Journal of the Korean Statistical Society (2001), 30: 3, pp 453–465

Automatic Selection of the Tuning Parameter in the Minimum Density Power Divergence Estimation†

Changkon Hong[1] and Youngseok Kim[1]

ABSTRACT

It is often the case that one wants to estimate parameters of the distribution which follows certain parametric model, while the data are contaminated. It is well known that the maximum likelihood estimators are not robust to contamination. Basu et al.(1998) proposed a robust method called the minimum density power divergence estimation. In this paper, we investigate data-driven selection of the tuning parameter α in the minimum density power divergence estimation. A criterion is proposed and its performance is studied through the simulation. The simulation includes three cases of estimation problem.

Keywords: Density power divergence, Density-based minimum divergence method, Robustness, Tuning parameter

1. Introduction

In parametric density estimation, density-based minimum divergence methods have long history. These methods include the maximum likelihood method, the minimum Hellinger distance method ($\int [\, \widehat{f}^{\,\frac{1}{2}}(x) - f^{\frac{1}{2}}(x\,;\,\theta)\,]^2\,dx$), and the minimum L_1-distance method ($\int |\, \widehat{f}\,(x) - f\,(x\,;\,\theta)\,|\,dx$). One of the practical drawbacks of maximum likelihood estimation is the nonrobustness to contamination and model misspecification. On the other hand, it was shown that minimum Hellinger distance method and minimum L_1-distance method have excellent robustness properties (see Beran (1977), Donoho and Liu (1988)). However, these methods also have drawback that they need the calculation of nonparametric estimator $\widehat{f}\,(x)$ with associated complication of smoothing parameter selection. To avoid the complication of nonparametric smoothing, Hjort (1994) and

†This work was supported by the research fund of Pusan National University

[1]Department of statistics, Pusan National University, Pusan 609-735, Korea

Scott (1999) hired the minimum L_2-distance method. Since the squared L_2-distance between the underlying density $g(x)$ and the assumed density $f(x;\theta)$ is

$$\int \{ g(x) - f(x;\theta) \}^2 dx$$
$$= \int g^2(x)\, dx - 2 \int f(x;\theta)\, g(x)\, dx + \int f^2(x;\theta)\, dx,$$

minimizing L_2-distance is equivalent to minimizing

$$\int f^2(x;\theta)\, dx - 2 \int f(x;\theta)\, g(x)\, dx.$$

The minimum L_2-distance estimator (called L_2E by Scott (1999)) is defined as the minimizer of the empirical version

$$\int f^2(x;\theta)\, dx - 2\frac{1}{n} \sum_{i=1}^{n} f(X_i;\theta)$$

which does not need any smooth nonparametric density estimation. The L_2E is robust, but under the assumed model it is inefficient. Basu et al. (1998) suggest a new family of density-based divergence measures called 'density power divergences'. The family is indexed by a single parameter α which controls the trade-off between robustness and asymptotic efficiency of the estimators. This family includes the Kullback-Leibler divergence (Kullback and Leibler (1951)) (when $\alpha = 0$) and L_2-distance (when $\alpha = 1$). Therefore, the minimum density power divergence estimator can be thought as a bridge between efficient-but-nonrobust estimator MLE and robust-but-inefficient estimator L_2E. Furthermore, it does not require nonparametric density estimation.

In this research, we will study the selection of the tuning parameter α. We will suggest a data-driven criterion and study its performance. In Section 2, we will reintroduce the density power divergence, summarize the results of Basu et al. (1998), and suggest a data-based criterion for selecting α. In Section 3, the performance is to be studied through simulation. The concluding remarks and further research will be given in Section 4.

2. The Minimum Density Power Divergence Estimator and The Suggested Criterion for α-Selection

2.1. The minimum density power divergence estimator

In this subsection, we briefly reintroduce the density power divergence and summarize the results of Basu et al. (1998). Consider a parametric family of distributions $\{F_t\}$, indexed by the unknown parameter $t \in \Omega \subset R^s$, having densities $\{f_t\}$ with respect to Lebesgue measure. Let G be a distribution function having density g with respect to Lebesgue measure.

