Bernoulli **23**(4A), 2017, 2746–2783
DOI: 10.3150/16-BEJ826

A generalized divergence for statistical inference

ABHIK GHOSH[1], IAN R. HARRIS[2], AVIJIT MAJI[3], AYANENDRANATH BASU[1,*] and LEANDRO PARDO[4]

[1]*Indian Statistical Institute, Kolkata, India. E-mail:* [*]*ayanbasu@isical.ac.in*
[2]*Southern Methodist University, Dallas, USA*
[3]*Indian Statistical Institute, Kolkata, India and Reserve Bank of India, Mumbai, India*
[4]*Complutense University, Madrid, Spain*

The power divergence (PD) and the density power divergence (DPD) families have proven to be useful tools in the area of robust inference. In this paper, we consider a superfamily of divergences which contains both of these families as special cases. The role of this superfamily is studied in several statistical applications, and desirable properties are identified and discussed. In many cases, it is observed that the most preferred minimum divergence estimator within the above collection lies outside the class of minimum PD or minimum DPD estimators, indicating that this superfamily has real utility, rather than just being a routine generalization. The limitation of the usual first order influence function as an effective descriptor of the robustness of the estimator is also demonstrated in this connection.

Keywords: breakdown point; divergence measure; influence function; robust estimation; *S*-divergence

1. Introduction

A density-based minimum divergence approach is a technique in parametric inference where the closeness of the data and the model is quantified by a suitable divergence measure between the data density and the model density. Inference methods based on this approach are useful because of the strong robustness properties that they inherently possess.

The use of minimum divergence procedures in robust statistical inference possibly originated with Beran's 1977 paper [8]. Since then, the literature has grown substantially, with monographs by Vajda [47], Pardo [36] and Basu *et al.* [7] serving as useful resources for the description of the research and developments in this field.

Several density-based minimum divergence estimators have very high asymptotic efficiency. The class of minimum disparity estimators [33] or minimum ϕ-divergence estimators [14], for example, have full asymptotic efficiency under the assumed parametric model. The power divergence (PD) family [13], a prominent subclass of disparities, produces robust and efficient parameter estimators. However, its application in continuous models is not so easy as it requires the use of a non-parametric density estimator such as the kernel density estimator that entails associated complications such as bandwidth selection. Basu *et al.* [2] developed an alternative family of density based divergences, namely the density power divergences (DPD), which avoids any non-parametric smoothing even under continuous models and generates robust estimators with only a slight loss in efficiency. Patra *et al.* [37] proved some interesting connections between these two divergence families.

In this paper, we describe the power divergence and density power divergence families as special cases of a superfamily of divergences which we term as the "S-divergence" family. The main aim of the current paper is, along with the description of the S-divergence family, to illustrate the wide scope of potential applications of this superfamily in case of parametric statistical inference with special focus on the robustness issues.

While we describe the development of the S-divergence in this paper and elaborate on the main statistical features of the divergence with special reference to the robustness of the corresponding procedures, many other research extensions involving this divergence are also being currently considered (or has been considered) by the authors which consist of applications or theoretical properties of this divergence. Ghosh [21] and Ghosh and Basu [22] have considered the asymptotic properties of the minimum S-divergence estimators under discrete and continuous models respectively. Possible applications of the S-divergence family in developing robust tests of hypothesis have been considered by [24,25]. A 2013 technical report [26] provides an extensive review of the PD and DPD families and their interconnection in the general framework of S-divergence. A similar discussion is also provided in a review article by [28], which also illustrates the use of the DPD measure in the context of robust reliability analysis. We will also point out several other interesting and potential areas of future research based on the S-divergences in the present paper.

Here we summarize the main achievements of the paper. These issues will be further elaborated upon when we discuss them in the main body of the paper.

1. We describe the development of a new family of divergences which is a superfamily containing the well known and well studied PD and DPD families. We explore the use of the corresponding minimum divergence estimators in statistical inference.
2. Many of these minimum divergence estimators generated by this class of divergences provide a high degree of stability under data contamination while attaining reasonable efficiency at the model.
3. Our analysis shows that in many cases involving data contamination, the most preferred minimum divergence estimator often lies outside the class of the union of minimum PD and minimum DPD estimators, demonstrating that the class of S-divergences has real utility, and the development of the class of S-divergences is not just for the sake of generalization.
4. We consider a data-driven method for the selection of "optimal" tuning parameters based on modifications of existing rules. This is also useful in establishing the issue discussed in the previous item.
5. Although we do not pursue it in any great length in this paper, we demonstrate that the development of the S-divergence also generates a generalized class of cross-entropies including the Boltzmann–Gibbs–Shannon entropy as a special case.
6. We provide a second order influence analysis which is more accurate than the first order influence analysis in predicting the bias of our estimators under contamination. We also demonstrate that the first order influence function may be deficient in predicting the bias irrespective of the truth; it may fail to indicate the robustness of highly stable estimators, and can also label highly unstable estimators as being strongly robust.
7. At least in the examples we have studied, the instability of the estimators under contamination is very accurately predicted by the violation of the conditions needed for the breakdown point results to hold.

The rest of the paper is organized as follows. In Section 2, we describe the PD and DPD families and study their interconnection; extensive discussions and examples about these divergences can be found in [7,26,38]. Section 3 ties in these families through the super-family of S-divergences. In Section 4, we discuss various properties of the minimum S-divergence estimators including the classical influence function and the asymptotic properties under both discrete and continuous models. A numerical analysis is presented in Section 5 to describe the performance of the proposed minimum S-Divergence estimators (MSDEs). Here we also discuss the limitation of the classical first order influence function analysis in describing the robustness of these estimators. As a remedy, we consider higher order influence function analysis and breakdown point analysis of the proposed minimum divergence estimators in Section 6 and Section 8.1, respectively. Combining the robustness perspectives and efficiency of the proposed estimators, some suggestions are made in Section 7 for data driven optimal choice of estimators to be used in specific situation. In Section 8.2, we discuss robust equivariant estimation of multivariate location and covariances using a particular subfamily of S-divergences. Brief comments about the testing of hypothesis based on the S-divergence family are provided in Section 9. Finally, Section 10 has some concluding remarks. For brevity of presentation, the assumptions required to prove the asymptotic distributions of the proposed minimum divergence estimators and some of the proofs are moved to the on-line supplement [27]. Some additional issues/examples are also included in the supplementary part.

Throughout the paper, we will use the term "density function" for both discrete and continuous models. We also use the term "distance" loosely, to refer to any divergence which is non-negative and is equal to zero if and only if its arguments are identically equal.

2. The PD and the DPD families

2.1. Minimum disparity estimation and the PD family

Let us consider a parametric model of discrete probability distributions $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$. Let $X_1, \ldots, X_n$ denote n independent and identically distributed (i.i.d.) observations from a discrete distribution G. Further, assume that both the true distribution G and the model family $\mathcal{F}$ have the support $\mathcal{X} = \{0, 1, 2, \ldots\}$, without loss of generality, and both belong to $\mathcal{G}$, the class of all distributions having densities with respect to some appropriate dominating measure. Let g and f_θ be the corresponding true and model density functions respectively. Suppose $d_n(x)$ represents the relative frequency of the value x in the above sample. We wish to estimate the parameter θ by minimizing the discrepancy between the data and the model quantified by the class of disparities between the probability vectors $d_n = (d_n(0), d_n(1), \ldots)^T$ and $f_\theta = (f_\theta(0), f_\theta(1), \ldots)^T$.

Definition 2.1 (Definition 2.1, [7]). *Let C be a thrice differentiable, strictly convex function on $[-1, \infty)$, with $C(0) = 0$. Let the Pearson residual at the value x be defined by $\delta(x) = \delta_n(x) = \frac{d_n(x)}{f_\theta(x)} - 1$. Then the disparity between d_n and f_θ generated by C is defined by*

$$\rho_C(d_n, f_\theta) = \sum_{x=0}^{\infty} C\big(\delta(x)\big) f_\theta(x). \tag{1}$$

The PD family [13] is a special case of this class and is defined in terms of a real parameter λ as

$$\mathrm{PD}_\lambda(d_n, f_\theta) = \frac{1}{\lambda(\lambda+1)} \sum d_n \left[\left(\frac{d_n}{f_\theta} \right)^\lambda - 1 \right] \tag{2}$$

$$= \sum \left\{ \frac{1}{\lambda(\lambda+1)} d_n \left[\left(\frac{d_n}{f_\theta} \right)^\lambda - 1 \right] + \frac{f_\theta - d_n}{\lambda+1} \right\}. \tag{3}$$

The second formulation makes all the terms in the summand non-negative. The $C(\cdot)$ function for the PD under this formulation is given by

$$C_\lambda(\delta) = \frac{(\delta+1)^{\lambda+1} - (\delta+1)}{\lambda(\lambda+1)} - \frac{\delta}{\lambda+1}.$$

For particular choices of $\lambda = 1, 0, -1/2, -1$, the PD family generates the Pearson's chi-square (PCS), the likelihood disparity (LD), the (twice, squared) Hellinger distance (HD) and the Kullback–Leibler divergence (KLD) respectively. Note that, the expressions for the PD corresponding to $\lambda = 0, -1$ can only be defined in terms of the continuous limit of the expression on the right-hand side of (2) as $\lambda \to 0, -1$. These cases result in the likelihood disparity (LD) and the Kullback–Leibler divergence (KLD), respectively.

Provided such a minimum exists, the minimum disparity estimator (MDE) $\hat{\theta}$ of θ based on the disparity ρ_C is defined by the relation

$$\rho_C(d_n, f_{\hat{\theta}}) = \min_{\theta \in \Theta} \rho_C(d_n, f_\theta). \tag{4}$$

The maximum likelihood estimator (MLE) belongs to the class of MDEs as it is the minimizer of the likelihood disparity.

Under differentiability of the model, the MDE solves the estimating equation

$$-\nabla \rho_C(d_n, f_\theta) = \sum_{x=0}^{\infty} \big(C'(\delta)(\delta+1) - C(\delta) \big) \nabla f_\theta = \sum_{x=0}^{\infty} A(\delta) \nabla f_\theta = 0, \tag{5}$$

where ∇ represents the gradient with respect to θ and $A(\delta) = C'(\delta)(\delta+1) - C(\delta)$. One can easily check that the above estimating equation (5) is an unbiased estimating equation at the assumed model. We can standardize the function $A(\delta)$, without changing the estimating properties of the disparity, so that $A(0) = 0$ and $A'(0) = 1$. This standardized function $A(\delta)$ is called the residual adjustment function (RAF) of the disparity and largely controls the robustness properties of the MDEs. For the PD family, the RAF is given by

$$A_\lambda(\delta) = \frac{(\delta+1)^{\lambda+1} - 1}{\lambda+1}. \tag{6}$$

In particular, the RAF for likelihood disparity is linear, given by $A_0(\delta) = A_{\mathrm{LD}}(\delta) = \delta$.

An observation x in the sample space with a large ($\gg 0$) Pearson residual δ may be considered a "probabilistic outlier" in the sense that the observed proportion $d_n(x)$ is significantly higher

than what is predicted by the model. An MDE will effectively control such a probabilistic outlier if the corresponding RAF $A(\delta)$ exhibits a strongly dampened response to increasing (positive) δ. In particular, the MDEs with negative values of the quantity $A_2 = A''(0)$, known as the estimation curvature of the disparity [33], are seen to be locally robust. For the PD family, this quantity equals the tuning parameter λ and all members of the PD family with $\lambda < 0$ are expected to be robust. However, the robustness of the MDEs is not adequately reflected by the first order influence function analysis which indicates that all MDEs have the same influence function as the MLE at the model.

However, the equivalence of the influence functions also indicates the asymptotic efficiency of the MDEs. See Lindsay [33] and Morales *et al.* [35] for the formal results in this context as well as for a more in depth discussion of the statements of the robustness properties of the MDE. All MDEs have the same asymptotic distribution as that of the MLE at the model.

The set-up and procedure as described above can also be generalized to the case of continuous models. However, here it is necessary to construct a continuous density estimate of the true density using appropriate non-parametric smoothing techniques which involves bandwidth selection and other problems. This is avoided by the density power divergence family described in the next subsection.

2.2. The density power divergence (DPD) family

We assume that the parametric set-up of Section 2.1 holds. Let $u_\theta(x) = \nabla \log f_\theta(x)$ be the likelihood score function where ∇ represents the gradient with respect to θ. Consider the estimating equations

$$\sum_{i=1}^{n} u_\theta(X_i) = 0 \quad \text{and} \quad \sum_{i=1}^{n} u_\theta(X_i) f_\theta^\alpha(X_i) = 0 \tag{7}$$

in the location model case, where $\alpha \in [0, 1]$. The first one is the likelihood score equation which corresponds to $\alpha = 0$. Clearly the second equation involves a density power downweighting compared to the likelihood equation, which indicates the robustness of the estimators resulting from this process. The degree of downweighting increases with α; $\alpha = 0$ indicates no downweighting. For more general models, the corresponding unbiased estimating equation at the model is

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta(X_i) f_\theta^\alpha(X_i) - \int u_\theta(x) f_\theta^{1+\alpha}(x)\, dx = 0, \qquad \alpha \in [0, 1]. \tag{8}$$

Basu *et al.* [2] used this form to construct the DPD family. Given densities g, f for distributions G and F in $\mathcal{G}$, the DPD with tuning parameter α is defined as

$$\mathrm{DPD}_\alpha(g, f) = \int \left[f^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) f^\alpha g + \frac{1}{\alpha} g^{1+\alpha} \right] \qquad \text{for } \alpha \geq 0. \tag{9}$$

When $\alpha = 0$, the divergence is defined as the continuous limit as $\alpha \downarrow 0$; this divergence equals $\int g \log(g/f)$ which is the likelihood disparity and is also the PD measure at $\lambda = 0$. For $\alpha = 1$, the corresponding divergence is the squared L_2-distance.

Now, consider again the parametric set up of Section 2.1 and define the minimum DPD functional $T_\alpha(G)$ at G by the relation

$$\mathrm{DPD}_\alpha(g, f_{T_\alpha(G)}) = \min_{\theta\in\Theta} \mathrm{DPD}_\alpha(g, f_\theta).$$

The functional is Fisher consistent by definition of the DPD. As $\int g^{1+\alpha}$ is independent of θ and $\int f_\theta^\alpha g$ can be empirically estimated as $\frac{1}{n}\sum_{i=1}^n f_\theta^\alpha(X_i)$, the objective function to be minimized for the estimation of the minimum DPD estimator (MDPDE) $\hat\theta_\alpha$ with tuning parameter α turns out to be

$$H_n(\theta) = \frac{1}{n}\sum_{i=1}^n V_\theta(X_i), \tag{10}$$

where $V_\theta(x) = \int f_\theta^{1+\alpha} - (1+\frac{1}{\alpha}) f_\theta^\alpha(x)$. Thus, the minimization of (10) does not require the use of a non-parametric density estimate for any α. In addition, expression (10) also shows that the MDPDE is in fact an M-estimator. Under differentiability of the model family, the minimization of $H_n(\theta)$ in (10) leads to the estimating equation (8).

