COMPUTATIONAL STATISTICS & DATA ANALYSIS

Computational Statistics & Data Analysis 45 (2004) 741–763

# Some variants of minimum disparity estimation

Ayanendranath Basu[a], Chanseok Park[b,*], Bruce G. Lindsay[c], Haihong Li[d]

[a] *Applied Statistics Unit, Indian Statistical Institute, Calcutta 700 108, India*
[b] *Department of Mathematical Sciences, Clemson University, Clemson SC 29634, USA*
[c] *Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA*
[d] *Department of Statistics, University of Florida, Gainesville, FL 32610, USA*

Received 15 September 2002; received in revised form 15 April 2003

**Abstract**

This paper proposes several variants of disparity-based inference (Ann. Statist. 22 (1994) 1081–1114). We introduce these modifications and explain the motivation behind them. Several of these estimators and tests have attractive efficiency and robustness properties. An extensive numerical and graphical investigation is presented to substantiate the theory developed and demonstrate the small sample properties of these methods. An empty cell penalty is found to greatly enhance the performance of some of these methods.

*Keywords:* Powered Pearson divergence; Robust likelihood disparity; Winsorized and trimmed divergences; Inflection point; Empty cell

## 1. Introduction

Consider the standard parametric setup of inference where we have a family of model distributions $\mathscr{F}_\Theta = \{F_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^p$. In reality, assumed models are almost never exactly true, and our goal is to estimate $\theta$ efficiently when the model is correct (i.e., when the true distribution $G \in \mathscr{F}_\Theta$) and robustly in case the true distribution is in the neighborhood of the model but not necessarily in it. In hypothesis testing problems we desire to have a procedure which has high power under the model while being fairly

[*] Corresponding author.
*E-mail address:* cspark@ces.clemson.edu (C. Park).

stable in terms of level and power when model assumptions are violated. Traditional parametric methods such as those based on maximum likelihood are generally poor performers from the robustness viewpoint, although they are usually optimal under model conditions. On the other hand, classical robust estimators such as the $M$-estimators usually attain their robustness at the cost of first-order efficiency (e.g., Hampel et al., 1986). Certain minimum divergence procedures, however, can attain these properties simultaneously. Among others, Beran (1977) and Tamura and Boos (1986) attempted to achieve the dual goals of efficiency and robustness by using the minimum Hellinger distance estimator for continuous models. Simpson (1987) studied minimum Hellinger distance estimation under discrete models. Simpson (1989) also discussed the robust hypothesis testing problem for general models using the Hellinger distance. Lindsay (1994) gave a general framework to density-based minimum divergence estimation through the construction of *disparities* and the description of a general class of estimators many of which are both robust and first-order efficient for discrete models. An extension to the continuous case was considered by Basu and Lindsay (1994).

While all the minimum disparity estimators are first-order efficient under standard regularity conditions, additional special features are required for their robustness (see Lindsay, 1994). In this paper we consider a new class of density-based divergences, many of which exhibit such robustness features very pronouncedly, and hence appear to be worth pursuing. The structure of these divergences are very much like those of disparities, but some of them do not have the convexity property of the defining function $C(\cdot)$ (Section 2). We investigate the efficiency and the robustness of the corresponding minimum divergence procedures, appropriately modified wherever necessary. The performance of the methods are illustrated through a large numerical study involving simulation results and real-data examples. To keep a clear focus in our investigations, we will restrict the present work to discrete models. We hope to consider the application of similar techniques to continuous models in a future paper. We trust that the paper illustrates, among other things (a more detailed discussion is given in the Concluding Remarks section) the possibilities of modifying other related divergences so that they can be analyzed within the framework of disparities.

The remainder of the paper is organized as follows: Section 2 contains a brief description of the class of disparities in general followed by a discussion of the proposed divergences. In Section 3 we discuss the asymptotic properties of the estimators, and their modifications, the breakdown point issue, and robust testing of hypotheses using the above divergences. Section 4 presents numerical results, where we also illustrate the effect of an empty cell penalty on the procedures. Section 5 presents some concluding remarks.

Throughout this paper we will refer to true distribution by $G$, which may or may not belong to $\mathscr{F}_\Theta$. We will assume that both $G$ and $\mathscr{F}_\Theta$ belong to $\mathscr{G}$, the class of all distributions having probability density functions (pdf's) with respect to a dominating measure. We will also denote the density function for each distribution with the corresponding lower case letter, e.g., the pdf's of $G$ and $F_\theta$ will be denoted by $g$ and $f_\theta$, respectively.

## 2. Minimum disparity inference and proposed methods

### 2.1. Disparities and residual adjustment function

Consider a parametric family of distributions $\mathscr{F}_\Theta$ having densities $f_\theta(\cdot)$ with a countable sample space. Without loss of generality, let the sample space be $\mathscr{X} = \{0,1,2,\ldots\}$. Let $d(x)$ be the empirical density at $x$ (relative frequency at $x$) based on a random sample of size $n$ from the true distribution $G(\cdot)$ which is modeled by the above parametric family of distributions. Our interest is in making inference about the unknown $\theta$. Following Lindsay (1994), we define a disparity—a measure of discrepancy between probability densities $d(\cdot)$ and $f_\theta(\cdot)$—given by a thrice differentiable convex function $C(\cdot)$ with $C(0)=0$ as

$$\rho_C(d,f_\theta)=\sum_{x\in\mathscr{X}} C(\delta(x))f_\theta(x), \qquad (1)$$

where the Pearson residual $\delta(x)$ is defined to be $\delta(x)=d(x)/f_\theta(x)-1$. The range of the Pearson residual is $[-1,\infty)$, and $\delta(x)=-1$ only when $d(x)=0$ (i.e., when the cell $x$ is empty), and equals 0 only when $d(\cdot)=f_\theta(\cdot)$. Under differentiability of the model, the minimization of the disparity measure (1) corresponds to solving an estimating equation of the form

$$-\nabla\rho_C=\sum_{x\in\mathscr{X}} A(\delta(x))\nabla f_\theta(x)=0, \qquad (2)$$

where $A(\delta)=(1+\delta)C'(\delta)-C(\delta)$ and $\nabla$ represents the gradient with respect to $\theta$. The function $A(\delta)$ can be centered and scaled, without changing the estimating properties of the disparity, so that $A(0)=0$ and $A'(0)=1$. We will call the centered and scaled function $A(\cdot)$ the residual adjustment function (RAF) of the disparity. Minimum disparity estimators have received wide attention in statistical inference because of their ability to reconcile the properties of robustness and asymptotic efficiency. See Lindsay (1994) for more details of the method, and Basu et al. (1997) for a comprehensive review including some of the later work. When $C(\cdot)$ is strictly convex, the disparity measure is nonnegative and equals 0 only when the densities $d(\cdot)$ and $f_\theta(\cdot)$ are identically equal. Through appropriate selection of $C(\cdot)$, a large family of important divergences and distances can be developed in this manner, including the power divergence family (Cressie and Read, 1984) which generates the Kullback-Leibler divergence and the Hellinger distance as special cases. The curvature parameter $A''(0)$, which is the second derivative of the RAF evaluated at $\delta=0$, is a measure of the tradeoff between robustness and second-order efficiency (Lindsay, 1994). Large negative values of $A''(0)$ correspond to stronger robustness properties (but also greater second-order deficiency), while $A''(0)=0$ corresponds to second-order efficient estimators in the sense of Rao (1961, 1962).

## 2.2. The powered Pearson divergence

Here we introduce a new family of divergences—the powered Pearson divergence family—between $d(\cdot)$ and $f_\theta(\cdot)$ which satisfy the general definition of a statistical distance in the sense that it is nonnegative and equal to zero if and only if $d(\cdot)=f_\theta(\cdot)$. In this paper we will consider the powered Pearson divergence (PPD) and appropriate modifications of it which have reasonable efficiency and robustness properties. Although the structures of the resulting estimating equations are similar to those of disparities, some of the PPDs as well as some of their modifications do not belong to the class of disparities. In addition, we present an extensive comparative study of the proposed methods with several robust modifications of the likelihood disparity as in Chakraborty et al. (2003), and show that the results are very similar in either case.

The $\mathrm{PPD}_\alpha$ indexed by a single parameter $\alpha \in (0,1]$ between two arbitrary discrete densities $g(\cdot)$ and $f(\cdot)$ on $\mathscr{X}$ is given by

$$\mathrm{PPD}_\alpha(g,f)=\frac{1}{2\alpha^2}\sum_{x\in\mathscr{X}}\left[\frac{g(x)^\alpha-f(x)^\alpha}{f(x)^\alpha}\right]^2 f(x),$$

and $g(\cdot)$ and $f(\cdot)$ are replaced by $d(\cdot)$ and $f_\theta(\cdot)$ under the parametric estimation setup. The PPD family includes Pearson's $\chi^2$ $(\alpha=1)$ and Hellinger distance $(\alpha=\frac{1}{2})$, and can be thought of as $L_2$ distance on the power transformed densities.

As in the case of disparities, one can write $\mathrm{PPD}_\alpha$ in the form (1), and arrive at an estimating equation of the form (2), where the $C(\cdot)$ function, its second derivative $C''(\cdot)$ and $A(\cdot)$ function are given by

$$C(\delta)=\frac{1}{2\alpha^2}[(\delta+1)^\alpha-1]^2, \tag{3}$$

$$C''(\delta)=\frac{1}{\alpha}(\delta+1)^{\alpha-2}[1-\alpha-(1-2\alpha)(\delta+1)^\alpha], \tag{4}$$

$$A(\delta)=\frac{1}{2\alpha^2}[(\delta+1)^\alpha-1][(2\alpha-1)(\delta+1)^\alpha+1]. \tag{5}$$

However $C''(\cdot)$ is always nonnegative only when $\alpha \geqslant \frac{1}{2}$, so the $C(\cdot)$ functions of the $\mathrm{PPD}_\alpha$ family are not convex on $[-1,\infty)$ when $\alpha < \frac{1}{2}$. But, since smaller values of $\alpha$ provide greater downweighting for larger outliers, these are the interesting values of $\alpha$ for robustness purposes.

One of our main objectives in this paper is to investigate the effect of this nonconvexity, and modify this family appropriately to obtain stable inference. Since $A'(\delta)=(\delta+1)C''(\delta)$ and $\delta \geqslant -1$, it can be seen from Eqs. (3), (4) and (5) that for $\alpha < \frac{1}{2}$, the RAF starts to redescend after a certain inflection point $\delta_1$ where $A'(\delta_1)$ becomes zero. Some simple algebra shows that this inflection point is given by $\delta_1=[(1-\alpha)/(1-2\alpha)]^{1/\alpha}-1$. Beyond $\delta > \delta_1$, the function $A(\cdot)$ steadily decreases, and moreover it becomes negative for $\delta > \delta_2$ with $\delta_2=[1/(1-2\alpha)]^{1/\alpha}-1$. This results in a negative impact of a big outlier, as compared to a large positive impact for methods like maximum likelihood, and minimal positive impact for good robust methods. When used just as it is, the estimation procedures resulting from the minimization of the PPD$_\alpha$ with $\alpha < \frac{1}{2}$ (particularly for very small values of $\alpha$) can lead to nonsensical results. Later on we will look at an example under the Poisson model where a very small value of $\alpha$ is shown to lead to a global but silly minimum at $\theta = 0$.