Basu et al. (1998) define the divergence $d_\alpha(g, f)$ between density functions g and f by

$$d_\alpha(g, f) = \int \left\{ f^{1+\alpha}(z) - \left(1 + \frac{1}{\alpha}\right) g(z) f^{\alpha}(z) + \frac{1}{\alpha} g^{1+\alpha}(z) \right\} dz , \qquad (2.1)$$

for $\alpha > 0$. When $\alpha = 0$, the integrand in expression (2.1) is undefined, and the divergence $d_0(g, f)$ can be defined as

$$d_0(g, f) = \lim_{\alpha \to 0} d_\alpha(g, f) = \int g(z) \log \{ g(z) / f(z) \} \, dz.$$

They call the family of divergence d_α, as a function of α, 'the class of density power divergence'. Note that $d_0(g, f)$ is the Kullback-Leibler divergence and $d_1(g, f)$ is the squared L_2-distance between g and f.

Theorem 1. *The quantity $d_\alpha(g, f)$ is a divergence in that it is nonnegative for all $g, f \in \mathcal{G}$ and is equal to zero if and only if $f \equiv g$ almost everywhere.*

The proof is given in Basu et al. (1998).

Now, suppose that random sample $X_1, \ldots, X_n$ are drawn from G. The minimum density power divergence estimator(MDPDE) $\widehat{\theta}_\alpha$ is defined to be the minimizer of

$$\int f_t^{1+\alpha}(z) \, dz - \left(1 + \frac{1}{\alpha}\right) n^{-1} \sum_{i=1}^{n} f_t^{\alpha}(X_i) \qquad (2.2)$$

with respect to t. In particular when $\alpha = 0$, $\widehat{\theta}_0$ is the maximum likelihood estimator if it exists; when $\alpha = 1$, the estimator $\widehat{\theta}_1$ is the minimum L_2-distance estimator and is called the L_2E estimator by Scott (1999). Note that the minimization of density power divergence does not require any smooth nonparametric

estimate of g, in contrast to work of Cao et al. (1995). Basu et al. (1998) show that the estimator $\widehat{\theta}_\alpha$ through the density power divergence becomes less and less efficient as α increases. Scott (1999) shows that L_2E estimator (when $\alpha = 1$) is robust but inefficient under assumed model. Therefore, in this paper we restrict our interest to the values of α between 0 and 1. For $0 < \alpha < 1$, the class of density power divergences provides a smooth bridge between the Kullback-Leibler divergence and the L_2-distance.

For general families of f_t, the minimizer $\widehat{\theta}_\alpha$ of (2.2) is obtained from the estimating equation

$$n^{-1} \sum_{i=1}^{n} u_t(X_i) f_t{}^{\alpha}(X_i) - \int u_t(z) f_t{}^{1+\alpha}(z)\, dz = 0, \tag{2.3}$$

where $u_t(z) = \partial \log f_t(z) / \partial t$ is the maximum likelihood score function. This shows that the minimum power divergence estimator $\widehat{\theta}_\alpha$ can be thought to be an M-estimator, that is, it solves an equation of the form $\sum_i \psi(X_i, t) = 0$ (see Hampel et al. (1986)).

Basu et al. (1998) show the following theorem holds. In the following, θ represents the best fitting value of the parameter, in the sense of minimizing the divergence $d_\alpha(g, f_t)$, whereas t denotes a generic element of Ω.

Theorem 2. *Under certain regularity conditions, there exists a solution $\widehat{\theta}_\alpha$ of the estimating equation (2.3) such that, as $n \to \infty$,*

(i) $\widehat{\theta}_\alpha$ is consistent for θ, and

(ii) $n^{\frac{1}{2}}(\widehat{\theta}_\alpha - \theta)$ is a asymptotically multivariate normal with vector mean zero and covariance matrix $J^{-1}KJ^{-1}$, where $J = J(\theta)$ and $K = K(\theta)$ are given by

$$\begin{aligned} J &= \int u_\theta(z)\, u_\theta{}^T(z)\, f_\theta{}^{1+\alpha}(z)\, dz \\ &\quad + \int \{ v_\theta(z) - \alpha\, u_\theta(z)\, u_\theta{}^T \} \{ g(z) - f_\theta(z) \} f_\theta{}^{\alpha}(z)\, dz \end{aligned} \tag{2.4}$$

and

$$K = \int u_\theta(z)\, u_\theta{}^T f_\theta{}^{2\alpha}(z)\, g(z)\, dz - \xi\, \xi^T. \tag{2.5}$$

