RASHI 2 (1) : 39-51 (2017)

Statistical Inference based on the Logarithmic power divergence

[1]Avijit Maji, [2]Saptarshi Chakraborty, [3]Ayanendranath Basu

[1]Reserve Bank of India and Indian Statistical Institute
[2]University of Florida, USA [3]Indian Statistical Institute

ABSTRACT

The power divergence family (PD; Cressie and Read, 1984) and the density power divergence family (DPD; Basu et al., 1998) are two of the most important constituents of the literature on the density-based minimum distance inference. The power divergence family is, arguably, the most prominent member of the class of -divergence family (Csiszár, 1963). Patra et al. (2013) demonstrated that a mathematical link exists between the PD and the DPD families. In this paper we have demonstrated that such links exist more generally between other variants of the DPD family and the generalized-divergences called the (h, ϕ) divergences (Pardo, 2006). In particular we look at the logarithmic power divergence (LPD) family, which may be obtained from a direct translation of the logarithmic density power divergence family (Jones et al., 2001) and explore its role in statistical inference. Several properties of the LPD family, including the asymptotic properties of the corresponding estimators and application of the divergence in performing tests of hypotheses are explored; the breakdown properties of the estimator and the corresponding disparity test statistic are discussed. The LPD is a version of a divergence considered by Renyi (1961) and we demonstrate that tests of parametric hypothesis or tests of goodness-of-fit based on the LPD family show competitive behavior compared to those based on the original PD family. The theory developed is substantiated with numerical studies involving simulated data.

1. Introduction

Minimum distance methods provide a natural technique for parametric statistical inference. Among the dierent types of the minimum distance methods available in the literature, the density-based minimum distance techniques have made a dierence because many members of this class possess full or very high asymptotic efficiency with strong robustness properties. Several authors, including Beran (1977), Tamura and Boos (1986), Simpson (1987), Lindsay (1994), Pardo (2006) and Basu *et al.* (2011) have contributed significantly to this area of research.

Cressie and Read (1984) proposed a generalized class of the density-based divergences and this family is known as the power divergence (PD) family. Although the primary intent of Cressie and Read in introducing this family was testing multinomial goodness-of-fit, the family of the power divergences have been extensively used in the robust minimum distance estimation. The power divergence family is a subclass of the family of ϕ-divergences (Csiszár, 1963).

Basu *et al.* (1998) introduced the family of the density power divergence (DPD). The divergences within this family also show attractive robustness properties. Although the estimators within this class do not have full asymptotic efficiency, several members within this class generate highly robust estimators with only nominal loss in asymptotic efficiency. Both the PD and the DPD families use downweighting based on powers of densities in their robustness scheme.

Recently, Patra *et al.* (2013) examined the mathematical structure of the PD and the DPD families and demonstrated that either family of these divergences can be obtained from the other by simply altering the degree of the density power downweighting. Their findings also indicate that the DPD family is the unique family within a large class of the divergences which allows the estimation of the parameters in the continuous case without any nonparametric density estimation, strengthening the already solid credentials of the DPD family in the parametric inference.

In this paper we have considered other related divergences and explored some other instances of the connection between families of the density-based divergences those originate from the alteration of degree of the density power downweighting. In particular, we will consider the logarithmic power divergence (LPD), which is a member of the (h, ϕ) divergences (Pardo, 2006) and is a version of the Renyi divergence (Renyi, 1961). The divergence was used by Pardo and his associates for testing multinomial goodness-of-fit tests (see Pardo, 2006). We are not aware of the application of this divergence in case of parametric hypothesis testing and we demonstrate here that the hypothesis testing results based on the logarithmic power divergence family are similar and competitive to those based on the ordinary power divergence family. It also lends itself to the improvements that result from the applications of inlier correction techniques (Mandal and Basu, 2013).

The rest of the paper is organized as follows. Section 2 describes the PD and the DPD families and describes the connection between them. Section 3 describes other variants of these divergences and explores similar connections between them. The logarithmic power

E-mail : ayanbasu@isical.ac.in

divergence family, which stands out in this exercise, is then chosen to explore its role in statistical inference. Section 4 discusses the penalized LPD to illustrate the role of inlier modication techniques whereas numerical results invoving simulation are provided in Section 5. Concluding remarks are presented in Section 6.

2. Background: The PD and the DPD Families

2.1 The Power Divergence Family

Let *G* represent the class of all distributions having densities of the appropriate measure. The power divergence (PD) family of Cressie and Read (1984) defines a density-based divergence between two densities *g* and *f*, as a function of a single real tuning parameter $\lambda \in \mathbb{R}$, as

$$\mathrm{PD}_\lambda(g,f) = \frac{1}{\lambda(\lambda+1)} \int g \left[\left(\frac{g}{f} \right)^\lambda - 1 \right] \quad (1)$$

The divergences corresponding to $\lambda = 0$ and $\lambda = -1$ cannot be directly obtained from Equation (1) and they have to be obtained using the continuous limits of the functional form in Equation (1) as $\lambda \to 0$ and $\lambda \to -1$, respectively. These divergences are given by

$$\mathrm{PD}_0(g,f) = \lim_{\lambda \to 0} \mathrm{PD}_\lambda(g,f) = \int g \log \left(\frac{g}{f} \right) \quad (2)$$

$$\mathrm{PD}_{-1}(g,f) = \lim_{\lambda \to -1} \mathrm{PD}_\lambda(g,f) = \int f \log \left(\frac{f}{g} \right) \quad (3)$$

The class of divergences defined by Equation (1) represents a rich class of the density-based divergences and includes several well-known divergences such as the Pearson's Chi-Square (PCS), the likelihood disparity (LD), the Hellinger distance (HD), the Kullback-Liebler divergence (KLD) and the Neyman's Chi-Square (NCS) as special cases, corresponding to $\lambda = 1, 0; -1/2; -1$ and -2, respectively. For the purpose of parametric estimation one replaces the density *f* in Equation (1) with f_θ, a member of a parametric family of densities $\mathcal{F}_\Theta = \{ f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p \}$ and finds the minimum distance estimator of θ as a function of the distribution *G* by minimizing $\mathrm{PD}_\lambda(g; f_\theta)$ over $\theta \in \Theta$, where *g* represents the density function corresponding to the distribution *G*. In actual practice, given a random sample $X_1, ..., X_n$ from the true data generating distribution *G*, one constructs a nonparametric density estimate $\hat{g}$ of *g* and minimizes $\mathrm{PD}_\lambda(\hat{g}; f_\theta)$ over $\theta \in \Theta$ to find the minimum divergence estimator of θ.

