The Korean Communications
in Statistics Vol. 9, No. 2, 2002
pp. 325-335

Adaptive M-estimation using Selector Statistics in Location Model [1)]

Sang Moon Han[2)]

Abstract

In this paper we introduce some adaptive M-estimators using selector statistics to estimate the center of symmetric and continuous underlying distributions. This selector statistics is based on the idea of Hogg(1983) and Hogg et. al. (1988) who used averages of some order statistics to discriminate underlying distributions. In this paper, we use the functions of sample quantiles as selector statistics and determine the suitable quantile points based on maximizing the distance index to discriminate distributions under consideration. In Monte Carlo study, this robust estimation method works pretty good in wide range of underlying distributions.

Keywords : robust, selector statistics, M-estimators

1. 서론

대칭인 오차분포군의 중심에 대한 추정법에서 정규가설을 이용한 표본평균의 사용은 많은 문제점을 가지고 있다는 사실은 많은 통계학자들이 오래전부터 인지하여왔다. 이 문제에 대한 해결책은 Huber(1972,1981)의 M-추정법에 의해 많이 해결되었다. Huber의 방법과 다른 각도로 Hogg(1967)는 기저 오차분포에 대한 형태를 선택통계량에 의해 파악하고, 이에 알맞은 추정량을 배분하는 적응 추정법을 제시하였다. 예컨대 선택통계량에 의해 오차분포가 정규분포 보다 이중지수분포에 가깝다는 사실을 알고 있다면 표본평균보다 중앙값을 중심에 대한 추정량으로 사용하는 것이 효율적일 것이다. 이러한 아이디어를 사용한 추정량들이 Andrews등의 프린스턴 로버스트 추정법 연구(1972)에서 가치를 인정 받았다. 그러나 그가 사용한 선택통계량은 표본 첨도(sample kurtosis)를 이용한 것으로 점근적 수렴속도가 느린 단점을 지니고 있어 이후 Hogg,Fisher ,Randles(1975). Hogg(1983), Hogg등(1988)에 의해 순서 통계량들의 평균들의 함수형태의 선택통계량들을 이용하여 이러한 단점을 보완하여 왔다.

논문에서는 Hogg(1983)가 제안한 선택통계량들을 이용하고 Hogg 등(1988)에 의해 제안된 두 분포의 거리를 최대로 하는 선택통계량의 구성방법을, 백분위수(quantile)의 함수들로 구성된 선택

1) This work was supported by the research fund of University of Seoul in the year of 2001

2) Professor, Department of Computer Science and Statistics, University of Seoul, 90 Cheonnong-dong, Dongdaemun-ku,130-743, Seoul, Korea
E-mail : smhan@uoscc.uos.ac.kr

통계량에 적용하여 새로운 선택통계량에 의한 추정량들을 제시하려한다. 먼저 Hogg(1983)가 제안한 선택통계량은 다음과 같다.

$\alpha+\beta+\gamma=0.5$ 라고 하자. 단 여기서 α, β 그리고 γ 는 음이 아닌 실수이다. 그리고 L_α , B_β , C_γ , D_γ, , E_β 그리고 U_α 는 각각 가장 작은 $n\alpha$ 순서통계량들의 평균, 다음으로 작은 $n\beta$ 순서통계량의 평균, 다음으로 작은 $n\gamma$ 순서통계량의 평균, 다음으로 작은 $n\gamma$ 순서통계량의 평균, 다음으로 작은 $n\beta$ 순서통계량의 평균이며 U_α 는 가장 큰 $n\alpha$ 순서통계량들의 평균이라 하자. 만약 $n\alpha$, $n\beta$ 그리고 $n\gamma$ 가 정수가 아니라면, 이러한 평균을 부분 관측치를 사용하여 계산한다. 그리고 이중지수분포(DE)와 같이 뾰족한 형태의 분포와 정규분포(NOR)와 같이 덜 뾰족한 분포를 구분하기 위한 선택통계량으로