Here $\xi = \int u_\theta(z) f_\theta{}^{\alpha}(z) g(z)\, dz$ and $v_\theta(z) = -\partial \{u_\theta(z)\} / \partial \theta$ is the information function.

2.2. Selection of the tuning parameter α

It is often the case that our interest lies in estimating the parameters of the true parametric distribution while the data are drawn from a contaminated distribution. To be more specific, suppose that the true distribution is F_{θ^*} but the data are drawn from a contiminated distribution $G = (1-\varepsilon)F_{\theta^*} + \varepsilon H$. Here H is a contaminating distribution and let us call G as 'underlying distribution'. Assume that it is our interest to estimate θ^* using the (possibly contaminated) data. As mentioned in the introduction, MLE is not robust to contamination and as a remedy for this the MDPDE is suggested by Basu et al. (1998). In MDPD estimation, Basu et al. (1998) uses prefixed α which is not far from 0. They achieved the robustness at the expense of efficiency. As was pointed out in Basu et al. (1998), there is no universal way of selecting an appropriate (prefixed) α. But when the data are not contaminated, it is advisable to use $\alpha = 0$(MLE), since MLE is asymptotically efficient in this case. And it is obvious that if the underlying distribution is close to the true one, then very small α is preferred. It is also suspected that as the underlying distribution is more contaminated, rather larger value of α may give better results. In this direction of argument, Basu et al. (1998) commented that "in some practical applications, 'prior motions of the extent of contamination' could be hired in determining α". In this paper, we want to suggest a data-driven criterion for selecting α without the 'prior motions of the extent of contaminations'.

To make the problem simple, we only study one parameter case. But the several parameter case can be treated in a similar way. Again, it should be emphasized that θ is the best fitting value of the parameter in parametric model and depends on both underlying distribution G and tuning parameter α. In general, θ is different from the true value θ^* of the parameter. How far is θ from the true parameter θ^*? Let us see this by examining an example of normal mean estimation problem. Let $\phi(z; a, b)$ be the pdf of $N(a, b)$ and suppose that the true distribution is $N(\theta^*, 1)$ while underlying distribuiton is $(1-\varepsilon)N(\theta^*, 1)+\varepsilon N(\theta_1, 1)$. Then $f_t(z) = \phi(z; t, 1)$ and $g(z) = (1-\varepsilon)\phi(z; \theta^*, 1)+\varepsilon\phi(z; \theta_1, 1)$. It can be shown that $\theta = (1-\varepsilon)\theta^* + \varepsilon\theta_1$ for $\alpha = 0$. But for $\alpha > 0$ with $\varepsilon > 0$, we cannot obtain a closed form expression of θ. It can be shown, however, that if $\varepsilon > 0$, θ goes to θ^* as α increases. Does this necessarily mean larger values of α would give better results for estimating the ture parameter θ^*? The answer is negative because too large α can result in too large variance of the estimator. In the simulation study of subsection 3.1 we will set $\theta^* = 0$, $\theta_1 = 10$, and here and now we calculate

the (approximate) numerical values of θ with these values of θ^* and θ_1. Table 1 shows the values of θ for various values of α with $\varepsilon = 0.1$. We can see that θ is quite close to the ture mean $\theta^* = 0$ for the values of α between 0.05 and 0.1. Since by theorem 2 MDPD estimator consistently estimates θ, it is quite robust to contamination in estimating θ^* for these values of α.