In particular when the data are generated from a discrete model, the vector of the relative frequencies represent the canonical choice for $\hat{g}$. Without loss of generality let the support of the random variable be $\chi = \{0; 1, ... \}$. Let $d_n(\chi)$ be the relative frequency of the value χ based on the sample. Then the minimum power divergence estimator $\hat{\theta}$ may be dened by the relation $\hat{\theta} = \arg\min_{\theta \in \Theta} \mathrm{PD}_\lambda(d_n ; f_\theta)$, where

$$\mathrm{PD}_\lambda(d_n, f_\theta) = \frac{1}{\lambda(\lambda+1)} \sum_{x=0}^{\infty} d_n(x) \left[\left(\frac{d_n(x)}{f_\theta(x)} \right)^\lambda - 1 \right]$$

It can be easily seen that the estimator $\hat{\theta}$ is a function of λ also, which we have suppressed for brevity.

2.2 The Density Power Divergence (DPD) Family

The minimum distance inference method described in the previous section for the discrete models can easily be extended to the continuous models also. However, for continuous models one has to use some nonparametric smoothing technique such as kernel density estimation to construct an estimate of the true density (see, e.g., Basu *et al.*, 2011). The estimation method based on minimizing the PD family inherits all the associated complications of kernel density estimation including and the slower rate of convergence of the kernel density estimate in higher dimensions. To overcome this issue, Basu *et al.* (1998) proposed the density power divergence (DPD) which successfully estimates the true density without using any kernel density technique. Given densities *g* and *f* for distributions *G* and *F* respectively, the density power divergence is defined in terms of a nonnegative tuning parameter $\alpha \geq 0$, as

$$\mathrm{DPD}_\alpha(g,f) = \int \left[f^{1+\alpha} - \left(1 + \frac{1}{\alpha} \right) f^\alpha g + \frac{1}{\alpha} g^{1+\alpha} \right] \quad (4)$$

This divergence is not directly defined for $\alpha = 0$ and needs to be obtained as the continuous limit of the above functional form as $\alpha \to 0$. This generates

$$\mathrm{DPD}_0(g,f) = \lim_{\alpha \to 0} \mathrm{DPD}_\alpha(g,f) = \int g \log \left(\frac{g}{f} \right) \quad (5)$$

which, incidentally, is identical to the divergence $\mathrm{PD}_0(g; f)$ given in Equation (2). This is the only divergence which is common to both the PD and the DPD families.

Patra *et al*. (2013) pointed out an interesting connection between these two families. We can express the power divergence between two densities *g* and *f* as

$$\mathrm{PD}_\lambda(g,f)=\int\left\{\frac{1}{\lambda(1+\lambda)}\left[\left(\frac{g}{f}\right)^{1+\lambda}-\left(\frac{g}{f}\right)\right]+\frac{1-g/f}{1+\lambda}\right\}f. \tag{6}$$

This is just a rewriting of Equation (1) and does not change the integral. In addition, the form given in Equation (6) makes the integrand (and not just the integral) non-negative. Suppose we wish to modify this divergence so that the modified form preserves the divergence properties and the corresponding minimum divergence estimator avoids nonparametric density estimation. To do this we need to eliminate the terms that contain a product of a nonlinear function of *g* with some function of *f* and for this we only need to adjust the term $(g/f)^{1+\lambda}$ in Equation (6). As the expression within the curly braces is nonnegative and equals zero only if $g = f$, we can replace the outer *f* term in Equation (6) by $f^{1+\lambda}$ and still get a valid divergence that simplies to

$$\begin{aligned}&\int\left\{\frac{\left[g^{1+\lambda}-gf^{\lambda}\right]}{\lambda(1+\lambda)}+\frac{f^{1+\lambda}-gf^{\lambda}}{1+\lambda}\right\}\\&=\frac{1}{1+\lambda}\int\left\{\frac{1}{\lambda}\left[g^{1+\lambda}-gf^{\lambda}\right]+f^{1+\lambda}-gf^{\lambda}\right\}\\&=\frac{1}{1+\lambda}\int\left\{f^{1+\lambda}-\left(1+\frac{1}{\lambda}\right)gf^{\lambda}+\frac{1}{\lambda}g^{1+\lambda}\right\}\end{aligned} \tag{7}$$

which is a scaled version of the divergence given in Equation (4) for $\lambda = \alpha$. Interestingly, the above operation generates strongly robust divergences starting from divergences that are highly non-robust and vice versa. For example, starting from the Pearson's chi-square divergence we derive the L_2-distance which is highly robust.

We can also reverse the above transformation to get the power divergence family from the density power divergence family by replacing the outer $f^{1+\alpha}$ term in

$$\mathrm{DPD}_\alpha(g,f)=\int\left\{1-\left(1+\frac{1}{\alpha}\right)\frac{g}{f}+\frac{1}{\alpha}\left(\frac{g}{f}\right)^{1+\alpha}\right\}f^{1+\alpha} \tag{8}$$

with *f*. After simplication, the divergence is easily seen to be equal to a scaled version of the PD_α family.

Patra *et al.* (2013) actually considered a more general class of divergences given by

$$\rho(g,f)=\int\eta(\delta+1)f^{\beta}dx \tag{9}$$

where $\beta > 1$, δ is the Pearson residual $\frac{g}{f}-1$ and $\eta(y) = \sum_{t\in T}a_t y^t$ for some finite set *T* with elements in *R* and real coefficients $\{a_t\}$ such that $\eta(.)$ is nonnegative on $[0, \infty)$ and $\eta(y) = 0$ only when $y = 1$. They found that the restrictions necessary for $\rho(g, f)$ to be a genuine statistical divergence as well as those necessary for avoiding nonparametric smoothing for the purpose of the estimation for continuous models; the authors demonstrated that this leads to the DPD family with parameter $\beta - 1$ as the unique solution.

