한국통계학회논문집 제4권 1호
1997년 4월 pp. 293-300

계산상의 단점들을 개선한 새로운 형태의 재하강 M-추정함수[1)]

박 노 진[2)]

요 약

재하강 M-추정함수는 여러 가지 장점을 갖고 있는 반면 이를 이용하여 추정치를 구할 경우 수치 해석상의 문제점을 갖고 있다. 물론, 그러한 문제들은 인위적인 방법에 의해 해결될 수 있으나 본 논문에서는 보다 근본적으로 문제를 해결하려 한다. 그 방법으로써 재하강 함수의 문제점을 갖고 있지 않은 Huber 형태에 유사하고 재하강 함수의 장점도 갖고 있는 새로운 형태의 재하강 M-추정함수를 유도하였다.

1. 배경

극단의 이상치들은 완전히 제거하고 보편적인 이상치들은 부드럽게 처리함을 통해 Huber 형태의 M-추정함수 (Huber, 1964)를 대체할 재하강 M-추정함수 (redescending M-estimation function; Andrews외 5인 (1972))가 제안되었고 널리 사용되고 있다. 그러나 재하강 M-추정함수를 사용하는 경우 계산상의 단점들이 지적되고 있다 (Huber (1980), Staudte 와 Sheather (1990)).

추정치 T_n의 계산은 뉴턴의 방법을 이용해 수치 해석적으로 이루어지는데, 즉,

$$T_n = T_n^{(0)} + \frac{S_n \sum_{i=1}^{n} \psi((x_i - T_n^{(0)})/S_n)}{\sum_{i=1}^{n} \psi'((x_i - T_n^{(0)})/S_n)}, \tag{1}$$

$T_n^{(0)}$은 초기치 그리고 S_n은 표준편차의 추정치를 나타낸다. (1)에서 문제는 ψ가 재하강 형태인 경우, 상승하는 부분의 미분 값과 하강하는 부분의 미분 값이 절대 값은 거의 같고 부호가 반대로 되고, 함수의 기울기가 양수에서 음수로 변하는 부분에 자료들이 많이 몰려 있다면, 식(1)의 두 번째 항의 분모가 0에 가까워지고, 결국 추정치는 비상식적인 값을 갖게 될 수가 있다. 또한, Huber (1980)가 지적하듯이 점근분산의 계산에서 이상치들이 하강하는 부분에 몰려 있다면, 점근분산의 분모에 지대한 영향을 미치게 된다. 즉, 점근분산

$$A(F, T) = \frac{\int \psi^2 dF}{\left(\int \psi' dF\right)^2} \tag{2}$$

1) 이 논문은 1996년도 기초과학연구소 학술연구조성비(BSRI-96-1439)에 의한 연구 입니다.
2) (300-173) 대전광역시 동구 용운동 대전 대학교 통계학과 조교수

의 계산에서 만일 이상치들이 몰려 있는 부근에서 (2)의 분모의 ψ'가 아주 큰 음의 값을 갖고 (2)의 분자의 ψ^2가 아주 큰 양의 값을 갖는다면 분산의 비정상적인 증가를 초래한다. 위의 단점들은 Huber의 M-추정함수에서는 일어나지 않는 현상들이다. 따라서, 본 논문에서는 재하강 형태를 유지하면서 Huber의 함수에 유사한 함수를 찾아내어 재하강 함수의 장점을 갖고 있으며 그 단점들을 어느 정도 제거한 새로운 재하강 M-추정함수를 도출하려 한다.

2. 새로운 재하강 M-추정함수의 유도

본 논문에서는 위치모수(location parameter)의 추정에 대해서만 논하기를 원한다. 먼저, $X_1, \cdots, X_n$ 을 확률 함수 $f(x;\theta)$에서 추출된 확률 표본이라 하자. 다음과 같이 자료와 모수를 포함하는 어떤 함수 $\rho(\cdot)$ 또는 그 함수의 미분 함수 $\psi(\cdot)$를 이용하여

$$\sum_{i=1}^{n} \rho(x_i - T_n) = \min!, \tag{3}$$

또는

$$\sum_{i=1}^{n} \psi(x_i - T_n) = 0, \text{ 여기서 } \psi(x;\theta) = (\partial/\partial\theta)\rho(x;\theta),$$

