Communications of the Korean Statistical Society
2009, Vol. 16, No. 6, 989–995

The Minimum Squared Distance Estimator and the Minimum Density Power Divergence Estimator

Ro Jin Pak[1,a]

[a]Department of Information and Statistics, Dankook University

Abstract

Basu *et al.* (1998) proposed the minimum divergence estimating method which is free from using the painful kernel density estimator. Their proposed class of density power divergences is indexed by a single parameter α which controls the trade-off between robustness and efficiency. In this article, (1) we introduce a new large class of the minimum squared distance which includes from the minimum Hellinger distance to the minimum L_2 distance. We also show that under certain conditions both the minimum density power divergence estimator(MDPDE) and the minimum squared distance estimator(MSDE) are asymptotically equivalent and (2) in finite samples the MDPDE performs better than the MSDE in general but there are some cases where the MSDE performs better than the MDPDE when estimating a location parameter or a proportion of mixed distributions.

Keywords: Asymptotic equivalence, density power divergence, Hellinger distance, L_2 distance.

1. Introduction

Basu *et al.* (1998) proposed a minimum divergence estimating method which used new density-based divergences. Unlike existing methods of this type such as minimum Hellinger distance estimation (Beran, 1977; Simpson, 1987; Tamura and Boos, 1986), Basu *et al.* (1998) avoid the use of nonparametric density estimation and associated complications such as bandwidth selection.

In this paper, a new class of squared distance and the associated estimator, called the minimum squared distance estimator, are introduced. The Hellinger distance and the ordinary L_2 distance are members of the new class. It is shown that the MDPDE and the MSDE are asymptotically equivalent under model conditions. Empirical studies show that the MDPDE performs better or similar to the MSDE in terms of mean squared error under most of the distributions considered, but when the true distribution is heavily contaminated, some advantages of the MSDE are identified.

2. Review of the Minimum Density Power Divergence Estimator

Define the divergence $d_\alpha(g, f)$ between density functions g and f to be

$$d_\alpha(g, f) = \int \left\{ f^{1+\alpha}(x) - \left(1 + \frac{1}{\alpha}\right) g(x) f^\alpha(x) + \frac{1}{\alpha} g^{1+\alpha}(x) \right\} dx, \quad (\alpha > 0).$$

When $\alpha = 0$, the divergence $d_0(g, f)$ is defined as

$$d_0(g, f) = \lim_{\alpha \to 0} d_\alpha(g, f) = \int g(x) \log \left\{ \frac{g(x)}{f(x)} \right\} dx.$$

This present research was conducted by the research fund of Dankook University in 2009.

[1] Professor, Department of Information and Statistics, Dankook University, Yongin 448-160, Korea.
E-mail: rjpak@dankook.ac.kr

Note that $d_0(g, f)$ is the Kullback-Leiber divergence. Consider a parametric family of models $\{F_\theta\}$, indexed by the unknown parameter θ (or, a vector of parameters), possessing densities $\{f_\theta\}$ with respect to Lebesgue measure. Given a random sample $X_1, \ldots, X_n$ from a distribution G (G may not belong to $\{F_\theta\}$) , the minimum density divergence estimator $\hat{\theta}$ is generated by minimizing

$$\int f_\theta^{1+\alpha}(x)dx - \left(1 + \frac{1}{\alpha}\right) n^{-1} \sum f_\theta^\alpha(X_i)$$

with respect to θ. Then the estimating equations have the form

$$U_n(\theta) \equiv n^{-1} \sum u_\theta(X_i) f_\theta^\alpha(X_i) - \int u_\theta(z) f_\theta^{1+\alpha}(z)dz, \tag{2.1}$$

where $u_\theta(z) = \partial \log f_\theta(z)/\partial\theta$ is the score function. The minimum density power divergence estimators are in fact M-estimators, and the corresponding ψ function is

$$\psi(x, \theta) = u_\theta(x) f_\theta^\alpha(x) - \int u_\theta(x) f_\theta^{1+\alpha}(x)dx.$$

The proposed class of 'density power divergences' is indexed by a single parameter α which controls the trade-off between robustness and efficiency. Choices of α near zero afford considerable robustness while retaining efficiency close to that of maximum likelihood.

Theorem 1. (Basu *et al.*, 1998) *Suppose that the true density belongs to the parametric family $\{f_\theta\}$, and under certain regularity conditions, there exists $\hat{\theta}$ such that, as $n \to \infty$,*

(i) $\hat{\theta}$ is consistent for θ.