2.3 (*h*, ϕ)-Divergence Family

Csiszár (1963) and Ali and Silvey (1966) provided a general class of divergences between two densities. Given densities *g* and *f*, this class is defined by

$$D_\phi(g,f)=\int\phi\left(\frac{g}{f}\right)f, \tag{10}$$

where $\phi(.)$ is a convex function such that $\phi(1) = 0$. The class of the power divergence family is a subclass of ϕ-divergences. Menéndez *et al*. (1995) described the (h, ϕ)-divergence between two densities *g* and *f* as

$$D_\phi^h(g,f)=h(D_\phi(g,f)), \tag{11}$$

where *h* is a real, increasing, differentiable function on the range of the ϕ divergence. Pardo (2006) provides an useful list of specic ϕ and (h, ϕ) divergences.

3 The Logarithmic Power Divergence and the Logarithmic Density Power Divergence Families

Jones *et al.* (2001) considered several variants of the DPD family, which allow for other forms of the density power downweighting without requiring any nonparametric smoothing. A class of divergences of this type is the logarithmic density power divergence (LDPD) family given by

$$\mathrm{LDPD}(g,f)=\log\int f^{1+\alpha}-\left(1+\frac{1}{\alpha}\right)\log\int f^{\alpha}g+\frac{1}{\alpha}\log\int g^{1+\alpha}, \tag{12}$$

where LDPD stands for the logarithmic density power divergence. This form has striking similarities with the DPD family, and may be considered to be a modication of the latter where the identity function is replaced by the logarithmic function. In the spirit of the connection between the PD and the DPD described in Equation (7), the LDPD family is also seen to be similarly connected in the same manner to an (h, ϕ) divergence; the h and ϕ functions are defined later in Section 3.2. In particular the LDPD family can be written as

$$\begin{aligned} \mathrm{LDPD}(g,f) &= \log\int f^{1+\alpha} - \left(1+\frac{1}{\alpha}\right)\log\int\left(\frac{g}{f}\right)f^{1+\alpha} \\ &+ \frac{1}{\alpha}\log\int\left(\frac{g}{f}\right)^{1+\alpha}f^{1+\alpha}. \end{aligned} \tag{13}$$

Replacing each $f^{1+\alpha}$ term with f in Equation (13) leads to the divergence

$$\frac{1}{\alpha}\log\int\frac{g^{1+\alpha}}{f^{\alpha}}. \tag{14}$$

We standardize this further to express this divergence as $\frac{1}{\alpha(\alpha+1)}\log f\frac{g^{1+\alpha}}{f^{\alpha}}$. In this form the divergence generates the divergences $\int g\log\left(\frac{g}{f}\right)$ and $\int f\log\left(\frac{f}{g}\right)$ as $\alpha\to 0$ and $\alpha\to -1$, respectively; these limiting divergences are the same as the corresponding limiting divergences in case of the PD family, as given in Equation (1). The family of divergences in Equation (14) will be called the logarithmic power divergences (LPD); using a different symbol for the tuning parameter, this divergence has the form

$$\mathrm{LPD}_{\beta}(g,f) = \frac{1}{\beta(\beta+1)}\log\int\frac{g^{1+\beta}}{f^{\beta}}, \beta\in\mathbb{R}. \tag{15}$$

3.1 The LPD as a divergence

The LPD has known forms for $\beta = 0$ or -1 (in the limiting sense). For other values of β we have the following theorem.

Theorem 1. *The* LPD *is a divergence for* $\beta\in\mathbb{R}\setminus\{0,-1\}$

Proof. Note that,

$$\begin{aligned} \mathrm{LPD}_{\beta}(g,f) &= \frac{1}{\beta(\beta+1)}\log\int\frac{g^{1+\beta}}{f^{\beta}} \\ &= \frac{1}{\beta(\beta+1)}\log\left[E_f\left(\frac{g^{1+\beta}}{f^{1+\beta}}\right)\right] \end{aligned}$$

Now let us consider the function $k(\chi) = \chi^{1+\beta}$, where $\chi > 0$. Then, $k'(\chi) = (1+\beta)\chi^{\beta}$ and $k''(\chi) = \beta(1+\beta)\chi^{\beta-1}$. So when $\beta > 0$ or $\beta < -1$, *i.e.*, $\beta(1+\beta) > 0$, $k''(\chi) > 0$, implying $k(\chi)$ is strictly convex, and when $-1 < \beta < 0$, *i.e.*, $\beta(1+\beta) < 0$, $k(\chi)$ is strictly concave. Therefore, when $\beta(1+\beta) > 0$, using Jensen's inequality we get,

$$\log\left[E_f\left(\frac{g^{1+\beta}}{f^{1+\beta}}\right)\right] \geq \log\left[\left\{E_f\left(\frac{g}{f}\right)\right\}^{1+\beta}\right] = 0.$$

On the other hand when $\beta(1+\beta) < 0$, Jensen's inequality gives

$$\log\left[E_f\left(\frac{g^{1+\beta}}{f^{1+\beta}}\right)\right] \leq \log\left[\left\{E_f\left(\frac{g}{f}\right)\right\}^{1+\beta}\right] = 0.$$

In either of the above two cases, the inequality becomes an equality if and only if $g \equiv f$, identically. Therefore, in either case we have,

$$\mathrm{LPD}_{\beta}(g,f) = \frac{1}{\beta(1+\beta)}\log\int\frac{g^{1+\beta}}{f^{\beta}} \geq 0,$$

with equality if and only if $g \equiv f$. This completes our proof.

3.2 The LPD as a (h, ϕ)-divergence

Pardo (2006) described the (h, ϕ)-divergence as

$$D_{\phi}^{h}(g,f) = h\left(\int \phi\left(\frac{g}{f}\right) f\right) \tag{16}$$

where h and ϕ have the usual properties. It can be easily shown that the PD family is a member of the ϕ-divergence family (for details see Basu *et al*., 2011). Now, the LPD family can be written as

$$\begin{aligned} \mathrm{LPD}_{\beta}(g,f) &= \frac{1}{\beta(1+\beta)} \log\left(\int \frac{g^{1+\beta}}{f^{\beta}}\right) \\ &= \frac{1}{\beta(1+\beta)} \log\left[\beta(1+\beta)\mathrm{PD}_{\beta}(g,f)+1\right] \end{aligned}$$

Now if we denote $\mathrm{PD}_{\beta}(g, f)$ as y, then $\mathrm{LPD}_{\beta}(g, f)$ can be denoted as $\Psi(y)$, where

$$\psi(y) = \frac{\log(\beta(1+\beta)y+1)}{\beta(1+\beta)}.$$

It is easy to check that $\Psi(y)$ is an increasing function of y. Thus the LPD defines a genuine (h, ϕ) divergence in the sense of Equation (16).

