Journal of Korean
Data & Information Science Society
2006, Vol. 17, No. 1, pp. 213 ~ 220

Reducing Bias of the Minimum Hellinger Distance Estimator of a Location Parameter

Ro Jin Pak[1)]

Abstract

Since Beran (1977) developed the minimum Hellinger distance estimation, this method has been a popular topic in the field of robust estimation. In the process of defining a distance, a kernel density estimator has been widely used as a density estimator. In this article, however, we show that a combination of a kernel density estimator and an empirical density could result a smaller bias of the minimum Hellinger distance estimator than using just a kernel density estimator for a location parameter.

Keywords : Density estimator, Efficiency, Hellinger distance, Robustness

1. Background and Motivation

Robustness procedures typically obtain robustness at the expense of not being optimal at the true model. However, Beran (1977) suggested the use of the minimum Hellinger distance (MHD) estimator which has certain robustness properties and is asymptotically efficient at the true model. Let $X_1, X_2, \cdots, X_n$ be a random sample from a population having a continuous probability which is a member of some postulated parametric family of densities $\{f_\theta; \theta \in \Theta\}$. The MHD estimator of θ is defined as a value of $\hat{\theta} = t(\hat{g}_n)$ which minimizes $\| f_\theta^{1/2} - \hat{g}^{1/2_n} \|$ ($\| \cdot \|$ is the usual L_2 norm) where $\hat{g}_n$ is a suitable nonparametric density estimator such as

1) Associate Professor, Department of Computer Sciences and Statistics, Dankook University, Seoul, 140-714, Korea.
E-mail : rjpak@dankook.ac.kr

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

where w is a density on R^k and $\{h_n\}$ is sequence of constants converging to zero. $t(\hat{g}_n)$ is a functional representation for $\hat{\theta}$ and note that $t(f_\theta) \equiv \theta$. Tamura and Boos (1986) provided a very important theorem which concluded with the asymptotic statement about the estimator for multivariate location and covariance as follows;

Theorem 1. (Tamura and Boos (1986)) Let $X_1, X_2, \cdots, X_n$ be a sample of independent and identically distributed k -vectors with density g and let $\hat{g}_n(x)$ be a kernel estimator. Suppose there exists the MHD estimator $\hat{\theta} = t(\hat{g}_n)$, then under some conditions,

$$\sqrt{n}(t(\hat{g}_n) - t(g) - B_n) \to N(0, \Sigma_g),$$

where

1. Let $s_\theta = f_\theta^{1/2}$ and let $\dot{s}_\theta$ and $\ddot{s}_\theta$ be the vector of the first and the second partial derivatives of s_θ with respect to θ.
2. $\psi_g(x) = -[\int \dot{s}_{t(g)}(x)\dot{s}_{t(g)}(x)^T dx]^{-1/2}\, \dot{s}_{t(g)}(x)/2g^{1/2}(x)$
3. $\Sigma_g = \int \psi_g(x)\psi_g(x)^T g(x)dx.$
4. $B_n = 2C_n^* \int \psi_g(x)\, \tilde{g}_n^{1/2}(x) g^{1/2}(x)dx$ with $C^* \to 1_{n\times n}$ (identity matrix) in probability.
5. $\tilde{g}_n = E[\hat{g}_n]$
6. When $g = f_\theta$, $t(f_\theta) = \theta$ and the covariance matrix $\Sigma_g = I_f$ (information matrix).

The existence of the bias term B_n in the above theorem is a critical drawback of the theorem, though the actual bias could become negligible when a data set is large. In this article, we are trying to find a way to reduce or remove the bias and we are able to find the solution in the form of a density estimator by taking the combination of an empirical density and a kernel density as a density estimator. It is shown that this new technique could reduce biases of MHD estimators under various distributions. Only the location parameter estimation is considered in this article - though results are not mentioned in this article, simulations for estimating a scale parameter were unsatisfactory and research is still on its way.