Unlike the MDEs, however, all MDPDEs have bounded influence functions for $\alpha > 0$. Also the DPD family allows a clear trade-off between robustness and efficiency, with larger values of α leading to greater robustness and smaller values of α providing greater asymptotic efficiency. Generally, the estimators corresponding to $\alpha > 1$ are considered too inefficient to be practically useful. See Basu *et al.* [2] for more details. Also see Dawid and Musio [15] and Dawid *et al.* [16].

2.3. Interconnection between the PD and the DPD families

The PD measures between the densities g and f_θ may be expressed as

$$\mathrm{PD}_\lambda(g, f_\theta) = \int \left\{ \frac{1}{\lambda(1+\lambda)} \left[\left(\frac{g}{f_\theta}\right)^{1+\lambda} - \left(\frac{g}{f_\theta}\right) \right] + \frac{1 - g/f_\theta}{1+\lambda} \right\} f_\theta. \tag{11}$$

As the expression within the parentheses is non-negative and equals zero only if $g = f_\theta$, the outer f_θ term in (11) can be replaced by $f_\theta^{1+\lambda}$ and one still gets a valid divergence that simplifies to

$$\begin{aligned} &\int \left\{ \frac{[g^{1+\lambda} - g f_\theta^\lambda]}{\lambda(1+\lambda)} + \frac{f_\theta^{1+\lambda} - g f_\theta^\lambda}{1+\lambda} \right\} \\ &= \frac{1}{1+\lambda} \int \left\{ \frac{1}{\lambda}\left[g^{1+\lambda} - g f_\theta^\lambda \right] + f_\theta^{1+\lambda} - g f_\theta^\lambda \right\} \\ &= \frac{1}{1+\lambda} \int \left\{ f_\theta^{1+\lambda} - \left(1 + \frac{1}{\lambda}\right) g f_\theta^\lambda + \frac{1}{\lambda} g^{1+\lambda} \right\}. \end{aligned} \tag{12}$$

The right-hand side of the above equation is just a scaled version of the DPD measure given in equation (9) for $\lambda = \alpha$. This modification, originally employed by Patra *et al.* [37] to reduce the divergence to an empirically estimable quantity, ends up generating the DPD.

The process can also be reversed to obtain the PD family starting from the DPD family.

3. A generalized divergence

We have seen in the previous section that the DPD measure at $\alpha = 1$ equals the squared L_2 distance while the limit $\alpha \to 0$ generates the likelihood disparity. Thus, the DPD family smoothly connects the likelihood disparity, a prominent member of the PD family, with the L_2 distance. A natural question is whether it is possible to construct a family of divergences which connect, in a similar way, other members of the PD family with the L_2 distance. In the following, we propose such a density-based divergence, indexed by two parameters α and λ, that connect each member of the PD family (having parameter λ) at $\alpha = 0$ to the L_2 distance at $\alpha = 1$. We denote this family as the S-divergence family; it is defined by

$$S_{(\alpha,\lambda)}(g, f) = \frac{1}{A}\int f^{1+\alpha} - \frac{1+\alpha}{AB}\int f^B g^A + \frac{1}{B}\int g^{1+\alpha}, \qquad \alpha \in [0, 1], \lambda \in \mathbb{R}, \tag{13}$$

with $A = 1 + \lambda(1 - \alpha)$ and $B = \alpha - \lambda(1 - \alpha)$. Clearly, $A + B = 1 + \alpha$. Also the above form is defined only when $A \neq 0$ and $B \neq 0$. If $A = 0$, then the corresponding S-divergence measure is defined by the continuous limit of (13) as $A \to 0$ which turns out to be

$$S_{(\alpha,\lambda:A=0)}(g, f) = \lim_{A\to 0} S_{(\alpha,\lambda)}(g, f) = \int f^{1+\alpha}\log\left(\frac{f}{g}\right) - \int \frac{(f^{1+\alpha} - g^{1+\alpha})}{1+\alpha}. \tag{14}$$

Similarly, for $B = 0$ the S-divergence measure is defined by

$$S_{(\alpha,\lambda:B=0)}(g, f) = \lim_{B\to 0} S_{(\alpha,\lambda)}(g, f) = \int g^{1+\alpha}\log\left(\frac{g}{f}\right) - \int \frac{(g^{1+\alpha} - f^{1+\alpha})}{1+\alpha}. \tag{15}$$

For $\alpha = 0$, the S-divergence family reduces to the PD family with parameter λ; for $\alpha = 1$, it gives the squared L_2 distance irrespective of the value of λ. For $\lambda = 0$, the S-divergence generates the DPD measure with parameter α.

Theorem 3.1. *Given two densities g and f, the function $S_{(\alpha,\lambda)}(g, f)$ represents a genuine statistical divergence for all $\alpha \geq 0$ and $\lambda \in \mathbb{R}$.*

The proof, which proceeds by showing that the integrand in the definition of the S-divergence itself is non-negative and equals zero if and only if the arguments are identical, is elementary and hence omitted. The S-divergence is not a proper distance metric for all values of the parameters α and λ as they are not symmetric in general. It is easy to see that $S_{(\alpha,\lambda)}(g, f) = S_{(\alpha,\lambda)}(f, g)$ if and only if $A = B$; this happens either if $\alpha = 1$ (which generates the L_2 squared divergence), or $\lambda = -\frac{1}{2}$. The latter case represents an interesting subclass of divergence measures which indeed corresponds to a proper distance and is defined by

$$S_{(\alpha,\lambda=-1/2)}(g, f) = \frac{2}{1+\alpha}\int \left(g^{(1+\alpha)/2} - f^{(1+\alpha)/2}\right)^2.$$

This is a generalized family of Hellinger type distances, which generates the (twice, squared) Hellinger distance at $\alpha = 0$. We will refer to this particular subfamily as the S-Hellinger Distance (SHD) family and study it further in Section 8.2.

Many other special cases are possible for particular values of λ. For example, the subfamily corresponding to $\lambda = 1$ gives the Pearson chi-square at $\alpha = 0$ and the subfamily corresponding to $\lambda = -2$ gives the Neyman chi-square divergence at $\alpha = 0$. However, they both give the (squared) L_2 divergence at $\alpha = 1$. Thus, these two subfamilies of S-divergence can be considered to be generalizations of the Pearson and Neyman chi-square divergences respectively; the robustness of corresponding minimum distance estimator increases with increasing α. These families are defined by the following expressions:

$$S_{(\alpha,1)}(g, f) = \frac{1}{(2-\alpha)} \int \left[f^{1+\alpha} - \frac{(1+\alpha)}{(2\alpha-1)} g^{2-\alpha} f^{2\alpha-1} + \frac{(2-\alpha)}{(2\alpha-1)} g^{1+\alpha} \right] \qquad (\alpha \neq 1/2),$$

$$S_{(\alpha,-2)}(g, f) = \frac{1}{(2\alpha-1)} \int \left[f^{1+\alpha} - \frac{(1+\alpha)}{(2-\alpha)} g^{2\alpha-1} f^{2-\alpha} + \frac{(2\alpha-1)}{(2-\alpha)} g^{1+\alpha} \right] \qquad (\alpha \neq 1/2),$$

and the corresponding divergences at $\alpha = 1/2$ are obtained as their continuous limits.

The S-divergence family has another interpretation from the information theory point of view. We can define a suitable cross-entropy function so that the divergence generated from that entropy is the S-divergence. The result is presented in the following remark.

Remark 3.1. Define the cross-entropy by

$$e(g, f) = -\frac{1+\alpha}{AB} \int g^A f^B + \frac{1}{A} \int f^{1+\alpha}.$$

Then the divergence induced by this cross entropy is obtained as

$$S(g, f) = -e(g, g) + e(g, f),$$

which is exactly the S-Divergence.

We will refer the cross-entropy $e(g, f)$ as the S-cross entropy. Interestingly, at $\lambda = \alpha = 0$, the S-cross entropy reduces to the Boltzmann–Gibbs–Shannon entropy; thus, we also get a generalized family of cross entropy measures as a by-product of the current work on S-divergence. It would be an interesting future work to explore the application of this general S-cross entropy in information theory.

4. Minimum S-divergence estimators (MSDEs)

Under the parametric set-up of Section 2, we are interested in the estimation of the parameter θ. The minimum S-divergence functional $T_{\alpha,\lambda}(G)$ at G is defined by the relation

$$S_{(\alpha,\lambda)}\big(g, f_{T_{\alpha,\lambda}}(G)\big) = \min_{\theta \in \Theta} S_{(\alpha,\lambda)}(g, f_\theta), \tag{16}$$

provided the minimum exists. From its definition, the functional $T_{\alpha,\lambda}(G)$ is Fisher Consistent under the assumption that the model is identifiable. When G is outside the model, $\theta^g_{\alpha,\lambda} = T_{\alpha,\lambda}(G)$

represents the best fitting parameter, and $f_{\theta g}$ is the model element closest to g in the S-divergence sense. For simplicity, we suppress the subscript α, λ for $\theta^g_{\alpha,\lambda}$.

Given the observed data, we estimate the parameter θ by minimizing the divergence $S_{(\alpha,\lambda)}(\hat{g}, f_\theta)$ over $\theta \in \Theta$, where $\hat{g}$ is some non-parametric estimate of g based on the sample. When the model is discrete, a simple choice for $\hat{g}$ is given by the relative frequencies; for continuous models there is no such simple choice and we need something like a non-parametric estimator $\hat{g}$ for g. The estimating equation for the minimum S-divergence estimator (MSDE) is given by

$$\begin{aligned} &\int f_\theta^{1+\alpha} u_\theta - \int f_\theta^B \hat{g}^A u_\theta = 0, \quad \text{or} \\ &\int K\big(\delta(x)\big) f_\theta^{1+\alpha}(x) u_\theta(x) = 0, \end{aligned} \tag{17}$$

where $\delta(x) = \delta_n(x) = \frac{\hat{g}(x)}{f_\theta(x)} - 1$ and $K(\delta) = \frac{[(\delta+1)^A - 1]}{A}$. For $\alpha = 0$ one gets $A = 1 + \lambda$, and the function $K(\cdot)$ coincides with the RAF of the PD family in equation (6). For any fixed α, the estimating equation of the MSDEs differ only in the form of the function $K(\cdot)$, so that the robustness properties of the estimators may at least be partially explained in terms of this function; this parallels the role of the RAF in minimum disparity estimation.

To perform minimum distance estimation using the family of S-divergences without any non-parametric smoothing, we need to choose α and λ so that the parameter A in equation (13) equals 1. This requires either $\lambda = 0$ or $\alpha = 1$. Thus the DPD family (corresponding to $\lambda = 0$) is the only subclass of S-divergences that allows parameter estimation without non-parametric smoothing. In fact, as shown by Patra *et al.* [37], the DPD family is the only family of divergences having this property over a larger class divergences. This property of the S-divergence measures can also be verified using the concept of decomposability introduced by Broniatowski, Toma and Vajda [10] for pseudo-distances, that is, for statistical divergences that do not satisfy the information processing property (e.g., PD, DPD families). The decomposability of a divergence allows defining the corresponding minimum divergence estimators without using any non-parametric density estimator. This is one of the fundamental difference between the PD and DPD families – the DPDs are decomposable but PDs are not; as a consequence the minimum divergence estimators based on DPDs do not use non-parametric smoothing, while those based on PDs inevitably require this. The family of S-divergence measures also satisfies the divergence properties but do not satisfy the information processing property and so they are indeed a family of pseudo-distances. Therefore, we can apply the results of Broniatowski, Toma and Vajda [10] to show that the S-divergences are decomposable only if $A = 1$, that is, only if $\lambda = 0$ or $\alpha = 1$. Therefore, the minimum S-divergence estimators avoiding the use of non-parametric density estimators are those based on the DPDs and the L_2-distance only, as noted earlier. Further, these decomposable members of the S-divergence family are the only divergences for which equation (17) defines an M-estimators as we can then rewrite the estimating equation in terms of sum of i.i.d. terms.

However, all members of the S-divergence family generate affine invariant estimators as presented in the following proposition. The proof is elementary and hence omitted.

Proposition 4.1. *Consider the transformation $Y = UX + v$ for some non-singular matrix U and vector v of the same dimension as that of the variable X. Then it easy to see that*

$$S_{(\alpha,\lambda)}(g_Y, f_Y) = kS_{(\alpha,\lambda)}(g_X, f_X),$$

where $k = |\mathrm{Det}(U)|^{1+\alpha} > 0$. Thus although the divergence $S_{(\alpha,\lambda)}(g, f)$ is not affine invariant the estimator that is obtained by minimizing this divergence is affine equivariant.

Considering the entropy interpretation as discussed in Remark 3.1, we can provide an alternative view for the proposed MSDEs. The minimization of the S-divergence measure $S_{(\alpha,\lambda)}(g, f_\theta)$ between the data g and model f_θ is equivalent to the minimization of the S-cross entropy between the corresponding uncertainty variables. The MSDEs are also the minimum S-cross entropy estimators and provide a generalized approach of the works of Shore and Johnson [40], Boer *et al.* [17], Wittenberg [49] and others. This may have potential application in information theory from the robustness perspective.

4.1. Influence functions

The influence function of an estimator is a common and useful indicator of its first-order robustness and efficiency. The influence function of a robust estimator should be bounded; nonboundedness of the influence function implies that the first order asymptotic bias of the estimators may diverge to infinity under contamination. In this section, we will examine the (first-order) robustness of the MSDEs in terms of its influence function.

Consider the minimum S-divergence functional $T_{\alpha,\lambda}(\cdot)$ as defined in (16). A straightforward differentiation of the estimating equation yields the following theorem.

Theorem 4.2. *Under above mentioned set-up, the influence function of the minimum S-divergence functional $T_{\alpha,\lambda}$ is given by*

$$\mathrm{IF}(y; T_{\alpha,\lambda}, G) = J^{-1}\big[Au_{\theta^g}(y) f_{\theta^g}^B(y) g^{A-1}(y) - \xi\big], \tag{18}$$

where $\xi = \xi(\theta^g)$, $J = J(\theta^g)$ with $\xi(\theta) = A\int u_\theta f_\theta^B g^A$ and

$$J(\theta) = A\int u_\theta^2 f_\theta^{1+\alpha} + \int \big(i_\theta - Bu_\theta^2\big)\big(g^A - f_\theta^A\big) f_\theta^B,$$

and $i_\theta(x) = -\nabla[u_\theta(x)]$.

