Annals of the Institute of Statistical Mathematics (2019) 71:1289–1322
https://doi.org/10.1007/s10463-018-0678-5

Robust statistical inference based on the C-divergence family

Avijit Maji[1] · Abhik Ghosh[2] · Ayanendranath Basu[2] · Leandro Pardo[3]

Received: 4 June 2017 / Revised: 30 March 2018 / Published online: 30 July 2018

Abstract

This paper describes a family of divergences, named herein as the C-divergence family, which is a generalized version of the power divergence family and also includes the density power divergence family as a particular member of this class. We explore the connection of this family with other divergence families and establish several characteristics of the corresponding minimum distance estimator including its asymptotic distribution under both discrete and continuous models; we also explore the use of the C-divergence family in parametric tests of hypothesis. We study the influence function of these minimum distance estimators, in both the first and second order, and indicate the possible limitations of the first-order influence function in this case. We also briefly study the breakdown results of the corresponding estimators. Some simulation results and real data examples demonstrate the small sample efficiency and robustness properties of the estimators.

Keywords C-Divergence · Density power divergence · Generalized power divergence · Power divergence

---

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-018-0678-5) contains supplementary material, which is available to authorized users.

✉ Avijit Maji
avijit.maji@hotmail.com

Abhik Ghosh
abhianik@gmail.com

Ayanendranath Basu
ayanbasu@isical.ac.in

Leandro Pardo
lpardo@mat.ucm.es

1 Department of Statistics and Information Management, Reserve Bank of India, South Gandhi Maidan, Patna 800001, India

2 Indian Statistical Institute, 203, B.T. Road, Kolkata 700108, India

3 Department of Statistics and O.R., Complutense University of Madrid, 28040 Madrid, Spain

1 Introduction

Statistical distances have a natural use in parametric estimation and parametric tests of hypotheses. By a statistical distance, we mean any measure of discrepancy between the data and the parametric model, which is nonnegative and equals zero if and only if the data fit the model perfectly. While we will usually refer to these measures as divergences, sometimes we will also loosely refer to them as distances.

The minimum distance branch of statistical inference has two primary subtypes. The first is based on divergences between distribution functions, while the second is based on divergences between densities. Both classes are known to produce highly robust inference; however, in this paper our attention will be entirely on density-based divergences. This is primarily due to the fact that in many cases the density-based approach leads to asymptotically highly efficient, sometimes even fully efficient, procedures. See Beran (1977), Tamura and Boos (1986), Simpson (1987, 1989), Lindsay (1994), Pardo (2006) and Basu et al. (2011) as general representatives providing the background to this area.

In this paper, we discuss a general family of density-based divergences which, while not entirely unknown in the literature, has certainly not received the attention it deserves. Here we will describe the divergences as the natural sequel of some of the existing divergence measures in the literature. We will study the inference procedures resulting from the minimization of these divergences and study the related issues from several angles. Our results will demonstrate that this rich class contains the power divergence, the density power divergence or even the S-divergence families as particular members of this new family. In fact, this new family also contains the recently proposed class of generalized S-divergences (Ghosh and Basu 2018). This new family will be referred to as the class of C-divergence measures.

This new collection of divergences, therefore, includes practically all the major density-based single integral (Jana and Basu 2018) divergences. It contains many new divergences which are not members of any established class of procedures. Yet, as we will see, these new divergences are often those which provide the best compromise between efficiency and robustness in real-life situations. While a lot of additional research has to be done to figure out which divergences are the most desirable in this respect, the description of the subsequent sections will clearly establish that the family of C-divergences allows many possible choices of optimal or near- optimal estimators in a variety of situations and a thorough exploration of the properties of this family of divergences will obviously be of significant value. However, it is worthwhile to note that some logarithmic divergences such as the logarithmic density power divergence (also referred to as the γ-divergence) of Jones et al. (2001) and Fujisawa and Eguchi., S. (2008) are not of the single-integral type and hence do not belong to our class of C-divergences.

The rest of the paper is organized as follows. Section 2 gives a description of some well-known divergences, and Sect. 3 introduces the family of C-divergences and explores some of their properties such as the influence function. Asymptotic properties of the estimators are established in Sects. 4 and 5 under discrete and continuous models, respectively; hypothesis testing is considered in Sect. 6. Some real data examples and simulation results are presented in Sect. 7, and concluding remarks are in Sect. 8.

Throughout this paper, we will refer to the true data generating distribution by G, having a density g with respect to an appropriate measure. The density g will be modeled by a parametric class of densities $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$. The distribution function corresponding to f_θ will be denoted by F_θ. Given a divergence measure $D(\cdot, \cdot)$, the best-fitting parameter (or the minimum divergence functional) at G will be denoted by $\theta^g = \arg\min_\theta D(g, f_\theta)$. When the true distribution belongs to the model, i.e., $G = F_\theta$ for some $\theta \in \Theta$, we have $\theta^g = \theta$. Estimation and tests about the unknown parameter will be based on a sample $X_1, X_2, \ldots, X_n$ of independently and identically distributed (i.i.d.) observations from the true distribution.

2 Background: two popular divergences

2.1 Cressie–Read family of power divergences

The power divergence (PD) family, originally proposed by Cressie and Read (1984) in the context of multinomial goodness-of-fit tests, is a particular subfamily of the class of disparities. A disparity measure between two densities g and f, both absolutely continuous with respect to a common dominating measure μ, is defined through the Pearson residual

$$\delta(x) = \frac{g(x)}{f(x)} - 1 \tag{1}$$

and a disparity generating function $N(\cdot)$ as

$$\rho_N(g, f) = \int N(\delta(x)) f(x) \mathrm{d}\mu(x). \tag{2}$$

We assume that the function $N(\cdot)$ is thrice differentiable and strictly convex on $[-1, \infty)$ with $N(0) = 0$, $N'(0) = 0$ and $N''(0) = 1$. It can be easily verified that $\rho_N(\cdot, \cdot)$ is a valid divergence. Throughout the rest of the paper, we will suppress the $\mathrm{d}\mu$ notation for brevity, but unless otherwise mentioned, all integrals are with respect to the dominating measure μ. The divergences defined in Eq. (2) are often described as $f-$divergences (or $\phi-$divergences). For details, see Csiszár (1963, 1967) and Ali and Silvey (1966).

Under the set up described in Sect. 1, minimum disparity estimation corresponding to the measure $\rho_N(\cdot, \cdot)$ involves the minimization of $\rho_N(\hat{g}_n, f_\theta)$ with respect to θ, where $\hat{g}_n$ is some nonparametric estimate of g based on the data. The corresponding Pearson residual, defined with $\hat{g}_n$ and f_θ in place of g and f, respectively, in Eq. (1), characterizes the potential outliers in the data probabilistically, with large positive values of the Pearson residual indicating outlying observations. These outliers can be down-weighted in the analysis by properly choosing the disparity generating function $N(\cdot)$ that down-weights large δ values. The Cressie–Read power divergence family is an important subclass of disparities having the disparity generating function

$$\xi_\lambda(\delta) = \frac{(\delta + 1)^{\lambda+1} - (\delta + 1)}{\lambda(\lambda + 1)} - \frac{\delta}{\lambda + 1}, \qquad \lambda \in (-\infty, \infty). \tag{3}$$

We reserve the specific symbol $\xi_\lambda(\cdot)$ for this family, while N is the generic symbol for a disparity generating function. The PD measure has the form

$$\rho_\lambda(g, f) = \int \xi_\lambda(\delta) f = \int \left[\frac{1}{\lambda(\lambda+1)} \left\{ g \left[\left(\frac{g}{f} \right)^\lambda - 1 \right] \right\} - \frac{g-f}{\lambda+1} \right]. \tag{4}$$

Several well-known divergences are members of this family. We can get the Pearson's Chi-square (PCS), the likelihood disparity (LD), the Hellinger distance (HD), the Kullback–Leibler divergence (KLD) and the Neyman's Chi-square (NCS) by putting $\lambda = 1, 0, -1/2, -1$ and -2, respectively, in Eq. (4).

In practice, it has been observed that minimum disparity estimators based on the PD family with $\lambda < 0$ down-weight large Pearson residuals and generate highly robust estimators, whereas estimators corresponding to $\lambda \geq 0$ fare poorly in terms of outlier stability. See Basu et al. (2011) and references therein for more detailed discussions.

In continuous models, the construction of the density estimate $\hat{g}_n$ requires the use of some nonparametric smoothing technique such as kernel density estimation. It is not an insurmountable barrier, of course, but if this step can be avoided, it makes the inference procedure simpler, both from the theoretical angle and from the point of view of implementation. However, the only divergence within the Cressie–Read class (or more generally within the class of disparities) which allows the minimization of the divergence without the density estimation component is the likelihood disparity which corresponds to the limiting case $\lambda \to 0$ in the PD family defined through Eq. (4). This divergence has the form $\rho_0(g, f_\theta) = \mathrm{LD}(g, f_\theta) = \int g \log\left(\frac{g}{f_\theta}\right)$. But

$$\mathrm{LD}(g, f_\theta) = \int g \log\left(\frac{g}{f_\theta}\right) = \int g \log(g) - \int g \log(f_\theta) = M - \int g \log(f_\theta).$$

The quantity M is independent of θ, and the maximum likelihood functional maximizes the expression $\int g \log(f_\theta)$ alone. One could write this as

$$\int g \log(f_\theta) = \int \log(f_\theta) \mathrm{d}G, \tag{5}$$

and when $\mathrm{d}G$ is replaced by dG_n, where G_n is the empirical distribution function based on an i.i.d. sample of size n, the quantity to be maximized is $n^{-1} \sum \log f_\theta(X_i)$, which is the log likelihood divided by n. Thus, one can avoid the construction of a nonparametric density estimator $\hat{g}$ in this case. However, this trick does not work with any other divergence within the class of disparities.

2.2 Density power divergence family

The density power divergence (DPD) family, introduced by Basu et al. (1998), represents another rich class of density-based divergences. The DPD measure between two densities g and f, both absolutely continuous with respect to some common dominating measure μ, is defined as

$$d_\alpha(g, f) = \int \left\{ f^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) g f^\alpha + \frac{1}{\alpha} g^{1+\alpha} \right\}, \alpha \geq 0. \tag{6}$$

Here the tuning parameter α controls the trade-off between robustness and efficiency. The value $\alpha = 0$ (in a limiting sense) produces the likelihood disparity which is highly efficient but non-robust estimator, whereas $\alpha = 1$ gives the L_2 distance leading to highly robust estimators.

Under the parametric estimation set up where f is replaced by f_θ, the third term in the right-hand side of Eq. (6) is independent of the parameter and has no role in the minimization. The rest of the objective function can be empirically modified using the same replacement trick as applied to Eq. (5), which generates the empirical objective function. Given a random sample $X_1, \ldots, X_n$ from the density g modeled by $\mathcal{F}$, the minimum DPD estimator (MDPDE) of θ is obtained by minimizing $d_\alpha(\hat{g}_n, f_\theta)$, or, equivalently, minimizing

$$\int f_\theta^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^{n} f_\theta^\alpha(X_i).$$

Thus, while the likelihood disparity is the only member of the disparity class which is “decomposable” (Broniatowski et al. 2012), this properly holds for all members of the DPD family.

3 The C-divergence family and robust parametric estimation

3.1 Formulation

Note that the DPD family in Eq. (6) can be rewritten as $d_\alpha(g, f) = \int \bar{\xi}_\alpha(\delta) f^{1+\alpha}$, where δ is as defined in Eq. (1) and a scaled version of

$$\bar{\xi}_\alpha(\delta) = \left\{ 1 - \left(1 + \frac{1}{\alpha}\right)(\delta + 1) + \frac{1}{\alpha}(\delta + 1)^{1+\alpha} \right\} \tag{7}$$

satisfies all the properties of a disparity generating function. In fact $\bar{\xi}_\alpha = (\alpha + 1)\xi_\alpha$, as is seen through a comparison of Eqs. (3) and (7). Based on this formulation, Patra et al. (2013) provided a nice connection between the DPD and the PD families. Motivated by this, we consider a general family of divergence measures, which we refer to as the C-divergence family, defined as

$$C(g, f) = \int N(\delta) f^{1+\alpha}, \tag{8}$$

where $N(\delta)$ is a regular disparity generating function. The divergence in Eq. (8) indeed defines a proper statistical distance measure for any $\alpha \geq 0$ by the properties of a disparity generating function. The function $N(\cdot)$ itself may depend on one or more tuning parameters. Note that the recently developed S-divergence family of Ghosh et al. (2013, 2017) becomes a particular subfamily of our general C-divergence family where

$$N(\delta) = N_{\alpha,\lambda}(\delta) = \frac{1}{A} - \frac{1+\alpha}{AB}(1+\delta)^A + \frac{1}{B}(1+\delta)^{1+\alpha}, \quad \alpha \geq 0, \lambda \in \mathbb{R}, \tag{9}$$

with $A = 1 + \lambda(1-\alpha)$ and $B = \alpha - \lambda(1-\alpha)$. Note that $N(0) = 0$ by default.

