Hellinger distance (Lindsay, 1994; Basu and Lindsay, 2004), defined by

$$\mathrm{BWHD}_\beta(d_n, m_\theta) = \frac{1}{2}\sum_{x\in\mathcal{X}} \frac{(d_n(x) - m_\theta(x))^2}{(\beta d_n^{1/2}(x) + \bar{\beta} m_\theta^{1/2}(x))^2}, \quad \beta \in [0, 1],\ \bar{\beta} = 1 - \beta. \tag{2.8}$$

The G function for this subfamily is given by $G_\beta(\delta) = \delta^2/\{2[\beta(\delta+1)^{1/2}+\bar{\beta}]^2\}$. 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), defined by

$$\mathrm{BWCS}_\tau(d_n, m_\theta) = \frac{1}{2}\sum_{x\in\mathcal{X}} \frac{(d_n(x) - m_\theta(x))^2}{\tau d_n(x) + \bar{\tau} m_\theta(x)}, \quad \tau \in [0, 1],\ \bar{\tau} = 1 - \tau. \tag{2.9}$$

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

Let $X_1, \ldots, X_n$ be independent and identically distributed observations from a discrete distribution modelled by $m_\theta(x)$, $\theta \in \Theta \subseteq \mathbb{R}^p$. Let the true distribution belong to the model with θ^0 being the true parameter. The following theorem sets the background of our paper.

Theorem 2.1 (*Theorem 33 Lindsay, 1994*). *Under the regularity conditions in* Sarkar and Basu (1995, p. 356), *the minimum disparity estimating equation* $\frac{\partial}{\partial\theta}\rho_G(d_n, m_\theta) = 0$ *has a consistent sequence of roots* $\hat{\theta}_n$, *that satisfy*

$$n^{1/2}(\hat{\theta}_n - \theta^0) \overset{a}{\sim} N_p\left(0, I^{-1}(\theta^0)\right),$$

where the $\overset{a}{\sim}$ *represents asymptotic distribution, and* $I(\theta)$ *is the Fisher information matrix at* m_θ.

Under proper regularity conditions all minimum disparity estimators are first order efficient and asymptotically there is nothing to distinguish between them. However, as we will see in the next section there are significant differences in their small sample performances at the model.

2.1. Disparity difference tests

As in robust estimation, the disparity statistics play an important role in the case of robust testing of hypothesis. One of the most popular tools in testing of hypothesis is the likelihood ratio test (LRT); see *e.g.*, Neyman and Pearson (1928). The LRT is routinely used in hypothesis testing problems and enjoys certain optimality properties under some standard regularity conditions.

However, just as the maximum likelihood estimator has many robustness problems, the performance of the LRT is also often significantly affected by model misspecification and the presence of outliers. In later sections we will see that the presence of even a very small proportion of extreme outliers can have a severe effect on the level and the power of the LRT. As an alternative, one can construct tests based on other disparities to get a better control over outliers. The class of these tests will be referred to as the class of disparity difference tests (DDTs) of which the LRT is a special case. Some members of this family possess strong robustness properties unlike the LRT. Moreover, the asymptotic null distributions of these DDTs are equivalent to that of the LRT when the model is correctly specified; under stronger assumptions, this equivalence holds for local alternatives as well.

As in the case of estimation, the methods based on the Hellinger distance have a special place in the class of the DDTs and historically developed earlier than the other tests. Simpson (1989) has discussed some robustness properties of the DDT using the Hellinger distance.

Suppose we have a random sample $X_1, X_2, \ldots, X_n$ of independent and identically distributed observations from the true probability mass function f with countable support, and let $\{m_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ be the parametric family of mass functions modelling it. Let us consider testing the following hypothesis

$$H_0 : \theta \in \Theta_0 \quad \text{against } H_1 : \theta \in \Theta - \Theta_0, \tag{2.10}$$

where Θ_0 is a proper subset of Θ. The LRT statistic can be expressed as

$$\mathrm{LRT} = 2\left[\log\left(\prod_{i=1}^n m_{\hat{\theta}_n^{\mathrm{ML}}}(X_i)\right) - \log\left(\prod_{i=1}^n m_{\hat{\theta}_n^{0\mathrm{ML}}}(X_i)\right)\right] = 2n\left[\mathrm{LD}(d_n, m_{\hat{\theta}_n^{0\mathrm{ML}}}) - \mathrm{LD}(d_n, m_{\hat{\theta}_n^{\mathrm{ML}}})\right],$$

where $\hat{\theta}_n^{\mathrm{ML}}$ and $\hat{\theta}_n^{0\mathrm{ML}}$ are the unrestricted MLE and the constrained MLE respectively. In analogy to the LRT, one may define the disparity difference test based on the disparity ρ_G as

$$\mathrm{DDT}_G = 2n\left[\rho_G(d_n, m_{\hat{\theta}_n^0}) - \rho_G(d_n, m_{\hat{\theta}_n})\right], \tag{2.11}$$

where $\hat{\theta}_n$ and $\hat{\theta}_n^0$ are the unrestricted MDE and the constrained MDE respectively, corresponding to the disparity ρ_G. Thus we may refer to the LRT as the DDT based on the likelihood disparity.

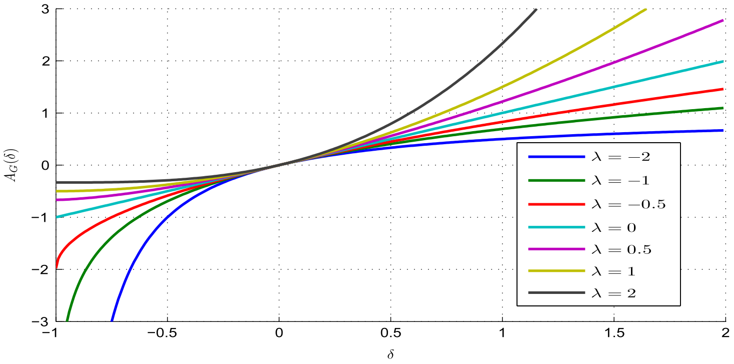

Fig. 1. Plot of the RAF $A_{G_\lambda}(\delta)$ for different values of λ in the case of the power divergence family.

Let Θ_0 be the subset of Θ with $r \le p$ restrictions on the vector θ such that $C_i(\theta) = 0,\ i = 1, 2, \dots, r$. We consider the hypothesis given in (2.10). Under this null, the set up may be described by the parameter $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_{p-r})^T$ with $p-r$ independent components, where $\gamma \in \Gamma \subseteq \mathbb{R}^{p-r}$. In this case there exists a function $\eta : \mathbb{R}^{p-r} \to \mathbb{R}^p$ such that $\theta = \eta(\gamma)$, where $\theta \in \Theta_0$ and $\gamma \in \Gamma$. If H_0 is true, then there exists a $\gamma^0 \in \Gamma$ such that $\theta^0 = \eta(\gamma^0)$, where $\theta^0 \in \Theta_0$ is the true value of θ. We assume that η has continuous second derivatives in an open set containing γ^0. Suppose the first derivative $\dot{\eta}(\gamma)$ of order $p \times (p-r)$ has full column rank at $\gamma = \gamma^0$. The next theorem forms the basis of our comparison of the different DDTs.

Theorem 2.2 (*Lemma 2.1 Sarkar and Basu, 1995*). *Under the regularity conditions considered in* Sarkar and Basu (1995, p. 356), *the null distribution of the disparity difference test statistic in (2.11) tends to a χ^2 distribution with r degrees of freedom as $n \to \infty$.*

In the following sections we will present an extensive numerical study to explore the properties of the MDEs and the corresponding DDTs as well as their inlier modified versions. We will primarily focus on the power divergence family, but will also use some other disparities for illustration. For a streamlined presentation, we use the following convention for the estimators and the tests. We denote the minimum disparity estimator as MDE, 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 MDE$_\lambda(2)$ will represent the MDE within the power divergence family with tuning parameter 2, MDE$_\beta(0.5)$ will represent the MDE within the blended weight Hellinger distance family with tuning parameter 0.5, and so on.

3. The small sample deficiency of the robust MDEs and DDTs

An observation $x \in \mathcal{X}$ having a large positive value of $\delta(x)$ may be called an *outlier* in the sense that it has a much larger observed proportion than what is predicted by the model; this is a probabilistic rather than a geometric characterization. To protect against outliers, one should choose such disparities which give small weights to observations having large positive values of δ. For such disparities the RAF $A_G(\delta)$ would exhibit a severely dampened response to increasing δ. For a qualitative description, one can take the RAF of the likelihood disparity $A_{\mathrm{LD}}(\delta)$ as the basis for comparison. For this disparity $A_{\mathrm{LD}}(\delta) = \delta$, and thus to compare the other minimum disparity estimators with the maximum likelihood estimator one must focus on how their RAFs depart from linearity for large positive δ. A graph of the RAFs of some of the common disparities within the power divergence family is given in Fig. 1. Disparities with large negative values of λ, for which the RAFs curve sharply down on the right hand side of the δ axis, are expected to perform better in terms of robustness and be more stable under data contamination.