We propose the following methods to remedy this problem. One way is to force the RAF to be equal to zero from the point where it dips below zero for the first time (at $\delta_2$), and the other is to extend the RAF at $\delta = \delta_1$, and hold the residual adjustment function constant at slope equal to zero beyond the inflection point. In the first case $A(\delta) = 0$ for $\delta > \delta_2$, and in the second case $A(\delta) = A(\delta_1) = 1/[2(1-2\alpha)]$ for $\delta > \delta_1$. We call the divergence based on the former modification the *trimmed* powered Pearson divergence (TPPD) and the divergence based on the latter the *Winsorized* powered Pearson divergence (WPPD). The TPPD and WPPD with $\alpha < \frac{1}{2}$ are given by

$$\mathrm{TPPD}_\alpha(d, f_\theta) = \frac{1}{2\alpha^2} \sum_{d/f_\theta < (1/(1-2\alpha))^{1/\alpha}} \left[\frac{d(x)^\alpha - f_\theta(x)^\alpha}{f_\theta(x)^\alpha}\right]^2 f_\theta(x)$$

$$+ 2(1-2\alpha)^{1/\alpha - 2} \sum_{d/f_\theta \geqslant (1/(1-2\alpha))^{1/\alpha}} d(x),$$

$$\mathrm{WPPD}_\alpha(d, f_\theta) = \frac{1}{2\alpha^2} \sum_{d/f_\theta < ((1-\alpha)/(1-2\alpha))^{1/\alpha}} \left[\frac{d(x)^\alpha - f_\theta(x)^\alpha}{f_\theta(x)^\alpha}\right]^2 f_\theta(x)$$

$$+ \sum_{d/f_\theta \geqslant ((1-\alpha)/(1-2\alpha))^{1/\alpha}} \left[\frac{(1-2\alpha)^{1/\alpha-2}}{(1-\alpha)^{1/\alpha-1}} d(x) - \frac{1}{2(1-2\alpha)} f_\theta(x)\right].$$

The $C(\cdot)$ functions are given by

$$C_{\mathrm{TPPD}}(\delta) = \begin{cases} \dfrac{1}{2\alpha^2}[(\delta+1)^\alpha - 1]^2, & \delta < \left(\dfrac{1}{1-2\alpha}\right)^{1/\alpha} - 1, \\ 2(1-2\alpha)^{1/\alpha-2}(\delta+1), & \delta \geqslant \left(\dfrac{1}{1-2\alpha}\right)^{1/\alpha} - 1, \end{cases} \tag{6}$$

$$C_{\mathrm{WPPD}}(\delta) = \begin{cases} \dfrac{1}{2\alpha^2}[(\delta+1)^\alpha - 1]^2, & \delta < \left(\dfrac{1-\alpha}{1-2\alpha}\right)^{1/\alpha} - 1, \\ \dfrac{(1-2\alpha)^{1/\alpha-2}}{(1-\alpha)^{1/\alpha-1}}(\delta+1) - \dfrac{1}{2(1-2\alpha)}, & \delta \geqslant \left(\dfrac{1-\alpha}{1-2\alpha}\right)^{1/\alpha} - 1. \end{cases} \tag{7}$$

The inflection points $\delta_1$ and the trimming points $\delta_2$ for each of several values of $\alpha$ are given in Table 1. Also we present the figures of the $C(\delta)$ and $A(\delta)$ functions of the PPD$_\alpha$, TPPD$_\alpha$ and WPPD$_\alpha$ families corresponding to $\alpha = 0.1$ in Fig. 1. The WPPD essentially replaces the remaining part of the $C(\delta)$ curve on the right with a line of slope equal to $k$ from the point where its derivative $C'(\delta)$ reaches its maximum value $k = C'(\delta_1)$ on the positive side of the axis (which is the inflection point). For TPPD

Table 1
Inflection and trimming points for $\mathrm{PPD}_\alpha$

| $\alpha$ | $\to 0$ | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | $\to \frac{1}{2}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\delta_1$ | $e-1$ | 1.95 | 2.25 | 2.65 | 3.21 | 4.06 | 5.46 | 8.11 | 14.59 | 43.18 | $\infty$ |
| $\delta_2$ | $e^2-1$ | 7.23 | 8.31 | 9.78 | 11.86 | 15.00 | 20.21 | 30.18 | 54.90 | 165.81 | $\infty$ |

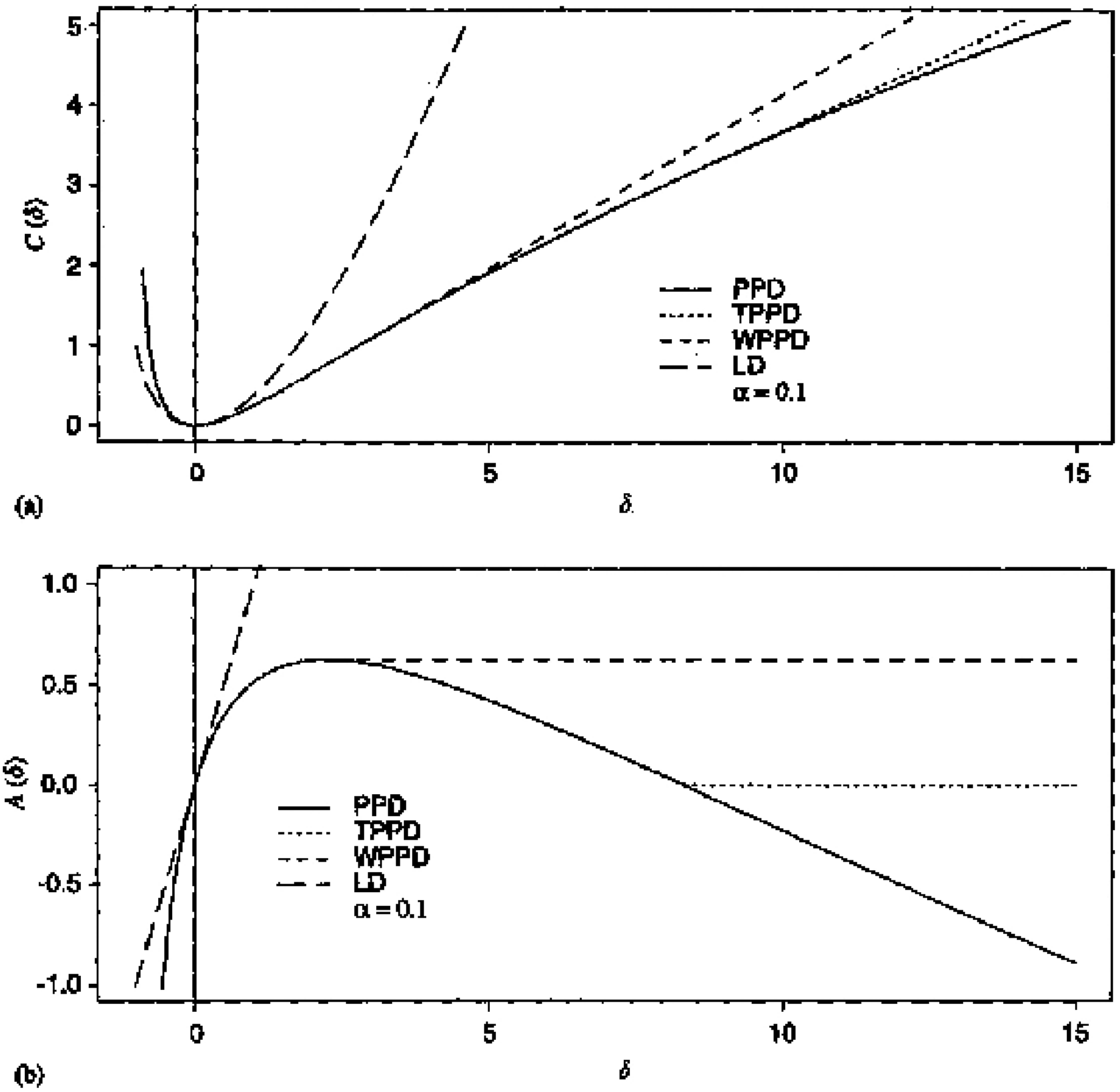

Fig. 1. $C(\cdot)$ and $A(\cdot)$ functions of PPD, TPPD, WPPD and LD with $\alpha = 0.1$.

the $C(\delta)$ function is linear beyond the trimming point $\delta_2$, with constant slope equal to $C'(\delta_2)$. Notice that the $C(\cdot)$ functions of WPPD are still convex (although not strictly convex) but the $C(\cdot)$ functions of TPPD are not. However both $C(\delta)$ functions have unique minimum (equal to 0) at $\delta = 0$.

Notice that for $\alpha \in [\frac{1}{2}, 1]$ no modification to $\mathrm{PPD}_\alpha$ is necessary since the defining function $C(\cdot)$ remains convex. Alternatively, the inflection and trimming points are at infinity for $\alpha \geqslant \frac{1}{2}$. Thus $\mathrm{PPD}_\alpha = \mathrm{WPPD}_\alpha = \mathrm{TPPD}_\alpha$ for $\alpha \geqslant \frac{1}{2}$. We will see later, however, that $\alpha = \frac{1}{2}$ is the only case of interest to us from the robustness viewpoint within the $\mathrm{PPD}_\alpha$, $\alpha \in [\frac{1}{2}, 1]$ class. For the rest of the paper, our interest will be on $\mathrm{WPPD}_\alpha$ and $\mathrm{TPPD}_\alpha$ families only for $\alpha \in (0, \frac{1}{2}]$.

We note that although we arrived at the $\mathrm{PPD}_\alpha$ family because weighted sums of squared differences seemed natural divergences to investigate, we have since noticed that Read and Cressie, (1988) arrived at this very same family as an approximation to the well-known Cressie Read power divergence (Read and Cressie, 1988, p. 95). However, unlike the $\mathrm{PPD}_\alpha$ family, the power divergences all correspond to convex $C(\cdot)$ functions.

In this paper we will compare the estimators generated by the $\mathrm{PPD}_\alpha$ and its modifications with those resulting from the Winsorized likelihood disparity (WLD) and the trimmed likelihood disparity (TLD) families which are modifications of the likelihood disparity similar in spirit to the modifications of the $\mathrm{PPD}_\alpha$ discussed earlier. The likelihood disparity (LD) between $d(\cdot)$ and $f_\theta(\cdot)$ is defined by

$$\mathrm{LD}(d, f_\theta) = \sum_{x \in \mathscr{X}} \left[ d(x) \log \frac{d(x)}{f_\theta(x)} - d(x) + f_\theta(x) \right],$$

which is minimized by the maximum likelihood estimator of $\theta$ in discrete models. The corresponding $C(\cdot)$ and $A(\cdot)$ functions are given by $C(\delta) = (\delta + 1)\log(\delta + 1) - \delta$ and $A(\delta) = \delta$ (for comparison we have presented the $C(\delta)$ and $A(\delta)$ functions of the LD in Fig. 1 also).