One prominent new member of the C-divergence family can be obtained by linking it with the PD family by choosing $N(\delta)$ as the function $\xi_\lambda(\delta)$ of Eq. (3). The resulting subfamily, which we refer to as the generalized power divergence ($\text{GPD}_{\alpha,\lambda}$) family, has the form

$$\text{GPD}_{\alpha,\lambda}(g, f) = \int \xi_\lambda(\delta) f^{1+\alpha}. \tag{10}$$

Note that, by substituting $\alpha = 0$ in Eq. (10), one gets the ordinary power divergence family in Eq. (4) as a particular member of the $\text{GPD}_{\alpha,\lambda}(g, f)$ family and $\alpha = \lambda$ gives a scaled version of the density power divergence family described by Eq. (6) with tuning parameter α. (Notice that by using the above scaling, we have also included all the divergences with negative values of the tuning parameter within the DPD class without disturbing the divergence properties). This shows that the class of C-divergences covers two significant families of divergences and we may try to exploit all their properties related to robustness and efficiency to the full extent. Both α and λ can vary over the real line, and it may also be noted that $\alpha = 0, \lambda = 0$ produces (in the limiting sense) the likelihood disparity. The N-function corresponding to the larger family of generalized S-divergences (Ghosh and Basu 2018) is given by

$$N_{\alpha,\gamma,\tau}(\delta) = \frac{1}{\tau\bar{\tau}(\alpha-\gamma)}\left[\left(\tau(\delta+1)^{1+\alpha}+\bar{\tau}\right) - \left(\tau(\delta+1)^{1+\gamma}+\bar{\tau}\right)^{\frac{1+\alpha}{1+\gamma}}\right], \quad \text{for } \alpha \geq 0,\ \tau \in [0,1],\ \gamma \in \mathbb{R}. \tag{11}$$

Note that the choices $\gamma \to -1$ or $\tau \to 0$ or $\tau \to 1$ recover the S-divergence as a special case of the generalized S-divergence.

Recently, Mattheou et al. (2009) and Vonta and Karagrigoriou (2010) have used this C-divergence family in applications such as model selection, survival analysis and reliability theory; however, the (asymptotic) distributional properties of the corresponding estimators and tests have not been studied, neither has the robustness properties of the corresponding estimators looked at. We are going to fill this gap through this present work and establish the inferential properties of this family in the case of independent and identically distributed data. Although we will discuss the theoretical results for the general C-divergence family, our examples and numerical illustrations will be primarily confined to the GPD subfamily throughout the present paper.

3.2 Estimating equation

Now, let us consider a random sample $X_1, \ldots, X_n$ from the true density g which we model by the parametric family $\mathcal{F} = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$. We are interested in estimating the parameter θ. Then, the minimum C-divergence functional $T(G)$ at G is defined as

$$C(g, f_{T(G)}) = \min_{\theta\in\Theta} C(g, f_\theta),$$

where $C(\cdot, \cdot)$ is as defined in Eq. (8). By the definition of the C-divergence, if $G = F_{\theta_0}$ then $T(F_{\theta_0}) = \theta_0$, implying the Fisher consistency of the minimum C-divergence functional $T(G)$.

Next, in order to estimate θ based on the observed data, we have to minimize $C(\hat{g}_n, f_\theta)$ with respect to θ, where $\hat{g}_n$ is the vector of relative frequencies or some continuous density estimate based on the data according to whether the setup is discrete or continuous. The estimating equation is then given by

$$-\int \{(1+\alpha)N(\delta_n(x)) - N'(\delta_n(x))(\delta_n(x)+1)\} f_\theta^{1+\alpha}(x) u_\theta(x) = 0, \qquad (12)$$

with $\delta_n(x) = \frac{\hat{g}_n(x)}{f_\theta(x)} - 1$ and $u_\theta(x) = \frac{\partial}{\partial\theta} \log f_\theta(x)$ is the likelihood score function. For simplicity, we rewrite Eq. (12) as $\int K(\delta_n(x)) f_\theta^{1+\alpha}(x) u_\theta(x) = 0$ with

$$K(\delta) = \{N'(\delta)(\delta+1) - (1+\alpha)N(\delta)\}. \qquad (13)$$

Figure 1 presents the plots of $K(\delta)$ versus δ for the GPD subfamily of C-divergences for a few values of α and λ. These give us an idea of how the large residuals are controlled by the $K(\delta)$ function, although the overall outlier controlling properties of the divergence also depend on the power $(1+\alpha)$ of the model density in Eq. (10). Note that when $\lambda = \alpha$, the curve becomes a straight line, since in this case $K(\delta) = \delta$, as may be verified through a simple calculation; in this case, the divergence reduces to the density power divergence with parameter α. The $K(\delta)$ curves for all the divergences satisfy the properties $K(0) = 0$ and $K'(0) = 1$, so that the curves are all tangential to the $K(\delta) = \delta$ line at $\delta = 0$. The curves also indicate that for each fixed α, the down-weighting strength is a decreasing function of λ, while for a fixed λ this power increases with α. Clearly, large positive α and large negative λ lead to the best results in terms of robustness. A moderately large negative value of α will require very high negative values of λ to offset the outlier sensitivity. Similarly, a very large positive value of λ will work from a robustness perspective only when α is a proportionately high positive value. All the above phenomena will be clearly illustrated in our numerical calculations.

In addition, all members of the C-divergence family generate affine invariant estimators as in the following proposition. The proof is straightforward and hence omitted.

Proposition 1 *Consider the transformation $Y = UX + v$ for some fixed non-singular matrix U and fixed vector v of the same dimension as that of the variable X. Let g_X*

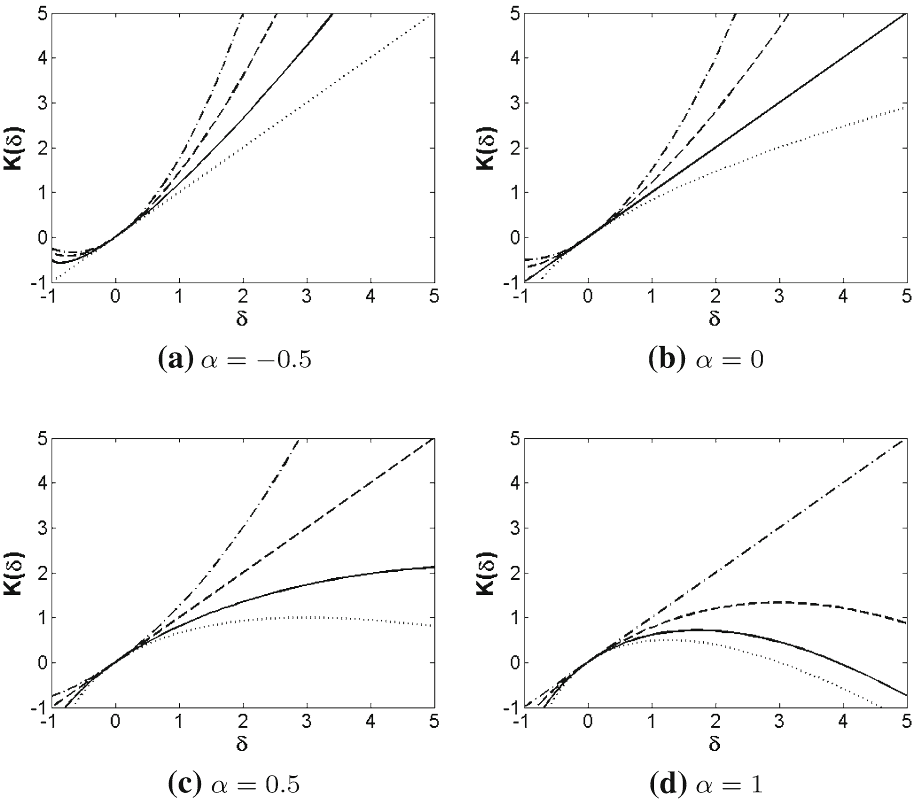


Fig. 1 Plot of $K(\delta)$ versus δ for different λ and α for the GPD family (dotted line: $\lambda = -0.5$; solid line: $\lambda = 0$; dashed line: $\lambda = 0.5$; dash-dotted line: $\lambda = 1$)

and f_X be two probability density functions for the random variable X, and let g_Y and f_Y be the respective probability density functions for the transformed variable Y. Then, some simple algebra shows that $C(g_Y, f_Y) = kC(g_X, f_X)$, where $k = |Det(U)|^{-(1+\alpha)} > 0$. Thus, the minimum C-divergence estimator is always affine equivariant, although the C-divergence itself is not necessarily affine invariant.

3.3 Influence function analysis

3.3.1 Classical first-order influence function

In order to check the robustness of the minimum C-divergence estimator, we first derive its classical first-order influence function. Consider the contaminated distribution $G_\epsilon = (1-\epsilon)G + \epsilon\wedge_y$, where G is the true distribution, $\epsilon \in [0, 1]$ is the contaminating proportion and $\wedge_y$ is the distribution degenerate at y. Then, the first-order influence function is defined as $\text{IF}(y; T, G) = T'(y) = \frac{\partial}{\partial\epsilon}T(G_\epsilon)\big|_{\epsilon=0}$. A straightforward differentiation of estimating Eq. (12) yields its influence function to be

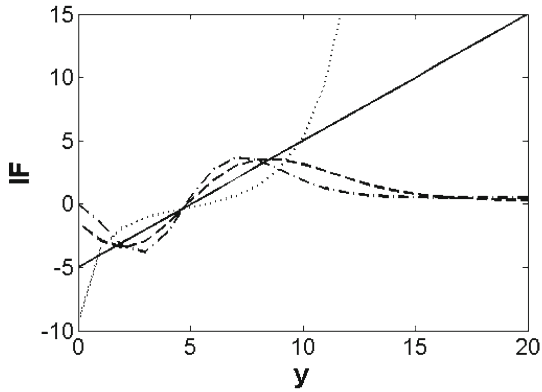


Fig. 2 The influence function of the MGPDE under the Poisson(θ) model at $\theta = 5$ (dotted line: $\alpha = -0.5$; solid line: $\alpha = 0$; dashed line: $\alpha = 0.5$; dash-dotted line: $\alpha = 1$)

$$\mathrm{IF}(y; T, G) = J_g^{-1} \left\{ u_{\theta^g}(y) K'(\delta_g^g(y)) f_{\theta^g}^{\alpha}(y) - \zeta_g \right\}, \tag{14}$$

where $\theta^g = T(G)$ represents the best-fitting parameter (indicating f_{θ^g} is the model element closest to g in the C-divergence sense), with $\delta_g^g(x) = \frac{g(x)}{f_{\theta^g}(x)} - 1$, $\zeta_g = E_g \left[u_{\theta^g}(X) K'(\delta_g^g(X)) f_{\theta^g}^{\alpha}(X) \right]$ and

$$\begin{aligned} J_g &= E_g \left[u_{\theta^g}(X) u_{\theta^g}^{\mathrm{T}}(X) K'(\delta_g^g(X)) f_{\theta^g}^{\alpha}(X) \right] \\ &\quad - \int K(\delta_g^g(x)) \left[\nabla u_{\theta^g}(x) + (1+\alpha) u_{\theta^g}(x) u_{\theta^g}^{\mathrm{T}}(x) \right] f_{\theta^g}^{1+\alpha}(x) \mathrm{d}x. \end{aligned} \tag{15}$$

In particular, at the model distribution $G = F_\theta$, we have $\theta^g = \theta$ and the above (first-order) influence function simplifies to

$$\mathrm{IF}(y; T, F_\theta) = \left[\int f_\theta^{1+\alpha} u_\theta u_\theta^{\mathrm{T}} \right]^{-1} \left\{ u_\theta(y) f_\theta^{\alpha}(y) - \int u_\theta f_\theta^{1+\alpha} \right\}, \tag{16}$$

which is the same as the influence function of the density power divergence under the model. This influence function at the model is also independent of the choice of N (and hence is independent of λ in the GPD subfamily). Figure 2 shows the above influence function of the minimum GPD estimator (MGPDE) for the Poisson model under different α; the true distribution is Poisson(5). The influence function for $\alpha = 0$ increases linearly, while for $\alpha > 0$ the influence functions are bounded and re-descending. The $\alpha = -0.5$ estimator is worse, by far, than even the $\alpha = 0$ case as it inflates the effect of the outlier faster. While this influence function is useful in that it can predict the increasing robustness with increasing α, it fails to capture the role of the λ parameter in the process as we will subsequently observe.