On the other hand, a point $x \in \mathcal{X}$ is an *inlier* if the corresponding Pearson residual $\delta(x)$ is negative; here the observed proportion is smaller than what is predicted by the model. Most naturally available disparities fail to provide a balanced treatment of outliers and inliers simultaneously. In Fig. 1 it is observed that within the power divergence family the RAFs which downweight large outliers end up magnifying the effect of large inliers. There is overwhelming empirical evidence to suggest that improper treatment of inliers adversely effects small sample efficiency. While the MHDE performs very satisfactorily in dealing with outlying observations relative to the MLE, it is less stable than the latter under the presence of inliers. Simpson (1989), Lindsay (1994), Basu et al. (1997) and Bhandari et al. (2006), among others, have highlighted this issue.

We now present some numerical studies to describe the efficiency problems faced by robust minimum disparity procedures. In the first experiment data are randomly generated from a Poisson distribution with mean 5. Samples of size 10–200 are drawn, the relevant MDE is calculated, and the exercise is replicated 1000 times at each sample size. The Poisson model is assumed, and six estimators of the population mean, the MLE, MDE$_\lambda(-0.7)$, MDE$_\lambda(-0.6)$, MDE$_\lambda(-0.5)$, MDE$_\beta(0.6)$ and MDE$_\tau(0.8)$, are chosen. The observed mean square errors (MSEs) of our estimators (multiplied by n) are plotted against the sample size, and presented in the same graph (Fig. 2) for comparison. Other than the MLE, all the other estimators are highly robust and strongly downweight the effect of larger outliers (while also magnifying the effect of large inliers). The figure shows that in this case the MSEs of all the MDEs considered are significantly higher compared to that of the MLE in small to moderate sample sizes. Even at a sample size of 200, (n times) the observed MSE of all the estimators appear to be significantly above 5, which is the common asymptotic limit of $n \times$ MSE for each of the estimators. This is a price a practitioner would hate to pay even when the gain in robustness is substantial.

Next we compare the observed levels of some robust DDTs. Here also we generate random samples from the Poisson(5) distribution, and assume a Poisson(θ) model. We have compared six tests, the LRT, $\text{DDT}_\lambda(-0.7)$, $\text{DDT}_\lambda(-0.6)$, $\text{DDT}_\lambda(-0.5)$, $\text{DDT}_\beta(0.6)$ and $\text{DDT}_\tau(0.8)$. We test the null hypothesis $H_0 : \theta = 5$ against the two sided alternative. The observed levels, the proportion of test statistics exceeding the χ^2 critical value, are plotted in Fig. 2 against the sample size using 2000 replications (nominal level is 5%). We note that the observed level of the robust DDTs are considerably high compared to that of the LRT (or the nominal level) even at sample sizes as high as 200. This severely limits the use of these tests, since it is hard to put any great value on the power of a test that cannot hold its level.

The poor small sample behaviour of the robust MDEs is unfortunate, since the estimators are otherwise desirable because of their robustness properties. As this deficiency appears to be primarily attributable to the improper handling of inliers, we aim to develop several classes of inlier modification techniques which improve the small sample performance without compromising the robustness properties of these procedures. We preserve the behaviour of these procedures on the outlier side, and apply suitable modifications to the inlier segment. We also provide some discussion on the choice of the appropriate parameters to get the best results.

4. Methods for inlier control

In this section we consider several inlier modification techniques. The underlying common theme in all of them is that they try to suitably modify the inlier part of such disparities which naturally provide a controlled treatment of outliers. 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, and subsequently we recover the form of the disparity generating function using (2.4).

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; Basu et al., 1996; Basu and Basu, 1998; Park et al., 2001; Basu et al., 2002; Pardo and Pardo, 2003; Alin, 2007; Basu et al., 2010), the combined disparities (*e.g.*, Park et al., 1995; Basu et al., 2002; Mandal et al., 2011), the coupled disparities, the ϵ-combined disparities and the inlier shrunk disparities (Patra et al., 2008).

It may be noted that the method based on the empty cell penalty has been studied in fair detail in the literature. The methods of combined and inlier shrunk disparities have been also proposed earlier, but rather insufficiently explored. The methods based on coupled and ϵ-combined disparities are entirely new proposals. However, none of these methods, including the method based on empty cell penalties have been studied in as much detail as in Mandal (2010), which provides a comprehensive comparison of these tools.

4.1. Penalized disparities

Let us assume the set up of Section 2. Suppose the disparity generating function G satisfies the conditions in (2.3) in addition to its usual properties. The disparity in (2.1) can be rewritten as

$$\rho_G(d_n, m_\theta) = \sum_{x:d_n(x)>0} G(\delta_{n\theta}(x))m_\theta(x) + G(-1)\sum_{x:d_n(x)=0} m_\theta(x). \tag{4.1}$$

This shows that the natural weight for the set $\{x : d_n(x) = 0\}$, *i.e.* the empty cells, is $G(-1)$; for the power divergence family this equals $1/(\lambda + 1)$ so that this is very large for values of λ close to but larger than -1. So those disparities give an unduly large weight to the empty cells. For $\lambda \leq -1$, these disparities are not defined even for one empty cell. For a parametric model m_θ with infinite support, the set $\{x : d_n(x) = 0\}$ is also infinite. Thus the proper control of the term $\sum_{x:d_n(x)=0} m_\theta(x)$ can lead to large benefits when the natural weight $G(-1)$ of the empty cell is too large.

The penalized disparity between the densities d_n and m_θ for the penalty weight h is defined as

$$\rho_{G_h}(d_n, m_\theta) = \sum_{x:d_n(x)>0} G(\delta_{n\theta}(x))m_\theta(x) + h\sum_{x:d_n(x)=0} m_\theta(x), \quad h > 0, \tag{4.2}$$

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 d_n and m_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 m_θ is independent of θ, the range of h can be enhanced to include $h = 0$.

The minimum penalized disparity estimator (MPDE) $\hat\theta_n^h$ is obtained by minimizing $\rho_{G_h}(d_n, m_\theta)$ over $\theta \in \Theta$, where ρ_{G_h} is as given in (4.2). So

$$\rho_{G_h}(d_n, m_{\hat\theta_n^h}) = \min_{\theta\in\Theta} \rho_{G_h}(d_n, m_\theta),$$

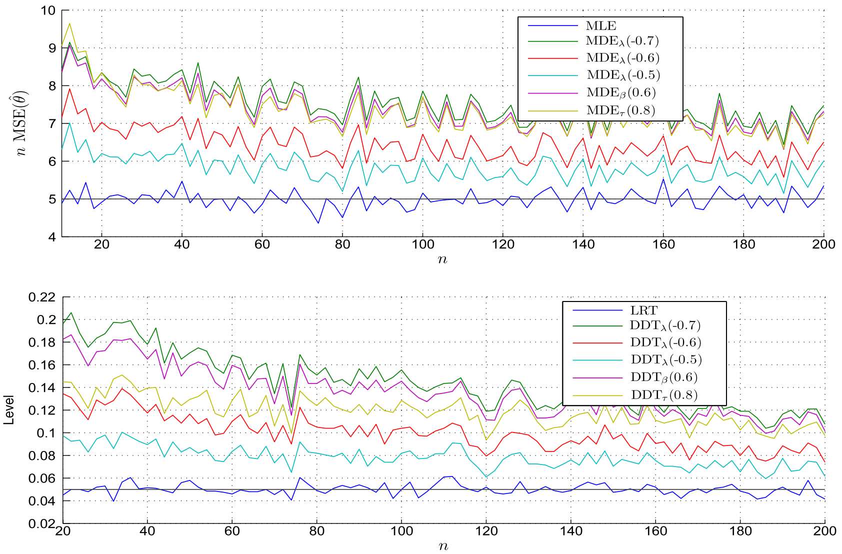

Fig. 2. Plot of (n times) observed MSE of different MDEs (above) and the observed level (nominal level is 5%) of different DDTs (below) against the sample size.

provided such a minimum exists. The results of Mandal et al. (2010) ensure that the minimum penalized disparity estimators are best asymptotically normal (BAN) and the asymptotic null distribution of the disparity difference test for (2.10) based on the penalized disparity is the same as that of the ordinary disparity difference test statistic.

