Ann Inst Stat Math (2019) 71:627–656
https://doi.org/10.1007/s10463-018-0665-x

Statistical inference based on bridge divergences

Arun Kumar Kuchibhotla[1] · **Somabha Mukherjee**[1] · **Ayanendranath Basu**[2]

Received: 27 June 2017 / Revised: 3 February 2018 / Published online: 17 May 2018

Abstract M-estimators offer simple robust alternatives to the maximum likelihood estimator. The density power divergence (DPD) and the logarithmic density power divergence (LDPD) measures provide two classes of robust M-estimators which contain the MLE as a special case. In each of these families, the robustness of the estimator is achieved through a density power down-weighting of outlying observations. Even though the families have proved to be useful in robust inference, the relation and hierarchy between these two families are yet to be fully established. In this paper, we present a generalized family of divergences that provides a smooth bridge between DPD and LDPD measures. This family helps to clarify and settle several longstanding issues in the relation between the important families of DPD and LDPD, apart from being an important tool in different areas of statistical inference in its own right.

Keywords Divergence · Robustness · M-estimators

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-018-0665-x) contains supplementary material, which is available to authorized users.

✉ Ayanendranath Basu
ayanbasu@isical.ac.in

Arun Kumar Kuchibhotla
arunku@wharton.upenn.edu

Somabha Mukherjee
somabha@wharton.upenn.edu

[1] University of Pennsylvania, 3730 Walnut St, Philadelphia, PA 19104, USA

[2] Indian Statistical Institute, 203, Barrackpore Trunk Road, Kolkata, West Bengal 700108, India

1 Introduction

Statistical procedures based on minimization of divergences are popular in the literature. In our context, a divergence is a distance like dissimilarity measure between two distributions which does not necessarily demand metric properties. Density-based minimum divergence methods are very useful as they combine high efficiency with strong robustness properties. Basu et al. (1998) proposed the class of density power divergences (DPD). These divergences, when constructed between an arbitrary density g and a parametric model density f_θ are indexed by a nonnegative robustness tuning parameter α and have the form

$$\rho_1^{(\alpha)}(g, f_\theta) = \int \left\{ f_\theta^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) g f_\theta^{\alpha} + \frac{1}{\alpha} g^{1+\alpha} \right\}. \tag{1}$$

The divergence at $\alpha = 0$ is defined as the continuous limit of the expression in Eq. (1) as $\alpha \to 0$. This generates the divergence

$$\rho_1^{(0)}(g, f_\theta) = \int g \log\left(\frac{g}{f_\theta}\right), \tag{2}$$

where log represents natural logarithm. Observe that the only term on the right hand side of Eq. (1) containing both g and f_θ has g in degree one. A divergence with such a property has been referred to as a decomposable divergence by some authors; see, for example, Broniatowski et al. (2012).

Another class of divergences with some similar properties was introduced by Jones et al. (2001) and was followed up by Fujisawa and Eguchi (2008), Broniatowski et al. (2012) and Fujisawa (2013) among others. In spite of its formal similarity with the DPD, this family, referred to herein as the logarithmic density power divergence (LDPD) family, was originally developed following the robust model fitting idea of Windham (1995). This class also has a nonnegative robustness tuning parameter α and is defined as

$$\rho_0^{(\alpha)}(g, f_\theta) = \log \int f_\theta^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \log \int g f_\theta^{\alpha} + \frac{1}{\alpha} \log \int g^{1+\alpha}. \tag{3}$$

The divergence $\rho_0^{(0)}$ is once again defined as a limiting case and some simple algebra shows that $\rho_0^{(0)}(g, f_\theta) = \rho_1^{(0)}(g, f_\theta)$, the common divergence being a version of the Kullback–Leibler divergence.

Jones et al. (2001) also provided a general form of divergence measures in terms of two tuning parameters α and ϕ given by

$$\rho_\phi^{(\alpha)}(g, f_\theta) = \frac{1}{\phi}\left(\int f_\theta^{1+\alpha}\right)^{\phi} - \frac{1}{\phi}\left(1 + \frac{1}{\alpha}\right)\left(\int g f_\theta^{\alpha}\right)^{\phi} + \frac{1}{\alpha\phi}\left(\int g^{1+\alpha}\right)^{\phi}, \tag{4}$$

for $0 \le \phi \le 1$ and $\alpha \ge 0$. The DPD and the LDPD measures are recovered from this general form for $\phi = 1$ and $\phi = 0$, the second one being obtained as the limiting case as $\phi \to 0$. Accordingly, the DPD and the LDPD families have also been referred to as type 1 and type 0 divergences in the literature. A scaled version of the LDPD has also

been called the γ-divergence by Fujisawa and Eguchi (2008). Both the DPD and the LDPD measures have proved to be useful additions to the literature of robust parameter estimation based on divergences. Both families, as well as the corresponding minimum divergence estimators, have been heavily cited in the literature and applied to many practical problems, and a thorough exploration of their relationship and hierarchy can help us develop better compromises. In either case, a simple density power down-weighting indicates the source of robustness of the resulting estimator. In both cases, the parameter α controls the trade-off between robustness and efficiency; smaller values of α lead to greater model efficiency and larger values of α lead to greater outlier stability. The minimum divergence estimator corresponding to $\alpha = 0$ is the maximum likelihood estimator in either case.

In terms of comparison between the minimum DPD and the minimum LDPD estimators, Jones et al. (2001) expressed a (weak) preference for the former. In particular, they observed that the presence of observations very close to zero could lead to spurious global minimum in case of the LDPD under the exponential model. On the other hand, Fujisawa and Eguchi (2008) and Fujisawa (2013) have claimed that the minimum LDPD estimators exhibit a greater relative stability under heavy contamination leading to smaller bias and smaller mean squared error. They argued that the small bias that the minimum LDPD estimator has is due to an approximate Pythagorean relation that the LDPD satisfies. In contrast, Broniatowski et al. (2012), based on their own simulation study, do not report any particular relative advantage for the minimum LDPD estimator.

These points raise certain unsettled issues involving these two useful classes of divergences and corresponding estimators which we hope to at least partially reconcile in the present paper. The generalized family of divergences in Eq. (4), useful as it is, does not generate minimum divergence estimators that are legitimate M-estimators except when $\phi = 0$ or 1. We will construct an alternative generalized class of divergences providing a bridge between these classes where several of the intermediate divergences lead to reasonable compromises between the positives of these two families. This new family will be called the family of *bridge density power divergences* and the generated estimators are all M-estimators. This will provide the user with the flexibility of choosing a suitable estimator from a larger class.

The new family of divergences will depend on two tuning parameters which are (1) the robustness parameter α and (2) the bridge parameter λ. It would be of interest to choose the tuning parameters adaptively with respect to the proportion of outliers so that the estimation is optimal in an appropriate sense. The robustness tuning parameter α should be close to zero for pure data and should assume a moderately large positive value in case of contaminated data. In this respect, Hong and Kim (2001) proposed the first method of choosing the tuning parameter by minimizing an estimate of the asymptotic variance and Warwick and Jones (2005) refined this process by using the mean squared error criterion together with a pilot estimate. We provide some justification for using a modified Hong and Kim (2001) procedure; see Sect. 8.

Before concluding this section, we summarize the main achievements of this paper.

1. We introduce a new family of divergences, the *Bridge Density Power Divergences* (BDPD) which produce a smooth link between the DPD and the LDPD; each

intermediate divergence is decomposable and leads to an M-estimator. Apart from helping to understand the relations between the DPD and the LDPD, this family is important in its own right, and in specific cases the performance of the intermediate divergences turn out to be superior than both the two marginal divergences (DPD and LDPD).

2. We demonstrate that the spurious root problem observed in case of the LDPD as noticed by Jones et al. (2001) is not an isolated problem for the exponential model, and is a more general phenomenon.
3. It is shown that the results and assertions of Fujisawa and Eguchi (2008) and Fujisawa (2013) are substantially correct, but only when a number of residual concerns are answered. In particular, the observed superior performance of the "minimum LDPD estimator", is, most of the time, achieved at a local minimum of the LDPD measure (rather than a global one), and has to be viewed as a weighted method of moments estimator rather than a minimum divergence estimator.
4. We demonstrate that the weighted method of moment equations corresponding to the LDPD (and indeed many other BDPD measures) can potentially throw up multiple roots; this is a real problem with practical implications. As the desired root is not necessarily the minimizer of the corresponding divergence, there is no automatic root selection strategy. As a choice of the incorrect solution can be disastrous, it is imperative that a clear answer to this question is provided in the literature, where this issue is so far unexplored.
5. We clearly define the target parameter when using the LDPD for parametric estimation. Our results in this paper show that it cannot, in general, be the global minimizer of the LDPD, nor any arbitrary root of the weighted method of methods equation. Our description of the target depends on an actual numerical algorithm for its selection (see item 6 below). This interpretation of the target parameter actually extends to the entire BDPD family, except the ordinary DPD.
6. Finally, we provide an algorithm for the selection of the suitable root of the weighted method of moment equations for any member of the BDPD family, including the LDPD. We also provide a method for the selection of "optimal" tuning parameters for analyzing the real data.

The rest of the paper is organized as follows. In Sect. 2, we introduce the basic parametric setup and also provide the construction leading to the family of bridge divergences. Various properties of the estimators including strong consistency, asymptotic normality and robustness are discussed in Sect. 3. In Sect. 4, we provide a heuristic explanation for the spurious minimum behavior of the LDPD measures in the case of small outliers. Section 5 rigorously proves that the LDPD is bound to fail in some specific examples. In Sect. 6, we address issues regarding multiple roots of the LDPD estimating equation and well-definedness of the estimators as discussed in the existing literature. We also introduce an algorithm for getting hold of a good root of the LDPD and all the other BDPD estimating equations. In Sect. 7, we provide the results of a simulation study conducted to support the claims made in previous sections. In Sect. 8, we give some directions on how to choose the tuning parameters. Finally, we conclude with a few remarks in Sect. 9.

2 The bridge density power divergence (BDPD) family

The problem of parameter estimation considered in this paper uses the following setup and notation. Let $\mathcal{G}$ denote the set of all probability distributions having densities with respect to some σ-finite base measure μ. We assume that the data generating distribution G and the model family $\mathcal{F}_\Theta = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ belong to $\mathcal{G}$. Let g and f_θ be the corresponding densities (with respect to μ). Let $X_1, X_2, \ldots, X_n$ be an independent and identically distributed (iid) random sample from G which is to be modeled by the family $\mathcal{F}_\Theta$. Our aim is to estimate the parameter θ by choosing the model density which gives the "closest fit" to the data. In this paper, we quantify the "closeness" using the density power divergences, the logarithmic density power divergences or their generalizations as mentioned in the previous section. However, we will see in the latter sections that this idea of closeness will require a refinement for most of the minimum BDPD estimators. All the integrals considered in this paper including those in the previous section are with respect to the measure μ.