Further, if the true distribution G belongs to the model family $\mathcal{F}$ with $g = f_\theta$, then the influence function of the minimum S-divergence functional simplifies to

$$\mathrm{IF}(y; F_\theta, T_{\alpha,\lambda}) = \frac{u_\theta(y) f_\theta^\alpha(y) - \int u_\theta f_\theta^{1+\alpha}}{\int u_\theta^2 f_\theta^{1+\alpha}}. \tag{19}$$

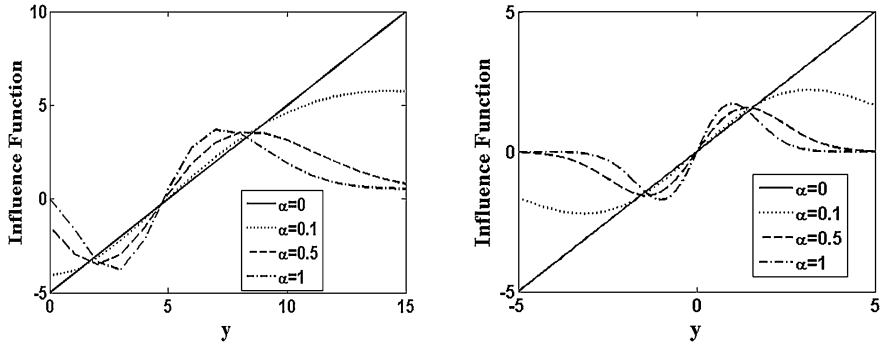


Figure 1. Influence function for the minimum S-divergence estimator of θ under the Poisson(θ) model at the Poisson(5) distribution (first panel) and the normal $N(\mu, 1)$ model at the $N(0, 1)$ distribution (second panel).

The remarkable observation here is that the above influence function of the MSDE at the model depends only on the parameter α and not on λ. Thus, the influence function analysis will predict similar behavior (in terms of first order robustness and efficiency) for all minimum S-divergence estimators with the same value of α irrespective of λ. In addition, this influence function is the same as that of the minimum DPD estimators for a fixed value of α; thus it has a bounded redescending nature except in the case where $\alpha = 0$. Figure 1 shows the nature of the influence functions for the Poisson-mean (discrete case) and Normal-mean (continuous case). This equivalence of the influence functions also indicates that the asymptotic variance of the minimum S-divergence estimators corresponding to any given (α, λ) pair should be the same as that of the corresponding minimum DPD estimator with the same value of α (irrespective of the value of λ).

4.2. Asymptotic properties: Discrete models

As noted earlier, our main aim in this paper is to propose a new family of divergences and indicate its potential uses in robust parametric inference. However, for the sake of completeness, we briefly mention the asymptotic properties of the proposed MSDEs – for the discrete models in this subsection and for the continuous models in the next subsection. Detailed proofs and discussions can be found in [21] and [22], respectively.

Consider the set-up and notations of discrete models as in Section 2.1. Consider the MSDE obtained by minimizing $S_{(\alpha,\lambda)}(d_n, f_\theta)$ over $\theta \in \Theta$, where d_n is the relative frequency. Define

$$J_g = E_g\big[u_{\theta g}(X)u_{\theta g}^T(X)K'\big(\delta_g^g(X)\big)f_{\theta g}^\alpha(X)\big] - \sum_{x=0}^{\infty} K\big(\delta_g^g(X)\big)\nabla_2 f_{\theta g}(x)$$

and $V_g = V_g[K'(\delta_g^g(X))f_{\theta g}^\alpha(X)u_{\theta g}(X)]$, where E_g and V_g represents the expectation and variance under g respectively, $K'(\cdot)$ denotes the first derivative of $K(\cdot)$ and θ^g is the best fitting parameter corresponding to the density g. Under the conditions (SA1)–(SA7) of [21] given in the on-line supplement, the MSDEs have the following asymptotic properties.

Theorem 4.3. *Under the above mentioned set-up and conditions* (SA1)–(SA7) *of* [21], *the following results hold*:

(a) *There exists a consistent sequence* θ_n *of roots to the minimum S-divergence estimating equation* (17).
(b) *The asymptotic distribution of* $n^{1/2}(\theta_n - \theta^g)$ *is p-dimensional normal with vector mean* 0 *and covariance matrix* $J_g^{-1} V_g J_g^{-1}$.

Corollary 4.4. *If the true distribution* $G = F_\theta$ *belongs to the model,* $n^{1/2}(\theta_n - \theta)$ *has an asymptotic* $N_p(0, J^{-1} V J^{-1})$ *distribution, where* $J = J_\alpha(\theta) = \int u_\theta u_\theta^T f_\theta^{1+\alpha}$, $V = V_\alpha(\theta) = \int u_\theta u_\theta^T f_\theta^{1+2\alpha} - \xi\xi^T$, *and* $\xi = \xi_\alpha(\theta) = \int u_\theta f_\theta^{1+\alpha}$. *This asymptotic distribution is the same as that of the DPD, and is independent of the parameter* λ.

4.3. Asymptotic properties: Continuous models

Now, let us consider the case of continuous models. In this case, we cannot simply use the relative frequency to estimate the data density g; instead we use the kernel estimator of the density given by

$$g_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} W(x, X_i, h_n) = \int W(x, y, h_n)\, dG_n(y), \tag{20}$$

where $W(x, y, h_n)$ is a smooth kernel function with bandwidth h_n and G_n is the empirical distribution function as obtained from the data. We estimate θ by minimizing the S-divergence measure between the kernel estimator g_n^* and the model density f_θ. Under suitable differentiability assumptions, the corresponding estimating equation is then given by (17) with $\hat{g}$ replaced by g_n^*. The rest of the procedure is similar to the discrete case; however the theoretical derivation of the asymptotic normality of the MSDEs and the description of their other asymptotic properties become far more complex due to the inclusion of the kernel density estimator. In particular, the choice of the sequence of kernel bandwidths h_n becomes critical.

In order to avoid such complications, Ghosh and Basu [22] derived the asymptotic properties of the MSDEs under the continuous model following an alternative smoothed model approach of [3]. They have considered the kernel integrated "smoothed" version of the model density defined as

$$f_\theta^*(x) = \int W(x, y, h_n)\, dF_\theta(y), \tag{21}$$

and have estimated θ by minimizing the S-divergence measure between g_n^* and f_θ^* over Θ taking $h_n = h$, independent of sample size. The resulting estimator, which they have referred to as the minimum S^*-divergence estimator (MSDE*) is, in general, not the same as the estimator obtained by minimizing $S_{(\alpha,\lambda)}(g_n^*, f_\theta)$. Ghosh and Basu have shown the $n^{1/2}$-consistency and asymptotic normality of the MSDE* under the conditions (SB1)–(SB7) of [22], which appear to be substantially simpler than what would be necessary to prove the corresponding properties for the MSDE itself.

Theorem 4.5. *Under the above mentioned set-up and conditions* (SB1)–(SB7) *of* [22], *the following results hold*:

(a) *There exists a consistent sequence* θ_n^* *of roots to the minimum* S^**-divergence estimating equation obtained by replacing* g *and* f_θ *by* g_n^* *and* f_θ^*, *respectively in* (17).

(b) *The asymptotic distribution of* $n^{1/2}(\theta_n^* - \theta^g)$ *is* p*-dimensional normal with mean* 0 *and covariance matrix* $[J_g^*]^{-1}V_g^*[J_g^*]^{-1}$, *where* θ^g *is now the best fitting parameter with respect to the* S*-divergence between* $g^* = \int W(\cdot, y, h)\,dG(y)$ *and* f_θ^*,

$$\begin{aligned} J_g^* &= J_{(\alpha,\lambda)}^*(g) \\ &= A\int \big(f_{\theta^g}^*\big)^{1+\alpha}\widetilde{u_{\theta^g}}\,\widetilde{u_{\theta^g}}^T + \int \big(\widetilde{i_{\theta^g}} - B\widetilde{u_{\theta^g}}\,\widetilde{u_{\theta^g}}^T\big)\big[\big(g^*\big)^A - \big(f_{\theta^g}^*\big)^A\big]\big(f_{\theta^g}^*\big)^B, \end{aligned} \tag{22}$$

and

$$V_g^* = V_{(\alpha,\lambda)}^*(g) = \mathrm{Var}\bigg[\int W(x, X, h)K'\big(\delta_g^{g^*}(x)\big)\big(f_{\theta^g}^*(x)\big)^\alpha \widetilde{u_{\theta^g}}(x)\,dx\bigg] \tag{23}$$

with $\widetilde{u}_\theta(x) = \nabla \log f_\theta^*(x)$ *and* $\widetilde{i}_\theta(x) = -\nabla[\widetilde{u_\theta}(x)]$.

In particular, at the model distribution, the above asymptotic distribution is independent of the parameter λ as in the discrete cases. Further, one can find suitable conditions on the kernel function so that the asymptotic variance (and all other first order asymptotic and robustness properties) of MSDE* have the same form as that of the MSDE under discrete models. See [22] for detailed proofs and discussions.

In this context, an interesting alternative solution to the above problem of kernel smoothing could be to transform the S-divergence measures and the defining equation of the corresponding minimum S-divergence estimator (MSDE) in such a way that the transformed problem avoids the use of non-parametric kernel density estimators yet yields robust estimators with similar properties as that of the MSDEs. Broniatowski and Keziou [9] and Toma and Broniatowski [45] proposed one such approach in the context of ϕ-divergence families (including the power divergence family) based on the dual form of the divergences. Their proposed dual ϕ-divergence estimators and minimum dual ϕ-divergence estimators are seen to be highly robust like the minimum ϕ-divergence estimators with no significant loss in asymptotic efficiency. It would be an interesting future work to study the properties of the dual form of the S-divergence measures and corresponding estimators in order to avoid the complications of kernel smoothing under continuous models.

5. Numerical studies: Limitations of the influence function

The classical first order influence function is generally a useful descriptor of the robustness of the estimator. However, the fact that the influence function of the MSDEs are independent of λ raises several questions. In actual practice the behavior of the MSDEs vary greatly over λ, and in this section we will give several numerical illustrations of this differential behavior. This

Table 1. The Empirical bias of the MSDEs under pure data from Poisson Model

λ	$\alpha=0$	$\alpha=0.1$	$\alpha=0.25$	$\alpha=0.4$	$\alpha=0.5$	$\alpha=0.6$	$\alpha=0.8$	$\alpha=1$
−1.0	–	−0.321	−0.122	−0.053	−0.029	−0.014	0.001	0.006
−0.7	−0.172	−0.111	−0.057	−0.027	−0.015	−0.007	0.003	0.006
−0.5	−0.093	−0.062	−0.033	−0.015	−0.008	−0.002	0.004	0.006
−0.3	−0.045	−0.030	−0.014	−0.005	−0.001	0.002	0.005	0.006
0.0	0.006	0.007	0.008	0.007	0.007	0.007	0.006	0.006
0.5	0.073	0.059	0.040	0.026	0.020	0.015	0.009	0.006
1.0	0.124	0.103	0.072	0.045	−0.024	0.022	0.011	0.006
1.5	0.161	0.139	0.102	0.065	0.045	0.032	0.014	0.006
2.0	0.189	0.167	0.129	0.087	0.060	0.039	0.016	0.006

also indicates that the influence function provides an inadequate description of the robustness of the minimum distance estimators within the S-divergence family. In the next section, we will demonstrate that a second order bias approximation (rather than the first order) gives a more accurate picture of reality, further highlighting the limitations of the first order influence function in this context.

5.1. A simple discrete model: Poisson model

For simplicity of presentation, we will first restrict our attention to the discrete case; in particular we will concentrate on the Poisson model. The numbers reported by [21,22] demonstrate that similar results are often obtained for other discrete or continuous models.

We consider a sample size of $n = 50$ and simulate data from a Poisson($\theta = 3$) distribution. We compute the MSDEs of θ for several combinations of α and λ and calculate the empirical bias and the MSE of each such estimator over 1000 replications. Our findings are reported in Tables 1 and 2.

Table 2. The Empirical MSE of the MSDEs under pure data from Poisson Model

λ	$\alpha=0$	$\alpha=0.1$	$\alpha=0.25$	$\alpha=0.4$	$\alpha=0.5$	$\alpha=0.6$	$\alpha=0.8$	$\alpha=1$
−1.0	–	0.203	0.086	0.071	0.071	0.073	0.079	0.086
−0.7	0.098	0.078	0.068	0.068	0.070	0.072	0.079	0.086
−0.5	0.070	0.065	0.064	0.067	0.069	0.072	0.079	0.086
−0.3	0.062	0.061	0.063	0.066	0.069	0.072	0.079	0.086
0.0	0.060	0.060	0.062	0.066	0.069	0.072	0.079	0.086
0.5	0.074	0.068	0.064	0.066	0.069	0.072	0.078	0.086
1.0	0.100	0.086	0.072	0.068	0.069	0.071	0.078	0.086
1.5	0.125	0.108	0.085	0.072	0.070	0.072	0.078	0.086
2.0	0.146	0.128	0.101	0.080	0.073	0.072	0.078	0.086

It is clear from the table that both the bias and MSE are quite small for all values of α and λ, although the MSE values do exhibit some increase with α, particularly for $\alpha > 0.5$. Simulation results done here and elsewhere indicate that under the model most minimum S-divergence estimators perform reasonably well. The parameter λ has, on the whole, marginal overall impact on the MSE values, although the values are less stable for very large or very small values of λ. More detailed simulation results, not presented here, demonstrate that the asymptotic convergence to the limiting distribution is slower for such values of λ. In particular, the MSE of the estimator is not available for the $(\lambda = -1, \alpha = 0)$ combination, as the observed frequencies of the cells show up in the denominator in this case, and the estimator is undefined for a single empty cell. Although the estimators do exist for positive α when $\lambda = -1$, the ($\lambda = -1$, α small) estimators remain somewhat unstable. The estimators corresponding to a very large positive λ and very small α also appear to be less stable than the other estimators. This is a manifestation of what is known as the inlier problem; see Basu *et al.*[7].

To explore the robustness properties of the minimum S-divergence estimators, we repeat the above study, but introduce a contamination in the data by (i) replacing the last observation of the sample with the value 50, or by (ii) randomly replacing 10% of the observations of the sample by Poisson($\theta = 12$) observations. We again compute the empirical bias and MSE for several values of α and λ against the target value of $\theta = 3$. We report findings for the contamination scheme (i) in Tables 3 and 4 and the findings of scheme (ii) are reported in Tables 5 and 6. The results in Tables 3 and 4 demonstrate that the MSDEs are robust to the outlying value for all $\alpha \in [0, 1]$ if $\lambda < 0$. For $\lambda = 0$, the estimators are largely unaffected for positive values of α, but at $\alpha = 0$ are adversely affected (note that $\alpha = 0$ and $\lambda = 0$ gives the MLE). For $\lambda > 0$, the corresponding estimators are highly sensitive to the outliers; this sensitivity decreases with α, and eventually the outlier has negligible effect on the estimator when α is very close to 1. The robustness of the estimators decrease sharply with increasing λ except when $\alpha = 1$ (in which case we get the L_2 divergence irrespective of the value of λ).

Tables 5 and 6 give the corresponding results for the second kind of contamination. While the general conclusions are similar, there are very important differences as well. We discuss some of these below.