$$HH_2 = \frac{E_\beta - B_\beta}{D_\gamma - C_\gamma}$$

를 사용하였다. 이것은 우리들의 상식에도 일치하는 것이다. 왜냐하면 오차분포가 뾰족할수록 $D_\gamma - C_\gamma$ 의 값은 작아지는 경향이 있을 것이고 따라서 HH_2의 값이 커지면 오차분포가 뾰족한 경향이 있을 것이기 때문이다. 그리고 덧붙여서 코우쉬분포(CA)와 같이 꼬리부분이 무거운 분포와 정규분포나 이중지수분포처럼 비교적 가벼운 꼬리부분을 가진 분포를 구별하기 위해

$$HH_3 = \frac{U_\alpha - L_\alpha}{E_\beta - B_\beta}$$

를 제안하였다. 그리고 이 또한 $U_\alpha - L_\alpha$ 가 오차분포가 무거운 꼬리를 가지로 있는 경우에는 커지는 경향이 있으므로 HH_3의 값은 커지면 오차분포가 무거운 꼬리를 가질 것이라고 판단되기 때문이다. 그리고 Hogg는 그의 논문(1983)에서 $\alpha=.05, \beta=.15, \gamma=.30$인 값을 제안하였다. 본 논문에서는 Hogg의 아이디어를 변형하여 HH_2 와 HH_3 통계량을 표본 백분위수의 함수로 표시하고 여기서 표본 백분위수의 함수로 표시된 새로운 선택통계량을 H_2 와 H_3라고 할 때의 이들의 점근적 성질을 규명하고, $H=H_2+H_3$ 통계량을 선택통계량으로 사용하여 NOR과 DE, 그리고 DE와 CA를 동시에 잘 구분시키는 α, β, γ의 값을 점근적으로 찾아 최종적인 선택통계량으로 사용하여 이 논문에서 제안할 적응 M-추정량에 사용할 것이다.

2. 선택통계량의 제안

이 절에서 이용될 정리를 간단히 소개하고 이것을 이용하여 제안하게 될 선택통계량들의 점근

적 성질을 규명하고자 한다.

정리 2.1 $0<p_1<p_2<\cdots<p_k<1$ 에 대해 $\hat{\xi}_{p_1}$, $\hat{\xi}_{p_2}$, $\cdots$, $\hat{\xi}_{p_k}$ 를 대응하는 표본 제 $100p_j$분위수, ξ_{p_1}, $\xi_{p_2}\cdots,\xi_{p_k}$를 대응하는 기저분포의 제 $100p_j$분위수($j=1,2,\cdots,,k$)라고 하자. 이때 $\sqrt{n}(\hat{\xi}_{p_1}-\xi_{p_1}),\sqrt{n}(\hat{\xi}_{p_2}-\xi_{p_2}),\cdots,\sqrt{n}(\hat{\xi}_{p_k}-\xi_{p_k})$ 의 결합분포는 점근적으로 평균 $\mathbf{0}$이고 분산-공분산행렬 $\Sigma=(\sigma_{ij})_{k\times k}$ 를 가지는 점근적 k 차원 정규분포를 따른다.

단, $\sigma_{ij}=p_i(1-p_j)/(f(\xi_{p_i})f(\xi_{p_j}))$이다. 단 $\hat{\xi}_p$ 는 $[np]+1$ 번째 순서통계량이다.

증명: David(1981) 255쪽 참조.

이 논문에서 제안하는 통계량은 Hogg(1983)가 제안한 순서통계량의 일부 표본들의 평균을 이용한 선택통계량을 표본 백분위수를 이용하여 다음과 같이 제안한다.

$$H_2=\frac{\hat{\xi}_{1-\beta}-\hat{\xi}_{\beta}}{\hat{\xi}_{1-\gamma}-\hat{\xi}_{\gamma}},\ H_3=\frac{\hat{\xi}_{1-\alpha}-\hat{\xi}_{\alpha}}{\hat{\xi}_{1-\beta}-\hat{\xi}_{\beta}},\quad H=H_2+H_3 \tag{2.1}$$

단, 여기서 $\alpha<\beta<\gamma<.5$ 이고 $\hat{\xi}_{\alpha},\hat{\xi}_{\beta},\hat{\xi}_{\gamma}$ 는 각각 표본의 $100\alpha,100\beta,100\gamma$ 백분위수이다. 이때, 오차분포의 밀도함수가 연속이며 대칭인 f(x)의 형태를 가질 때 제안된 선택통계량인 H_2 와 H_3 는 점근적인 분포는 다음과 같다.