In Tables 1 and 2, we have presented the bias and the MSE of the minimum GPD$_{\alpha,\lambda}$ estimator (MGPDE, with different α and λ) of the parameter θ under the Poisson(`) model based on 1000 replications of random samples of size $n = 50$ simulated from the 0.9 Poisson(5) + 0.1 Poisson(25) mixture. The target parameter is the mean of the major component (which equals 5 in this case). The behavior of the estimators

Table 1 Bias of the MGPDEs of the Poisson mean under contaminated data (sample size 50) for various values of λ and α

$\lambda \downarrow \alpha \rightarrow$	−0.9	−0.7	−0.5	−0.3	−0.1	0	0.1	0.3	0.5	0.7	0.9
−0.9	7.88	7.69	7.12	5.04	0.47	−0.37	−0.34	−0.26	−0.2	−0.15	−0.11
−0.7	7.9	7.74	7.33	5.98	1.25	−0.18	−0.2	−0.13	−0.09	−0.06	−0.03
−0.5	7.91	7.76	7.39	6.31	1.91	−0.1	−0.15	−0.08	−0.05	−0.03	0
−0.3	7.91	7.77	7.45	6.56	2.79	−0.01	−0.13	−0.05	−0.03	−0.01	0.02
−0.1	7.92	7.8	7.53	6.86	4.31	0.44	−0.16	−0.04	−0.01	−0.01	0.03
0	7.92	7.81	7.58	7.02	5.17	1.99	−0.17	−0.04	0	−0.01	0.03
0.1	7.93	7.83	7.63	7.17	5.87	4.04	0.39	−0.05	0	0.02	0.03
0.3	7.93	7.86	7.71	7.42	6.75	6.06	4.78	0.05	0	0.02	0.04
0.5	7.94	7.88	7.77	7.59	7.21	6.87	6.35	3.91	0.04	0.02	0.04
0.7	7.94	7.89	7.81	7.69	7.45	7.26	7	5.99	3.23	0.04	0.04
0.9	7.94	7.9	7.84	7.75	7.59	7.47	7.31	6.78	5.64	2.67	0.05

Table 2 MSE of the MGPDEs of the Poisson mean under contaminated data (sample size 50) for various values of λ and α

$\lambda \downarrow \alpha \rightarrow$	−0.9	−0.7	−0.5	−0.3	−0.1	0	0.1	0.3	0.5	0.7	0.9
−0.9	65.58	62.56	54.33	29.16	0.56	0.39	0.37	0.32	0.29	0.27	0.26
−0.7	65.71	63.09	56.92	39.19	2.13	0.19	0.2	0.19	0.19	0.19	0.2
−0.5	65.77	63.29	57.72	42.88	4.52	0.14	0.15	0.15	0.16	0.18	0.19
−0.3	65.83	63.54	58.53	45.88	9.18	0.13	0.14	0.14	0.15	0.17	0.19
−0.1	65.93	63.92	59.71	49.72	20.55	0.43	0.15	0.13	0.15	0.17	0.18
0	65.99	64.15	60.43	51.98	28.94	4.88	0.16	0.13	0.15	0.16	0.18
0.1	66.04	64.4	61.19	54.24	36.9	18.16	0.36	0.13	0.14	0.16	0.18
0.3	66.13	64.86	62.55	58.05	48.39	39.42	25.12	0.13	0.14	0.16	0.18
0.5	66.18	65.19	63.54	60.58	54.92	50.17	43.21	17.44	0.14	0.15	0.17
0.7	66.19	65.41	64.18	62.15	58.58	55.78	51.92	38.76	12.41	0.15	0.17
0.9	66.19	65.55	64.6	63.12	60.7	58.91	56.54	49.01	34.78	8.87	0.17

observed in Tables 1 and 2 is in perfect accordance with our observations in Sect. 3.2 based on the shape of the $K(\delta)$ functions. Large positive α and large negative λ generally lead to more robust solutions. When read with Tables S1–S4 of the Online Supplement, it is clear that there are several members within the GPD class which provide excellent compromise between efficiency at the model and stability under data contamination. These include the estimators at $\alpha = 0.2$ and $-0.3 \leq \lambda \leq 0$; $\alpha = 0.3$ and $-0.2 \leq \lambda \leq 0.2$; and $\alpha = 0.4$ and $0.2 \leq \lambda \leq 0.3$. Most of these minimum distance estimators do not belong to any of the previously existing family of divergences. Thus, there are many new divergences within the GPD family which can be quite competitive with the existing standard in minimum distance inference. And this is only the case of GPD. While we have not provided the detailed analysis, outside this family some other members of the C-divergence class are also quite desirable in these respect. In short, further in-depth study of the C-divergence family is likely to produce many other divergences having excellent potential in statistical inference.

It is also clear that the (first-order) influence function analysis is inadequate to predict the behavior of these estimators. For example, the estimators with $\alpha = 0$ and moderately large negative values of λ provide highly stable estimators in contradiction to what is predicted by their unbounded influence functions. On the other hand, the bias and mean square errors of several estimators with positive α and large positive λ are literally huge, belying the behavior expected in view of their bounded influence functions.

So we need to consider a second-order influence analysis to get a better understanding of the robustness of the minimum C-divergence estimators.

3.3.2 Higher-order influence function

Consider again the contaminated distribution G_ϵ defined in Sect. 3.3.1; the second-order influence function of the functional $T(G)$ can be defined as

$$\mathrm{IF}_2(y; T, G) = T''(y) = \frac{\partial^2}{\partial \epsilon^2} T(G_\epsilon)\bigg|_{\epsilon=0},$$

whose general form can be obtained by differentiating the estimating equation twice. We present the form of this second-order influence function at the model for the minimum GPD estimator of a scalar parameter θ in the following theorem. The derivation involves routine differentiation and is omitted.

Theorem 2 *Suppose the true distribution belongs to the model family with $g = f_\theta$ with a scalar parameter θ. Then, the second-order influence function of the minimum GPD estimator of θ is*

$$T''(y) = (N_0^p D_0 - N_0 D_0^p)/D_0^2,$$

where

$$\begin{aligned} N_0 &= f_\theta^\alpha(y) u_\theta(y) - c_1, \qquad D_0 = c_2, \\ N_0^p &= T'(y) p_1 + (\alpha - \lambda) p_2, \quad D_0^p = T'(y) q_1 + q_2, \end{aligned}$$

with $u'_\theta(y) = \frac{\partial}{\partial\theta} u_\theta(y)$, $c_i = \int f_\theta^{1+\alpha} u_\theta^i$ and $d_i = \int f_\theta^{1+\alpha} u_\theta^i u'_\theta$ for $i = 0, 1, 2$, and

$$\begin{aligned} p_1 &= (2\alpha - \lambda) f_\theta^\alpha(y) u_\theta^2(y) + f_\theta^\alpha(y) u'_\theta(y) - (2\alpha - \lambda) c_2 - d_0, \\ p_2 &= 2 f_\theta^\alpha(y) u_\theta(y) - f_\theta^{\alpha-1}(y) u_\theta(y) - c_1, \\ q_1 &= (1 - \lambda + 3\alpha) c_3 + 3 d_1, \\ q_2 &= (2\alpha - \lambda)(c_2 - f_\theta^\alpha(y) u_\theta^2(y)) + d_0 - f_\theta^\alpha(y) u'_\theta(y). \end{aligned}$$

Notice that unlike the first-order influence function, the second- order influence function of the minimum GPD estimator is quite critically dependent on the value of λ. In order to interpret the second-order influence function, let us consider the Taylor series expansion of the minimum GPD functional $T(\cdot)$ (or, in general, any minimum C-divergence estimator) as

$$T(G_\epsilon) = T(G) + \epsilon T'(y) + \frac{\epsilon^2}{2} T''(y),$$

which should give a better approximation to the bias $[T(G_\epsilon) - T(G)]$ under contamination. A measure of the limitation of the first-order bias approximation (given by the linear approximation involving the first-order influence function $T'(y)$ only) can be described by the ratio

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

This specifically shows that if ϵ is larger than $\epsilon_{\text{critical}} = \left|\frac{T'(y)}{T''(y)}\right|$, the second-order approximation will differ by more than 50% compared to the first-order approximation. Smaller values of $\epsilon_{\text{critical}}$ demonstrate that the second-order approximation deviates faster from the first, indicating greater inadequacy of the latter in predicting bias under contamination.

Example: Poisson Mean

We now present a numerical example to show the performance of the second-order influence analysis through its application in the case of the Poisson model with mean θ. We compute the first and second-order bias approximations using their respective expressions. For brevity, we will only present the results of some particular cases with $\theta = 4$, the contamination point $y = 12$ and some specific (α, λ) combinations. The corresponding bias plots are shown in Figs. 3, 4 and 5, respectively, for $\lambda = 0$, $\lambda > 0$ and $\lambda < 0$. The findings in the figures are consistent with our observations in Table 1.

Comments on Fig. 3 ($\lambda = 0$): As expected, both the first-order and second-order influence functions for $\alpha = 0$ generate the same straight line (the influence function of the first two orders are identical). For $\alpha > 0$, the second-order influence function predicts a smaller bias compared to the first-order one as for all these cases the inequality $\alpha > \lambda$ holds. The bias approximation does not vary much over the various values of α.

Comments on Fig. 4 ($\lambda > 0$): As we have already noted large positive values of λ lead to poor behavior in terms of robustness. This figure demonstrates that values of α smaller than λ are usually not enough to offset this lack of robustness. The present figure shows that as long as $\alpha < \lambda$ holds, the second-order bias approximation is larger compared to the first order.

Comments on Fig. 5 ($\lambda < 0$): This figure highlights a behavior which complements the observations of Fig. 4. In this case, we have chosen the α values to be larger than the value of λ and now the second-order approximation consistently predicts a smaller bias compared to the first- order approximation.

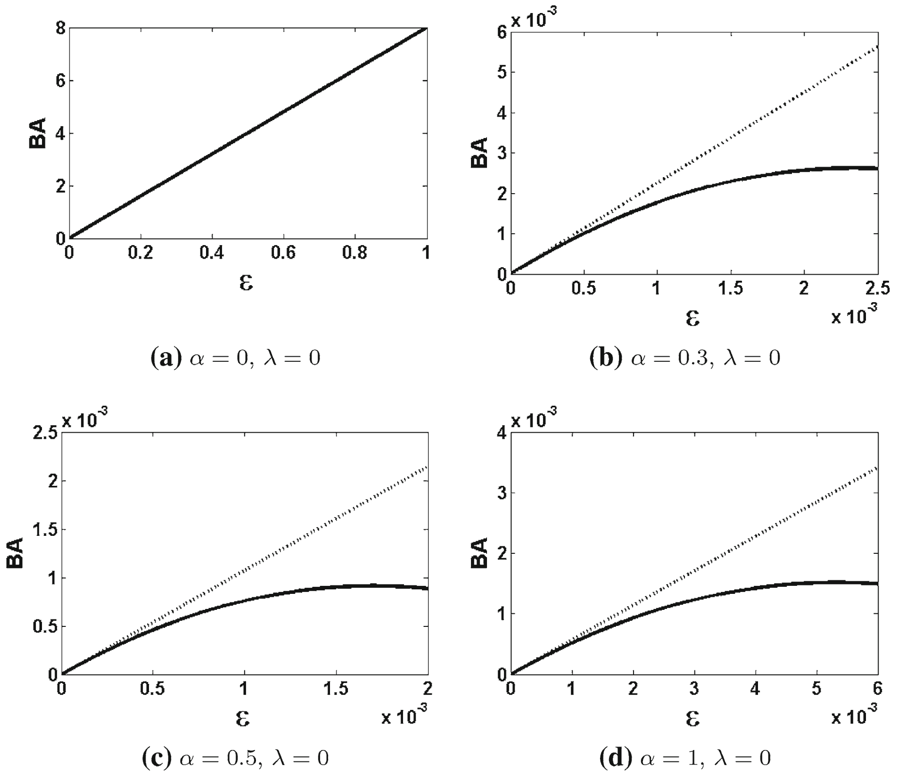


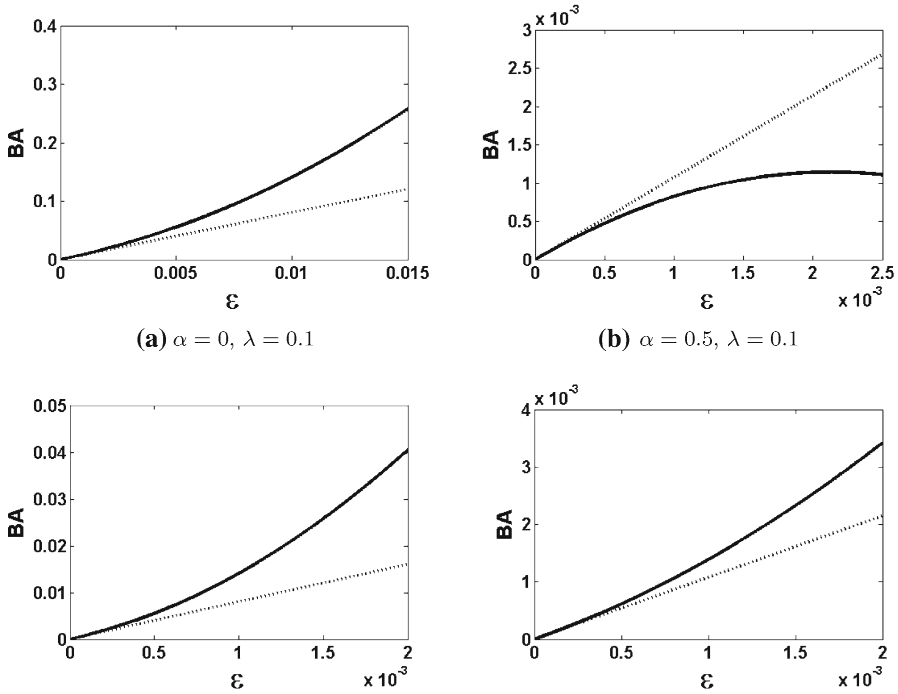
Fig. 3 Bias approximations (BA; solid line: second order; dashed line: first order) for the MGPDEs with $\lambda = 0$ over contamination proportion ϵ

3.4 Breakdown point analysis

In this section, we will derive the asymptotic breakdown point of the minimum C-divergence estimator under the location model $\mathcal{F}_\theta = \{f_\theta(x) = f(x-\theta) : \theta \in \Theta\}$. The reason for using the location family for this purpose is simply that the location family enjoys the property

$$\int \{f(x-\theta)\}^{1+\alpha} \mathrm{d}x = \int \{f(x)\}^{1+\alpha} \mathrm{d}x = M_f^\alpha < \infty, \quad \text{(say)}$$

so that the breakdown point proof avoids the problem of having the above integral depending on the parameter. Since we are interested in the breakdown within the model family, we will be primarily interested in the case where the true density g belongs to the model, i.e., $g = f_{\theta^g}$. Let $c(g, f) = N(\frac{g}{f} - 1) f^{1+\alpha}$, and define $c(g, 0) = \lim_{f\to 0} \mathrm{c}(g, f) = \lim_{f \to 0} N\left(\frac{g}{f} - 1\right) f^{1+\alpha}$. The following discussion will assume that the class of divergences considered here satisfies the condition given below.



