



## Some Variants of Minimum Disparity Estimation

By AYANENDRANATH BASU, CHANSEOK PARK, BRUCE G. LINDSAY and  
HAIHONG LI

Technical Report #01-03-15

2000

---

**Center for Likelihood Studies**  
DEPARTMENT OF STATISTICS  
THE PENNSYLVANIA STATE UNIVERSITY  
UNIVERSITY PARK, PA 16802

# Some Variants of Minimum Disparity Estimation

Ayanendranath Basu<sup>a,\*</sup>, Chanseok Park<sup>b</sup>, Bruce G. Lindsay<sup>b,†</sup> and Haihong Li<sup>b</sup>

<sup>a</sup>*Applied Statistics Unit, Indian Statistical Institute, Calcutta 700 035, India*

<sup>b</sup>*Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA*

Running Title: Minimum disparity estimation

---

## Abstract

This paper proposes several variants of disparity based inference (Lindsay 1994). 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  $\mathcal{F}_\Theta = \{F_\theta, \theta \in \Theta\}$ ,  $\Theta \subseteq \mathbb{R}^p$ . Throughout this paper we will refer to true distribution by  $G$ , which may or may not belong to  $\mathcal{F}_\Theta$ . We will assume that both  $G$  and  $\mathcal{F}_\Theta$  belong to  $\mathcal{G}$ , the class of all distributions having probability density functions (pdf's) with respect to a dominating measure (*e.g.*, Lebesgue measure for continuous models and counting measure for discrete ones). 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  $g$  and  $f_\theta$  respectively.

---

\*His work was done while he was visiting the Department of Statistics, Pennsylvania State University.

†His research was partially supported by NSF Grant DMS 9870193.

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  $G \in \mathcal{F}_\Theta$ ) 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 stable in terms of level and power when model assumptions are violated. In estimation problems, Beran (1977) and Tamura and Boos (1986) attempted to achieve the above goal 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 that are both robust and first order efficient for discrete models. An extension to the continuous case was considered by Basu and Lindsay (1994).

In this paper we consider a new class of density based divergences which generates first order efficient estimators, and several members of which have good robustness properties. 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 efficiency and robustness of these estimators. We will show that one must exercise caution in the creation of such first order efficient disparities as several of the estimators needed modifications in order to perform well. We also illustrate the performance of the method 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.

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.

## 2. Minimum disparity inference and proposed methods

### 2.1. Disparities and Residual Adjustment Function

Consider a parametric family of distributions  $F_\theta, \theta \in \Theta$ , having densities  $f_\theta(\cdot)$  with a countable sample space. Without loss of generality, let the sample space be  $\mathcal{X} = \{0, 1, 2, \dots\}$ . 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 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 convex function  $C(\cdot)$  as

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

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

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

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  $f_\theta(\cdot)$  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 and 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 efficiency 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 non-negative 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 Basu *et al.* (2000), and show that the results are very similar in either case.

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

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

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-square ( $\alpha = 1$ ) and Hellinger distance ( $\alpha = 1/2$ ), and can be thought of as  $L_2$  distance on the powered transformed distances.

As in the case of disparities, one can write  $\text{PPD}_\alpha$  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

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

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

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

However  $C''(\cdot)$  is always non-negative only when  $\alpha \geq \frac{1}{2}$ , so the  $C(\cdot)$  functions of the  $\text{PPD}_\alpha$  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  $\delta \geq -1$ , it can be seen from equations (3), (4) and (5) that for  $\alpha < 1/2$ , 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 the  $\text{PPD}_\alpha$  with  $\alpha < 1/2$  (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

$$\begin{aligned} \text{TPPD}_\alpha(d, f_\theta) &= \frac{1}{2\alpha^2} \sum_{\frac{d}{f_\theta} < (\frac{1}{1-2\alpha})^{1/\alpha}} f_\theta(x)^{1-2\alpha} (d(x)^\alpha - f_\theta(x)^\alpha)^2 \\ &\quad + 2(1-2\alpha)^{1/\alpha-2} \sum_{\frac{d}{f_\theta} \geq (\frac{1}{1-2\alpha})^{1/\alpha}} d(x) \\ \text{WPPD}_\alpha(d, f_\theta) &= \frac{1}{2\alpha^2} \sum_{\frac{d}{f_\theta} < (\frac{1-\alpha}{1-2\alpha})^{1/\alpha}} f_\theta(x)^{1-2\alpha} (d(x)^\alpha - f_\theta(x)^\alpha)^2 \\ &\quad + \sum_{\frac{d}{f_\theta} \geq (\frac{1-\alpha}{1-2\alpha})^{1/\alpha}} \left[ \frac{(1-2\alpha)^{1/\alpha-2}}{(1-\alpha)^{1/\alpha-1}} d(x) - \frac{1}{2(1-2\alpha)} f_\theta(x) \right]. \end{aligned}$$

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

$$C_{\text{TPPD}}(\delta) = \begin{cases} \frac{1}{2\alpha^2} [(\delta+1)^\alpha - 1]^2 & : \delta < (\frac{1}{1-2\alpha})^{1/\alpha} - 1 \\ 2(1-2\alpha)^{1/\alpha-2} (\delta+1) & : \delta \geq (\frac{1}{1-2\alpha})^{1/\alpha} - 1 \end{cases} \quad (6)$$

$$C_{\text{WPPD}}(\delta) = \begin{cases} \frac{1}{2\alpha^2} [(\delta+1)^\alpha - 1]^2 & : \delta < (\frac{1-\alpha}{1-2\alpha})^{1/\alpha} - 1 \\ \frac{(1-2\alpha)^{1/\alpha-2}}{(1-\alpha)^{1/\alpha-1}} (\delta+1) - \frac{1}{2(1-2\alpha)} & : \delta \geq (\frac{1-\alpha}{1-2\alpha})^{1/\alpha} - 1 \end{cases} \quad (7)$$

The inflection points  $\delta_1$  and the trimming points  $\delta_2$  for each of several values of  $\alpha$  are given in Table 1.

Table 1: The inflection and trimming points for  $\text{PPD}_\alpha$ .

$\alpha$	$\rightarrow 0$	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	$\rightarrow \frac{1}{2}$
$\delta_1$	$e-1$	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	$\infty$

Also we present the figures of the  $C(\delta)$  and  $A(\delta)$  functions of the  $\text{PPD}_\alpha$ ,  $\text{TPPD}_\alpha$  and  $\text{WPPD}_\alpha$  families corresponding to  $\alpha = 0.1$  in Figure 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 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  $\delta = 0$ .