정리 2.2 H_2 와 H_3가 (2.1)과 같이 정의되고, $\xi_{\alpha},\xi_{\beta},\xi_{\gamma}$ 가 각각 오차분포의 100α, 100β, 100γ 백분위수일 때, $n\to\infty$일 때

$$\sqrt{n}(H_2-\mu_{H_2})\to N(0,\ \boldsymbol{a}'A\,\boldsymbol{a}) \tag{2.2}$$

$$\sqrt{n}(H_3-\mu_{H_3})\to N(0,\ \boldsymbol{b}'B\,\boldsymbol{b}) \tag{2.3}$$

이다.

단, $\mu_{H_2}=\frac{\xi_{1-\beta}-\xi_{\beta}}{\xi_{1-\gamma}-\xi_{\gamma}}$, $\mu_{H_3}=\frac{\xi_{1-\alpha}-\xi_{\alpha}}{\xi_{1-\beta}-\xi_{\beta}}$ 이고 $\boldsymbol{a}$ 벡터는 H_2 를 각각 $\hat{\xi}_{1-\beta},\hat{\xi}_{\beta},\hat{\xi}_{1-\gamma},\hat{\xi}_{\gamma}$ 에 대해 편미분하여 $\xi_{1-\beta},\xi_{\beta},\xi_{1-\gamma},\xi_{\gamma}$ 에서 계산한 열 벡터이고,

A 는 $\hat{\xi}_{1-\beta},\hat{\xi}_{\beta},\hat{\xi}_{1-\gamma},\hat{\xi}_{\gamma}$ 인 4개의 표본백분위수가 구성하는 분산-공분산 행렬이다. 마찬가지로, $\boldsymbol{b}$ 와 B 도 정의된다.

증명: 먼저 (2.2)식을 증명하기 위해 $\hat{\xi}_{1-\beta},\hat{\xi}_{\beta},\hat{\xi}_{1-\gamma},\hat{\xi}_{\gamma}$ 를 각각 X_1,X_2,X_3,X_4라고 하고

$\xi_{1-\beta}, \xi_\beta, \xi_{1-\gamma}, \xi_\gamma$를 각각 $\mu_1, \mu_2, \mu_3, \mu_4$하자. 그리고 $H_2(\boldsymbol{X}) = H_2(X_1, X_2, X_3, X_4)$ 라 놓으면 $H_2(\boldsymbol{X}) = (X_1 - X_2)/(X_3 - X_4)$이고, 이 식을 $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)'$ 근방에서 다변량 Taylor전개하면,

$$H_2(\boldsymbol{X}) - H_2(\boldsymbol{\mu}) = \sum_{i=1}^{4}(X_i - \mu_i)\partial H_2/\partial X_i \mid_{X=\mu} + \frac{1}{2!}\left\{\sum_{i=1}^{4}(X_i - \mu_i)^2 \partial^2 H_2/\partial X_i^2 \mid_{X=\mu} + 2\sum_{i<}^{3}\sum_{j}^{4}(X_i - \mu_i)(X_j - \mu_j)\partial^2 H_2/\partial X_i \partial X_j \mid_{X=\mu}\right\} + \cdots$$

그런데 $i = 1, 2, 3, 4$에 대해 $\sqrt{n}(X_i - \mu_i) \to N(0, \xi_i(1-\xi_i))$ 이고,

$\sqrt{n}(X_i - \mu_i)^2 = \sqrt{n}(X_i - \mu_i)(X_i - \mu_i)$에서 $X_i - \mu_i \xrightarrow{p} 0$ 이므로,

$\sqrt{n}(X_i - \mu_i)^2 \xrightarrow{p} 0$이다. 마찬가지로, $\sqrt{n}(X_i - \mu_i)(X_i - \mu_i) \xrightarrow{p} 0$. 그리고 3차항 이상의 항에 대해서도 모두 $\xrightarrow{p} 0$이다.