2. Bias of the MHDE

We begin by considering how to remove or reduce the bias B_n in asymptotic

$$\widehat{g}_n \to E[\widehat{g}_n] = \frac{1}{nh_n}\sum_{i=1}^{n} E_{X_i}\left[w\left(\frac{x-X_i}{h_n}\right)\right] = \frac{1}{h_n}\int w\left(\frac{x-y}{h_n}\right)dG(y),$$

sense. Since we have by the law of large number as $n\to\infty$, $nh_n\to\infty$, and at the model $\widehat{g}_n \to \tilde{f}_\theta \neq f_\theta$ ($\widehat{g}_n$ is not an asymptotically unbiased estimator for f_θ),

$$\tilde{f}_\theta(x) \equiv \frac{1}{h_n}\int w\left(\frac{x-y}{h_n}\right)dF_\theta(y),$$

where
if F_θ is the cumulative distribution function of f_θ.

Comment:

In addition to $n\to\infty$, $nh_n\to\infty$, if we have $h_n\to 0$, then $\tilde{f}_\theta \to f_\theta$ and finally

$$\begin{aligned} B_n &\to -C_n^*\left[\int \dot{s}_\theta(x) f^{1/2}{}_\theta(x)\right]^{-1/2}\int \dot{s}_\theta(x)\tilde{f}^{1/2}{}_\theta(x)\,dx \\ &\to -C_n^*\left[\int \dot{s}_\theta(x) f^{1/2}{}_\theta(x)\right]^{-1/2}\int \dot{s}_\theta(x) f_\theta^{1/2}(x)\,dx \qquad (x)\ f^{1/2}{}_\theta(x)\,dx = 0. \\ &\propto \int \frac{1}{2}\dot{f}_\theta(x)\ f^{-1/2}{}_\theta \end{aligned}$$

Conjecture: By the comment in the above, we know that to get rid of the bias term B_n, h_n should go to 0. Therefore, 'smoothness' of the density estimator may need to be dropped for reducing bias. We conjecture that MHD with an empirical density estimator may perform better than MHD with a kernel density estimator. Consider an empirical density estimator,

$$\widehat{g}_e(x) = \frac{1}{n}\sum_{i=1}^{n} w(X_i, x),$$

by putting $w(t,x) = 1/b(x)$, where $b(t)$ is the width of the bin containing t, if x and t fall in the same bin, or $w(t,x) = 0$ otherwise.

Though we use an empirical density in stead of smoothed kernel densities, Theorem 1 still works.

Simulations: In order to verify that conjecture, we run simulations as follows:

- Generate 500 samples of size 10, 20 and 50 from $N(0,1)$, a contaminated normal distribution with 10% contamination of $N(3,1)$ (denoted by $10\%3N$), a contaminated normal distribution with 10% contamination of $N(10,1)$ (denoted by $10\%10N$) and t-distribution with three degrees of freedom.
- Calculate (i) MHD estimators with an empirical density (MHDe) and (ii) MHD estimators with the Gaussian kernel density (MHDk). $b_{opt} = 6^{1/3} R(f)^{-1/3} n^{-1/3}$

and $h_{opt} = k_2^{-4/5} R(k)^{1/5} R(f')^{-1/5} n^{-1/5}$ are used as the optimal bin width and bandwidth, respectively, which minimize the approximate mean integrated square error with an empirical density (AMISEe) and the approximate mean integrated square error with a kernel density (AMISEk) as below. For this simulation, we use the basic optimal values;

$$b_{opt} = 3.5\sigma n^{-1/3} \text{ and } h_{opt} = 1.06\,\sigma n^{-1/5}.$$

With an empirical density estimator $\hat{g}_e$, we have

$$AMISE_e = \frac{1}{nb} + \frac{1}{12} b^2 R(f),$$

where the roughness function is defined by $R(\phi) = \int \phi^2(x)dx$ (Scott, 1992). With a kernel density estimator $\hat{g}_k$,

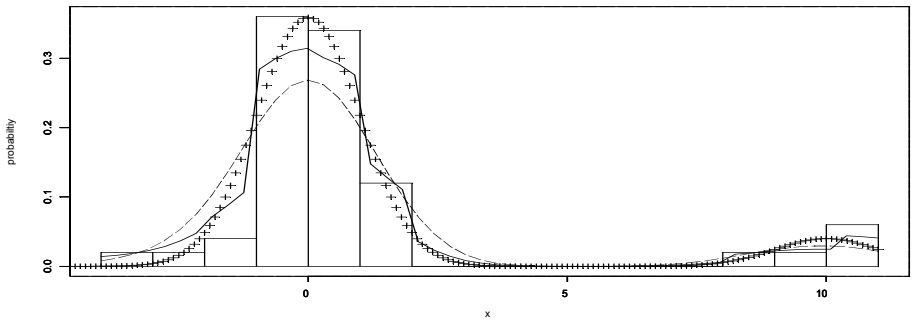
$$AMISE_k = \frac{R(k)}{nh} + \frac{1}{4} h^4 k_2^4 R(f'),$$

where h is a bandwidth, $k(\cdot)$ is a given kernel function and $k_2^2 = \int t^2 k(t)dt$ (Silverman, 1986).