In order to estimate θ based on an iid sample, one needs to construct an empirical estimate of the divergence. In this regard, note that the first terms of Eqs. (1) and (3), depending only on the model density f_θ, need no estimation. The third terms are independent of θ and so do not figure in any minimization process over θ. For the second terms, note that the integral involved can be written as $\mathbb{E}_g f_\theta^\alpha(X)$ and so can be estimated by the average of $f_\theta^\alpha(X_i)$, $1 \le i \le n$. This is a consequence of the decomposability property referred to in Sect. 1. In light of this discussion, the parameter estimates based on the divergences $\rho_1^{(\alpha)}$ and $\rho_0^{(\alpha)}$ are given by

$$\hat{\theta}_{n1}^\alpha := \arg\min_{\theta\in\Theta} \left[\int f_\theta^{1+\alpha} - \left(1+\frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^n f_\theta^\alpha(X_i) \right], \tag{5}$$

and

$$\hat{\theta}_{n0}^\alpha := \arg\min_{\theta\in\Theta} \left[\log\left(\int f_\theta^{1+\alpha} \right) - \left(1+\frac{1}{\alpha}\right) \log\left(\frac{1}{n} \sum_{i=1}^n f_\theta^\alpha(X_i) \right) \right]. \tag{6}$$

Under certain regularity conditions allowing the interchange of the derivative and the integral, the estimating equations corresponding to the above divergences are given, respectively, by

$$\int f_\theta^{1+\alpha} u_\theta = \frac{1}{n} \sum_{i=1}^n f_\theta^\alpha(X_i) u_\theta(X_i), \tag{7}$$

and

$$\left(\frac{1}{n} \sum_{i=1}^n f_\theta^\alpha(X_i) \right) \int f_\theta^{1+\alpha} u_\theta = \left(\frac{1}{n} \sum_{i=1}^n f_\theta^\alpha(X_i) u_\theta(X_i) \right) \int f_\theta^{1+\alpha}. \tag{8}$$

Here, $u_\theta(x)$ represents the likelihood score function given by $u_\theta(x) = \nabla \log f_\theta(x)$. Throughout this manuscript, we use the symbols ∇, ∇_2 to denote the first and the second derivatives with respect to θ. Equations (7) and (8) demonstrate that both the minimum DPD and the minimum LDPD estimators are legitimate M-estimators.

Later in this section and in Sect. 6, we will consider the weighted method of moments representation of these estimating equations.

In this paper, we will show that the DPD and the LDPD families can be combined in a larger super-family of divergences where each intermediate divergence class leads to an M-estimator unlike the generalization given in Eq. (4). For notational simplicity in the following derivation, define

$$t_1(\theta) := \int f_\theta^{1+\alpha}, \quad t_2(\theta) := \int g f_\theta^{\alpha}, \quad t_3(\theta) := \int g^{1+\alpha}.$$

We consider the density versions (in terms of the true density g) of Eqs. (7) and (8) rather than the versions based on the empiricals. These equations are given by

$$\int f_\theta^{1+\alpha} u_\theta = \int f_\theta^{\alpha} g u_\theta, \tag{9}$$

and

$$\int f_\theta^{\alpha} g \int f_\theta^{\alpha} u_\theta = \int f_\theta^{\alpha} g u_\theta \int f_\theta^{\alpha}. \tag{10}$$

Rewriting the estimating Eqs. (9) and (10) in terms of t_1 and t_2, we get

$$\begin{aligned} \left[\frac{\nabla t_1(\theta)}{\alpha+1} - \frac{\nabla t_2(\theta)}{\alpha}\right] &= 0, \\ \left[\frac{t_2(\theta)\nabla t_1(\theta)}{\alpha+1} - \frac{t_1(\theta)\nabla t_2(\theta)}{\alpha}\right] &= 0. \end{aligned}$$

For $\lambda \in [0, 1]$, consider an estimating equation which equates a convex combination of the above two estimating functions to zero; it is given explicitly by

$$\lambda\left[\frac{t_1'(\theta)}{\alpha+1} - \frac{t_2'(\theta)}{\alpha}\right] + (1-\lambda)\left[\frac{t_2(\theta)t_1'(\theta)}{\alpha+1} - \frac{t_1(\theta)t_2'(\theta)}{\alpha}\right] = 0. \tag{11}$$

Rearranging the terms on the left hand side leads to the differential equation,

$$\begin{aligned} \frac{t_1'(\theta)}{1+\alpha}\left[\lambda + \bar{\lambda} t_2(\theta)\right] &= \frac{t_2'(\theta)}{\alpha}\left[\lambda + \bar{\lambda} t_1(\theta)\right], \\ \frac{t_1'(\theta)}{\lambda + \bar{\lambda} t_1(\theta)} &= \left(\frac{1+\alpha}{\alpha}\right)\frac{t_2'(\theta)}{\lambda + \bar{\lambda} t_2(\theta)}, \end{aligned} \tag{12}$$

where $\bar{\lambda} = 1 - \lambda$. It is now easy to derive the corresponding objective function which turns out to be

$$\frac{1}{\bar{\lambda}}\log(\lambda + \bar{\lambda} t_1(\theta)) - \frac{1}{\bar{\lambda}}\left(\frac{1+\alpha}{\alpha}\right)\log(\lambda + \bar{\lambda} t_2(\theta)) + C,$$

for some constant C independent of θ. Imposing the condition that the divergence has to be zero when $g = f_\theta$, the constant C is recovered to be $\frac{1}{\bar\lambda}\frac{1}{\alpha}\log(\lambda+\bar\lambda t_3(\theta))$. Hence the new divergence can be written as

$$\rho^{(\alpha,\lambda)}(g,\, f_\theta) = \frac{1}{\bar\lambda}\log(\lambda+\bar\lambda t_1(\theta)) - \frac{1}{\bar\lambda}\left(\frac{1+\alpha}{\alpha}\right)\log(\lambda+\bar\lambda t_2(\theta)) + \frac{1}{\bar\lambda}\frac{1}{\alpha}\log(\lambda+\bar\lambda t_3(\theta)). \tag{13}$$

This is our class of *bridge density power divergences*, defined for $\alpha \geq 0$ and $\lambda \in [0, 1]$. For $\lambda = 0$, this reduces to the class of logarithmic density power divergences; as $\lambda \to 1$, we recover the class of density power divergences. It is also easy to verify that as $\alpha \to 0$, the limiting divergence is the Kullback–Leibler divergence as given in Eq. (2) irrespective of the value of λ. Due to the consideration of efficiency, we will, in the rest of the paper, restrict the robustness parameter α to the range $[0, 1]$.

The estimating Eq. (12) for the bridge divergence with parameters (α, λ) can be written as:

$$\frac{\int f_\theta^{1+\alpha} u_\theta}{\lambda + \bar\lambda \int f_\theta^{1+\alpha}} = \frac{\frac{1}{n}\sum_{i=1}^n f_\theta^\alpha(X_i) u_\theta(X_i)}{\lambda + \bar\lambda \frac{1}{n}\sum_{i=1}^n f_\theta^\alpha(X_i)}.$$

This may be referred to as the BDPD moment equations. It is immediately obvious that the above equation represents a very rich class of score moment equations. When $\alpha = 0$, the above represents the ordinary likelihood score equation irrespective of the value of λ. More generally it represents a score moment equation where the scores are variably weighted by powers of densities depending on the parameter α, and the weights are variably normalized depending on the parameter λ. When $\lambda = 0$, the weights are perfectly normalized (as in the case of the LDPD), while we get the non-normalized equations for $\lambda = 1$, as in the case of the DPD. For intermediate values of λ we get partially normalized estimating equations, with the degree of normalization dropping off with increasing λ. We will observe the effect of this normalization throughout the rest of the paper, most notably in Sect. 4, where we will give some indication of its role in the occurrence of the spurious roots.

Remark 1 The BDPD family was constructed using an appropriate combination of the DPD and the LDPD estimating equations with the same tuning parameter α; there is no specific reason, apart from mathematical convenience, for considering the same α in both the estimating equations. However, if estimating equations with different values of α are combined it does not appear to lead to a genuine objective function for such an estimating equation.

Remark 2 Equation (11) cannot be obtained by starting directly with a convex combination of the DPD and the LDPD with the same value of α.

Remark 3 It is possible to combine the DPD and the LDPD families in ways other than the BDPD and Eq. (4). See, for example, Kanamori and Fujisawa (2014) who provide two other combined families containing these two divergences, those based on the Bregman scores and the Hölder scores. Based on the form of the potential function, these authors classify the Bregman scores into two types, the separable Bregman scores and the nonseparable Bregman scores, the former containing the density power

score and the latter containing the γ-score as special cases. The density power score and the γ-score are also special cases of the Hölder scores. The density power score corresponds to the DPD, and the γ-score corresponds to the LDPD.

3 Properties of minimum BDPD estimators

Based on the bridge density power divergence defined in the previous section and an iid random sample $X_1, X_2, \ldots, X_n$ from a distribution G, an estimator of θ is given by

$$\hat{\theta}_n^{(\alpha,\lambda)} := \arg\min_{\theta\in\Theta} \frac{1}{\bar{\lambda}} \log(\lambda + \bar{\lambda} t_1(\theta)) - \frac{1}{\bar{\lambda}} \left(\frac{1+\alpha}{\alpha}\right) \log\left(\lambda + \bar{\lambda} \frac{1}{n} \sum_{i=1}^{n} f_\theta^\alpha(X_i)\right).$$

By definition the BDPD leads to an M-estimator and so its asymptotic and robustness properties can be derived from the well-established M-estimation theory. Therefore, we present the results for consistency and asymptotic normality of the bridge divergence estimator without proofs. For this, we use the following notation.

$$P_{i,\gamma} = \int g^i f_\theta^\gamma, \quad Q_{i,\gamma} = \int g^i f_\theta^\gamma u_\theta u_\theta^\top, \quad R_{i,\gamma} = \int g^i f_\theta^\gamma u_\theta, \quad S_{i,\gamma} = \int g^i f_\theta^\gamma \nabla u_\theta.$$

We now present a consistency theorem with conditions in alignment with those in Wald's consistency theorem (Ferguson (1996, Chapter 17)) for the maximum likelihood estimator and as the proof is essentially the same, it is moved to the Supplementary Material.

Theorem 1 (Consistency) *Suppose that the following assumptions hold.*

(C1) Θ *is a compact metric space;*

(C2) There exists a function $K(x)$ *(independent of* θ*) such that* $|\varphi_\alpha(x)| \le K(x)$ *and* $K(X)$ *has finite expectation (with respect to* G*). Here,* $\varphi_\alpha(x) = f_\theta^\alpha(x)$ *for* $\alpha > 0$ *and* $\varphi_0(x) = \log f_\theta(x)$*;*

(C3) For each x *and any sequence* $\theta_n \to \theta$*,*

$$\lim_{n\to\infty} f_{\theta_n}(x) = f_\theta(x),$$

for all x*, except possibly on a set (which might depend on* θ *but not on the sequence* $\{\theta_n\}$*) of* μ*-measure zero. Also,* $\theta_g^{(\alpha,\lambda)}$ *is the unique minimizer of the bridge density power divergence* $\rho^{(\alpha,\lambda)}(g, f_\theta)$*.*

Then the minimum BDPD estimator $\hat{\theta}_n^{(\alpha,\lambda)}$ *is strongly consistent for* $\theta_g^{(\alpha,\lambda)}$ *for any fixed* $\alpha \in [0, 1]$ *and* $\lambda > 0$*.*

Remark 4 Note that the conditions are independent of the value of λ and match those of Wald's consistency theorem when $\alpha = 0$. The consistency theorem only requires Θ to be a metric space and not an Euclidean space. The proof can be extended to settings

other than iid samples using various generalizations of the uniform strong law of large numbers (USLLN).