(ii) At the assumed model f_θ, $n^{1/2}(\hat{\theta} - \theta)$ is asymptotically multivariate normal with vector mean zero and covariance matrix $J^{-1}KJ^{-1}$, where

$$J = \int u_\theta(z) u_\theta^T(z) f_\theta^{1+\alpha}(z)dz,$$
$$K = \int u_\theta(z) u_\theta^T(z) f_\theta^{2\alpha+1}(z)dz - \eta\eta^T, \quad \eta = \int u_\theta(z) f^{\alpha+1}(z)dz.$$

Note that, in the limit $\alpha \to 0$, J and K both become equal to the Fisher information $I^{-1}(\theta)$, and the covariance matrix becomes $I^{-1}(\theta)$.

3. Minimum Squared Distance Estimator: Definition and Asymptotic Results

With G, g, F_θ and f_θ being defined in the preceding section, suppose a random sample $X_1, X_2, \ldots, X_n$ from G, and then we begin this section with a definition of the new family of distances.

Definition 1. *Define the squared distance between g and f_θ to be*

$$d_\beta(g, f_\theta) = \int \left\{ g^{\frac{1}{\beta}}(x) - f_\theta^{\frac{1}{\beta}}(x) \right\}^2 dx,$$

indexed by $\beta \in [1, 2]$. The minimum squared distance estimator $\hat{\theta}$, as the value at $\hat{g}_n$ of a functional T, is defined as

$$\hat{\theta} = T(\hat{g}_n) = \min_\theta \int \left\{ \hat{g}_n^{\frac{1}{\beta}}(x) - f_\theta^{\frac{1}{\beta}}(x) \right\}^2 dx,$$

where $\hat{g}_n$ is a suitable density estimator for g

$$\hat{g}_n = (nc_ns_n)^{-1}\sum_{i=1}^{n} w\left[(c_ns_n)^{-1}(x - X_i)\right],$$

$\{c_n\}$ being a sequence of constants to zero at an appropriate rate, s_n being a scale estimator, and w being a smooth density on the real line. If $\beta = 2$, $d_\beta(g, f_\theta)$ is the Hellinger distance and if $\beta = 1$, it becomes L_2-distance.

Theorem 2. *Suppose*

(i) w is absolutely continuous and has compact support; w' and w'' are bounded.

(ii) g is uniformly and absolutely continuous and g'' is bounded and $g > 0$.

(iii) $\lim_{n\to\infty} c_n = 0$, $\lim_{n\to\infty} n^{1/2}c_n = \infty$ *and* $\lim_{n\to\infty} n^{1/2}c_n^2 = 0$.

(iv) As $n \to \infty$, $s_n \to_p s$ a positive finite constant depending on g.

(v) $u_\theta, u'_\theta, u''_\theta$ are bounded by the functions of x , whose expectation w. r. t. f_θ should be finite,

then $n^{1/2}(\hat{\theta}_n - \theta)$ is asymptotically multivariate normal with mean zero and covariance matrix $J(g)^{-1}$ $K(g)J(g)^{-1}$, where

$$K(g) = \int u_\theta(x)u_\theta^T(x)f^{\frac{4}{\beta}-2}(x)g(x)dx - \eta_\theta\eta_\theta^T \quad \textit{and} \quad \eta_\theta = \int u_\theta(x)f^{\frac{2}{\beta}-1}(x)g(x)dx.$$

and

$$J(g) = \int \frac{-2}{\beta^2}u_\theta(x)u_\theta^T(x)f_\theta^{\frac{2}{\beta}}(x)dx + \int \frac{-2}{\beta}\left\{u'_\theta(x) + \frac{1}{\beta}u_\theta(x)u_\theta^T(x)\right\} f_\theta^{\frac{1}{\beta}}(x)\left\{g^{\frac{1}{\beta}}(x) - f_\theta^{\frac{1}{\beta}}(x)\right\} dx.$$

Furthermore, at the model $n^{1/2}(\hat{\theta}_n - \theta)$ is asymptotically multivariate normal with mean zero and covariance matrix $J^{-1}KJ^{-1}$, where

$$J = \int u_\theta(x)u_\theta^T(x)f_\theta^{\frac{2}{\beta}}(x)dx \tag{3.1}$$

$$K = \int u_\theta(x)u_\theta^T(x)f^{\frac{4}{\beta}-1}(x)dx - \eta_\theta\eta_\theta^T, \quad \eta_\theta = \int u_\theta(x)f^{\frac{2}{\beta}}(x)dx. \tag{3.2}$$

In the limit $\beta \to 2$, J and K both become the Fisher Information $I^{-1}(\theta)$, the covariance matrix becomes $I^{-1}(\theta)$.