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

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

Fig. 4 Bias approximations (BA; solid line: second order; dashed line: first order) for the MGPDEs with $\lambda > 0$ over contamination proportion ϵ

Assumption (C) $N(-1)$ and $N'_\alpha(\infty)$ are both bounded, where $N'_\alpha(\infty) := \lim_{\delta\to\infty} \frac{N(\delta)}{\delta^{1+\alpha}}$.

It is important to note that these assumptions are satisfied by several important members of the C-divergence family, for example, the S-divergence family—see Eq. (9)—with $B > 0$; this includes the DPD family in Eq. (6) with $\alpha > 0$. These assumptions also hold for the GPD family in Eq. (10) with $\lambda < \alpha$.

To derive the breakdown results for the above setup, we consider the contamination model $H_{\epsilon,n} = (1-\epsilon)G+\epsilon V_n$, where $\{V_n\}$ is a sequence of contaminating distributions. Let $h_{\epsilon,n}$, g and v_n be the corresponding densities. The functional T is said to break down at ϵ level of contamination (see Simpson 1987) if there exists a sequence $\{v_n\}$ of densities such that

$$|T(H_{\epsilon,n}) - T(G)| \to \infty \quad \text{as} \quad n \to \infty.$$

We write below $\theta_n = T(H_{\epsilon,n})$. Given the model family $\mathcal{F}_\theta$, we will consider the following assumptions for our contamination sequences for the derivation of the breakdown point result in Theorem 3; the proof is given in the Online Supplement. Park and Basu (2004) and Ghosh et al. (2017), among others, have utilized similar assumptions.

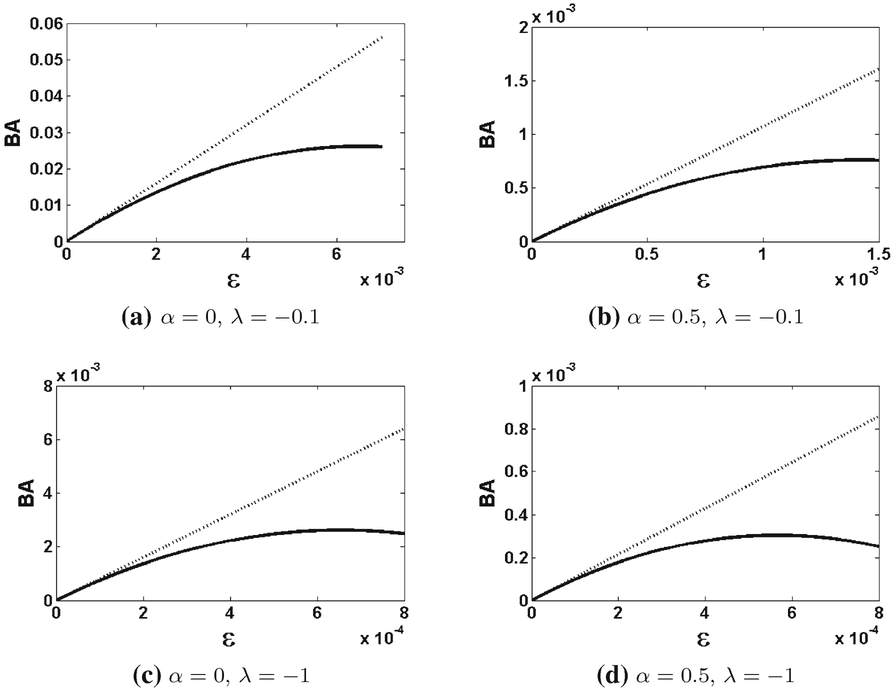


Fig. 5 Bias approximations (BA; solid line: second order; dashed line: first order) for the MGPDEs with $\lambda < 0$ over contamination proportion ϵ

Assumption (BP) The contaminating sequence of densities $\{v_n\}$, the data density $g(x)$ and the model $f_\theta(x)$ satisfy the following.

1. $\int \min\{f_\theta(x), v_n(x)\} \to 0$ as $n \to \infty$ uniformly for $|\theta| \le a$ for any fixed a. That is, the contamination distribution is asymptotically singular to specified models.
2. $\int \min\{g(x), f_{\theta_n}(x)\} \to 0$ as $n \to \infty$ if $|\theta_n| \to \infty$ as $n \to \infty$. That is, large values of the parameter θ give distributions which become asymptotically singular to the true distribution.
3. For any $\theta \in \Theta$ and $0 < \epsilon < 1$, we have $C(\epsilon v_n, f_\theta) \ge C(\epsilon f_\theta, f_\theta)$ and $\limsup_{n\to\infty} \int v_n^{1+\alpha} \le M_f^\alpha$.

Theorem 3 *Under* Assumptions (C) *and* (BP), *the asymptotic breakdown point* ϵ^* *of the minimum* C*-divergence functional is at least* $\frac{1}{2}$ *at the model.*

Remark 1 It follows from Rousseeuw and Leroy (1987) that the maximum asymptotic breakdown point for any affine equivariant location estimator is 1/2. As the minimum C-divergence estimator is affine equivariant (see Proposition 1), the above theorem also establishes that the asymptotic breakdown point of the minimum C-divergence estimator of a location parameter is exactly 1/2 at the model.

Remark 2 Theorem 3 above provides a generalization of the available breakdown point result for the minimum S-divergence estimators as derived in Ghosh et al. (2017), which requires both A and B to be positive along with Assumption (BP). But our result is valid for all the S-divergence measures with only $B > 0$, since this itself implies Assumption (C).

4 Asymptotic properties of the minimum C-divergence estimators under discrete models

Let us consider the setup of discrete models where the true density g and the model density f_θ are supported, without loss of generality, on $\chi = \{0, 1, 2, \ldots\}$ with μ being the counting measure over χ under the notation of the previous sections. Based on n independent and identically distributed observations $X_1, X_2, \ldots, X_n$ from g, let the relative frequency at $x \in \chi$ be denoted as $d_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i = x)$ where $I(A)$ is the indicator function for the event A.

Then, the minimum C-divergence estimator has to be obtained by minimizing the C-divergence measure between the data $\mathbf{d}_n = (d_n(0), d_n(1), \ldots)^{\mathrm{T}}$ and the model probability $\mathbf{f}_\theta = (f_\theta(0), f_\theta(1), \ldots)^{\mathrm{T}}$ with respect to θ. The corresponding estimating equation is then given by $\sum_x K(\delta_n(x)) f_\theta^{1+\alpha}(x) u_\theta(x) = 0$, where now we have $\delta_n(x) = \frac{d_n(x)}{f_\theta(x)} - 1$ and $K(\cdot)$ is as defined in Eq. (13).

Let $u_{j\theta}$, $u_{jk\theta}$ and $u_{jkl\theta}$ be the first- order, second-order and third-order partial derivatives of $\log f_\theta$ with respect to θ, $1 \le j, k, l \le p$. Further, let $\theta^g = T(G)$ denote the best-fitting parameter and so f_{θ^g} is the model element closest to g in the C-divergence sense. Consider the matrix J_g as defined in Eq. (15) and define

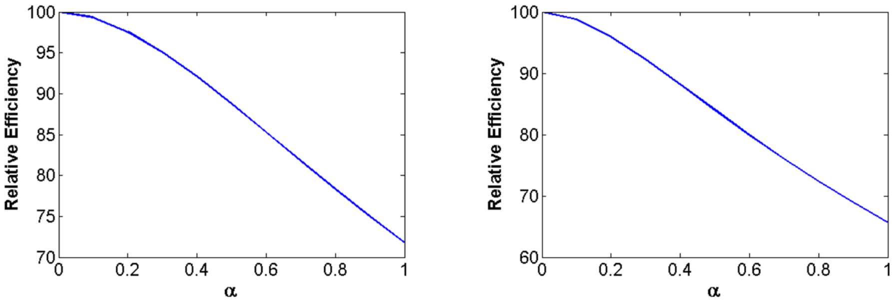
$$V_g = V_g\left[K'(\delta_g^g(X)) f_{\theta^g}^\alpha(X) u_{\theta^g}(X)\right], \tag{17}$$

with $\delta_g^g(x) = \frac{g(x)}{f_{\theta^g}(x)} - 1$.

In order to obtain the asymptotic properties of the minimum C-divergence estimators (MCDEs) under this discrete model setup, we will make the following assumptions on the model family and the C-divergence generating function $N(\cdot)$.

Assumptions

- (A1) The model family $\mathcal{F}$ is identifiable meaning that different values of the parameter must generate different probability distributions of the observable variables.
- (A2) The model density f_θ has common support $\{x : f_\theta(x) > 0\}$ independently of θ; the true density g has also the same support.
- (A3) There exists an open subset $\omega \subset \Theta$ such that θ^g is an interior point of ω and for almost all x, $f_\theta(x)$ admits all third-order partial derivatives for all $\theta \in \omega$.
- (A4) The matrix J_g defined in Eq. (15) is positive definite.
- (A5) The quantities $\sum_x g^{1/2}(x) f_\theta^\alpha(x)|u_{j\theta}(x)|$, $\sum_x g^{1/2}(x) f_\theta^\alpha(x)|u_{j\theta}(x)||u_{k\theta}(x)|$ and $\sum_x g^{1/2}(x) f_\theta^\alpha(x)|u_{jk\theta}(x)|$ are bounded for all j, k and all $\theta \in \omega$.
- (A6) For almost all x, there exist functions $M_{jkl}(x)$, $M_{jk,l}(x)$, $M_{j,k,l}(x)$ that dominate, in absolute value, $f_\theta^\alpha(x) u_{jkl\theta}(x)$, $f_\theta^\alpha(x) u_{jk\theta}(x) u_{l\theta}(x)$ and $f_\theta^\alpha(x) u_{j\theta}(x) u_{k\theta}(x) u_{l\theta}(x)$, respectively, for all j, k, l and all these dominating functions are uniformly bounded in expectation with respect to g and f_θ for all $\theta \in \omega$.



(a) Binomial(5,0.6) **(b)** Poisson(3)

Fig. 6 Relative efficiency of the MGPDE over various α

(A7) Function $K(\cdot)$ is twice differentiable. Also, for each $\theta \in \omega$, $K'(\delta_\theta(x))$ and $K''(\delta_\theta(x))(1+\delta_\theta(x))$ are bounded uniformly in x, where $\delta_\theta(x) = \frac{g(x)}{f_\theta(x)} - 1$.

The first three assumptions are standard ones for any asymptotic derivation and hold for most common parametric model families. Assumptions (A4)–(A6) are also quite common in the literature of minimum distance methods and can be seen to hold at the model for standard parametric families like Bernoulli, Poisson, geometric; see Ghosh (2015) for some such illustrations. Finally, Assumption (A7) is indeed exactly the same as that required for disparity family (Assumption A7, Basu et al. 2011, p.61) with $K(\cdot)$ playing a role analogous to the residual adjustment function.

The following theorem proves the consistency and asymptotic normality of the minimum C-divergence estimator. The proof is presented in the Online Supplement.

Theorem 4 *Under* Assumptions (A1)–(A7)*, there exists a consistent sequence θ_n of roots to the minimum C-divergence estimating* Eq. (12) *and asymptotically*

$$\sqrt{n}(\theta_n - \theta^g) \sim N_p\left(0, J_g^{-1} V_g J_g^{-1}\right).$$

Corollary 5 *When the true distribution G belongs to the model family with $g = f_\theta$ for some $\theta \in \Theta$, then $\theta^g = \theta$ and $\sqrt{n}(\theta_n - \theta)$ has an asymptotic $N_p(0, J^{-1}VJ^{-1})$ distribution, where*

$$J = J_\alpha(\theta) = E_g[u_\theta(X)u_\theta(X)^{\mathrm{T}} f_\theta^\alpha(X)] = \sum u_\theta(x)u_\theta^{\mathrm{T}}(x) f_\theta^{1+\alpha}(x) \tag{18}$$

$$V = V_\alpha(\theta) = V_g[u_\theta(X) f_\theta^\alpha(X)] = \sum u_\theta(x)u_\theta^{\mathrm{T}}(x) f_\theta^{1+2\alpha}(x) - \iota\iota^{\mathrm{T}}, \tag{19}$$

$$\iota = \iota_\alpha(\theta) = E_g[u_\theta(X) f_\theta^\alpha(X)] = \sum u_\theta(x) f_\theta^{1+\alpha}(x). \tag{20}$$

Note that the asymptotic distribution at the model is again independent of the choice of $N(\cdot)$ and hence is independent of the parameter λ in the GPD family. This implies that the asymptotic efficiency of the minimum C-divergence estimator at the model also depends only on the parameter α and in fact coincides with that of the minimum DPD estimators. Figure 6 shows the plot of the asymptotic relative efficiency over

α for the binomial and Poisson models; the panel labels (a) and (b) indicate the true underlying distributions. Clearly, there is a loss in efficiency with increasing α as the cost for higher robustness, but this loss is not very substantial at small positive values of α.