The $\mathrm{WLD}_\lambda$ and the $\mathrm{TLD}_\lambda$ for $\lambda$ any fixed number in $(0, 1]$ and $\bar{\lambda} = 1 - \lambda$ are of the form.

$$\mathrm{WLD}_\lambda(d, f_\theta) = \sum_{d/f_\theta < 1/\bar{\lambda}} \left[ d(x) \log\left( \frac{d(x)}{f_\theta(x)} \right) + f_\theta(x) - d(x) \right] - \sum_{d/f_\theta \geqslant 1/\bar{\lambda}} \left[ d(x) \log \bar{\lambda} + \frac{\lambda}{\bar{\lambda}} f_\theta(x) \right],$$

$$\mathrm{TLD}_\lambda(d, f_\theta) = \sum_{d/f_\theta < 1/\bar{\lambda}} \left[ d(x) \log\left( \frac{d(x)}{f_\theta(x)} \right) + f_\theta(x) - d(x) \right] - \sum_{d/f_\theta \geqslant 1/\bar{\lambda}} [d(x)(\log \bar{\lambda} + \lambda)].$$

The $\mathrm{WLD}_\lambda$ is a form of the robustified likelihood disparity (RLD) considered by Chakraborty et al. (2003). It is easy to see that the $C(\cdot)$ functions for $\mathrm{WLD}_\lambda$ are convex (although not strictly convex) while the $C(\cdot)$ functions for $\mathrm{TLD}_\lambda$ are not. Also $\mathrm{WLD}_{\lambda=1} = \mathrm{TLD}_{\lambda=1} = \mathrm{LD}$.

For better understanding the robustness of these methods, we also present the combined weight function $w_c(\delta_c)$ (Park et al., 2002) for the $\mathrm{TPPD}_\alpha$, $\mathrm{WPPD}_\alpha$, $\mathrm{TLD}_\lambda$, and $\mathrm{WLD}_\lambda$ families for different values of $\alpha$ and $\lambda$. The combined weight function $w_c(\delta_c)$

represents the relative impact of the observation in the estimating equation compared to maximum likelihood. Here we define a combined residual $\delta_c$ as

$$\delta_c(x) = \begin{cases} \delta_P(x), & d \leqslant f_\theta, \\ \delta_N(x), & d > f_\theta \end{cases}$$

with the Neyman residual $\delta_N(x) = [d(x) - f_\theta(x)]/d(x)$. The combined weight function $w_c(\delta_c)$ is

$$w_c(\delta_c) = \begin{cases} \dfrac{A(\delta_c)}{\delta_c}, & -1 \leqslant \delta_c < 0, \\ A'(0), & \delta_c = 0, \\ \dfrac{1-\delta_c}{\delta_c} A\left(\dfrac{\delta_c}{1-\delta_c}\right), & 0 < \delta_c < 1, \\ A'(\infty), & \delta_c = 1. \end{cases} \tag{8}$$

On the positive side of the $\delta_c$ axis, this amounts to looking at the weights as a function of the Pearson residuals but in the Neyman scale. For better outlier robustness, it is desirable that the weight functions converge to 0 as $\delta_c \to 1$. If we restrict TPPD$_\alpha$ and WPPD$_\alpha$ to $\alpha \in (0, \frac{1}{2}]$, all the four families of weight functions satisfy this property, but the TPPD$_\alpha$ and WPPD$_\alpha$ appear to do this more smoothly. The PPD$_\alpha$, TPPD$_\alpha$ and WPPD$_\alpha$ families coincide for $\alpha \geqslant \frac{1}{2}$, and their behavior for large values of $\delta$ makes them highly nonrobust (except for $\alpha = \frac{1}{2}$) which is demonstrated here only for $\alpha = 0.6$ in Fig. 2(b), but is actually true for all $\alpha > \frac{1}{2}$.

## 3. Asymptotic distributions and tests of hypotheses

### 3.1. Asymptotic distributions

While we emphasize the numerical results in this paper, we present brief remarks about the asymptotic behavior of the estimators. Notice that under the model the estimators corresponding to the minimizers of PPD$_\alpha$, WPPD$_\alpha$ and TPPD$_\alpha$ are all Fisher consistent, which implies that they are all weakly consistent under the model as well (e.g., Cox and Hinkley, 1974, p. 288). For the asymptotic normality of the functional under the model, notice that for the PPD$_\alpha$, the general proof for asymptotic normality would work directly provided Lindsay's boundedness assumptions on $A(\cdot)$ were satisfied (Assumption 24, Lindsay, 1994). Unfortunately, the boundedness conditions referred to above do not hold generally for the members of the PPD$_\alpha$ family (except for $\alpha = 1$). Hence we need to refine Lindsay's proof.

The failure of the PPD$_\alpha$ family to satisfy the conditions imposed by Lindsay is due to the fact that $A'$ and $A''$ become unbounded as $\delta \to -1$, i.e. $\delta = -1$ is the only aberrant point. However, this is also the feature of the Hellinger distance and the asymptotic normality of the minimum Hellinger distance estimator is well established (e.g., Simpson, 1987), showing that the boundedness requirement is not strictly necessary, although

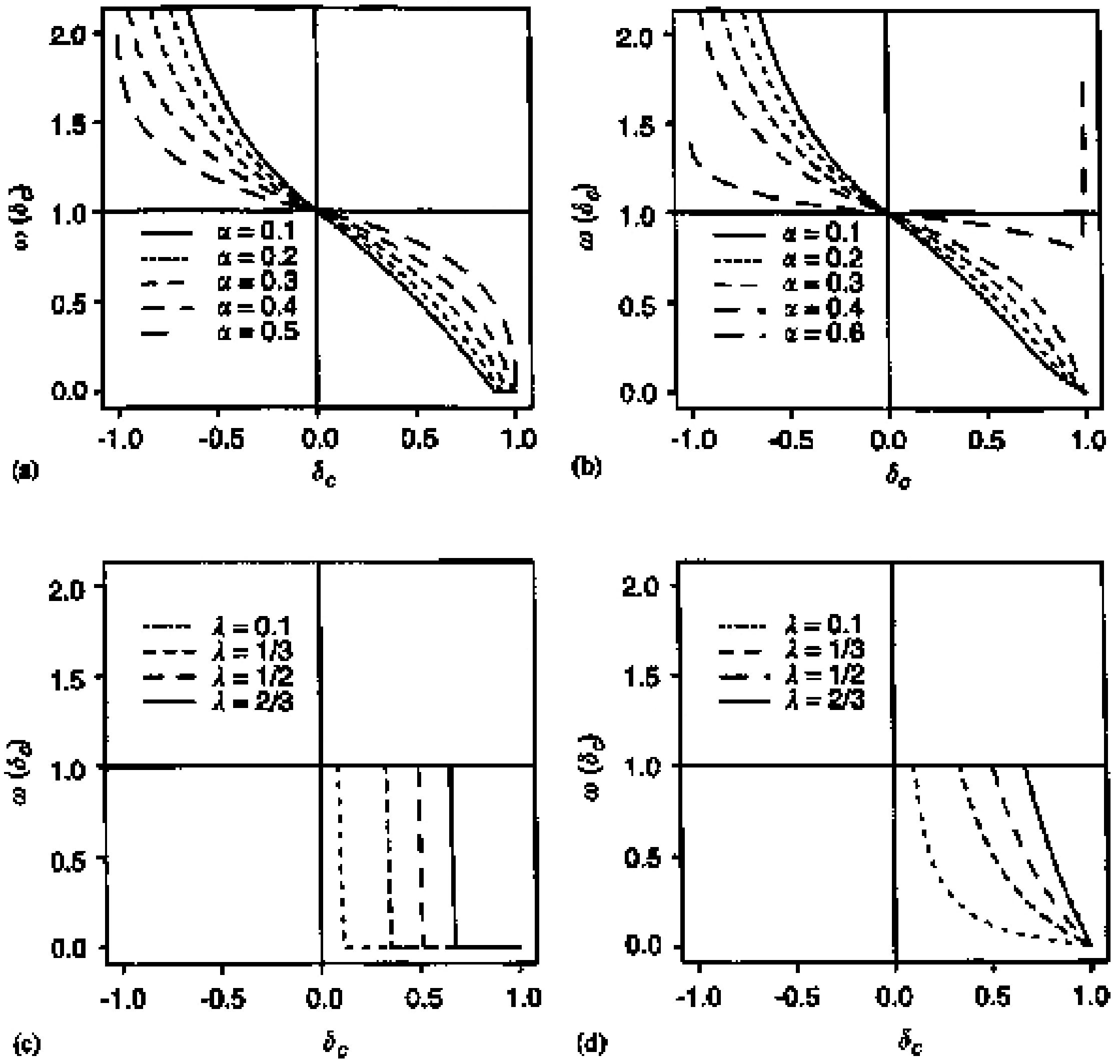

Fig. 2. Combined weight functions of (a) $\text{TPPD}_\alpha$, (b) $\text{WPPD}_\alpha$, (c) $\text{TLD}_\lambda$, (d) $\text{WLD}_\lambda$.

sufficient. In order to bypass this problem, we next establish the following Lemma, which helps us refine the asymptotic normality proof (with the true distribution being in the model). For this purpose, we modify Assumption 24 of Lindsay.

**Assumption 1.** $A(-1)$ is finite and $A'(\delta)$ and $A''(\delta)\delta$ are bounded in a neighborhood of zero and as $\delta \to \infty$.

Notice that we have imposed the boundedness condition on a smaller set rather than on $\delta \in [-1, \infty)$, thus excluding the point where Lindsay's condition is violated by the $\text{PPD}_\alpha$ family.

**Lemma 1.** *If $A(\delta)$ satisfies Assumption 1, then for all nonnegative $r$*

$$|A(r^2 - 1) - (r^2 - 1)| \leqslant B \times (r - 1)^2.$$

*for some positive constant $B$.*

**Proof.** As

$$h(r) = \frac{A(r^2 - 1) - (r^2 - 1)}{(r - 1)^2}$$

is a continuous function, it suffices to show that $h(r)$ is bounded at $r = 0$, as $r \to 1$, and as $r \to \infty$. This is easily seen to be true by applications of L'Hospital's rule and Assumption 1. □

Notice that the above lemma essentially establishes the result

$$|A(r^2 - 1) - A(s^2 - 1) - (r^2 - s^2)A'(s^2 - 1)| \leqslant B \times (r - s)^2$$

in Lindsay (Lemma 25), restricted to the case where the true distribution belongs to the model. It follows from our Lemma 1, above, that

$$|A(\delta(x)) - \delta(x)| \leqslant B \times \left\{ \left( \frac{d(x)}{f_\theta(x)} \right)^{1/2} - 1 \right\},$$

the analog of the bound in Lindsay, Eq. (30). Thus the most important step in the derivation of the asymptotic normality is obtained without requiring the boundedness of $A'$ or $A''$ at $\delta = -1$.

We make the following additional assumption on the model.

**Assumption 2.** Let $f_\theta(x)$ be the density of the true distribution, and let $u_\theta(x) = (\partial/\partial\theta) \log f_\theta(x)$ represent the likelihood score function. Then $\sum f_\theta^{1/2}(x)|u_\theta(x)| < \infty$.

Then, for the true distribution belonging to the model, the results in Theorem 23 and Lemma 29 of Lindsay continue to hold, under the weaker conditions in Assumptions 1 and 2 above.