Proof: $\| \hat{g}_n(x) - g(x) \| \overset{p}{\to} 0$ as $n \to \infty$ by the proof of Theorem 2 in Beran (1977). Note that by the Taylor we have the approximation

$$\hat{g}_n^{\frac{2}{\beta}}(x) \approx g^{\frac{2}{\beta}}(x) + (\hat{g}_n(x) - g(x))\frac{2}{\beta}g^{\frac{2}{\beta}-1}(x),$$

so that

$$\limsup_{n\to\infty}\int\left|\hat{g}_n^{\frac{2}{\beta}}(x)-g^{\frac{2}{\beta}}(x)\right|dx=\limsup_{n\to\infty}\int|\hat{g}_n(x)-g(x)|\frac{2}{\beta}g^{\frac{2}{\beta}-1}(x)dx$$

$$\leq\limsup_{n\to\infty}\int|\hat{g}_n(x)-g(x)|dx\to 0.$$

That is, $\| \hat{g}_n^{2/\beta} - g^{2/\beta} \| \xrightarrow{p} 0$ as $n \to \infty$. Rewrite $g^{1/\beta}(x)$ as $(g^{2/\beta}(x))^{1/2}$ and let $g^{2/\beta}(x)$ and $\hat{g}_n^{2/\beta}(x)$ play a role of $g(x)$ and $\hat{g}_n$ in the proof of Theorem 3 in Beran (1977). Therefore, we have the results. □

Remark 1. By the Theorem 1 and the Theorem 2, both the the MDPDE and the MSDE have the same asymptotic distribution if $\alpha = 2/\beta - 1$.

4. Simulations

4.1. Location parameter

The performance of the the MDPDE and the MSDE for estimating a location parameter μ of $N(\mu, 1)$ in a small sample Monte Carlo study are presented in Table 1. The model distribution was standard normal, and the true distributions, where samples were generated, were

- standard normal distribution, denoted by $N(0, 1)$,
- a contaminated standard normal distribution with 10% contamination of $N(3, 1)$, denoted by $10\%3N$,
- a contaminated standard normal distribution with 30% contamination of $N(3, 1)$, denoted by $30\%3N$,
- a contaminated standard normal distribution with 10% contamination of Uniform distribution $U(-3, 3)$, denoted by $10\% \pm 3U$,
- t-distribution with 5 *d.f.*, denoted by $t(5)$ and
- Uniform distribution $U(-3, 3)$.

The Epanechnikov kernel, $(3/4)(1 - t^2/5)/\sqrt{5}I_{-\sqrt{5},\sqrt{5}}(t)$ (Silverman, 1986) is used for calculating the MSDE. For simulations, $\beta = 3/2$, which is in the middle of $[1, 2]$, is chosen for MSDE, therefore MDPDE is calculated with $\alpha = 1/3$. The 500 random samples of size 10, 20 and 50 were generated from the true distributions, and the MDPDE and the MSDE for various window widths (h = 1.0, 1.5 and 2.0) were calculated. Both the MDPDE and MSDE perform similar in terms of bias and/or mean square error. However, under asymmetric distributions such as $30\%3N$ and $10\% \pm U$ we can find the cases where the MSDE performs better than the MDPDE. It can be said when the data are from an heavily asymmetric true distribution, a kernel density estimator is more appropriate in constituting a distance with the model density $N(0, 1)$, that is, the effect of smoothing data seems very useful in reducing biases and/or mean square in most cases (Table 1).