Table 3. The Empirical Bias of the MSDEs with contaminated data (one outlier at $x = 50$)

λ	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.25$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1$
−1.0	–	−0.304	−0.107	−0.037	−0.012	0.004	0.022	0.031
−0.7	−0.159	−0.097	−0.042	−0.011	0.003	0.012	0.024	0.031
−0.5	−0.080	−0.049	−0.018	0.002	0.010	0.017	0.026	0.031
−0.3	−0.033	−0.016	0.001	0.012	0.017	0.021	0.027	0.031
0.0	0.957	0.021	0.023	0.024	0.025	0.026	0.028	0.031
0.5	15.039	14.094	9.584	0.043	0.038	0.034	0.031	0.031
1.0	15.832	15.579	14.706	11.364	0.316	0.042	0.033	0.031
1.5	16.025	15.911	15.559	14.501	12.073	9.135	0.036	0.031
2.0	16.100	16.033	15.844	15.339	14.363	10.807	0.038	0.031

Table 4. The Empirical MSE of the MSDEs with contaminated data (one outlier at $x = 50$)

λ	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.25$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1$
−1.0	–	0.221	0.095	0.077	0.075	0.077	0.083	0.090
−0.7	0.107	0.084	0.073	0.072	0.073	0.076	0.082	0.090
−0.5	0.076	0.070	0.068	0.070	0.072	0.075	0.082	0.090
−0.3	0.066	0.065	0.066	0.069	0.072	0.075	0.082	0.090
0.0	0.976	0.063	0.065	0.068	0.071	0.074	0.082	0.090
0.5	226.217	198.686	91.878	0.068	0.071	0.074	0.081	0.090
1.0	250.719	242.759	216.292	129.174	0.171	0.073	0.081	0.090
1.5	256.899	253.246	242.149	210.318	145.791	90.100	0.080	0.090
2.0	259.291	257.160	251.120	235.340	206.341	116.826	0.080	0.090

Table 5. The Empirical MSE of the MSDE with contaminated data (10% outlier from Poisson(12))

λ	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.25$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1$
−1	–	−0.224	−0.047	0.015	0.037	0.051	0.069	0.0796
−0.7	−0.054	−0.010	0.027	0.047	0.056	0.063	0.072	0.0796
−0.5	0.064	0.064	0.065	0.066	0.068	0.070	0.074	0.0796
−0.3	0.216	0.154	0.107	0.087	0.080	0.077	0.076	0.0796
0	0.861	0.492	0.220	0.128	0.103	0.089	0.080	0.0796
0.5	2.119	1.760	1.008	0.335	0.176	0.119	0.085	0.0796
1	2.584	2.383	1.903	1.081	0.404	0.186	0.092	0.0796
1.5	2.779	2.656	2.352	1.774	1.125	0.776	0.100	0.0796
2	2.875	2.793	2.588	2.185	1.690	0.900	0.111	0.0796

Table 6. The Empirical MSE of the MSDE with contaminated data (10% outlier from Poisson(12))

λ	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.25$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1$
−1	–	0.215	0.102	0.086	0.086	0.088	0.095	0.1034
−0.7	0.111	0.095	0.086	0.085	0.086	0.089	0.095	0.1034
−0.5	0.098	0.091	0.087	0.086	0.088	0.090	0.096	0.1034
−0.3	0.136	0.109	0.094	0.089	0.089	0.091	0.096	0.1034
0	0.812	0.318	0.131	0.099	0.094	0.093	0.096	0.1034
0.5	4.703	3.283	1.125	0.194	0.116	0.100	0.097	0.1034
1	6.932	5.934	3.850	1.311	0.249	0.122	0.099	0.1034
1.5	7.983	7.318	5.798	3.386	1.431	0.807	0.100	0.1034
2	8.527	8.066	6.971	5.048	3.100	0.957	0.103	0.1034

The contamination in case (i) represents an absolutely extreme outlier, which normally would be identified and discounted by most robust estimation methods. This is essentially what is observed in Tables 3 and 4. That is why, even the weakly robust estimators generated by the S-divergences, such as the one corresponding to $\alpha = 0.1$, $\lambda = 0$ perform admirably in this case. There is, in fact, little reason to venture outside the fold of DPD generated estimators in this case.

The contamination in case (ii), on the other hand, will generate observations of which some may be legitimately confused with the observations coming from the major Poisson(3) component. Some others represent mild outliers, while some would be moderate to extreme outliers. In such a situation, it is clear from Tables 5 and 6 that the minimum divergence estimators within the PD and DPD classes no longer contain the divergence which minimizes the mean square error. In fact, the best performance in Table 6 ($\alpha = 0.4$, $\lambda = -0.7$) is well separated from both the PD and DPD family boundaries. Roughly, the zone of best performance is an elliptical (or circular) subset of the tuning parameter space, with one axis extending roughly from $\alpha = 0.1$ to $\alpha = 0.6$ and the other roughly from $\lambda = -0.3$ to $\lambda = -1$.

The above study shows that there are many useful divergences in the S-divergence class which lie outside the PD–DPD families. The above study also clearly illustrates that the robustness properties of the MSDEs are critically dependent on the value of λ for each given value of α. Yet, as we have seen, the canonical (first order) influence functions of the MSDE are independent of λ, and this index would fail to make any distinction between the different estimators for a fixed value of α; this property severely limits the usefulness of the influence function in assessing the robustness credentials of these estimators. In practice, estimators with $\alpha = 0$ and negative λ appear to have excellent outlier resistant properties, while those corresponding to small positive values of α and large positive λ perform poorly at the model in terms of robustness; in either case the these behaviors are contrary to what would be expected from the usual influence function approach.

5.2. Mixture normal models

We now consider the more complex example of normal mixture models, where we assume that the number of mixing components is known. In particular, we will focus on a two-component mixture of normal distributions with density

$$f_\theta = \pi\phi\big(\cdot, \mu_1, \sigma_1^2\big) + (1-\pi)\phi\big(\cdot, \mu_2, \sigma_2^2\big), \qquad \theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \pi)^T, \tag{24}$$

where $\phi(\cdot, \mu, \sigma^2)$ denotes a normal density with mean μ and variance σ^2 and π is the mixing proportion. We are interested in the estimation of the parameter vector θ. The approach can be easily extended to the mixture of more than two distributions.

A robust solution to this problem can be easily obtained using the minimum S-divergence estimators. However, since we have a continuous model in this case we will follow the Basu–Lindsay approach of smoothed model as discussed in Section 4.3. Following [22], the Gaussian kernel with $W(x, y, h)$ being the $N(y, h^2)$ density at x gives a reasonable choice under the normal model while using the Basu–Lindsay approach. Note that, the smoothed model with

respect to the Gaussian kernel is then given by

$$f_\theta^* = \pi\phi\big(\cdot, \mu_1, \sigma_1^2 + h^2\big) + (1-\pi)\phi\big(\cdot, \mu_2, \sigma_2^2 + h^2\big).$$

Then we can easily derive the proposed MSDE* of θ under this mixture normal model by minimizing the S-divergence between the smoothed model f_θ^* and the kernel density estimate of the true density with respect to the same Gaussian kernel.

Here we will present some simulation results to illustrate the performance of our proposed MSDE* under this relatively complex model situation. As in the previous example, we will take the sample size $n = 50$ and generate 1000 samples from the assumed model (24) with the true parameter value $\theta = (0, 10, 3, 3, 0.5)$. Here, for simplicity, we have taken $\sigma_1 = \sigma_2 = 3$ as known and numerically compute the MSDE* of the parameters (μ_1, μ_2, π). For the computation purpose, we have taken the bandwidth $h = h_n$ as per the normal reference rule [39] yielding

$$h_n = 1.06\sigma n^{-1/5},$$

with σ being the common standard deviation (which is assumed to be known and equals 3). Next, for examining the robustness, we contaminate $100\epsilon\%$ of the sample by observations from a $N(20, 3)$ distribution which are potential outliers with respect to the assumed two-component mixture model. Based on the 1000 such Monte-Carlo samples (each of size $n = 50$), we have computed the bias and MSE of the MSDE* of the three parameters for different choices of the contamination proportion ϵ. For simplicity of presentation, we have presented only the average absolute bias and the average MSE of the three parameter estimates under the 10% contamination scenario in Tables 7 and 8, respectively.

We notice that the general pattern of the performance of the estimators remains the same. In this case, the system requires larger values of α for the estimators to achieve reasonable degrees of robustness. Notice that the model is substantially more complicated than the Poisson model, involving three parameters, an additional tuning parameter (the bandwidth) and a measure which is the combination of the three absolute biases or the MSEs. However in this case also the best solution lies well outside the PD–DPD families.

Table 7. The Empirical average absolute bias of the MSDE*s for different values of α and λ (10% contamination in Normal Mixture Model)

λ	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.25$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1$
−1.0	0.756	0.611	0.492	0.420	0.391	0.353	0.309	0.308
−0.7	0.872	0.739	0.534	0.451	0.409	0.381	0.299	0.308
−0.5	0.975	0.800	0.567	0.470	0.430	0.365	0.322	0.308
−0.3	1.037	0.868	0.624	0.503	0.443	0.395	0.325	0.308
0	1.207	0.889	0.690	0.527	0.466	0.416	0.338	0.308
0.5	1.612	1.395	0.937	0.563	0.516	0.443	0.348	0.308
1	1.844	1.669	1.329	0.862	0.584	0.480	0.349	0.308
1.5	1.953	1.829	1.571	1.134	0.786	0.526	0.361	0.308
2	2.038	1.940	1.691	1.342	1.031	0.616	0.374	0.308

Table 8. The Empirical average MSE of the MSDE*s for different values of α and λ (10% contamination in Normal Mixture Model)

λ	$\alpha=0$	$\alpha=0.1$	$\alpha=0.25$	$\alpha=0.4$	$\alpha=0.5$	$\alpha=0.6$	$\alpha=0.8$	$\alpha=1$
−1	2.890	2.512	1.976	1.658	1.477	1.470	1.366	1.409
−0.7	3.306	2.882	2.182	1.642	1.606	1.525	1.436	1.409
−0.5	3.729	3.136	2.331	1.797	1.583	1.486	1.385	1.409
−0.3	4.000	3.259	2.482	1.940	1.774	1.499	1.382	1.409
0	4.786	3.299	2.690	2.003	1.791	1.613	1.373	1.409
0.5	7.056	5.662	3.546	2.319	2.021	1.614	1.356	1.409
1	8.635	7.282	5.216	3.121	2.326	1.714	1.383	1.409
1.5	9.546	8.451	6.460	4.246	2.840	2.022	1.455	1.409
2	10.191	9.249	7.345	5.214	3.761	2.293	1.485	1.409

5.3. An errors-in-variables model

In our last numerical example, we will consider the more complex errors-in-variables model. One popular and classical version of the errors-in-variables model, as given in [20], may be expressed by the relations

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \qquad X_i = x_i + u_i, \qquad i = 1, \ldots, n, \tag{25}$$

where we only observe (X_i, Y_i) and x_i is the unobserved true values of the covariates. We also assume that the error variables e_i and u_i are i.i.d. normal with means 0 and variances σ_e and σ_u, respectively. It is also assumed that the covariate x_i's are stochastic and independent of the errors; such a model is generally referred as the structural model. In particular, we will assume that the covariates are i.i.d. normal with mean μ_x and variance σ_x. Then the observed variables (Y_i, X_i) form an i.i.d. random sample from a bivariate normal distribution with mean $\mu = (\beta_0+\beta_1\mu_x, \mu_x)^T$ and variance-covariance matrix $\Sigma = \begin{pmatrix} \beta_1^2\sigma_x+\sigma_e & \beta_1\sigma_x \\ \beta_1\sigma_x & \sigma_x+\sigma_u \end{pmatrix}$. Since a bivariate normal distribution is characterized by only 5 parameters and the model (25) contains 6 unknown parameters, it is not identifiable and we need to impose some restrictions on these parameters to make it identifiable. A common such restriction is to assume that the ratio σ_e/σ_u is known. As in any other regression model, here also the main quantity of interest is the regression coefficient β_1.

There have been some attempts to develop robust parameter estimation under the above errors-in-variables model; for example, by Brown [11], Carroll and Gallo [12] and Zamar [50]. See Basu and Sarkar [6] for a brief discussion, who had demonstrated the superiority of the minimum disparity estimators over the others under contamination using the Hellinger distance. As we have seen in the case of previous two examples, some member of the proposed S-divergence family may lead to better robustness compared to the Hellinger distance (which is a member of PD family at $\lambda = -1/2$ and $\alpha = 0$), we can examine the same in case of the contaminated Gaussian errors-in-variables model also.

In this spirit, we have conducted a simulation study similar to that presented in [6]. In particular we generate 100 samples each of size $n = 20$ from the model (25) with true parameter values $\mu_x = 0$, $\sigma_x = 1$, $\sigma_u = \sigma_e = 0.25$, $\beta_0 = 0$ and $\beta_1 \sim \text{Uniform}[-5, 5]$. The values of β_1 has been chosen randomly for each sample following [50], which takes into account that the robustness of the resulting estimator might depend on the true value of β_1. The 5% contamination has been introduced into both the error variables u_i and e_i by observations from $N(0, \sigma^2)$ and $N(0, \tau^2)$ respectively. Also, for model identifiability, we assume that $\sigma_e/\sigma_u = 1$ is known.

For each sample, we compute the MSDE* of the 5 unknown parameters following the smoothed model approach. As in [6], we also use the bivariate Gaussian kernel $N_2(0, h^2 I)$ with the value of bandwidth $h = 0.5$. Since our model is bivariate normal with mean μ and variance matrix Σ, the resulting smoothed model is also a bivariate normal distribution with mean μ and variance matrix $(\Sigma + h^2 I)$. In particular, we consider the estimator $\hat{\beta}_{1j}$ of β_1 obtained from jth sample ($j = 1, \ldots, 100$) and derive the following performance measure of the estimator [50]

$$m = \sum_{j=1}^{100} \left(1 - \frac{|1 + \hat{\beta}_{1j}\beta_{1j}|}{(1 + \hat{\beta}_{1j}^2)^{1/2}(1 + \beta_{1j}^2)^{1/2}} \right),$$

where β_{1j} is the true value of β_1 in the jth sample. Clearly, smaller values of m indicate greater robustness of the estimator. We have considered several choices of the (σ, τ) combinations as in Zamar [50] and compute the above performance measure m for the MSDE* with different values of α and λ. However, for brevity in presentation, we only present the results for the choice $(\sigma, \tau) = (2, 2)$ in Table 9; the results for all other choices of (σ, τ) are generally similar except for changes in magnitude.

It is once again clear that the estimators with moderately large positive values of α and moderately large negative values of λ perform better or competitively in terms of robustness compared to the other estimators.