따라서, $\sqrt{n}(H_2(\boldsymbol{X}) - H_2(\boldsymbol{\mu}))$은 점근적으로 $\sqrt{n}\sum_{i=1}^{4}(X_i - \mu_i)\partial H_2/\partial X_i \mid_{X=\mu}$와 동일한 분포를 따른다. 즉

$$\sqrt{n}\sum_{i=1}^{4}(X_i - \mu_i)\partial H_2/\partial X_i \mid_{X=\mu}$$

는 정리2.1에 의해 점근적으로 평균이 0이고 분산이 $\boldsymbol{a}'\boldsymbol{A}\boldsymbol{a}$을 가지므로, (2.2)식은 증명되었다. 마찬가지 방법으로 (2.3)식도 증명된다.

다음으로 본 논문에서 최종적인 선택통계량으로 사용될 $H = H_2 + H_3$ 통계량의 점근적 분포는 다음과 같다.

따름정리 2.3 $H = H_2 + H_3$ 이고 $\xi_\alpha, \xi_\beta, \xi_\gamma$가 각각 오차분포의 $100\alpha, 100\beta, 100\gamma$ 백분위수일 때, $n \to \infty$이면

$$\sqrt{n}(H - \mu_H) \to N(0, \boldsymbol{c}' C \boldsymbol{c}) \ , \tag{2.4}$$

이다.

단, 여기서 $\mu_H = \frac{\xi_{1-\beta} - \xi_\beta}{\xi_{1-\gamma} - \xi_\gamma} + \frac{\xi_{1-\alpha} - \xi_\alpha}{\xi_{1-\beta} - \xi_\beta}$ 이고 $\boldsymbol{c}$벡터는 통계량 H를 각각 $\widehat{\xi_{1-\alpha}}, \widehat{\xi_\alpha}, \widehat{\xi_{1-\beta}}, \widehat{\xi_\beta}, \widehat{\xi_{1-\gamma}}, \widehat{\xi_\gamma}$ 에 대해 편미분하여 $\xi_{1-\alpha}, \xi_\alpha, \xi_{1-\beta}, \xi_\beta, \xi_{1-\gamma}, \xi_\gamma$ 에서 계산한 열 벡터이고, C는 $\widehat{\xi_{1-\alpha}}, \widehat{\xi_\alpha}, \widehat{\xi_{1-\beta}}, \widehat{\xi_\beta}, \widehat{\xi_{1-\gamma}}, \widehat{\xi_\gamma}$ 인 6개의 표본 백분위수가 구성하는 분산-공분산 행렬이다.

증명 정리 2.2 의 증명과 마찬가지로, $\widehat{\xi_{1-\alpha}}, \widehat{\xi_{\alpha}}, \widehat{\xi_{1-\beta}}, \widehat{\xi_{\beta}}, \widehat{\xi_{1-\gamma}}, \widehat{\xi_{\gamma}}$ 를 각각 $X_1, X_2, X_3, X_4, X_5, X_6$ 라고 하고 $\xi_{1-\alpha}, \xi_{\alpha}, \xi_{1-\beta}, \xi_{\beta}, \xi_{1-\gamma}, \xi_{\gamma}$ 를 각각 $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6$ 라고 하자. 그리고 $H(\boldsymbol{X})=H(X_1, X_2, X_3, X_4, X_5, X_6)$ 라 놓고 $H(\boldsymbol{X})$ 를 $\boldsymbol{\mu}=(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6)'$ 근방에서 다변량 Taylor 전개하면, 정리2.1의 증명과 동일한 방법으로 증명된다.

본 절에서는 선택통계량 H 가 오차분포들을 잘 구분(discrimate)하여 주도록 α, β 와 γ 의 값을 결정해 주고자 한다. 많은 분포에 대해 모두 고려 할 수는 없으므로, 가벼운 꼬리를 가진 NOR 과 뾰족한 중심을 가진 DE, 그리고 극단적으로 무거운 꼬리를 가진 CA분포에 대해 이 세 개의 분포를 동시에 잘 구분하는 α, β 와 γ 의 값을 정하고자 한다. 이를 위해 $\sqrt{n}(H-\mu_H)$ 가 각각의 오차분포NOR, DE, CA에서 서로 다른 점근적 정규분포를 따르기 때문에 각각의 평균과 분산을 각각 $(0, \sigma^2_{NOR}), (0, \sigma^2_{DE}), (0, \sigma^2_{CA})$ 라고 하자. 이 점근적 분포들은 이것들이 많이 떨어져 있을수록 선택통계량 H 에 의해 잘 구분이 될 것이다. 여기서 두 개의 분포의 거리를 다음과 같이 정의하자. 예컨대 오차분포가 NOR과 DE 일 때, 이것들의 거리를 다음과 같이 정의하자.