를 만족하는 추정치 T_n을 M-추정치라 한다. 이러한 방법을 통해 추정치를 구하는 추정법을 M-추정법이라 한다. 따라서, M-추정법의 궁극적인 과제는 여러 가지 바람직한 성질을 갖는 추정치를 도출해 내는 $\psi(\cdot)$를 찾는데 있다고 할 수 있다. 그러한 ψ-함수들 중에 1절에서 언급한 재하강 형태의 함수들이 고안되었고 여러 가지 장점들과 더불어 위에서 언급한 단점들도 있다. 그러한 단점들이 함수의 하강 부분의 기울기에 기인한다고 볼 때 재하강하지만 급하게 기울지 않는 함수를 도출함으로 그 단점들을 해소 내지 감소시키고자 한다. 그 방법으로 재하강 추정함수의 기울기를 완만하게 하기 위해 (3)의 식 대신에

$$\sum_{i=1}^{n} (x_i - T_n) \cdot \psi(x_i - T_n) = \min!,$$

즉,

$$\sum_{i=1}^{n} \psi(x_i - T_n) + (x_i - T_n) \cdot \psi'(x_i - T_n) = 0$$

을 만족하는 T_n으로 θ의 추정량을 삼는다. 이 제안에서 우리가 얻고자 하는 것은 로우버스트한 추정치를 생산하는 재하강 M-추정함수 $\psi(x_i - T_n)$와 기울기가 상승하는 직선 함수 $(x_i - T_n)$의 곱을 통하여 위에서 언급한 재하강하나 기울기가 급격하지 않은 M-추정함수를 고안하자는 것이다. 이 과정을 보다 일반화하면, M-추정함수 $\psi(x)$이 주어졌을 때 새로운 M-추정함수 $\psi^*(x)$을 다음과 같이 정의한다.

$$\psi^*(x) = \begin{cases} \Lambda(x), & 0 \le |x| \le p \\ \psi(x) + x\psi'(x), & p < |x| \le c, \text{ 여기서, } \psi(c) + c\psi'(c) = 0 \text{ 그리고 } c \neq 0 \\ 0, & \text{다른곳에서,} \end{cases} \tag{4}$$

여기서, $\Lambda=-f'/f=(-\ln f)'$ 그리고 $0<p<c$에 대하여 $\Lambda(p)=\psi(p)+p\psi'(p)$를 만족한다. 위의 정의로부터 원점 근처에서 Hampel외 3인 (1986)의 연구에 따라 점근분산을 최소화 하고자 하는 의도에서 함수 Λ를 사용한다.

우리는 본 논문에서 재하강 M-추정함수 중 정규 분포 하에서 점근분산을 작게 함에 있어서 최적인(optimal) Tanh-함수를 예로 들어 새로운 함수의 특징을 설명하고자 한다.

Tanh-함수 (Hampel외 3인, 1986): 주어진 상수 A,B,k,r,p에 대하여

$$\psi_{r,k}(x)=\begin{cases}\Lambda(x) & 0\le|x|\le p\\ (A(k-1))^{1/2}\tanh[(1/2)((k-1)B^2/A)^{1/2}(r-|x|)]\,\mathrm{sign}(x) & p\le|x|\le r\\ 0 & r\le|x|\end{cases},$$

여기서, $0<p<r$은 $\Lambda(p)=(A(k-1))^{1/2}\tanh[(1/2)((k-1)B^2/A)^{1/2}(r-p)]$ 를 만족한다.

만일 실효성을 비교하기 위해 모델이 정규 분포를 따른다고 할 때 $\Lambda(x)=x$가 된다. 따라서, 새로이 제안된 추정함수 $\psi^*_{r,k}(x)$는 식 (4)의 $\psi(x)$, $\psi'(x)$ 대신 $\psi_{r,k}(x)$, $\psi'_{r,k}(x)$를 대신 넣은 다소 복잡한 식으로 표현된다. 상수 A,B,r,k와 (4)의 조건을 만족하는 상수 p,c에 대하여,

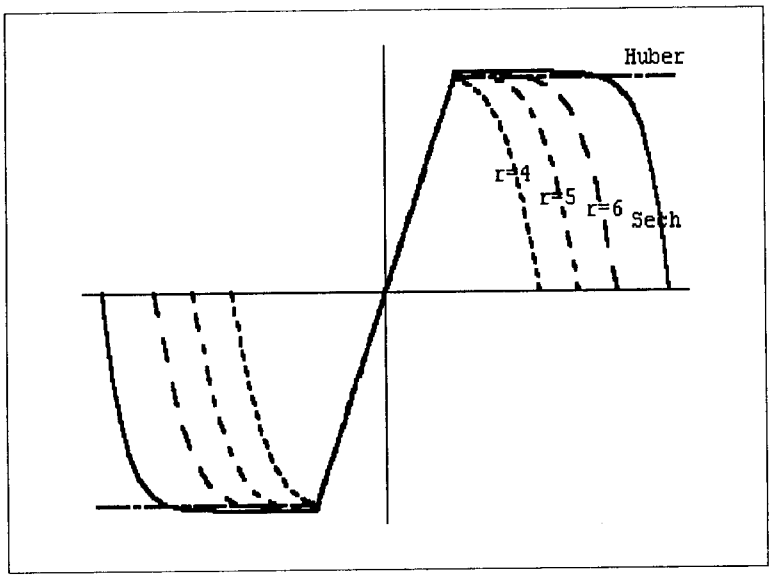
$$\psi^*_{r,k}(x)=\begin{cases}\Lambda(x) & 0\le|x|\le p\\ (A(k-1))^{1/2}\tanh[(1/2)((k-1)B^2/A)^{1/2}(r-|x|)]\,\mathrm{sign}(x)-x(A(k-1))^{1/2}\\ ((1/2)((k-1)B^2/A)^{1/2})\ \mathrm{sech}^2[(1/2)((k-1)B^2/A)^{1/2}(r-|x|)] & p\le|x|\le c\\ 0 & \text{다른곳에서,}\end{cases}$$

