COMPARISION OF THREE ROBUST DIVERGENCE

Introduction

Divergences are measures of dissimilarities between probability distributions.The $\phi - divergence$ which was eastablished by Ali and Silvey [2] and Csiszar [1] separately is very popularly known because it is a natural generalisation of the kullback-Leibler divergence which corresponds to likelihood estimation.

The power divergence is considered as the important family under the class of $\phi - divergence$ because it is robust against outliers.In 1998 Basu et.al [3] presented another family of divergences named as density power divergence, having the similar kind of properties to those of power divergence with some advantages over power divergence .Later Jones et.al in 2001 [5] proposed another set of divergence termed as log density power divergence with more power to downweight the effect of outliers.Here our main objective is to bring all the divergences under the same class i.e $\phi - divergence$

1 The Three Divergence family

Let us consider the two probability density function g(x) and $f_\theta(x)$ where g represent continous distribution function based on the observed data for continous data satisfying the property of probability density function i.e ,

$$\int_{-\infty}^{\infty} g(x)dx = 1$$

similarly for descrete dataset it represent the relative frequency of the independent and identically distributed observation and the expression reduces to

$$\sum_{i=1}^{\infty} g_n(i) = 1$$

whereas $f_\theta(x)$ represent parametric distribution which is used to outline the true distribution function satisfying

$$\int_{-\infty}^{\infty} f_\theta(x)dx = 1$$

$$\sum_{i=1}^{\infty} f_\theta(i) = 1$$

respectively for continous and discreate data.

Considering the general form of $\phi - divergence$ [1,2] we will try to deduce all the three divergences.

$$D_\phi(g(x), f_\theta(x)) = \int_{-\infty}^{\infty} f_\theta(x)\phi(\frac{g(x)}{f_\theta(x)})dx \tag{1}$$

For the given expression we need to define $\phi(u)$ for general case such that function should be strictly convex and should satisfy $\phi(1) = 0$ as asssumed in equation (1) where u is equal to $\frac{g(x)}{f_\theta(x)}$ General case:-

$$\phi(u) = 1 + \frac{1}{\alpha - 1}u - \frac{\alpha}{\alpha - 1}u^{\frac{1}{\alpha}} \tag{2}$$

where α is the parameter which is greater than 1 and hence $\phi(u)$ satisfy all the properties of $\phi(u)$ divergence.

Putting the appropriate value of u in the expression of $\phi(u)$ and hence changing simultaneously $D_\phi(g(x), f_\theta(x))$ we can get all the three divergences .We can see in the expression of $\phi(u)$ if we put $u = \frac{g(x)}{f_\theta(x)}$ we will get expression (1)

$$
\begin{aligned}
D_\phi(g(x), f_\theta(x)) &= \int_{-\infty}^{\infty} f_\theta(x)\Big[1 + \frac{1}{\alpha-1}u - \frac{\alpha}{\alpha-1}\Big(\frac{g(x)}{f_\theta(x)}\Big)^{\frac{1}{\alpha}}\Big]dx \\
&= 1 + \frac{1}{\alpha-1}\int_{-\infty}^{\infty} g(x) - \frac{\alpha}{\alpha-1}\int_{-\infty}^{\infty} g^{\frac{1}{\alpha}} f_\theta^{1-\frac{1}{\alpha}} \\
&= 1 + \frac{1}{\alpha-1} - \frac{\alpha}{\alpha-1}\int_{-\infty}^{\infty} g^{\frac{1}{\alpha}} f_\theta^{1-\frac{1}{\alpha}}
\end{aligned}
$$

Here we can see that the above equation is minimisation of the following equation

$$
\min_{\forall\theta\in\Theta} D_\phi(g(x), f_\theta(x)) = \min_{\forall\theta\in\Theta} -\frac{\alpha}{\alpha-1}\int_{-\infty}^{\infty} g^{\frac{1}{\alpha}} f_\theta^{1-\frac{1}{\alpha}} \tag{3}
$$

Hence it is seen that minimisation of the above equation represents the minimisation of power divergence with parameter α which is greater than 1 [4].

