Journal of the Korean Statistical Society (1998), 27: 4, pp 507–514

M-Estimation Functions Induced From Minimum L_2 Distance Estimation[†]

Ro Jin Pak[1]

ABSTRACT

The minimum distance estimation based on the L_2 distance between a model density and a density estimator is studied from M-estimation point of view. We will show that how a model density and a density estimator are incorporated in order to create an M-estimation function. This method enables us to create an M-estimating function reflecting the natures of both an assumed model density and a given set of data. Some new types of M-estimation functions for estimating a location and scale parameters are introduced.

Keywords: M-estimation; Minimum distance estimation; Robustness

1. INTRODUCTION

Least squares estimation is the most widely used estimation method in many areas. However, the estimator by this method is very sensitive to extra-ordinary observations. This drawback is due to the fact that the least squares method deals with the square of the original quantity. Another reason is that least squares method does not include any probabilistic structure on which data depend for instance, whether it is skewed or not. In order to cure such problems, some robust estimation techniques have been developed, and one of which is the minimum distance estimation so-called. Minimum distance (MD) estimators are obtained by minimizing a distance, or a metric, between a probabilistic structure induced from data and an assumed probabilistic model. (For a complete list of various minimum distance estimation methods, see Donoho and Liu (1988).) Among the

[†]This research was supported in part by the Basic Science Research Institute Program, Ministry of Education, the Republic of Korea, 1997, BSRI-97-1439

[1]Department of Statistics, Taejon University, Taejon,300-716, Korea.

many types of minimum distance estimation methods, we are interested in those based on two distances:

$$\text{Cramér von Mises} \quad : \quad \mu(F,G) = \int (F(x) - G(x))^2 dG$$

$$\text{Hellinger} \quad : \quad \mu(p,g) = \int (\sqrt{p(x)} - \sqrt{g(x)})^2 dx,$$

where F is an empirical distribution function, G is a distribution function, g is a density function, and p is a density estimator. According to Donoho and Liu's concept of 'Automatic' robustness (1988), the estimators based on both distances are robust. The estimator based on the Hellinger distance is asymptotically efficient (Beran, 1977), and the estimator by the Cramér von Mises distance is consistent but not fully efficient (Par and Schucany, 1980). The minimum Hellinger distance estimators certainly have desirable properties, but it is not easy to deal with the square root terms in the Hellinger distance. On the other hand, an empirical distribution function in the Cramér von Mises distance is rather too primitive to handle data unlike a kernel density estimator. Consider the distance

$$\mu(p,g) = \int (p(x) - g(x))^2 dx, \tag{1.1}$$

which is a hybrid of the Cramér von Mises and Hellinger distances, and call it L_2 distance. We will consider an MD estimator as an statistical quantity minimizing L_2 distance, and call it the minimum L_2 distance estimator (ML2D estimator). In this article, we want to claim that the ML2D estimation method provides us with simple and robust estimators. Robustness and asymptotic properties of ML2D estimators will be investigated from the M-estimator's point of view. Though the ML2D estimator are robust, they lose a little bit of efficiency, just as usual M-estimators do. However, we will show that loss in efficiency can be minimized by adjusting the degree of smoothness of kernel density functions.

2. ML2D ESTIMATION AS M-ESTIMATION

2.1. Derivation of M-estimation function for a location parameter

Let g_θ be a family of probability densities indexed by θ. Based on the distance given in (1.1), an ML2D estimator $\hat{\theta}$ is defined by a statistical quantity minimizing ML2D, which is a solution to

$$\nabla_\theta \mu(p, g_\theta) = \nabla_\theta \int (p(x) - g_\theta(x))^2 dx = 0, \tag{2.1}$$

where we assume $p(x), g_\theta(x) \in L_2$ and ∇_θ represent a derivative with respect to θ. The equation (2.1) can be written as

$$\int (p(x) - g_\theta(x)) \nabla_\theta g_\theta(x) dx = 0. \tag{2.2}$$

Since we have $\int g_\theta(x) \nabla_\theta g_\theta(x) dx = (1/2)\nabla_\theta \int g_\theta^2(x) dx = 0$, if θ is a location parameter, the equation (2.2) becomes

$$\int p(x) \nabla_\theta g_\theta(x) dx = \nabla_\theta \int p(x) g_\theta(x) dx = 0. \tag{2.3}$$