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 $\rho_{G_c}(d_n, m_\theta) = \sum_{x\in\mathcal{X}} G_c(\delta_{n\theta}(x))m_\theta(x)$, where

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

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

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

When using the combined disparity approach, our aim will be to combine the RAF of an outlier robust disparity on the positive side of the δ axis with a RAF which provides a controlled treatment of inliers on the negative side of the δ axis. To achieve our general purpose sometimes it is easier to begin with the residual adjustment function of an outlier stable RAF and modify it on the inlier side to arrest its sharp natural decline. 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). Similarly the asymptotic null distribution the disparity difference test for (2.10) based on the combined disparity is the same as that of ordinary disparity difference test statistic.

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

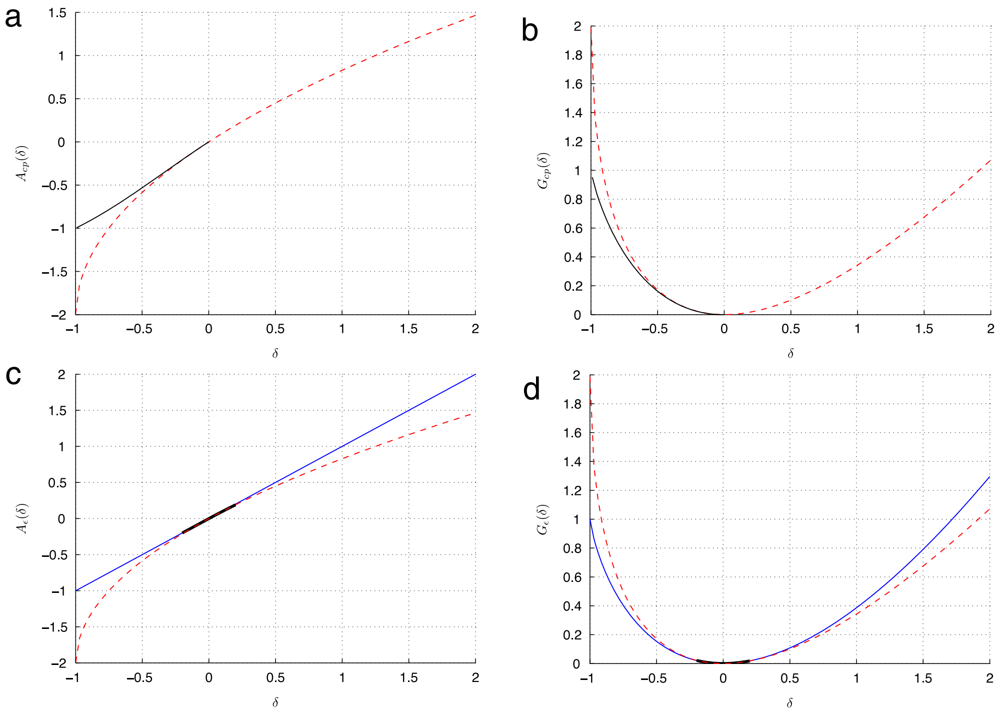

Fig. 3. (a) and (b) A_{cp} and G_{cp} functions respectively of the coupled disparity where the Hellinger distance (dashed line) is on the outlier part and $k_0 = -1$. (c) and (d) A_ϵ and G_ϵ functions respectively of the ϵ-combined disparity where the Hellinger distance (dashed line) is on the outlier part and the likelihood disparity (straight line) is on the inlier part and $\epsilon = 0.2$.

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. Proper choices of k_0 can prevent the RAF from dipping too far down on the inlier side. The residual adjustment function for the likelihood disparity satisfies $A_{\text{LD}}(-1) = -1$, and generally a choice of k_0 around -1 works satisfactorily in controlling the inliers.

Conditions 1–3 lead to four algebraic constraints on $A_{\text{cp}}(\delta)$, *viz.* $A_{\text{cp}}(0) = 0$, $A'_{\text{cp}}(0) = 1$, $A''_{\text{cp}}(0) = A''_G(0)$ and $A_{\text{cp}}(-1) = k_0$. We therefore assume that $A_{\text{cp}}(\delta)$ 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

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

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

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

Using Eq. (2.4), the corresponding reconstructed G function is given by

$$G_{\text{cp}}(\delta) = \frac{1}{4}(A_2 - 2k_0 - 2)\,\delta^3 - \frac{1}{4}(A_2 - 6k_0 - 6)\,\delta^2 - \frac{1}{2}(A_2 - 6k_0 - 4)\,(\delta - (1 + \delta)\log(1 + \delta)),$$

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 $k_0 \le -\frac{1}{6}(A_2 + 4)$ (see Mandal, 2010), and in that case the coupled disparity satisfies all the properties of a regular disparity.

4.4. ϵ-combined disparities

In this case 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 (see Fig. 3). This smooth joining allows the combined function to retain the second order smoothness properties at $\delta = 0$. The ϵ-combined disparity is similar in spirit to the combined disparity approach but avoids the resulting lack of smoothness caused by the brute force merger of the combined disparity case. The residual adjustment function of the ϵ-combined disparity is defined by

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

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 for δ in $[-\epsilon, \epsilon]$ serves our 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 distributions of the resulting minimum ϵ-combined disparity estimator and the corresponding disparity difference test statistic again follow from existing results. However, also see the comments in Remark 1 later.

4.5. *The inlier-shrunk disparities*

Let G be a function satisfying the disparity conditions in Eq. (2.3). Define the corresponding inlier-shrunk class of disparity generating functions indexed by the inlier shrinkage 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. Clearly there is no shrinkage for the case $\gamma = 0$; on the other hand, as the parameter γ increases, the inlier component is subjected to greater shrinkage. 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 distributions of the minimum inlier shrunk disparity estimator and the corresponding disparity difference test will still be as given in Theorems 2.1 and 2.2.

In Section 2 we have already presented a convention for expressing our estimators and tests in a streamlined fashion. It is necessary to build on this convention and to add an inlier component to this notation. A second argument will represent the inlier modification parameter. In the following the term MDE will represent a generic minimum disparity estimator. In addition, the letters ‘P’, ‘C’, ‘Cp’ and ‘IS’ will represent the penalized, combined, coupled and inlier shrunk disparities or the corresponding estimators/tests respectively. Thus $\text{MPDE}_\lambda(-0.5, 1)$ will represent the minimum penalized disparity estimator in the power divergence family with tuning parameter -0.5 and penalty weight 1. Similarly $\text{MCpDE}_\beta(0.6, -1)$ will represent the estimator obtained by minimizing the coupled disparity in the BWHD family with tuning parameter 0.6 and intercept parameter -1 and so on. The corresponding disparity difference tests with be denoted by $\text{PDDT}_\lambda(-0.5, 1)$, CpDDT_β $(0.6, -1)$ *etc.*

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 (2.3) 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 distributions of the corresponding estimator and test statistic continue to hold as long as conditions of Theorem 2.1 are satisfied. In the 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 Theorems 2.1 and 2.2 will hold for the inlier shrunk disparities.

5. Simulation results

Although theoretical indicators can give some idea about the behaviour of our procedures, extensive numerical work is also necessary to illustrate their performance and supplement the theory developed. In this section we consider and undertake appropriate numerical studies for this purpose.

We have presented the results in the context of the penalized, combined, coupled and inlier shrunk disparities in this section. We have not provided additional results for the ϵ-combined disparity since it appears to provide results that are very close to those of the combined disparity case when ϵ is small.

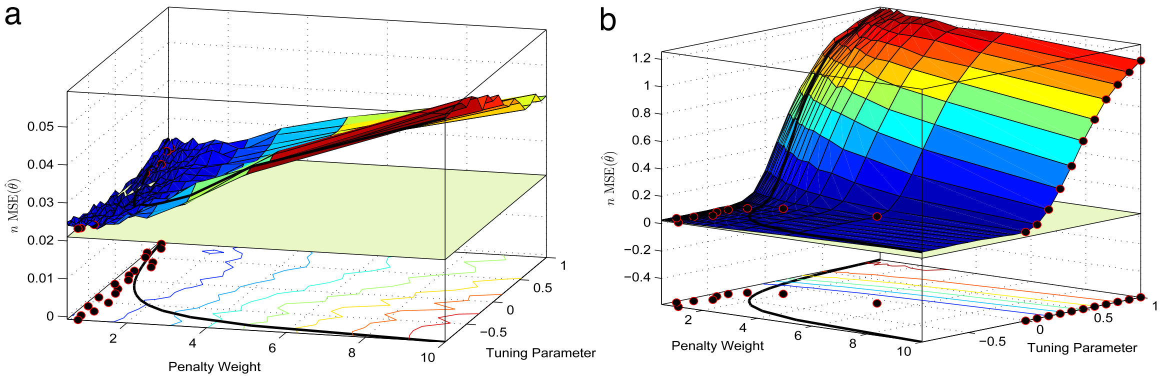

Fig. 4. Observed MSE surfaces (multiplied by n) of the MPDE$_\lambda$ in the case of (a) the pure and (b) the contaminated binomial models for different values of the tuning parameter and the empty cell penalty weight. The minimum MSEs over the penalty weight for fixed values of the tuning parameter are marked by black dots (both on the surface and the base). The black line represents the ordinary MDEs.