We will now see that the density power divergence proposed by Basu et.al [3] also can be brought under the $\phi - Divergence$ by modifying the expression of $D_\phi(g(x), f_\theta(x))$.

$$
\begin{aligned}
D_\phi(g^\alpha(x), f_\theta^\alpha(x)) &= \int_{-\infty}^{\infty} f_\theta^\alpha(x)\phi\Big(\frac{g^\alpha(x)}{f_\theta^\alpha(x)}\Big)dx \\
&= \int_{-\infty}^{\infty} f_\theta^\alpha(x)\Big[1 + \frac{1}{\alpha-1}\frac{g^\alpha(x)}{f_\theta^\alpha(x)} - \frac{\alpha}{\alpha-1}\Big(\frac{g^\alpha(x)}{f_\theta^\alpha(x)}\Big)^{\frac{1}{\alpha}}\Big]dx \\
&= \int_{-\infty}^{\infty} f_\theta^\alpha(x)dx - \int_{-\infty}^{\infty} g^\alpha(x)dx - \frac{\alpha}{\alpha-1}\int_{-\infty}^{\infty} g(x)f_\theta^{\alpha-1}(x)dx
\end{aligned}
$$

Here also $\int_{-\infty}^{\infty} g^\alpha(x)dx$ is irrelavant to our minimisation objective hence the governing equation for minimisation will leads to :

$$
\min_{\forall\theta\in\Theta} D_\phi(g^\alpha(x), f_\theta^\alpha(x)) = \min_{\forall\theta\in\Theta}\Big[\int_{-\infty}^{\infty} f_\theta^\alpha(x)dx - \frac{\alpha}{\alpha-1}\int_{-\infty}^{\infty} g(x)f_\theta^{\alpha-1}(x)\Big] \tag{4}
$$

so fron the equation 4 it is clearly seen that minimisation of $D_\phi(g^\alpha(x), f_\theta^\alpha(x))$ is same as minimisation of $DPD_\alpha(g(x), f_\theta(x))$ [4]

Similar to density power divergence we will now redefine $D_\phi(g(x), f_\theta(x))$ to $D_\phi(g\prime(x), f_\theta\prime(x))$ by introducing new probability density functions $g\prime(x)$ and

$f_{\theta}\prime(x)$ and get log density power divergence from the new equation. where

$$g\prime(x) = \frac{g^{\alpha}(x)}{\int_{-\infty}^{\infty} g^{\alpha}(y)dy}$$

$$f_{\theta}\prime(x) = \frac{f_{\theta}^{\alpha}(x)}{\int_{-\infty}^{\infty} f_{\theta}^{\alpha}(y)dy}$$

Hence calculating the modified $D_{\phi}(g\prime(x), f_{\theta}\prime(x))$ we get,

$$\begin{aligned}
D_{\phi}(g\prime(x), f_{\theta}\prime(x)) &= \int_{-\infty}^{\infty} f_{\theta}\prime(x)\phi(\frac{g\prime(x)}{f_{\theta}\prime(x)}dx \qquad (5) \\
&= \int_{-\infty}^{\infty} f_{\theta}\prime(x)\Big[1 + \frac{1}{\alpha-1}\frac{g\prime(x)}{f_{\theta}\prime(x)} - \frac{\alpha}{\alpha-1}\Big(\frac{g\prime(x)}{f_{\theta}\prime(x)}\Big)^{\frac{1}{\alpha}}\Big]dx \\
&= 1 + \frac{1}{\alpha-1} - \frac{\alpha}{\alpha-1}\int_{-\infty}^{\infty}\Big(\frac{g^{\alpha}(x)}{\int_{-\infty}^{\infty} g^{\alpha}(y)dy}\Big)^{\frac{1}{\alpha}}\Big(\frac{f_{\theta}^{\alpha}(x)}{\int_{-\infty}^{\infty} f_{\theta}^{\alpha}(y)dy}\Big)^{1-\frac{1}{\alpha}}dx \\
&= 1 + \frac{1}{\alpha-1} + \frac{\alpha}{1-\alpha}\int_{-\infty}^{\infty}\Big(\frac{g(x)}{||g(x)||}\Big)\Big(\frac{f_{\theta}(x)}{||f_{\theta}(x)||}\Big)^{\alpha-1}dx \qquad (6)
\end{aligned}$$