$$nD^2=(\mu_{NOR}-\mu_{DE})^2(\frac{1}{2}(\sigma^2_{NOR}+\sigma^2_{DE}))^{-1} \tag{2.5}$$

식 (2.4)는 동일한 분산-공분산행렬 $\boldsymbol{\Sigma}$ 를 가지는 두 개의 다변량 분포 $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ 사이의 Mahalanobs 거리가 $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$로 정의되고(1981, Mardia등), 이것의 변형으로 분산-공분산 행렬이 Σ_1 과 Σ_2 로 같지 않을 때. Nakanish와 Sato(1985) 및 Hogg 등(1988)은 다음과 같이 두 개의 다변량 분포의 거리를 정의하였다.

$$nD^2=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'(\frac{1}{2}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2))^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \tag{2.6}$$

(2.5)식은 (2.6)식의 두 개의 분포에 대한 점근적 거리의 1차원적인 표현이다. (2.5)식을 이용하여 선택통계량 H 에 의한 (NOR,DE) 와 (DE,CA) 사이의 점근적 거리를 각각 d_1 , d_2 이라 하고 $0<\alpha<\beta<\gamma<0.5$에 대해 $\alpha=0.01$, $\beta=0.02$, 그리고 $\gamma=0.03$ 으로부터 시작하여 0.01씩 값을 증가시켜 각 분포들간의 거리를 <표 2.1> 과 같이 구하였다.

<표 2.1>에서 보는 바와 같이 d_1 , d_2 의 값을 동시에 최대가 되게 만드는 α, β, γ 의 값이 없다. 따라서 d_1+d_2 의 값을 최대로 하는 $\alpha=0.01$, $\beta=0.04$, $\gamma=0.22$ 의 값을 최종적인 값으로 하여 선택통계량을 결정하였다. 즉

$$H_2 = \frac{\widehat{\xi_{0.96}} - \widehat{\xi_{0.04}}}{\widehat{\xi_{0.78}} - \widehat{\xi_{0.22}}}, H_3 = \frac{\widehat{\xi_{0.99}} - \widehat{\xi_{0.01}}}{\widehat{\xi_{0.96}} - \widehat{\xi_{0.04}}}, \quad H = H_2 + H_3 \tag{2.7}$$

<표 2.1> 선택통계량 H 를 사용한 두분포들간의 점근적 거리

α 값	$\alpha=0.01$						$\alpha=0.02$					
β 값	$\beta=0.04$		$\beta=0.05$		$\beta=0.06$		$\beta=0.04$		$\beta=0.05$		$\beta=0.06$	
점근 거리	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
$\gamma=0.18$	1111	587	1156	456	1219	375	807	706	801	758	817	697
$\gamma=0.20$	1104	629	1138	496	1197	403	831	701	817	775	827	738
$\gamma=0.22$	1086	663	1108	535	1157	431	845	691	822	778	825	766
$\gamma=0.24$	1056	688	1065	572	1103	460	846	675	817	761	812	781
$\gamma=0.26$	1016	703	1014	605	1038	490	837	656	801	750	789	781
$\gamma=0.28$	966	708	953	633	966	519	817	633	775	723	756	766

3. 추정량의 제안

본 논문에서 제안할 추정량은 Huber와 Tukey의 M-추정량의 조절상수(tuning constant)를 선택통계량 H 의 값에 따라 결정하는 적응 M-추정량이다. 먼저, Huber와 Tukey의 M-추정량에 대해 간단히 언급하면 다음과 같다. $\hat{\theta}$ 을 중심에 대한 예비추정량(일반적으로 표본 중앙값)이라고 하고 MAD를 $MAD=median_i\{|X_i - med_j(X_i)|\}$로 정의하면, 조절상수 k를 가진 Huber의 M-추정량 $Hub(k)$와 Tukey의 M-추정량 $Tuk(k)$는 각각

$$Hub(k) = \sum_{i=1}^{n} w_i X_i / \sum_{i=1}^{n} w_i,$$

단 $w_i = 1$, $|X_i - \hat{\theta}| \le kMAD/.6745$ 인 경우

$= .6745kMAD/|X_i - \hat{\theta}|$, 다른 경우

$$Tuk(k) = \sum_{i=1}^{n} w_i X_i / \sum_{i=1}^{n} w_i,$$

단 $w_i = \{1 - [(X_i - \hat{\theta})/kMAD]^2\}^2$, $|X_i - \hat{\theta}| \le kMAD$ 인 경우