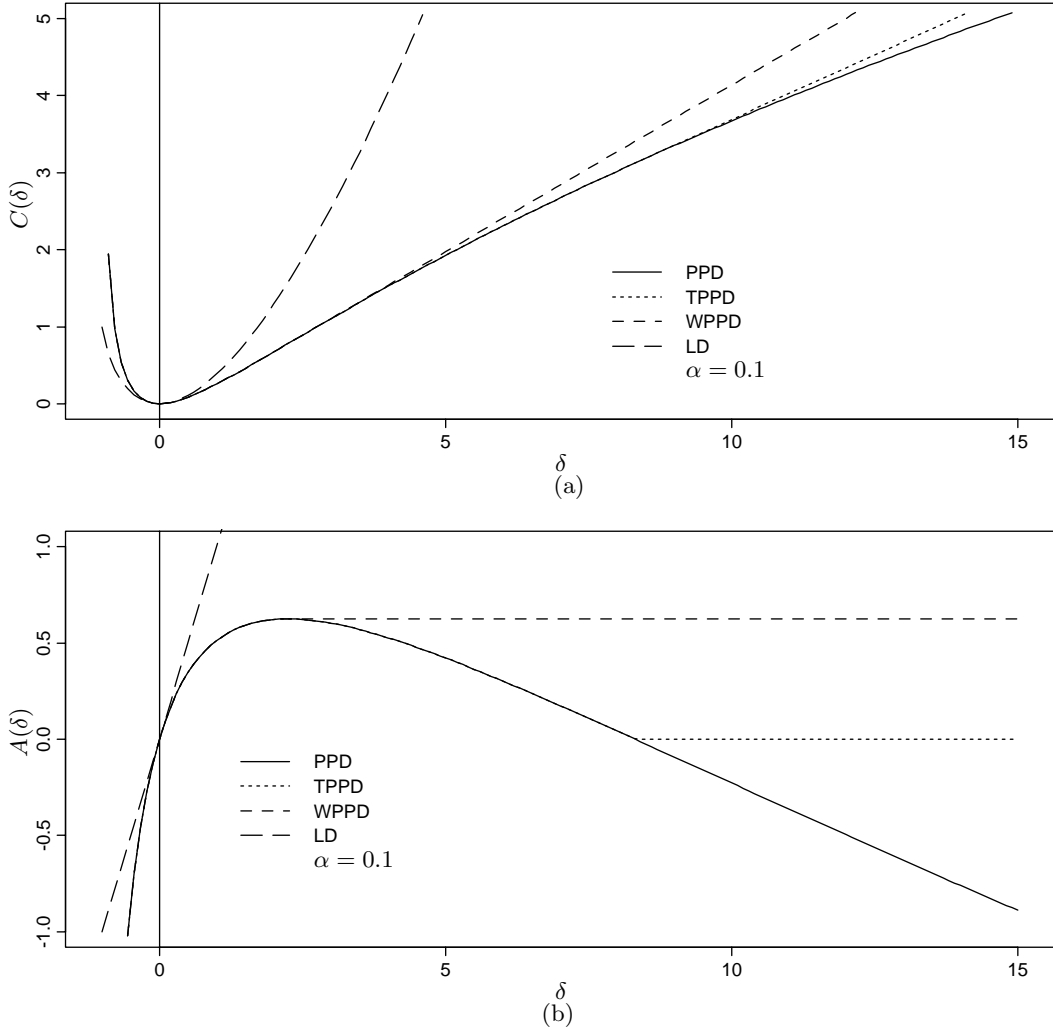


Figure 1:  $C(\cdot)$  and  $A(\cdot)$  functions of PPD, TPPD, WPPD and LD with  $\alpha = 0.1$ .

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

We note that although we arrived at the  $\text{PPD}_\alpha$  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  $\text{PPD}_\alpha$  family, the power divergences all correspond to convex  $C(\cdot)$  functions.

In this paper we will compare the estimators generated by the  $\text{PPD}_\alpha$  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  $\text{PPD}_\alpha$  discussed earlier. The likelihood disparity (LD) between  $d(\cdot)$  and  $f_\theta(\cdot)$  is defined by

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

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 Figure 1 also).

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

$$\begin{aligned} \text{WLD}_\lambda(d, f_\theta) &= \sum_{d/f_\theta < 1/\bar{\lambda}} [d(x) \log (d(x)/f_\theta(x)) + f_\theta(x) - d(x)] \\ &\quad - \sum_{d/f_\theta \geq 1/\bar{\lambda}} [d(x) \log \bar{\lambda} + \frac{\lambda}{\bar{\lambda}} f_\theta(x)] \\ \text{TLD}_\lambda(d, f_\theta) &= \sum_{d/f_\theta < 1/\bar{\lambda}} [d(x) \log (d(x)/f_\theta(x)) + f_\theta(x) - d(x)] \\ &\quad - \sum_{d/f_\theta \geq 1/\bar{\lambda}} [d(x)(\log \bar{\lambda} + \lambda)] \end{aligned}$$

The  $\text{WLD}_\lambda$  is a form of the robustified likelihood disparity (RLD) considered by Basu *et al.* (2000). It is easy to see that the  $C(\cdot)$  functions for  $\text{WLD}_\lambda$  are convex (although not strictly convex) while the  $C(\cdot)$  functions for  $\text{TLD}_\lambda$  are not. Also  $\text{WLD}_{\lambda=1} = \text{TLD}_{\lambda=1} = \text{LD}$ .

For better understanding the robustness of these methods, we also present the combined weight function  $w_c(\delta_c)$  (Park *et al.* 2000) for the  $\text{TPPD}_\alpha$ ,  $\text{WPPD}_\alpha$ ,  $\text{TLD}_\lambda$ , and  $\text{WLD}_\lambda$  families for different values of  $\alpha$  and  $\lambda$ . The combined weight function  $w_c(\delta_c)$  represent the relative impact of the observation in the estimating equation compared to maximum likelihood. Here we define a combined

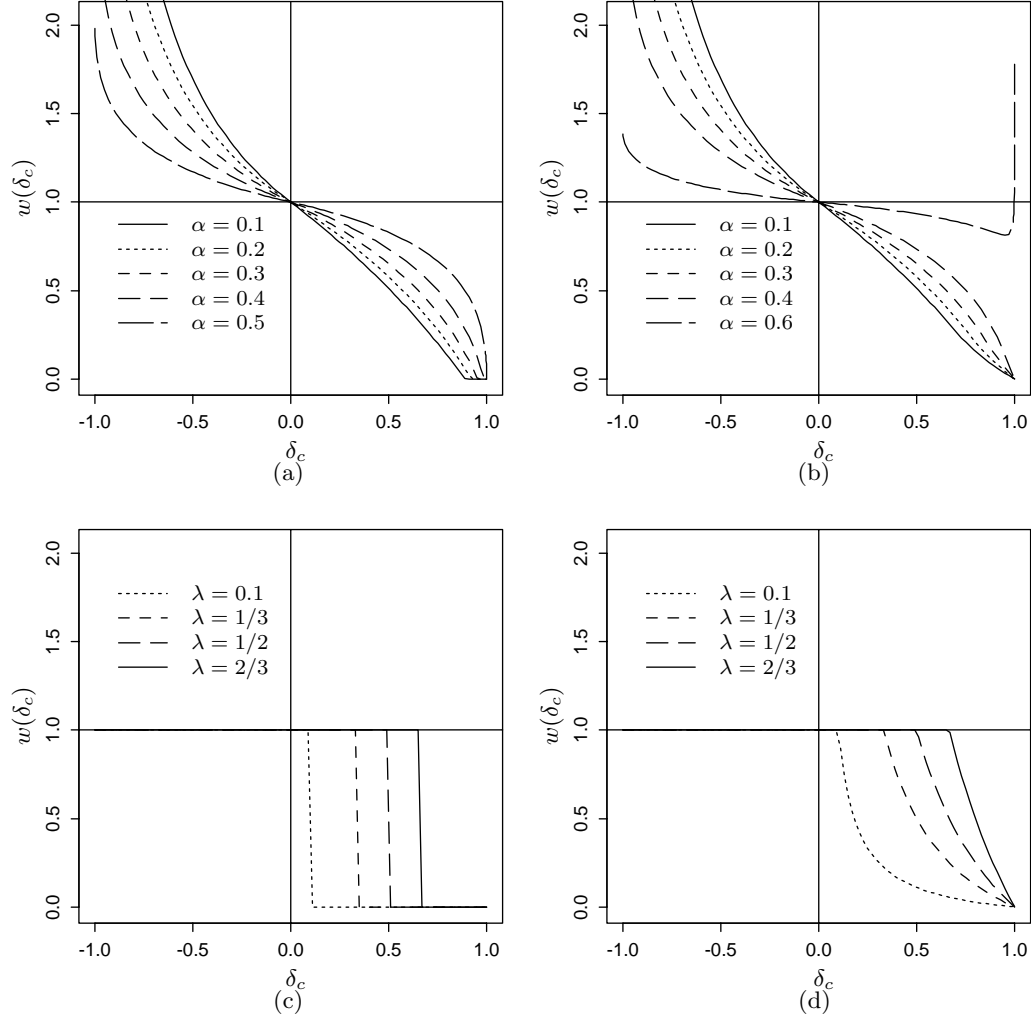


Figure 2: Combined weight functions of (a)  $\text{TPPD}_\alpha$ ; (b)  $\text{WPPD}_\alpha$ ; (c)  $\text{TLD}_\lambda$ ; (d)  $\text{WLD}_\lambda$ .