Table 1. Values of θ in normal mean, with $\theta^* = 0, \theta_1 = 10$, and $\varepsilon = 0.1$,

α	0.0	0.005	0.01	0.02	0.03	0.04
θ	1.0	0.83	0.68	0.436	0.276	0.172
α	0.05	0.06	0.07	0.08	0.09	0.1
θ	0.11	0.07	0.046	0.03	0.02	0.014

When there is no contamination ($\varepsilon = 0$), $\theta = \theta^*$ by theorem 1 and the asymptotic variance of the MDPD estimator $\hat{\theta}_\alpha$ is minimized at $\alpha = 0$. This is closely related to the fact that MLE is asymptotically efficient when there is no contamination. This motivates our seggestion of estimated asymptotic variance as a criterion for selecting α.

The asymptotic variance $J^{-2}K$ is a function of α as well as a function of both the underlying distribution G and θ. Let us denote the variance $V \equiv V(\alpha; \theta, G)$. Fortunately, it is (heterogeneously) linear in g and we don't need smooth nonparametric estimation of g. We can get a natural estimator $\widehat{V}$ of V by replacing θ and G by $\widehat{\theta}_\alpha$ and G_n, respectively, where G_n is the empirical distribution. Since $\widehat{\theta}_\alpha$ can be obtained from the estimating equation (2.3) for given α the estimator $\widehat{V}$ can be expressed as a function of α only. Let us denote $\widehat{V}$ by $\widehat{V}(\alpha)$. We suggest $\widehat{V}(\alpha)$ as a criterion for selecting α.

Suppose that the data $X_1, \ldots, X_n$ are generated from the underlying distribution G. Then the estimated J and K can be obtained as follows

$$\begin{aligned}\widehat{J}(\alpha) &= (1+\alpha)\int \{ u_{\widehat{\theta}_\alpha}(z) \}^2 f_{\widehat{\theta}_\alpha}^{1+\alpha}(z)\, dz - \int v_{\widehat{\theta}_\alpha}(z) f_{\widehat{\theta}_\alpha}^{1+\alpha}(z)\, dz \\ &\quad + \int \left[v_{\widehat{\theta}_\alpha}(z) - \alpha \{ u_{\widehat{\theta}_\alpha}(z) \}^2 \right] f_{\widehat{\theta}_\alpha}^{\alpha}(z)\, dG_n\end{aligned}$$

$$= (1+\alpha)\int \{ u_{\widehat{\theta}_\alpha}(z) \}^2 f_{\widehat{\theta}_\alpha}^{1+\alpha}(z)\,dz - \int v_{\widehat{\theta}_\alpha}(z) f_{\widehat{\theta}_\alpha}^{1+\alpha}(z)\,dz + \frac{1}{n}\sum_{i=1}^{n}\left[v_{\widehat{\theta}_\alpha}(X_i) - \alpha \{ u_{\widehat{\theta}_\alpha}(X_i) \}^2 \right] f_{\widehat{\theta}_\alpha}^{\alpha}(X_i),$$

$$\widehat{K}(\alpha) = \int \{ u_{\widehat{\theta}_\alpha}(z) \}^2 f_{\widehat{\theta}_\alpha}^{2\alpha}(z)\,dG_n - \left[\int u_{\widehat{\theta}_\alpha}(z) f_{\widehat{\theta}_\alpha}^{\alpha}(z)\,dG_n \right]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} \{ u_{\widehat{\theta}_\alpha}(X_i) \}^2 f_{\widehat{\theta}_\alpha}^{2\alpha}(X_i) - \left[\frac{1}{n}\sum_{i=1}^{n} u_{\widehat{\theta}_\alpha}(X_i) f_{\widehat{\theta}_\alpha}^{\alpha}(X_i) \right]^2,$$

where, $u_{\widehat{\theta}_\alpha}(z) = \frac{\partial}{\partial\theta} \log f_\theta(z) \Big|_{\theta=\widehat{\theta}_\alpha}$ and $v_{\widehat{\theta}_\alpha}(z) = -\frac{\partial}{\partial\theta}\{u_\theta(z)\} \Big|_{\theta=\widehat{\theta}_\alpha}$. And we can obtain the criterion $\widehat{V} = \widehat{J}(\alpha)^{-2}\,\widehat{K}(\alpha)$.