3.3 The Minimum LPD estimator and its Estimating Equation

The minimum LPD estimator $\hat{\theta}$ of θ at a density g is defined by the relation

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \mathrm{LPD}_{\beta}(g, f_{\theta}).$$

The estimating equation of the minimum LPD estimator at a density g is given by

$$-\nabla \mathrm{LPD}_{\beta}(g, f_{\theta}) = 0,$$

where ∇ represents derivative with respect to θ. The estimating equation has the form

$$\begin{aligned} &\frac{1}{\beta(1+\beta)} \left(\int \frac{g^{1+\beta}}{f_{\theta}^{\beta}}\right)^{-1} \nabla\left(\int \frac{g^{1+\beta}}{f_{\theta}^{\beta}}\right) = 0 \\ &\Rightarrow \frac{1}{\beta(1+\beta)} \nabla\left(\int \frac{g^{1+\beta}}{f_{\theta}^{\beta}}\right) = 0. \end{aligned} \tag{17}$$

Following the same terminology we can write the PD family as

$$\mathrm{PD}_{\beta}(g, f_{\theta}) = \frac{1}{\beta(1+\beta)} \left(\int \frac{g^{1+\beta}}{f_{\theta}^{\beta}} - 1\right) \tag{18}$$

The corresponding estimating equation will be

$$-\nabla \mathrm{PD}_{\beta}(g, f_{\theta}) = 0,$$

which gives

$$\frac{1}{\beta(1+\beta)} \nabla\left(\int \frac{g^{1+\beta}}{f_{\theta}^{\beta}}\right) = 0. \tag{19}$$

It is clear that Equations (17) and (19) are identical equations. This implies that though the PD and the LPD families have different functional forms, they have identical estimating equations which implies both the minimum LPD and the minimum PD estimators are the same. This is expected, since the LPD is an increasing function of the PD.

3.4 Asymptotic Distribution of the minimum LPD estimator

As shown in the Section 3.3, the minimum PD estimator and the minimum LPD estimators are the same; consequently they will have the same asymptotic distribution. Under some regularity conditions, there exists a consistent sequence θ_n of roots to the minimum LPD estimating equation and the asymptotic distribution of $n^{1/2}(\theta_n - \theta_g)$ is multivariate normal with mean vector 0 and covariance matrix $J_g^{-1}V_gJ_g^{-1}$ where θ^g is the best fitting parameter as defined in Section 2.3, Basu *et al.* (2011), and J_g and V_g are as defined in Basu *et al.* (2011, Theorem 2.19).

3.5 The Minimum LPD Estimator : The Discrete Model

Let $X_1, X_2, \ldots\ldots X_n$ be n independently and identically distributed observations from a discrete population G, modeled by the parametric family

$F_{\theta} = \{f_{\theta} : \theta \in \Theta \subseteq \mathbb{R}^p\}$. Without loss of generality, we can assume that the support of the distribution G is $\chi = \{0, 1, 2, \ldots\ldots\}$. Let $d_n(\chi)$ be the relative frequency of the value χ in the random sample. Then the minimum LPD estimator $\hat{\theta}$ of θ is defined by the relation

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \mathrm{LPD}_{\beta}(d_n, f_{\theta}).$$

3.6 Using the LPD in Testing of Hypothesis

Using a discrete model and the parametric set up of Section 2.1, consider the null hypothesis

$$H_0 : \theta \in \Theta_0 \tag{20}$$

where Θ_0 is a proper subset of the parameter space Θ. Let r be the number of independent restrictions imposed by the null hypothesis. The deviance test statistic based

on the LPD is defined by

$$D_{\mathrm{LPD}_\beta} = 2n(\mathrm{LPD}_\beta(d_n, f_{\hat{\theta}_0}) - \mathrm{LPD}_\beta(d_n, f_{\hat{\theta}})), \tag{21}$$

where $\hat{\theta}_0$ and $\hat{\theta}$ are the estimates of under the null hypothesis and without any restrictions respectively. Although the minimum PD and LPD estimators are identical, the same does not hold for the deviance statistic based on the PD and the LPD. We will now find the asymptotic null distribution of the test statistic given in Equation (21). Though the PD and the LPD deviance statistics are not themselves identical, they are asymptotically equivalent and we will derive the asymptotic null distribution of the deviance test statistic based on the LPD by demonstrating the same. The result corresponding to the PD is derived, for example, in Basu *et al.* (2011). The LPD can be written as

$$\begin{aligned}\mathrm{LPD}_\beta(d_n, f_\theta) &= \frac{1}{\beta(1+\beta)} \log\left(\sum \frac{d_n^{1+\beta}}{f_\theta^\beta}\right) \\ &= \frac{1}{\beta(1+\beta)} \log\left[\beta(1+\beta)\mathrm{PD}_\beta(d_n, f_\theta) + 1\right].\end{aligned}$$