where $||f_{\theta}(x)|| = \Big(\int_{-\infty}^{\infty} f_{\theta}^{\alpha}(y)dy\Big)^{\frac{1}{\alpha}}$ represent norm of $f_{\theta}(x)$ and similarly $||g(x)|| = \Big(\int_{-\infty}^{\infty} g^{\alpha}(y)dy\Big)^{\frac{1}{\alpha}}$ represents the norm of g(x). Hence the minimisation of $D_{\phi}(g\prime(x), f_{\theta}\prime(x))$ represents the maximisation of$\int_{-\infty}^{\infty}\Big(\frac{g(x)}{||g(x)||}\Big)\Big(\frac{f_{\theta}(x)}{||f_{\theta}(x)||}\Big)^{\alpha-1}dx$ which is simalar as minimisation of $LDPD_{\alpha}(g, f)$ for α greater than equal to 1

One of the major drawback of the power divergence is that we cannot prevent ourself by the use of non parametric density estimate [6] case of continous model hence the calculation becomes extremely difficult .Apart from this there is no such case for log density power divergence and density power divergence where g(x) appears linearly and we can replace it with its emperical distribution function.Hence for the case of log density power divergence the equation reduces to

$$\begin{aligned}
\min_{\forall\theta\in\Theta} D_{\phi}(g\prime(x), f_{\theta}\prime(x)) &= \min_{\forall\theta\in\Theta}\frac{\alpha}{1-\alpha}\int_{-\infty}^{\infty}\Big(\frac{g(x)}{||g(x)||}\Big)\Big(\frac{f_{\theta}(x)}{||f_{\theta}(x)||}\Big)^{\alpha-1}dx \\
&= \min_{\forall\theta\in\Theta}\frac{\alpha}{1-\alpha}\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{||g(x)||}\Big)\Big(\frac{f_{\theta}(x)}{||f_{\theta}(x)||}\Big)^{\alpha-1} \qquad (7)
\end{aligned}$$

similarly for the case of density power divergence equation 4 becomes

$$D_\phi(g^\alpha(x), f_\theta^\alpha(x)) = \int_{-\infty}^{\infty} f_\theta^\alpha(x)dx - \frac{\alpha}{\alpha-1}\frac{1}{n}\sum_{i=1}^{n} f_\theta^{\alpha-1}(x) \tag{8}$$

2 Data Example

In the given example we have studied the difference between various families like Power divergence, Density power divergence and Log density power divergence and their ability to down weight the contaminated observation. We will demonstrate this difference with a real data example which represents Cysts in embryonic mice kidney. Simple cysts usually do not cause harm but there are some cysts that are causing symptoms or blocking the flow of blood or urine through the kidney. The above experimental was analysed by Chan et.al and published by American Psychological society where a group of 111 kidneys of the mice were subjected to particular steroid where 19 is the large outlier and with some handful of moderate outliers. The second dataset (Stigler, 1977, Table 5)is experimented by Simon Newcombs measured the amount of time required by light to travel a distance of 7442 metres measured in millionth of second and hence calculated the speed of light.

For the first table the equation will reduced as follows:

For Power Divergence:

$$PD_\alpha(g(x), f_\theta(x)) = \min_{\forall\theta\in\Theta} D_\phi(g(x), f_\theta(x)) = \min_{\forall\theta\in\Theta} -\frac{\alpha}{\alpha-1}\int_{-\infty}^{\infty} g^{\frac{1}{\alpha}} f_\theta^{1-\frac{1}{\alpha}}$$

for poissons distribution f_θ can be replaced by :

$$f_\theta = \frac{e^{-\lambda}\lambda^x}{x!} \tag{9}$$

$$PD_\alpha(g(x), f_\theta(x)) = \min_{\forall\theta\in\Theta} -\frac{\alpha}{\alpha-1}\sum_{1}^{n} g(x)^{\frac{1}{\alpha}}\left[\frac{e^{-\lambda}\lambda^x}{x!}\right]^{1-\frac{1}{\alpha}} \tag{10}$$