3. Simulation Study

In this section, we will study the performance of the criterion through the simulation. The data are generated from the contaminated distribution $G = (1-\varepsilon)F_{\theta^*} + \varepsilon H$ for ε =0.00, 0.05, 0.10, 0.15,0.20, where F_{θ^*} is the true parametric distribution and H is a contaminating distribution. In subsection 3.1, the mean of the univariate normal distribution will be estimated. Estimating the standard deviation of the normal distribution will be treated in subsection 3.2. The subsection 3.3 is on the estimation of the exponential distribution. For each ε-contaminated underlying model, 100 samples of sample size n = 100 will be generated and the optimal α's, say $\widehat{\alpha}$'s, and the resulting estimators $\widehat{\theta}_{\widehat{\alpha}}$ are computed.

3.1. Mean of the univariate normal distribution

In this subsection, we consider the problem of estimating the mean θ^* of the assumed normal distribution, $N(\theta^*, \sigma^2)$, with σ^2 known, while the underlying distribution is a mixture of two normal distributions $N(\theta^*, \sigma^2)$ and $N(\theta_1, \sigma^2)$. In this case, $G = (1-\varepsilon)N(\theta^*, \sigma^2) + \varepsilon N(\theta_1, \sigma^2)$. Without loss of generality, we assume σ^2=1. Since

$$\widehat{u}(\alpha) = u_{\widehat{\theta}_\alpha}(z) = z - \widehat{\theta}_\alpha, \qquad \widehat{v}(\alpha) = v_{\widehat{\theta}_\alpha}(z) = 1,$$

the quantities $\widehat{J}(\alpha)$ and $\widehat{K}(\alpha)$ are as follows;

$$\begin{aligned}
\widehat{J}(\alpha) &= (1+\alpha)\int \{\widehat{u}(\alpha)\}^2 f_{\widehat{\theta}_\alpha}^{1+\alpha}(z)\,dz \\
&\quad - \int \widehat{i}(\alpha) f_{\widehat{\theta}_\alpha}^{1+\alpha}(z)\,dz + \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{i}(\alpha) - \alpha\{\widehat{u}(\alpha)\}^2\right] f_{\widehat{\theta}_\alpha}^{\alpha}(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{1-\alpha(X_i-\widehat{\theta}_\alpha)^2\right\}\left[\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(X_i-\widehat{\theta}_\alpha)^2}{2}\right\}\right]^{\alpha}, \\
\widehat{K}(\alpha) &= \frac{1}{n}\sum_{i=1}^{n}\{\widehat{u}(\alpha)\}^2 f_{\widehat{\theta}_\alpha}^{2\alpha}(X_i) - \left[\frac{1}{n}\sum_{i=1}^{n}\widehat{u}(\alpha) f_{\widehat{\theta}_\alpha}^{\alpha}(X_i)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\widehat{\theta}_\alpha)^2\left[\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(X_i-\widehat{\theta}_\alpha)^2}{2}\right\}\right]^{2\alpha} \\
&\quad - \left\{\frac{1}{n}\sum_{i=1}^{n}(X_i-\widehat{\theta}_\alpha)\left[\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(X_i-\widehat{\theta}_\alpha)^2}{2}\right\}\right]^{\alpha}\right\}^2.
\end{aligned}$$

We generate the data from G with $\theta^* = 1$, $\theta_1 = 10$. For each ε ($=0.00,\ 0.05,\ 0.10,\ 0.15,\ 0.20$), 100 samples of sample size n=100 are drawn. For each sample, we compute $\widehat{\alpha}$ as follows;

i) First, fix $\alpha \in [0,1]$ and compute $\widehat{\theta}_\alpha$.

ii) Compute $\widehat{V}(\alpha)$.

iii) Repeat i) and ii) for the values of $\alpha = 0.000, 0.001, \cdots, 1$.

iv) Find $\widehat{\alpha}$ at which the minimum of $\widehat{V}(\alpha)$ is achieved.