Given a random sample, $X_1, X_2, \cdots, X_n$, having a density of $g_\theta(x)$, let $p(x)$ be a density estimator for $g_\theta(x)$, such as

$$p(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right),$$

where $K(\cdot)$ is a kernel and h is the window width (Silverman, 1986). The equation (2.3) can be written as

$$\sum_{i=1}^{n} \nabla_\theta \int \frac{1}{h} K\left(\frac{x - X_i}{h}\right) g_\theta(x) dx = 0. \tag{2.4}$$

If we follow Huber (1981), and if we denote T_n as an estimate of θ, we have

$$\psi(X_i; T_n) = \nabla_\theta \int \frac{1}{h} K\left(\frac{x - X_i}{h}\right) g_\theta(x) dx \Big|_{\theta = T_n}. \tag{2.5}$$

Example 2.1: Let $X_1, \cdots, X_n$ be independent and identically distributed with a normal density, $N(\theta, \sigma^2)$, where σ^2 is known. A kernel is $K(t) = (1/\sqrt{2\pi}) \exp\{-t^2/2\}$, where h is a window width. We have

$$\begin{aligned}
&\nabla_\theta \int \frac{1}{h} K\left(\frac{x - X_i}{h}\right) g_\theta(x) dx \\
&= \frac{d}{d\theta} \int \frac{1}{\sqrt{2\pi} h} \exp\left\{-\frac{(x - X_i)^2}{2h^2}\right\} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \theta)^2}{2\sigma^2}\right\} dx \\
&= \frac{d}{d\theta} \left[\frac{1}{\sqrt{2\pi}\sqrt{h^2 + \sigma^2}} \exp\left\{-\frac{(X_i - \theta)^2}{2(h^2 + \sigma^2)}\right\}\right].
\end{aligned}$$

After dropping unnecessary constants, we have

$$\psi(X_i - T_n) = (x_i - \theta) \exp\left\{-\frac{(x_i - \theta)^2}{2(h^2 + \sigma^2)}\right\} \Big|_{\theta = T_n}.$$

In Table 2.1, we list ψ-functions according to each kernel function used in estimating μ when the model is $N(\mu, \sigma^2)$. Most of the ψ-functions are similar to the well-known ψ-functions, but the ψ-functions corresponding to Biweight and Gaussian kernels are new types. These ψ-functions in Table 2.1 can be generalized, for example, by replacing $h^2 + \sigma^2$ in the ψ-function for Gaussian kernel by r^2, such as $s \exp\{-s^2/r^2\}$. In a similar manner, ψ-functions for Biweight kernel can be generalized as $\psi(s) = r^2 s + s^3$ or $-r^2 s + s^3$ for $|s| < r$.

2.2. Robustness and Asymptotic Variance

Suppose we are interested in estimating a location parameter of $g_\theta(x) = g(x - \theta)$. Writing $Z_i = (X_i - \theta)/h$, the integral in (2.4) can be expanded as follows when $h \to \infty$:

$$\begin{aligned}
&\int \frac{1}{h} K\left(\frac{x - X_i}{h}\right) g(x-\theta)dx \\
&= \int \frac{1}{h} K\left(\frac{x-\theta}{h} - \frac{X_i - \theta}{h}\right) g(x-\theta)dx \\
&= \int \frac{1}{h} K\left(u - Z_i\right) hg(hu)du \\
&= \int \frac{1}{h} K(Z_i - u) hg(hu)du = \int \frac{1}{h} K(u - Z_i) hg(hu)du \\
&= \int \left[K(Z_i) - uK'(Z_i) + \frac{u^2}{2} K''(Z_i) + \cdots\right] g(hu)du \\
&= \frac{1}{h} K(Z_i) + \frac{1}{2h^3} K''(Z_i) \int v^2 g(v)dv + O\left(\frac{1}{h^5}\right),
\end{aligned}$$