Remark 5 The main ingredient in the proof of Theorem 1 is Theorem 5.7 of van der Vaart (1998) that uses uniform convergence of the sample-based objective functions to their population counterparts. For $\lambda > 0$, the function $\xi \mapsto \log(\lambda + \bar{\lambda}\xi)$ is Lipschitz on any bounded closed interval I that does not contain 0. This fact allows one to conclude the following implication (under (C2)):

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} f_\theta^\alpha(X_i) - \int g f_\theta^\alpha\right| = o_p(1), \quad \text{(follows from uniform SLLN)}$$
$$\Rightarrow \sup_{\theta\in\Theta}\left|\log\left(\lambda + \bar{\lambda}\frac{1}{n}\sum_{i=1}^{n} f_\theta^\alpha(X_i)\right) - \log\left(\lambda + \bar{\lambda}\int g f_\theta^\alpha\right)\right| = o_p(1).$$

At $\lambda = 0$ the Lipschitz property fails and the above implication breaks down. This is the reason for restricting the range of λ to $\lambda > 0$. Thus the above proof does not automatically cover the DPD family. To extend the consistency claim to $\lambda = 0$, one needs to make the additional assumption $\inf_{\theta\in\Theta}\int g f_\theta^\alpha > 0$.

The proof of asymptotic normality of the minimum BDPD estimator can be obtained through a quadratic approximation of the objective function. Using Cramér-Rao type regularity conditions, we also observe that there exists a sequence of roots of the estimating equation which is consistent and asymptotically normal. We do not repeat the conditions here but refer the reader to Theorem 2 of Basu et al. (1998). The proof is similar to Lehmann's proof (Lehmann and Casella (1998)) of consistency and asymptotic normality of the MLE, and is hence omitted.

Theorem 2 (Asymptotic Normality) *Under certain regularity conditions, there exists a sequence of roots θ_n of the bridge divergence estimating equation which is consistent and $\sqrt{n}(\theta_n - \theta_g^{(\alpha,\lambda)})$ has an asymptotically normal distribution with mean 0 and variance given by $J(\theta_g)^{-1}K(\theta_g)J(\theta_g)^{-1}$, where $\theta_g := \theta_g^{(\alpha,\lambda)}$,*

$$\begin{aligned}
K(\theta) &= (1-\lambda)^2 P_{1,2\alpha} R_{0,\alpha+1} R_{0,\alpha+1}^\top - \bar{\lambda}\left[\lambda + \bar{\lambda}P_{0,\alpha+1}\right] R_{0,\alpha+1} R_{1,2\alpha}^\top \\
&\quad - \bar{\lambda}\left[\lambda + \bar{\lambda}P_{0,\alpha+1}\right] R_{1,2\alpha} R_{0,\alpha+1}^\top + \left[\lambda + \bar{\lambda}P_{0,\alpha+1}\right]^2 Q_{1,2\alpha} \\
&\quad - \bar{\lambda}^2 P_{1,\alpha}^2 R_{0,\alpha+1} R_{0,\alpha+1}^\top + \bar{\lambda}\left[\lambda + \bar{\lambda}P_{0,\alpha+1}\right] P_{1,\alpha} R_{1,\alpha} R_{0,\alpha+1}^\top \\
&\quad + \bar{\lambda}\left[\lambda + \bar{\lambda}P_{0,\alpha+1}\right] P_{1,\alpha} R_{0,\alpha+1} R_{1,\alpha}^\top - \left[\lambda + \bar{\lambda}P_{0,\alpha+1}\right]^2 R_{1,\alpha} R_{1,\alpha}^\top \\
J(\theta) &= (\alpha+1) Q_{0,\alpha+1}\left[\lambda + \bar{\lambda}P_{1,\alpha}\right] + S_{0,\alpha+1}\left[\lambda + \bar{\lambda}P_{1,\alpha}\right] \\
&\quad - \alpha Q_{1,\alpha}\left[\lambda + \bar{\lambda}P_{0,\alpha+1}\right] - S_{1,\alpha}\left[\lambda + \bar{\lambda}P_{0,\alpha+1}\right] \\
&\quad + \bar{\lambda}\alpha R_{1,\alpha} R_{0,\alpha+1}^\top - \bar{\lambda}(\alpha+1) R_{0,\alpha+1} R_{1,\alpha}^\top.
\end{aligned}$$

3.1 Pythagorean relation

Fujisawa and Eguchi (2008) have claimed that under heavy contamination, the minimum LDPD estimator achieves a smaller bias (compared to the minimum DPD estimator); it is also suggested that this phenomenon can be partially explained by an approximate Pythagorean relation which the LDPD satisfies. We present a sequence of results which shows that such a Pythagorean relation holds in general for all BDPD (except the DPD), so that the LDPD result is a special case of the more general result. For any $0 \leq t \leq 1$, let $\bar{t} = 1 - t$. Define the cross-entropy between any two densities g and f for $\lambda, \alpha \in [0, 1]$ by,

$$d_{\lambda,\alpha}(g, f) = \frac{1}{\bar{\lambda}} \frac{1}{1+\alpha} \log\left(\lambda + \bar{\lambda} \int f^{1+\alpha}\right) - \frac{1}{\alpha\bar{\lambda}} \log\left(\lambda + \bar{\lambda} \int g f^{\alpha}\right).$$

The divergence induced by this cross-entropy is given by

$$D_{\lambda,\alpha}(g, f) = -d_{\lambda,\alpha}(g, g) + d_{\lambda,\alpha}(g, f).$$

which is a scaled version of $\rho^{\lambda,\alpha}(g, f)$, satisfying $D_{\lambda,\alpha}(\cdot, \cdot) = \rho^{\lambda,\alpha}(\cdot, \cdot)/(1+\alpha)$.

Theorem 3 *Let f and δ be given probability density functions and let $0 \leq \varepsilon \leq 1$. If $g(\cdot) = (1-\varepsilon) f(\cdot) + \varepsilon\delta(\cdot)$ and h is any positive function, then for any $\alpha \in [0, 1]$ and $\lambda \in [0, 1)$,*

$$d_{\lambda,\alpha}(g, h) = d_{\lambda,\alpha}(f, h) - \frac{1}{\alpha\bar{\lambda}} \log(1-\varepsilon) + O(T_{\epsilon,\delta}),$$

where

$$T_{\varepsilon,\delta} := \frac{\varepsilon}{1-\varepsilon} \left[\lambda + \bar{\lambda} \int \delta h^{\alpha}\right] \Big/ \alpha\bar{\lambda} \left[\lambda + \bar{\lambda} \int f h^{\alpha}\right].$$

For $\lambda = 1$, we have

$$d_{1,\alpha}(g, h) = d_{1,\alpha}(f, h) + \frac{\varepsilon}{\alpha} \left[\int \{f - \delta\} h^{\alpha}\right].$$

Proof See Section S.2 of the supplementary material for a proof. □

Remark 6 This theorem can be used to prove a Pythagorean relation (for $0 \leq \lambda < 1$) similar to Theorem 3.2 of Fujisawa and Eguchi (2008). The statement of the result is given below for completeness; however, the proof is very similar to the one in Fujisawa and Eguchi (2008) and is hence omitted. The Fujisawa and Eguchi (2008) result thus becomes a particular case of the following theorem.

Theorem 4 (Pythagorean Relation) *Let f and δ be given probability density functions and let $0 \leq \varepsilon \leq 1$. If $g(\cdot) = (1-\varepsilon) f(\cdot) + \varepsilon\delta(\cdot)$ and h is any positive function, then for any $0 \leq \lambda < 1$,*

$$\Delta(g, f, h) := D_{\lambda,\alpha}(g, h) - D_{\lambda,\alpha}(g, f) - D_{\lambda,\alpha}(f, h) = O(\varepsilon\nu),$$

where

$$\nu = \lambda + \bar{\lambda} \max\left\{\int \delta f^{\alpha}, \int \delta h^{\alpha}\right\}.$$

Remark 7 As was pointed out by one of the referees, Theorem 4 gives a practically useful relation only for $\nu \approx 0$. This holds only when $\lambda \approx 0$, which corresponds to the LDPD or a divergence close to it in the λ scale. However, the equality as stated is always true and might provide information about how large $\Delta(g, f, h)$ can be.

Remark 8 Theorem 4 is one of the main theoretical reasons for the behavior of the minimum BDPD estimators to be observed in the subsequent sections over various choices of tuning parameters and various contaminating distributions. Note that the function

$$\xi(\lambda) = \frac{1}{\bar{\lambda}} \frac{\lambda + \bar{\lambda} a}{\lambda + \bar{\lambda} b},$$

for fixed positive values of a and b, is necessarily increasing in $\lambda \in [0, 1)$ if and only if $a \leq b$. This implies that when

$$\int \delta h^{\alpha} \leq \int f h^{\alpha}, \tag{14}$$

the error term $T_{\varepsilon,\delta}$ is an increasing function of $\lambda \in [0, 1)$ and so the bias of the minimum BDPD estimator is expected to be an increasing function of $\lambda \in [0, 1)$; however, as the Pythagorean relation does not exist for the DPD, calculations based on $\xi(\lambda)$ are not helpful in theoretically comparing the bias of the minimum DPD estimator. The condition (14) with $h = f_{\theta}$ and $f = f_{\theta_0}$ for $\theta, \theta_0 \in \Theta$ is implied by the usual approximate singularity condition in the robustness literature; the latter condition dictates that the contaminating distribution (in this case represented by δ) in the gross error model is approximately singular with the parametric family so that $\int \delta h^{\alpha} \approx 0$ and $\int f h^{\alpha}$ is relatively large. In the numerical simulations, we will see that this reasoning generally fits in well with the observed behavior of the estimators except for very small values of α; in the latter case the quantities [in Eq. (14)] can be very close and the above observations may not hold.

But when the contaminating distribution δ is concentrated at (or around) the mode of the model density $h = f_{\theta_0}$, the integral $\int \delta h^{\alpha}$ is often at least as large as $\int f h^{\alpha}$ so that the changing behavior of $\xi(\lambda)$ over λ with $a = \int \delta h^{\alpha}$ and $b = \int f h^{\alpha}$, is less predictable. We will observe in the rest of the paper that the performance of the

minimum LDPD estimator (and several other minimum BDPD estimators) may suffer badly in this case.

In the following sections, we will consider two kinds of contamination for the parametric model. The first case will correspond to the approximate singularity idea of the gross error model where the contaminating (minor) component is well-separated from the target (major) component. We will refer to this as *outer contamination*; in this case the contaminating values will be *surprising observations* in the sense of Lindsay (1994). In the second case of contamination, we will choose the contaminating component near the mode of the major component so that these observations are no longer surprising observations but will nevertheless distort the shape of the distribution relative to the parametric model. We will refer to this case as *inner contamination*.

4 Spurious behavior under inner contamination

Jones et al. (2001) reported that in case of the exponential distribution, *small outliers* (near the origin) made the minimum LDPD estimator nonrobust. These small outliers are essentially what we have described as constituting inner contamination in Sect. 3. Let $\eta_\theta(x)$ denote the density of the exponential distribution with mean θ. The nonrobustness of the minimum LDPD estimator under small outliers was demonstrated by Jones et al. (2001) in simulation studies and was confirmed by determining the LDPD between the densities η_θ and the $0.85\eta_1 + 0.15\delta_{x_0}$ mixture with $x_0 = 0.0001$ where δ_{x_0} represents the indicator function at $x = x_0$. In Figs. 1 and 2, we have exhibited the BDPD objective function between the above two densities over θ for $\alpha = 0.5$ and several values of λ in $[0, 1]$. It is clear that the global minimizer of the objective function remains stuck at a value very close to zero at least up to $\lambda = 0.7$. It is noteworthy that this spurious minimum may easily be missed by any gradient descent root search algorithm since the minimum is obtained as a needle sharp notch over a very limited region. It might very well be possible (and indeed it is the case in Figs. 1 and 2) that a local minimizer of the LDPD may serve as a reasonable robust solution even in this case. But the asymptotic results for a sequence of roots of the estimating equation present in the literature do not prescribe the method to *choose* a suitable sequence of roots in case of multiple roots. In any case, it is clear that the global minimizer of the LDPD up to (at least) $\lambda = 0.7$ is a nonsensical value.