The mean(MEAN), standard deviation(STD) and the first and third quartiles (Q1 and Q3) of MLE and $\widehat{\theta}_{\widehat{\alpha}}$ are given in Table 2. For the uncontaminated data sets($\varepsilon = 0$), our estimation procedure chooses very small α's and the statistics for $\widehat{\theta}_{\widehat{\alpha}}$ and MLE are almost the same. (For more than half of the simulated data sets with $\varepsilon = 0$, $\hat{\alpha} = 0$ and thus $\widehat{\theta}_{\widehat{\alpha}}$ = MLE). For the contaminated data sets($\varepsilon > 0$), $\widehat{\theta}_{\widehat{\alpha}}$ shows better robustness to the contamination and has smaller standard deviation than MLE for each $\varepsilon > 0$.

3.2. Standard deviation of the univariate normal distribution

Now we investigate the case of estimating the standard deviation θ^* of the assumed normal distribution, $N(0,\ \theta^{*2})$, while the underlying distribution is a mixture of two normal distributions. Since

Table 2. Mean, standard deviation and first and third quartiles of MLE and $\widehat{\theta}_{\widehat{\alpha}}$ for the normal mean.

		ε				
Estimator	Statistic	0.00	0.05	0.10	0.15	0.20
MLE	MEAN	0.014	0.536	1.048	1.515	1.967
	STD	0.109	0.261	0.313	0.308	0.452
	Q1	-0.067	0.351	0.830	1.284	1.679
	Q3	0.090	0.708	1.232	1.687	2.243
$\widehat{\theta}_{\widehat{\alpha}}$	MEAN	0.013	0.005	0.003	0.007	0.028
	STD	0.108	0.106	0.092	0.108	0.123
	Q1	-0.070	-0.070	-0.060	-0.075	-0.050
	Q3	0.090	0.085	0.065	0.080	0.100
$\widehat{\alpha}$	MEAN	0.051	0.147	0.165	0.178	0.158
	STD	0.092	0.139	0.131	0.162	0.095
	Q1	0.000	0.094	0.111	0.122	0.129
	Q3	0.062	0.142	0.160	0.155	0.155

$$\widehat{u}(\alpha) = u_{\widehat{\theta}_\alpha}(z) = \frac{z^2}{\widehat{\theta}_\alpha^3} - \frac{1}{\widehat{\theta}_\alpha} \quad \text{and} \quad \widehat{v}(\alpha) = v_{\widehat{\theta}_\alpha}(z) = \frac{3z^2}{\widehat{\theta}_\alpha^4} - \frac{1}{\widehat{\theta}_\alpha^2},$$

$\widehat{J}(\alpha)$ and $\widehat{K}(\alpha)$ are given by

$$\begin{aligned}\widehat{J}(\alpha) &= \frac{\alpha}{\widehat{\theta}_\alpha^{2+\alpha}\sqrt{(2\pi)^\alpha(1+\alpha)}} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{3X_i^2}{\widehat{\theta}_\alpha^4} - \frac{1}{\widehat{\theta}_\alpha^2} - \alpha\left(\frac{X_i^2}{\widehat{\theta}_\alpha^3} - \frac{1}{\widehat{\theta}_\alpha}\right)^2\right\} \\ &\qquad \times \left\{\frac{1}{\sqrt{2\pi}\,\widehat{\theta}_\alpha}\exp\left(-\frac{X_i^2}{2\widehat{\theta}_\alpha^2}\right)\right\}^\alpha,\end{aligned}$$

$$\begin{aligned}\widehat{K}(\alpha) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i^2}{\widehat{\theta}_\alpha^3} - \frac{1}{\widehat{\theta}_\alpha}\right)^2\left\{\frac{1}{\sqrt{2\pi}\,\widehat{\theta}_\alpha}\exp\left(-\frac{X_i^2}{2\widehat{\theta}_\alpha^2}\right)\right\}^{2\alpha} \\ &\quad - \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i^2}{\widehat{\theta}_\alpha^3} - \frac{1}{\widehat{\theta}_\alpha}\right)\left\{\frac{1}{\sqrt{2\pi}\,\widehat{\theta}_\alpha}\exp\left(-\frac{X_i^2}{2\widehat{\theta}_\alpha^2}\right)\right\}^\alpha\right]^2.\end{aligned}$$

Table 3. Mean, standard deviation and first and third quartiles of MLE and $\widehat{\theta}_{\widehat{\alpha}}$ for the standard deviation of the normal distribution.