5.1. Inlier modified estimators

In this section we will calculate the mean square error (MSE) of different inlier modified estimators for pure as well as contaminated data. Subsequently we will provide, based on our numerical results, some guidelines about choosing the tuning parameters and the inlier parameters to get good efficiency in small sample sizes with little or no loss in robustness.

5.1.1. Minimum penalized disparity estimator (MPDE)

Here we present 3D plots of the MSE surface of the MPDEs for different combinations of the tuning parameter of the power divergence statistics and the empty cell penalty weight. The MSE contour is plotted over the two dimensional (tuning parameter, penalty weight) plane. The minimum MSEs (over the penalty weight for fixed values of the tuning parameter) are marked by dots on the surface as well as on the base on the plot. The black lines on the surface and the base of the plot describe the points which correspond to the ordinary minimum disparity estimators.

For the example with the minimum penalized disparity estimators we have chosen the Binomial (10, θ) model where θ is the parameter of interest; the data are generated from the Binomial (10, 0.3) distribution. Plot (a) of Fig. 4 presents (n times) the observed MSE of MPDEs of θ around the true value 0.3 in 2000 replications. The sample size in each replication is $n = 25$. It is clear from the plot that very high weights on the empty cells increase the MSE of the estimators. That is why the small sample efficiency of the robust estimators, *i.e.* those corresponding to disparities with large negative values of the tuning parameter, is poor. Although the optimal value of the penalty seems to depend on the model, the parameters, and the particular disparity used, choice of penalty weights between 0.5 and 1 appear to be reasonable for most of our robust estimators. The figure also shows that the amount of improvement due to the application of the penalty can be really huge for the more robust members of the class, *i.e.* for those with tuning parameter close to −1. In particular (n times) the MSE of MDE$_\lambda(-0.9)$ equals 0.0583, while those of MPDE$_\lambda(-0.9, 1)$ and MLE equal 0.0264 and 0.0215 respectively.

The improvement of the small sample performance at the model due to the application of the penalty will mean very little if the modified estimator is not able to retain its robustness characteristics under data contamination. To verify this, we next estimate the parameter θ under the binomial model when the data are actually simulated from the $(1-\epsilon)\text{Bin}(10, 0.3)+\epsilon\chi_{10}$ distribution, where Bin(n, θ) represents the binomial distribution with the indicated parameters, and χ_{10} is a unit mass distribution at the fixed point 10. Bin(10, 0.3) is our target distribution, and the MSE of the estimator is still computed against the target value of 0.3. The values of the contamination proportion ϵ and the sample size n are 0.1 and 25 respectively; the results are based on 2000 replications. Fig. 4(b) shows the MSE contours in this case over the (tuning parameter, penalty weight) plane. Clearly the inlier modified estimators corresponding to large negative values of the tuning parameter and penalty weights in the range (0.5, 1) hold up their own against data contamination in this example. For example, (n times) the MSEs for MDE$_\lambda(-0.9)$ and MPDE$_\lambda(-0.9, 1)$ are 0.0663 and 0.0287 respectively. On the other hand, the estimators corresponding to large positive values of the tuning parameter perform miserably, irrespective of the value of the penalty weight; in particular, (n times) the MSE of the MLE, *i.e.* MDE$_\lambda(0)$, is 0.1864.

Here, as well as in the other figures of this section, the scales of the vertical axis for the pure data and the contaminated data plots are different to accommodate the full impact of the distortion due to contamination in the latter case.

5.1.2. Minimum combined disparity estimator (MCDE)

To illustrate the performance of the MCDE we have chosen the Poisson (θ) model; we are interested in estimating the parameter θ. The true data generating distribution is Poisson with mean parameter 3. In Fig. 5(a) we present 3D plots of (n times) the observed MSE surface of the MCDEs of θ for several combined disparities. These combined disparities are generated by combining two different power divergence statistics; the parameters for the outlier and the inlier parts are

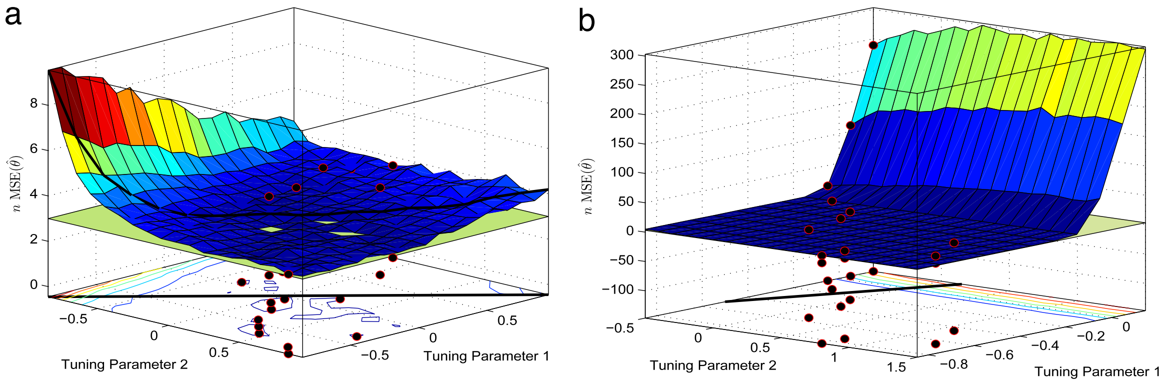

Fig. 5. Observed MSE surfaces (multiplied by n) of the MCDE$_\lambda$ in the case of (a) the pure and (b) the contaminated Poisson models for different values of the outlier parameter (tuning parameter 1) and the inlier parameter (tuning parameter 2). The minimum MSEs over tuning parameter 2 for fixed values of tuning parameter 1 are marked by black dots. The black line represents the ordinary MDEs.

referred to as the tuning parameter 1 and tuning parameter 2 respectively. The number of replications is 2500, and the sample size in each replication is $n = 20$. The minimum MSEs (over tuning parameter 2 for fixed values of tuning parameter 1) are marked by dots on the surface as well as on the base of the plot. The black lines on the surface and the base of the plot indicate the points where the two tuning parameters are equal, representing to the ordinary minimum disparity estimators. The figure clearly demonstrates the huge benefits due to inlier modification. For example, while (n times) the MSE of the ordinary estimator MDE$_\lambda(-0.75)$ equals 6.6253, the combined estimator MCDE$_\lambda(-0.75, 0)$ has (n times) a MSE of only 3.4232; here -0.75 and 0 represent the tuning parameters 1 and 2 respectively. The maximum likelihood estimator MDE$_\lambda(0)$ has (n times) a MSE of 3.1733. It appears that the estimators corresponding to combined disparities work well when tuning parameter 2 is close to the negative of tuning parameter 1.

Data are then generated from the contaminated Poisson mixture $(1-\epsilon)$Poiss$(3) + \epsilon\chi_{20}$, where Poiss(θ) represents the Poisson distribution with the indicated parameter; the results presented here correspond to $\epsilon = 0.1$. The model is still assumed to be a Poisson, and (n times) the observed MSE surface of the MCDEs of the Poisson mean parameter (calculated against the target value of 3) are presented in Fig. 5(b). It is seen that when the value of tuning parameter 1 is greater than or equal to zero the MSE values are clearly unacceptable. On the other hand, the inlier modification appears to have very little effect on the MSE of the robust estimators within the range of interest. For example, (n times) the MSE of MDE$_\lambda(-0.75)$ is 4.1597 in this case, while that for MCDE$_\lambda(-0.75, 0)$ is 3.5584. In comparison, the maximum likelihood estimator MDE$_\lambda(0)$ has (n times) a MSE of 165.1721. Clearly the inlier modification has not led to any concession in terms of robustness in this case.