To make the asymptotic argument for the WPPD$_\alpha$ and TPPD$_\alpha$ families, let us write $d(\cdot) = d_n(\cdot)$ for the data density at sample size $n$, and let $\theta_n$ be the corresponding estimator. For the WPPD$_\alpha$ case, once again our Assumption 1 continues to hold. Notice that for the inflection point $\delta_1$ the second derivative $A''(\cdot)$ is not defined; however, the smoothness conditions are satisfied in an interval of $\delta$ around 0, and the results follow by noticing that $\{x : d_n(x)/f_{\theta_n}(x) - 1 = \delta_1\}$ converges to a set of probability zero under a true distribution which belongs to the model. Similarly, $A'(\cdot)$ and $A''(\cdot)$ do not exist for the TPPD$_\alpha$ at the trimming point $\delta = \delta_2$, but the probability of the set $\{x : d_n(x)/f_{\theta_n}(x) - 1 = \delta_2\}$ goes to zero under the model.

### 3.2. Robust tests of hypotheses

Given a parametric hypothesis $H_0 : \theta = \theta_0$ (or more generally $H_0 : \theta \in \Theta_0 \subset \Theta$), one can define robust tests of hypothesis for the above using the TPPD$_\alpha$ and WPPD$_\alpha$. Given the empirical density $d(\cdot)$, the WPPD$_\alpha$ test statistic for the above hypothesis is given by

$$2n[\mathrm{WPPD}_\alpha(d, f_{\hat\theta_0}) - \mathrm{WPPD}_\alpha(d, f_{\hat\theta})],$$

where $\hat{\theta}_0$ and $\hat{\theta}$ are the minimizers of $\mathrm{WPPD}_\alpha(d, f_\theta)$ over $\Theta_0$ and the unrestricted parameter space $\Theta$, respectively. One can generate a corresponding statistic using the $\mathrm{TPPD}_\alpha$ in place of $\mathrm{WPPD}_\alpha$. Similar statistics can be defined for $\mathrm{WLD}_\lambda$ and $\mathrm{TLD}_\lambda$. Combined with the result of Section 3.1, it follows from Theorem 6 of Lindsay (1994) that the null distributions of the $\mathrm{WPPD}_\alpha$ and $\mathrm{TPPD}_\alpha$ statistics have the same $\chi^2$ limit as the $-2\times\log$ likelihood ratio.

### *3.3. Breakdown points*

The asymptotic breakdown points of the estimators corresponding to the $\mathrm{WPPD}_\alpha$ and $\mathrm{TPPD}_\alpha$ have been established elsewhere by Basu et al. (2001). In particular, the results show that the estimators have 50% breakdown point under the model for any outlier sequence.

## 4. Numerical studies

### *4.1. Preliminaries*

We perform an extensive numerical study to investigate the properties of the minimum divergence estimators and the corresponding tests of hypotheses for the proposed families and compare them to the methods based on the $\mathrm{WLD}_\lambda$ family (those based on $\mathrm{TLD}_\lambda$ were very similar). We chose the Poisson and geometric models (which are the two most common count data models) to base our investigations upon. Since the results are very similar, we concentrate primarily on the Poisson model in our presentations to make our point more succinct.

First, to demonstrate the peculiarities of the $\mathrm{PPD}_\alpha$ method, we consider a part of an experiment originally reported by Woodruff et al. (1984), and analyzed by Simpson (1987). The frequencies of frequencies of daughter flies carrying a recessive lethal mutation on the X-chromosome are considered where the male parents have been exposed to a certain degree of a chemical. Roughly 100 daughter flies were sampled for each male. This particular experiment resulted in $(x_i, f_i) = (0,23), (1,7), (2,3), (91,1)$, where $x_i$ is the number of daughters carrying the recessive lethal mutation and $f_i$ is the number of male parents having $x_i$ such daughters. We will refer to this as the *Drosophila Data I*.

The estimators of $\theta$ under a parametric Poisson $(\theta)$ model corresponding to $\alpha = 0.1, 0.2, 0.3, 0.4$ for the Drosophila Data I are presented in Table 2 for the $\mathrm{PPD}_\alpha$, $\mathrm{TPPD}_\alpha$, $\mathrm{WPPD}_\alpha$ and $\mathrm{pWPPD}_\alpha$ families. The $\mathrm{pWPPD}_\alpha$ method is a modification of $\mathrm{WPPD}_\alpha$ to be introduced later in this section. For this model it can be shown that $\mathrm{PPD}_\alpha(d, f_\theta)$ converges to $c_1 = (1/2\alpha^2)(d(0)^\alpha - 1)^2$ as $\theta \to 0$. In this example it appears (Fig. 3) that $\theta = 0$ is the global minimum of $\mathrm{PPD}_\alpha$ for $\alpha = 0.1$. It means that the estimator tends to "implode" toward 0 in this case. On the other hand as $\theta \to 0$ $\mathrm{TPPD}_\alpha(d, f_\theta)$ converges to $c_2 = (1/2\alpha^2)(d(0)^\alpha - 1)^2 + 2(1-2\alpha)^{1/\alpha - 2}\sum_{x>0} d(x)$, and $\mathrm{WPPD}_\alpha(d, f_\theta)$ converges to $c_3 = (1/2\alpha^2)(d(0)^\alpha - 1)^2 + [(1-2\alpha)^{1/\alpha-2}/(1-\alpha)^{1/\alpha-1}]\sum_{x>0} d(x)$. Notice that $c_1 \leqslant c_2 \leqslant c_3$ for $\alpha < \frac{1}{2}$, and at least for the Drosophila Data I example, $c_2$, $c_3$

Table 2
The estimated parameters under the Poisson model for the Drosophila Data I

| $\alpha$ | $\mathrm{PPD}_\alpha$ | $\mathrm{TPPD}_\alpha$ | $\mathrm{WPPD}_\alpha$ | $\mathrm{pWPPD}_\alpha$ | $\mathrm{PPD}_{\alpha=0.5}(\mathrm{HD})$ | $\mathrm{PPD}_{\alpha=0.51}$ |
|---|---|---|---|---|---|---|
| 0.1 | 0 | 0.153 | 0.160 | 0.352 | 0.364 | 11.018 |
| 0.2 | 0.246 | 0.246 | 0.246 | 0.360 | | |
| 0.3 | 0.302 | 0.302 | 0.302 | 0.368 | | |
| 0.4 | 0.339 | 0.339 | 0.339 | 0.376 | | |

The estimated parameters under the LD are $\hat{\theta} = 3.059$ and 0.394 with and without the outlier, respectively. The same are $\hat{\theta} = 32.565$ and 0.424 when $\hat{\theta}$ is the minimum Pearson $\chi^2$ estimator.

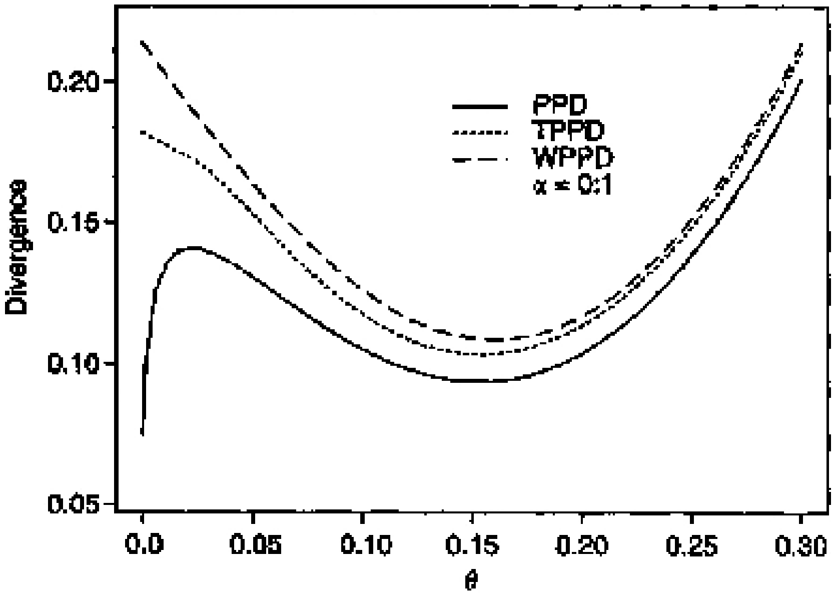

Fig. 3. Values of the divergences over $\theta$ for the Drosophila Data I example.

are not the global minima of the corresponding divergences at $\alpha = 0.1$. See Fig. 3 for a graph of the three divergences as a function of $\theta$ when $\alpha = 0.1$. This example demonstrates the possible pitfalls of the $\mathrm{PPD}_\alpha$ for small $\alpha$, and the need to modify it. A similar imploding behavior towards zero has also been noticed by Jones et al. (2001) in another density-based minimum divergence estimator for a different model.

A second concern was the small sample efficiency of the proposed estimators. Notice that the value of $C(-1)$ for the $\mathrm{PPD}_\alpha$ and the derived families is $1/(2\alpha^2)$, so that families with very small values of $\alpha$ put a huge weight on an empty cell ($\delta = -1$, i.e., $d(x) = 0$), and this can lead to the small sample performance of the methods to be quite inefficient at the model, although their outlier robustness properties make them otherwise attractive. A similar phenomenon for the Hellinger distance and some of its relations was observed, among others, by Lindsay (1994), Harris and Basu (1994) and Basu et al. (1996). We show here that an empty cell penalty as developed in Harris and Basu (1994) can lead to dramatic improvements in the method. The penalized versions of $\mathrm{TPPD}_\alpha$ and $\mathrm{WPPD}_\alpha$ are obtained by modifying the weight of an empty cell

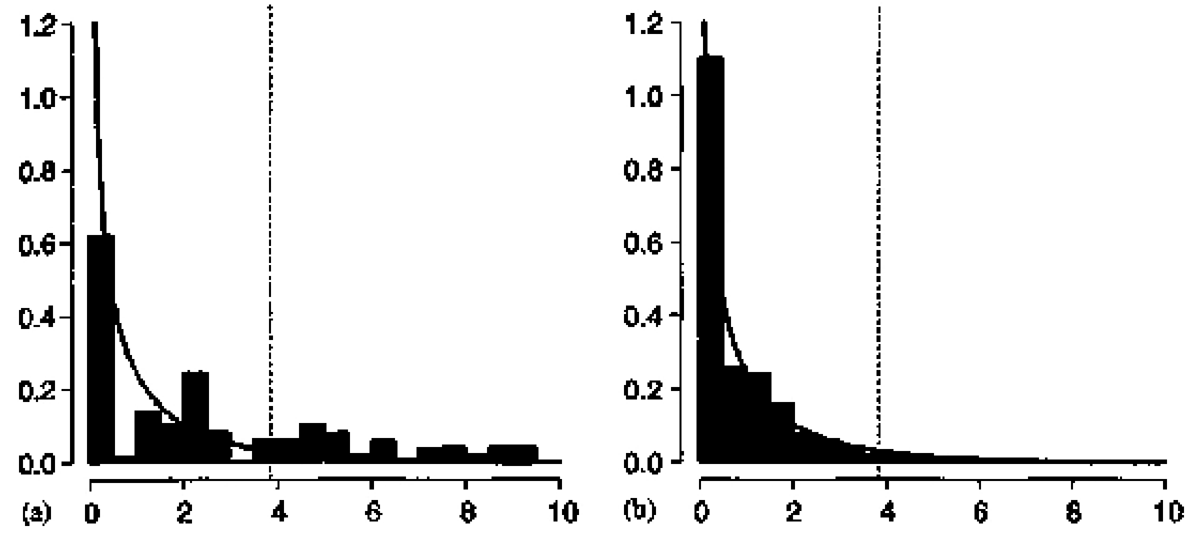

Fig. 4. Histogram of the null distribution of $\mathrm{TPPD}_{0.3}$ and $\mathrm{pTPPD}_{0.3}$ test statistics. Sample size $n=20$ with 100 replications: (a) ordinary test statistic; (b) penalized test statistic.