5 Asymptotic properties of the minimum *C*-divergence estimator under continuous models using the Basu–Lindsay approach

When the objective function relating to the minimization of a divergence represents an i.i.d. sum over the observed data points, establishing the asymptotic properties of the minimum divergence estimator is simple as the estimator then represents an M-estimator. Such divergences have been called "decomposable pseudodistances" in the literature; see, e.g., Broniatowski et al. (2012). As already observed in Section 3.1, the GPD subclass with the restriction $\alpha = \lambda$ represents a family of decomposable pseudodistances. However, when the divergence is not decomposable and so cannot be approximated by the empirical mean of some loss function, one does need a nonparametric estimate of the true unknown density to reconstruct the empirical divergence. When the model is discrete, there is a natural estimate of this density given by the vector of relative frequencies d_n, which we have exploited in Section 4 to develop the asymptotic distribution of the corresponding estimator. In the continuous case, however, there is no simple and natural estimate of the population density. The sampled data, even if the model is a continuous one, are always discrete. To construct a divergence between two densities under the same measure, therefore, one has to first construct a continuous density estimate of the true unknown data generating density using methods such as kernel density estimation. This adds an extra layer to the parameter estimation scheme. In spite of this additional complication, the method has been successfully used in both parametric estimation and parametric hypothesis testing by several authors including Beran (1977), Tamura and Boos (1986), Simpson (1989) and Park and Basu (2004) and others; also see Broniatowski and Vajda (2012). In the subsequent description, we describe a similar approach for performing minimum divergence estimation involving non-decomposable divergences under continuous models.

Suppose that the true density g and the model family $\mathcal{F}$ represent continuous densities with respect to the Lebesgue measure. Under the above construct, the minimum C-divergence estimator of θ may be obtained by minimizing $C(g_n^*, f_\theta)$, where g_n^* is a kernel density estimate of the true density g based on the sample data $X_1, \ldots, X_n$ given by

$$g_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} W(x, X_i, \nu_n) = \int W(x, y, \nu_n)\mathrm{d}G_n(y), \tag{21}$$

with $W(x, y, \nu_n)$ being a smooth kernel function with bandwidth ν_n. Usual choices for the kernel W are given in terms of a symmetric nonnegative density function $w(\cdot)$ as $W(x, X_i, \nu_n) = \frac{1}{\nu_n}\, w\left(\frac{x - X_i}{\nu_n}\right)$. This approach of density-based minimum distance estimation under continuous models does depend, sometimes critically, on the choice of the bandwidth sequence. The alternative approach of Basu and Lindsay (1994), which we are going to follow here, attempts to nullify the effect of the bandwidth sequence by also introducing the same distortion in the model by convoluting it with the same kernel. Let f_θ^* be the kernel smoothed model density given by $f_\theta^*(x) = \int W(x, y, \nu_n)\, dF_\theta(y)$. In the Basu–Lindsay approach, we consider the minimization of $C(g_n^*, f_\theta^*)$ in order to estimate the parameter θ. The corresponding estimating equation is then given by

$$\int K(\delta_n^*)(f_\theta^*)^{1+\alpha}\widetilde{u_\theta} = 0, \tag{22}$$

where $\widetilde{u_\theta} = \nabla \log f_\theta^* = \frac{\nabla f_\theta^*}{f_\theta^*}$, $\delta_n^* = \frac{g_n^*}{f_\theta^*} - 1$ and $K(\delta)$ is as defined in Eq. (13). We will denote the resulting estimator as the minimum C^*-divergence estimator (MCDE*) which is, in general, not the same as the MCDE obtained by minimizing $C(g_n^*, f_\theta)$. However, it is possible to impose some restrictions on the kernel density estimator so that the asymptotic properties of both MCDE and MCDE* become equivalent when the true density g belongs to the model family $\mathcal{F}$. Unlike the unsmoothed scheme, model smoothing allows the corresponding estimator to be consistent for fixed bandwidths $\nu_n = \nu$.

For the true, unknown, data generating density $g(x)$, let

$$g^*(x) = \int W(x, y, \nu)g(y)\mathrm{d}y$$

be its kernel smoothed version. We assume that a random sample $X_1, \ldots, X_n$ is drawn from the true distribution G (having density function g). A typical density in the model family is denoted by f_θ, while its kernel smoothed version is represented by f_θ^*. Let θ^g represent the best-fitting parameter which satisfies $C(g^*, f_{\theta^g}^*) = \min_{\theta\in\Theta} C(g^*, f_\theta^*)$. The Pearson residuals in this context are defined as $\delta_n^*(x) = g_n^*(x)/f_\theta^*(x) - 1$ and $\delta_g^*(x) = g^*(x)/f_\theta^*(x) - 1$. Corresponding to $\widetilde{u_\theta}(x)$, we will also define the smoothed partials $\widetilde{u_{j\theta}}(x) = \nabla_j \log f_\theta^*(x)$, and $\widetilde{u_{jk\theta}}(x) = \nabla_{jk} \log f_\theta^*(x)$.

The influence function for the minimum C^* estimator, given in the following lemma, results from a straightforward differentiation of Eq. (22).

Lemma 6 *The influence function for the minimum C^* functional T^* at the distribution G is given by*

$$IF(y; G, T^*) = [J^{*g}(\theta^g)]^{-1} u_{\theta^g}^{*g}(y), \tag{23}$$

where $u_\theta^{*g}(y) = \int \widetilde{u_\theta}(x)K'(\delta_g^*(x))\{f_\theta^*(x)\}^\alpha\{W(x, y, \nu) - g^*(x)\}dx$ and

$$\begin{aligned} J^{*g}(\theta) = &\int \widetilde{u_\theta}(x)\widetilde{u_\theta}(x)^{\mathrm{T}} K'(\delta_g^*(x))\{f_\theta^*(x)\}^\alpha g^*(x)dx \\ &- \int K(\delta_g^*(x))[\nabla\widetilde{u_\theta}(x) + (1+\alpha)\widetilde{u_\theta}(x)\widetilde{u_\theta}(x)^{\mathrm{T}}\{f_\theta^*(x)\}^{1+\alpha}]dx. \end{aligned} \tag{24}$$

Corollary 7 *When the true density g belongs to the model family $\mathcal{F}$, i.e., $g = f_\theta$, the influence function for the minimum C^* estimator functional T^* at the distribution $G = F_\theta$ is given by*

$$IF(y; F_\theta, T^*) = [J^*(\theta)]^{-1} u_\theta^*(y) \tag{25}$$

where $u_\theta^*(y) = \int \{f_\theta^*(x)\}^\alpha \widetilde{u_\theta}(x)\{W(x, y, \nu) - f_\theta^*(x)\}dx$ and

$$J^*(\theta) = \int \widetilde{u_\theta}(x)\widetilde{u_\theta}(x)^{\mathrm{T}}\{f_\theta^*(x)\}^{1+\alpha} dx. \tag{26}$$

We now provide the regularity conditions leading to the main theorem of this section.

(B1) The model family $\mathcal{F}$ is identifiable.

(B2) The probability density function f_θ of the model distribution has common support so that the set $\chi = \{x : f_\theta(x) > 0\}$ is independent of θ. Also the true distribution g has the same support.

(B3) There exists an open subset $\omega \subset \Theta$ such that θ^g is an interior point of ω, and for almost all x, $f_\theta^*(x)$ admits all third-order partial derivatives for all $\theta \in \omega$.

(B4) The matrix J^{*g} is positive definite.

(B5) The quantities $\int g^{*1/2}(x) f_\theta^{*\alpha}(x)|\widetilde{u_{j\theta}}(x)|$, $\int g^{*1/2}(x) f_\theta^{*\alpha}(x)|\widetilde{u_{j\theta}}(x)||\widetilde{u_{k\theta}}(x)|$ and $\int g^{*1/2}(x) f_\theta^{*\alpha}(x)|\widetilde{u_{jk\theta}}(x)|$ are bounded $\forall j, k$ and $\forall \theta \in \omega$.

(B6) For almost all x, there exist functions $M_{jkl}(x)$, $M_{jk,l}(x)$, $M_{j,k,l}(x)$ that dominate, in absolute value, $f_\theta^{*\alpha}(x)\widetilde{u_{jkl\theta}}(x)$, $f_\theta^{*\alpha}(x)\widetilde{u_{jk\theta}}(x)\widetilde{u_{l\theta}}(x)$ and $f_\theta^{*\alpha}(x)$ $\widetilde{u_{j\theta}}(x)\widetilde{u_{k\theta}}(x)\widetilde{u_{l\theta}}(x)$, respectively, $\forall j, k, l$ and that are uniformly bounded in expectation with respect to g^* and f_θ^* for all $\theta \in \omega$.

(B7) The function $K(\cdot)$ is twice differentiable. Also, for each $\theta \in \omega$, $K'(\delta_g^*(x))$ and $K''(\delta_g^*(x))(1 + \delta_g^*(x))$ are bounded uniformly in x, where $\delta_g^*(x) = \frac{g^*(x)}{f_\theta^*(x)} - 1$.

Theorem 8 *Under the above set of assumptions, there exists a consistent sequence θ_n^* of roots to the minimum C^* estimating equation (22). Also, the asymptotic distribution of $\sqrt{n}(\theta_n^* - \theta^g)$ is p-dimensional normal with mean 0 and variance $[J^{*g}(\theta^g)]^{-1} V^*(\theta^g)[J^{*g}(\theta^g)]^{-1}$, where $J^{*g}(\theta)$ is as defined in (24) and $V^*(\theta) = Var\left[\int W(x, X, \nu)K'(\delta_g^*(x))(f_\theta^*(x))^\alpha \widetilde{u_\theta}(x)dx\right]$.*

Proof It follows along the lines of Theorem 3.19 of Basu et al. (2011). □

Corollary 9 *When the true density g belongs to the model family $\mathcal{F}$, i.e., $g = f_\theta$, then the asymptotic distribution will be the same with the variance having the simpler form $[J^*(\theta)]^{-1} V_0^*(\theta)[J^*(\theta)]^{-1}$ with $J^*(\theta)$ as defined in Eq. (26) and $V_0^*(\theta) = Var\left[\int W(x, X, \nu)(f_\theta^*(x))^\alpha \widetilde{u_\theta}(x)dx\right]$.*

6 Testing parametric hypotheses using *C*-divergence measures

Let us now move to the other important domain of statistical inference, i.e., testing of hypothesis. Following the Basu et al. (2013) approach, we will consider two specific cases of one- and two-sample problems in this section.

6.1 One-sample problem

We consider a parametric family of densities $\mathcal{F} = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ as in the previous sections. Based on an observed random sample $X_1, \ldots, X_n$ from a density modeled by $\mathcal{F}$, we want to test for the simple null hypothesis

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \neq \theta_0. \tag{27}$$

It is well known that there is no uniformly most powerful test available for this problem for most common parametric models and the widely used likelihood ratio test (LRT) is highly non-robust. We can alternatively perform a robust test for the hypothesis in Eq. (27) by using the C-divergence measure between f_{θ_0} and $f_{\hat{\theta}}$, where $\hat{\theta}$ is the minimum C-divergence estimator of θ based on the given sample. However, for theoretical and computational ease, we will introduce certain modifications in the process. Recalling that the asymptotic distribution of the minimum C-divergence estimator depends only on the parameter α in Eq. (8), here we identify the C-divergence between the densities g and f simply by the parameter α as $C_\alpha(g, f)$, even though we allow a general $N(\cdot)$ in Eq. (8). We then consider the test statistic given by

$$T_{C_\gamma}(\hat{\theta}_\alpha, \theta_0) = 2nC_\gamma(f_{\hat{\theta}_\alpha}, f_{\theta_0}), \tag{28}$$

where $\hat{\theta}_\alpha$ is the MDPDE of θ with tuning parameter α. The use of the estimator $\hat{\theta}_\alpha$ allows us to avoid the use of any nonparametric smoothing without disturbing the asymptotic null distribution. Notice that in the test statistic defined in Eq. (28) we allow the use of two distinct indices γ and α, although in the implementation part we often use $\gamma = \alpha$. In order to prove the asymptotic results related to the above test statistic, we will make the standard assumptions on model densities as given by Assumptions (A)–(D) of Lehmann (1983, p. 429), to be referred to herein as the "Lehmann conditions"; these conditions are provided in the Online Supplement. If the true data generating density is f_{θ_0}, we have

$$\sqrt{n}\left(\hat{\theta}_\alpha - \theta_0\right) \underset{n\to\infty}{\overset{\mathcal{D}}{\rightarrow}} N_p\left(0, J_\alpha(\theta_0)^{-1}V_\alpha(\theta_0)J_\alpha(\theta_0)^{-1}\right),$$

where $J_\alpha(\theta)$ and $V_\alpha(\theta)$ are as in Corollary 5. We then have the following theorem.

Theorem 10 *Suppose the model densities satisfy the Lehmann conditions. Then, under the null hypothesis $H_0 : \theta = \theta_0$, the test statistic $T_{C_\gamma}(\hat{\theta}_\alpha, \theta_0)$ has the same asymptotic distribution as the distribution of $\sum_{i=1}^r \zeta_i^{\gamma,\alpha}(\theta_0)Z_i^2$ where $Z_1, \ldots, Z_r$ are independent standard normal variables, $\zeta_1^{\gamma,\alpha}(\theta), \ldots, \zeta_r^{\gamma,\alpha}(\theta)$ are the nonzero eigenvalues of $J_\gamma(\theta)J_\alpha^{-1}(\theta)V_\alpha(\theta)J_\alpha^{-1}(\theta)$ and*

$$r = rank(J_\alpha^{-1}(\theta_0)V_\alpha(\theta_0)J_\alpha^{-1}(\theta_0)J_\gamma(\theta_0)J_\alpha^{-1}(\theta_0)V_\alpha(\theta_0)J_\alpha^{-1}(\theta_0)).$$

Notice that the asymptotic null distribution of the test statistic depends only on the two parameters γ and α. We now give an approximation to the power function of the proposed test based on the statistic in Eq. (28). Although the asymptotic null distribution is independent of the choice of $N(\cdot)$ of the C-divergence family (and hence the parameter λ of the GPD family), we will see that the approximate (asymptotic) power function depends on $C(\cdot,\cdot)$, and therefore obviously on $N(\cdot)$ (equivalently on λ for the GPD family).