For Density Power Divergence

$$DPD_\alpha(g(x), f_\theta(x)) = D_\phi(g^\alpha(x), f_\theta^\alpha(x)) = \min_{\forall\theta\in\Theta}\left[\sum_{1}^{n} f_\theta^\alpha(x) - \frac{\alpha}{\alpha-1}\frac{1}{n}\sum_{i=1}^{n} f_\theta(x)^{\alpha-1}\right]$$

$$DPD_\alpha(g(x), f_\theta(x)) = \min_{\forall\theta\in\Theta}\left[\sum_{1}^{n}\left(\frac{e^{-\lambda}\lambda^x}{x!}\right)^\alpha - \frac{\alpha}{\alpha-1}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{e^{-\lambda}\lambda^x}{x!}\right)^{\alpha-1}\right] \tag{11}$$

For Log Density Power Divergence

$$LDPD_\alpha(g(x), f_\theta(x)) = \min_{\forall\theta\in\Theta} D_\phi(g\prime(x), f_\theta\prime(x)) = \min_{\forall\theta\in\Theta}\left[\frac{\alpha}{1-\alpha}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{||g(x)||}\right)\left(\frac{f_\theta(x)}{||f_\theta(x)||}\right)^{\alpha-1}\right]$$

$$LDPD_\alpha(g(x), f_\theta(x)) = \min_{\forall\theta\in\Theta}\left[\frac{\alpha}{1-\alpha}\frac{1}{n}\frac{1}{||g(x)||||f_\theta(x)||}\sum_{i=1}^{n}\left(\frac{e^{-\lambda}\lambda^x}{x!}\right)^{\alpha-1}\right] \tag{12}$$

For the second case the equation will reduced to the following form

For Power Divergence

Here as we have seen earlier that we have to use non parametric density estimate for continous model as we cannot approximate our true data generating equation with relative frequency as we have seen in the case of discreate model.Hence we approximate our true data generating equation by taking proper kernel.

$$\hat{g}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x_i - x}{h}\Big) \tag{13}$$

Where h represents the bandwidth.Here for the data example we have taken Epanechinikov kernel which is defined as [6]:

$$K(u) = \begin{cases} \frac{3}{4}(1-u^2), & \text{if } |u| \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

and satisfied the property that

$$\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K(u)du = 1 \tag{15}$$

Hence the minimisation of Power Divergence will reduced to

$$\begin{aligned} PD_\alpha(g(x), f_\theta(x)) &= \min_{\forall\theta\in\Theta} -\frac{\alpha}{\alpha-1}\int_{-\infty}^{\infty} \hat{g}^{\frac{1}{\alpha}} f_\theta^{1-\frac{1}{\alpha}} \\ PD_\alpha(g(x), f_\theta(x)) &= \min_{\forall\theta\in\Theta} -\frac{\alpha}{\alpha-1}\int_{-\infty}^{\infty} \hat{g}(x)^{\frac{1}{\alpha}} \Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\Big)^{1-\frac{1}{\alpha}} \end{aligned} \tag{16}$$

For Density Power Divergence

$$DPD_\alpha(g(x), f_\theta(x)) = \min_{\forall\theta\in\Theta} D_\phi(g^\alpha(x), f_\theta^\alpha(x)) = \min_{\forall\theta\in\Theta}\Big[\int_{-\infty}^{\infty} f_\theta^\alpha(x)dx - \frac{\alpha}{\alpha-1}\frac{1}{n}\sum_{i=1}^{n} f_\theta^{\alpha-1}(x)\Big] \tag{17}$$

We will reduce this equation to its simplest form for normal distribution.