to be equal to that of LD as

$$\mathrm{pTPPD}_\alpha(d,f_\theta)=\frac{1}{2\alpha^2}\sum_{0<d/f_\theta<(1/(1-2\alpha))^{1/\alpha}}\left[\frac{d(x)^\alpha-f_\theta(x)^\alpha}{f_\theta(x)^\alpha}\right]^2 f_\theta(x)$$

$$+2(1-2\alpha)^{1/\alpha-2}\sum_{d/f_\theta\geqslant(1/(1-2\alpha))^{1/\alpha}}d(x)+\sum_{d=0}f_\theta(x),$$

$$\mathrm{pWPPD}_\alpha(d,f_\theta)=\frac{1}{2\alpha^2}\sum_{0<d/f_\theta<((1-\alpha)/(1-2\alpha))^{1/\alpha}}\left[\frac{d(x)^\alpha-f_\theta(x)^\alpha}{f_\theta(x)^\alpha}\right]^2 f_\theta(x)$$

$$+\sum_{d/f_\theta\geqslant((1-\alpha)/(1-2\alpha))^{1/\alpha}}\left[\frac{(1-2\alpha)^{1/\alpha-2}}{(1-\alpha)^{1/\alpha-1}}d(x)-\frac{1}{2(1-2\alpha)}f_\theta(x)\right]$$

$$+\sum_{d=0}f_\theta(x).$$

While the improved performance of the penalized estimators and tests will be self-evident in the simulations, here we present a small graphical investigation of the nature of improvement using the test statistics and their asymptotic limits. We take the Poisson $(\theta)$ model, generate data from Poisson (5), and consider testing $H_0:\theta=5$ versus $H_1:\theta\neq 5$. For illustration we choose $\alpha=0.3$. In Figs. 4(a) and (b) we present the histograms of the test statistics for the $\mathrm{TPPD}_{0.3}$ and $\mathrm{pTPPD}_{0.3}$ methods. The sample size was $n=20$ with 100 replications. We also superimpose the $\chi^2(1)$ density on it, which is its asymptotic limit. Clearly, the $\chi^2$ curve provides a far superior approximation for the histogram of the penalized test statistic—particularly the tail part. The vertical line represents the 5% critical point of $\chi^2(1)$.

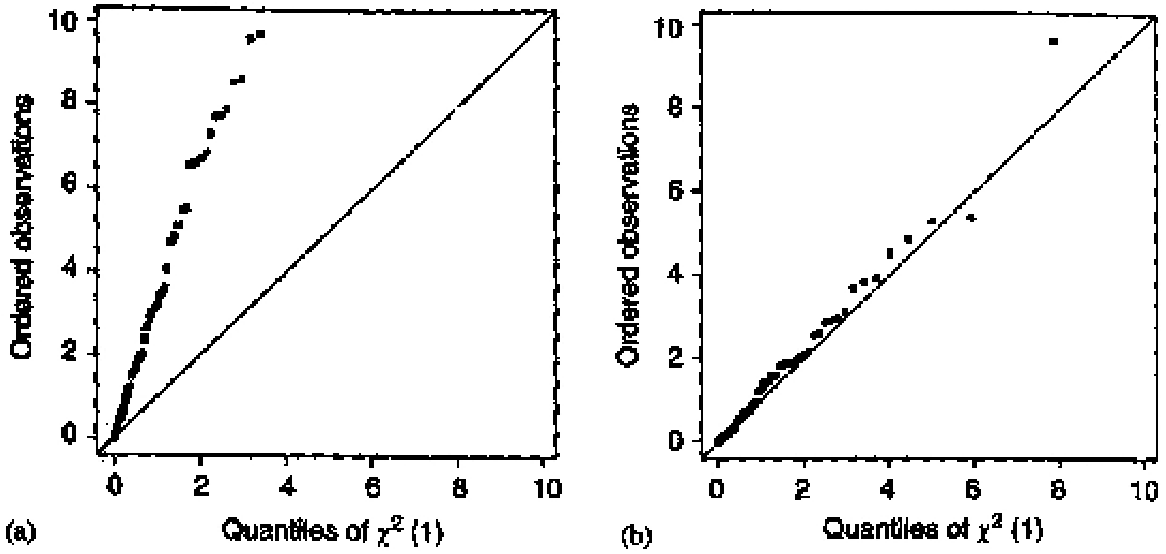

Fig. 5. $\chi^2(1)$ Q–Q plot of WPPD$_{0.3}$ and pWPPD$_{0.3}$ test statistics. Sample size $n = 100$ with 100 replications: (a) ordinary test statistic; (b) penalized test statistic.

For the same hypotheses and same true distribution, in Figs. 5(a) and (b) we present the probability plots (Wilk and Gnanadesikan, 1968) of the quantiles of the ordinary and penalized version of the WPPD$_{0.3}$ test statistics against the quantiles of the $\chi^2(1)$ distribution. A sample size $n = 100$ with 100 replications was used. The significant improvement due to penalty is apparent.

### 4.2. Examples

We applied the methods proposed in this paper to some real data sets. The first example involves the incidence of peritonitis on $n = 390$ kidney patients (Table 3). A glance at the data suggests that a *geometric* model with $\theta$ around $\frac{1}{2}$ may fit the data well. The data set, provided by Prof. P.W.M. John, was previously analyzed by Basu and Basu (1998). The observed frequency ($O_k$) of the number of cases of peritonitis ($k$) is modeled by the geometric distribution with success probability $\theta$. For an estimate $\hat{\theta}$, the expected frequencies are then obtained as $E_k = n\hat{\theta}(1-\hat{\theta})^k$. The largest number of cases of peritonitis is $k = 12$, so we merged all the expected frequencies for $k \geqslant 12$. To assess the goodness-of-fit of the model, we use the log likelihood ratio statistic which is given for this data as

$$G^2 = 2\sum_{k=0}^{12} O_k \log(O_k/E_k).$$

In this example the fit provided by the MLE is excellent; those for the estimators based on the penalized divergence are almost as good, and certainly much better than those for the estimators based on the ordinary divergence. The two marginally large observations at 10 and 12 have little impact since the sample size is so large. This example shows that when the data roughly follows the model the penalized methods are close to likelihood based ones in performance.

Table 3
The observed frequencies ($O_k$) of the number of cases ($k$) of peritonitis for each of 390 kidney patients and the expected frequencies under different methods with the goodness-of-fit likelihood ratio statistics ($G^2$)

| $\alpha$ | $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12+ | $G^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $O_k$ | 199 | 94 | 46 | 23 | 17 | 4 | 4 | 1 | 0 | 0 | 1 | 0 | 1 | — |
| | ML | 193.5 | 97.5 | 49.1 | 24.7 | 12.5 | 6.3 | 3.2 | 1.6 | 0.8 | 0.4 | 0.2 | 0.1 | 0.1 | 10.4 |
| $TPPD_\alpha$ | | | | | | | | | | | | | | | |
| 0.1 | | 237.8 | 92.8 | 36.2 | 14.1 | 5.5 | 2.2 | 0.8 | 0.3 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 52.4 |
| 0.2 | | 216.8 | 96.3 | 42.8 | 19.0 | 8.4 | 3.7 | 1.7 | 0.7 | 0.3 | 0.1 | 0.1 | 0.0 | 0.0 | 21.9 |
| 0.3 | | 207.8 | 97.1 | 45.3 | 21.2 | 9.9 | 4.6 | 2.2 | 1.0 | 0.5 | 0.2 | 0.1 | 0.0 | 0.0 | 14.8 |
| 0.4 | | 202.9 | 97.3 | 46.7 | 22.4 | 10.8 | 5.2 | 2.5 | 1.2 | 0.6 | 0.3 | 0.1 | 0.1 | 0.1 | 12.3 |
| 0.5 | | 199.1 | 97.5 | 47.7 | 23.4 | 11.4 | 5.6 | 2.7 | 1.3 | 0.7 | 0.3 | 0.2 | 0.1 | 0.1 | 11.1 |
| $pTPPD_\alpha$ | | | | | | | | | | | | | | | |
| 0.1 | | 200.6 | 97.4 | 47.3 | 23.0 | 11.2 | 5.4 | 2.6 | 1.3 | 0.6 | 0.3 | 0.1 | 0.1 | 0.1 | 11.6 |
| 0.2 | | 200.0 | 97.4 | 47.5 | 23.1 | 11.3 | 5.5 | 2.7 | 1.3 | 0.6 | 0.3 | 0.2 | 0.1 | 0.1 | 11.4 |
| 0.3 | | 199.3 | 97.5 | 47.6 | 23.3 | 11.4 | 5.6 | 2.7 | 1.3 | 0.7 | 0.3 | 0.2 | 0.1 | 0.1 | 11.2 |
| 0.4 | | 198.3 | 97.5 | 47.9 | 23.6 | 11.6 | 5.7 | 2.8 | 1.4 | 0.7 | 0.3 | 0.2 | 0.1 | 0.1 | 11.0 |
| 0.5 | | 196.7 | 97.5 | 48.3 | 23.9 | 11.9 | 5.9 | 2.9 | 1.4 | 0.7 | 0.4 | 0.2 | 0.1 | 0.1 | 10.7 |
| $WPPD_\alpha$ | | | | | | | | | | | | | | | |
| 0.1 | | 237.7 | 92.8 | 36.2 | 14.2 | 5.5 | 2.2 | 0.8 | 0.3 | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 | 52.2 |
| 0.2 | | 216.6 | 96.3 | 42.8 | 19.0 | 8.5 | 3.8 | 1.7 | 0.7 | 0.3 | 0.1 | 0.1 | 0.0 | 0.0 | 21.8 |
| 0.3 | | 207.7 | 97.1 | 45.4 | 21.2 | 9.9 | 4.6 | 2.2 | 1.0 | 0.5 | 0.2 | 0.1 | 0.0 | 0.0 | 14.8 |
| 0.4 | | 202.8 | 97.3 | 46.7 | 22.4 | 10.8 | 5.2 | 2.5 | 1.2 | 0.6 | 0.3 | 0.1 | 0.1 | 0.1 | 12.3 |
| 0.5 | | 199.1 | 97.5 | 47.7 | 23.4 | 11.4 | 5.6 | 2.7 | 1.3 | 0.7 | 0.3 | 0.2 | 0.1 | 0.1 | 11.1 |
| $pWPPD_\alpha$ | | | | | | | | | | | | | | | |
| 0.1 | | 200.4 | 97.4 | 47.4 | 23.0 | 11.2 | 5.4 | 2.6 | 1.3 | 0.6 | 0.3 | 0.1 | 0.1 | 0.1 | 11.5 |
| 0.2 | | 199.9 | 97.4 | 47.5 | 23.2 | 11.3 | 5.5 | 2.7 | 1.3 | 0.6 | 0.3 | 0.2 | 0.1 | 0.1 | 11.3 |
| 0.3 | | 199.2 | 97.5 | 47.7 | 23.3 | 11.4 | 5.6 | 2.7 | 1.3 | 0.7 | 0.3 | 0.2 | 0.1 | 0.1 | 11.2 |
| 0.4 | | 198.3 | 97.5 | 47.9 | 23.6 | 11.6 | 5.7 | 2.8 | 1.4 | 0.7 | 0.3 | 0.2 | 0.1 | 0.1 | 11.0 |
| 0.5 | | 196.7 | 97.5 | 48.3 | 23.9 | 11.9 | 5.9 | 2.9 | 1.4 | 0.7 | 0.4 | 0.2 | 0.1 | 0.1 | 10.7 |