$= 0$, 다른 경우

와 같이 정의된다. 그리고 기저 오차분포의 형태가 f(x) 일 때, M-추정량의 점근적 분산은 $E(\psi)^2 / (E(\psi'))^2$ 와 같은 형태를 가진다는 사실은 잘 알려져 있다.(Huber(1981)). 단,

$\rho = d[-\log f(x)]/dx$ 이다. <표 3.1>과 <표 3.2>는 위의 공식을 사용하여 조절상수의 값에 따른 점근적 분산을 각각 구한 것이다. 상기 표에서 보는 바와 같이 NOR처럼 가벼운 꼬리를 가진 분포에서는 Huber와 Tukey의 M-추정량 모두 조절상수의 값이 커질수록 점근적 분산이 작아지는 경향이 있고, DE나 CA처럼 뾰족한 중심부분을 가진 분포나 극단적으로 무거운 꼬리를 가진 분포에서는 조절 상수의 값이 작아질수록 점근적 분산이 작아지는 경향이 있다.

그리고 확률 0.9로 표준정규분포를 따르고, 확률 0.1로 평균이 0, 분산이 9인 분포를 따르는 오염분포(CON), 로지스틱분포(LOG), 자유도 3인 t-분포(T(3))처럼 중간정도의 두께를 가진 분포에서는 일정한 규칙이 없다는 사실을 알 수 있다. 그리고 (2.6)에서 제안된 선택통계량 H 를 사용한 추정량을 제안하기위해 n=40개의 표본을 사용하여 4000번의 반복에 의한 모의실험에서 H 값의 변화에 따른 빈도수를 가벼운 꼬리를 가진 오차분포(NOR), 중간정도의 꼬리를 가진 오차분포(LOG,CON등),뾰족한 가운데 부분을 가진 오차분포(DE) 및 무거운 꼬리를 가진 분포(CA)에 대해 모의실험하였다. <표 3.3>에서 보는 바와 같이 NOR나 LOG처럼 가벼운 꼬리나 중간정도의 두께를 가진 오차분포 하에서는 선택 통계량 H 의 값이 작아지는 경향이 있고, DE나 CA처럼 뾰족한 중심부분을 가진 분포나 극단적으로 무거운 꼬리를 가진 분포에서는 H 의 값이 작아지는 경향이 있다. 이와 같은 모의실험의 결과 본 논문에서 제안하는 두 가지 형태의 적응 M-추정량은 다음과 같다.

$$AH = \begin{cases} Hub(1.5), & H \le 4.0 \text{ 경우} \\ Hub(0.8), & 4.0 < H < 5.0 \text{ 경우} \\ Hub(0.5), & H > 5.0 \text{ 경우} \end{cases} \tag{3.1}$$

$$AT = \begin{cases} Tuk(7.5), & \text{if } H \le 4.0 \text{ 경우} \\ Tuk(6.0), & \text{if } 4.0 < H < 5.0 \text{ 경우} \\ Tuk(4.5), & \text{if } H \ge 5.0 \text{ 경우} \end{cases} \tag{3.2}$$

여기서 AH와 AT는 각각 Huber와 Tukey의 M-추정량을 이용한 적응 M-추정량이다.