residual  $\delta_c$  as

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

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

$$w_c(\delta_c) = \begin{cases} \frac{A(\delta_c)}{\delta_c} & : -1 \leq \delta_c < 0 \\ A'(0) & : \delta_c = 0 \\ \frac{1 - \delta_c}{\delta_c} A\left(\frac{\delta_c}{1 - \delta_c}\right) & : 0 < \delta_c < 1 \\ A'(\infty) & : \delta_c = 1 \end{cases} . \quad (8)$$

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  $\delta_c \rightarrow 1$ . If we restrict  $\text{TPPD}_\alpha$  and  $\text{WPPD}_\alpha$  to  $\alpha \in [0, 1/2]$ , all the four families of weight functions satisfy this property, but the  $\text{TPPD}_\alpha$  and  $\text{WPPD}_\alpha$  appear to do this more smoothly. The  $\text{PPD}_\alpha$ ,  $\text{TPPD}_\alpha$  and  $\text{WPPD}_\alpha$  families coincide for  $\alpha \geq 1/2$ , and their behavior for large values of  $\delta$  makes them highly nonrobust which is demonstrated here only for  $\alpha = 0.6$  in Figure 2 (b), but is actually true for all  $\alpha > 1/2$ . For the rest of this paper, we focus our attention on the  $\text{TPPD}_\alpha$  and  $\text{WPPD}_\alpha$  families for  $\alpha \in (0, 1/2]$ .

### 3. Asymptotic results, breakdown issues, and Tests

#### 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  $\text{PPD}_\alpha$ ,  $\text{WPPD}_\alpha$  and  $\text{TPPD}_\alpha$  are all Fisher consistent, which implies that they are all weakly consistent under the model as well (*e.g.*, Cox and Hinkley, 1974, pp 288). For the asymptotic normality of the functional under the model, notice that for the  $\text{PPD}_\alpha$ , Lindsay's assumptions on  $A(\cdot)$  hold (Assumption 24, Lindsay 1994) so that the general proof for asymptotic normality works. To make the asymptotic argument for the  $\text{WPPD}_\alpha$  and  $\text{TPPD}_\alpha$  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  $\text{WPPD}_\alpha$  case, once again the assumptions of Lindsay (1994) hold except for the inflection point  $\delta_1$  where 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 the true distribution which belongs to the model. Similarly,  $A'(\cdot)$  and  $A''(\cdot)$  do not exist for the  $\text{TPPD}_\alpha$  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. Breakdown points

The breakdown point of a statistical functional is roughly the smallest fraction of contamination in the data that may cause an arbitrarily extreme value in the estimate. Here we establish the breakdown point of the  $\text{TPPD}_\alpha$  and  $\text{WPPD}_\alpha$  estimation functionals under the following set up.

Let  $T_\alpha(G)$  be the  $\text{TPPD}_\alpha$  or  $\text{WPPD}_\alpha$  estimation functional at the true distribution  $G$ . For  $\epsilon \in (0, 1)$ , consider the contamination model,

$$H_{\epsilon,m} = (1 - \epsilon)G + \epsilon K_m,$$

where  $\{K_m\}$  is a sequence of contaminating distributions, and  $h_{\epsilon,m}$ ,  $g$  and  $k_m$  are the corresponding densities with respect to a dominating measure  $\mu$  (e.g., counting measure). Given a contamination sequence  $\{K_m\}$  we will say that there is breakdown in  $T_\alpha$  for  $\epsilon$  level contamination if  $\lim_{m \rightarrow \infty} |T_\alpha(H_{\epsilon,m})| = \infty$ , in which case we are interested in  $\gamma = \inf\{\epsilon : \lim_{m \rightarrow \infty} |T_\alpha(H_{\epsilon,m})| = \infty\}$ . We write below  $\theta_m = T_\alpha(H_{\epsilon,m})$ , suppressing the  $\alpha$  and  $\epsilon$  subscripts for brevity.

We develop the following conditions for the breakdown point analysis. The conditions put appropriate structure on the model and on the contamination sequence which allows us to determine the behavior of the divergences under extreme forms of contamination. The proofs of this section, however, are not limited to discrete models and are expressed more generally in the form involving integrals. The following conditions **A1** - **A3** are conveniently expressed in terms of densities.

**Definition 1.** *A contaminating sequence of densities  $\{k_m\}$  will be called an outlier sequence relative to truth  $g(x)$  and model  $f_\theta(x)$  if:*

**A1.**  $\int \min(g(x), k_m(x)) d\mu(x) \rightarrow 0$  as  $m \rightarrow \infty$ . *That is, the contamination distribution becomes asymptotically singular to the true distribution.*

**A2.**  $\int \min(f_\theta(x), k_m(x)) d\mu(x) \rightarrow 0$  as  $m \rightarrow \infty$  uniformly for  $|\theta| \leq c$ , for any fixed  $c$ . *That is the contamination distribution is asymptotically singular to the specified models.*

*Finally we assume that*

**A3.**  $\int \min(g(x), f_{\theta_m}(x))d\mu(x) \rightarrow 0$  as  $m \rightarrow \infty$  if  $|\theta_m| \rightarrow \infty$  as  $m \rightarrow \infty$ . That is, large values of the parameter  $\theta$  give distributions which become singular to the true distribution.

Hereafter  $x$  will be suppressed from the integral notation. Intuitively, outlier sequences represent the worst possible type of contamination sequences. Before proving the asymptotic breakdown results, we first derive a boundedness result for the divergences.

**Theorem 1.** For given densities  $g(\cdot)$  and  $f(\cdot)$ , the  $\text{TPPD}_\alpha$  and  $\text{WPPD}_\alpha$  are bounded for  $0 \leq \alpha \leq \frac{1}{2}$ , that is,

$$\begin{aligned} 0 \leq \text{TPPD}_\alpha(g, f) &\leq \frac{1}{2\alpha^2} + 2(1 - 2\alpha)^{1/\alpha-2}, \\ 0 \leq \text{WPPD}_\alpha(g, f) &\leq \frac{1}{2\alpha^2} + \frac{(1 - 2\alpha)^{1/\alpha-2}}{(1 - \alpha)^{1/\alpha-1}}. \end{aligned}$$

The left equality holds if and only if  $g(\cdot) \equiv f(\cdot)$ , and the right equality holds when the supports of the two densities are disjoint almost everywhere, i.e.,  $\{x : f(x) > 0\} \cap \{x : g(x) > 0\}$  is a set of measure 0 with respect to the dominating measure.

**Proof.** Denote  $D_1(g, f) = (2\alpha^2)^{-1}f^{1-2\alpha}(g^\alpha - f^\alpha)^2$  and  $D_2(g, f) = 2(1 - 2\alpha)^{1/\alpha-2}g$ . Then the  $\text{TPPD}_\alpha$  can be rewritten as

$$\text{TPPD}_\alpha(g, f) = \int_{g < f} D_1(g, f)d\mu + \int_{(1-2\alpha)^{1/\alpha}g < f < g} D_1(g, f)d\mu + \int_{0 \leq f \leq (1-2\alpha)^{1/\alpha}g} D_2(g, f)d\mu.$$

First, for fixed  $f$  and  $g \in (0, f)$ , look at  $D_1(g, f)$  as a function of  $g$ .

$$\frac{\partial}{\partial g} D_1(g, f) = \frac{1}{\alpha} f^{1-2\alpha} (g^\alpha - f^\alpha) g^{\alpha-1} < 0, \quad \text{for } \forall g \in (0, f).$$

Since  $D_1(\cdot, f)$  is strictly decreasing for  $g \in (0, f)$  and right-continuous at  $g = 0$ , we have  $D_1(g, f) \leq D_1(0, f) = (2\alpha^2)^{-1}f$  for  $g \in (0, f)$  with the equality only when  $g = 0$ .