5.1.3. Minimum coupled disparity estimator (MCpDE)

The coupled disparities in this section are constructed with a suitable member of the power divergence family on the outlier side with an appropriate intercept parameter k_0 for the inlier part. Here we have taken a Poisson model, and the data generating distribution is Poisson with mean 5. Fig. 6(a) presents (n times) the observed MSE of MCpDEs of the mean parameter θ under the Poisson model in 2000 replications; (n times) the MSE contour is plotted over the (tuning parameter, intercept parameter) plane. The sample size n in each replication is 25. The figure shows that for MCpDEs with large negative values of the tuning parameter, efficiency increases with the intercept parameter becoming optimal at some value between $(-1, -0.5)$ depending on the tuning parameter. The dots on the surface (as well as on the base) of the plot represent the optimal intercept parameter for fixed values of the tuning parameter in terms of controlling the MSE. The black line on the surface and the base of the plots indicates the coupled disparities whose intercepts are the same as those of the ordinary disparities. In general, a value of -1 appears to be a good default value of the intercept parameter for large negative values of the tuning parameter. In particular (n times) the MSE of MDE$_\lambda(-0.55)$ equals 6.0503, while those of MCpDE$_\lambda(-0.55, -1)$ and MLE equal 5.4940 and 4.6139 respectively.

Next we examine the effect of the intercept parameter on robustness of MCpDE$_\lambda$. Data are simulated from the $(1-\epsilon)$ Poiss$(5) + \epsilon\chi_{20}$ distribution, and we use $\epsilon = 0.1$ in the calculations below. The sample size is 20 and the experiment is replicated 2000 times for the different combinations of the tuning parameter and intercept parameter. For this case, Plot (b) of Fig. 6 shows that while the MSE of the MCpDE increases rapidly for tuning parameter values greater than zero, it is remarkably stable for negative values of the tuning parameter. Also, the MSE has little variation over the intercept parameter for any value of the tuning parameter in this case. Thus the intercept parameter has a secondary role compared to the tuning parameter under data contamination (although, as we have seen, it has a major effect on the small sample efficiency of the MCpDE under pure data). In the contaminated scenario, (n times) the MSE of MDE$_\lambda(-0.55)$ is 6.9931 while that for MCpDE$_\lambda(-0.55, -1)$ is 6.1950. In comparison, the (n times) MSE of the maximum likelihood estimator MDE$_\lambda(0)$ in this case is 127.2874.

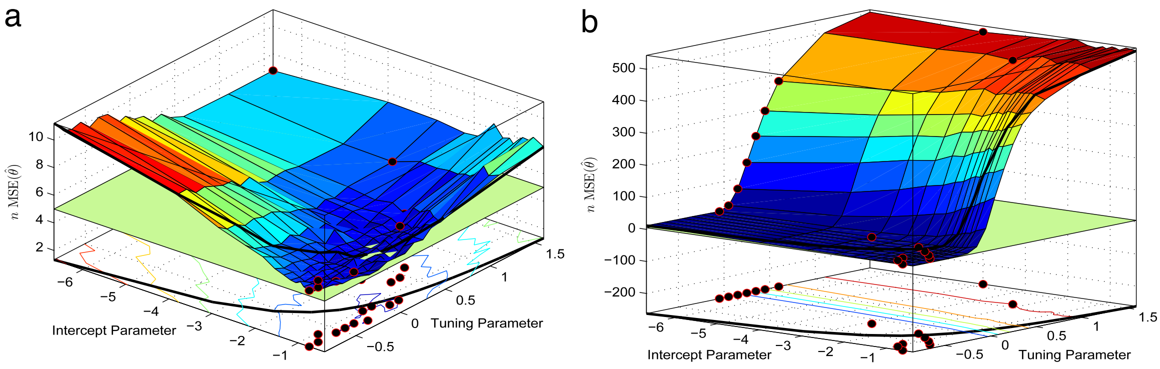

Fig. 6. Observed MSE surfaces (multiplied by n) of the MCpDE$_\lambda$ in the case of (a) the pure and (b) the contaminated Poisson models for different values of the tuning parameter and the intercept parameter. The minimum MSEs over the intercept parameter for fixed values of the tuning parameter are marked by black dots. The black line represents the ordinary MDEs.

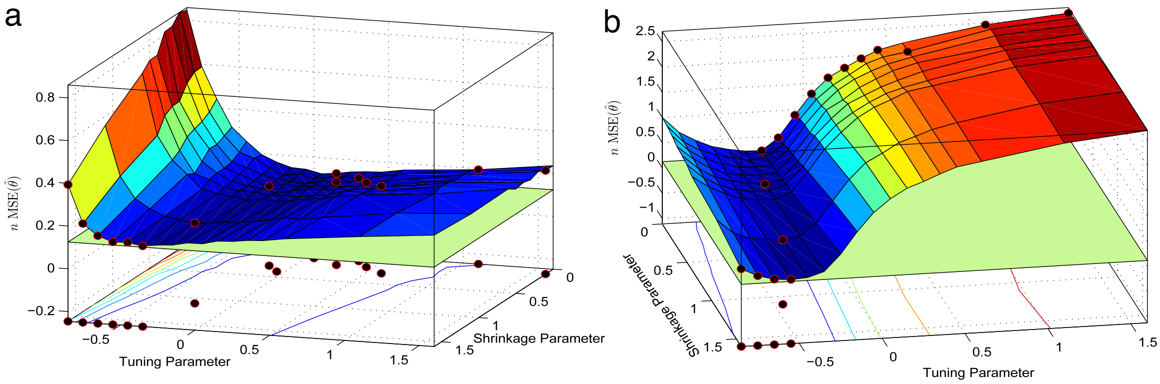

Fig. 7. Observed MSE surfaces (multiplied by n) of the MISDE$_\lambda$ in the case of (a) the pure and (b) the contaminated geometric models for different values of the tuning parameter and the inlier shrinkage parameter. The minimum MSEs over the inlier shrinkage parameter for fixed values of the tuning parameter are marked by black dots.

5.1.4. Minimum inlier shrunk disparity estimator (MISDE)

In this section we present the observed (n times) MSE surface of the MISDEs for different combinations of the tuning parameter and the inlier shrinkage parameter. The inlier shrunk disparities are formed by combining different members of the power divergence family on the outlier side with different inlier shrinkage parameters controlling the inlier component.

Here we have taken a geometric model with parameter θ, and the data are generated from the Geo(0.5) distribution; Geo(θ) represents the geometric distribution with the indicated parameter. Plot (a) of Fig. 7 presents (n times) the observed MSE of MISDEs of θ in 2000 replications. The sample size in each replication is 20. The figure shows that for MISDEs with large negative values of the tuning parameter, efficiency increases with the inlier shrinkage parameter; in general, 1.5 may be a good default value of the inlier shrinkage parameter for such values of the tuning parameter. The dots on the surface (as well as on the base) represent the optimal value of inlier shrinkage parameter for fixed values of the tuning parameter in terms of controlling the MSE. In particular, (n times) the MSE of MDE$_\lambda(-0.55)$ equals 0.3153, while those of MISDE$_\lambda(-0.55, 1.5)$ and MLE equal 0.1400 and 0.1415 respectively.

Next we examine the effect of the inlier shrinkage parameter on robustness of the inlier shrunk disparities. We have simulated data from the distribution $(1-\epsilon)\text{Geo}(0.5) + \epsilon\chi_{20}$, estimated the parameter θ under the Geo(θ) model, and computed the MSE of the estimate around the target value 0.5. The sample size is 20, $\epsilon = 0.1$, and the experiment is replicated 2000 times for every combination of the tuning and inlier shrinkage parameters. Fig. 7(b) shows that while the MSE of the MISDE$_\lambda$ increases rapidly for values of the tuning parameter larger than zero, it is very stable for negative values of the tuning parameter. The inlier modification shows no detrimental effect on the robustness of the estimators. For example, (n times) the MSE of MDE$_\lambda(-0.55)$ is 0.3654 in this case, while that for MISDE$_\lambda(-0.55, 1.5)$ is 0.1617. In comparison, the maximum likelihood estimator MDE$_\lambda(0)$ has (n times) a MSE of 1.2950.

In order to provide general recommendations for the appropriate choice of tuning parameters and inlier modification parameters that lead to best results, large extensive numerical studies are needed. We have undertaken a fair number of such studies apart from those presented in Sections 5.1.1–5.1.4. On the basis of our empirical observations in these studies, we make the following recommendations.