Theorem 11 *Suppose the model densities satisfy the Lehmann conditions. An approximation to the power function of test statistic* (28) *for testing* (27) *at the significance level α_0 is given by*

$$\pi_{n,\alpha_0}(\theta^*) = 1 - \Phi\left(\frac{\sqrt{n}}{\sigma(\theta^*)}\left(\frac{t_{\alpha_0}^{\alpha,\gamma}}{2n} - C_\gamma(f_{\theta^*}, f_{\theta_0})\right)\right), \quad \theta^* \neq \theta_0 \tag{29}$$

where Φ is the standard normal distribution function, $t_{\alpha_0}^{\alpha,\gamma}$ is the $(1-\alpha_0)$th quantile of the asymptotic null distribution of $T_{C_\gamma}(\hat{\theta}_\alpha, \theta_0)$, and $\sigma^2(\theta)$ is defined as $\sigma^2(\theta) = M_{C_\gamma}(\theta)^{\mathrm{T}} J_\alpha^{-1}(\theta) V_\alpha(\theta) J_\alpha^{-1}(\theta) M_{C_\gamma}(\theta)$ with $M_{C_\gamma}(\theta) = \nabla C_\gamma(f_{\theta_0}, f_\theta)$.

6.2 Two-sample problem

We will now consider the case of two independent random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ of sizes n and m, respectively. We will assume that the corresponding population densities belong the parametric family $\mathcal{F} = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ with parameter values θ_1 and θ_2, respectively. Our objective is to test for the homogeneity of the two samples based on the observed data, which is equivalent to testing the hypothesis

$$H_0 : \theta_1 = \theta_2 \quad \text{against} \quad H_1 : \theta_1 \neq \theta_2. \tag{30}$$

As a natural extension of the previous cases, we will use the statistic

$$S_{C_\gamma}({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha) = \frac{2nm}{n+m} C_\gamma(f_{{}^{(1)}\hat{\theta}_\alpha}, f_{{}^{(2)}\hat{\theta}_\alpha}), \tag{31}$$

where ${}^{(1)}\hat{\theta}_\alpha$ and ${}^{(2)}\hat{\theta}_\alpha$ are the MDPDEs of θ_1 and θ_2, respectively. The next theorem presents the asymptotic null distribution of the proposed test statistic.

Theorem 12 *Suppose the model densities satisfy the Lehmann conditions. Then, under the null hypothesis $H_0 : \theta_1 = \theta_2$, the test statistic $S_{C_\gamma}({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha)$ has the same asymptotic distribution as the distribution of $\sum_{i=1}^r \zeta_i^{\gamma,\alpha}(\theta_1) Z_i^2$ where $Z_1, \ldots, Z_r$ and $\zeta_1^{\gamma,\alpha}(\theta), \ldots, \zeta_r^{\gamma,\alpha}(\theta)$ are as defined in* Theorem 10 *with*

$$r = rank(J_\alpha^{-1}(\theta_1) V_\alpha(\theta_1) J_\alpha^{-1}(\theta_1) J_\gamma(\theta_1) J_\alpha^{-1}(\theta_1) V_\alpha(\theta_1) J_\alpha^{-1}(\theta_1).$$

Once again, one can also derive an approximation to the power for the two-sample test based on $S_{C_\gamma}({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha)$; this approximation is given in the next theorem. The proof is similar to that in the one-sample case.

Theorem 13 *Suppose the model densities satisfy the Lehmann conditions. An approximation to the power function of the test statistic* $S_{C_\gamma}({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha)$ *for testing the hypothesis in* Eq. (30) *at the significance level* α_0 *is given by*

$$\pi_{m,n,\alpha_0}(\theta_1, \theta_2) = 1 - \Phi\left(\frac{\sqrt{\frac{nm}{n+m}}}{\bar{\sigma}(\theta_1, \theta_2)}\left(\frac{s_{\alpha_0}^{\alpha,\gamma}}{2}\frac{n+m}{nm} - C_\gamma(f_{\theta_1}, f_{\theta_2})\right)\right), \theta_1 \neq \theta_2$$

where Φ *is the standard normal distribution function,* $s_{\alpha_0}^{\alpha,\gamma}$ *is the* $(1-\alpha_0)th$ *quantile of the asymptotic null distribution of* $S_{C_\gamma}({}^{(1)}\hat{\theta}_\alpha, {}^{(2)}\hat{\theta}_\alpha)$ *and* $\bar{\sigma}(\theta_1, \theta_2)$ *is defined as*

$$\begin{aligned}\bar{\sigma}^2(\theta_1, \theta_2) = \omega G_\gamma^{\mathrm{T}} J_\alpha^{-1}(\theta_1) V_\alpha(\theta_1) J_\alpha^{-1}(\theta_1) G_\gamma \\ +(1-\omega) H_\gamma^{\mathrm{T}} J_\alpha^{-1}(\theta_2) V_\alpha(\theta_2) J_\alpha^{-1}(\theta_2) H_\gamma \end{aligned} \qquad (32)$$

with $G_\gamma = (g_1^\gamma, \ldots, g_p^\gamma)^{\mathrm{T}}$, $H_\gamma = (h_1^\gamma, \ldots, h_p^\gamma)^{\mathrm{T}}$ *and* $g_i^\gamma = \frac{\partial C(f_{\theta_1}, f_{\theta_2})}{\partial \theta_{1i}}$, $h_i^\gamma = \frac{\partial C(f_{\theta_1}, f_{\theta_2})}{\partial \theta_{2i}}$ *for* $i = 1, \ldots, p$, *and* $\omega \in (0, 1)$ *is the limit of* $\frac{m}{n+m}$ *as* $n, m \to \infty$.

Remark 3 While we have described the development of the test statistics with the continuous models in mind (where avoiding the nonparametric smoothing component provides a major benefit), we can continue with the actual minimizers of the C-divergence (rather than the MDPDEs) in discrete models. A two-sample test of hypothesis presented in Sect. 7.4 based on the Poisson model and the GPD family makes use of the MGPDEs, rather than just the MDPDEs. This does not change the asymptotic null distribution of the test statistics.

Remark 4 We have considered the simple null hypothesis when developing the test statistics in this section. With some additional machinery, these can be extended to handle composite nulls. For brevity, we do not consider that situation in this paper.

7 Examples

In this section, we present some examples and numerical simulations for the GPD family. Other possible choices of $N(\delta)$ are explored in Sect. 7.5.

7.1 Drosophila data: estimation problem

We consider the data presented by Woodruff et al. (1984) involving a sex-linked recessive lethal test in Drosophila (fruit flies). The first two rows of Table 3 show the frequencies of number of recessive lethal mutations observed among the daughters of male flies exposed to certain doses of a chemical to be screened. The remaining rows of the table present the estimated frequencies for several MGPDEs of θ under the Poisson(θ) model. A more detailed description of the estimators (without the frequencies) are given in Table 4. The observations at $x = 3$ and $x = 4$ appear to

Table 3 Fits of the Poisson model to the Drosophila data using MLE and the GPD methods with several λ and α. ML+D denotes the outlier deleted MLE

Values observed frequency	Recessive lethal count						$\hat{\theta}$
	0	1	2	3	4	≥ 5	
	23	3	0	1	1	0	
Fits based on MLE							
MLE	19.59	7.00	1.25	0.15	0.01	–	0.3571
ML+D	24.95	2.88	0.17	0.01	–	–	0.1154
Fits based on MGPDE with parameter λ, α (existing members)							
$\lambda = \alpha = -0.5$ ($DPD_{-0.5}$)	13.06	9.96	3.80	0.96	0.18	0.03	0.7624
$\lambda = -0.5, \alpha = 0$ ($PD_{-0.5}$)	24.70	3.09	0.19	0.01	–	–	0.1252
$\lambda = 0.5, \alpha = 0$ ($PD_{0.5}$)	15.58	9.13	2.68	0.52	0.08	0.01	0.5862
$\lambda = 0.5, \alpha = 0.5$ ($DPD_{0.5}$)	24.11	3.60	0.27	0.01	–	–	0.1493
$\lambda = 1, \alpha = 1$ (DPD_1)	23.78	3.88	0.32	0.02	–	–	0.1633
Fits based on MGPDE with parameter λ, α (new members)							
$\lambda = -0.5, \alpha = 0.3$	25.02	2.81	0.16	0.01	–	–	0.1125
$\lambda = -0.3, \alpha = 0.1$	25.12	2.72	0.15	0.01	–	–	0.1084
$\lambda = -0.1, \alpha = 0.2$	25.08	2.76	0.15	0.01	–	–	0.1101
$\lambda = 0, \alpha = 0.2$	24.81	3.00	0.18	0.01	–	–	0.1211
$\lambda = 0.1, \alpha = 0.3$	24.84	2.98	0.18	0.01	–	–	0.1199

Table 4 The MGPDEs of θ for various values of λ and α for the Drosophila data

$\lambda \downarrow \alpha \rightarrow$	−0.9	−0.7	−0.5	−0.3	−0.1	0	0.1	0.3	0.5	0.7	0.9
−0.9	0.65	0.62	0.46	0.24	0.1	0.07	0.07	0.09	0.11	0.12	0.14
−0.7	0.82	0.79	0.67	0.42	0.16	0.1	0.09	0.11	0.13	0.14	0.15
−0.5	0.9	0.87	0.76	0.55	0.23	0.13	0.1	0.11	0.13	0.14	0.15
−0.3	0.94	0.91	0.82	0.65	0.34	0.18	0.11	0.11	0.13	0.14	0.15
−0.1	0.97	0.94	0.85	0.71	0.46	0.29	0.15	0.11	0.13	0.15	0.16
0	0.98	0.95	0.86	0.73	0.51	0.36	0.2	0.12	0.13	0.15	0.16
0.1	0.99	0.96	0.88	0.75	0.55	0.42	0.26	0.12	0.13	0.15	0.16
0.3	1.01	0.97	0.9	0.79	0.62	0.51	0.39	0.17	0.13	0.15	0.16
0.5	1.02	0.98	0.91	0.81	0.67	0.59	0.48	0.27	0.15	0.15	0.16
0.7	1.03	0.99	0.92	0.83	0.71	0.64	0.56	0.37	0.21	0.15	0.16
0.9	1.04	0.99	0.93	0.85	0.75	0.68	0.61	0.45	0.29	0.18	0.16

represent moderate outliers. The rows corresponding to ML and ML+D represent the analysis based on the method of maximum likelihood on the full data and the outlier deleted data, respectively. It may be seen that the estimators corresponding to $\alpha > 0$, $\lambda < 0$, or small positive λ combined with moderately large positive α have the highest outlier stability and are closest to the outlier deleted MLE in magnitude. Note also that the five estimators in the lower block of Table 3 do not belong to any existing family of minimum distance estimators but yet are extremely close to the outlier deleted MLE.

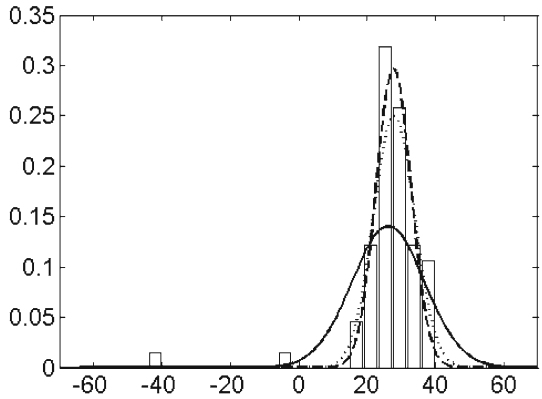


Fig. 7 Various fits on Newcomb's data (solid line: normal fit based on the MLE for full data, dashed line: normal fit based on the MLE for outlier deleted data, dotted line: normal fit using $\mathrm{GPD}_{\lambda=-0.5,\alpha=0.5}$ for full data)

7.2 Newcomb's Data

This example has been taken from Stigler (1977, Table 5) and involves data on speed of light as reported by S. Newcomb. The data clearly show (see histogram in Fig. 7) that there is a dominant bell-shaped structure, which is, however, blemished by two large outliers. The maximum likelihood estimates of the location parameter and the scale parameter for the full data under the $N(\mu, \sigma^2)$ model come out to be 26.21 and 10.66, respectively; without these two outliers, the estimates shift to 27.75 and 5.04, respectively. To calculate the minimum GPD estimates, we have used the kernel density estimator with the Gaussian kernel for the construction of the divergence. The bandwidth ν_n has been taken as $\nu_n = 1.06\sigma_n n^{-1/5}$ where $\sigma_n = \mathrm{median}_i\, |X_i - \mathrm{median}_j\, X_j|/0.6745$. We have used the smoothed model approach (Basu and Lindsay 1994) where the model is smoothed with the same kernel function. The estimates of μ and σ are given in Tables 5 and 6. The observations match our previous findings in that the (α, λ) combinations which we have so far been found to be more robust continue to provide superior performance in this case as well.

7.3 The hypothesis testing problem: a simulation exercise

In this section, we consider an extensive simulation study where we compare the test statistics based on GPD with some other possible competitors. Consider the set up presented in Sect. 6.1, and the hypothesis considered in Eq. (27). For the GPD, the test statistic presented in Eq. (28) reduces to

$$T_{\mathrm{GPD}_{\alpha,\lambda}}(\hat{\theta}_{\alpha,\lambda}, \theta_0) = 2n\mathrm{GPD}_{\alpha,\lambda}(f_{\hat{\theta}_{\alpha,\lambda}}, f_{\theta_0}) \tag{33}$$

where $\hat{\theta}_{\alpha,\lambda}$ is the MGPDE of θ with parameters (α, λ). For comparison, we consider Wald-type test statistics based on the popular M-estimators given by

$$W_\psi = n(\widehat{\theta}_\psi - \theta_0)^{\mathrm{T}}(\Sigma_\psi(\widehat{\theta}_\psi))^{-1}(\widehat{\theta}_\psi - \theta_0), \tag{34}$$

where $\hat{\theta}_\psi$ is an M-estimator of θ defined as the solution of the estimating equation $\sum_{i=1}^n \psi(X_i, \theta) = 0$, for some suitable ψ-function (see, e.g., Hampel et al. 1986).