Table 1: MDPDE and MSDE for a location under various distributions

No. of Obs.	estimator	statistics	distribution					
			$N(0,1)$	$10\%3N$	$30\%3N$	$10\% \pm 3U$	t_5	$\pm 3U$
10	MDPDE	bias	−.0016	−.0071	.4362	.0175	.0345	−.0248
		m.s.e.	.0976	.1099	.3754	.1341	.1290	.1290
	MSDE	bias	.0056	−.0038	.2806	.0224	.0278	.0402
	$h = 1.0$	m.s.e.	.1181	.1274	.4416	.1529	.1733	1.6438
	MSDE	bias	.0013	.0040	.3310	.0184	.0340	.0370
	$h = 1.5$	m.s.e.	.1062	.1158	.4311	.1425	.1823	1.3397
	MSDE	bias	−.0136	.0016	.3590	.0166	.0338	.0322
	$h = 2.0$	m.s.e.	.1193	.1117	.4327	.1367	.1766	1.2471
20	MDPDE	bias	.0026	.0264	.4755	.0235	−.0071	−.0200
		m.s.e.	.0540	.0632	.3240	.1031	.0691	.4415
	MSDE	bias	.0038	−.0148	.3660	.0284	−.0078	.0204
	$h = 1.0$	m.s.e.	.0588	.0654	.2850	.0803	.0728	1.0554
	MSDE	bias	.0033	.0132	.3888	.0278	.0098	.0240
	$h = 1.5$	m.s.e.	.0559	.0637	.2929	.0726	.0719	.9706
	MSDE	bias	.0030	.0110	.4044	.0340	−.0086	.0158
	$h = 2.0$	m.s.e.	.0552	.0633	.3014	.0687	.0723	.9083
50	MDPDE	bias	.0112	.0340	.5064	.0216	−.0023	.0216
		m.s.e.	.0212	.0286	.3002	.0286	.0288	.2392
	MSDE	bias	.0121	.0156	.3522	.0134	−.0072	.0218
	$h = 1.0$	m.s.e.	.0222	.0278	.1837	.0265	.0307	.6023
	MSDE	bias	.0116	.0210	.3636	.0124	−.0066	.0362
	$h = 1.5$	m.s.e.	.0218	.0283	.1907	.0262	.0311	.5549
	MSDE	bias	.0114	.0242	.3798	.0122	−.0072	.0404
	$h = 2.0$	m.s.e.	.0217	.0285	.2007	.0259	.0305	.5204

Table 2: MDPDE and MSDE for a proportion under various distributions; for example, $30\%1N$ implies a mixture density as $0.7N(0,1) + 0.3N(1,1)$.

No. of Obs.	estimator	statistics	mixing density					
			$30\%1N$	$50\%1N$	$30\%3N$	$50\%3N$	$30\%5N$	$50\%5N$
10	MDPDE	bias	−.2326	−.2513	−.0869	−.0907	−.0379	−.0408
		m.s.e.	.0559	.0964	.0295	.0116	.0023	.0026
	MSDE	bias	.0634	.0322	.0096	.0023	−.0010	.0005
	$h = 1.0$	m.s.e.	.0681	.0851	.0101	.0111	.0040	.0042
	MSDE	bias	.0531	.0150	.0115	.0036	−.0007	.0002
	$h = 1.5$	m.s.e.	.0643	.0773	.0075	.0080	.0022	.0030
	MSDE	bias	.0558	−.0048	.0107	−.0026	−.0014	.0001
	$h = 2.0$	m.s.e.	.0645	.0209	.0072	.0071	.0022	.0021
20	MDPDE	bias	.0073	−.0138	.0026	.0050	.0014	.0017
		m.s.e.	.0410	.0490	.0037	.0036	.0005	.0007
	MSDE	bias	.0309	−.0053	.0066	.0010	−.0035	.0010
	$h = 1.0$	m.s.e.	.0411	.0518	.0041	.0043	.0013	.0013
	MSDE	bias	.0340	−.0010	.0074	.0034	−.0008	.0010
	$h = 1.5$	m.s.e.	.0386	.0495	.0038	.0043	.0009	.0011
	MSDE	bias	.0387	−.0063	.0103	.0040	.0006	.0011
	$h = 2.0$	m.s.e.	.0367	.0418	.0033	.0036	.0008	.0009
50	MDPDE	bias	.0044	.0055	.0003	−.0031	.0010	−.0003
		m.s.e.	.0204	.0209	.0014	.0016	.0001	.0003
	MSDE	bias	.0184	.0059	.0018	.0026	.0001	.0003
	$h = 1.0$	m.s.e.	.0195	.0209	.0015	.0017	.0003	.0004
	MSDE	bias	.0205	.0044	.0053	−.0006	−.0002	.0010
	$h = 1.5$	m.s.e.	.0175	.0196	.0014	.0016	.0004	.0004
	MSDE	bias	.0298	.0040	.0111	.0011	.0016	−.0010
	$h = 2.0$	m.s.e.	.0161	.0182	.0014	.0017	.0002	.0004