The PD deviance statistic is defined by

$$D_{\mathrm{PD}_\beta}(d_n, f_\theta) = 2n(\mathrm{PD}_\beta(d_n, f_{\hat{\theta}_0}) - \mathrm{PD}_\beta(d_n, f_{\hat{\theta}})], \tag{22}$$

which can be written as $2n\left(y_1^{(n)} - y_2^{(n)}\right)$, where $y_1^{(n)} = \mathrm{PD}_\beta(d_n, f_{\hat{\theta}_0})$ and $y_2^{(n)} = \mathrm{PD}_\beta(d_n, f_{\hat{\theta}})$, and the corresponding statistic in Equation (21) will be of the form $2n\left(\psi\left(y_1^{(n)}\right) - \psi\left(y_2^{(n)}\right)\right)$, where $\psi(y) = \frac{\log(\beta(1+\beta)y+1)}{\beta(1+\beta)}$. Now, using Taylor's expansion, we get,

$$\psi\left(y_1^{(n)}\right) - \psi\left(y_2^{(n)}\right) = \left(y_1^{(n)} - y_2^{(n)}\right)\psi'\left(y_2^{(n)}\right) + \frac{1}{2}\left(y_1^{(n)} - y_2^{(n)}\right)^2 \psi''(\xi),$$

for some ξ between $y_1^{(n)}$ and $y_2^{(n)}$. Now $\psi'(y) = \frac{1}{\beta(\beta+1)y+1}$ and $\psi''(y) = -\frac{\beta(\beta+1)}{(\beta(\beta+1)y+1)^2}$. Under the null hypothesis, as $n \to \infty$, $\psi'\left(y_2^{(n)}\right) \to \psi'(0) = 1$ and for any fixed β, $\psi''(y)$ is a bounded finite term since $y \geq 0$. Thus

$$\begin{aligned}2n\left(\psi\left(y_1^{(n)}\right) - \psi\left(y_2^{(n)}\right)\right) &= 2n\left(y_1^{(n)} - y_2^{(n)}\right)\psi'\left(y_2^{(n)}\right) \\ &\quad + n \times \left(y_1^{(n)} - y_2^{(n)}\right) \times a_n,\end{aligned} \tag{23}$$

where $a_n = \left(y_1^{(n)} - y_2^{(n)}\right) \times \psi''(\xi)$. It follows from Basu et al. (2011) that

$$2n\left(y_1^{(n)} - y_2^{(n)}\right) \xrightarrow{d} W \sim \chi_r^2. \tag{24}$$

Hence $n\left(y_1^{(n)} - y_2^{(n)}\right) = O_p(1)$, and thus $\left(y_1^{(n)} - y_2^{(n)}\right) \xrightarrow{p} 0$. Replacing these results in Equation (23) gives us

$$2n(\psi(y_1^{(n)}) - \psi(y_2^{(n)})) = 2n(y_1^{(n)} - y_2^{(n)}) + o_p(1). \quad (25)$$

Hence

$$D_{\mathrm{LPD}_\beta}(d_n, f_\theta) \xrightarrow{d} W \sim \chi^2_r. \quad (26)$$

Thus the null distribution of the deviance statistic based on the LPD follows the limiting χ^2 distribution.

3.6.1 Testing of hypothesis using the Rao and Wald Statistics

The test statistic in Equation (21) may be considered to be the analogue of the likelihood ratio test. One can also perform tests of hypothesis for the null hypothesis in Equation (20) using the Rao and the Wald statistics based on the LPD. TheWald statistic (Wald, 1943) for the null hypothesis in Equation (20) has the form

$$W = n\left(\hat{\theta} - \hat{\theta}_0\right)^T I\left(\hat{\theta}\right)\left(\hat{\theta} - \hat{\theta}_0\right) \quad (27)$$

which has the same asymptotic null distribution as the deviance statistic defined in Equation (21). Here $\hat{\theta}$ and $\hat{\theta}_0$ are the unrestricted minimum LPD estimators of θ, and the estimator under the null, respectively.

Under the discrete set-up, the Rao test (Rao, 1973) for the simple null hypothesis can be performed through the statistic

$$S = n\varrho_n^T(\theta_0) I^{-1}(\theta_0) \varrho_n(\theta_0), \quad (28)$$

where

$$\varrho_n(\theta) = -n^{\frac{1}{2}} \nabla \zeta_C(d_n, f_\theta)$$

and $\varsigma_C(d_n, f_\theta)$ is any divergence within the LPD class.

The Wald statistic involves the divergence only through the parametric estimate and for that reason both the PD and the LPD families will generate the same Wald Statistic. The derivation in Section 3.3 shows that the Rao statistics for the two cases are asymptotically equivalent under the null hypothesis.

3.7 Multinomial Goodness-of-Fit Testing using the LPD

Multinomial goodness-of-fit testing is one of the oldest classical techniques in statistics and date goes back to more than a century (Pearson, 1900). Several books including Bishop, Fienberg and Holland (1975), Fienberg (1980), Agresti (1984), Freeman (1987), Read and Cressie (1988) and more recently, Pardo (2006) and Basu *et al.* (2011) consider different aspects of this problem. The LPD can be used for the multinomial goodness-of-fit testing problem as well. Suppose we have a k-cell multinomial with probability vector $\pi = (\pi_1, \pi_2, ... \pi_k)$. Based on a multinomial sample which leads to a frequency of n_i in the i-th cell, $i = 1, ...k$, $\sum n_i = n$, we are interested in testing the null hypothesis

$$H_0 : \pi = \pi(\theta)$$

for some unknown θ which is an s-dimensional parameter ($s < k-1$) taking values in the set Θ. Then the goodness-of-fit statistic based on the LPD divergence is given by

$$T = 2n \min_{\theta \in \Theta} \mathrm{LPD}(\hat{p}, \pi(\theta)) = 2n\mathrm{LPD}(\hat{p}, \pi(\hat{\theta})),$$

where $\hat{p} = (\hat{p}_1 ....., \hat{p}_k)$, $p_i = \frac{n_i}{n}$ being the observed proportion of the i-th cell, $i = 1, ..., k$ and $\hat{\theta}$ is the minimum LPD estimator of θ defined by $\min_{\theta \in \Theta}$ LPD $\left(\hat{p}, \pi(\theta)\right)$ = LPD $\left(\hat{p}, \pi\left(\hat{\theta}\right)\right)$. Under H_0 the LPD test statistic has an asymptotic χ^2_{k-s-1} distribution (Pardo, 2006). When the null fully specifies the probability vector, the degrees of freedom of the χ^2 distribution is $(k-1)$.

3.8 Breakdown Properties

3.8.1 Breakdown point of the minimum LPD Estimator

As discussed in the Section 3.3, both the minimum LPD estimator and the minimum PD estimator are the same and so we can conclude that both have identical breakdown properties. See Basu *et al.* (2011) for a discussion of the breakdown properties of the minimum PD estimator.