Table 5 The MGPDEs of μ for the Newcomb data for various values of λ and α

$\lambda \downarrow \alpha \rightarrow$	0	0.1	0.2	0.3	0.4	0.5
−0.9	27.7249	27.6968	27.6551	27.6146	27.5754	27.5374
−0.8	27.7282	27.7015	27.6589	27.6179	27.5784	27.5402
−0.7	27.7309	27.7064	27.6628	27.6212	27.5814	27.5430
−0.6	27.7327	27.7115	27.6668	27.6244	27.5843	27.5457
−0.5	27.7331	27.7172	27.6710	27.6277	27.5872	27.5484
−0.4	27.7300	27.7235	27.6756	27.6312	27.5902	27.5512
−0.3	27.7171	27.7306	27.6808	27.6348	27.5932	27.5539
−0.2	27.6739	27.7379	27.6870	27.6389	27.5963	27.5567
−0.1	27.5418	27.7408	27.6942	27.6436	27.5997	27.5595
0	26.2121	27.7173	27.7013	27.6491	27.6034	27.5625
0.1	26.9916	27.5759	27.7001	27.6552	27.6077	27.5657
0.2	26.7532	26.7253	27.6489	27.6582	27.6125	27.5693
0.3	26.2675	26.5729	26.7820	27.6381	27.6162	27.5732

Table 6 The MGPDEs of σ for the Newcomb data for various values of λ and α

$\lambda \downarrow \alpha \rightarrow$	0	0.1	0.2	0.3	0.4	0.5
−0.9	4.9647	4.9291	4.9406	4.9387	4.9234	4.9008
−0.8	4.9786	4.9347	4.9465	4.9450	4.9296	4.9065
−0.7	4.9920	4.9382	4.9512	4.9505	4.9353	4.9121
−0.6	5.0059	4.9393	4.9543	4.9553	4.9406	4.9174
−0.5	5.0229	4.9376	4.9556	4.9591	4.9455	4.9226
−0.4	5.0490	4.9327	4.9545	4.9617	4.9498	4.9274
−0.3	5.1026	4.9246	4.9504	4.9627	4.9534	4.9320
−0.2	5.2405	4.9161	4.9423	4.9615	4.9560	4.9362
−0.1	5.6150	4.9234	4.9302	4.9572	4.9571	4.9398
0	10.6636	5.0142	4.9196	4.9491	4.9560	4.9424
0.1	15.2984	5.4804	4.9426	4.9395	4.9521	4.9435
0.2	17.1436	12.6237	5.1509	4.9448	4.9460	4.9426
0.3	18.2700	15.3856	11.0103	5.0495	4.9472	4.9399

Under the null hypothesis and some regularity conditions, $\sqrt{n}(\widehat{\theta_\psi} - \theta_0)$ is asymptotically p-variate normal with variance $\Sigma_\psi(\theta_0)$ and hence the Wald-type test statistics W_ψ has an asymptotic χ^2 distribution with p degrees of freedom. In this paper, we consider two particular M-estimators—(i) the optimal B-robust M-estimator corresponding to the Huber's ψ-function given by

$$\psi_c(x,t) = \begin{cases} -c & \text{for } \ x \leq \sqrt{t}(\beta - c), \\ +c & \text{for } \ x \geq \sqrt{t}(\beta + c), \\ \frac{x}{\sqrt{t}} - \beta & \text{for } \ \text{otherwise}, \end{cases} \tag{35}$$

Table 7 Simulated levels for the GPD tests with pure data under the Poisson model (sample size 20) for various values of λ and α; the nominal level is 5%

$\lambda \downarrow \alpha \rightarrow$	−0.4	−0.3	−0.2	−0.1	0	0.1	0.2	0.3	0.4
−0.9	0.182	0.315	0.322	0.296	0.262	0.228	0.189	0.165	0.138
−0.8	0.068	0.192	0.211	0.206	0.187	0.156	0.137	0.122	0.104
−0.7	0.037	0.119	0.148	0.146	0.123	0.117	0.107	0.097	0.089
−0.6	0.024	0.087	0.096	0.099	0.096	0.098	0.087	0.076	0.069
−0.5	0.021	0.076	0.083	0.08	0.081	0.075	0.075	0.068	0.061
−0.4	0.021	0.064	0.079	0.074	0.071	0.066	0.059	0.06	0.055
−0.3	0.027	0.063	0.072	0.067	0.062	0.057	0.053	0.048	0.044
−0.2	0.032	0.068	0.069	0.062	0.057	0.05	0.045	0.045	0.039
−0.1	0.037	0.074	0.069	0.062	0.056	0.049	0.041	0.037	0.033
0	0.045	0.087	0.078	0.06	0.052	0.042	0.037	0.035	0.031

and (ii) the optimal V-robust re-descending M-estimator corresponding to the hyperbolic tangent (tanh) ψ-function given by

$$\psi_c(x,t) = c \tanh\left(\frac{x-\beta}{c\sqrt{t}}\right), \tag{36}$$

where $\beta = \beta(t)$ is obtained from $\int \psi_c(x,t)dF_t = 0$ in both cases; Simpson et al. (1987) studied the asymptotic properties of these estimators under discrete models including our example of the Poisson model.

In addition, we provide some divergence-based competitors of the GPD based tests. In particular, we choose two members of the tests based on the S-divergence family (see Ghosh et al. 2015, and Ghosh and Basu 2016 for details of the use of S-divergence in hypothesis testing problems). The form of the S-divergence based on tuning parameters (λ, α) is based on the $N(\cdot)$ function given by Eq. (9), and the actual tuning parameters used in relation to this divergence in described are the relevant table.

We have considered the Poisson model in our simulations. Let θ represent the mean parameter of interest, and we are interested in testing the null hypothesis $H_0 : \theta = 5$ against $H_1 : \theta \neq 5$. The entire exercise was based on samples of size 20 with 1000 replications. For the first study, data were generated from the Poisson(5) distribution. The empirical level of the test for this scenario with pure data was computed as the proportion of test statistics (out of the 1000 replications) that exceed the upper 5% critical value of the $\chi^2(1)$ distribution. For the GPD-based tests, these levels are presented in Table 7. Contaminated data were then generated from the 0.9 Poisson(5) + 0.1 Poisson(25) mixture, and the same set of hypotheses were tested with these contaminated data. The empirical levels of the GPD tests, under these contaminated data, were computed as in the previous cases and are presented in Table 8. Pure data were then generated from the Poisson(3) distribution, and the same hypotheses were again tested to determine the empirical power of the tests under pure data. The empirical powers for the GPD are presented in Table 9. Contaminated data were

Table 8 Simulated levels for the GPD tests with contaminated data under the Poisson model (sample size 20) for various values of λ and α; the nominal level is 5%

λ ↓ α →	−0.4	−0.3	−0.2	−0.1	0	0.1	0.2	0.3	0.4
−0.9	0.881	0.873	0.748	0.267	0.311	0.272	0.236	0.206	0.177
−0.8	0.868	0.879	0.814	0.389	0.225	0.205	0.183	0.165	0.143
−0.7	0.865	0.873	0.845	0.511	0.177	0.162	0.145	0.124	0.11
−0.6	0.867	0.874	0.851	0.597	0.128	0.127	0.11	0.099	0.091
−0.5	0.871	0.871	0.86	0.662	0.104	0.101	0.09	0.082	0.073
−0.4	0.872	0.872	0.864	0.736	0.085	0.089	0.076	0.07	0.061
−0.3	0.873	0.874	0.868	0.788	0.078	0.078	0.065	0.059	0.057
−0.2	0.873	0.878	0.873	0.834	0.093	0.074	0.057	0.049	0.051
−0.1	0.875	0.878	0.877	0.858	0.207	0.082	0.049	0.044	0.046
0	0.876	0.88	0.878	0.872	0.707	0.105	0.048	0.039	0.039

Table 9 Simulated powers for the GPD tests with pure data under the Poisson model (sample size 20) for various values of λ and α; the nominal level is 5%

λ ↓ α →	−0.4	−0.3	−0.2	−0.1	0	0.1	0.2	0.3	0.4
−0.9	0.999	1	0.998	0.998	0.994	0.99	0.983	0.979	0.969
−0.8	0.998	1	1	0.999	0.995	0.99	0.989	0.98	0.969
−0.7	0.989	1	1	1	0.997	0.991	0.989	0.985	0.971
−0.6	0.984	1	1	1	0.998	0.992	0.991	0.987	0.973
−0.5	0.979	0.999	1	0.998	0.998	0.995	0.992	0.988	0.978
−0.4	0.97	0.996	1	0.998	0.998	0.996	0.992	0.986	0.978
−0.3	0.96	0.993	0.999	0.998	0.997	0.996	0.992	0.986	0.981
−0.2	0.95	0.991	0.997	0.998	0.997	0.996	0.993	0.987	0.982
−0.1	0.939	0.99	0.994	0.998	0.996	0.995	0.991	0.987	0.982
0	0.934	0.988	0.993	0.996	0.997	0.995	0.991	0.986	0.983

Table 10 Simulated powers for the GPD tests for contaminated data under the Poisson model (sample size 20) for various values of λ and α; the nominal level is 5%

λ ↓ α →	−0.4	−0.3	−0.2	−0.1	0	0.1	0.2	0.3	0.4
−0.9	0.513	0.593	0.746	0.992	0.994	0.991	0.984	0.976	0.966
−0.8	0.563	0.589	0.635	0.969	0.996	0.993	0.986	0.979	0.967
−0.7	0.586	0.598	0.586	0.919	0.995	0.994	0.992	0.981	0.969
−0.6	0.614	0.608	0.561	0.831	0.995	0.994	0.992	0.981	0.966
−0.5	0.64	0.631	0.563	0.728	0.989	0.995	0.991	0.982	0.964
−0.4	0.665	0.645	0.567	0.637	0.986	0.995	0.991	0.98	0.968
−0.3	0.687	0.672	0.583	0.536	0.977	0.992	0.991	0.981	0.968
−0.2	0.703	0.702	0.612	0.513	0.938	0.989	0.989	0.98	0.967
−0.1	0.727	0.731	0.648	0.514	0.812	0.984	0.988	0.978	0.966
0	0.742	0.759	0.684	0.578	0.535	0.962	0.982	0.975	0.965

Table 11 Simulated levels and powers with pure and contaminated (Cont.) data under the Poisson model using different competitive tests (sample size 20); the nominal level is 5%

Test		S(−1, 0.5)	S(−0.5, 0.3)	MH(1)	MH(1.5)	MH(2)	MT
Level	Pure data	0.072	0.065	0.055	0.045	0.039	0.105
	Cont. data	0.087	0.080	0.057	0.070	0.119	0.116
Power	Pure data	0.953	0.994	0.875	0.894	0.903	0.992
	Cont. data	0.926	0.987	0.672	0.675	0.668	0.984

S(λ, α): S-divergence-based tests with tuning parameters (λ, α)
MH(c): M-estimator-based Wald-type tests for the Huber ψ-function with different c
MT: M-estimator-based Wald-type tests for the tan-hyperbolic ψ-function with $c = 1$

Table 12 Frequencies of the number of recessive lethal daughters for the Drosophila data used in the two-sample testing problem

x	0	1	2	3	4	5	6	7
Observed (control)	159	15	3	0	0	0	0	0
Observed (treated)	110	11	5	0	0	0	1	1

then generated from the 0.9 Poisson(3) + 0.1 Poisson(15) mixture, and the empirical powers under these contaminated data were computed. For the GPD-based test, these powers are presented in Table 10. The pure and contaminated levels and powers of the S-divergence-based tests and the M-estimator-based tests are presented in Table 11.

A quick comparison shows that several members of the GPD-based tests, particularly those with small positive values of α and small negative values of λ provide excellent compromise between pure data efficiency and stability under contaminated data. These tests have minimal inflation in level and minimal drop in power under contamination and clearly have superior performance compared to the S-divergence-based tests and the M-estimator- based tests.

7.4 A two-sample hypothesis testing problem with Drosophila data

In this section, we will consider a two-sample Drosophila dataset, where the male flies in the treated group were exposed to 2000 µg butyraldehyde; the other group was exposed only to control conditions. As in Sect. 7.1, the observations in Table 12 represent the frequencies of the number of recessive lethal mutations among daughter flies. See Woodruff et al. (1984) and Simpson (1989) for more details of the experimental set up. The responses are assumed to be Poisson with means θ_1 (control group) and θ_2 (treated group), respectively. The frequencies corresponding to $x = 6, 7$ for the treated group can be considered as outliers. We will demonstrate that our methods provide stable inference discounting the effects of these outliers.