The main reason for this phenomenon in the LDPD and the bridge divergences close to it is the behavior of the log function at 0, where it diverges to $-\infty$. If one writes the LDPD between the η_θ and $0.85\eta_1 + 0.15\delta_{0.0001}$ densities, we can easily check that there is a sharp drop in the objective function around a value very close to zero because of the above logarithm effect (although at 0, the objective function is positive infinity). This effect slowly smooths out as the λ parameter increases since the presence of the additive λ term in each of the logarithms eventually forces the argument to be bounded away from zero. The validity of this reasoning is confirmed by examining the behavior of the LDPD and the BDPD close to it in case of the $N(0, \sigma^2)$ model, where it can be expected that the estimate of σ will be driven to zero if there are some outliers near 0 in the sample. For that matter, this behavior can also be seen in case of the $N(\mu, \sigma^2)$ model with some outliers near the true mean. Figure 3 exhibits the LDPD at

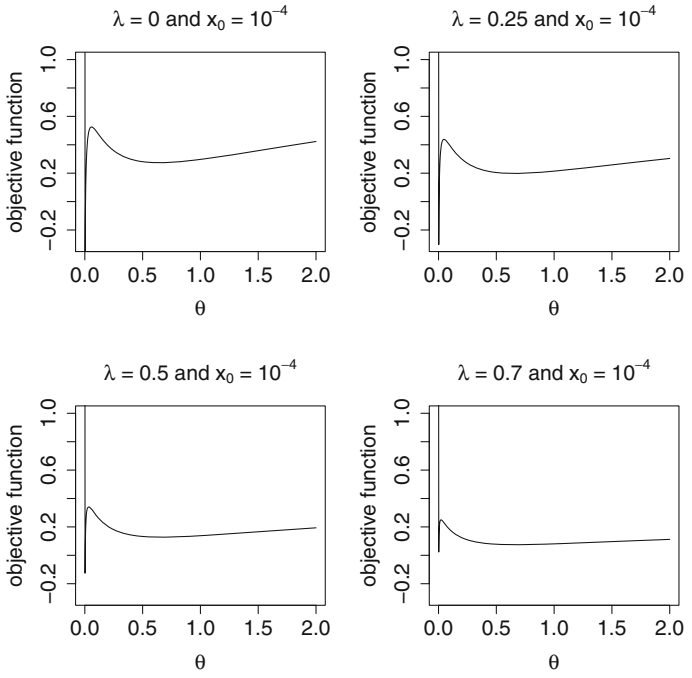

Fig. 1 Plots of the BDPD objective function ($\lambda = 0.0, 0.25, 0.5, 0.7$)

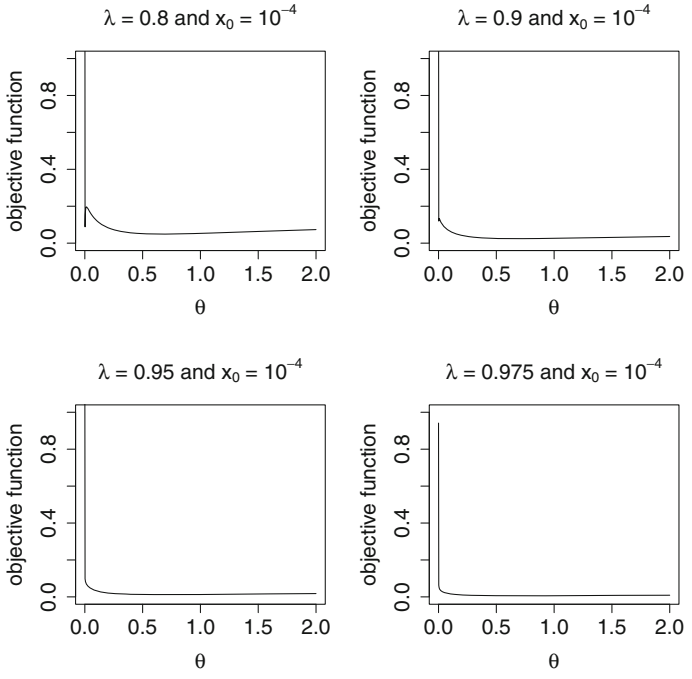

Fig. 2 Plots of the BDPD objective function ($\lambda = 0.8, 0.9, 0.95, 0.975$)

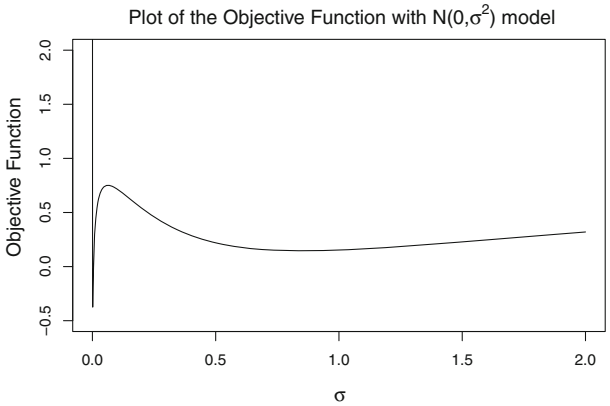

Fig. 3 Plot of the LDPD objective function under the $N(0, \sigma^2)$ model

$\alpha = 0.5$ as a function of σ when computed between the densities of the $N(0, \sigma^2)$ and the mixture $0.85N(0, 1) + 0.15\delta_{0.001}$. Clearly, there is a reasonable local minimum around $\sigma = 0.8$; but it is beaten hands down by the useless global minimum in the neighborhood of zero.

While the behavior observed in Figs. 1, 2 and 3 represent the patterns in the divergences between the actual densities, such behavior can also be observed under pure data (although it is relatively rare in scale models). An actual random sample of size 20 from $N(0, 1)$ with seed 129 was obtained in R as

$$\begin{aligned} &-1.120900, -0.989724, -1.374697, -1.355645, 1.996755, \quad 0.695870, \\ &0.771968, -0.002847, \quad 1.008549, -0.990280, 1.131772, -0.244929, \\ &1.186625, -1.671537, -0.081999, -1.831365, 0.358867, \quad 0.891639, \\ &0.489801, \quad 0.000010. \end{aligned}$$

The exact value of the last observation is 1.048653×10^{-5}, which forces a spurious behavior in the LDPD objective function plotted in Fig. 4 for $\alpha = 0.8$. To be precise, if $f_\sigma(x)$ denotes the probability density of $N(0, \sigma^2)$ with respect to the Lebesgue measure, then

$$\int f_\sigma^{\alpha+1}(x)dx = \frac{1}{\sqrt{\alpha+1}(\sqrt{2\pi}\sigma)^\alpha},$$

and the LDPD objective function satisfies

$$\begin{aligned} M_n(\sigma) &\leq \log\left(\frac{1}{\sqrt{\alpha+1}(\sqrt{2\pi}\sigma)^\alpha}\right) - \left(\frac{\alpha+1}{\alpha}\right)\log\left(\frac{1}{n}f_\sigma^\alpha(X_{20})\right) \\ &= \log\left(n^{1+1/\alpha}\sigma\right) + \frac{1}{2}\log\left(\frac{2\pi}{\alpha+1}\right) + \left(\frac{\alpha+1}{2}\right)\frac{X_{20}^2}{\sigma^2}. \end{aligned}$$

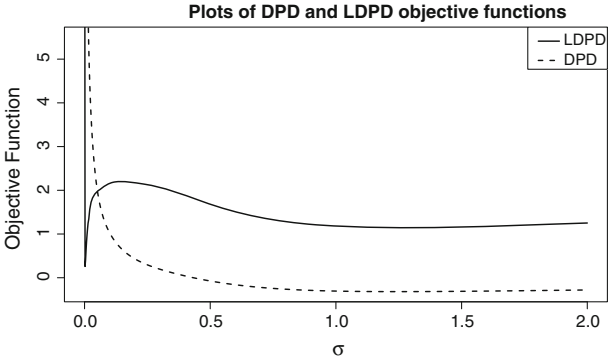

Fig. 4 Plots of the sample-based DPD and LDPD objective functions under the $N(0, \sigma^2)$ model based on a sample of size 20 from $N(0, 1)$

where X_{20} is the last observation which is close to zero and $n = 20$. This inequality confirms our reasoning. As σ slides down toward zero, for a while the first term involving the logarithm dominates, pulling the objective function down. However, as σ gets really close to zero, the third term takes over and there is an extremely sharp rise in the objective function, which generates a minimum with a razor sharp notch. Figure 4 shows the spurious (global) minimum near zero (at $\sigma = 0.00001309906$ to be exact), although there is a reasonable local minimum at $\sigma = 1.265882$. For this sample, the spurious global minimum phenomenon is observed for all $\alpha \geq 0.500144205$ and all $\lambda \in [0, 0.228827308]$. In contrast, the global minimum of the DPD objective function (also plotted in Fig. 4) for $\alpha = 0.8$ is obtained at $\sigma = 1.20833$. If the last observation 1.048653×10^{-5} were removed from this data set, this spurious minimum behavior of the LDPD objective function disappears. The minimum LDPD estimator now equals $\sigma = 1.321807$ at $\alpha = 0.8$, an entirely reasonable value. (The corresponding minimum DPD estimator is 1.298228). Thus one single observation can bring about an absolutely drastic change in the minimum LDPD estimator which is against the spirit of stability that robust estimators should have; so far we (or indeed anybody else), have not detected such spurious behavior in any scenario involving the DPD.

Thus the spurious root issue in case of the LDPD is not a isolated problem limited to the case of the exponential distribution. It is, in fact, a more serious problem in the case of the location-scale model (compared to just the scale model), as we will see in the next section.

5 Unboundedness of the bridge divergences

In the previous section, we explained the reason for observing a spurious minimum with LDPD in the case of inner contamination and argued that it might happen even in case of real data generated from the pure model. Intuitively, this may be explained by the fact that the probability of observing a value near the mode of the majority distribution is not too small at moderate sample sizes.

Here, we will formally prove that the sample version of the BDPD objective function (given on the right hand side of Eq. (13)) is unbounded below in case the parametric model family is a location-scale family $\mathcal{G}_{f,\Theta}$, where f is a probability density function, $\Theta = \mathbb{R} \times \mathbb{R}^+$, and

$$\mathcal{G}_{f,\Theta} = \left\{ f_\theta(\cdot) : \ f_\theta(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) \text{ for some } \theta := (\mu, \sigma) \in \Theta \right\}.$$

This result proved in Theorem 5 implies that one cannot fit a location-scale family using the minimizer of a BDPD, except the DPD.