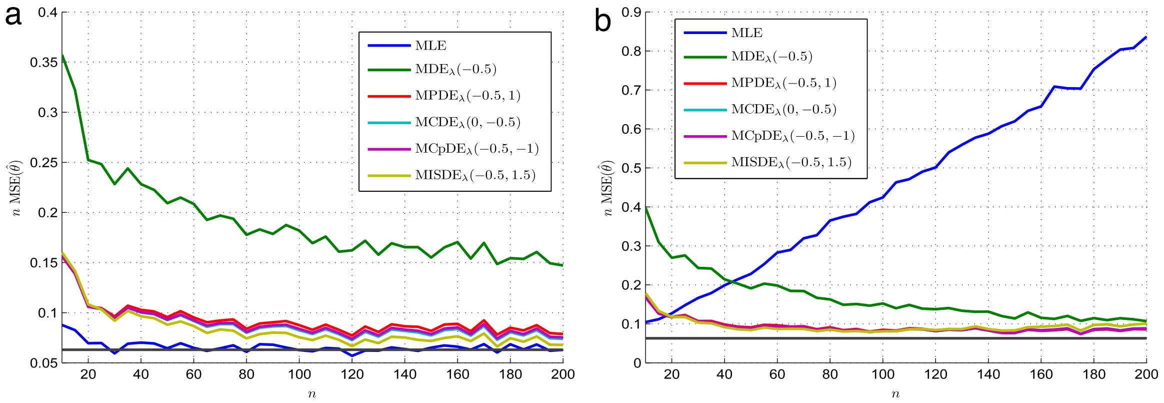

Fig. 8. MSE curves (multiplied by n) of several different ordinary and inlier modified statistics for different sample sizes in the case of (a) the pure and (b) the contaminated geometric model.

1. For the penalized families, combining power divergence statistics with tuning parameter in the range $[-0.6, -0.4]$ and penalty weight in $[0.5, 1]$ appear to provide the best result.
2. For combined families, combination of two different power divergence statistics with tuning parameter 1 in $[-0.6, -0.4]$ and tuning parameter 2 in $[0, 0.5]$ seem to be the best.
3. For coupled disparities, combining power divergence statistics with tuning parameter in $[-0.6, -0.4]$ and inlier intercept parameter in $[-1.1, -0.8]$ appear to be the best.
4. For inlier shrunk disparities, outlier tuning parameter in $[-0.6, -0.4]$ within the power divergence family combined with inlier shrinkage parameter in $[1.2, 1.6]$ seem to provide the best choice.
5. In general there is little to distinguish between the inlier modified estimators in these ranges, although, in rare cases, the penalized statistic is slightly poor compared to the other three statistics.

5.2. *The role of the sample size in estimation*

We will present a simulation example demonstrating the behaviour of different estimators in the geometric model as a function of increasing sample size. Data are randomly generated from a geometric distribution with parameter $\theta = 0.3$. Here we compare the empirical MSEs of the (a) MLE, (b) MDE$_\lambda(-0.5)$, (c) MPDE$_\lambda(-0.5, 1)$, (d) MCDE$_\lambda(-0.5, 0)$, (e) MCpDE$_\lambda(-0.5, -1)$ and (f) MISDE$_\lambda(-0.5, 1.5)$. As the value of θ is taken to be 0.3 the Fisher information $I(\theta)$ in this case is $1/(0.3^2(1-0.3)) = 15.873$, and n times the asymptotic variances of all the eight estimators are equal to $1/I(\theta) = 0.063$.

The empirical MSEs of each of these estimates are computed for 1000 replications at each sample size between 10 and 200 around the target value of 0.3. These observed MSEs (times n) are plotted as a function of the sample size in plot (a) of Fig. 8. The horizontal straight line in this plot is n times the theoretical asymptotic variance of the estimators. The most remarkable point in the graph is the extremely poor small sample performance of the ordinary MDE. This estimator appears unsatisfactory, from the efficiency standpoint, even at a sample size of 200. Clearly the MSEs of the estimators based on the inlier modified disparities are substantially closer to that of the maximum likelihood estimator compared to that of MDE$_\lambda(-0.5)$. The deficiency in the ordinary minimum disparity estimator appears to be almost completely eliminated by our inlier modification schemes.

Next we simulate data from the $(1-\epsilon)\text{Geo}(0.3) + \epsilon\chi_{20}$ distribution. In this simulation the value of ϵ is taken to be 0.05, and the experiment is replicated 1000 times for every n. The MSE of each estimator is calculated, under the geometric model, around the target value of 0.3. Plot (b) of Fig. 8 shows that the MSE of the MLE diverges as n increases, but the MSE of the other estimators remain fairly stable. Within the latter group, the performance of the inlier modified estimators appear to be significantly better than that of the ordinary minimum disparity estimator. Thus the inlier modified estimators stand out in terms of their performance under both pure and contaminated data.

5.3. *Inlier modified test statistics*

In this section we calculate the level and power of different inlier modified tests under pure and contaminated data, and examine the role of the inlier parameters. For brevity we restrict ourselves to the penalized disparity difference test statistics in Section 5.3.1, but similar improvements are also noticeable under the other inlier modified disparity difference tests.

5.3.1. *Penalized disparity difference test (PDDT)*

Here we present 3D plots of the level and power surfaces of the penalized disparity difference tests (PDDTs) for different combinations of its parameters. The penalized disparities are formed by taking a set of members from the power divergence

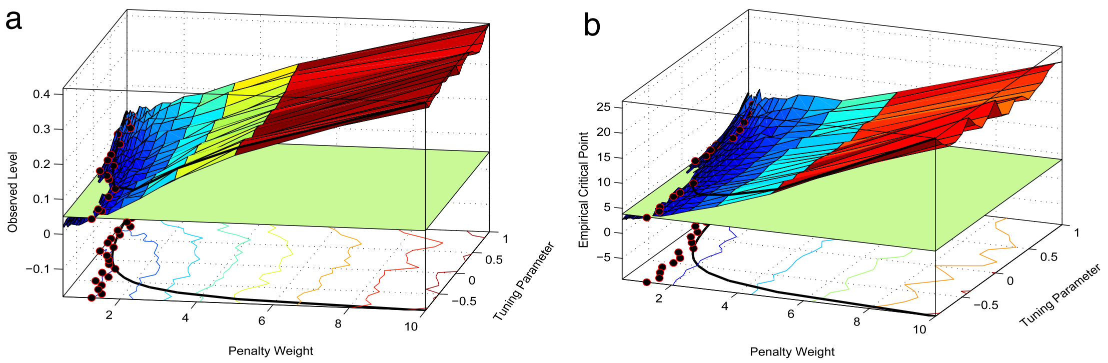

Fig. 9. (a) Level surface and (b) empirical critical points of PDDT$_\lambda$ for different values of the tuning parameter and the penalty weight in the case of the binomial model. The black dots indicate the optimum (a) observed level and (b) empirical critical points over the penalty weight when the tuning parameter is fixed. The black line represents the ordinary DDTs.

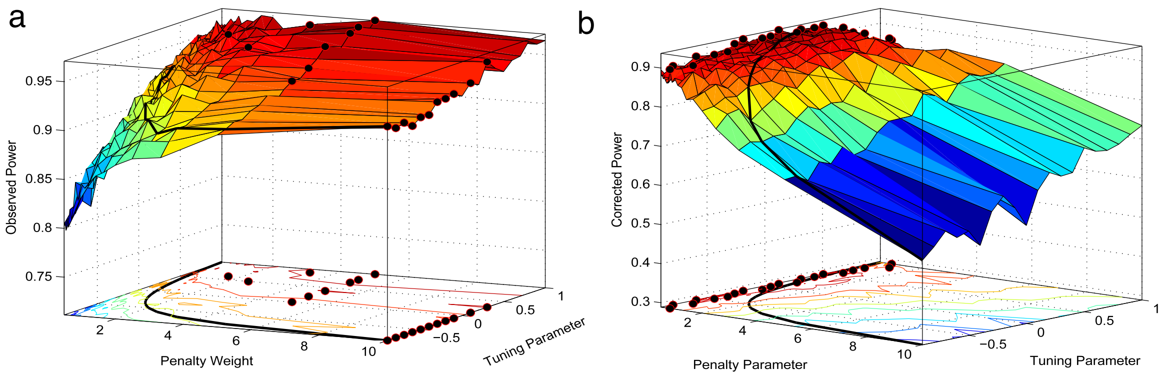

Fig. 10. (a) Observed and (b) corrected powers of the PDDT$_\lambda$ for different values of the tuning parameter and penalty weight combination for pure binomial data. The black dots indicate the optimum (a) observed and (b) corrected powers over the penalty weight when the tuning parameter is fixed. The black line represents the ordinary DDTs.

statistics (corresponding to different tuning parameters) with different empty cell penalty weights. The level and power contours are plotted over the two dimensional (tuning parameter, penalty weight) plane. The black dots on the plot indicate the optimum observed value over the penalty weight when the tuning parameter is fixed; depending on the type of the plot optimality is measured in terms of the closeness of the observed quantities to the nominal quantities, or in terms of the maximized power. The black lines on the surface and the base of the figures indicate the points which represent the ordinary disparity difference tests.