3.8.2 Power Breakdown of the LPD based deviance statistic

We have already observed that while the minimum PD and minimum LPD estimators are identical, the test statistics for testing parametric hypothesis based on the PD and LPD are not so, although they are asymptotically equivalent under the null. Here we will briefly study the power influence function of the LPD based deviance tests. Consider the hypothesis and the set-up of Section 3.6. We will follow the approach of Simpson (1989) to determine the power breakdown point of the disparity difference tests. Consider the functional

$$\phi(G) = \mathrm{LPD}(g, f_{\hat{\theta}_0(g)}) - \mathrm{LPD}(g, f_{\hat{\theta}(g)}) \tag{29}$$

as a function of the distribution point G (having density g), where

$$\hat{\theta}_0(g) = \arg\min_{\theta\in\Theta_0} \mathrm{LPD}(g, f_\theta) \text{ and } \hat{\theta}(g) = \arg\min_{\theta\in\Theta} \mathrm{LPD}(g, f_\theta).$$

Let

$$\epsilon(G;t) = \inf\left\{\epsilon : \inf_{V\in\mathcal{G}} t[(1-\epsilon)G + \epsilon V] = t_{\min}\right\},$$

where t is a generic functional, $G \in g$, V is a contaminating distribution and $t_{\min} = \inf_{F\in G} t(\mathrm{F})$. From the functional ϕ as defined in (29), we will refer $\in(G; \phi)$ as the power-breakdown point, since for a contamination proportion below $\in(G; \phi)$ the deviance test will be consistent, so that it will eventually reject a false hypothesis and so asymptotically there will be no power breakdown. For the true distribution G and the contaminating distribution V, let $u = (1- \in)g + \in v$ represent the density of the contaminated distribution U, where v represent the contaminating density and u is the density of U, the mixture of the target and the contamination. To find the power breakdown at G one needs to analyze the behaviour of the quantity

$$\phi(U) = \mathrm{LPD}(u, f_{\hat{\theta}_0(u)}) - \mathrm{LPD}(u, f_{\hat{\theta}(u)}). \tag{30}$$

Following Simpson (1989), we will determine a lower bound for LPD $\left(u, f_{\hat{\theta}_0(u)}\right)$ and an upper bound for LPD $\left(u, f_{\hat{\theta}_0(u)}\right)$ for all possible distributions $V \in G$. It is clear that if the lower bound derived is strictly greater than the upper bound then we can conclude that power breakdown will not occur.

The LPD(g; f_θ) family has the form

$$\mathrm{LPD}(g, f_\theta) = \frac{1}{\beta(\beta+1)} \log \int \left(g^{1+\beta} f_\theta^{-\beta}\right).$$

Table 1: Bound for Power Breakdown of the LPD Disparity Tests

$\beta\downarrow$ $0_1 \rightarrow$	4	8
– 0:25	0.1942	0.4160
– 0:50	0.1341	0.3974
– 0:75	0.0066	0.1566

We then obtain the necessary upper and lower bounds as

$$\mathrm{LPD}(u, f_{\hat{\theta}(u)}) \leq \mathrm{LPD}(u, f_{\hat{\theta}(g)}) \leq \frac{1}{\beta(\beta+1)} \log \int \left[(1-\epsilon)^{1+\beta} g^{1+\beta} f_{\hat{\theta}(g)}^{-\beta}\right]$$

and

$$\begin{aligned} \mathrm{LPD}(u, f_{\hat{\theta}_0(u)}) &\geq \mathrm{LPD}(u, f_{\hat{\theta}_0(g)}) \\ &\geq \frac{1}{\beta(\beta+1)} \log \int \left[(1-\epsilon)^{1+\beta} g^{1+\beta} f_{\hat{\theta}_0(g)}^{-\beta} + \epsilon^{1+\beta}\right]. \end{aligned}$$

After some routine simplifications based on the bounds, we see that break-down cannot occur, *i.e.* $\phi(U)$ in Equation (30) is greater than zero, for

$$\epsilon < \left[1 + \left(\int g^{1+\beta} f_{\hat{\theta}(g)}^{-\beta} - \int g^{1+\beta} f_{\hat{\theta}_0(g)}^{-\beta}\right)^{-1/(1+\beta)}\right]^{-1} \tag{31}$$

Example: Let f_θ denote the Poisson density with mean. We want to test the hypothesis

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta \neq \theta_0$$

Table 1 gives the results for the breakdown bound in Equation (31) for several combinations of β and θ, when $\theta_0 = 1$. The true density g is the Poisson density f_{θ_1}. The results corresponding to $\beta = -1/2$ match with the results of Simpson (1989).

4 Inlier Modication

One of the main attractive features of the minimum divergence procedures such as those based on the PD and the LPD families is that many members within these families have automatic robustness properties under data contamination. Lindsay (1994), Basu and Lindsay (1994) and several later authors have characterized the outliers probabilistically in this context, rather than geometrically, although the concepts often coincide. Formally, given the observed data density vector d_n and the model density f_θ, let $\delta(\chi) = \frac{d_n(\chi)}{f_\theta(\chi)} - 1$ be a standardized residual. An observation x will be called an outlier if $\delta(\chi)$ is a large positive value. Clearly the robust minimum divergence procedures will have to strongly downweight the large δ observations to keep the impact of these observations under control. On the other hand, many of the robust minimum divergence methods (including those within the PD and the LPD families) provide an improper treatment of inliers, which are cells that have less data than are predicted by the model. Many of these divergences put unduly large weights on the inliers which appear to adversely affect their small sample efficiencies. To overcome this problem several inlier modification techniques are available in the literature. The most commonly used method in this context is the technique of penalized disparities (see Harris and Basu, 1994; Basu *et al.*, 1996; Basu and Basu, 1998; Park *et al.*, 2001; Basu *et al.*, 2002; Pardo and Pardo, 2003; Alin, 2007; Basu *et al.*, 2010). Here we will discuss use of the penalized disparity in controlling inlier problem in the LPDs.