		ε				
Estimator	Statistic	0.00	0.05	0.10	0.15	0.20
MLE	MEAN	1.008	1.873	2.530	3.191	3.647
	STD	0.150	0.683	0.829	0.877	0.964
	Q1	0.891	1.419	1.881	2.585	2.978
	Q3	1.118	2.105	3.074	3.799	4.266
$\widehat{\theta}_{\widehat{\alpha}}$	MEAN	1.002	1.029	1.074	1.094	1.137
	STD	0.075	0.096	0.085	0.113	0.112
	Q1	0.945	0.962	1.013	1.023	1.061
	Q3	1.054	1.102	1.125	1.162	1.195
$\widehat{\alpha}$	MEAN	0.038	0.319	0.406	0.468	0.531
	STD	0.102	0.146	0.104	0.090	0.096
	Q1	0.000	0.279	0.356	0.423	0.485
	Q3	0.000	0.421	0.470	0.530	0.585

The data are generated from $G = (1-\varepsilon)\,N(0\,,\,1) + \varepsilon\,N(0\,,\,4)$. (Here $\theta^* = 1$ and $\theta_1 = 2$). In the same way as in the previous subsection, $\widehat{\alpha}$'s and $\widehat{\theta}_{\widehat{\alpha}}$'s are computed. The results are given in Table 3. From the results we can again see that the estimator $\widehat{\theta}_{\widehat{\alpha}}$ and MLE show similar behavior when $\varepsilon = 0$, and $\widehat{\theta}_{\widehat{\alpha}}$ has better robustness and stability than MLE.

3.3. Exponential distribution

In this subsection, we will estimate the parameter θ^* of the exponential distribution $\mathcal{E}(\theta)$. Since the true density is $f_\theta^*(z) = (1/\theta^*)\exp(-z/\theta^*)$, $z > 0$,

$$\widehat{u}(\alpha) = u_{\widehat{\theta}_\alpha}(z) = \frac{z}{\widehat{\theta}_\alpha^{\,2}} - \frac{1}{\widehat{\theta}_\alpha}, \qquad \widehat{v}(\alpha) = v_{\widehat{\theta}_\alpha}(z) = \frac{2z}{\widehat{\theta}_\alpha^{\,3}} - \frac{1}{\widehat{\theta}_\alpha^{\,2}}$$

and the quantities $\widehat{K}(\alpha)$ and $\widehat{J}(\alpha)$ are given by

$$\widehat{J}(\alpha) = \frac{\alpha}{(1+\alpha)\,\widehat{\theta}_\alpha^{2+\alpha}} + \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{2X_i}{\widehat{\theta}_\alpha^3} - \frac{1}{\widehat{\theta}_\alpha^2} - \alpha\left(\frac{X_i}{\widehat{\theta}_\alpha^2} - \frac{1}{\widehat{\theta}_\alpha}\right)^2\right\} \times \left\{\frac{1}{\widehat{\theta}_\alpha}\exp\left(-\frac{X_i}{\widehat{\theta}_\alpha}\right)\right\}^\alpha,$$

$$\widehat{K}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i}{\widehat{\theta}_\alpha^2} - \frac{1}{\widehat{\theta}_\alpha}\right)^2\left\{\frac{1}{\widehat{\theta}_\alpha}\exp\left(-\frac{X_i}{\widehat{\theta}_\alpha}\right)\right\}^{2\alpha} - \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i}{\widehat{\theta}_\alpha^2} - \frac{1}{\widehat{\theta}_\alpha}\right)\left\{\frac{1}{\widehat{\theta}_\alpha}\exp\left(-\frac{X_i}{\widehat{\theta}_\alpha}\right)\right\}^\alpha\right]^2.$$

Again 100 samples of sample size n = 100 are generated from $G = (1-\varepsilon)\,\mathcal{E}(1) + \varepsilon\,\mathcal{E}(5)$, for each ε. Here $\theta^* = 1$ and $\theta_1 = 5$.