$$\begin{aligned} \int_{-\infty}^{\infty} f_\theta^\alpha(x)dx &= \Big[\int_{-\infty}^{\infty}\Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\Big)^\alpha\Big]^{\frac{1}{\alpha}} \\ \int_{-\infty}^{\infty} f_\theta^\alpha(x)dx &= \frac{1}{(\sqrt{2\pi\sigma^2})^{\alpha-1}(\sqrt{\alpha})}\Big[\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi(\frac{\sigma}{\sqrt{\alpha}})^2}} e^{-\frac{(x-\mu)^2}{2(\frac{\sigma}{\sqrt{\alpha}})^2}}\Big] \\ \int_{-\infty}^{\infty} f_\theta^\alpha(x)dx &= \frac{1}{(\sqrt{2\pi\sigma^2})^{\alpha-1}(\sqrt{\alpha})} \end{aligned} \tag{18}$$

Hnece the above equation will reduce 17 will reduce to

$$DPD_\alpha(g(x), f_\theta(x)) = \min_{\forall\theta\in\Theta}\Big[\frac{1}{(\sqrt{2\pi\sigma^2})^{\alpha-1}(\sqrt{\alpha})} - \frac{\alpha}{\alpha-1}\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\Big)^{\alpha-1}\Big] \tag{19}$$

For Log Density Power Diveregence

$$
\begin{aligned}
LDPD_\alpha(g(x), f_\theta(x)) &= \min_{\forall\theta\in\Theta} \frac{\alpha}{1-\alpha}\frac{1}{n}\frac{1}{||g(x)||||f_\theta(x)||}\sum_{i=1}^{n} f_\theta^{\alpha-1}(x) \\
LDPD_\alpha(g(x), f_\theta(x)) &= \min_{\forall\theta\in\Theta} \frac{\alpha}{1-\alpha}\frac{1}{n}\frac{1}{||g(x)||||f_\theta(x)||}\sum_{i=1}^{n} \Big(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\Big)^{\alpha-1} \quad (20)
\end{aligned}
$$

The above table consists of the observed frequency and the estimated parameter after considering various values of λ.

Table 1: fit of the poisons model to the cysts in embryonic mice using several estimation method

	No of cysts in kidney														
	0	1	2	3	4	5	6	7	8	9	10	11	12-18	19	$\hat{\theta}$
Obs.	65	14	10	6	4	2	2	2	1	1	1	2	0	1	
PD($\alpha = 1.2$)	42.89	40.78	19.38	6.1	1.46	0.2	0.04	0	0	0	0	0	0	0	0.95
DPD($\alpha = 1.2$)	57.33	37.87	12.51	2.75	0.45	0.006	0	0	0	0	0	0	0	0	0.66
LDPD($\alpha = 1.2$)	58.51	37.46	11.9	2.5	0.4	0.02	0	0	0	0	0	0	0	0	0.64
PD($\alpha = 1.5$)	53.2	39.12	14.38	3.52	0.64	0.09	0.01	0	0	0	0	0	0	0	0.73
DPD($\alpha = 1.5$)	71.32	31.5	6.97	1.02	0.13	0.001	0	0	0	0	0	0	0	0	0.44
LDPD($\alpha = 1.5$)	77.43	27.8	5.01	0.6	0.05	0	0	0	0	0	0	0	0	0	0.36
LD	23.57	36.5	28.2	14.6	5.6	1.7	0.4	0.1	0	0	0	0	0	0	1.54
LD+D	51.3	39.6	15.2	3.9	0.7	0.1	0	0	0	0	0	0	0	0	0.77

Table 2: Fit of normal distribution to the newcombs speed of light dataset using several estimation method

Family	$\hat{\mu}$	$\hat{\sigma}$
LD/PD($\alpha = 1$)	26.212	10.66
DPD($\alpha = 1$)	26.21	10.66
LDPD($\alpha = 1$)	26.21	10.66
LD+D	27.750	5.044
PD($\alpha = 2$)	27.738	5.127
DPD($\alpha = 2$)	27.29	4.67
LDPD($\alpha = 2$)	27.24	4.33
PD($\alpha = 0.5$)	14.40	22.242
DPD($\alpha = 0.5$)	15.40	11.25
LDPD($\alpha = 0.5$)	16.54	18.81