Theorem 5 *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed observations from a density g, which is modeled by the densities in the location-scale family of distributions $\mathcal{G}_{f,\Theta}$ with a fixed f satisfying $f(0) > 0$. Then, for any $0 \le \lambda < 1$ and $\alpha > 0$,*

$$\inf_{\theta \in \Theta} \left[\log\left(\lambda + \bar{\lambda} \int f_\theta^{1+\alpha}(x)dx\right) - \left(\frac{1+\alpha}{\alpha}\right) \log\left(\lambda + \frac{\bar{\lambda}}{n} \sum_{i=1}^{n} f_\theta^\alpha(X_i)\right)\right] = -\infty. \tag{15}$$

Proof Set the objective function on the left hand side of Eq. (15) as $M_n(\theta)$. Then for $\theta = (\mu, \sigma)$,

$$M_n(\theta) = \log\left(\lambda + \frac{\bar{\lambda}}{\sigma^\alpha} \int f^{1+\alpha}(x)dx\right) - \frac{1+\alpha}{\alpha} \log\left(\lambda + \frac{\bar{\lambda}}{n\sigma^\alpha} \sum_{i=1}^{n} f^\alpha\left(\frac{X_i - \mu}{\sigma}\right)\right).$$

We will now show that for each $1 \le j \le n$, $M_n(X_j, \sigma)$ (that is, $\theta = (X_j, \sigma)$) converges to $-\infty$ as $\sigma \downarrow 0$. Fix $1 \le j \le n$. Since $f(x) \ge 0$ for all x, we get, by taking $\mu = X_j$,

$$\left(\frac{1+\alpha}{\alpha}\right) \log\left(\lambda + \frac{\bar{\lambda}}{n\sigma^\alpha} \sum_{i=1}^{n} f^\alpha\left(\frac{X_i - \mu}{\sigma}\right)\right) \ge \left(\frac{1+\alpha}{\alpha}\right) \log\left(\frac{\bar{\lambda}}{n\sigma^\alpha} f^\alpha(0)\right).$$

Substituting this bound in $M_n(\theta) = M_n(X_j, \sigma)$, we obtain

$$M_n(X_j, \sigma) \le \log\left(n^{\frac{1+\alpha}{\alpha}} \left(\lambda\sigma^{\alpha+1} + \bar{\lambda}\sigma \int f^{1+\alpha}\right)\right) - C_f,$$

where

$$C_f = (1+\alpha) \log\left(\bar{\lambda}^{\frac{1}{\alpha}} f(0)\right).$$

Letting σ tend to zero implies $M_n(X_j, \sigma) \to -\infty$ proving Eq. (15). □

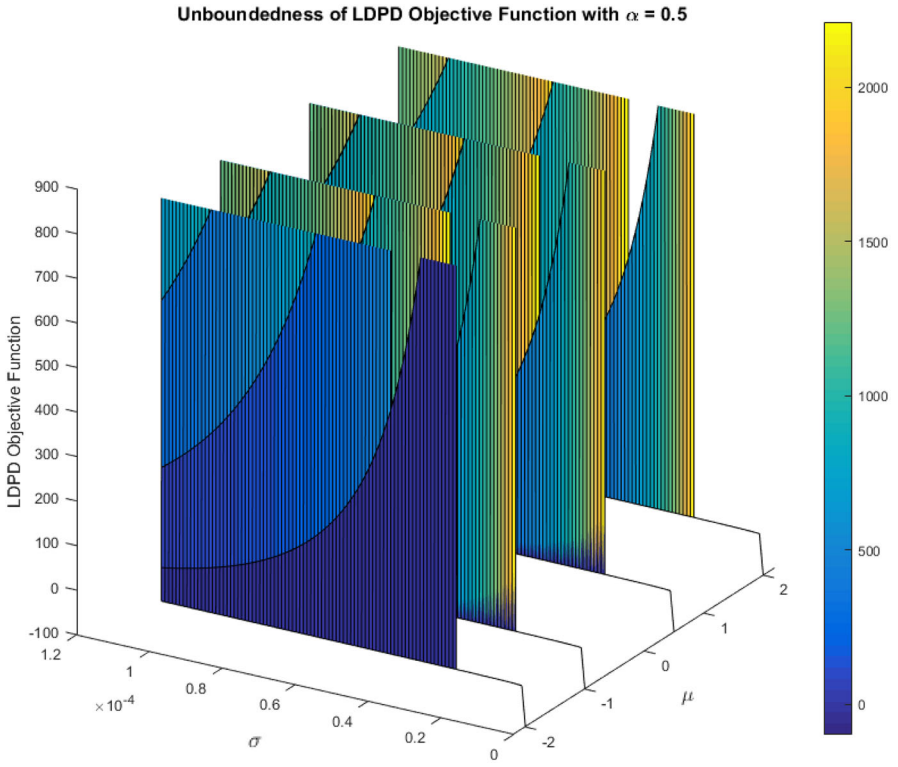

Fig. 5 LDPD objective function for the $N(\mu, \sigma^2)$ model with an artificial dataset of four observations

Remark 9 From the proof, it is seen that the global minimizer of the bridge divergence in case of any sample is at the extreme point ($\sigma = 0$) and the global minimum overall (μ, σ) combinations, is attained for at least n points namely $(X_j, 0)$, $1 \leq j \leq n$. This is similar in spirit to the classical example of likelihood inference for Gaussian mixture modeling as explained in Example 2.4.5 of Bickel and Doksum (2015). Note that the theorem does not cover the case of DPD which is given below.

In order to illustrate the phenomenon discussed above, we plot the LDPD objective with $\alpha = 0.5$ in Fig. 5, based on an artificial sample of size 4, where the sample observations are $-2, -1, 0.5, 2$. The underlying model family is assumed to be $\{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}$. We observe that the LDPD objective drops down sharply to $-\infty$ as σ approaches 0, when μ is either of the four sample points. See Section S.2 of supplementary materials for more details.

Theorem 6 *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed observations from a density g, which is modeled by the densities of the location-scale family $\mathcal{G}_{f,\Theta}$ with a fixed f. Suppose that $\lim_{|x|\to\infty} f(x) = 0$. Then, for any $\alpha > 0$, the following holds for all n if μ does not belong to the set $\{X_1, ..., X_n\}$, and for all $n > \frac{(1+\alpha) f^{\alpha}(0)}{\alpha \int f^{1+\alpha}}$ if μ is equal to some X_i in the above set:*

$$\lim_{\sigma \to 0} \left(\int f_{(\mu,\sigma)}^{1+\alpha}(x)dx - \left(\frac{1+\alpha}{\alpha}\right) \frac{1}{n} \sum_{i=1}^{n} f_{(\mu,\sigma)}^{\alpha}(X_i) \right) = \infty. \tag{16}$$

Proof First, note that the quantity on the left hand side of (16) whose limit needs to be evaluated as $\sigma \to 0$, can be written as

$$\sigma^{-\alpha} \left(\int f^{1+\alpha} - \left(\frac{1+\alpha}{\alpha n}\right) \sum_{i=1}^{n} f^{\alpha}\left(\frac{X_i - \mu}{\sigma}\right) \right). \tag{17}$$

Observe that, if μ does not belong to the set $\{X_1, ..., X_n\}$, then

$$\lim_{\sigma \to 0} f^{\alpha}\left(\frac{X_i - \mu}{\sigma}\right) = 0 \quad \text{for all} \quad 1 \le i \le n.$$

Consequently, (17) goes to ∞ as $\sigma \to 0$. On the other hand, if μ equals X_j for some $1 \le j \le n$, then (17) can be written as:

$$\sigma^{-\alpha} \left(\int f^{1+\alpha} - \left(\frac{1+\alpha}{\alpha n}\right) f^{\alpha}(0) - \left(\frac{1+\alpha}{\alpha n}\right) \sum_{i \neq j} f^{\alpha}\left(\frac{X_i - \mu}{\sigma}\right) \right) \tag{18}$$

It is now easy to see that for $n > \frac{(1+\alpha) f^{\alpha}(0)}{\alpha \int f^{1+\alpha}}$, (18) goes to ∞ as $\sigma \to 0$. This completes the proof. □

Remark 10 Theorem 6 shows that the DPD objective function for the location-scale family (for $\alpha > 0$) diverges to ∞ as $\sigma \to 0$, for all n, if μ is different from all the observations. On the other hand, if μ equals one of the observations, the situation differs in the sense that there exists a_f such that if $n > a_f$, the DPD objective function diverges to ∞ and if $n < a_f$, it diverges to $-\infty$. The case of $n = a_f$ (when a_f is a natural number) depends on the rate of convergence of $f(x)$ to zero as $|x| \to \infty$.

Remark 11 Theorem 5, in particular, covers the case of $N(\mu, \sigma^2)$ parametric family and out of entire BDPD family the DPD alone stands out as a practically valid method of inference. Fujisawa and Eguchi (2008) provided an iterative algorithm to obtain the *minimizer* of the LDPD in case the parametric family is one of the exponential families, applicable to $N(\mu, \sigma^2)$ family. However, while their iterative algorithm is monotone decreasing, it does not in general guarantee convergence to the global minimizer. In fact, the small bias property shown for the scale parameter in their numerical study (Table 2, Fujisawa and Eguchi (2008)) suggests, in view of our discussion in Sect. 4, that their algorithm mostly converges to a local minimizer, at least in the normal family. This is also very similar to the case of likelihood inference for Gaussian mixture modeling wherein one uses the EM algorithm to get hold of the useful local minimizer of the log-likelihood. This fact is also consistent with the asymptotic normality result presented in Section 5 of Fujisawa and Eguchi (2008). The asymptotic normality holds for a sequence of roots of the estimating equation while the global minimizer does not

exist (in the interior of the parameter space). Note that in this case the global minimizer does not appear as a root of the estimating equation owing to the fact that any global minimum is attained on the boundary of the parameter space.

Remark 12 Some fixes are available in the literature to avoid unboundedness of the likelihood in the case of Gaussian mixtures. For example, introduction of the additional constraint that the variances of the components of the Gaussian mixture model are all equal, or placing a positive lower bound on the componentwise variance mitigates the problem of unbounded likelihood. Chen and Tan (2009) proposed a penalized likelihood method for estimating the mixing distribution, and proved consistency properties of their estimator. They also explored a convenient EM algorithm for computing the maximum penalized likelihood estimator. It may be possible to use some such fix to avoid unboundedness of the LDPD (or any other intermediate BDPD).

Another possibility for avoiding the unboundedness of the objective function is to consider the divergence between the smoothed data and the smoothed model as introduced in Basu and Lindsay (1994) and further refined by Seo and Lindsay (2013). Sections 4 and 5 of Seo and Lindsay (2013) prove uniform consistency of their smoothed maximum likelihood approach. We expect a similar smoothing approach to work in this case without restricting the parameters. For example, in case of a Gaussian kernel with bandwidth h, the variance of the smoothed model density becomes $(\sigma^2 + h^2)$, the inverse of which remains bounded even when $\sigma \to 0$.