To illustrate the performance of the tests we have taken a Binomial $(20, \theta)$ model, and we test the null hypothesis $H_0 : \theta = 0.3$ at 5% level of significance. First we have drawn data from $\mathrm{Bin}(20, 0.3)$. Fig. 9(a) presents the observed level of the different penalized disparity difference tests over 2000 replications at a sample size of 25; the observed levels are calculated as the proportion of the test statistics that exceed the chi-square critical value. In Fig. 9(b) we present the observed critical points (the critical values which make the observed level equal to the nominal level) of PDDTs over these 2000 replications. It is clear from the plots that small weights on the empty cells ensure the closeness of the observed level to its nominal level as well as the closeness of the observed critical points to the asymptotic chi-square critical point.

Next we have drawn data from the $\mathrm{Bin}(20, 0.4)$ distribution (using the same sample size and number of replications), and construct the observed power of the tests for the hypothesis $H_0 : \theta = 0.3$ using the asymptotic chi-square and the observed critical points respectively. Fig. 10(a) shows that the observed power of the tests decrease with the penalty weight when chi-square critical values are used. This is only because the actual size of the tests are very high. But the corrected powers (obtained using the observed critical points) of the tests reveal that, in fact, the powers of the tests increase if we impose smaller empty cell weights (Fig. 10(b)).

Our next aim is to study the robustness properties of the penalized disparity difference test. For this purpose we have simulated data from the $(1-\epsilon)\mathrm{Bin}(20, \theta)+\epsilon\mathrm{Bin}(20, 0.2)$ distribution, where $\mathrm{Bin}(20, 0.2)$ is the contaminating component; we have taken $\epsilon = 0.1$. First we generate data from this contaminated distribution with $\theta = 0.3$ to compare the distortion in the observed level for testing $H_0 : \theta = 0.3$. In Fig. 11(a) we display the observed levels of PDDT with respect to the observed critical values presented in Fig. 9(b). The sample size is 25 and the experiment is replicated 2000 times for every combination of the tuning parameter and the penalty weight. Next we have taken samples from the contaminated distribution where the

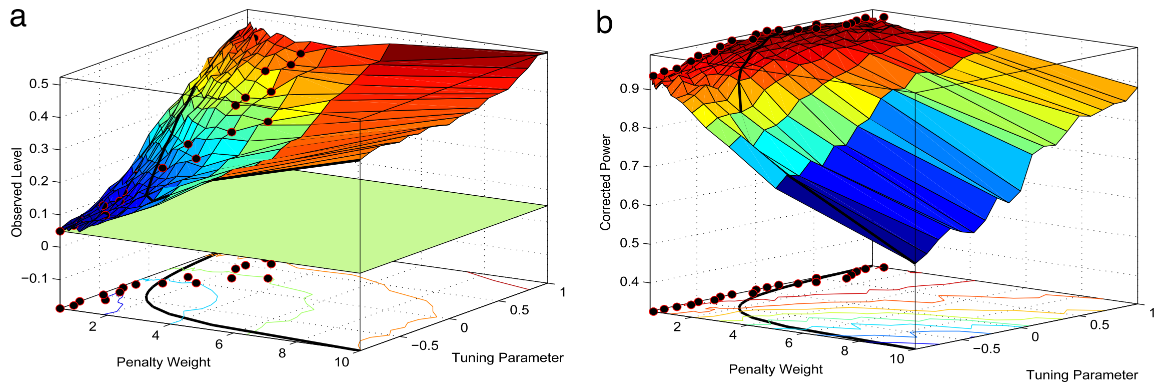

Fig. 11. (a) Observed level and (b) corrected power of the PDDT$_\lambda$ for different values of the tuning parameter and penalty weight combination for contaminated binomial data. The black dots indicate the optimum (a) observed level and (b) corrected power over the penalty weight when the tuning parameter is fixed. The black line represents the ordinary DDTs.

true value of θ is 0.4, but we still test $H_0 : \theta = 0.3$. It is expected that the presence of the contaminating component may lead to a loss in power for classical tests like likelihood ratio. Plot (b) of Fig. 11 shows the corrected power of the tests. It may be observed that both the level and the corrected power are optimized for smaller values of the penalty weight. In practice, a value of the penalty weight in the interval described by our recommendations in Section 5.1 may be a reasonable choice.

5.4. *The role of the sample size in testing of hypothesis*

In this section we will illustrate the behaviour of the various disparity difference tests in the geometric model as a function of increasing sample size.

For the first study, data are randomly generated from a geometric distribution with parameter $\theta = 0.2$. Here as well as in the rest of this subsection we test the null hypothesis

$$H_0 : \theta = 0.2 \tag{5.1}$$

at 5% level of significance. In this study we compare the observed levels, computed as the proportion of test statistics exceeding the chi-square critical value, of the (a) LRT, (b) DDT$_\lambda(-0.5)$, (c) PDDT$_\lambda(-0.5, 1)$, (d) CDDT$_\lambda(-0.5, 0)$ (e) CpDDT$_\lambda(-0.5, -1)$ and (f) ISDDT$_\lambda(-0.5, 1.5)$. These are computed over 1000 replications at each sample size between 10 and 200 and are presented in the same graph in plot (a) of Fig. 12 (the same range for the sample size and the same number of replications are used in each of the other graphs of this subsection). The horizontal straight lines in the figures represent the nominal level ($\alpha = 0.05$) of the tests. From Fig. 12(a) it is quite obvious that the levels of the inlier modified disparity difference tests are very close to that of the LRT (and to the nominal value of 0.05). The observed level of the ordinary disparity difference test based on the Hellinger distance remains significantly higher than the nominal level even at a sample size of 200, while such deficiencies are almost entirely eliminated by the appropriate choice of the inlier parameter.

Next we explore the robustness features of the tests. For this purpose we simulate data from the $0.95\text{Geo}(0.2) + 0.05\chi_{20}$ mixture. The first component is our target distribution, and the second component represents the contamination. Plot (b) of Fig. 12 shows that the observed level of the LRT increases without bounds as the sample size increases, while the level of the other tests remain stable near the nominal level. Within the latter group, the ordinary disparity difference test is clearly inferior compared to the others.

Next we investigate the power behaviour of the test statistics. Data are now generated from the geometric distribution with parameter $\theta = 0.3$. Plots (c) and (d) of Fig. 12 present the observed and corrected powers of the tests of the null hypothesis in (5.1). The corrected power is calculated by using the empirical critical values instead of the asymptotic chi-square critical points. The observed power of DDT$_\lambda(-0.5)$ is high compared to the other tests because the actual size of the test is very high. The corrected power of all the tests are, however, more or less equal.

Finally, we check the power of the tests at $\theta = 0.3$ under data contamination. Data are now generated from the $0.95\text{Geo}(0.3) + 0.05\chi_{20}$ mixture. Plots (a) and (b) of Fig. 13 present the observed and corrected powers of the test. It is obvious that the power of the LRT breaks down in the presence of contamination in the data, but the other tests perform well.