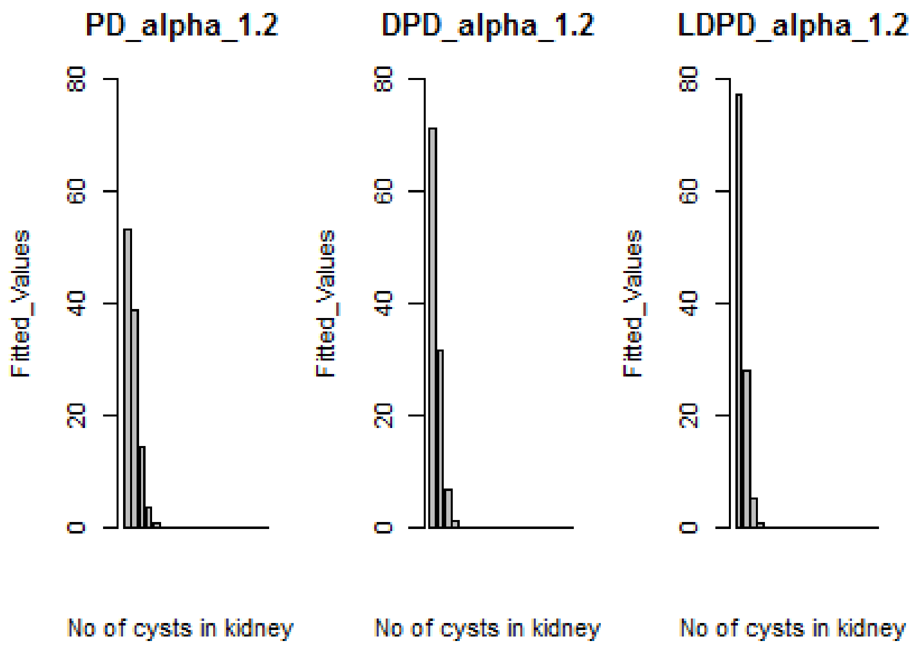

Figure 1: Barplots of three types of divergences corresponding to $\alpha = 1.5$ for cysts in embryonic mice dataset

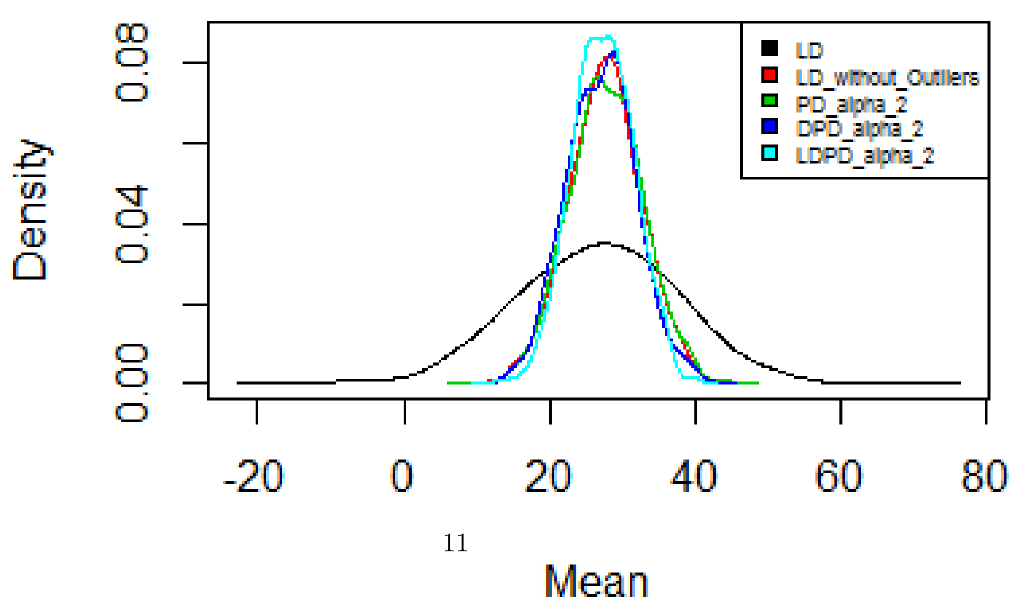

Figure 2: Density Plots for of three types of divergences corresponding to $\alpha = 2$ and LD for Newcombs speed of light dataset