Table 9. The Empirical performance measure m for estimating β_1 under the error-in-variable model with $(\sigma, \tau) = (2, 2)$

λ	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.25$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1$
−1	0.136	0.133	0.106	0.089	0.086	0.088	0.080	0.080
−0.7	0.129	0.137	0.096	0.082	0.081	0.082	0.077	0.080
−0.5	0.147	0.128	0.091	0.088	0.084	0.090	0.081	0.080
−0.3	0.197	0.127	0.120	0.102	0.080	0.086	0.083	0.080
0	0.647	0.368	0.165	0.106	0.088	0.081	0.089	0.080
0.5	0.526	0.577	0.599	0.177	0.110	0.089	0.082	0.080
1	0.471	0.513	0.544	0.578	0.306	0.116	0.075	0.080
1.5	0.496	0.439	0.490	0.574	0.586	0.268	0.080	0.080
2	0.433	0.514	0.504	0.556	0.629	0.625	0.082	0.080

6. Higher order influence analysis

Lindsay [33] had observed that the usual first order influence function failed to capture the robustness of the minimum disparity estimators with large negative values of λ. The description of the previous section has demonstrated that this phenomenon can be a more general one, and is not restricted to estimators which have unbounded influence functions. It may fail to predict the strength of robustness of highly robust estimators, and it may also declare extremely unstable estimators as having good stability. As in [33], we consider a second order influence function analysis of the MSDEs and show that this provides a significantly improved prediction of the robustness of these estimators. See [33] and [7] for a general description of second order influence analysis in minimum disparity estimation.

Let G and $G_\epsilon = (1-\epsilon)G + \epsilon \wedge_y$ represent the true distribution and the contaminated distribution respectively, where ϵ is the contaminating proportion, y is the contaminating point, and $\wedge_y$ is a degenerate distribution with all its mass on the point y; let $T(G_\epsilon)$ be the value of the functional T evaluated at G_ϵ. The influence function of the functional $T(\cdot)$ is given by $T'(y) = \frac{\partial T(G_\epsilon)}{\partial \epsilon}|_{\epsilon=0}$. Viewed as a function of ϵ, $\Delta T(\epsilon) = T(G_\epsilon) - T(G)$ quantifies the amount of bias under contamination; under the first-order Taylor expansion the bias may be approximated as $\Delta T(\epsilon) = T(G_\epsilon) - T(G) \approx \epsilon T'(y)$. From this approximation, it follows that the predicted bias up to the first order will be the same for all functionals having the same influence function. Thus for the minimum S-divergence estimators, the first order bias approximation is not sufficient for predicting the true bias under contamination and hence not sufficient for describing the robustness of such estimators.

We consider the second order Taylor series expansion to get a second-order prediction of the bias curve as $\Delta T(\epsilon) = \epsilon T'(y) + \frac{\epsilon^2}{2} T''(y)$. The ratio of the second-order (quadratic) approximation to the first (linear) approximation, given by

$$\frac{\text{quadratic approximation}}{\text{linear approximation}} = 1 + \frac{[T''(y)/T'(y)]\epsilon}{2}$$

can serve as a simple measure of adequacy of the first-order approximation. Often when the first order approximation is inadequate, the second order approximation can give a more accurate prediction. If ϵ is larger than $\epsilon_{\text{crit}} = |\frac{T'(y)}{T''(y)}|$, the second-order approximation may differ by more than 50% compared to the first-order approximation. When the first order approximation is inadequate, such discrepancies will occur for fairly small values of ϵ.

In the following theorem, we will present the expression of our second order approximation $T''(y)$; for simplicity we will deal with the case of a scalar parameter. The next straightforward corollary gives the special case of the one parameter exponential family having unknown mean parameter.

Theorem 6.1. *Let $F_\theta \in \mathcal{F}$ with θ being a scalar parameter. Assume that the true distribution belongs to the model $\mathcal{F}$. For the minimum divergence estimator defined by the estimating equation* (17) *where the function $K(\delta)$ satisfies $K(0) = 0$ and $K'(0) = 1$, we have*

$T''(y) = T'(y)(\int u_\theta^2 f_\theta^{1+\alpha})^{-1}[m_1(y) + K''(0)m_2(y)]$, *where*

$$m_1(y) = 2\nabla u_\theta(y) f_\theta^\alpha(y) + 2\alpha u_\theta^2(y) f_\theta^\alpha - 2\int \nabla u_\theta f_\theta^{1+\alpha} - 2\alpha \int u_\theta^2 f_\theta^{1+\alpha} - T'(y)\left[(1+2\alpha)\int u_\theta^3 f_\theta^{1+\alpha} + 3\int u_\theta \nabla u_\theta f_\theta^{1+\alpha}\right],$$

$$m_2(y) = T'(y)\int u_\theta^3 f_\theta^{1+\alpha} - 2u_\theta^2(y) f_\theta^\alpha(y) + \frac{u_\theta(y) f_\theta^{\alpha-1}(y) - \int u_\theta f_\theta^{1+\alpha}}{u_\theta(y) f_\theta^\alpha(y) - \int u_\theta f_\theta^{1+\alpha}}.$$

In particular for the minimum S-divergence estimator, we have $K''(0) = A - 1$.

Corollary 6.2. *For the one parameter exponential family with mean* θ, *the above theorem simplifies to* $T''(y) = T'(y)[K''(0)Q(y) + P(y)]$ *with*

$$Q(y) = \left(\frac{u_\theta(y) f_\theta^{\alpha-1}(y) - \int u_\theta f_\theta^{1+\alpha}}{u_\theta(y) f_\theta^\alpha(y) - \int u_\theta f_\theta^{1+\alpha}}\right) + f_\theta^\alpha\left[\frac{(y-\theta)c_3}{c_2^2} - \frac{2(y-\theta)^2}{c_2}\right],$$

$$P(y) = \left(\frac{2c_0}{c_2} - 2\alpha\right) - f_\theta^\alpha\left[\frac{2}{c_2} + \frac{2\alpha(y-\theta)c_3}{c_2^2} - \frac{2\alpha(y-\theta)^2}{c_2}\right],$$

where $c_i = \int u_\theta^i f_\theta^{1+\alpha}$ *for* $i = 0, 1, 2, 3$. *If* y *represents an extreme observation with a small probability, the leading term of* $Q(y)$ *is dominant*.

Example 6.1 (Poisson mean). For a numerical illustration of the second order influence analysis, we consider the Poisson model with mean θ. This is a one-parameter exponential family so that we can compute the exact values of the second order bias approximation by using the above corollary. Also we can compute the first order approximation of bias by the expression of influence function from equation (18). For all our simulation results explained below, we have considered the true value of θ to be 4 and put a contamination at the point $y = 10$ which lies at the boundary of the 3σ limit for the mean parameter $\theta = 4$.

We have examined the relation between these two bias approximations for several different values of α and λ. In the following, we present some of our crucial findings through some graphical representations of the predicted biases. Figures 2, 3 and 4 contain the approximate bias plots for different α and $\lambda = 0$, $\lambda > 0$, $\lambda < 0$ respectively. In all the plots, the ϵ-axis runs from 0 to $\min(0.1, \epsilon_{\text{crit}})$, where ϵ_{crit} is the smallest value of ϵ where the second order approximation is double or half of the quantity predicted by the first order influence function for the first time. As this depends on the particular situation, all the plots in this section have different axis ranges.

Comments on Figure 2 ($\lambda = 0$): Clearly the two approximations coincide when $\alpha = 0$ and $\lambda = 0$ which generates the maximum likelihood estimator. This is expected from the theory of the MLE. However the difference between the predicted biases increase as α increases up to 0.5 and then the difference falls again and almost vanishes at $\alpha = 1$. In addition, the actual bias

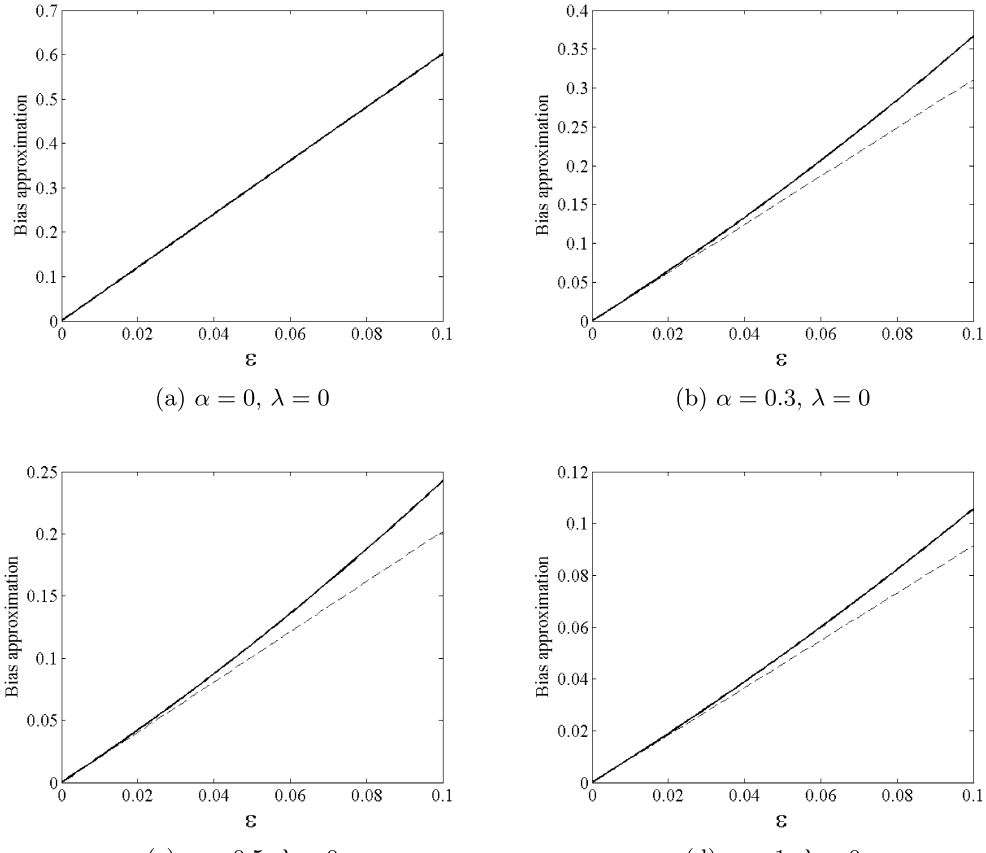
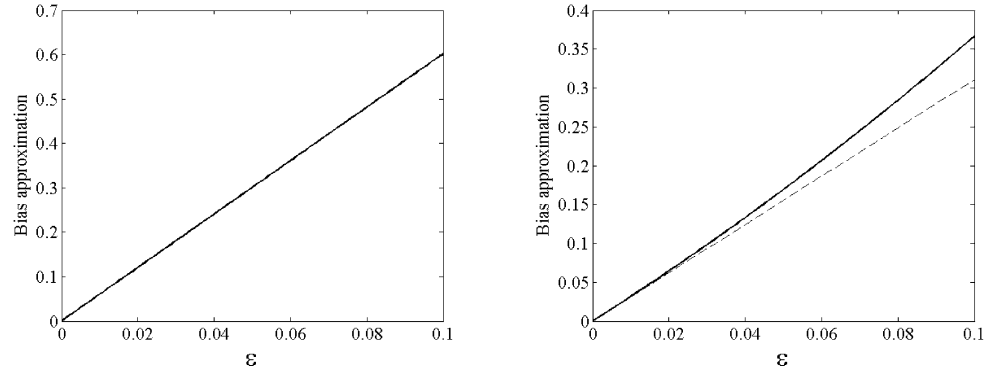


Figure 2. Bias approximations (dashed line: first order; solid line: second order) for different α and $\lambda = 0$ for the Poisson mean $\theta = 4$ and contamination at $y = 10$.

approximation generally drops with increasing α for both the approximations which should also be expected.

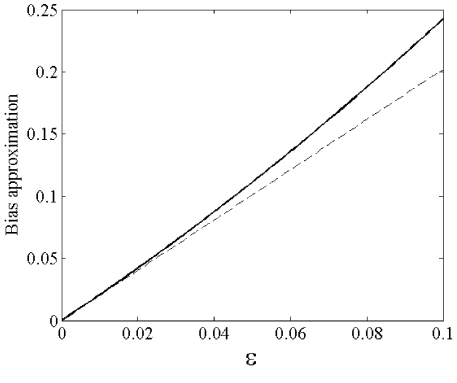
Comments on Figure 3 ($\lambda > 0$): For positive λ, the bias approximations are very different even for small values of α. As α increases the difference between the two bias approximations decreases. Here, $\epsilon = \epsilon_{\text{crit}}$ is the value of ϵ where the quadratic approximation is exactly double of the linear approximation for the first time. These estimators have weak stability properties in the presence of outliers, but the first order influence function approximation gives a false, conservative picture. We also note that this critical value of ϵ (ϵ_{crit}) also increases as α increases or λ decreases.

Comments on Figure 4 ($\lambda < 0$): Here also the plots are shown up to $\epsilon = \epsilon_{\text{crit}}$; in this case ϵ_{crit} is the value where the quadratic approximation drops to half of that of the linear approximation

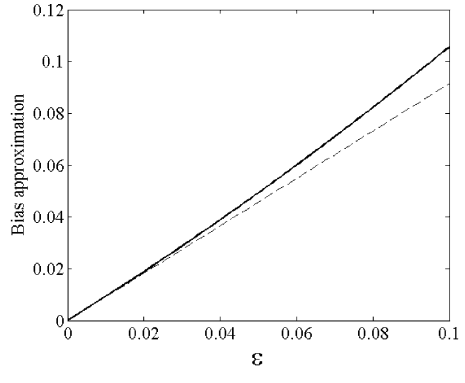


(a) $\alpha = 0, \lambda = 0.1$

(b) $\alpha = 0.5, \lambda = 0.1$



(c) $\alpha = 0, \lambda = 1$

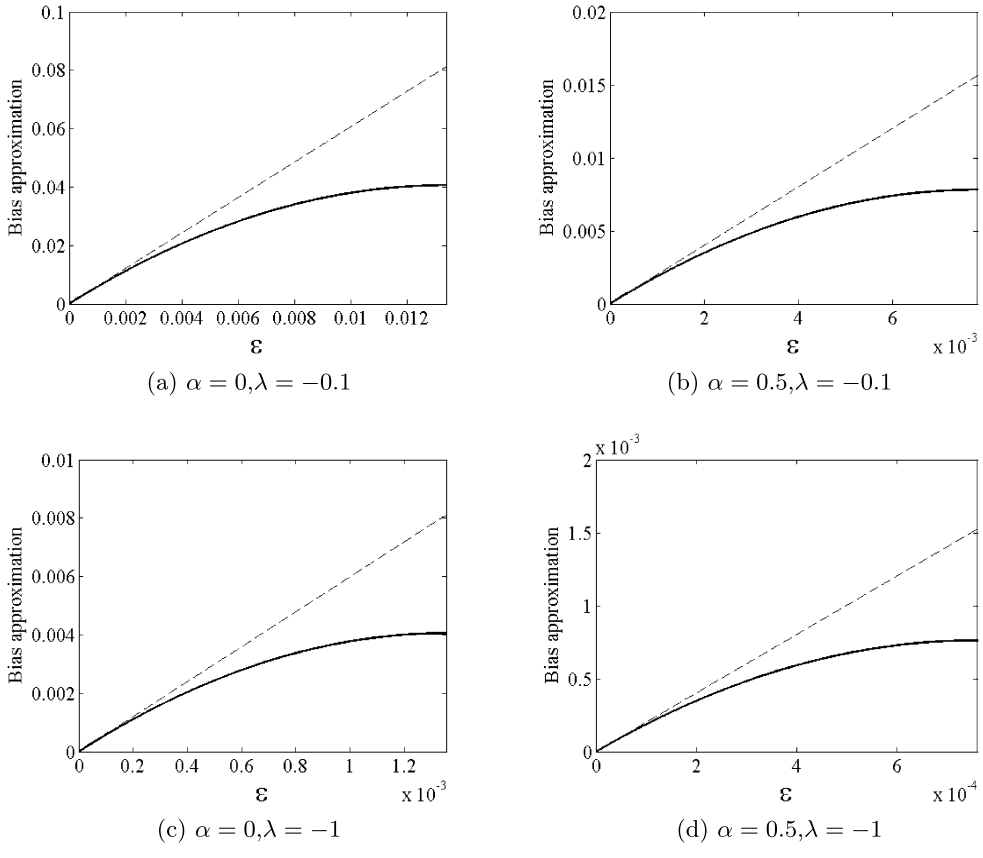


(d) $\alpha = 0.5, \lambda = 1$

Figure 3. Bias approximations (dashed line: first order; solid line: second order) for different α and $\lambda > 0$ for the Poisson model with mean $\theta = 4$ and contamination at $y = 10$.

for the first time. Here the estimators have strong robustness properties, but the influence function gives a distorted negative view. Contrary to the positive λ case, here this critical value ϵ_{crit} increases as both α or λ increases.