6. Conclusion

In this paper we have proposed several inlier correction methods which help to improve the small sample efficiency of robust minimum distance estimators. Our numerical studies show that over a broad range of scenarios this is achieved without compromising the robustness properties of the estimators. In the hypothesis testing scenario, similarly, the inlier corrections appear to make the tests more accurate in terms of their adherence to the nominal level, without affecting their

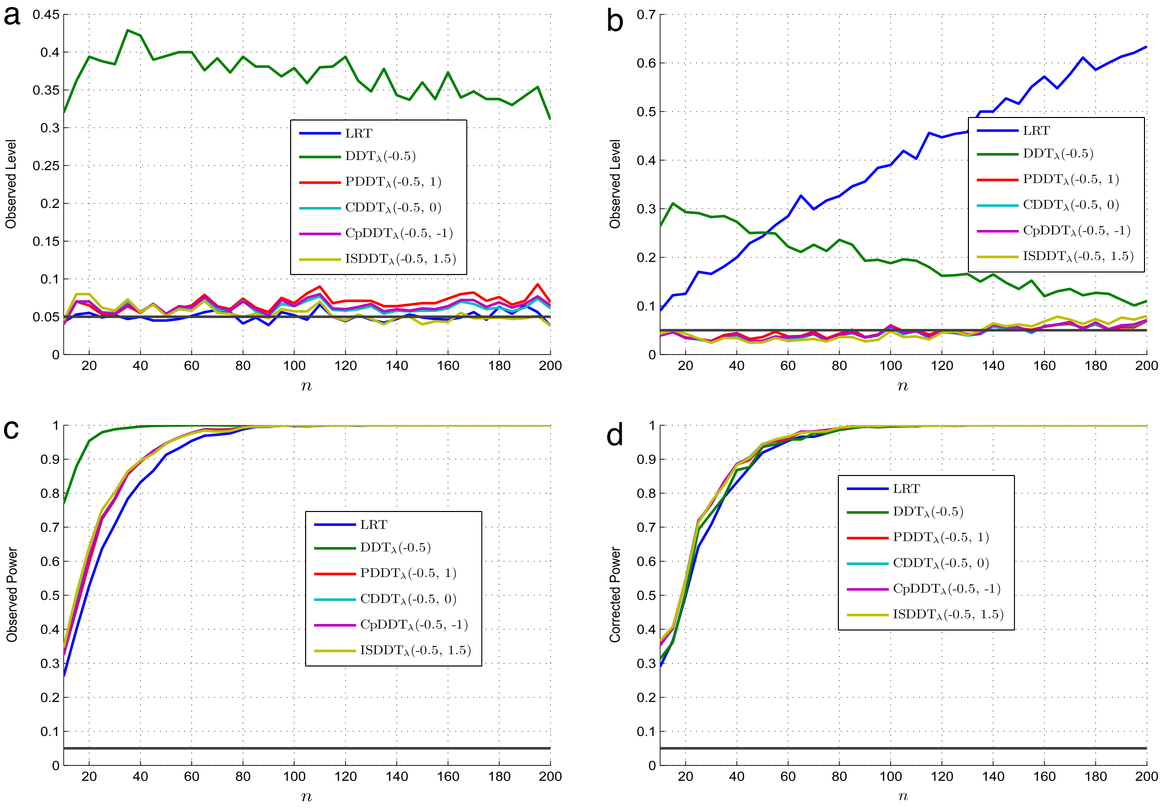

Fig. 12. Observed level of different tests in (a) the pure and (b) the contaminated geometric model for different values of *n*. (c) Observed and (d) corrected powers of different tests in the pure geometric model for different values of *n*.

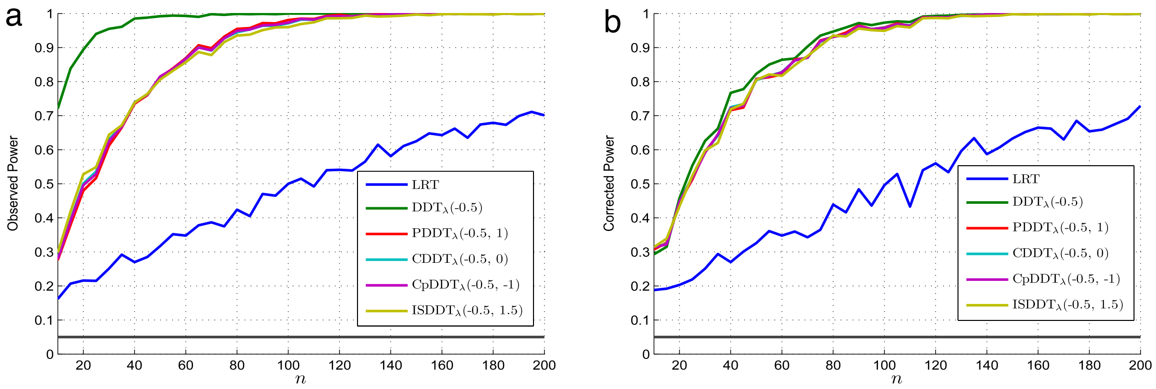

Fig. 13. (a) Observed and (b) corrected powers of different tests in the contaminated geometric model for different values of *n*.

power and level robustness. We trust that our proposed inlier modifications will substantially increase the practical value of the disparity based minimum distance estimators and tests which are already popular on account of their robustness properties.

As mentioned earlier, all the techniques described in this paper are applicable in the continuous case with the exception of the penalized distance technique. The inlier modified distances can also be used to construct weighted likelihood functions as considered in Markatou et al. (1998) and Agostinelli and Markatou (2001).

In all the studies with contaminated data we have taken the contaminating component to be far off from the target component. All our robust estimators essentially ignore these highly discrepant contaminating values. If we choose the contaminating values to close to the target values and move them away sequentially then all the robust methods (including the inlier modified ones) get initially affected by the outlier but eventually begin to discount them as they keep moving away. For brevity we simply report this observation, rather than present another set of graphs to illustrate this.

Acknowledgements

Most of the material presented in this paper is a part of the Ph.D. Thesis of the first author and the research is partially supported by the Department of Science and Technology, Govt. of India, New Delhi under Project No. SR/S4/MS:516/07 dated 21st April 2008. The authors gratefully acknowledge the suggestions of two anonymous referees which led to an improved version of the paper.

References

Agostinelli, M., Markatou, C., 2001. Test of hypotheses based on the weighted likelihood methodology. Statist. Sinica 11 (2), 499–514.

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

Basu, A., Basu, S., 1998. Penalized minimum disparity methods for multinomial models. Statist. Sinica 8 (3), 841–860.

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

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

Basu, A., Lindsay, B.G., 2004. The iteratively reweighted estimating equation in minimum distance problems. Comput. Statist. Data Anal. 45 (2), 105–124.

Basu, A., Mandal, A., Pardo, L., 2010. Hypothesis testing for two discrete populations based on the Hellinger distance. Statist. Probab. Lett. 80, 206–214.

Basu, A., Ray, S., Park, C., Basu, S., 2002. Improved power in multinomial goodness-of-fit tests. Statistician 51 (3), 381–393.

Beran, R., 1977. Minimum Hellinger distance estimates for parametric models. Ann. Statist. 5 (3), 445–463.

Bhandari, S.K., Basu, A., Sarkar, S., 2006. Robust inference in parametric models using the family of generalized negative exponential dispatches. Aust. N. Z. J. Stat. 48 (1), 95–114.

Cressie, N., Read, T.R.C., 1984. Multinomial goodness-of-fit tests. J. R. Stat. Soc. Ser. B 46 (3), 440–464.

Csiszár, I., 1963. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen Ketten. Magyar. Tud. Akad. Mat. Kutató Int. Közl. 8, 85–108.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. In: Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons Inc., New York.

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

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

Mandal, A., 2010. Minimum disparity inference: strategies for improvement in efficiency. Ph.D. Thesis. Indian Statistical Institute, Kolkata, India.

Mandal, A., Bhandari, S.K., Basu, A., 2011. Minimum disparity estimation based on combined disparities: asymptotic results. J. Statist. Plann. Inference 141 (2), 701–710.

Mandal, A., Pardo, L., Basu, A., 2010. Minimum disparity inference and the empty cell penalty: asymptotic results. Sankhyā Ser. A 72 (2), 376–406.

Markatou, M., Basu, A., Lindsay, B.G., 1998. Weighted likelihood equations with bootstrap root search. J. Amer. Statist. Assoc. 93 (442), 740–750.

Neyman, J., Pearson, E.S., 1928. On the use and interpretation of certain test criteria for purposes of statistical inference: part I. Biometrika 20 (1), 175–240.

Pardo, L., 2006. Statistical Inference Based on Divergence Measures. In: Statistics: Textbooks and Monographs, vol. 185. Chapman & Hall/CRC, Boca Raton, FL.

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

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

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

Patra, R.K., Mandal, A., Basu, A., 2008. Minimum Hellinger distance estimation with inlier modification. Sankhyā Ser. B 70 (2), 310–322.

Pearson, K., 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Phil. Mag. Ser. 5 50 (302), 157–175.

Rao, C.R., 1957. Theory of the method of estimation by minimum chi-square. Bull. Inst. Internat. Statist. 35 (2), 25–32.

Sarkar, S., Basu, A., 1995. On disparity based robust test for two discrete populations. Sankhyā Ser. B 57 (3), 353–364.

Simpson, D.G., 1989. Hellinger deviance tests: efficiency, breakdown points, and examples. J. Amer. Statist. Assoc. 84 (405), 107–113.

Vajda, I., 1989. Theory of Statistical Inference and Information. Kluwer Academic Pub.