The second example also involves data from Woodruff et al. (1984). The responses now are the frequencies of frequencies of daughter flies having a recessive lethal mutation on the X-chromosome where the male parent was either exposed to a dose of chemical or to control conditions. This data set, also analyzed by Simpson (1989, Table 5) will be referred to as the *Drosophila Data II*. The responses are modeled as Poissons with mean $\theta_1$ (control), and $\theta_2$ (exposed) respectively. For testing $H_0 : \theta_1 \geqslant \theta_2$ against $H_1 : \theta_1 < \theta_2$, a two sample signed divergence is appropriate. Suppose that random samples of size $n_i$ are available from the population with density $f_{\theta_i}(\cdot)$ and let $d_i(\cdot)$ be the empirical density of $i$th sample, $i = 1, 2$. For a divergence $\rho(\cdot)$ between two densities, define the overall divergence for the two sample case as

$$D = D(\theta_1, \theta_2) = \frac{1}{n_1 + n_2}(n_1\rho(d_1, f_{\theta_1}) + n_2\rho(d_2, f_{\theta_2})).$$

Table 4
The signed divergence statistics and their $p$-values for the Drosophila Data II

| Divergence | $\alpha$ | All observations | | Outliers deleted | |
|---|---|---|---|---|---|
| | | signed div. | $p$-value | signed div. | $p$-value |
| LD | | 2.595 | 0.002 | 1.099 | 0.136 |
| HD | | 0.698 | 0.243 | 0.743 | 0.229 |
| pHD | | 0.707 | 0.240 | 0.750 | 0.227 |
| TPPD$_\alpha$ | 0.1 | 0.028 | 0.489 | 0.187 | 0.426 |
| | 0.2 | 0.105 | 0.458 | 0.226 | 0.411 |
| | 0.3 | 0.244 | 0.404 | 0.326 | 0.372 |
| | 0.4 | 0.448 | 0.327 | 0.507 | 0.306 |
| pTPPD$_\alpha$ | 0.1 | 0.027 | 0.489 | 0.187 | 0.426 |
| | 0.2 | 0.104 | 0.459 | 0.225 | 0.411 |
| | 0.3 | 0.244 | 0.404 | 0.326 | 0.372 |
| | 0.4 | 0.451 | 0.326 | 0.509 | 0.305 |
| WPPD$_\alpha$ | 0.1 | 0.162 | 0.436 | 0.247 | 0.402 |
| | 0.2 | 0.171 | 0.432 | 0.264 | 0.396 |
| | 0.3 | 0.245 | 0.403 | 0.327 | 0.372 |
| | 0.4 | 0.448 | 0.327 | 0.507 | 0.306 |
| pWPPD$_\alpha$ | 0.1 | 0.162 | 0.436 | 0.248 | 0.402 |
| | 0.2 | 0.171 | 0.432 | 0.264 | 0.396 |
| | 0.3 | 0.245 | 0.403 | 0.327 | 0.372 |
| | 0.4 | 0.451 | 0.326 | 0.509 | 0.305 |

Given the ordinary divergence test statistic $t_n = 2n(\hat{D}_0 - \hat{D})$, where $\hat{D}_0$ and $\hat{D}$ are the minimizers of $D(\cdot,\cdot)$ under the null and without any restrictions respectively, the signed divergence statistic is given by $t_n^{1/2}\,\mathrm{sign}(\hat{\theta}_2 - \hat{\theta}_1)$ where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the unrestricted minimum divergence estimators of the parameters; for both the ordinary divergence and the penalized divergence, the signed divergence test is asymptotically equivalent to the signed likelihood ratio test. For the full data and the reduced data (after removing the two large observations from the treated group) the signed divergences and the associated $p$-values using the standard normal approximation are given in Table 4. The results for the Hellinger distance (HD) and the penalized Hellinger distance (pHD) are also provided.

The presence or absence of the two large counts in the treated group has little effect on the robust methods. The null hypothesis, that the mean number for the control group is no smaller than the treated group is supported in either case. The conclusions, however, are opposite when one uses the signed likelihood ratio test. The outliers cause the result to be significant in this case. Also, the $p$-values for the ordinary and penalized statistics are very close, indicating that the robustness property has not been compromised by the use of the penalty in this case.

Table 5
Estimated biases and mean square errors of the estimators under consideration

| $\alpha$ | $\mathrm{TPPD}_\alpha$ | | $\mathrm{pTPPD}_\alpha$ | | $\mathrm{WPPD}_\alpha$ | | $\mathrm{pWPPD}_\alpha$ | |
|---|---|---|---|---|---|---|---|---|
| | Bias | MSE | Bias | MSE | Bias | MSE | Bias | MSE |
| *Sample size* $n = 20$ | | | | | | | | |
| 0.1 | −0.6163 | 1.1902 | −0.1031 | 0.3891 | −0.6144 | 1.1832 | −0.0958 | 0.3733 |
| 0.2 | −0.5182 | 0.9108 | −0.0996 | 0.3521 | −0.5154 | 0.9026 | −0.0935 | 0.3421 |
| 0.3 | −0.4038 | 0.6596 | −0.0884 | 0.3181 | −0.4003 | 0.6492 | −0.0852 | 0.3136 |
| 0.4 | −0.2774 | 0.4402 | −0.0692 | 0.2882 | −0.2764 | 0.4384 | −0.0682 | 0.2872 |
| 0.5 | −0.1587 | 0.3126 | −0.0404 | 0.2633 | −0.1587 | 0.3126 | −0.0404 | 0.2633 |
| MLE | 0.0079 | 0.2493 | | | | | | |
| *Sample size* $n = 50$ | | | | | | | | |
| 0.1 | −0.5978 | 0.7866 | −0.0701 | 0.1471 | −0.5955 | 0.7805 | −0.0677 | 0.1450 |
| 0.2 | −0.4078 | 0.4098 | −0.0648 | 0.1343 | −0.4063 | 0.4074 | −0.0631 | 0.1330 |
| 0.3 | −0.2751 | 0.2414 | −0.0567 | 0.1232 | −0.2742 | 0.2403 | −0.0557 | 0.1226 |
| 0.4 | −0.1772 | 0.1601 | −0.0450 | 0.1137 | −0.1769 | 0.1599 | −0.0447 | 0.1135 |
| 0.5 | −0.1009 | 0.1206 | −0.0287 | 0.1059 | −0.1009 | 0.1206 | −0.0287 | 0.1059 |
| MLE | 0.0020 | 0.1005 | | | | | | |
| *Sample size* $n = 100$ | | | | | | | | |
| 0.1 | −0.5510 | 0.5780 | −0.0572 | 0.0673 | −0.5493 | 0.5745 | −0.0560 | 0.0668 |
| 0.2 | −0.3165 | 0.2206 | −0.0512 | 0.0627 | −0.3157 | 0.2198 | −0.0504 | 0.0624 |
| 0.3 | −0.1947 | 0.1138 | −0.0437 | 0.0586 | −0.1943 | 0.1135 | −0.0433 | 0.0585 |
| 0.4 | −0.1197 | 0.0742 | −0.0344 | 0.0551 | −0.1196 | 0.0741 | −0.0343 | 0.0551 |
| 0.5 | −0.0673 | 0.0579 | −0.0225 | 0.0523 | −0.0673 | 0.0579 | −0.0225 | 0.0523 |
| MLE | 0.0003 | 0.0503 | | | | | | |

5000 random samples were drawn from Poisson (5) with sample size $n = 20, 50, 100$.

### 4.3. Simulation results

In the first study, the data are generated from the Poisson distribution with mean 5, and modeled as the Poisson ($\theta$) distribution. Next, data are generated from the 0.9 Poisson (5)+0.1 Poisson (15) mixture, and the assumed model is Poisson ($\theta$). Here, as well as in the rest of the paper, three sample sizes $n = 20, 50, 100$ are considered. In Tables 5 and 6, we have presented the bias and the mean square errors of the estimators of $\theta$ (against the target value of 5) obtained by minimizing the $\mathrm{WPPD}_\alpha$ and $\mathrm{TPPD}_\alpha$ and their penalized versions for several values of $\alpha$ for pure and contaminated Poisson data respectively. It is clear that the small sample efficiency at the model is an increasing function of $\alpha$. The performance of the penalized versions are remarkably better. At sample size $n = 100$, the efficiency of the $\mathrm{pWPPD}_{0.5}$ estimator is over 95% compared to the MLE. The performance of the $\mathrm{TPPD}_\alpha$ and $\mathrm{WPPD}_\alpha$ estimators are very close. For contaminated data, more robust methods (those with smaller values of $\alpha$) start doing better.

For comparison, corresponding values for $\mathrm{WLD}_\lambda$ are presented in Tables 7 and 8 for several values of $\lambda$. The efficiencies are now increasing in $\lambda$ under the model, while

Table 6
Estimated biases and mean square errors of the estimators under consideration

| $\alpha$ | $\text{TPPD}_\alpha$ | | $\text{pTPPD}_\alpha$ | | $\text{WPPD}_\alpha$ | | $\text{pWPPD}_\alpha$ | |
|---|---|---|---|---|---|---|---|---|
| | Bias | MSE | Bias | MSE | Bias | MSE | Bias | MSE |
| *Sample size $n = 20$* | | | | | | | | |
| 0.1 | −0.5112 | 1.1982 | 0.0106 | 0.5027 | −0.5094 | 1.1930 | 0.0260 | 0.4908 |
| 0.2 | −0.4218 | 0.9493 | 0.0115 | 0.4556 | −0.4179 | 0.9423 | 0.0264 | 0.4495 |
| 0.3 | −0.3107 | 0.7126 | 0.0253 | 0.4223 | −0.3051 | 0.7060 | 0.0387 | 0.4217 |
| 0.4 | −0.1782 | 0.5150 | 0.0603 | 0.4032 | −0.1724 | 0.5127 | 0.0705 | 0.4058 |
| 0.5 | 0.0281 | 0.4263 | 0.2002 | 0.4577 | 0.0281 | 0.4263 | 0.2002 | 0.4577 |
| MLE | 1.0038 | 1.7522 | | | | | | |
| *Sample size $n = 50$* | | | | | | | | |
| 0.1 | −0.3854 | 0.7158 | 0.0212 | 0.1861 | −0.3826 | 0.7114 | 0.0322 | 0.1859 |
| 0.2 | −0.2406 | 0.3923 | 0.0333 | 0.1708 | −0.2360 | 0.3901 | 0.0443 | 0.1717 |
| 0.3 | −0.1315 | 0.2472 | 0.0531 | 0.1608 | −0.1254 | 0.2461 | 0.0635 | 0.1625 |
| 0.4 | −0.0317 | 0.1827 | 0.0907 | 0.1601 | −0.0258 | 0.1828 | 0.0989 | 0.1621 |
| 0.5 | 0.1337 | 0.1868 | 0.2212 | 0.2089 | 0.1337 | 0.1868 | 0.2212 | 0.2089 |
| MLE | 1.0099 | 1.3133 | | | | | | |
| *Sample size $n = 100$* | | | | | | | | |
| 0.1 | −0.1495 | 0.4935 | 0.0460 | 0.0900 | −0.1451 | 0.4919 | 0.0559 | 0.0909 |
| 0.2 | −0.0499 | 0.1953 | 0.0609 | 0.0870 | −0.0436 | 0.1954 | 0.0709 | 0.0884 |
| 0.3 | 0.0140 | 0.1183 | 0.0833 | 0.0867 | 0.0214 | 0.1190 | 0.0928 | 0.0887 |
| 0.4 | 0.0781 | 0.0993 | 0.1228 | 0.0933 | 0.0848 | 0.1007 | 0.1305 | 0.0955 |
| 0.5 | 0.2189 | 0.1365 | 0.2517 | 0.1458 | 0.2189 | 0.1365 | 0.2517 | 0.1458 |
| MLE | 1.0061 | 1.1591 | | | | | | |

5000 random samples were drawn from 0.9 Poisson (5)+0.1 Poisson (15) with sample size $n = 20, 50, 100$.