<표 3.1> Huber M-추정량의 점근적 분산

	Hub(.5)	Hub(1.0)	Hub(1.5)	Hub(2.0)
NOR	1.2625	1.1073	1.0371	1.0140
CON	1.4789	1.3330	1.2959	1.3237
LOG	3.4818	3.1946	3.2090	3.0177
T(3)	1.5695	1.5207	1.5824	1.6862
DE	1.1652	1.3226	1.4653	1.5888
CA	2.1553	2.5465	2.9927	3.5208

<표 3.2> Tukey M-추정량의 점근적 분산

	Tuk(3)	Tuk(6)	Tuk(9)	Tuk(12)
NOR	1.2930	1.0160	1.0040	1.0013
CON	1.5156	1.2780	1.3760	1.4802
LOG	5.5375	3.1815	3.0296	3.0542
T(3)	1.7098	1.5904	1.7489	1.8799
DE	1.4037	1.4946	1.6383	1.7531
CA	2.2245	2.5891	3.2365	3.9847

<표 3.3> H 값의 변화에 따른 빈도수

H값 / 분포	4.0<	4.0-4.5	4.5-5.0	5.0-5.5	5.5-6.0	6.0-6.5	6.5-7.0	7.0-7.5	7.5-8.0	8.0-8.5	8.5>
NOR	2884	793	247	63	8	4	1	0	0	0	0
LOG	1747	1208	649	257	85	37	8	8	0	1	0
D.E.	231	367	490	540	505	431	335	251	194	150	506
CA	5	13	39	62	102	124	148	157	172	165	3013

4. 모의실험

분포가 7개이며 나머지 2개는 로지스틱분포(LOGISTICS)와 자유도 3인 t-분포(T(3))를 사용하였다. 그리고 $S=Y/Z$ 형태의 확률변수에서 Y, Z 는 서로 독립이고 Y 는 표준정규분포의 확률변수이고, Z 는 다음과 같다.

(1) Normal(NOR) : $Z=1$

(2) Slate (TE) : $Z=U^{\frac{1}{10}}$, 여기서 U 는 표준균일분포를 따르는 확률변수.

(3) Slacu (CU) : $Z=U^{\frac{1}{3}}$

(4) Slash (SH) : $Z=U$

(5) Coutaminated (Con) : $Z=\begin{cases}1 & , \text{확률 } 0.9\\ 1/3 & , \text{확률 } 0.1\end{cases}$

(6) Double Exponential (DE) : $Z=1/\sqrt{W}$, 여기서 W는 자유도 2인 카이제곱 확률변수.

(7) Cauchy (CA) : $Z=|V|$, 여기서 V는 표준정규분포를 따르는 확률변수.

그리고 본 논문에서 사용되는 추정량은 모두 7개로 구성되어 있는데 표본평균(Mean), 표본중앙

값(Med), Huber의 Hub(1.25), Tukey의 Tuk(4.82), p=1.277, x_1=1.344, r=4를 적용한 Huber-Collins의 M-추정량 HC (Hampel 등(1986) 참조)과 위의 (3.1)과 (3.2)에서 정의한 선택통계량을 사용한 두 개의 적응 M-추정량 AH와 AT로 되어있다. 모의실험의 반복횟수는 4000번이고, 평균제곱오차(MSE)* 10^4의 크기로 <표 4.1>에 결과가 주어져있다. 그리고 모의실험에서 사용된 scale값은 MAD/.6745를 사용하였다. 위의 모의실험 결과를 요약하면 다음과 같다.

(1) 적응 M-추정량인 AH와 AT는 광범위한 분포군에 대해 좋은 효율을 가짐을 알 수 있다. 특히 AT는 비 적응 M-추정량인 Tuk(4.82)에 비해 본 모의실험에서 제시된 모든 분포군에 대해 좋은 효율을 가짐을 알 수 있다. 그러나 AH는 비적응 추정량인 Hub(1.25)에 비해 NOR, TE등 가벼운 꼬리를 가진 분포나 SH,CA등 극단적로 무거운 꼬리를 가진 분포에 대해서는 우월하나, LOG,CU등 중간정도의 꼬리를 가진 분포에 대해서는 효율이 약간 떨어짐을 알 수 있다.

(2) 비적응 추정량인 Hub(1.25),Tuk(4.82), HC중에 가벼운 꼬리 혹은 중간정도의 꼬리를 가진 분포의 위치모수추정에서는 Hub(1.25)가 가장 좋은 추정량이고, 모든 분포군에 대한 무난한 추정량으로는 Huber-Collins 추정량 HC임을 알수 있다.