Remark 13 The result as stated in Theorem 5 does not apply to a scale family with a fixed location (say, 0). But if one of the observations is exactly 0, then the proof as presented above can be employed to get the same conclusion. More precisely, we have, (taking $\sigma = X_{(1)}$),

$$\inf_{\sigma>0}\left[\log\int f_\sigma^{1+\alpha} - \left(1+\frac{1}{\alpha}\right)\log\left(\frac{1}{n}\sum_{i=1}^{n} f_\sigma^{\alpha}(X_i)\right)\right] \le \log(n^{1+1/\alpha}X_{(1)}) + D_f,$$

where $D_f = \log\int f^{1+\alpha} - (1+\alpha)\log f(1)$ under the assumption $f(1) > 0$. Here, $X_{(1)}$ denotes the first order statistic in the sample $X_1, \ldots, X_n$ and $f_\sigma(x) = f(x/\sigma)/\sigma$ for a density f on $\mathbb{R}^+$. This indicates that if we have inner contamination (that is, contamination near the mode 0) so that $n^{1+1/\alpha}X_{(1)}$ is small enough, then there is a possibility of a spurious global minimum near zero. Observe that the true distribution is a mixture described as

$$X|Z = 0 \sim f_{\sigma_0} \quad \text{and} \quad X|Z = 1 \sim \delta_0,$$

where $Z \sim \text{Bernoulli}(p)$ and δ_0 represents a point mass at 0. This represents a simple case inner contamination model with contamination level p. Then in a sample of size n, we have for any fixed $\varepsilon > 0$,

$$\mathbb{P}(X_{(1)} > n^{-1-1/\alpha}\varepsilon) = \left(1 - p - (1-p)F\left(n^{-1-1/\alpha}\varepsilon/\sigma_0\right)\right)^n \le (1-p)^n,$$

implying convergence with probability one of $n^{1+1/\alpha} X_{(1)}$ to zero as $n \to \infty$ for any fixed $p > 0$. Here, F represents the cumulative distribution function of X. In a given sample though, this spurious minimum may disappear (or may not be significantly pronounced) depending on the proportion of contamination. In this case, the BDPD do not behave so badly as in the location-scale family. As already observed, the divergences close to the DPD will lead to reasonable estimators.

6 Toward the desired estimator

As we have seen, the analysis using the LDPD can be beset with different kinds of problems, both theoretical and computational. In this section, we add to the discussion of the possible flaws in the current state of the analysis based on the LDPD, eventually giving a recipe for consolidating the existing knowledge and overcoming the present difficulties so that one can arrive at a best compromise. We do not exactly refute the findings of Fujisawa and Eguchi (2008) and Fujisawa (2013) as we find a substantial part of this research to be valuable and useful. However, there are too many loose ends that remain unaccounted which limit the practical applications of the method without further consolidation.

Fujisawa and Eguchi (2008) and Fujisawa (2013) define the latent bias of an estimator as the difference between the target parameter and the limit of the estimator. To describe the basic flaw with the minimized LDPD estimator or in general the normalized estimating equations, we rework the arguments of Fujisawa (2013) which gives the hint of a very small latent bias under heavy contamination for the minimum LDPD estimator. Let $f(x) = f_{\theta^*}(x)$ be the target density within our parametric family and $\delta(x)$ be the contamination density so that the data generating density g is given by

$$g(x) = (1 - \varepsilon) f(x) + \varepsilon \delta(x).$$

Let $\ell(x; \theta) = \log f_\theta(x)$ and $u_\theta(x) = \nabla \ell(x; \theta)$ be the usual log-likelihood and the score, respectively. A general normalized estimating equation is given by

$$\frac{\mathbb{E}_g[\xi(\ell(X;\theta))u_\theta(X)]}{\mathbb{E}_g[\xi(\ell(X;\theta))]} = \frac{\mathbb{E}_{f_\theta}[\xi(\ell(X;\theta))u_\theta(X)]}{\mathbb{E}_{f_\theta}[\xi(\ell(X;\theta))]}, \tag{19}$$

where $\xi(\cdot)$ is a nonnegative weight function satisfying $\xi(a) \to 0$ as $a \to -\infty$ and $\mathbb{E}_h[\cdot]$ represents the expectation with respect to the density h. Assume that

$$\mathbb{E}_\delta[\xi(\ell(X;\theta))] \approx 0, \quad \text{and} \quad \mathbb{E}_\delta[\xi(\ell(X;\theta))u_\theta(X)] \approx 0, \tag{20}$$

in a neighborhood of $\theta = \theta^*$, which is usually satisfied under the classical outlier model formulation (or, as we call it, outer contamination). Substituting the form of the density g, we get

$$\frac{(1-\varepsilon)\mathbb{E}_{f_{\theta^*}}[\xi(\ell(X;\theta))u_\theta(X)] + \varepsilon\mathbb{E}_\delta[\xi(\ell(X;\theta))u_\theta(X)]}{(1-\varepsilon)\mathbb{E}_{f_{\theta^*}}[\xi(\ell(X;\theta))] + \varepsilon\mathbb{E}_\delta[\xi(\ell(X;\theta))]} = \frac{\mathbb{E}_{f_\theta}[\xi(\ell(X;\theta))u_\theta(X)]}{\mathbb{E}_{f_\theta}[\xi(\ell(X;\theta))]}.$$

Whenever the approximations in (20) hold, we get the following approximate equality

$$\frac{\mathbb{E}_{f_{\theta^*}}[\xi(\ell(X;\theta))u_\theta(X)]}{\mathbb{E}_{f_{\theta^*}}[\xi(\ell(X;\theta))]} \approx \frac{\mathbb{E}_{f_\theta}[\xi(\ell(X;\theta))u_\theta(X)]}{\mathbb{E}_{f_\theta}[\xi(\ell(X;\theta))]}.$$

Thus, θ^* is an approximate solution of the normalized estimating Eq. (19). For concreteness, consider the case where $f(x) = f_0(x)$ (i.e., $\theta^* = 0$), where f_μ denotes the density of the normal distribution with mean μ and variance 1 and $\delta(x) = f_{10}(x)$. Take $\theta = 10$ in the normalized estimating Eq. (19). It is easy to see that

$$\mathbb{E}_{f_0}[\xi(\ell(X;\theta))] \approx 0, \quad \text{and} \quad \mathbb{E}_{f_0}[\xi(\ell(X;\theta))u_\theta(X)] \approx 0,$$

in a neighborhood of $\theta = 10$. Therefore, as above, the equality

$$\frac{\mathbb{E}_{f_{10}}[\xi(\ell(X;\theta))u_\theta(X)]}{\mathbb{E}_{f_{10}}[\xi(\ell(X;\theta))]} = \frac{\mathbb{E}_{f_\theta}[\xi(\ell(X;\theta))u_\theta(X)]}{\mathbb{E}_{f_\theta}[\xi(\ell(X;\theta))]},$$

holds approximately. Thus, not only $\theta^* = 0$ but $\theta = 10$ is also an approximate root of this estimating equation. The latent bias argument for normalizing estimating equations would be one sided if one were to use it only for θ^*, and not for the root corresponding to the contaminating component. One probably can have an estimator with small latent bias in this case, but one would have to appropriately choose between the multiple roots to ensure this small bias.

To validate our argument, we generated 20 observations randomly from the mixture distribution $0.9N(10, 1) + 0.1N(0, 1)$. We present the entire sample for the future reproducibility of our results.

10.73402487,	10.07316778,	12.06112801,	8.52976810,	9.66834408,
10.69923160,	10.65200828,	9.87964160,	8.76851845,	10.91329572,
8.12296270,	8.57642728,	8.52115355,	12.42173904,	10.45406538,
9.86883487,	−0.89282090,	−0.06273843,	1.34072270,	−0.65001187.

The parametric family is taken to be $\mathcal{F} = \{N(\mu, 1) : \mu \in \mathbb{R}\}$. The LDPD objective function in (6) with $\alpha = 0.5$ is plotted against θ in Fig. 6. It is seen to have a (local) minimum near 0 and a (possibly global) minimum around 10, which in this case is the mean of the larger component of the mixture. Note that while there are at least two obvious roots of the corresponding estimating equation, a clear root selection strategy is unavailable. While it so happens that in this case the divergence is well behaved, in many standard cases it is not (as we have seen in Sects. 4 and 5).

We are now in a position to give an analysis of the minimum LDPD estimator in particular (and the minimum BDPD estimator in general) by consolidating the different bits of results that we have presented so far in this paper. In Sect. 3 we have proved the consistency of the minimum BDPD estimator. What it means in layman's terms is that the global minimizer of the empirical LDPD objective function converges to the global minimizer of the theoretical objective function obtained by replacing the empirical distribution function by the true one. However, the usefulness

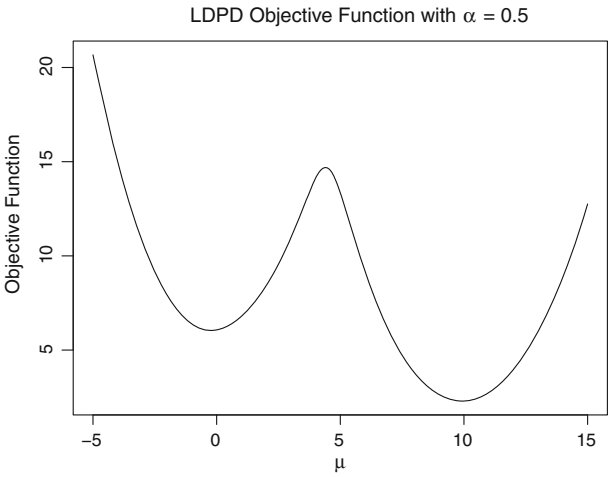

Fig. 6 LDPD objective function for the $N(\mu, 1)$ model under normal mixture data

of these consistency results are immediately challenged by our findings in Sects. 4 and 5 as well as the previous part of this section. First of all, it is quite possible, even in case of pure data, that the global minimizer of the empirical objective function is a value which is entirely useless to the statistician for all practical purposes. Yet, the procedure might actually have a very reasonable local minimizer close to our target. This is what we have observed in Fig. 4, for example. Fujisawa and Eguchi (2008) or Fujisawa (2013)—or, indeed, anybody else for that matter—do not give any root selection criteria in such cases; and in the absence of any such criteria, one would be hard put to justify the choice of a local minimizer, when a perfectly legitimate (although statistically useless) global minimizer exists. Our results in Sect. 5 show that the global minimizers for the LDPD would always be at the boundary of the parameter space and will be of very little practical value to us under location-scale models. However, in almost every such case there exists—as we have seen in our simulations—reasonable roots of the estimating equations, representing appropriate local minima of the objective function. The bias and mean square error reported in Figures 2–5 of Fujisawa (2013) correspond to the root generated by this reasonable local (but not global) minimum.

Secondly, an equally serious problem is the extreme shift in the target parameter itself under inner contamination when LDPD is the divergence of choice. As shown in Fig. 3, a small proportion of inner contamination is enough to shift the target parameter (the global minimizer of the divergence between the theoretical densities) to such an extreme degree that the new target does not characterize the original major component in any way at all. Yet, once again, the divergence has a local minimizer which reasonably characterizes the major component. Once again, however, there is no conceivable reason for considering a local minimizer as the target in the presence of a global minimizer (which, unfortunately, does not characterize the component of interest).

It is clear, therefore, that usual inference based on the LDPD alone can lead the inference astray, if one proceeds with the global minimizer. Other reasonable roots

do exist, but root selection strategies do not. Existing literature claims the existence of "a root" which has the desired properties, without any recipe for arriving at that root. Such inadequacy exists in case of many other members of the BDPD class. The only member within this class which are entirely free—as far as it is observed in our explorations—from the anomalies of spurious roots and ill-defined targets is the DPD. In every model and every case that we have looked at, the DPD has a well-defined minimizer that would reasonably represent, in theory, a properly described target, and would generate a suitable estimator for the same target under randomly generated data. Acknowledging that the LDPD does have certain roots which are superior in dealing with the bias under heavy contamination, we propose to utilize the global minimizer of the DPD to arrive at the desired root of the LDPD. The next subsection describes the chain algorithm which we propose for this purpose.