if $g(v)$ and $K(t)$ are symmetric about 0. Therefore, we have

$$\psi(X_i - T_n) = \frac{1}{h} \nabla_\theta K\left(\frac{X_i - \theta}{h}\right) + \frac{1}{2h^3} \nabla_\theta K''\left(\frac{X_i - \theta}{h}\right) \int v^2 g(v)dv + O\left(\frac{1}{h^5}\right)\Bigg|_{\theta = T_n}$$

In terms of the definition of the influence function

$$IF_{\theta_\psi} = \frac{\psi(X - \theta_\psi)}{E_F[\psi'(X - \theta_\psi)]},$$

we can claim that the shape of an influence function of an ML2D estimator is determined by the shape of the kernel function, and the sup-norm of an influence function, called "gross-error sensitivity", is somehow determined by the moments

of the model density. If the derivatives of a kernel function are bounded and a model density has finite moments, a ψ-function induced from ML2D estimation is the so-called B-robust at F (Hampel, et al., 1986), or just robust. Robustness of an ML2D estimator can be interpreted in terms of robustness of an M-estimator, and certainly helps us to understand the characteristics of an ML2D estimator.

Since we have realized that the ML2D estimator can be considered as an usual M-estimator, they have the same asymptotic properties as the M-estimator. For example, suppose we have

$$\psi(s) = s \exp\left\{-\frac{s^2}{2(h^2+\sigma^2)}\right\},$$

which is the M-estimating function when a model is the normal density and a kernel is Gaussian. The asymptotic variance turns out

$$V(\psi_\mu, \Phi) = \frac{\int \psi^2 dF}{(\int \psi' dF)^2} = \left(\frac{2\sigma^2+h^2}{\sigma^2+h^2}\right)^3 \left(\frac{\sigma^2+h^2}{3\sigma^2+h^2}\right)^{3/2} \sigma^2,$$

where h is a window width. The asymptotic variance is controlled by the value of h, the smoothness of a kernel. It is always slightly greater than σ^2 and converges to σ^2 as h increases Of course, when we are trying to estimate μ if a model is the normal with mean μ and variance σ^2, the least possible value of the asymptotic variance becomes σ^2. The larger values of h for a density estimator are practically meaningless, so that the M-estimator based on that $\psi(s)$ is not fully asymptotically efficient.

2.3. Additional Comments

In section 2.1, we assume $\int g_\theta(x) \nabla_\theta g_\theta(x) dx = 0$, which is true for a location case, but that equality does not always hold for any parameter θ. Thus, instead of (2.5) we should have

$$\psi(X_i;\theta) = \nabla_\theta \int \frac{1}{h} K\left(\frac{x-X_i}{h}\right) g_\theta(x) dx - \nabla_\theta \int g_\theta^2(x) dx. \tag{2.6}$$

However, the very last term in (2.6) will be adjusted in order for ψ to satisfy the condition, as $\int \psi(x) dG(x) = 0$ if $G(x)$ is a model distribution function. Therefore, without loss of generality, we can let

$$\psi(X_i;\theta) = \nabla_\theta \int \frac{1}{h} K\left(\frac{x-X_i}{h}\right) g_\theta(x) dx.$$

Table 2.1: ψ-functions for estimating a location parameter when a model density is $N(\mu, \sigma^2)$, σ^2 is known.