While the three figures are drawn for zero, positive and negative values of λ, comparing of specific panels over the three different figures may show the predicted behavior of the procedure over different λ for the same value of α. For example, consider Figures 2(c), 3(b) and 4(b). As the first order influence function is independent of λ, the corresponding predicted behavior in these three panels are identical with α fixed at 0.5. As we have seen in Tables 3–6, the actual behaviors of these estimators are widely different. This is recognized to a large extent in the second order influence curves, which shrinks the first order prediction for $\lambda < 0$ and inflates it for $\lambda \geq 0$ (more so for $\lambda > 0$). The first order prediction, however, makes no such distinction.



(a) $\alpha = 0, \lambda = -0.1$ (b) $\alpha = 0.5, \lambda = -0.1$

(c) $\alpha = 0, \lambda = -1$ (d) $\alpha = 0.5, \lambda = -1$

Figure 4. Bias approximations (dashed line: first order; solid line: second order) for different α and $\lambda < 0$ for the Poisson model with mean $\theta = 4$ and contamination at $y = 10$.

We trust that the above gives a fairly comprehensive picture of the limitation of the first order influence function in the present context. We can say that for any $\lambda \neq 0$, this critical value ϵ_{crit} of ϵ where the quadratic approximation is double or half of the linear approximation for the first time increases as α increases or $|\lambda|$ decreases. Table 10 presents the value of ϵ_{crit} for several combinations of λ and α. These values are increasing with α in either case.

7. On the choice of tuning parameters α and λ

In the previous sections, we have observed that the robustness of the proposed minimum S-divergence estimators (MSDE) differ widely for different values of the tuning parameters α and λ, although their first order influence function is independent of λ. This is clearly observed in all the simulation exercises presented in this paper and other relevant numerical illustrations provided in [21,22]. In fact, the MSDEs at $\alpha = 0$ or low values close to zero are seen to be highly

Table 10. The minimum values of the contamination proportion ϵ for which the ratio of the second order bias approximation over the first order is 2 (for $\lambda > 0$) or $\frac{1}{2}$ (for $\lambda < 0$)

α	$\lambda = -1$	$\lambda = -0.5$	$\lambda = -0.1$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$
0	0.0020	0.004	0.020	0.040	0.008	0.004
0.1	0.0020	0.004	0.023	0.043	0.009	0.005
0.2	0.0025	0.005	0.027	0.048	0.010	0.005
0.3	0.0030	0.006	0.032	0.056	0.012	0.006
0.4	0.0035	0.007	0.040	0.067	0.015	0.008
0.5	0.0050	0.009	0.052	0.087	0.019	0.010
0.6	0.0070	0.014	0.073	0.121	0.026	0.130
0.7	0.0110	0.022	0.114	0.191	0.041	0.021
0.8	0.0200	0.040	0.211	0.363	0.077	0.039

affected by the choice of the value of λ under the presence of outliers. However, the asymptotic relative efficiency of the MSDEs as obtained from their asymptotic variances is independent of λ and decreases as α increases. Therefore, a proper choice of the tuning parameters α and λ is necessary when applying the proposed MSDEs in any practical problem.

We note that the tuning parameters that appear to produce the most accurate estimators are not exactly the same in the different set ups and models considered in Section 5. In general, we observe that for a simple unidimensional model like the Poisson and for the type of contaminations considered, values of α in the neighborhood of 0.25 and values of λ in the range $[-0.3, -0.7]$ appear to provide the best compromise. For the normal mixture model and the errors in variables model, slightly larger values of α are necessary to generate estimators with a reasonable level of robustness. As a fixed overall recommendation, we feel values of α around 0.5 and values of λ in $[-0.3, -0.7]$ might provide the best overall compromise.

However, an adaptive method, on the basis of observed data, that provides an optimal choice of the couple (α, λ) would be interesting and useful for the practitioners for implementing our proposal in real life. One such popular adaptive method in the related context of the optimum selection of α in the minimum DPD estimation suggests the minimization of the estimated mean square error (MSE) of the estimator [23,31,48]. We call it the Warwick and Jones approach. We can easily extend that approach to choose the optimum (α, λ) couple in our proposed minimum S^* divergence estimators. Suppose θ^0 denotes the true target parameter value. Suppose we have the contaminated density $g(x) = (1-\epsilon) f_{\theta^0}(x) + \epsilon \delta_y(x)$, where $\delta_y(x)$ is the Dirac delta function at the point y. Let us define $\theta_{\alpha,\lambda} = \arg\min_\theta S_{(\alpha,\lambda)}(g, f_\theta)$ for any given (α, λ) combination. Given the observed sample, let the MSDE of the parameter θ with tuning parameter (α, λ) be denoted as $\hat{\theta}_{\alpha,\lambda}$. Then, in the Warwick and Jones [48] approach, we should chose the optimum (α, λ) combination by minimizing an estimate of the asymptotic approximation of $E[(\hat{\theta}_{\alpha,\lambda} - \theta^0)^T (\hat{\theta}_{\alpha,\lambda} - \theta^0)]$, the summed MSE of the MSDE $\hat{\theta}_{\alpha,\lambda}$ of θ. Using the asymptotic distribution of the MSDE under discrete model as discussed in Section 4.2, we can easily obtain the required asymptotic approximation as

$$E\left[\left(\hat{\theta}_{\alpha,\lambda} - \theta^0\right)^T \left(\hat{\theta}_{\alpha,\lambda} - \theta^0\right)\right] = \left(\theta_{\alpha,\lambda} - \theta^0\right)^T \left(\theta_{\alpha,\lambda} - \theta^0\right) + \frac{1}{n} \operatorname{trace}\left\{J_g^{-1} V_g J_g^{-1}\right\}, \tag{26}$$

where J_g and V_g are the matrices as defined in Section 4.2. In case of continuous models, we should use J_g^* and V_g^* defined in Section 4.3 in place of J_g and V_g respectively, which was also indicated briefly in [22]. Note that, if we ignore the first bias term in the above approximation (26), the procedure of selecting the optimum tuning parameter coincide with the proposal of Hong and Kim [31].

Clearly, one can easily estimate the second component in the approximation (26) by putting $\hat{\theta}_{\alpha,\lambda}$ for $\theta_{\alpha,\lambda}$ and the empirical estimates of the true density g from the observed data (through the relative frequency or the kernel density estimator for discrete and continuous models respectively). For the first component in (26), one can again estimate $\theta_{\alpha,\lambda}$ by the observed MSDE $\hat{\theta}_{\alpha,\lambda}$; but there is clearly no direct estimate for θ^0. Warwick and Jones [48] took various possible estimators θ^0 which they referred to as the "pilot estimators" and compared these choices through an extensive simulation study for the MDPDE. They recommended the choice of the "pilot estimator" which is the MSDE (or the MDPDE) corresponding to $\alpha = 1$, that is, the minimum L_2-divergence estimator. We have followed the Warwick and Jones proposal in this paper. In repeated simulations involving the Poisson model, we have observed the following:

(a) When the data come from the pure distributions (as used in Section 5) we have found that a large proportion of cases choose $\alpha = 0$, $\lambda = 0$ as the optimal parameter set. Ghosh and Basu [22] reported similar simulation findings in case of the choice of optimal tuning parameters for the use of the S-divergence in continuous models.
(b) When the data are chosen from contaminated Poisson mixtures, very often the optimal tuning parameter set has a moderately large values of α and a moderately large negative value of λ. This again indicates the usefulness of minimum S-divergence estimators outside the PD–DPD class.

It is also worthwhile to note that for Short's data and Newcomb's data, two famous data sets presented in [43] which contain moderate to large outliers, the optimal (α, λ) combinations turn out to be $(0.8, -0.3)$ and $(0.6, -0.4)$, as reported by in [22] for the normal model. In either case, the optimal set is far off from the PD–DPD class.

For the set ups in Sections 5.2 and 5.3, we believe we will need more research to make more clear cut and pinpointed recommendations about the choice of the tuning parameter. Apart from the complicated nature of the models, this is also due to the presence of a third tuning parameter, the bandwidth. The actual choice of the bandwidth in an optimal manner will probably require a substantial amount of additional research. However, we are able to confirm that larger values of the bandwidth appear to generate more efficient estimators, while smaller values produce more robust ones.

8. Two interesting special cases

8.1. Asymptotic breakdown point of the MSDEs under the location model

Now we will establish the breakdown point of the minimum S-divergence functional $T_{\alpha,\lambda}(G)$ under the location family of densities $\mathcal{F}_L = \{f_\theta(x) = f(x-\theta) : \theta \in \Theta\}$. Note that $\int f^{1+\alpha}(x-\theta)\,dx = \int f^{1+\alpha}(x)\,dx = M_f^\alpha$, say, which is independent of the parameter θ. Recall that we can

write the S-divergence as $S_{(\alpha,\lambda)}(g, f) = \int f^{1+\alpha} C_{(\alpha,\lambda)}(\delta)$ where $\delta = \frac{g}{f} - 1$ and $C_{(\alpha,\lambda)}(\delta) = \frac{1}{AB}[B - (1+\alpha)(\delta+1)^A + A(\delta+1)^{1+\alpha}]$. Now $C_{(\alpha,\lambda)}(-1) = \frac{1}{A}$ which is clearly bounded for all $A \neq 0$. Define $D_{(\alpha,\lambda)}(g, f) = f^{1+\alpha} C_{(\alpha,\lambda)}(\frac{g}{f} - 1)$. Then note that whenever $A > 0$ and $B > 0$, we have $D_{(\alpha,\lambda)}(g, 0) = \lim_{f\to 0} D_{(\alpha,\lambda)}(g, f) = \frac{1}{B} g^{1+\alpha}$. Our subsequent results will be based on the next lemma which follows from Holder's inequality.

Lemma 8.1. *Assume that the two parameters α and λ are such that both A and B are positive. Then for any two densities g, h in the location family $\mathcal{F}_L$ and any $0 < \epsilon < 1$, the integral $\int D_{(\alpha,\lambda)}(\epsilon g, h)$ is minimized when $g = h$.*

Consider the contamination model $H_{\epsilon,n} = (1-\epsilon)G + \epsilon K_n$, where $\{K_n\}$ is a sequence of contaminating distributions. Let $h_{\epsilon,n}$, g and k_n be the corresponding densities. We say that there is breakdown in $T_{\alpha,\lambda}$ for ϵ level contamination if there exists a sequence K_n such that $|T_{\alpha,\lambda}(H_{\epsilon,n}) - T(G)| \to \infty$ as $n \to \infty$. We write below $\theta_n = T_{\alpha,\lambda}(H_{\epsilon,n})$ and assume that the true distribution belongs to the model family, i.e., $g = f_{\theta^g}$. We make the following assumptions:

(BP1) $\int \min\{f_\theta(x), k_n(x)\} \to 0$ as $n \to \infty$ uniformly for $|\theta| \leq c$ for any fixed c. That is, the contamination distribution is asymptotically singular to the true distribution and to specified models within the parametric family.

(BP2) $\int \min\{f_{\theta^g}(x), f_{\theta_n}(x)\} \to 0$ as $n \to \infty$ if $|\theta_n| \to \infty$ as $n \to \infty$, that is, large values of θ give distributions which become asymptotically singular to the true distribution.

(BP3) The contaminating sequence $\{k_n\}$ is such that

$$S_{(\alpha,\lambda)}(\epsilon k_n, f_\theta) \geq S_{(\alpha,\lambda)}(\epsilon f_\theta, f_\theta) = C_{(\alpha,\lambda)}(\epsilon - 1) M_f^\alpha$$

for any $\theta \in \Theta$ and $0 < \epsilon < 1$ and $\limsup_{n\to\infty} \int k_n^{1+\alpha} \leq M_f^\alpha$.

Theorem 8.2. *Assume that the two parameters α and λ are such that both A and B are positive. Then under the assumptions* (BP1)–(BP3) *above, the asymptotic breakdown point ϵ^* of the minimum S-divergence functional $T_{\alpha,\lambda}$ is at least $\frac{1}{2}$ at the (location) model.*

Corollary 8.3 (Density power divergence). *The well-known Density Power divergence (DPD) belongs to the S-divergence family (for $\lambda = 0$) for which $A = 1 > 0$ and $B = \alpha > 0$ for all $\alpha > 0$. Thus under the assumptions* (BP1)–(BP3), *the MDPDE for $\alpha > 0$ has breakdown point of $\frac{1}{2}$ at the location model of densities.*

Remark 8.1. Whenever the contaminating densities $\{k_n\}$ belong to the location model under consideration then one can easily check that Assumption (BP3) holds; its second part holds trivially whereas the first part holds by Lemma 8.1.

Further, if $k_n = f_{\theta_n}$ with $|\theta_m| \to \infty$ as $n \to \infty$, then Assumption (BP2) implies assumption (BP1). In particular, for the normal location model, Assumption (BP2) holds trivially and so under model contamination Assumption (BP1) also holds.