Second, for fixed  $g$  and  $f \in ((1 - 2\alpha)^{1/\alpha}g, g)$ , look at  $D_1(g, f)$  as a function of  $f$ .

$$\frac{\partial}{\partial f} D_1(g, f) = \frac{1}{2\alpha^2} (f^\alpha - g^\alpha) (f^\alpha - (1 - 2\alpha)g^\alpha) f^{-2\alpha},$$

and  $\frac{\partial}{\partial f} D_1(g, f) = 0$  at  $f = f^* = (1 - 2\alpha)^{1/\alpha}g$ . An inspection of the signs of the derivative  $\frac{\partial}{\partial f} D_1(g, f)$  on either side of  $f = f^*$  shows  $D_1(g, f) \leq D_1(g, f^*) = 2(1 - 2\alpha)^{1/\alpha-2}g$  for  $f \in ((1 - 2\alpha)^{1/\alpha}g, g)$ .

Next,  $f \in [0, (1 - 2\alpha)^{1/\alpha}g]$ , and it is easy to see  $D_1(g, f) \leq D_2(g, f) = 2(1 - 2\alpha)^{1/\alpha-2}g$ . Thus we have  $D_1(g, f) \leq D_2(g, f) = 2(1 - 2\alpha)^{1/\alpha-2}g$  for  $f < g$ .

It follows that

$$\begin{aligned} \text{TPPD}_\alpha(g, f) &\leq \int_{g < f} D_1(0, f)d\mu + \int_{f < g} D_2(g, f)d\mu \\ &\leq \int D_1(0, f)d\mu + \int D_2(g, f)d\mu = \frac{1}{2\alpha^2} + 2(1 - 2\alpha)^{1/\alpha-2}. \end{aligned}$$

It is easily shown that the equality holds only when the two densities have disjoint support almost everywhere. Similar arguments establish the result for  $\text{WPPD}_\alpha(g, f)$ .  $\square$

**Lemma 2.** Denote  $g_\eta^*(x) = \eta g(x)$  and for any  $\eta \in [0, 1]$ ,  $\text{TPPD}_\alpha(g^*, f) \geq C_{\text{TPPD}}(\eta - 1)$  and  $\text{WPPD}_\alpha(g^*, f) \geq C_{\text{WPPD}}(\eta - 1)$ , where  $C_{\text{TPPD}}(\cdot)$  and  $C_{\text{WPPD}}(\cdot)$  are the defining functions for the  $\text{TPPD}_\alpha$  and  $\text{WPPD}_\alpha$  in (6) and (7) respectively.

**Proof.** Even though  $g_\eta^*(\cdot)$  is not a density, one can formally calculate  $\text{TPPD}_\alpha(g_\eta^*, f)$ . For brevity, denote  $C(\cdot) = C_{\text{TPPD}}(\cdot)$ . For any  $\delta \geq -1$  and any  $\eta \in [0, 1]$ , we have  $C(\delta) \geq C'(\eta - 1)(\delta - (\eta - 1)) + C(\eta - 1)$ . It follows that for any  $\eta \in [0, 1]$ ,

$$C\left(\frac{\eta g}{f} - 1\right) \geq C'(\eta - 1)\left(\frac{\eta g}{f} - 1 - (\eta - 1)\right) + C(\eta - 1).$$

If we integrate both sides above with respect to  $F(\cdot)$ , we have  $\text{TPPD}_\alpha(\eta g, f) \geq C(\eta - 1)$ .

Note that  $\text{WPPD}_\alpha(\eta g, f) \geq C_{\text{WPPD}}(\eta - 1)$  holds directly by Jensen's inequality since  $C_{\text{WPPD}}(\cdot)$  is convex.  $\square$

**Theorem 3.** Let  $\{k_m\}$  be an outlier sequence of densities with respect to the true distribution and the model (i.e.  $\{k_m\}$  satisfies conditions **A1** and **A2**). In addition suppose that the model satisfies condition **A3** in relation to the true distribution. If the true distribution belongs to the model, then, for any  $\epsilon < 1/2$ ,  $\limsup_{m \rightarrow \infty} |T_\alpha(H_{\epsilon, m})| < \infty$  where  $T_\alpha$  is the  $\text{TPPD}_\alpha$  or  $\text{WPPD}_\alpha$  estimation functional. When the true distribution does not belong to the model,  $\limsup_{m \rightarrow \infty} |T_\alpha(H_{\epsilon, m})| < \infty$  whenever  $\epsilon < \epsilon^*$ , where  $\epsilon^* = \inf\{\epsilon : b(\epsilon) \leq \gamma(\epsilon)\}$  and for the  $\text{TPPD}_\alpha$  functional,  $b(\epsilon) = 2(1 -$

$\epsilon)(1 - 2\alpha)^{1/\alpha-2} + C_{\text{TPPD}}(\epsilon - 1)$  and  $\gamma(\epsilon) = 2\epsilon(1 - 2\alpha)^{1/\alpha-2} + \text{TPPD}_\alpha((1 - \epsilon)g, f_{\theta^*})$ , and  $\theta^*$  is the minimizer of  $\text{TPPD}_\alpha((1 - \epsilon)g, f_\theta)$ ; for the  $\text{WPPD}_\alpha$  functional,  $b(\epsilon) = (1 - \epsilon)(1 - 2\alpha)^{1/\alpha-2}/(1 - \alpha)^{1/\alpha-1} + C_{\text{WPPD}}(\epsilon - 1)$  and  $\gamma(\epsilon) = \epsilon(1 - 2\alpha)^{1/\alpha-2}/(1 - \alpha)^{1/\alpha-1} + \text{WPPD}_\alpha((1 - \epsilon)g, f_{\theta^*})$ , and  $\theta^*$  is the minimizer of  $\text{WPPD}_\alpha((1 - \epsilon)g, f_\theta)$ .

**Proof.** Let  $T_\alpha$  be the  $\text{TPPD}_\alpha$  estimation functional, and denote

$$D(g, f) = \begin{cases} (2\alpha^2)^{-1} f^{1-2\alpha} (g^\alpha - f^\alpha)^2 & : g/f < \left(\frac{1}{1-2\alpha}\right)^{1/\alpha} \\ 2(1 - 2\alpha)^{1/\alpha-2} g & : g/f \geq \left(\frac{1}{1-2\alpha}\right)^{1/\alpha} \end{cases}.$$

For brevity, denote  $C(\cdot) = C_{\text{TPPD}}(\cdot)$ . Given a level  $\epsilon$  of contamination, suppose, if possible, breakdown occurs, that is there exists a sequence  $\{k_m\}$  such that  $|\theta_m| \rightarrow \infty$  as  $m \rightarrow \infty$ , where  $\theta_m = T_\alpha(H_{\epsilon, m})$ . Define  $A_m = \{x : g(x) > \max(k_m(x), f_{\theta_m}(x))\}$ , so that

$$\text{TPPD}_\alpha(h_{\epsilon, m}, f_{\theta_m}) = \int_{A_m} D(h_{\epsilon, m}, f_{\theta_m}) d\mu + \int_{A_m^c} D(h_{\epsilon, m}, f_{\theta_m}) d\mu. \quad (9)$$

We start by determining the limit in  $m$  of the first term in the right of (9). Let  $\mathbb{I}_A(\cdot)$  be the indicator function of the set  $A$ . Notice that  $\int_{A_m} k_m d\mu = \int \mathbb{I}_{A_m} k_m d\mu \leq \int \min(g, k_m) d\mu \rightarrow 0$  as  $m \rightarrow \infty$  by **A1**, and thus the set  $A_m$  converges to a set of zero probability under  $k_m$ . Similarly it follows from **A3** that the set  $A_m$  converges to a set of zero probability under  $f_{\theta_m}$ . Next notice that  $A_m^c \subset B_{1, m} \cup B_{2, m}$  where  $B_{1, m} = \{x : g(x) \leq k_m(x)\}$  and  $B_{2, m} = \{x : g(x) \leq f_{\theta_m}(x)\}$ . The integrals  $\int \mathbb{I}_{B_{1, m}} g d\mu$  and  $\int \mathbb{I}_{B_{2, m}} g d\mu$  converge to zero as  $m \rightarrow \infty$  from conditions **A1** and **A3** respectively. Thus under  $g$ , the set  $A_m^c$  converges to a set of zero probability. Thus

$$\mathbb{I}_{A_m} D(h_{\epsilon, m}, f_{\theta_m}) \rightarrow D((1 - \epsilon)g, 0)$$

as  $m \rightarrow \infty$ .