6.1 The chain algorithm

Under repeated simulations, we have observed, under many standard parametric models, that as far as the global minimizers of the BDPDs are concerned over all $\lambda \in [0, 1]$ given a specific value α, the minimum DPD estimator is almost always the best in terms of latent bias (and not the minimum LDPD estimator). At the same time we acknowledge that the minimum LDPD estimating equations (or, more generally, the minimum BDPD estimating equations) may have roots which represent some local minimum of the divergences which might improve upon the performance of the minimum DPD estimator in terms of latent bias. As indicated by the heuristics, and as we have seen in our simulations, the LDPD (more generally the BDPD) has a local minimizer that is close to the true θ^*. This we believe is at least partially due to the Pythagorean relation that we presented in Sect. 3. Since the members of the BDPD family form a literal bridge between the DPD and LDPD, we have an opportunity for getting at a root of the LDPD estimating equations that is consistent and asymptotically normal. A completely rigorous proof is unavailable at this time, but we provide a strong heuristic argument and extensive simulations to support this claim.

From the form of the bridge divergence, for a fixed α and θ, we obtain the limit

$$\rho^{(\alpha,\lambda_k)}(g, f_\theta) \to \rho^{(\alpha,\lambda_0)}(g, f_\theta) \quad \text{as} \quad \lambda_k \to \lambda_0. \tag{21}$$

Heuristically this indicates that $\rho^{(\alpha,\lambda)}(g, f_\theta)$ has at least a local minimizer close to $\hat{\theta}^\alpha_{n1}$ [defined in Eq. (5)] if λ is close to one. For example, one can take $\lambda = 1 - n^{-1/2}$ which is sufficiently close to 1 for this root to be reasonably good. We shall now present a chain algorithm for getting hold of a good root of the LDPD estimating equation (indeed, all bridge estimating equations). The chain algorithm proceeds in the following steps: Fix $\alpha \in [0, 1]$.

1. First choose a sequence $\{\lambda_i\}$ satisfying

$$1 = \lambda_K > \lambda_{K-1} > \cdots > \lambda_2 > \lambda_1 > \lambda_0 = 0 \quad \text{and} \quad \max_{1 \le i \le K} |\lambda_i - \lambda_{i-1}| \le r_n,$$

with $r_n \to 0$ at some rate. Define a function $\lambda \mapsto \hat{\theta}^\alpha(\lambda)$.

2. Solve the DPD problem completely, that is, find the global minimizer $\hat{\theta}^{\alpha}_{n1}$ of the sample estimate of the DPD. Set $\hat{\theta}(\lambda_K) = \hat{\theta}^{\alpha}_{n1}$. Note that $\lambda = 1$ in the bridge divergence expression in Eq. (13) corresponds to the DPD (in the limit).
3. For $i = K-1, K-2, \ldots, 0$, find the local minimizer of the sample estimate of the bridge divergence with parameters (α, λ_i) that is closest to $\hat{\theta}(\lambda_{i+1})$. Set this closest local minimizer as $\hat{\theta}(\lambda_i)$.
4. Return $\{\hat{\theta}(\lambda_i) : 0 \leq i \leq K\}$.

Step 3 above can be solved by using any off-the-shelf algorithm for minimization with $\hat{\theta}(\lambda_{i+1})$ as a starting point. To draw analogue with existing algorithms of this type, we note that the LASSO that has taken over the literature in the past few years is solved by using a chain/path algorithm as above; there the issue is about getting fast algorithm not distinguishing local and global minimizers (it is a convex problem).

Conjecture 1 In any model under the assumptions that guarantee the asymptotic normality of $\hat{\theta}^{\alpha}_{n1}$, $\hat{\theta}^{\alpha}(\lambda_0)$ is the consistent root of the LDPD estimating equation that is closest to θ^*. And this is the solution claimed by Fujisawa and Eguchi (2008) to have a small latent bias.

The simulations using this chain algorithm are very encouraging and we think that a proof of the above conjecture can be obtained by an application of the Taylor series and using (21) with an appropriate choice of rate of convergence of $\lambda_k - \lambda_0$ to zero. In many *nicely behaved* parametric densities, simply choosing the root of LDPD estimating equation closest to the minimum DPD might be good enough. However, in general cases, the difference between LDPD and DPD estimating equations can be drastic for this simple trick to work (at least for a theoretical guarantee).

Remark 14 The chain algorithm described here is very similar to the homotopy method available for global optimization. See Dunlavy (2005) and Dunlavy and O'Leary (2005) for more details about the homotopy method. In our case, the bridge parameter λ, the DPD and the LDPD are, respectively, the analogues of the homotopy parameter λ, the initial function f^0 and the objective function f^1 in Section 2.1 of Dunlavy and O'Leary (2005). Often the goal of the homotopy method is to obtain the global optima of the objective function. Here, the goal of our chain algorithm is to define our target of interest.

In the following section, we will frequently use the term "minimum bridge divergence estimator". It is to be understood that this will refer to the local minimizer obtained by using the chain algorithm which uses the global minimizer of the DPD as the starting value. In addition, we will also frequently use the term"root of the estimating equation". In this case, it will be understood that the correspondence is with a local minimizer, and not an intervening local maximizer (or saddle point), which may also produce a legitimate root of the estimating equation.

7 Simulation study

In this section, we apply the chain algorithm discussed in Sect. 6 on simulated data. We consider the exponential and the normal scale families, the former to illustrate the case

of outer contamination, and the latter to illustrate the case of inner contamination. The estimators are computed from 1000 replications. The bias and mean squared error over these 1000 replications are calculated against the target component of the underlying majority distribution.

The chain algorithm is applied for each pair (α, ϵ), where the robustness parameter α runs from 0 to 1 in steps of 0.2 and the contamination level ϵ are 0, 0.05 or 0.2. For each (α, ϵ) combination, the bridge parameter λ runs from 1 to 0 in steps of 0.1 as the chain algorithm proceeds. Since the parameter α is somewhat better understood in the literature and our introduced parameter λ needs to be analyzed more deeply, the latter is varied through finer steps than the former.

7.1 Exponential scale model

Here, the model is the class of exponential distributions with mean σ over $\sigma \in (0, \infty)$. Data are simulated from the mixture $(1 - \epsilon)\text{Exp}(1) + \epsilon\text{Unif}(6 - 10^{-4}, 6 + 10^{-4})$, where the first component is exponential with rate 1, and the second is uniform over the indicated range. The contamination level ϵ is taken to be 0, 0.05 and 0.2. For step 2 of the chain algorithm which involves solving the DPD problem completely, 25 starting points are drawn uniformly from the interval [0, 0.1] and the remaining 75 uniformly from the interval (0.1, 10].

7.2 Normal scale model

Here, the model is the class of normal distributions with mean 5 and variance σ^2 over $\sigma \in (0, \infty)$. Data are simulated from the mixture $(1 - \epsilon)N(5, 1) + \epsilon\text{Unif}(5 - 10^{-5}, 5 + 10^{-5})$, where the first and the second components are normal and uniform, respectively, with parameters as indicated. The contamination level ϵ is taken to be 0, 0.05 and 0.2. For step 2 of the chain algorithm which involves solving the DPD problem completely, 25 starting points are drawn uniformly from the interval [0, 0.1] and the remaining 75 uniformly from the interval (0.1, 10].

7.3 Description of the results

For the sake of brevity, we will only include the graphs of the mean squared errors (scaled by sample size $n = 100$) of the minimum bridge divergence estimators in the exponential scale model case, for contamination levels 0.05 and 0.2, respectively (in Fig. 7), in the main article. Tables presenting exact values of the bias and MSE of the minimum bridge divergence estimators for both the exponential and normal scale families, for all the three contamination levels 0, 0.05 and 0.2, are given in the online supplement.

It is observed that for the 5% contamination level, the MSE of the minimum bridge divergence estimators decreases as the chain algorithm proceeds (λ goes from 1 to 0) for α up to 0.6, and, roughly speaking, increases as the chain algorithm proceeds for $\alpha = 0.8$ and 1. On the other hand, for the 20% contamination level, the MSE of

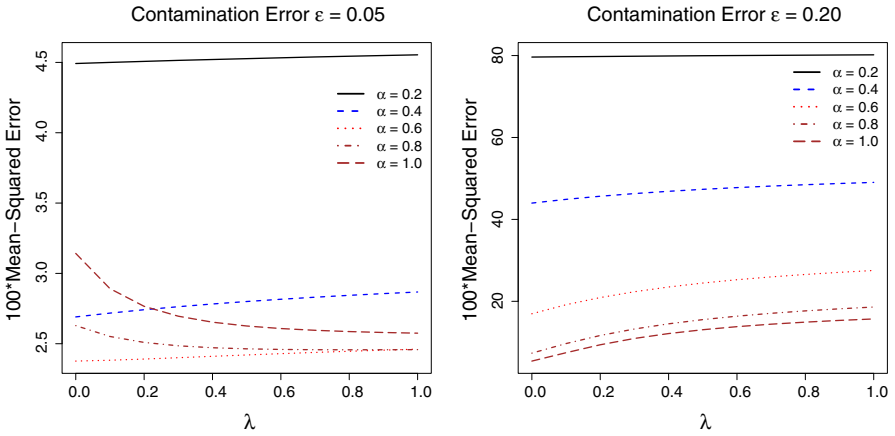

Fig. 7 Scaled MSE of the minimum bridge divergence estimators for the Exp(σ) model at 5 and 20% Contaminations

the minimum bridge divergence estimators decreases as the chain algorithm proceeds from $\lambda = 1$ to $\lambda = 0$ for all values of α. As a function of α (with λ held fixed), for 5% contamination, the MSE decreases as α increases roughly up to $\alpha = 0.6$ and then increases with α, whereas for 20% contamination, a completely decreasing trend of the MSE is observed as α increases.

Note that the values reported here are not, except for the DPD ($\lambda = 1$ case) case, based on the global minimizers of the divergences; rather they are the chain algorithm solutions. What this shows is that for heavy contamination (at the level of 20%) the LDPD solution (not necessarily the global minimizer) as obtained from the chain algorithm does seem to dominate the other BDPD solutions, and the mean square errors decrease uniformly in the direction $1 \to 0$ over λ for each α. This provides partial confirmation of the results of Fujisawa and Eguchi (2008) and Fujisawa (2013). However, the results are now stronger and more useful in that we identify the root for which it works; it does not work for just any root, and certainly not for the global minimizer.

The situation observed in the normal distribution case is somewhat different. Here, the contamination is an inner contamination, which leads to spurious global minimizers more often for the BDPDs. The chain algorithm does guide the process to a sensible root (that is also a local minimizer) in each case. But in this scenario, the minimum DPD estimator remains the best in terms of both the bias and mean squared error for practically all values of α under both mild ($\epsilon = 0.05$) and heavy ($\epsilon = 0.2$) contaminations.

For the data of size 20 presented in Sect. 4 from $N(0, 1)$, the minimum bridge divergence estimators of σ under the $N(0, \sigma^2)$ model with $\alpha = 0.8$ and varying λ from 0 to 1 are presented in Table 1. For each value of λ, the true global minimizers of the bridge divergence objective functions are presented in the parentheses. It is remarkable that even with λ as high as 0.9, the spurious root phenomenon is observed.