Table 4. Mean, standard deviation and first and third quartiles of MLE and $\widehat{\theta}_{\widehat{\alpha}}$ for the exponential distribution.

		ε				
Estimator	Statistic	0.00	0.05	0.10	0.15	0.20
MLE	MEAN	0.994	1.218	1.412	1.556	1.738
	STD	0.111	0.177	0.237	0.225	0.271
	Q1	0.923	1.093	1.239	1.391	1.535
	Q3	1.070	1.323	1.558	1.685	1.892
$\widehat{\theta}_{\widehat{\alpha}}$	MEAN	0.986	1.067	1.144	1.196	1.226
	STD	0.113	0.128	0.139	0.172	0.188
	Q1	0.917	0.966	1.029	1.095	1.104
	Q3	1.067	1.171	1.237	1.299	1.376
$\widehat{\alpha}$	MEAN	0.035	0.223	0.320	0.392	0.518
	STD	0.113	0.159	0.206	0.199	0.228
	Q1	0.000	0.135	0.198	0.272	0.365
	Q3	0.017	0.286	0.422	0.450	0.642

Table 4 shows the statistics for MLE, $\widehat{\theta}_{\widehat{\alpha}}$, and $\widehat{\alpha}$. Again, these results show that $\widehat{\theta}_{\widehat{\alpha}}$ is almost the same as MLE when $\varepsilon = 0$. When $\varepsilon > 0$, $\widehat{\theta}_{\widehat{\alpha}}$ is robust to the contamination and has smaller variation than MLE.

4. Concluding Remarks

In this paper, we investigate how to select the tuning parameter α in MDPD estimation when estimating the true parameter θ^* is the object of estimation while the data are contaminated. We suggest a data-driven criterion $\widehat{V}(\alpha)$ for α-selection and study its performance through simulation. The simulation study includes three cases, the normal mean, the standard deviation of normal distribution, and the exponential distribution. In all of the three cases, MLE and $\widehat{\theta}_{\widehat{\alpha}}$ show almost the same performance for the uncontaminated data. For the ε-contaminated data, as ε increases $\theta_{\widehat{\alpha}}$ shows better robustness and stability than MLE.

Since the object of estimation is to get a good estimator of the true parameter θ^* not of θ and MDPD estimator does not consistently estimates θ^* but θ, more reasonable criterion would be $(\theta - \theta^*)^2 + V$. One of the referees pointed out this. But it is thought to be quite impossible to get reasonable estimate of $(\theta - \theta^*)^2$ because the data are contaminated. Fortunately, if there is no contamination, $\theta = \theta^*$ and MLE($\alpha = 0$) minimizes V. One of our research purposes is automatic selection of $\alpha \simeq 0$ when data are not contaminated, which was shown to be nearly achieved by our criterion through simulation. We haven't studied the asymptotic properties of $\widehat{\alpha}$ and $\widehat{\theta}_{\widehat{\alpha}}$, yet. It is expected to be comlicated to derive the asymptotic properties and that would be our future research.

Acknowledgement

The authors would like to express their thanks to the referees. The referees' pinpoint comments were very helpful to make this paper much more presentable.

References

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

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

Cao, R., Cuevas, A., and Fraiman, R. (1995). "Minimum distance density-based estimation," *Computational Statistics & Data Analysis,* **20**, 611-631.

Donoho, D. L. and Liu, R. C. (1988). "The automatic robustness of minimum distance functionals," *The Annals of Statistics,* **16**, 552-586.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions,* Wiley: New York.

Hjort, N. L. (1994). "Minimum L_2 and robust Kullback-Leibler estimation," *Proceedings of the 12th Prague Conference on Information Theory, Statistical Decision Functions and Random Process*, Ed. P. Lachout and J. Á. Víšek, 102-105. Prague: Academy of Sciences of the Czech Republic.

Kullback, S. and Leibler, R. A. (1951). "On information and sufficiency," *The Annals of Mathematical Statistics,* **22**, 79-86.

Scott, D. (1999). "Parametric modelling by minimum L_2 error," *Manuscript.*