Next notice from the proof of Theorem 1 that

$$\left| \mathbb{I}_{A_m} D(h_{\epsilon, m}, f_{\theta_m}) \right| \leq \left| D(h_{\epsilon, m}, f_{\theta_m}) \right| = D(h_{\epsilon, m}, f_{\theta_m}) \leq (2\alpha^2)^{-1} f_{\theta_m} + 2(1 - \epsilon)(1 - 2\alpha)^{1/\alpha-2} h_{\epsilon, m}$$

and  $\int [(2\alpha^2)^{-1} f_{\theta_m} + 2(1 - \epsilon)(1 - 2\alpha)^{1/\alpha-2} h_{\epsilon, m}] d\mu$  equals  $(2\alpha^2)^{-1} + 2(1 - \epsilon)(1 - 2\alpha)^{1/\alpha-2}$  for all  $m$ .

Thus by a generalized version of the dominated convergence theorem (Royden, 1988, page 92)

$$\left| \int_{A_m} D(h_{\epsilon, m}, f_{\theta_m}) d\mu - \int D((1 - \epsilon)g, 0) d\mu \right| \rightarrow 0, \quad (10)$$

which gives the limit to the first term in (9), since  $\int D((1-\epsilon)g, 0)d\mu = 2(1-\epsilon)(1-2\alpha)^{1/\alpha-2}$ .

Similarly, we have

$$\left| \int_{A_m^c} D(h_{\epsilon,m}, f_{\theta_m})d\mu - \int D(\epsilon k_m, f_{\theta_m})d\mu \right| \rightarrow 0 \quad (11)$$

which gives us the second term in (9). Putting (10) and (11) together, we get,

$$\liminf_{m \rightarrow \infty} \text{TPPD}_\alpha(h_{\epsilon,m}, f_{\theta_m}) = 2(1-\epsilon)(1-2\alpha)^{1/\alpha-2} + \liminf_{m \rightarrow \infty} \int D(\epsilon k_m, f_{\theta_m})d\mu$$

Notice that  $\int D(\epsilon k_m, f_{\theta_m})d\mu \geq C(\epsilon-1)$  by Lemma 2. It follows that

$$\liminf_{m \rightarrow \infty} \text{TPPD}_\alpha(h_{\epsilon,m}, f_{\theta_m}) \geq 2(1-\epsilon)(1-2\alpha)^{1/\alpha-2} + C(\epsilon-1). \quad (12)$$

We will have a contradiction to our assumption of the existence of a sequence  $\{k_m\}$  for which breakdown occurs if we can show that there exists a constant value  $\theta^*$  in the parameter space such that

$$\limsup_{m \rightarrow \infty} \text{TPPD}(h_{\epsilon,m}, f_{\theta^*}) < b(\epsilon) \quad (13)$$

where  $b(\epsilon)$  is the right hand side of equation (12) as then the  $\{\theta_m\}$  sequence above could not minimize  $\text{TPPD}_\alpha$  for every  $m$ . We will show that this is true for all  $\epsilon < 1/2$  under the model where  $\theta^*$  is the minimizer of  $\int D((1-\epsilon)g, f_\theta)d\mu$ .

Using analogous techniques, assumptions **A1** and **A3**, and Lemma 2 we obtain, for fixed any  $\theta$ ,

$$\begin{aligned} \lim_{m \rightarrow \infty} \text{TPPD}_\alpha(h_{\epsilon,m}, f_\theta) &= 2\epsilon(1-2\alpha)^{1/\alpha-2} + \int D((1-\epsilon)g, f_\theta)d\mu \\ &\geq 2\epsilon(1-2\alpha)^{1/\alpha-2} + \inf_{\theta} \int D((1-\epsilon)g, f_\theta)d\mu. \end{aligned} \quad (14)$$

with equality for  $\theta = \theta^*$ . Notice from (14) that among all fixed  $\theta$  the divergence  $\text{TPPD}_\alpha(h_{\epsilon,m}, f_\theta)$  is minimized in the limit by  $\theta^*$ .

If  $g(\cdot) = f_{\theta_0}(\cdot)$ , that is the true distribution belongs to the model,  $\int D((1-\epsilon)f_{\theta_0}, f_{\theta_0})d\mu = C(-\epsilon)$  which is also the lower bound (over  $\theta \in \Theta$ ) for  $D((1-\epsilon)f_{\theta_0}, f_\theta)$ . Thus in this case  $\theta^* = \theta_0$ , and from (14),

$$\lim_{m \rightarrow \infty} \text{TPPD}_\alpha(h_{\epsilon,m}, f_{\theta^*}) = \lim_{m \rightarrow \infty} \text{TPPD}_\alpha(h_{\epsilon,m}, f_{\theta_0}) = 2\epsilon(1-2\alpha)^{1/\alpha-2} + C(-\epsilon). \quad (15)$$

As a result asymptotically there is no breakdown for  $\epsilon$  level contamination when  $a(\epsilon) < b(\epsilon)$ , where  $a(\epsilon)$  is the right hand sides of equation (15). Note that  $a(\epsilon)$  and  $b(\epsilon)$  are strictly increasing and decreasing respectively in  $\epsilon$ , and  $a(1/2) = b(1/2)$ , so that asymptotically there is no breakdown and  $\limsup_{m \rightarrow \infty} |T_\alpha(H_{\epsilon,m})| < \infty$  for  $\epsilon < 1/2$ .

The proof also shows that under  $g = f_{\theta_0}$  as  $m \rightarrow \infty$ ,  $\theta_0$  is the minimizer of  $\text{TPPD}_\alpha(h_{\epsilon,m}, f_\theta)$  when  $\epsilon < 1/2$  for any outlier sequence  $\{k_m\}$ .

More generally, when  $g$  does not belong to the model, there is no breakdown for any outlier sequence whenever  $\epsilon < \epsilon^*$ , where

$$\epsilon^* = \inf\{\epsilon : b(\epsilon) \leq \gamma(\epsilon)\},$$

where  $\gamma(\epsilon)$  is the right hand side of (14).

The asymptotic breakdown results for the  $\text{WPPD}_\alpha$  estimation functional, including a breakdown point of 1/2 or more at the model, are similarly established.

□

### 3.3. 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  $\text{TPPD}_\alpha$  and  $\text{WPPD}_\alpha$ . Given the empirical density  $d(\cdot)$ , the  $\text{WPPD}_\alpha$  test statistic for the above hypothesis is given by

$$2n \left[ \text{WPPD}_\alpha(d, f_{\hat{\theta}_0}) - \text{WPPD}_\alpha(d, f_{\hat{\theta}}) \right],$$

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

## 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  $WLD_\lambda$  family (those based on  $TLD_\lambda$  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  $PPD_\alpha$  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*.

Table 2: The estimated parameters under the Poisson model for the *Drosophila Data I*. The estimated parameters under the LD are  $\hat{\theta} = 3.059$  and  $\hat{\theta} = 0.394$  with and without the outlier respectively. The same are  $\hat{\theta} = 32.565$  and  $\hat{\theta} = 0.424$  when  $\hat{\theta}$  is the minimum Pearson chi-square estimator.