4.1 The Penalized LPD

Following the approach of Basu *et al.* (1996), we will establish the form of the penalized version of the LPD. The LPD has the form

$$\mathrm{LPD}_\beta(d_n, f) = \frac{1}{\beta(1+\beta)} \log\left(\sum_x \frac{d_n^{1+\beta}}{f^\beta}\right) \tag{32}$$

Now for $\delta = \frac{d_n}{f} - 1$ we get

$$\begin{aligned}
\sum_x \frac{d_n^{1+\beta}}{f^\beta} &= \sum_x \frac{d_n^{1+\beta}}{f^{1+\beta}} f \\
&= \sum_x \left[\{(\delta+1)^{1+\beta} - (\delta+1)\} - \beta\delta\right] f + 1 \\
&= \sum_x [G^*(\delta)]\, f + 1 \\
&= \sum_{x:\delta(x)>-1} [G^*(\delta)]\, f + \sum_{x:\delta(x)=-1} [G^*(\delta)]\, f + 1 \\
&= \sum_{x:\delta(x)>-1} [G^*(\delta)]\, f + G(-1) \sum_{x:\delta(x)=-1} f + 1,
\end{aligned}$$

where $G^*(\delta) = (1 + \delta)^{1+\beta} - (1 + \delta) - \beta\delta$. Now to control the inlier we need to modify the coefficient of f when $\delta(\chi) = -1$. For deffining the form of the penalized $\text{LPD}_\beta(d_n, f)$ we will consider two cases.

Case 1 ($\beta < -1$ or $\beta > 0$) : We would like to maintain the non-negativity of each of the terms while changing the coefficient of the second term and that can be done easily just by a positive constant (k, say). Therefore, the penalized $\text{LPD}_\beta(d_n, f)$ is defined as

$$\text{PLPD}_{\beta,k}(d_n, f) = \frac{1}{\beta(1+\beta)} \log \left(\sum_{x:\delta(x)>-1} G^*(\delta) f + k \sum_{x:\delta(x)=-1} f + 1 \right), \quad (33)$$

for $k > 0$.

Case 2 $(-1 < \beta < 0)$: The penalized LPD will take the form

$$\text{PLPD}_{\beta,k}(d_n, f) = \frac{1}{\beta(1+\beta)} \log \left(\sum_{x:\delta(x)>-1} G^*(\delta) f + k \sum_{x:\delta(x)=-1} f + 1 \right), \quad (34)$$

for $k > \beta$ to maintain the non-negativity of each terms.

4.1.1 Estimating Equation

The estimating equation of the penalized LPD would be

$$\frac{1}{\beta(1+\beta)} \left[\nabla \left(\sum_{x:\delta(x)>-1} G^*(\delta) f \right) + k \sum_{x:\delta(x)=-1} \nabla f \right] = 0. \quad (33)$$

Following Mandal and Basu (2013), the penalized PD has the form

$$\text{PPD}_{\beta,\mu}(d_n, f) = \frac{1}{\beta(1+\beta)} \left(\sum_{x:\delta(x)>-1} G^*(\delta) f \right) + \mu \sum_{x:\delta(x)=-1} f \quad (34)$$

which has the following estimating equation

$$\frac{1}{\beta(1+\beta)} \nabla \left(\sum_{x:\delta(x)>-1} G^*(\delta) f \right) + \mu \nabla \sum_{x:\delta(x)=-1} f = 0. \quad (37)$$

Comparing (35) and (37), we can say that both equations are identical when $\mu = \frac{k}{\beta(1+\beta)}$

4.2 Goodness-of-Fit using the inlier modified LPD

Consider the goodness-of-fit testing problem as a k-cell multinomial for the equi-probable null

$$H_0 : \pi_i = \frac{1}{k}, \quad \forall_i = 1(1)k$$

Table 2: Comparison of the observed levels of the testing methods based on the two families at nominal level 0:05

Method	∈	*n*	PHD	HD	PLHD	LHD
		20	0.0474	0.1068	0.0464	0.1104
	0	50	0.05	0.0846	0.049	0.088
Deviance		100	0.0518	0.0808	0.0498	0.0842
		20	0.054	0.1238	0.0536	0.1326
	0.1	50	0.0584	0.093	0.0608	0.1068
		100	0.0644	0.0768	0.0696	0.0868
		20	0.0272	0.1118	0.0382	0.1426
	0	50	0.0414	0.0862	0.047	0.097
Rao		100	0.045	0.0836	0.0472	0.0894
		20	0.0228	0.1114	0.0442	0.1614
	0.1	50	0.0396	0.0846	0.059	0.1136
		100	0.0482	0.0722	0.065	0.091

Table 3: Comparison of the observed powers of the testing methods based on the two families at nominal level 0:05

Method	∈	*n*	PHD	HD	PLHD	LHD
		20	0.8004	0.908	0.8282	0.9404
	0	50	0.9954	0.9984	0.9956	0.9984
Deviance		100	1	1	1	1
		20	0.7134	0.8676	0.7644	0.9186
	0.1	50	0.9792	0.9926	0.98	0.994
		100	0.9998	0.9998	0.9998	0.9998
		20	0.7014	0.9142	0.7614	0.9318
	0	50	0.9942	0.9984	0.9952	0.999
Rao		100	1	1	1	1
		20	0.531	0.8516	0.6592	0.9032
	0.1	50	0.9596	0.9896	0.9734	0.9926
		100	0.9996	1	1	1

against the alternative hypothesis

$$H_1 : \pi_i = \begin{cases} \frac{1-\frac{\gamma}{k-1}}{k} & \text{if } i = 1, 2, \cdots, k-1, \\ \frac{1+\gamma}{k} & \text{if } i = k, \end{cases} \tag{38}$$

where $-1 \leq \gamma \leq (k-1)$ is a constant.

The goodness-of-fit test statistic has the form

$$T = 2n \min_{\theta \in \Theta} \mathrm{PLPD}_{\beta,k}(d_n, f_\theta) = 2n\mathrm{PLPD}_{\beta,k}(d_n, f_{\hat{\theta}})$$

The asymptotic distribution corresponding to the penalized LPD closely follows the approach of Mandal and Basu (2013). We omit the proof here, but the asymptotic null distribution is $\chi^2(k-1)$.

5 Simulation Example

In this section we will give a simulation scheme using the LPD and its penalized version with the corresponding member of the PD family.