2.0.1 Description

Description: In the first case robustness is measured by taking two values of parameter and we will compare this result with maximum likelihood with outliers and without outliers. According to definition of outlier we will consider beyond Median(+/-) 1.5IQR as outliers hence we will take value after 5 as outliers. From the table it can be seen from the respective parameters of LD and LD+D that the difference is relevant. Here we have noted that all the three divergences have the substantial ability to downweight the effect of outliers for $\alpha > 1$ and it can be seen that PD($\alpha = 1.5$) nearly resembles to Maximum likelihood with deleted outliers.From the Barplots in figure 1 it can be seen LDPD is more steeper than DPD which is again more steeper than PD for the same value of α and hence we can say LDPD is more powerful in down-weighing the value which are far from the bulk values i.e outliers due to its steeply nature therefore we can say it is more robust than the other two divergences .

For the second case in table 2 we have taken some of the values from the book named Statistical inferences written by Ayanendranath Basu,Shioya and C.park [6] where it can be seen from table 2 that PD($\alpha = 1$),DPD($\alpha = 1$)and LDPD($\alpha = 1$) all have the same value of $\hat{\mu}$ and $\hat{\sigma}$ since the general form becomes same form all the three types of divergences at the value of tuning parameter equal to 1.

$$D_\phi(g(x), f_\theta(x)) = D_\phi(g^\alpha(x), f_\theta^\alpha(x)) = D_\phi(g\prime(x), f_\theta\prime(x)),\ \alpha = 1 \tag{21}$$
$$\tag{22}$$

As seen in the previous case it can be seen from the respective parameters of LD and LD+D that the difference is relevant and hence we can conclude that the effect of the outliers is significant on the dataset.Here we have find the parameters at $\alpha = 2$ which is similar to finding parameter by minimizing the Hellinger distance. [6] which is known for it's property of robustness and belongs to Power Divergence.

$$\min_{\forall\theta\in\Theta} D_\phi(g(x), f_\theta(x)) = \min_{\forall\theta\in\Theta} 2\sum\left(\sqrt{g(x)} - \sqrt{f_\theta(x)}\right)^2 \tag{23}$$

From the table 2 and also from figure 2 we can see that Hellinger distance nearly resembles LD+D and hence turned out as the robust distance.Similarly the other two divergences are also robust enough to down-weight the effect of the outliers.At last we have find the values at 0.5 which is out of range of α which is defined for $\alpha > 1$ it can be clearly seen from the table that robustness disappears and hence they fail to lower the effect of outliers.

References

[1] I. Csiszár, "Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizität von Markhoffschen Ketten," *Publ. Math. Inst. Hungar. Acad. Sci.*, vol. 8, pp. 85–108, January 1963.

[2] Ali, S. M.; Silvey, S. D. ,"A general class of coefficients of divergence of one distribution from another" .*Journal of the Royal Statistical Society* , Series B ,vol. 28(1) , pp. 131142 ,1966

[3] A. Basu, I. R. Harris, N. L. Hjort and M. C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998.

[4] S. Patra, A. Maji, L. Padro and A. Basu, "The power divergence and the density power divergence families: the mathematical connection," *Sankhya, Series B*, Vol. 75, pp. 16–28, 2013.

[5] M. C. Jones, N. L. Hjort, I. R. Harris, and A. Basu, "A comparison of related density based minimum divergence estimators," *Biometrika*, vol. 88, no. 3, pp. 865–873, 2001.

[6] A. Basu, H. Shioya and C. Park, "*Statistical Inference: The Minimum Distance Approach*," Chapman & Hall/ CRC Monographs on Statistics and Applied Probability, vol. 120, CRC Press, Boca Raton, Florida, USA, June 2011.