$\alpha$	$PPD_\alpha$	$TPPD_\alpha$	$WPPD_\alpha$	$pWPPD_\alpha$	$PPD_{\alpha=0.5}(\text{HD})$	$PPD_{\alpha=0.51}$
0.1	0	0.153	0.160	0.352		
0.2	0.246	0.246	0.246	0.360	0.364	11.018
0.3	0.302	0.302	0.302	0.368		
0.4	0.339	0.339	0.339	0.376		

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  $PPD_\alpha$ ,  $TPPD_\alpha$ ,  $WPPD_\alpha$  and  $pWPPD_\alpha$  families. The  $pWPPD_\alpha$  method is a modification of  $WPPD_\alpha$  to be introduced later in this section. For this model it can be shown that  $PPD_\alpha(d, f_\theta)$  converges to  $c_1 = \frac{1}{2\alpha^2}(d(0)^\alpha - 1)^2$  as  $\theta \rightarrow 0$ . In

this example it appears (Figure 3) that  $\theta = 0$  is the global minimum of  $\text{PPD}_\alpha$  for  $\alpha = 0.1$ . It means that the estimator tends to “implode” toward 0 in this case. On the other hand as  $\theta \rightarrow 0$   $\text{TPPD}_\alpha(d, f_\theta)$  converges to  $c_2 = \frac{1}{2\alpha^2}(d(0)^\alpha - 1)^2 + 2(1 - 2\alpha)^{1/\alpha-2} \sum_{x>0} d(x)$ , and  $\text{WPPD}_\alpha(d, f_\theta)$  converges to  $c_3 = \frac{1}{2\alpha^2}(d(0)^\alpha - 1)^2 + \frac{(1-2\alpha)^{1/\alpha-2}}{(1-\alpha)^{1/\alpha-1}} \sum_{x>0} d(x)$ . Notice that  $c_1 \leq c_2 \leq c_3$  for  $\alpha < \frac{1}{2}$ , and at least for the Drosophila Data I example,  $c_2, c_3$  are not the global minima of the corresponding divergences at  $\alpha = 0.1$ . See Figure 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  $\text{PPD}_\alpha$  for small  $\alpha$ , and the need to modify it. A similar imploding behavior towards zero has also been noticed by Jones *et al.* (2000) in another density based minimum divergence estimator for a different model.

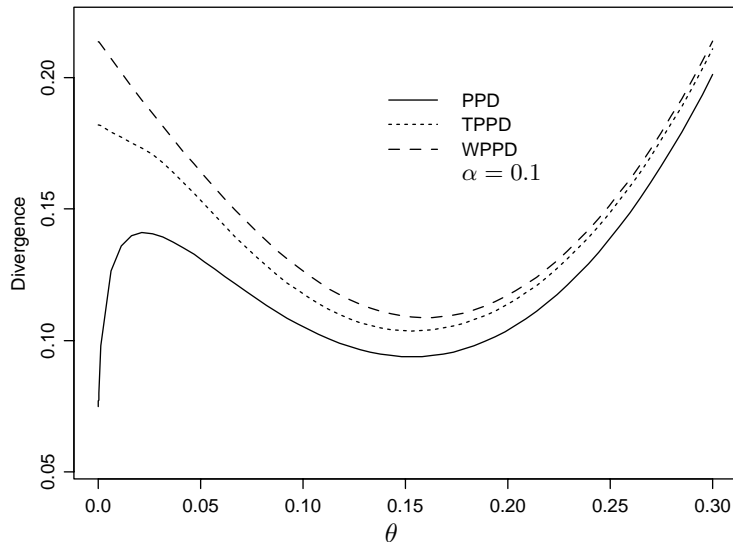


Figure 3: Values of the divergences over  $\theta$  for the Drosophila Data I example.

A second concern was the small sample efficiency of the proposed estimators. Notice that the value of  $C(-1)$  for the  $\text{PPD}_\alpha$  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.*,  $d(x) = 0$ ), 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, Harris and Basu (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  $\text{TPPD}_\alpha$  and  $\text{WPPD}_\alpha$  are obtained by modifying the weight of an empty cell to be equal to that of LD as

$$\begin{aligned} \text{pTPPD}_\alpha(d, f_\theta) &= \frac{1}{2\alpha^2} \sum_{0 < \frac{d}{f_\theta} < (\frac{1-\alpha}{1-2\alpha})^{1/\alpha}} f_\theta(x)^{1-2\alpha} (d(x)^\alpha - f_\theta(x)^\alpha)^2 \\ &\quad + 2(1-2\alpha)^{1/\alpha-2} \sum_{\frac{d}{f_\theta} \geq (\frac{1-\alpha}{1-2\alpha})^{1/\alpha}} d(x) + \sum_{d=0} f_\theta(x) \\ \text{pWPPD}_\alpha(d, f_\theta) &= \frac{1}{2\alpha^2} \sum_{0 < \frac{d}{f_\theta} < (\frac{1-\alpha}{1-2\alpha})^{1/\alpha}} f_\theta(x)^{1-2\alpha} (d(x)^\alpha - f_\theta(x)^\alpha)^2 \\ &\quad + \sum_{\frac{d}{f_\theta} \geq (\frac{1-\alpha}{1-2\alpha})^{1/\alpha}} \left[ \frac{(1-2\alpha)^{1/\alpha-2}}{(1-\alpha)^{1/\alpha-1}} d(x) - \frac{1}{2(1-2\alpha)} f_\theta(x) \right] + \sum_{d=0} f_\theta(x). \end{aligned}$$

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 Figures 4 (a) and (b) we present the histograms of the test statistics for the  $\text{TPPD}_{0.3}$  and  $\text{pTPPD}_{0.3}$  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  $\chi^2(1)$ .

For the same hypotheses and same true distribution, in Figures 5 (a) and (b) we present the probability plots (Wilk and Gnanadesikan 1968) of the quantiles of the ordinary and penalized version of the  $\text{WPPD}_{0.3}$  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  $1/2$  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  $k \geq 12$ . 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.

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  $H_0 : \theta_1 \geq \theta_2$  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

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

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  $t_n^{1/2} \text{sign}(\hat{\theta}_2 - \hat{\theta}_1)$  where  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are the unrestricted minimum divergence estimators of the

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 ( $G^2$ ).

$k$	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
$\alpha$	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
$\alpha$	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
$\alpha$	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
$\alpha$	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

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 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

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

Divergence	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 $_{\alpha}$	$\alpha = 0.1$	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 $_{\alpha}$	$\alpha = 0.1$	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 $_{\alpha}$	$\alpha = 0.1$	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 $_{\alpha}$	$\alpha = 0.1$	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

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.

### 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 WPPD $_{\alpha}$  and 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  $\text{pWPPD}_{0.5}$  estimator is over 95% compared to the MLE. The performance of the  $\text{TPPD}_\alpha$  and  $\text{WPPD}_\alpha$  estimators are very close. For contaminated data, more robust methods (those with smaller values of  $\alpha$ ) start doing better.

Table 5: Estimated biases and mean square errors of the estimators under consideration. 5000 random samples were drawn from Poisson (5) with sample size  $n = 20, 50, 100$ .

$\alpha$	$\text{TPPD}_\alpha$		$\text{pTPPD}_\alpha$		$\text{WPPD}_\alpha$		$\text{pWPPD}_\alpha$	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
<i>Sample Size n = 20</i>								
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						
<i>Sample Size n = 50</i>								
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						
<i>Sample Size n = 100</i>								
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						

For comparison corresponding values for  $\text{WLD}_\lambda$  are presented in Tables 7 and 8 for several values of  $\lambda$ . The efficiencies are now increasing in  $\lambda$  under the model, while 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 equating the Winsorizing point gives  $\lambda = 1 - (\frac{1-2\alpha}{1-\alpha})^{1/\alpha}$ . 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

Table 6: Estimated biases and mean square errors of the estimators under consideration. 5000 random samples were drawn from  $0.9\text{Poisson}(5) + 0.1\text{Poisson}(15)$  with sample size  $n = 20, 50, 100$ .

$\alpha$	TPPD $_{\alpha}$		pTPPD $_{\alpha}$		WPPD $_{\alpha}$		pWPPD $_{\alpha}$	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
<i>Sample Size n = 20</i>								
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						
<i>Sample Size n = 50</i>								
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						
<i>Sample Size n = 100</i>								
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						

better calibration.

We now turn our attention to problems of hypothesis testing. Here again we looked at the TPPD $_{\alpha}$ , WPPD $_{\alpha}$ , WLD $_{\lambda}$  and TLD $_{\lambda}$  families in detail. However, as the results are very similar, we only present the results for the 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-square 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 pWPPD $_{0.5}$  were practically equal to those of LRT.

Table 7: Estimated biases and mean square errors of the  $WLD_\lambda$  estimators. 5000 random samples were drawn from Poisson (5) with sample size  $n = 20, 50, 100$ .

$\lambda$	$n = 20$		$n = 50$		$n = 100$	
	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

Table 8: Estimated biases and mean square errors of the  $WLD_\lambda$  estimators. 5000 random samples were drawn from  $0.9\text{Poisson}(5) + 0.1\text{Poisson}(15)$  with sample size  $n = 20, 50, 100$ .