5.1 Testing using the Penalized LPD

As shown in the earlier sections, the estimators under the minimum PD method and the corresponding minimum LPD method will be same and for that we will only show the testing result to compare among various divergences. We have generated samples from the

(1– $\in$) Poisson (2) + $\in$ Poisson (15) mixture for $\in$ = 0; 0.1 and various sample sizes n = 20, 50, 100. Each experiment is replicated 5000 times and all penalized methods have been implemented by reducing the penalty weights to 50 per cent of their normal weights. Here we have the taken β = –1/2 as the specific case of study in our investigation leading to the four different variants of the Hellinger distance, *viz*., the Hellinger distance (HD; PD with β = –1/2), the penalized Hellinger distance (PHD; PPD with β = –1/2), the logarithmic Hellinger distance (LPD with β = –1/2) and the penalized logarithmic Hellinger distance (PLHD; PLPD with β = –1/2). Table 2 gives us the observed levels while testing H_0 : θ = 2 and the powers given in table 3 considering the testing problem H_0 : θ = 3. It is obvious that in general the PLHD leads to improvements of the same order over the ordinary LHD, as the PHD does over the ordinary HD. The PHD and the PLHD are generally competitive in their performance, and both can serve as excellent choices for parametric hypothesis testing; both have very good small sample efficiencies and strong robustness properties. However, in case of the Rao test the PHD is overly conservative, particularly in small samples, while the PLHD provides more reasonable levels in this case.

CONCLUSION

The logarithmic power divergence family can be an useful candidate in minimum distance inference. In this paper we have shown that this family is closely related to the ordinary power divergence, and generate very similar inference as the ordinary power divergence; we have demonstrated this in the paper both theoretically and through simulations. Thus the inference methods based on the minimum LPD family can be very useful tools for the practitioner. Our limited study clearly indicates that there may be many benets of using the inference methods based on the LPD or PLPD families. In addition, our results give some general indication of the possibility of developing new and useful divergences using the (h, ϕ) route.

ACKNOWLEDGEMENT

This work is part of the doctoral thesis of the first author currently ongoing at the Indian Statistical Institute, India.

REFERENCES

Agresti, A. 1984. Analysis of Ordinal Categorical Data. Wiley, New York.

Ali, S. M. and Silvey, S. D. 1966. A general class of coefficients of divergence of one distribution from another. J. Royal Statistical Soc.,**28 :** 131-42.

Alin, A. 2007. A note on penalized power-divergence test statistics. International Journal of Mathematical, Computational Science and Engineering (World Academy of Science, Engineering and Technology), **1:** 209-15.

Basu, A. and Basu, S. 1998. Penalized minimum disparity methods for multinomial models. *Statistica Sinica*, **8**: 841-60.

Basu, A., Harris, I. and Basu, S. 1996. Tests of hypotheses in discrete models based on the penalized Hellinger distance. *Statistics and Probability Letters*, **27:** 367-73.

Basu, A., Harris, I., Hjort, N. L. and Jones, M. C. 1998. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85**: 549-59.

Basu, A. and Lindsay, B. G. 1994. Minimum disparity estimation for continuous models: Eciency, distributions and robustness. Annals of the Institute of Statistical Mathematics, **46**: 683-705.

Basu, A., Mandal, A. and Pardo, L. 2010. Hypothesis testing for two discrete populations based on the Hellinger distance. Statistics and Probability Letters, **80**:206-14.

Basu, A., Ray, S., Park, C. and Basu, S. 2002. Improved power in multinomial goodness-of-t tests. *Statistician*, **51**:381-93.

Basu, A., Shioya, H. and Park, C. 2011. Statistical Inference: The Minimum Distance Approach. Chapman & Hall/CRC, Boca Raton, FL.

Beran, R. J. 1977. Minimum Hellinger distance estimates for parametric models. Annals of Statistics, **5:** 445-63.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. 1975. Discrete Multivariate Analysis: Theory and Practice. The M. I. T. Press, Cambridge, MA.

Cressie, N. and Read, T. R. C. 1984. Multinomial goodness-of-t tests. Journal of the Royal Statistical Society B, **46:** 440-64.

Csiszár, I. 1963. Eine informations theoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoschen Ketten. Publ. Math. Inst. Hungar. *Acad. Sci.*, **3:** 85-107.

Fienberg, S. E. 1980. The Analysis of Cross Classied Categorical Data. The M. I. T. Press, Cambridge, MA.

Freeman, D. H. 1987. Applied Categorical Data Analysis. Mercel Dekker, New York.

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. L., Harris, I. R. and Basu, A. (2001). A comparison of related density based minimum divergence estimators. Biometrika, 88, 865-873.

Lindsay, B. G. (1994). Eciency versus robustness: The case for minimum Hellinger distance and related methods. Annals of Statistics, 22, 1081-1114.

Mandal, A., Bhandari, S. K. and Basu, A. (2011). Minimum disparity estimation based on combined disparities: asymptotic results. Journal of Statistical Planning and Inference, 141, 701-710.

Men–endez, M. L., Morales, D., Pardo, L. and Vajda, I. (1995). Divergence based estimation and testing of statistical models of classication. Journal of Multivariate Analysis, 54, 329-354.

Pardo, L. (2006). Statistical Inference based on Divergences. CRC/Chapman-Hall.

Pardo, L. and Pardo, M. C. (2003). Minimum power-divergence estimator in three-way contingency tables. Journal of Statistical Computation and Simulation, 73, 819-831.

Park, C., Basu, A. and Harris, I. R. (2001). Tests of hypotheses in multiple samples based on penalized disparities. Journal of Korean Statistical Society, 30, 347-366.

Patra, S., Maji, A., Basu, A. and Pardo, L. (2013). The Power Divergence and the Density Power Divergence Families: the Mathematical Connection. Sankhya B, 75, 16-28.

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

Rao, C. R. (1973). Linear Statistical Inference and Its Applications. John Wiley & Sons, New York.

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

Renyi, A. (1961). On measures of entropy and information. In Proceedings of Fourth Berkeley Symposium on Mathematical Statistics and Probability, volume I, pages 547{561. University of California.

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: Eciency, breakdown points, and examples. Journal of the American Statistical Association, 84, 107-113.

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.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society, 54, 426-482.