Our breakdown results of this section provide a remarkable match with the empirical findings about the performance of the minimum S-divergence estimators as observed in Section 5. Notice

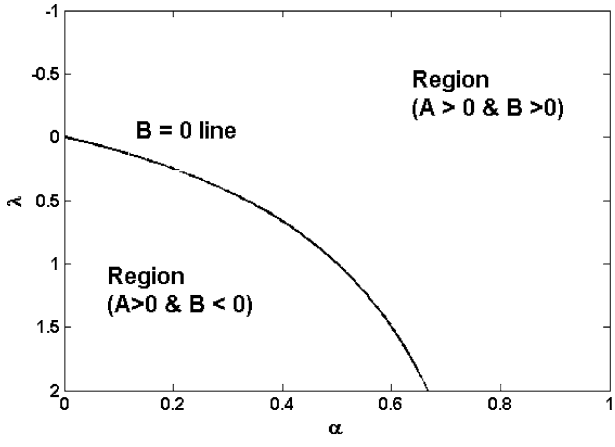


Figure 5. The region of positive A and B in the (α, λ)-plane.

that the breakdown result of Theorem 8.8 requires that both A and B should be strictly greater than zero. In Figure 5, we have represented the rectangular span of the combination of tuning parameters α and λ given by $[0, 1] \times [-1, 2]$. This has been the region of our interest in the simulations of Section 5. The subregion of this rectangle where $A > 0$, $B > 0$ condition is violated is given in the lower left hand corner of this region. What is striking is that this subregion matches exactly with the region of instability observed under data contamination in Tables 3–6 in Section 5.

8.2. The *S*-hellinger distance: Estimating multivariate location and covariance

Beran [8] considered minimum Hellinger distance estimation for univariate parametric models and established the efficiency and demonstrated the robustness of the corresponding minimum distance estimator. Tamura and Boos [44] extended his work on minimum Hellinger distance estimation to the case of multivariate location and covariance estimation under the class of elliptically symmetric models. In the subsequent literature, there has been many attempts to generalize the Hellinger distance suitably to extract some more desirable properties along with greater robustness; see, for example, [1,41]. However, most, if not all, such generalizations are in the case of univariate models.

The S-divergence family developed here provides us one such generalization of the Hellinger distance, namely the S-Hellinger family, briefly described in Section 3; it is the subfamily of S-divergences corresponding to the parameter $\lambda = -\frac{1}{2}$ that connects the ordinary Hellinger distance to the L_2-divergence smoothly through the parameter α.

Here we explore the properties of the S-Hellinger distance family described briefly in Section 3, to generate a class of robust estimators of multivariate location and covariances. Thus, our work will generalize the work of Tamura and Boos [44] from the Hellinger distance to the case

of the general family of S-Hellinger distances. Note that just as the ordinary Hellinger distance represents the self adjoint member of the PD family in the sense of [32], any other cross section of the class of S-divergences for a fixed value α also has a self adjoint member in the S-Hellinger family $S_{(\alpha,-1/2)}$.

8.2.1. *The S-hellinger distance* (*SHD*) *family*

The S-Hellinger Distance (SHD) family between two generic probability density functions f and g is defined in terms of the tuning parameter α as

$$\begin{aligned}\mathrm{SHD}_\alpha(g, f) = S_{(\alpha,-1/2)} &= \frac{2}{1+\alpha}\left[\int f^{1+\alpha} - 2\int f^{\frac{1+\alpha}{2}} g^{\frac{1+\alpha}{2}} + \int g^{1+\alpha}\right] \\ &= \frac{2}{1+\alpha}\int \left(g^{\frac{1+\alpha}{2}} - f_\theta^{\frac{1+\alpha}{2}}\right)^2 = \frac{2}{1+\alpha}\left\| g^{\frac{1+\alpha}{2}} - f_\theta^{\frac{1+\alpha}{2}}\right\|_2^2,\end{aligned} \tag{27}$$

where $\|\cdot\|_2$ represents the using the L_2-norm. Although $\mathrm{SHD}_\alpha(g, f)$ in (27) is not itself a distance, it corresponds to the distance $[\frac{1+\alpha}{2}\mathrm{SHD}_\alpha(g, f)]^{\frac{1}{2}}$. Note that, for $\alpha = 0$ the measure reduces to the ordinary (twice squared) Hellinger distance.

We will use the S-Hellinger distance for doing multivariate location and covariance estimation because it is easily amenable to this unlike other members of S-divergence. These properties mainly come from the fact that it corresponds to a distance metric.

Lemma 8.4. *Let f_1, f_2, f_3 be three densities with respect to some common dominating measure. Then for any $0 \le s \le 1$, we have*

$$\mathrm{SHD}_\alpha\big((1-s)f_1 + sf_2, f_3\big) \le (1-s)\mathrm{SHD}_\alpha(f_1, f_3) + s\mathrm{SHD}_\alpha(f_2, f_3).$$

Lemma 8.5. *Let f_1, f_2, f_3 be three densities with respect to some common dominating measure. Then we have*

$$\mathrm{SHD}_\alpha(f_1, f_2) \le 2\big[\mathrm{SHD}_\alpha(f_1, f_3) + \mathrm{SHD}_\alpha(f_2, f_3)\big].$$

8.2.2. *The minimum SHD estimator* (*MSHDE*) *of multivariate location and covariance*

Now let us consider the usual (multivariate) set-up of parametric estimation; we have n independent and identically distributed observations $X_1, X_2, \ldots, X_n$ from a (multivariate) distribution G, having density g with respect to a dominating measure ν. We will model the true density by a parametric (multivariate) model $\mathcal{F}_d = \{f_\theta : \theta \in \Theta_0 \subseteq \mathbb{R}^p\}$; our interest here is to estimate the parameter θ based on the observed data. We will define the measure of discrepancy between the sample data and the model family by the S-Hellinger distance measure between the model density f_θ and a non-parametric density estimator $\hat{g}_n$. Thus the Minimum SHD Estimator (MSHDE) is given by

$$\begin{aligned}\hat{\theta}_n = T(\hat{G}_n) &= \arg\min_{\theta\in\Theta} \frac{2}{1+\alpha}\left\| \hat{g}_n^{\frac{1+\alpha}{2}} - f_\theta^{\frac{1+\alpha}{2}}\right\|_2^2 \\ &= \arg\min_{\theta\in\Theta}\left\| \hat{g}_n^{\frac{1+\alpha}{2}} - f_\theta^{\frac{1+\alpha}{2}}\right\|_2,\end{aligned} \tag{28}$$

where $\hat{g}_n$ is the kernel density estimator given by

$$\hat{g}_n(x) = \frac{1}{nh_n^p} \sum_{i=1}^{n} w\left(\frac{x - X_i}{h_n}\right), \tag{29}$$

with $w(\cdot)$ being a p-variate density function and $\{h_n\}$ being a suitable sequence of bandwidths; $\hat{G}_n$ is the distribution function corresponding to the density $\hat{g}_n$.

In this section, we will consider only the minimum SHD estimator of multivariate location and covariance under suitable assumptions and describe some important properties of it. We will restrict ourselves only to parametric families having elliptically symmetric distribution with density function given by

$$f_\theta(x) \propto |\Sigma|^{-1/2} \psi\big\{(x-\mu)^T \Sigma^{-1}(x-\mu)\big\}, \tag{30}$$

where ψ is an univariate density symmetric around zero, $\theta = (\mu, \Sigma)$, $\Theta = \mathbb{R}^p \times \mathcal{S}$, with $\mathcal{S}$ being the set of all $p \times p$ positive definite matrices.

The influence function of the minimum SHD functional can be derived directly from the results of Section 4.1. As a particular example, we will consider the case of The elliptically symmetric density of p-variate normal with mean vector μ and covariance matrix Σ. The influence function of the MSHDE of $\theta = (\mu, \Sigma)$ can be written in terms of that of the individual parameters, namely $T_\alpha^\mu(F_\theta)$ and $T_\alpha^\Sigma(F_\theta)$ of μ and Σ, as $\mathrm{IF}(y; T_\alpha, F_\theta) = (\mathrm{IF}(y; T_\alpha^\mu, F_\theta), \mathrm{IF}(y; T_\alpha^\Sigma, F_\theta))$. Also, it is easily seen that the influence function of the MSHDE of location μ is given by

$$\mathrm{IF}\big(y; T_\alpha^\mu, F_\theta\big) = (1+\alpha)^{\frac{p}{2}+1}(y-\mu)e^{-\frac{\alpha}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)},$$

and that of the covariance Σ is given by

$$\begin{aligned}&\mathrm{IF}\big(y; T_\alpha^\Sigma, F_\theta\big)\\ &\quad= \zeta_{\alpha,p}\big[(1+\alpha)^{\frac{p}{2}+1}\big\{(y-\mu)(y-\mu)^T\Sigma^{-1} - p\big\}e^{-\frac{\alpha}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)} - \big(1-p(1+\alpha)\big)\big]\Sigma,\end{aligned}$$

with $\zeta_{\alpha,p} = 2(1+\alpha)[3 - 2p(1+\alpha) + p^2(1+\alpha)^2]^{-1}$.

8.2.3. *Equivariance*

An estimator of (multivariate) location and covariance can potentially be used to infer about the orientation and shape of data-points in a multi-dimensional space. Since there is no single universal way of measuring data, a minimal requirement for these estimators is that they should be independent of the coordinate system. So for the estimation of multivariate location and covariance, we prefer estimators that are affine equivariant and affine covariant respectively; see [44] for relevant definitions. We can derive some sufficient conditions, under which the Minimum SHD estimator of multivariate location and covariance satisfies the requirement of equivariance.

Lemma 8.6. *Suppose that*:

1. *the model family is chosen to be elliptical having density of the form* (30) *and*
2. *the kernel density estimator satisfies the condition* $\hat{g}_{AX+b,n}(x) = \hat{g}_{X,n}(A^{-1}(x-b))/|A|$.

Then the minimum SHD estimator of multivariate location and covariance are affine equivariant and affine covariant, respectively.

The proof is straightforward and is omitted. In a practical situation, it is necessary to choose a kernel density estimator satisfying condition (8.6). We can construct a density estimate of the radial type satisfying (8.6), whenever we have an initial affine covariant estimate $\hat{\Sigma}^0$ of covariance based on the observed data; this will be of the form

$$\hat{g}_n(x) = \frac{1}{nh_n^p |\hat{\Sigma}^0|^{1/2}} \sum_{i=1}^{n} w\big(h_n^{-1} \|x - X_i\|_{\hat{\Sigma}^0}\big), \tag{31}$$

where $\|x\|_{\Sigma}^2 = x^T \Sigma^{-1} x$. Then the MSHDE of the location and covariance using this kernel density estimate will be both affine equivariant and affine covariant within the multivariate elliptical family (30).

8.2.4. *Consistency*

Suppose the observed sample data $X_1, X_2, \ldots, X_n$ come from the true common density g. Then the minimum SHD estimator can be shown to be consistent under suitable conditions on the Kernel density estimator $\hat{g}_n$ and the model family $f_\theta, \theta \in \Theta$.

Theorem 8.7. *Suppose that, with probability one, the minimum SHD estimator $T(\hat{G}_n)$ exists for all sufficiently large n and as $n \to \infty$,*

$$\big\| \hat{g}_n^{\frac{1+\alpha}{2}} - f_\theta^{\frac{1+\alpha}{2}} \big\|_2 \to 0 \quad \textit{and} \tag{32}$$

$$\big\| f_{\theta_n}^{\frac{1+\alpha}{2}} - f_\theta^{\frac{1+\alpha}{2}} \big\|_2 \to 0 \quad \Rightarrow \quad \theta_n \to \theta, \tag{33}$$

for any sequence $\{\theta_n : \theta_n \in \Theta\}$. Then we have, with probability one,

$$T(\hat{G}_n) \to \theta \qquad \textit{as } n \to \infty. \tag{34}$$

Note that condition (32) needed for the consistency of the minimum SHD estimator, simply says that the kernel density estimator $\hat{g}_n$ converges to the true density g in the metric corresponding to the SHD. This is a very intuitive assumption for the consistency and other asymptotic properties of the any minimum distance estimator. However, in this case, it can be proved using the uniform convergence or by L_1 convergence of the kernel estimator under some mild assumptions on the corresponding bandwidth sequence. Condition (33) needed for the consistency of the minimum SHD estimator, is a model specific condition. It is expected to hold for many elliptically symmetric densities. We have checked that this holds for the multivariate normal model.

Deriving the complete asymptotic results for the minimum S-Hellinger distance estimator in case of multivariate location and covariance could pose some challenges as it would involve the convergence properties of a kernel in a multivariate setting. We will take up the actual asymptotic distribution of these estimators in our future work. In case of the DPD, the corresponding

minimum divergence estimators are multivariate M-estimators of location and covariance which should be relatively easier to handle. Some relevant results in this connection are presented in [29]. Toma and Leoni-Aubin [46] present some divergence based estimators of multivariate location and covariance. Note, however, that the DPD and the S-Hellinger distance families have no overlap except at the L_2 distance.

Using robust data reduction techniques prior to the analysis of multivariate location and covariance could be another approach to tackle this problem. See the recent monograph on robust dimension reduction by [19] for a general discussion of this area.

A small simulation example illustrating the robustness of the minimum S-Hellinger distance estimators has been provided in the Supplementary Material [27].

8.2.5. *Breakdown point*

We now consider the breakdown point of the minimum S-Hellinger distance functional in the case of the estimation of multivariate location and covariance as a measure of its robustness. Let X represent the original data set of a given size n and Y be a contaminating data set of size m ($m \leq n$). The estimator $\hat{\theta}_n$ will be said to break down if, through proper choice of the elements of the data set Y, the difference $\hat{\theta}_{m+n}(X \cup Y) - \hat{\theta}_n(X)$ can be made arbitrarily large. If m^* is the smallest number of the contaminating values for which the estimator breaks down, then the breakdown point of the corresponding estimator at X is $\frac{m^*}{(m+n)}$. In the case of multivariate location and covariance, the breakdown point of the joint estimation of location and covariance has been defined by [18] through the following measure of discrepancy

$$B(\theta_1, \theta_2) = \operatorname{tr}\bigl(\Sigma_1 \Sigma_2^{-1} + \Sigma_1^{-1} \Sigma_2\bigr) + \|\mu_1 - \mu_2\|^2 \tag{35}$$

between parameter values $\theta_1 = (\mu_1, \Sigma_1)$ and $\theta_2 = (\mu_2, \Sigma_2)$, where tr represents the trace of a matrix and $\|\cdot\|$ represents the Euclidean norm. The joint estimate of multivariate location and covariance will break down when the supremum of the discrepancy, as given in the above equation, between the pure data estimate at X and the contaminated data estimate at $X \cup Y$, is infinite.