(3) AT는 Hub(1.25)에 비해 중간정도의 꼬리를 가진 분포군의 위치모수추정에는 약간 효율이 떨어지나, 가벼운 꼬리를 가진 분포군에서는 약간 효율이 우월하고, 특히 무거운 꼬리를 가진 분포군에서는 월등히 효율이 좋았다.

(4) 잘 알려진 사실이지만, 표본평균(Mean)은 다른 모든 M-추정량에 비해 가벼운 꼬리를 가진 분포군의 위치모수 추정에는 나은 효율을 가지나, 다른 경우에는 못한 효율을 가지며, 특히 무거운 꼬리를 가진 분포군의 위치모수 추정에는 아주 못한 효율을 가진다. 그리고 표본중앙값(Med)은 DE처럼 극단적으로 뾰족한 중심을 가진 분포나, SH나 CA처럼 극단적으로 무거운 꼬리를 가지는 분포 외에는 다른 모든 M-추정량에 비해 못한 효율을 가진다. 결론적으로 위의 모의실험 결과 선택통계량을 이용한 M-추정법은 기존의 M-추정법에 비해 분명히 우월하다고는 할 수 없지만, 하나의 대안으로 사용될 수 있는 추정법임을 확인 할 수 있다. 앞으로 연구가 더

이루어져야 하는 부분은 선택통계량에 대한 좀더 정교한 연구가 이루어져야 하겠고, 선택통계량을 이용한 비대칭인 오차분포에 대한 위치모수 추정법의 연구가 이루어져야 하겠다.

< 표 4.1> 추정량들의 평균제곱오차

분포 / 추정량	NOR	TE	LOG	CU	CON	T(3)	DE	SH	CA
Mean	2408*	3020*	8249	7221	4365	7384	4919	-	-
Med	3662	4432	9653	6563	4265	4409	3190*	15598	6793
Hub(1.25)	2629	3217	7586*	5062*	3186*	3795	3289	16473	8836
Tuk(4.82)	3066	3726	8412	5538	3480	3893	3224	12986	6174
HC	2743	3352	7849	5137	3203	3787*	3301	13462	6931
AH	2625	3216	7625	5117	3327	3802	3304	14779	7712
AT	2612	3205	7770	5158	3229	3840	3315	12956*	6101*

1) MSE값은 실제 얻어진 MSE값을 10000배 한 것이다.
2) 표본평균의 MSE가 SH와 CA의 경우 너무 큰 값이 나오므로 - 로 표시하였다.
3) 제안된 추정량들 중에 MSE값이 가장 작은 추정량을 *로 표시하였다

References

[1] Andrews, D.F. , Bickel, P.J. , Hampel, F.R. , Huber P.J. , Roger W.H. , Tukey, J.W. (1972). *Robust Estimation of Location: Survey and Advances*. Princeton Univ. Press, Princeton, NJ.

[2] David. H.A. (1981). *Order Statistics*. John Wiley and Sons.

[3] Hogg, R.V. (1967). Some oberservations on robust estimation, *J. Amer. Stat.* Assoc. 62, 1179-1186.

[4] _____, Fisher D.M., Randles R.H. (1975). A two-sample adaptive distribution-free test. *J. Amer. Stat. Assoc.* 70, 656-661.

[5] _____, (1983). On adaptive statistical inference, *Comm. Statist.* 11, 2531-2542.

[6] _____, Bril, G.K. , S.M. Han, L. Yuh (1988). An Argument for Adaptive Robust Estimation, Probability and Statistics, *Essay in Honor of Graybill*, F.A., North Holland, 135-148.

[7] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.M., Stahel, W.A.(1986). *Robust Statistics, the Approach Based on Influence Functions*. John Wiley and Sons.

[8] Huber, P.J. (1972). Robust statistics: *A review. Ann. Math. Stat. 43*, 1041-1967.

[9] _____, (1981). *Robust statistics*. John Wiley and Sons.

[10] Mardia, K.V., Kent, J.T. , Bibby, J.M. (1980). *Multivariate Analysis*. Academic Press.

[11] Nakanish, H., Sato, Y. (1985), The performance of the linear and quadratic discriminant functions for three types of non-normal distributions, *Comm. Statist.* 14, 1181-1200.

[2001년 12월 접수, 2002년 5월 채택]