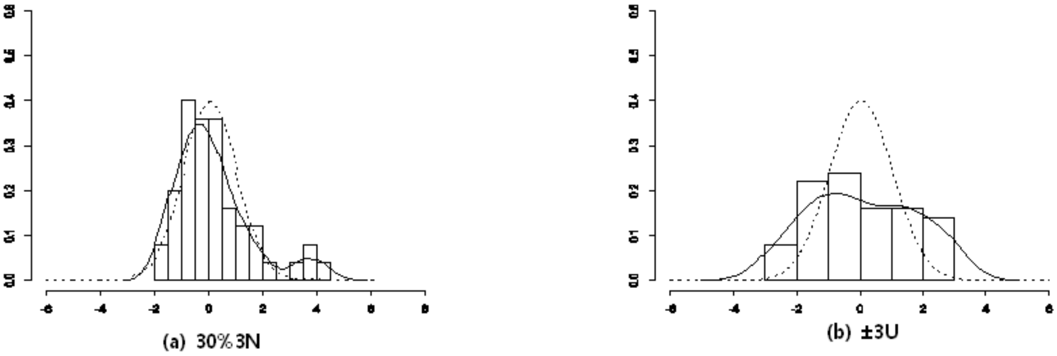

Figure 1: *Empirical density, kernel density and standard normal density (dotted line): a model density is* $N(0, 1)$*, a true density is (a)* $30\%3N$ *and (b)* $\pm 3U$*.*

4.2. Mixture proportion

Consider the estimation of the proportions $\theta_1, \ldots, \theta_m$ in the mixture density

$$f_\theta(x) = \theta_1 f_1(x) + \theta_2 f_2(x) + \cdots + \theta_m f_m(x),$$

where $\theta = \{\theta_1, \ldots, \theta_m\}$ and $f_1(x), \ldots, f_m(x)$ are densities. Here, once again $\alpha = 1/3$ and $\beta = 3/2$ are used for the MDPDE and the MDSE, respectively. In order to see the performance of MDPDE and MSDE for estimating proportions, simulations were carried out with 500 random samples of sizes, 10, 20 and 50, generated from 30% and 50% mixtures of two normals, $N(0, 1)$ with $N(1, 1), N(3, 1)$ and $N(5, 1)$, respectively (Table 2).

We have discovered that the MSDE is consistently better than the MDPDE in terms of having smaller bias and mean squared error when the level of contamination is high like 50% and the mixing distribution is far from $N(0, 1)$ like $N(5, 1)$.

5. Conclusion

The minimum density power divergence estimators and the minimum squared distance estimators are asymptotically equivalent with $\alpha = 2/\beta - 1$. Though the former performs better than the latter in general, some advantages of the latter over the former are identified when the true density is heavily contaminated or asymmetric. The magnitude of asymptotic statistics, say $K(g)$ and $J(g)$, depend on the closeness of a model density and a density estimator for a true density. Since the true density is estimated by an empirical density for the MDPDE and by an smoothed (kernel) density estimator for the MSDE, if a true density is better estimated by a smoothed (kernel) density estimator than by an empirical density the MSDE would perform better than the MDPDE. If, for example,a true density is $30\%3N$, a kernel density is closer to the normal density near the interval where contaminations take place (Figure 1(a)), but if a true density is $\pm 3U$, a kernel density overestimate on both ends while an empirical density fits the normal density within it's boundary (Figure 1(b)). Hence, we can claim that the MSDE performs better than MDPDE when the true density is $30\%3N$ while it is reversed when the true density is $\pm 3U$. However, we have to admit that the MSDE may not prevail the MDPDE as long as we consider the cost of choosing a bandwidth. Though we have to pay the cost of choosing a bandwidth, it is also should be admitted that there are still some cases where smoothing is mattered. As a future research, it remains to identify and verify the conditions which a smoothed (kernel) density estimator outperform an empirical density in estimating a true density.

References

Basu, A., Harris, I. R., Jort, 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 estimators for parametric models, *Annals of Statistics*, **5**, 445–463.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, 34–43.

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

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

Received September 2009; Accepted October 2009