We will derive the breakdown point of the minimum SHD estimators under specific assumptions on the true and model densities. Let $\mathcal{G}^*$ denote a subclass of all probability densities satisfying

$$\mathrm{SHD}_\alpha(g, f) \leq \kappa(\alpha), \qquad \text{for all } g, f \in \mathcal{G}^*, \tag{36}$$

where $\kappa(\alpha)$ is a finite positive bound depending only on α. Then, we will assume that the model family $\mathcal{F}_d$ and true density g both belong to the family $\mathcal{G}^*$. Also let $\hat{g}_n(c_n) = \hat{g}_n(c_n, x)$ be the kernel density having bandwidth c_n and $f_{\theta_n}(c_n)$ denotes the model density nearest to the above kernel estimator in terms of the S-Hellinger distance. We assume that such a model density exists. Define

$$\begin{aligned} a_{n,m} &= \mathrm{SHD}_\alpha\bigl(\hat{g}_n(c_{n+m}), f_{\theta_n}(c_n)\bigr), \quad \text{and} \\ v^* &= \liminf_{\theta_1, \theta_2} \mathrm{SHD}_\alpha(f_{\theta_1}, f_{\theta_2}), \qquad \text{where the limit is taken as } B(\theta_1, \theta_2) \to \infty. \end{aligned} \tag{37}$$

Theorem 8.8. *Assume that the model family $\mathcal{F}_d$ and true density g both belong to $\mathcal{G}^*$. Then, the minimum SHD estimator $\theta_n(c_n)$ has the breakdown point $\varepsilon^*(\theta_n(c_n))$ satisfying*

$$\varepsilon^*\big(\theta_n(c_n)\big) \geq \left[\frac{\frac{v^*}{4} - a_{n,m}}{\kappa(\alpha) - a_{n,m}}\right]. \tag{38}$$

Corollary 8.9. *Let g be the true density. If $\mathrm{SHD}_\alpha(g, f_{\theta_n}(c_n)) \to a$ and $\hat{g}_{n+m}(c_{n+m}) \to g(x)$, for each x, we have $a_{a,m} \to a$ almost surely. Then, by the theorem,*

$$\liminf_{n\to\infty} \varepsilon^*\big(\theta_n(c_n)\big) \geq \left[\frac{\frac{v^*}{4} - a}{\kappa(\alpha) - a}\right]. \tag{39}$$

Further, if the true distribution belongs to the model family, that is, $g = f_\theta$ for some $\theta \in \Theta$, we get $a = 0$ and then

$$\liminf_{n\to\infty} \varepsilon^*\big(\theta_n(c_n)\big) \geq \frac{v^*}{4\kappa(\alpha)}. \tag{40}$$

Remark. Whenever $v^* = \kappa(\alpha)$ and the true distribution belongs to the model family, the corollary yields

$$\liminf_{n\to\infty} \varepsilon^*\big(\theta_n(c_n)\big) \geq \frac{1}{4}.$$

That is, breakdown cannot occur for the minimum SHD estimator in this case for $\varepsilon < \frac{1}{4}$. The result is remarkable since the breakdown bound is independent of the dimension of the data unlike the shrinking bounds offered by the M-estimator.

9. Role of the *S*-divergence in robust hypothesis testing

The S-divergence measure and the minimum S-divergence estimators can also be applied to construct robust tests for statistical hypothesis. Such divergence based tests have been attempted in case of the power divergences [42] and the density power divergences [4,5], which can be further extended to generate a larger class of test statistics using the family of S-divergences.

As a brief introduction, let us consider the set-up of Section 4 with an i.i.d. sample of size n. Fix $\theta_0 \in \Theta$. Consider the problem of testing the simple null hypothesis given by

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \neq \theta_0. \tag{41}$$

Under H_0, if the model is correctly specified, f_{θ_0} is the true data generating density. Therefore, a class of test statistics for testing (41) can be constructed by using the S-divergence measure between f_{θ_0} and $f_{\hat{\theta}}$ for any estimator $\hat{\theta}$ of θ based on the observed sample. And, an ideal choice for $\hat{\theta}$ in this case should be a minimum S-divergence estimator so that the test statistics has the form

$$T^{(1)}_{\gamma,\lambda}(\hat{\theta}_{\beta,\tau}, \theta_0) = 2n S_{(\gamma,\lambda)}(f_{\hat{\theta}_{\beta,\tau}}, f_{\theta_0}),$$

where $\hat{\theta}_{\beta,\tau}$ is the minimum S-divergence estimator (MDPDE) of θ with tuning parameter β and τ and $S_{(\gamma,\lambda)}(\cdot,\cdot)$ is the S-divergence measure with parameter γ and λ.

Ghosh, Basu and Pardo [25] proved several asymptotic and robustness properties of the above S-divergence based test statistics with $\tau = 0$. The main advantage of taking $\tau = 0$ is that the MDPDE $\hat{\theta}_{\beta,0}$ does not require the consideration of kernel density estimator and hence avoids all the associated complications in asymptotic derivations. However, they have argued that, even if the general MSDE $\hat{\theta}_{\beta,\tau}$ were used, under the set-up of discrete models the asymptotic null distribution would remain the same. See [25] for more detailed derivations and examples. Interestingly, the robustness properties of the proposed test procedure have been observed to depend directly on the robustness of the minimum S-divergence estimator used.

Similar classes of S-divergence based test statistics can also be constructed for testing composite hypothesis problems or for testing the two sample problems extending the ideas of [4,5]. This problem has also been considered by the first author in his doctoral dissertation and are available in [24].

The robustness properties of similar test statistics derived from robust M-estimators have been pointed out in [30]. The works of [15], [16] and [34] related to scoring rules also suggest some other robust tests of parametric statistical hypothesis which are similar in spirit to our work.

10. Concluding remarks

In this paper, we have described the development of a large family of density based divergences which includes both the classes of power divergences and density power divergences as special cases. The family gives the data analyst a large number of choices of possible divergences to apply in the minimum distance estimation context. Several members of the family are distinguished by their strong robustness properties, and many of them generate estimators with high asymptotic efficiency. The best performer within the minimum divergence class induced by the S-divergences is often outside the PD–DPD class, underscoring the utility of the family. The performance of the method is demonstrated on very simple everyday use models like the Poisson, as well as more complicated and sophisticated models like the normal mixture model and normal errors-in-variables model. The family is indexed by two parameters, only one of which shows up in the influence function and the asymptotic efficiency expressions. Yet both the tuning parameters have important roles in actual finite sample efficiencies and the robustness of the estimators. The behavior of the estimators within this family clearly show the limitation of the influence function as a measure of robustness; we demonstrate that a second order influence analysis could be a much more accurate predictor of the robustness of these estimators (or lack of it). The instability of the estimators under contamination also appears to be linked closely to the violation of the breakdown conditions. A particular subfamily of this class can be used to construct robust high breakdown equivariant estimators of multivariate location and covariances.

Acknowledgements

This research forms a part of the doctoral dissertation of the first author. The authors thank the three anonymous referees as well as the members of the editorial board for useful suggestions which led to an improved version of the paper.

Supplementary Material

Supplement to "A generalized divergence for statistical inference" (DOI: 10.3150/16-BEJ826SUPP; .pdf). Supplement contains all the assumptions required for the asymptotic derivations and proofs of all the technical results presented in the paper. It also contains some remarks on the computation of MSDE and a simulation study illustrating the performance of the MSHDE under the bivariate normal model.

References

[1] Basu, A., Basu, S. and Chaudhuri, G. (1997). Robust minimum divergence procedures for count data models. *Sankhyā Ser. B* **59** 11–27. MR1733377

[2] Basu, A., Harris, I.R., Hjort, N.L. and Jones, M.C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85** 549–559. MR1665873

[3] Basu, A. and Lindsay, B.G. (1994). Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Ann. Inst. Statist. Math.* **46** 683–705. MR1325990

[4] Basu, A., Mandal, A., Martin, N. and Pardo, L. (2013). Testing statistical hypotheses based on the density power divergence. *Ann. Inst. Statist. Math.* **65** 319–348. MR3011625

[5] Basu, A., Mandal, A., Martin, N. and Pardo, L. (2013). Density power divergence tests for composite null hypotheses. Preprint. Available at arXiv:1403.0330.

[6] Basu, A. and Sarkar, S. (1994). Minimum disparity estimation in the errors-in-variables model. *Statist. Probab. Lett.* **20** 69–73. MR1294806

[7] Basu, A., Shioya, H. and Park, C. (2011). *Statistical Inference*: *The Minimum Distance Approach*. *Monographs on Statistics and Applied Probability* **120**. Boca Raton, FL: CRC Press. MR2830561

[8] Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5** 445–463. MR0448700

[9] Broniatowski, M. and Keziou, A. (2009). Parametric estimation and tests through divergences and the duality technique. *J. Multivariate Anal.* **100** 16–36. MR2460474

[10] Broniatowski, M., Toma, A. and Vajda, I. (2012). Decomposable pseudodistances and applications in statistical estimation. *J. Statist. Plann. Inference* **142** 2574–2585. MR2922007

[11] Brown, M. (1982). Robust line estimation with errors in both variables. *J. Amer. Statist. Assoc.* **77** 71–79.

[12] Carroll, R.J. and Gallo, P.P. (1982). Some aspects of robustness in the functional errors-in-variables regression model. *Comm. Statist. Theory Methods* **11** 2573–2585. MR0681779

[13] Cressie, N. and Read, T.R.C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46** 440–464. MR0790631

[14] Csiszár, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **3** 85–107.

[15] Dawid, A.P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron* **72** 169–183. MR3233147

[16] Dawid, A.P., Musio, M. and Ventura, L. (2016). Minimum scoring rule inference. *Scand. J. Stat.* **43** 123–138.

[17] de Boer, P.-T., Kroese, D.P., Mannor, S. and Rubinstein, R.Y. (2005). A tutorial on the cross-entropy method. *Ann. Oper. Res.* **134** 19–67. MR2136658

[18] Donoho, D. and Huber, P.J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P. Bickel, K. Doksum and J. Hodges, Jr., eds.). *Wadsworth Statist./Probab. Ser.* 157–184. Belmont, CA: Wadsworth. MR0689745
[19] Farcomeni, A. and Greco, L. (2015). *Robust Methods for Data Reduction*. Londom: Chapman & Hall.
[20] Fuller, W.A. (1987). *Measurement Error Models*. New York: Wiley. MR0898653
[21] Ghosh, A. (2015). Asymptotic properties of minimum S-divergence estimator for discrete models. *Sankhya A* **77** 380–407. MR3400120
[22] Ghosh, A. and Basu, A. (2015). The minimum S-divergence estimator under continuous models: The Basu–Lindsay approach. *Statist. Papers*. To appear. DOI:10.1007/s00362-015-0701-3.
[23] Ghosh, A. and Basu, A. (2015). Robust estimation for non-homogeneous data and the selection of the optimal tuning parameter: The density power divergence approach. *J. Appl. Stat.* **42** 2056–2072. MR3371040
[24] Ghosh, A. and Basu, A. (2016). Testing composite null hypotheses based on S-divergences. *Statist. Probab. Lett.* **114** 38–47.
[25] Ghosh, A., Basu, A. and Pardo, L. (2016). On the robustness of a divergence based test of simple statistical hypotheses. *J. Statist. Plann. Inference* **161** 91–108. MR3316553
[26] Ghosh, A., Harris, I.R., Maji, A., Basu, A. and Pardo, L. (2013). A Generalized Divergence for Statistical Inference. Technical Report, BIRU/2013/3, Bayesian and Interdisiplinary Research Unit. Kolkata, India: Indian Statistical Institute.
[27] Ghosh, A., Harris, I.R., Maji, A., Basu, A. and Pardo, L. (2016). Supplement to "A generalized divergence for statistical inference." DOI:10.3150/16-BEJ826SUPP.
[28] Ghosh, A., Maji, A. and Basu, A. (2013). Robust Inference Based on Divergences in Reliability Systems. In *Applied Reliability Engineering and Risk Analysis. Probabilistic Models and Statistical Inference* (I. Frenkel, A. Karagrigoriou, A. Lisnianski and A. Kleyner, eds.). New York: Wiley.
[29] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics*: *The Approach Based on Influence Functions*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. New York: Wiley. MR0829458
[30] Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *J. Amer. Statist. Assoc.* **89** 897–904. MR1294733
[31] Hong, C. and Kim, Y. (2001). Automatic selection of the tuning parameter in the minimum density power divergence estimation. *J. Korean Statist. Soc.* **30** 453–465. MR1895987
[32] Jiménez, R. and Shao, Y. (2001). On robustness and efficiency of minimum divergence estimators. *TEST* **10** 241–248. MR1881138
[33] Lindsay, B.G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Statist.* **22** 1081–1114. MR1292557
[34] Mameli, V. and Ventura, L. (2015). Higher-order asymptotics for scoring rules. *J. Statist. Plann. Inference* **165** 13–26.
[35] Morales, D., Pardo, L. and Vajda, I. (1995). Asymptotic divergence of estimates of discrete distributions. *J. Statist. Plann. Inference* **48** 347–369. MR1368984
[36] Pardo, L. (2006). *Statistical Inference Based on Divergence Measures*. *Statistics*: *Textbooks and Monographs* **185**. Boca Raton, FL: Chapman & Hall/CRC. MR2183173
[37] 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. MR3082808
[38] Read, T.R.C. and Cressie, N.A.C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer. MR0955054
[39] Scott, D.W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics* **43** 274–285. MR1943184

[40] Shore, J.E. and Johnson, R.W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inform. Theory* **26** 26–37. MR0560389
[41] Simpson, D.G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *J. Amer. Statist. Assoc.* **82** 802–807. MR0909985
[42] Simpson, D.G. (1989). Hellinger deviance tests: Efficiency, breakdown points, and examples. *J. Amer. Statist. Assoc.* **84** 107–113. MR0999667
[43] Stigler, S.M. (1977). Do robust estimators work with real data? *Ann. Statist.* **5** 1055–1098. MR0455205
[44] Tamura, R.N. and Boos, D.D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *J. Amer. Statist. Assoc.* **81** 223–229. MR0830585
[45] Toma, A. and Broniatowski, M. (2011). Dual divergence estimators and tests: Robustness results. *J. Multivariate Anal.* **102** 20–36. MR2729417
[46] Toma, A. and Leoni-Aubin, S. (2015). Robust portfolio optimization using pseudodistances. *PLoS ONE* **10** e0140546. DOI:10.1371/journal.pone.0140546.
[47] Vajda, I. (1989). *Theory of Statistical Inference and Information*. Dordrecht: Kluwer Academic.
[48] Warwick, J. and Jones, M.C. (2005). Choosing a robustness tuning parameter. *J. Stat. Comput. Simul.* **75** 581–588. MR2162547
[49] Wittenberg, M. (1962). An introduction to maximum entropy and minimum cross-entropy estimation using stata. *Stata J.* **10** 315–330.
[50] Zamar, R.H. (1989). Robust estimation in the errors-in-variables model. *Biometrika* **76** 149–160. MR0991433

Received December 2014 and revised February 2016