$\lambda$	$n = 20$		$n = 50$		$n = 100$	
	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

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-square critical values and empirically determined critical values. The results for the nominal level  $\gamma = 0.05$  are presented in Figure 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-squared critical values are used, some of the powers of the ordinary 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 (Figure 6b). However most of these 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-square 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-square critical values leads to results almost identical to the true powers of the tests. While results for other levels of significance and other sample sizes are not reported, they were very similar.

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

$\alpha$	0	0.1	0.2	1/4	0.3	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	$\infty$

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 Figure 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.

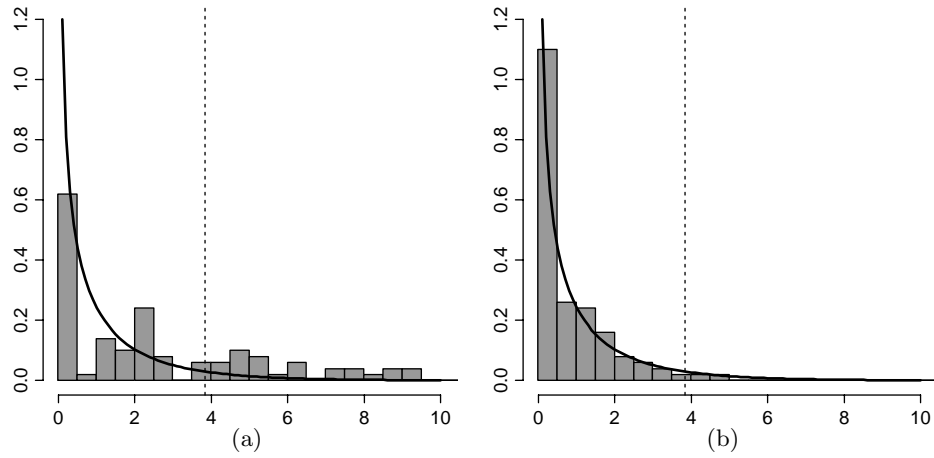


Figure 4: Histogram of the null distribution of  $\text{TPPD}_{0.3}$  and  $\text{pTPPD}_{0.3}$  test statistics. Sample size  $n = 20$  with 100 replications. (a) ordinary test statistic; (b) penalized test statistic.

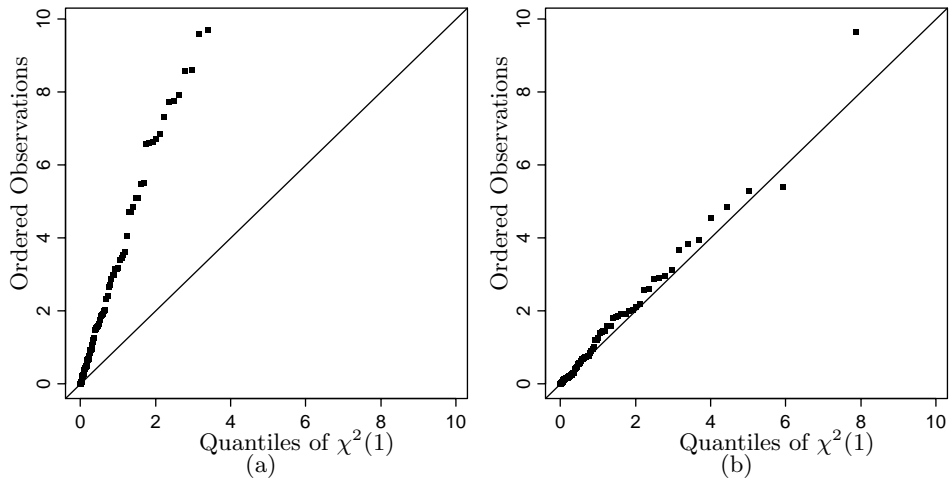


Figure 5:  $\chi^2(1)$  Q-Q plot of  $\text{WPPD}_{0.3}$  and  $\text{pWPPD}_{0.3}$  test statistics. Sample size  $n = 100$  with 100 replications. (a) ordinary test statistic; (b) penalized test statistic.

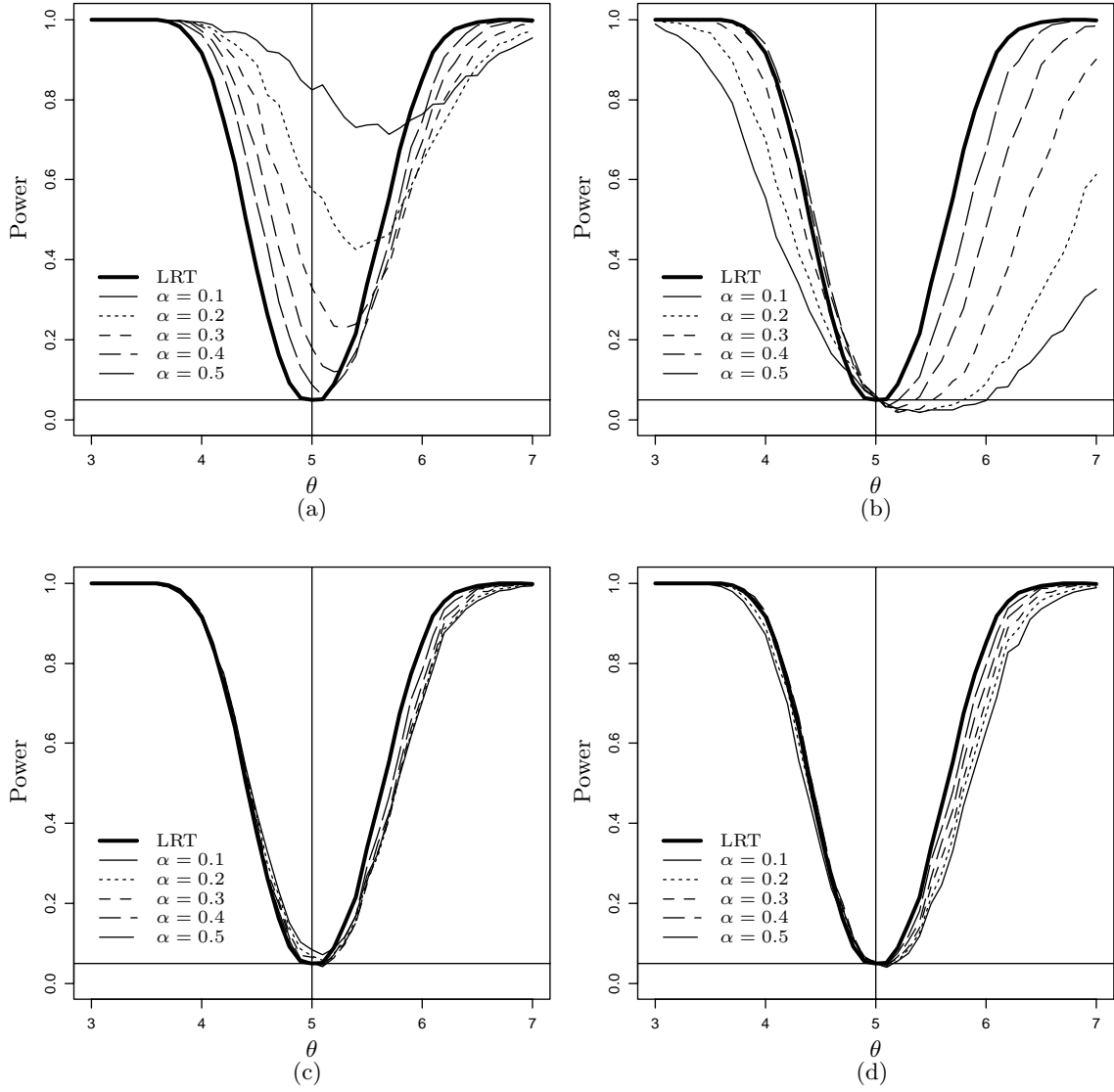


Figure 6: Estimated powers for the tests under consideration testing  $H_0 : \theta = 5$  versus  $H_1 : \theta \neq 5$  with level  $\gamma = 0.05$ . 1,000 random samples were drawn from Poisson( $\theta$ ) with sample size  $n = 50$ . (a) WPPD $_{\alpha}$  based on chi-square critical value; (b) WPPD $_{\alpha}$  based on empirical critical value; (c) pWPPD $_{\alpha}$  based on chi-square critical value; (d) pWPPD $_{\alpha}$  based on empirical critical value.