Kernel	$K(t)$	$\psi(s)$	Shape
Epanechnikov	$\frac{3}{4}(1-\frac{1}{5}t^2)/\sqrt{5}$ for $\lvert t\rvert < \sqrt{5}$, 0, otherwise	s for $\lvert s\rvert < h\sqrt{5}$, 0, otherwise	Huber
Biweight	$\frac{15}{16}(1-t^2)^2$ for $\lvert t\rvert < 1$,	$h^2\sigma^2(3h^{-2}\sigma^2-1)s + s^3$ for $\lvert s\rvert < h$, 0, otherwise	Tukey's biweight
Triangular	$1-\lvert t\rvert$ for $\lvert t\rvert < 1$; 0, otherwise	$\text{sign}(s)1_{[-h,h]}(s)$	Skipped median
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}t^2\}$	$s\exp\{-\frac{1}{2(\sigma^2+h^2)}s^2\}$	New type
Rectangular	$\frac{1}{2}$ for $\lvert t\rvert < 1$; 0, otherwise	0 for all t	Not useful

Table 2.2: χ-functions for estimating a scale parameter when a model density is $N(\mu, \sigma^2)$, μ is known.

Kernel	$K(t)$	$\psi(s)$	Shape
Epanechnikov	$\frac{3}{4}(1-\frac{1}{5}t^2)/\sqrt{5}$ for $\lvert t\rvert < \sqrt{5}$, 0, otherwise	s^2 for $\lvert s\rvert < h\sqrt{5}$, 0, otherwise	Huber
Biweight	$\frac{15}{16}(1-t^2)^2$ for $\lvert t\rvert < 1$,	$h^2\sigma^2(3h^{-2}\sigma^2-1)s^3+s^4$ for $\lvert s\rvert < h$, 0, otherwise	New type
Triangular	$1-\lvert t\rvert$ for $\lvert t\rvert < 1$; 0, otherwise	$\lvert s\rvert$ for $\lvert s\rvert < h$	New type
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}t^2\}$	$(-1+\frac{s^2}{\sigma^2+h^2})\exp\{-\frac{1}{2(\sigma^2+h^2)}s^2\}$	New type
Rectangular	$\frac{1}{2}$ for $\lvert t\rvert < 1$; 0, otherwise	0 for all t	Not useful

Example 2.2: Estimating σ when the model is $N(\mu, \sigma^2)$ and the kernel is Gaussian. We first have

$$\nabla_\sigma \int K\left(\frac{x - X_i}{h}\right) g_\sigma\left(\frac{x}{\sigma}\right) dx = \frac{1}{\sqrt{2\pi}(h^2+\sigma^2)^{3/2}}\left(-1 + \frac{(X_i-\mu)^2}{h^2+\sigma^2}\right) \exp\left\{-\frac{(X_i-\mu)^2}{2(h^2+\sigma^2)}\right\}.$$

After dropping unnecessary terms and adjusting ψ in order to make $\int \psi(x)dG(x) = 0$, redefine ψ as

$$\chi\left(\frac{X_i}{T(n)}\right) = \left(-1 + \frac{(X_i-\mu)^2}{h^2+\sigma^2}\right) \exp\left\{-\frac{(X_i-\mu)^2}{2(h^2+\sigma^2)}\right\} + \left(\frac{h^2+\sigma^2}{h^2+2\sigma^2}\right)^{3/2}\Bigg|_{\sigma=T_n}.$$

In Table 2.2, we list χ-functions according to kernels when the model is $N(\mu, \sigma)$, where μ is known. Of course, these functions should be adjusted to satisfy the condition that the expectation of each function should be zero. Most of them in Table 2.2 are newly proposed M-estimation functions for scale parameters.

References

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

Donoho, D. L. and Liu, R. C. (1988). "The 'Automatic' robustness of minimum distance functionals," *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* New York: John Wiley & Sons.

Huber, P. J. (1981). *Robust Statistics* New York: John Wiley & Sons.

Par, W. C. and Schucany, W. R. (1980). "Minimum distance and robust estimation," *Journal of the American Statistical Association.* **75**, 616-624.

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