Table 1 Bridge divergence estimators of σ with $\alpha = 0.8$

λ	$\hat{\theta}_{\alpha,\lambda}$	λ	$\hat{\theta}_{\alpha,\lambda}$	λ	$\hat{\theta}_{\alpha,\lambda}$
0.9	1.245808 (1.4411e−5)	0.6	1.248243 (1.4124e−5)	0.3	1.252942 (1.4085e−5)
0.8	1.246477 (1.4218e−5)	0.5	1.249401 (1.4106e−5)	0.2	1.255756 (1.4078e−5)
0.7	1.247269 (1.4155e−5)	0.4	1.250926 (1.4094e−5)	0.1	1.259926 (1.4073e−5)

The global minimizers are presented in the parentheses

7.4 Summary of simulation results

So far in this section, we have described the findings of the selected simulation results for the exponential scale and the normal scale problem, some of which have been expanded upon in the online supplement. In this subsection, we give a brief summary of the comparative performance of the different estimators including those cases for which detailed simulation results have not been presented.

1. The spurious behavior of the global minimizer of the LDPD is present in all kinds of situations including outer contamination, inner contamination and no contamination in both the normal and the exponential scale models. This behavior is also present in many of the other bridge divergences.
2. If the global minimizer of the divergence is the estimator of choice, then the DPD is the only divergence without spurious behavior within the bridge divergence family, at least for the models considered here.
3. For heavy outer contamination ($\epsilon \geq 0.2$), the reasonable roots of the bridge divergences (obtained using the chain algorithm) appear to provide mean squared errors which are increasing in λ. Thus the chain algorithm root for the LDPD is the most desirable estimator in this system of estimators. This has been our consistent observation in both the normal and the exponential scale models.
4. For heavy inner contamination, however, the situation is different. Here, there is little to choose between the performance of the bridge divergence estimators (obtained using the chain algorithm) for α roughly up to 0.5. But for larger values of α, the mean squared error has a clearly declining pattern with increasing λ and the minimum DPD estimator appears to dominate the others. Once again this holds for both the normal and the exponential scale models.
5. The performance of the LDPD estimator would have been substantially inferior had we chosen the global minimizer of the divergence instead of the chain algorithm root. For example, Fujisawa (2013) considered a case of outer contamination in the $N(\mu, \sigma^2)$-model using a sample of size 40 replicated 500 times, and reported (Fig. 5d) a mean squared error of about 0.24 for the estimate of σ. However, the global minimizer of the LDPD has an MSE of 1 in this case.

Based on our experience in the theoretical calculations, numerical work and simulations, we recommend that the estimators based on the LDPD and the other bridge divergences whenever they are used should be obtained using the chain algorithm starting with the minimum DPD estimator rather than direct minimization of the objective function.

8 Choice of tuning parameters

Most of the robust estimators derive their robustness from down-weighting the observations suspected as outliers. As can be seen from the BDPD objective functions, the observations are proportionally weighted by powers of the model density which demonstrates that the observations are down-weighted for being outliers with respect to the density f_θ. It is this outlier stability property which makes the corresponding estimators desirable from the point of view of robustness. If there are no outliers in the data, then the tuning parameter α must ideally be set to zero in order to get optimal inference. But if the data do contain outliers, then the tuning parameters α and λ should be chosen so as to partially or fully eliminate the effect of the outlying observations. However, the choice of the tuning parameter must be an automatic procedure where a data-driven choice of the parameters α and λ is generated by the method. One of the first methods of choosing tuning parameters in case of the DPD was proposed in Hong and Kim (2001), where the tuning parameter is chosen by minimizing the trace of an estimate of the asymptotic covariance matrix. Warwick and Jones (2005) refined this method to find an "optimal" tuning parameter by minimizing the trace of an estimate of the MSE of the estimator. Here, the MSE is computed by separately estimating the bias and the variance components. From the asymptotic variance formula, one can easily estimate the variance. But bias estimation involves the use of a pilot estimate, and the estimated mean square error has a strong dependence on it. Warwick and Jones (2005) used the minimum L_2 distance estimator as the pilot; however, under the current state of the art, there is no universally acceptable choice of the pilot, on which the process critically depends.

We acknowledge that the method of Warwick and Jones (2005) would be perfect if we could eliminate its dependence on the pilot estimator, or find an independent estimate of the bias as a function of the tuning parameter. However, here we present a modified version of the approach of Hong and Kim (2001) and give an (almost theoretical) justification of the same. Firstly, one should not use the trace of an estimate of the asymptotic covariance matrix (although some of the present authors have done so in the past in the absence of a better strategy) for the construction of the objective function to be minimized. This leads to adding quantities with different units. Consider the $N(\mu, \sigma^2)$ example in which case the asymptotic variance of $(\hat{\mu}, \hat{\sigma}^2)$ or any of the variance estimates have the units of X_i and X_i^2 that should not be added. One could, of course, think of estimating (μ, σ) where adding the variance estimates would be more sensible. But this remedy does not work in general parameter spaces.

Why is the asymptotic variance adequate? An intuitive explanation is as follows. It is well-known that the delete-d jack-knife estimator is a consistent estimator of the asymptotic variance. The delete-d jack-knife estimator is obtained by calculating the difference between the estimator calculated based on n observations and the estimator calculated based on $n - d$ observations. By the definition of a good robust estimator, this difference should be small for a robust estimator but for a nonrobust estimator like the maximum likelihood estimator this difference would be large if the d observations deleted contain some outliers. This implies that robust estimators should have "smaller" asymptotic variance than nonrobust estimators. Of course, if there is no contamination then it is known that the maximum likelihood estimator would have

"smaller" variance for large enough sample size. Using a bootstrap asymptotic variance estimator also gives a similar conclusion. We do ignore the bias, but the question clearly goes in favor of a robust estimator. By definition, they are closer to the true parameter based on the majority of the data.

A more concrete explanation is offered by using the closeness measure introduced in Zhang and He (2016). Let $\mathcal{V}_\Lambda$ define a set of variance matrices V_α indexed by a (possibly vector) parameter $\alpha \in \Lambda$.

Definition 1 A nonnegative real-valued matrix function $\mathfrak{m}(\cdot)$ defined on the set $\mathcal{V}_\Lambda$ is called a closeness measure if and only if the following conditions hold.

1. Consistency. For any $V_1, V_2 \in \mathcal{V}_\Lambda$, if $V_1 \succeq V_2$, then $\mathfrak{m}(V_1) \geq \mathfrak{m}(V_2)$, where the equality holds only if $V_1 = V_2$.
2. Continuity. For any matrix sequence $\{V_n\} \subseteq \mathcal{V}_\Lambda$ with $V_n \succeq B$. As $n \to \infty$, $\|V_n - B\|_\infty \to 0$ if and only if $\mathfrak{m}(V_n) \to \mathfrak{m}(B)$.

Here, by $A \succeq B$, we mean $A - B$ is nonnegative definite. Definition 1 implies that if there is an efficient variance matrix in the set $\mathcal{V}_\Lambda$, then minimizing $\mathfrak{m}(V_\alpha)$ over $\alpha \in \Lambda$ leads to such a matrix. Zhang and He (2016) additionally prove that the trace and the Frobenious norm are valid closeness measures. We use the determinant of the matrix as the measure of closeness (see Section S.5 of the supplementary material). Note that determinant is a valid quantity to consider from the point of view of units and it also has a practical interpretation as the volume of a confidence set. These arguments give a justification for minimizing a closeness measure of an estimate of the asymptotic variance.

Summing up all these arguments, we claim that minimizing the determinant of an estimate of the asymptotic variance provides an asymptotically optimal estimator whenever such an estimator exists in the family of estimators under consideration. A detailed analysis of this procedure would be taken up in a future paper. Just as a remark, for the normal data of size 20 presented in Sect. 4, the optimal λ parameter with $\alpha = 0.8$ using this procedure turned out to be $\lambda = 1$. A plot of the asymptotic variance over all λ is given in the supplementary material.

9 Concluding remarks

In this paper, the competing families of divergences, DPD and LDPD, are critically examined for their strengths and deficiencies. The bridge divergence family introduced in this paper is an attempt at combining the good of both and nullify the disadvantages of either. Unlike the DPD, the LDPD estimating equation admits roots with small latent bias. However, these roots may not be the global minimizers of the LDPD objective. The phenomenon of spurious global minimizers of the LDPD is rigorously proved in specific parametric families where the DPD provably works. The bridge divergences partly suppress this problem for certain tuning parameter (λ) values along the bridge.

However, the bridge divergence is not a perfect solution in that it also faces the same problem as LDPD in some cases. The point made here is that one cannot expect the global minimizer of the LDPD to generate a good estimator and the DPD estimator

is a safe bet in all the cases, but whenever possible some members of the minimum bridge divergence estimators can help reduce the latent bias of the DPD estimator.

Acknowledgements The authors gratefully acknowledge the comments of two anonymous referees as well as the members of the editorial board which led to a significantly improved version of the paper. The authors are indebted to Srijata Samanta of University of Florida for her contribution toward Remark 13.

References

Basu, A., Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, *46*(4), 683–705.

Basu, A., Harris, I. R., Hjort, N. L., Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, *85*(3), 549–559.

Bickel, P. J., Doksum, K. A. (2015). *Mathematical statistics–basic ideas and selected topics* (Vol. 1). Texts in Statistical Science Series, second edition. Boca Raton, FL: CRC Press.

Broniatowski, M., Toma, A., Vajda, I. (2012). Decomposable pseudodistances and applications in statistical estimation. *Journal of Statistical Planning and Inference*, *142*(9), 2574–2585.

Chen, J., Tan, X. (2009). Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, *100*, 1367–1383.

Dunlavy, D. M. (2005). *Homotopy Optimization Methods and Protein Structure Prediction*. PhD dissertation, The Graduate School of the University of Maryland, College Park.

Dunlavy, D. M., O'Leary, D. P. (2005). *Homotopy optimization methods for global optimization*. Technical report, Sandia National Laboratories, SAND2005-7495, Albuquerque, New Mexico 87185 and Livermore, California 94550.

Ferguson, T. S. (1996). *A course in large sample theory. Texts in Statistical Science Series*. London: Chapman & Hall.

Fujisawa, H. (2013). Normalized estimating equation for robust parameter estimation. *Electronic Journal of Statistics*, *7*, 1587–1606.

Fujisawa, H., Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, *99*(9), 2053–2081.

Hong, C., Kim, Y. (2001). Automatic selection of the tuning parameter in the minimum density power divergence estimation. *Journal of the Korean Statistical Society*, *30*(3), 453–465.

Jones, M. C., Hjort, N. L., Harris, I. R., Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika*, *88*(3), 865–873.

Kanamori, T., Fujisawa, H. (2014). Affine invariant divergences associated with proper composite scoring rules and their applications. *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, *20*(4), 2278–2304.

Lehmann, E. L., Casella, G. (1998). *Theory of point estimation*. Springer Texts in Statistics, second edition. New York: Springer-Verlag.

Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *The Annals of Statistics*, *22*(2), 1081–1114.

Seo, B., Lindsay, B. G. (2013). A universally consistent modification of maximum likelihood. *Statistica Sinica*, *23*(2), 467–487.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge: Cambridge University Press.

Warwick, J., Jones, M. C. (2005). Choosing a robustness tuning parameter. *Journal of Statistical Computation and Simulation*, *75*(7), 581–588.

Windham, M. P. (1995). Robustifying model fitting. *Journal of the Royal Statistical Society. Series B. Methodological*, *57*(3), 599–609.

Zhang, S., He, X. (2016). Inference based on adaptive grid selection of probability transforms. *Statistics. A Journal of Theoretical and Applied Statistics*, *50*(3), 667–688.