Table 7
Estimated biases and mean square errors of the $\text{WLD}_\lambda$ estimators

| $\lambda$ | $n = 20$ | | $n = 50$ | | $n = 100$ | |
|---|---|---|---|---|---|---|
| | Bias | MSE | Bias | MSE | Bias | MSE |
| 0.5 | −0.1095 | 0.3152 | −0.0515 | 0.1156 | −0.0306 | 0.0540 |
| 0.632 | −0.0657 | 0.2850 | −0.0329 | 0.1093 | −0.0196 | 0.0524 |
| 0.692 | −0.0505 | 0.2777 | −0.0264 | 0.1073 | −0.0158 | 0.0517 |
| 0.763 | −0.0352 | 0.2702 | −0.0194 | 0.1052 | −0.0118 | 0.0512 |
| 0.802 | −0.0283 | 0.2660 | −0.0157 | 0.1042 | −0.0095 | 0.0510 |
| 0.845 | −0.0210 | 0.2617 | −0.0118 | 0.1031 | −0.0070 | 0.0508 |
| 0.875 | −0.0159 | 0.2587 | −0.0089 | 0.1025 | −0.0054 | 0.0507 |
| 0.936 | −0.0045 | 0.2530 | −0.0032 | 0.1016 | −0.0025 | 0.0505 |
| 1 | 0.0079 | 0.2493 | 0.0020 | 0.1005 | 0.0003 | 0.0503 |

5000 random samples were drawn from Poisson (5) with sample size $n = 20, 50, 100$.

smaller values of $\lambda$ are better for robustness. It appears that one can get similar degrees of small sample efficiency and robustness for $\text{WPPD}_\alpha$ and $\text{WLD}_\lambda$ by suitable choice of index parameters $\alpha$ and $\lambda$. Exact calibration of the $\alpha$ and $\lambda$ values are difficult, but

Table 8
Estimated biases and mean square errors of the $\mathrm{WLD}_\lambda$ estimators

| $\lambda$ | $n=20$ | | $n=50$ | | $n=100$ | |
|---|---|---|---|---|---|---|
| | Bias | MSE | Bias | MSE | Bias | MSE |
| 0.5 | −0.0013 | 0.4050 | 0.0693 | 0.1522 | 0.1132 | 0.0887 |
| 0.632 | 0.0535 | 0.3952 | 0.1075 | 0.1589 | 0.1507 | 0.0995 |
| 0.692 | 0.0777 | 0.3959 | 0.1279 | 0.1652 | 0.1713 | 0.1076 |
| 0.763 | 0.1105 | 0.4067 | 0.1572 | 0.1777 | 0.2012 | 0.1211 |
| 0.802 | 0.1336 | 0.4214 | 0.1773 | 0.1878 | 0.2217 | 0.1314 |
| 0.845 | 0.1621 | 0.4421 | 0.2047 | 0.2035 | 0.2491 | 0.1466 |
| 0.875 | 0.1868 | 0.4625 | 0.2289 | 0.2193 | 0.2738 | 0.1622 |
| 0.936 | 0.2619 | 0.5323 | 0.3060 | 0.2795 | 0.3519 | 0.2213 |
| 1 | 1.0038 | 1.7522 | 1.0099 | 1.3133 | 1.0061 | 1.1591 |

5000 random samples were drawn from 0.9 Poisson (5)+0.1 Poisson (15) with sample size $n=20, 50, 100$.

Table 9
Corresponding tuning parameters $\alpha$ and $\lambda$ obtained by equating the Winsorizing points

| $\alpha$ | $\to 0$ | 0.1 | 0.2 | $\frac{1}{4}$ | 0.3 | $\frac{1}{3}$ | 0.4 | 0.5 |
|---|---|---|---|---|---|---|---|---|
| $\lambda$ | 0.632 | 0.692 | 0.763 | 0.802 | 0.845 | 0.875 | 0.936 | 1 |
| $\delta_1(\alpha)=\delta_1(\lambda)$ | 1.718 | 2.247 | 3.214 | 4.063 | 5.458 | 7 | 14.588 | $\infty$ |

equating the Winsorizing point gives $\lambda = 1-(1-2\alpha/1-\alpha)^{1/\alpha}$. The resulting $\lambda$ values for several values of $\alpha$ are given Table 9. This however is just a crude correspondence. Visual inspection shows somewhat smaller values of $\lambda$ than given by the above relation will give better calibration.

We now turn our attention to problems of hypothesis testing. Here again we looked at the $\mathrm{TPPD}_\gamma$, $\mathrm{WPPD}_\alpha$, $\mathrm{WLD}_\lambda$ and $\mathrm{TLD}_\lambda$ families in detail. However, as the results are very similar, we only present the results for the $\mathrm{WPPD}_\alpha$ case. Once again we looked at the Poisson $(\theta)$ model, generated data from the Poisson (5) distribution and tested $H_0:\theta=5$ against $H_1:\theta\neq 5$. Since the distributions of the ordinary test statistics are very far off from the limiting $\chi^2$ distributions, we computed the empirical critical values for each of the test statistics at our true null distribution based on 5000 replications of the test statistic for all the three sample sizes considered. We have not presented these empirical critical values here, but by the time $n$ equaled 100, the critical values of $\mathrm{pWPPD}_{0.5}$ were practically equal to those of the LRT.

Next, we generated data from Poisson distributions with $\theta$ in the range (3,7), and determined the power of each of the tests for the same set of hypotheses based on both the $\chi^2$ critical values and empirically determined critical values. The results for the nominal level $\gamma=0.05$ are presented in Fig. 6 and are based on sample size 50 with 1000 replications. The thick solid line represents the likelihood ratio test for each case. Notice that when the $\chi^2$ critical values are used, some of the powers of the ordinary

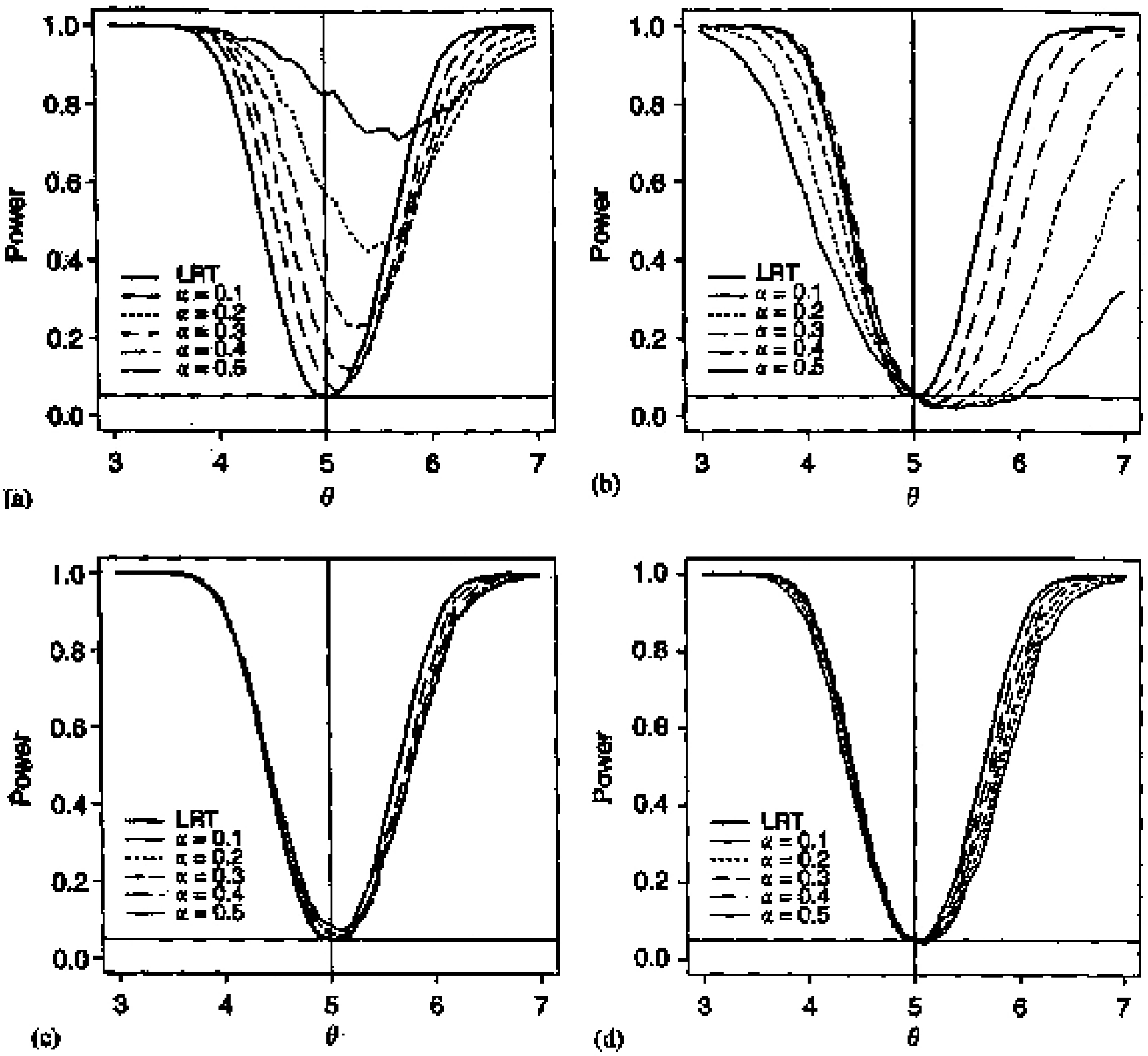

Fig. 6. Estimated powers for the tests under consideration testing $H_0 : \theta = 5$ versus $H_1 : \theta \neq 5$ with level $\gamma = 0.05$. 1000 random samples were drawn from Poisson ($\theta$) with sample size $n = 50$: (a) WPPD$_\alpha$ based on $\chi^2$ critical value; (b) WPPD$_\alpha$ based on empirical critical value; (c) pWPPD$_\alpha$ based on $\chi^2$ critical value; (d) pWPPD$_\alpha$ based on empirical critical value.