Observation: In Table 1, we can observe that the MHDe tends to have smaller bias than the MHDk under true distributions such as $N(0,1)$ and $t(3)$, but under mixture distributions like $10\%3N$ and $10\%10N$ with few exception where $n=10$ the biases of the MHDk are smaller than those of the MHDe. The simulation implies that using either only an empirical density estimator or a

kernel density estimator as usual does not always produce promising results.

We can argue that for the mixed distribution such as $10\%3N$ and $10\%10N$ empty bins like the ones in Figure 2 could cause increase the discrepancy between a histogram and a model density, that is, a smoothed density estimator like a kernel density estimator is better to approximate a model density than an empirical density estimator. Therefore, since we do not know a exact form of model density in practice, it is safe working with a density estimator, which is smooth on the intervals of empty bins and at time same ensures smaller bias like an empirical density.



<Figure 1> A histogram, a kernel density (dot) and a combination of a histogram and a kernel density with $\alpha=0.5$ (solid) based on a random sample of size 50 from $10\%10N$ (curve with +)

3. Bias reduction for the MHDE

In the previous section, we know that $\hat{g}_n \to f_\theta \Rightarrow B_n \to 0$, that is, the closeness of $\hat{g}_n$ and f_θ are is a key factor to reduce bias of MHD estimator. We also know that closeness of a density estimator and a model density can be measured by the approximate mean integrated square error. In Figure 2 shows, the AMISE with an empirical density are smaller/larger than that with a kernel estimator depending on the width of a bin or a bandwidth (the Epanechnikov kernel is used). Since we do not know a true distribution in practice, we argue that using either only an empirical density estimator or only a kernel density estimator is not a good idea in reducing bias of an MHD estimator.

We propose to use a combined density estimator such as $\hat{g}_{ek}=\alpha\ \hat{g}_e+(1-\alpha)\ \hat{g}_k$. Simple manipulations give for $\alpha\in[0,1]$

$$AMISE_{ek} = \alpha^2 AMISE_e + (1-\alpha)^2 AMISE_k + higher-order\ terms\ \text{in}\ h.$$

Figure 2 shows that AMISE with combined density estimator ($\alpha = 0.5$) is between two curves (AMISE with an empirical density estimator and a kernel density estimator). We need a properly chosen α to form $\hat{g}_{ek}$, which could be found as the value of α minimizing $AMISE_{ek}$, that is, to solve the following equations;

$$\partial AMISE_{ek}/\partial b = 0, \quad \partial AMISE_{ek}/\partial h = 0, \quad \text{and} \quad \partial AMISE_{ek}/\partial \alpha = 0,$$

where $\hat{b}$ and $\hat{h}$ be the solutions from the first and second equations, respectively. By plugging $\hat{b}$ and $\hat{h}$ into the third equation, we get an optimal value for α,

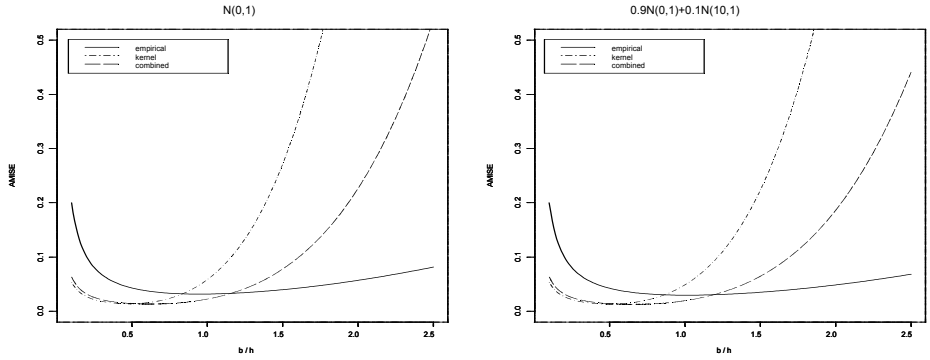
$$\hat{\alpha} = \frac{AMISE_k}{(AMISE_e + AMISE_k)} \Big|_{b=\hat{b},\, h=\hat{h}}.$$