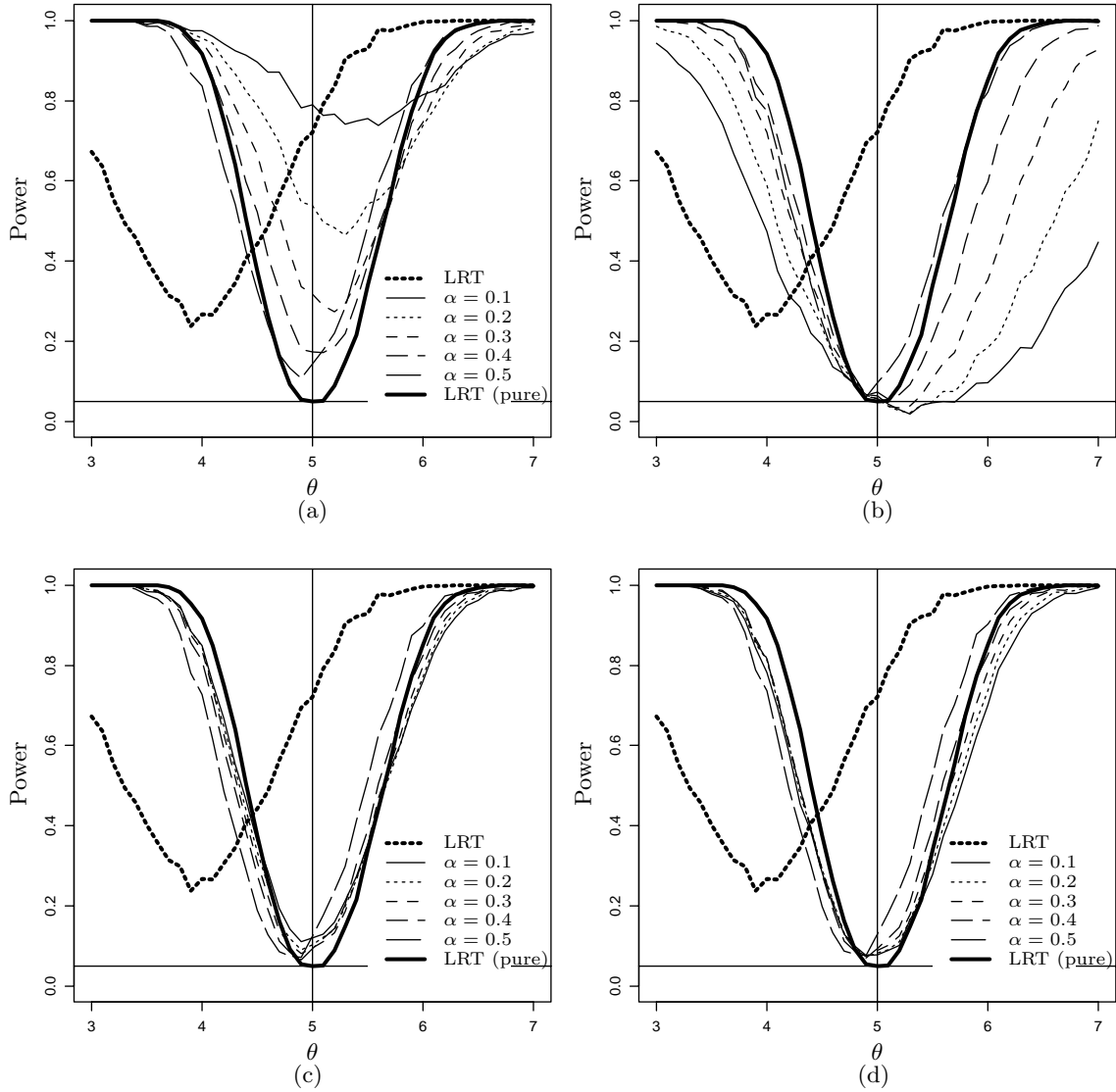


Figure 7: Estimated powers for the tests under consideration testing  $H_0 : \theta = 5$  versus  $H_1 : \theta \neq 5$  with level  $\gamma = 0.05$ . 1,000 random samples were drawn from  $0.9 \text{Poisson}(\theta) + 0.1 \text{Poisson}(15)$  with sample size  $n = 50$ . (a) WPPD $_{\alpha}$  based on chi-square critical value; (b) WPPD $_{\alpha}$  based on empirical critical value; (c) pWPPD $_{\alpha}$  based on chi-square critical value; (d) pWPPD $_{\alpha}$  based on empirical critical value.

## 5. Concluding Remarks

We have shown that modification of the  $\text{PPD}_\alpha$  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  $\text{PPD}_\alpha$  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  $\delta \rightarrow \infty$ . 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  $\delta \rightarrow \infty$ , together with  $A(-1)$  being finite). Notice that these conditions are not satisfied by the  $\text{PPD}_\alpha$  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  $\delta \rightarrow -1$  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.

## *References*

- Basu, A. and Basu, S. (1998). Penalized minimum disparity methods for multinomial models. *Statistica Sinica*, **8**, 841–860.
- Basu, A. and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, **46**, 683–705.
- Basu, A., Harris, I. R., and Basu, S. (1996). Tests of hypotheses in discrete models based on the penalized Hellinger distance. *Statistics and Probability Letters*, **27**, 367–373.
- Basu, A., Harris, I. R., and Basu, S. (1997). Minimum distance estimation: the approach using density based distances. In G. S. Maddala and C. R. Rao, editors, *Handbook of Statistics Vol. 15, Robust Inference*, pages 21–48. Elsevier Science, New York, NY.
- Basu, A., Chakraborty, B., and Sarkar, S. (2000). Robustification of the MLE without loss in efficiency. Unpublished Manuscript.
- Beran, R. J. (1977). Minimum Hellinger distance estimates for parametric models. *Annals of Statistics*, **5**, 445–463.
- Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall.
- Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society B*, **46**, 440–464.
- Harris, I. R. and Basu, A. (1994). Hellinger distance as a penalized log likelihood. *Communications in Statistics: Simulation and Computation*, **23**, 1097–1113.

- Jones, M. C., Hjort, N., Harris, I. R., and Basu, A. (2000). A comparison of related density based minimum divergence estimators. Technical Report ASD/2000/20, Applied Statistics Division, Indian Statistical Institute, Calcutta, India.
- Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Annals of Statistics*, **22**, 1081–1114.
- Park, C., Basu, A., and Lindsay, B. G. (2000). The residual adjustment function and weighted likelihood: A graphical interpretation of robustness of minimum disparity estimators. Technical Report ASD/2000/21, Applied Statistics Division, Indian Statistical Institute, Calcutta, India.
- Rao, C. R. (1961). Asymptotic efficiency and limiting information. In *Proc. Fourth Berkeley Symp.*, volume 1, pages 531–546, Berkeley. University of California Press.
- Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples (with discussion). *Journal of the Royal Statistical Society B*, **24**, 46–72.
- Read, T. R. C. and Cressie, N. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York.
- Simpson, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *Journal of the American Statistical Association*, **82**, 802–807.
- Simpson, D. G. (1989). Hellinger deviance test: efficiency, breakdown points, and examples. *Journal of the American Statistical Association*, **84**, 107–113.
- Stigler, S. M. (1977). Do robust estimators work with real data? *Annals of Statistics*, **5**, 1055–1098.
- Tamura, R. N. and Boos, D. D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, **81**, 223–229.
- Wilk, M. B. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, **55**, 1–17.
- Woodruff, R. C., Mason, J. M., Valencia, R., and 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. *Environmental Mutagenesis*, **6**, 189–202.