test statistics, particularly those for the lower values of $\alpha$ are very high, but that means very little because these tests are not even close to being level 0.05 test. When the true levels of the ordinary test statistics are held at 0.05 by using the empirically determined critical values, the actual power of the more robust divergences are easily found to be quite poor (Fig. 6b). However, most of theses problems are resolved by using the penalized divergences. Notice that the application of the penalty makes the performance of the methods based on empirically determined critical values and $\chi^2$ critical values dramatically closer. This is particularly encouraging since in actual practice when one wants to use these tests determining empirical critical values for each individual case is obviously not practical. Our results show that for the penalized tests the use of the $\chi^2$ critical values leads to results almost identical to the true powers

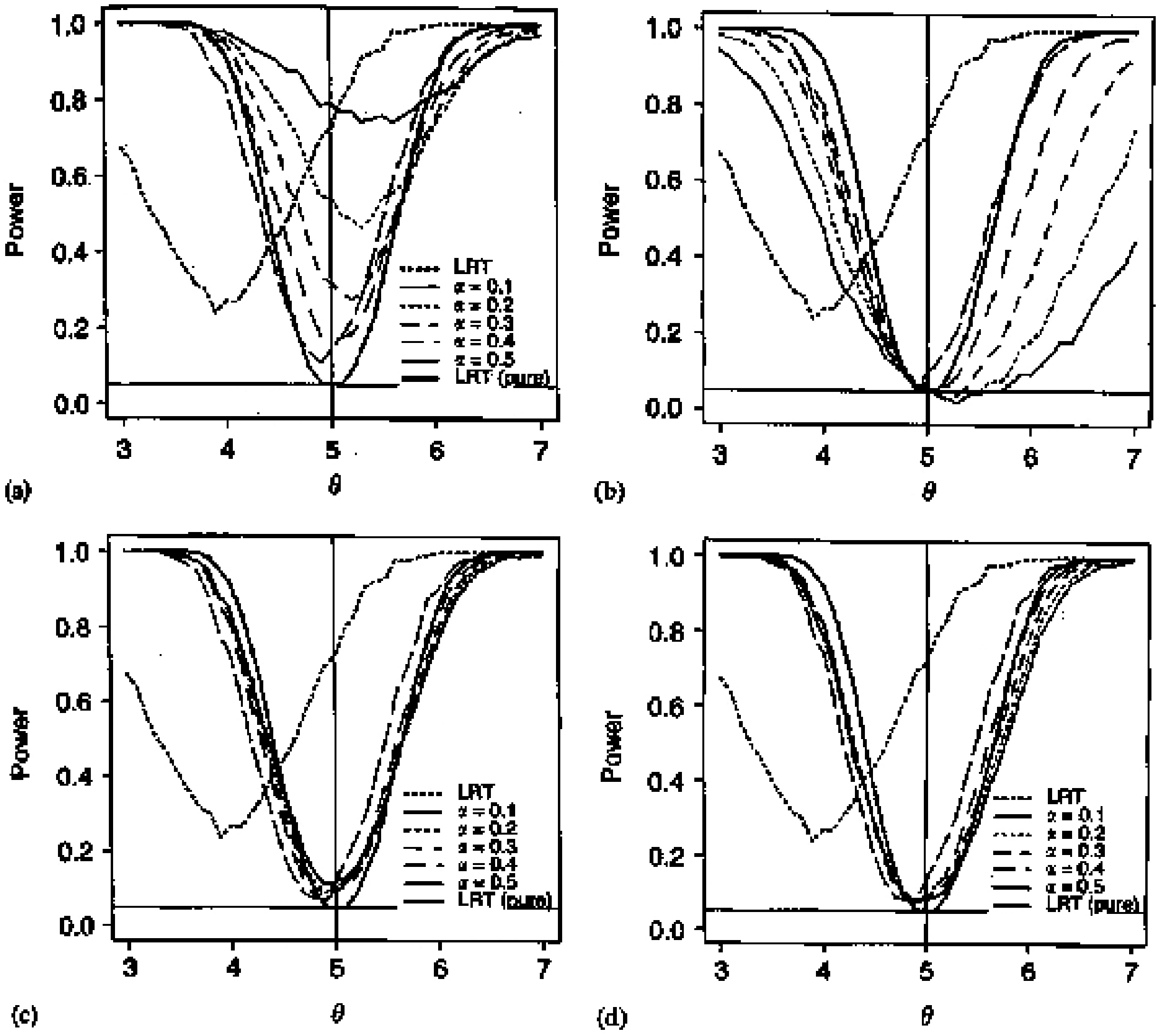

Fig. 7. Estimated powers for the tests under consideration testing $H_0:\theta=5$ versus $H_1:\theta\neq5$ with level $\gamma=0.05$. 1000 random samples were drawn from 0.9 Poisson $(\theta)$ + 0.1 Poisson (15) with sample size $n=50$: (a) WPPD$_\alpha$ based on $\chi^2$ critical value; (b) WPPD$_\alpha$ based on empirical critical value; (c) pWPPD$_\alpha$ based on $\chi^2$ critical value; (d) pWPPD$_\alpha$ based on empirical critical value.

of the tests. While results for other levels of significance and other sample sizes are not reported, they were very similar.

We next looked at the powers of the methods for the same set of hypotheses under contamination. Data are now generated from 0.9 Poisson $(\theta)$+0.1 Poisson (15) mixture. The results for the nominal level $\gamma=0.05$ are presented in Fig. 7 and are based on sample size 50 with 1000 replications. For comparison purposes, the power curve of the likelihood ratio test for the no contamination case is presented with the other graphs as the thick solid line. While the power curve of the likelihood ratio test under contamination shows a dramatic shift with substantial loss of power at several cases, the other curves are largely unchanged in comparison, demonstrating the relative stability of these test statistics under contamination.

## 5. Concluding remarks

We have shown that modification of the PPD$_\alpha$ leads to nice results, but we also believe that the story this investigation is telling about minimum disparity inference is more important than the story of PPD$_\alpha$ itself. We can summarize the lessons of this investigation as follows. There are infinitely many ways to construct distance measures for discrete models in such a way that the resulting estimators are first-order efficient. However, if one wishes to obtain reasonable statistical behavior in a wider sense, then:

(1) One should avoid residual adjustment functions $A(\cdot)$, or equivalently distance kernels $C(\cdot)$, that grow too fast as $\delta \to \infty$. In fact Lindsay (1994) showed that outlier stability follows from conditions such as $A(\delta) = O(\delta^{1/2})$ (or more generally $A(\delta) = O(\delta^{(k-1)/k})$ for $k > 0$ as $\delta \to \infty$, together with $A(-1)$ being finite). Notice that these conditions are not satisfied by the PPD$_\alpha$ family for $\alpha > 0.5$.

(2) At the other extreme, one should avoid an $A(\cdot)$ which is decreasing for some range of $\delta$. We have seen that when unmodified, the estimators from decreasing $A(\cdot)$ functions can lead to strange results. On the other hand, when modified to preserve their increasing nature, natural and meaningful results follow.

(3) One should also be careful about $A(\cdot)$ at the lower end of $\delta$'s range, as the behavior of $A(\cdot)$ when $\delta \to -1$ is also very important. In discrete model $\delta(x) = -1$ corresponds to the cell $x$ having no data, so $d(x) = 0$. If the RAF gives too large a weight to these cells, then the estimator become hypersensitive in small samples, and so has a large variance. Empty cells are extreme cases of inliers which represent values with less observed data than expected under the model. Notice that the MLE, while not outlier robust, is inlier robust. Our empty cell penalty essentially mimics the treatment of the empty cells by the MLE. We have shown how this simple empty-cell modification of $A(\cdot)$ can greatly improve statistical behavior.

Another lesson of the paper is that one can develop appropriate modifications of natural divergences for the purpose of improving the robustness and efficiency properties of the corresponding estimators and tests. In this particular paper we have experimented with the powered Pearson divergence and shown that the proposed modifications can lead to attractive inference procedures. In general, however, such improvements can be effected with many other well-known disparities and divergences. We have compared the modifications of the powered Pearson divergences to those of the likelihood disparity. The modifications of either divergence considered here appear to provide stable, satisfactory, and similar inference.

## Acknowledgements

Dr. Basu's research was done while he was visiting the Department of Statistics, Pennsylvania State University. Dr. Park's research was partially supported by

Clemson RGC award. Dr. Lindsay's research was partially supported by NSF Grant DMS 0104443. The authors thank the referees for useful suggestions.

## References

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

Basu, A., Lindsay, B.G., 1994. Minimum disparity estimation for continuous models: efficiency, distributions and robustness. Ann. Inst. Statist. Math. 46, 683–705.

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

Basu, A., Harris, I.R., Basu, S., 1997. Minimum distance estimation: the approach using density based distances. In: Maddala, G.S., Rao, C.R. (Eds.), Handbook of Statistics, Vol. 15, Robust Inference. Elsevier Science, New York, NY, pp. 21–48.

Basu, A., Park, C., Lindsay, B.G., Li, H., 2001. The powered Pearson divergence and estimation breakdown. Technical Report #01-04-25, Center for Likelihood Studies, Pennsylvania State University.

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

Chakraborty, B., Basu, A., Sarkar, S., 2003. Robustification of the MLE without loss in efficiency. Unpublished manuscript.

Cox, D.R., Hinkley, D.V., 1974. Theoretical Statistics. Chapman & Hall, London.

Cressie, N., Read, T.R.C., 1984. Multinomial goodness-of-fit tests. J. Roy. Statist. Soc. B 46, 440–464.

Hampel, F.R., Ronchetti, E., Rousseeuw, P.J., Stahel, W., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.

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

Jones, M.C., Hjort, N., Harris, I.R., Basu, A., 2001. A comparison of related density based minimum divergence estimators. Biometrika 88, 865–873.

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

Park, C., Basu, A., Lindsay, B.G., 2002. The residual adjustment function and weighted likelihood: a graphical interpretation of robustness of minimum disparity estimators. Comput. Statist. Data Anal. 39, 21–33.

Rao, C.R., 1961. Asymptotic efficiency and limiting information. In: Proceedings of the Fourth Berkeley Symposium, Vol. 1, University of California Press, Berkeley, pp. 531–546.

Rao, C.R., 1962. Efficient estimates and optimum inference procedures in large samples (with discussion). J. Roy. Statist. Soc. B 24, 46–72.

Read, T.R.C., Cressie, N., 1988. Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer, New York.

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

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

Tamura, R.N., Boos, D.D., 1986. Minimum Hellinger distance estimation for multivariate location and covariance. J. Amer. Statist. Assoc. 81, 223–229.

Wilk, M.B., Gnanadesikan, R., 1968. Probability plotting methods for the analysis of data. Biometrika 55, 1–17.

Woodruff, R.C., Mason, J.M., Valencia, R., Zimmering, A., 1984. Chemical mutagenesis testing in drosophila—I: Comparison of positive and negative control data for sex-linked recessive lethal mutations and reciprocal translocations in three laboratories. Environ. Mutagenesis 6, 189–202.