With $\hat{b} = b_{opt}$ and $\hat{h} = h_{opt}$, mentioned in section 2, we have

$$\hat{\alpha} = \frac{(5/4)[k_2 R(k)]^{4/5} R(f')^{1/5} n^{-4/5}}{(3/4)^{2/3} R(f)^{1/3} n^{-2/3} + (5/4)[k_2 R(k)]^{4/5} R(f')^{1/5} n^{-4/5}}.$$

Table 1 shows that the biases of MHDek with $\hat{\alpha}$ are between those of MHDe and MHDk. In practice, since we don't know the exact underlying distribution, it would be useful and safe considering MHDek along with MHDk. The theoretically sound estimator for α will be obtained by replacing $R(f)$ and $R(f')$ by corresponding data driven estimators such as ones in Sheather & Jones (1991). We would like to reserve detail discussion for the next time, until the materials in this article are proven to be reasonable.

However, with help of a fast computer, we can calculate the Hellinger distance on a grid of (α, μ) and pick up the α at which the distance is minimum by searching it on a fine grid. For example, here we have a sample of size 10 from $10\%10N$: { 10.986162084 0.566889305 0.003544275 -0.404087363 0.511053799, -0.183609944 -1.244355509 0.091889855 -0.166264207 1.514067179}, the Hellinger distance is minimized when $\alpha = 0.345$ and the corresponding MHD estimate is 0.1866 (the true mean is 0). The other estimates are as follows; the mean is 1.1675, the median is .0477, the Huber estimator (with 95% efficiency at normal distribution) is 1.1417.



<Figure 2> AMISE for a bin and a bandwidth; an empirical density estimator, a kernel density estimator and a combined estimator with $\alpha = 0.5$.

<Table 1> Biases of the MHD estimators

No of Obs.	density estimator	distribution			
		$N(0,1)$	$10\%3N$	$10\%10N$	$t(3)$
10	empirical(b_{opt})	-.0260	.0376	.0315	-.0073
	kernel(h_{opt})	-.0240	.0340	-.0521	-.0174
	mixed($\hat{\alpha}$)	-.0237	.0336	-.0411	-.0111
	mean	-.0050	.2900	.9851	-.0057
	median	-.0140	.1178	.1380	.0116
20	empirical(b_{opt})	-.0332	.0387	-.1030	-.0639
	kernel(h_{opt})	-.0457	.0261	-.0462	-.0712
	mixed($\hat{\alpha}$)	-.0394	.0367	-.0453	-.0715
	mean	-.0160	.2816	1.0590	-.0112
	median	-.0140	.1133	.1431	-.0246
50	empirical(b_{opt})	-.0206	.1018	-.0681	-.0242
	kernel(h_{opt})	-.0327	.0874	-.0511	-.0593
	mixed($\hat{\alpha}$)	-.0287	.0916	-.0516	-.0493
	mean	-.0021	.2948	1.0001	-.0012
	median	.0091	.1258	.1420	.0019

3. Further researches and conclusions

We proposed a new type of a density estimator, which is a combination of an empirical and a kernel density estimator, for the MHD estimation. On the condition that we do not know the true underlying distribution in practice, the MHD with newly proposed density estimator is also better recommended than just keep using only a kernel density estimator. Mathematical exposition was carried out but still numerical problems are remained to be more investigated.

References

1. Beran, R. (1977). Minimum Hellinger distance estimations for parametric models, *The Annals of Statistics*, **5**, 445-463.
2. Scott, David W. (1992). *Multivariate Density Estimation: Theory. Practice, and Visualization*, John Wiley & Sons, Inc., New York.
3. Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society* B, **53**, 683-690.
4. Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.
5. Tamura, R. N. and Boos, D. D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance, J*ournal of the American Statistical Association*, **81**, 223-229.

[received date : Dec. 2005, accepted date : Feb. 2006]