Let $p(\theta)$ denote the Poisson probability mass function with parameter θ, n and m are the sample sizes from the two populations, ${}^{(1)}\hat{\theta}_{\alpha,\lambda}$ and ${}^{(2)}\hat{\theta}_{\alpha,\lambda}$ are the MGPDEs of the two population parameters corresponding to α and λ, and ${}^{(0)}\hat{\theta}_{\alpha,\lambda}$ is the common

Table 13 The MGPDEs of the Poisson parameters for the two-sample Drosophila data (the numbers in the bracket show the corresponding estimates after deleting the two outliers)

	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$
${}^{(1)}\hat{\theta}_{\alpha,\lambda=0.1}$	0.1201	0.1135	0.1044	0.1003	0.0995	0.1003
${}^{(2)}\hat{\theta}_{\alpha,\lambda=0.1}$	0.3967	0.1745	0.1266	0.118	0.1179	0.1221
	(0.1716)	(0.1545)	(0.1272)	(0.1144)	(0.1125)	(0.1152)
${}^{(0)}\hat{\theta}_{\alpha,\lambda=0.1}$	0.2568	0.1375	0.1136	0.1077	0.1072	0.1095
	(0.1415)	(0.1303)	(0.1137)	(0.1063)	(0.105)	(0.1066)

MGPDE under the null. We consider the hypothesis

$$H_0 : \theta_1 = \theta_2 \quad \text{against} \quad H_1 : \theta_1 \neq \theta_2. \tag{37}$$

As described in Remark 3, the test statistic for testing hypothesis (37) is given, under obvious modifications of the symbols in Theorem 12, by

$$^{*}S_{\gamma,\lambda}({}^{(1)}\hat{\theta}_{\alpha,\lambda}, {}^{(2)}\hat{\theta}_{\alpha,\lambda}) = \frac{1}{\zeta({}^{(0)}\hat{\theta}_{\alpha,\lambda})} \frac{2nm}{n+m} \mathrm{GPD}_{\gamma,\lambda}(p({}^{(1)}\hat{\theta}_{\alpha,\lambda}), p({}^{(2)}\hat{\theta}_{\alpha,\lambda})),$$

where $\zeta({}^{(0)}\hat{\theta}_{\alpha,\lambda}) = A_{\gamma}({}^{(0)}\hat{\theta}_{\alpha,\lambda}) V_{\alpha}({}^{(0)}\hat{\theta}_{\alpha,\lambda}) J_{\alpha}^{-2}({}^{(0)}\hat{\theta}_{\alpha,\lambda})$ with $A_{\gamma}(\theta) = \sum_{x=0}^{\infty} \left(\frac{e^{-\theta}\theta^x}{x!}\right)^{1+\gamma} \left(\frac{x}{\theta} - 1\right)^2$, $J_{\alpha}(\theta) = \sum_{x=0}^{\infty} \left(\frac{e^{-\theta}\theta^x}{x!}\right)^{1+\alpha} \left(\frac{x}{\theta} - 1\right)^2$ and $V_{\alpha}(\theta) = \sum_{x=0}^{\infty} \left(\frac{e^{-\theta}\theta^x}{x!}\right)^{1+2\alpha} \left(\frac{x}{\theta} - 1\right)^2 - \left[\sum_{x=0}^{\infty} \left(\frac{e^{-\theta}\theta^x}{x!}\right)^{1+\alpha} \left(\frac{x}{\theta} - 1\right)\right]^2$. We have used $\gamma = \alpha$ in our simulations. The change in the estimators over α is displayed for some specific cases in Table 13. The Chi-square (1 degree of freedom) p values for various values of λ and α with and without outliers are shown in Tables 14 and 15, respectively. The numbers again show that the $\alpha > 0$ and $\lambda < 0$ combinations are most successful in providing stable decisions with or without the outliers.

7.5 Inference with other forms of $N(\delta)$: combined disparities

While the results of our paper are general, all our illustrations so far have been with respect to the GPD, where the $N(\delta)$ function coincides with the function $\xi_{\lambda}(\delta)$ of the power divergence family. In this section, we consider other functions for $N(\delta)$ beyond the ordinary power divergence family. While many forms may be contemplated, we have taken a combined function approach to enhance the scope of applicability of our methods where we technically go beyond the GPD class, but use the nature of the functions within the GPD class in producing divergences with improved performances. We consider the general form of the C-divergence as given in Eq. (8), but to differentiate between the treatment of positive (outliers) and negative (inliers) Pearson residuals, we propose to combine two distinct disparity generating functions within the GPD family corresponding to two different values of λ. It is clear that a negative value

Table 14 The p values of the GPD-based tests for different tuning parameters with $\gamma = \alpha$ for the two-sample Drosophila data with outliers

$\lambda \downarrow \alpha \rightarrow$	0	0.1	0.3	0.5	0.7	0.9
−0.9	0.664	0.778	0.802	0.759	0.685	0.611
−0.7	0.551	0.709	0.775	0.751	0.682	0.608
−0.5	0.459	0.644	0.741	0.739	0.677	0.605
−0.3	0.359	0.587	0.694	0.721	0.67	0.601
−0.1	0.115	0.547	0.636	0.696	0.66	0.595
0	0.006	0.472	0.609	0.679	0.655	0.593
0.1	0	0.179	0.587	0.661	0.648	0.59
0.3	0	0	0.461	0.618	0.627	0.581
0.5	0	0	0.004	0.549	0.602	0.57
0.7	0	0	0	0.055	0.563	0.555
0.9	0	0	0	0	0.193	0.534

Table 15 The p values of the GPD-based tests for different tuning parameters with $\gamma = \alpha$ for the two-sample Drosophila data without outliers

$\lambda \downarrow \alpha \rightarrow$	0	0.1	0.3	0.5	0.7	0.9
−0.9	0.665	0.764	0.86	0.858	0.828	0.797
−0.7	0.554	0.688	0.835	0.852	0.826	0.796
−0.5	0.463	0.609	0.803	0.844	0.825	0.795
−0.3	0.383	0.526	0.759	0.829	0.819	0.794
−0.1	0.315	0.444	0.698	0.807	0.813	0.791
0	0.288	0.406	0.664	0.792	0.809	0.789
0.1	0.262	0.371	0.627	0.775	0.803	0.788
0.3	0.222	0.312	0.549	0.73	0.786	0.783
0.5	0.194	0.266	0.476	0.674	0.762	0.774
0.7	0.174	0.232	0.412	0.61	0.73	0.761
0.9	0.159	0.208	0.359	0.546	0.687	0.744

of λ is more appropriate for dealing with outliers. However, there is a large body of the literature which indicates that such values of λ are not adequate in dealing with inliers, observations with less data than what is expected under the model; see, e.g., Basu et al. (2011). The inlier concerns are better handled by positive values of λ. To accommodate all these considerations, one could consider, for example, the C-divergence corresponding to a disparity generating function $N(\delta)$ of the form

$$N(\delta) = \begin{cases} \xi_{\lambda=-0.5}(\delta) & \text{if } \delta \geq 0, \\ \xi_{\lambda=1}(\delta) & \text{if } \delta < 0, \end{cases} \tag{38}$$

and some fixed value of α. We would not expect the minimum divergence estimator generated by the above divergence (which we will loosely refer to as the combined GPD) to be any inferior (in terms of robustness) to the minimum GPD estimator corresponding to $\lambda = -0.5$ and the same α, as the modification in Eq. (38) has not tampered with the outlier controlling capability of the divergence. However, we do

Table 16 The MSEs of the estimators corresponding to different values of α for the ordinary GPD at $\lambda = -0.5$, and the corresponding combined divergence defined in Eq. (38)

α	0	0.2	0.4	0.6	0.8
Pure data ($\epsilon = 0$)					
Ordinary GPD ($\lambda = -0.5$)	0.2873	0.3313	0.3806	0.4351	0.5023
Combined GPD	0.2439	0.2764	0.3241	0.3832	0.4588
Contaminated data ($\epsilon = 0.1$)					
Ordinary GPD ($\lambda = -0.5$)	0.3524	0.4152	0.4804	0.5492	0.6063
Combined GPD	0.2978	0.3129	0.37	0.4493	0.5159

expect such modifications to improve the small sample efficiencies of the minimum divergence estimators, as they handle the inliers better. We choose data from the $(1-\epsilon)\text{Poisson}(5) + \epsilon\text{Poisson}(25)$ mixture and, assuming that the data come from a pure Poisson model, estimate the value of the mean parameter θ. We do this for $\epsilon = 0, 0.1$, so that the performances are evaluated at pure data as well as at a moderate level of contamination. In Table 16, we present the MSEs of our estimators against the target parameter 5, using the GPD at $\lambda = -0.5$ and several values of α, as well as the combined GPD as defined in (38) with the same values of α. The sample size is 20, and the number of replications is 1000. It may be easily seen that in each case there is a very substantial improvement in the mean square error of the estimator due to the use of the combined disparity generating function.

8 Conclusion

Statistical divergences, as many authors have noticed, have a natural role in robust inference. Density-based divergences, in particular, often combine a high degree of efficiency with strong robustness properties. The power divergence, the density power divergence, the class of disparities, the family of S-divergences and the class of generalized S-divergences all come under the larger umbrella of the C-divergence family described in this paper. In this sense, the latter class provides a very large scheme of density-based minimum divergence methods, which essentially subsumes all the major existing single-integral density-based procedures. Here we have studied many important properties of the inference procedures based on the class of C-divergences. Extensive numerical evidence has been provided to substantiate the theoretical properties studied in the paper. On the whole, we expect that this consolidates much of the previous work done on density-based divergences and could present a wide range of options for the practitioner.

Acknowledgements We wish to thank the editor, the associate editor and the anonymous referees for their careful reading of the manuscript and several constructive suggestions which have significantly improved the paper.

References

Ali, S. M., Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society B*, *28*, 131–142.

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

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

Basu, A., Mandal, A., Martin, N., Pardo, L. (2013). Testing statistical hypotheses based on the density power divergence. *Annals of the Institute of Statistical Mathematics*, *65*, 319–348.

Basu, A., Shioya, H., Park, C. (2011). *Statistical inference*: *The minimum distance approach*. Boca Raton: Chapman & Hall/CRC.

Beran, R. J. (1977). Minimum Hellinger distance estimates for parametric models. *Annals of Statistics*, *5*, 445–463.

Broniatowski, M., Vajda, I. (2012). Several applications of divergence criteria in continuous families. *Kybernetika*, *48*, 600–636.

Broniatowski, M., Toma, A., Vajda, I. (2012). Decomposable pseudodistances and applications in statistical estimation. *Journal of Statistical Planning and Inference*, *142*, 2574–2585.

Cressie, N., Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society B*, *46*, 440–464.

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

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, *2*, 299–318.

Fujisawa, H., Eguchi., S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, *99*, 2053–2081.

Ghosh, A. (2015). Asymptotic properties of minimum *S*-divergence estimator for discrete models. *Sankhya A*, *77*, 380–407.

Ghosh, A., Basu, A. (2016). Testing composite null hypotheses based on S-divergences. *Statistics and Probability Letters*, *114*, 38–47.

Ghosh, A., Basu, A. (2018). A new family of divergences originating from model adequacy tests and application to robust statistical inference. *IEEE Transactions on Information Theory*. https://doi.org/10.1109/TIT.2018.2794537.

Ghosh, A., Basu, A., Pardo, L. (2015). On the robustness of a divergence based test of simple statistical hypotheses. *Journal of Statistical Planning and Inference*, *161*, 91–108.

Ghosh, A., Harris, I. R., Maji, A., Basu, A., Pardo, L. (2017). A generalized divergence for statistical inference. *Bernoulli*, *23*, 2746–2783.

Ghosh, A., Maji, A., Basu, A. (2013). Robust inference based on divergences in reliability systems. In I. Frenkel, A. Karagrigoriou, A. Lisnianski, A. Kleyner (Eds.), *Applied reliability engineering and risk analysis. probabilistic models and statistical inference. Dedicated to the Centennial of the birth of Boris Gnedenko*. New York: Wiley.

Hampel, F. R., Ronchetti, E., Rousseeuw, P. J., Stahel, W. (1986). *Robust statistics*: *The approach based on influence functions*. New York: Wiley.

Jana, S., Basu, A. (2018). A characterization of all single-integral, non-kernel divergence estimators. *Technical Report, ISRU/2018/1*. Kolkata: Indian Statistical Institute.

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

Lehmann, E. L. (1983). *Theory of point estimation*. New York: Wiley.

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

Mattheou, K., Leeb, S., Karagrigoriou, A. (2009). A model selection criterion based on the BHHJ measure of divergence. *Journal of Statistical Planning and Inference*, *139*, 228–235.

Pardo, L. (2006). *Statistical inference based on divergences*. Boca Raton: CRC/Chapman-Hall.

Park, C., Basu, A. (2004). Minimum disparity estimation: Asymptotic normality and breakdown point results. *Bulletin of Informatics and Cybernetics*, *36*, 19–33. (special issue in Honor of Professor Takashi Yanagawa).

Patra, S., Maji, A., Basu, A., Pardo, L. (2013). The power divergence and the density power divergence families: The mathematical connection. *Sankhya B*, *75*, 16–28.

Rousseeuw, P. J., Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Simpson, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *Journal of the American Statistical Association*, *82*, 802–807.

Simpson, D. G. (1989). Hellinger deviance test: Efficiency, breakdown points, and examples. *Journal of the American Statistical Association*, *84*, 107–113.

Simpson, D. G., Carroll, R. J., Ruppert, D. (1987). M-estimation for discrete data: Asymptotic distribution theory and implications. *Annals of Statistics*, *15*, 657–669.

Stigler, S. M. (1977). Do robust estimators work with real data? *Annals of Statistics*, *5*, 1055–1098. (with discussion).

Tamura, R. N., Boos, D. D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, *81*, 223–229.

Vonta, F., Karagrigoriou, A. (2010). Generalized measures of divergences in survival analysis and reliability. *Journal of Applied Probability*, *47*, 216–234.

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