이때, 새로이 제안된 함수를 Sech-함수라고 하자. 예를 들어 gross-error-sensitivity가 1.97762에 근사하도록 구한 기존의 Tanh-함수와 새로이 제안된 추정함수를 <표 1>과 <그림 1>에 정리해 보았다.

<표 1> 기존의 Tanh-추정 함수들과 새로 제안된 추정 함수들의 예

	r	k	A	B	p	γ^*
$\psi_{r,k}(x)$	4.0	4.9917	0.857044	0.911135	1.801123	1.97757
	5.0	4.8340	0.893243	0.937508	1.842257	1.9775
	6.0	4.8092	0.941556	0.941556	1.849375	1.97762
$\psi^*_{r,k}(x)$	9.0	4.8092	0.885063	0.894478	1.848315	1.97762

<그림 1> Tanh-함수 (r = 4, 5, 6), 새로 제안된 추정함수 (Sech) 그리고 Huber-함수의 예; 원점 근처에서 함수들이 겹쳐 있다.



3. 제안된 추정함수의 성질

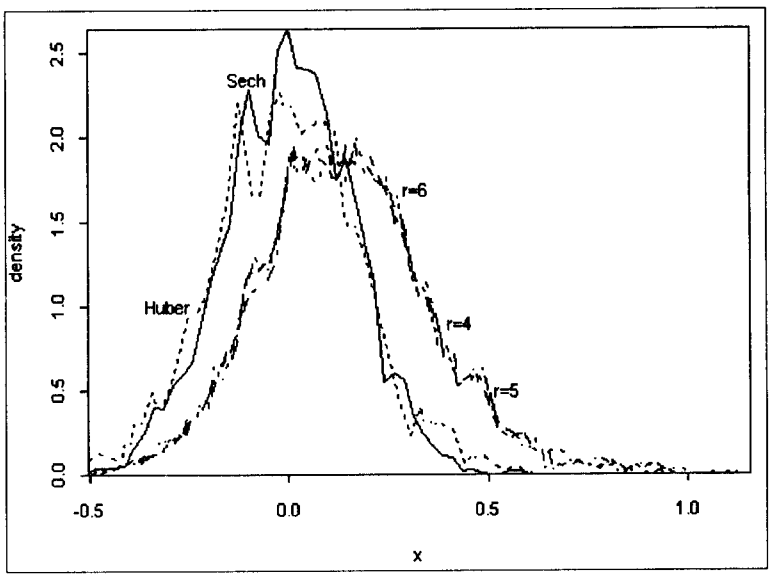
<표 1>에서 언급한 추정함수들을 <그림 1>에 그려보았다. 우리가 최초에 원했듯이 새로이 제안된 추정함수는 원점 근처에서 상승하고 어느 정도 Huber의 함수처럼 수평을 유지하고 하강하는 형태를 갖고 있다. 우리가 원했던 효과를 갖고 있는지 알아보기 위해 간단한 모의실험을 실시하여 보았다. 먼저 표준 정규 분포에서 크기가 40인 1000개의 표본을 택하여 위의 예의 네 가지 함수들과 Huber-함수를 적용하여 추정치를 계산하여 <그림 2-(1)>의 경험적 분포 함수를 그려보았다. 그리고, 표준 정규 분포에서 30개, 평균이 -4와 4이고 분산이 1인 정규 분포에서 각각 5개씩, 총 크기가 40인 1000개의 표본을 택하여 추정치를 구하여 경험적 분포 함수를 <그림 2-(2)>에 그려보았다. Huber-함수는 gross-error-sensitivity가 1.97762가 되도록 b = 1.8506인 경우를 사용했다. <그림 2-(2)>에서 볼 수 있듯이 기존의 Tanh-함수가 하강하는 -4와 4 근처에서 일부 자료를 택하여 모의실험을 한 결과 추정치들이 0을 크게 벗어나 오른쪽으로 치우쳐 있음을 알 수 있다. 즉, 계산식 (1)의 분모가 표준 정규 분포 하에서 분자의 값에 비해 상대적으로 작아짐으로 그 값이 커지게 된다. 그리고 상승하는 부분에 더 많은 자료들이 위치하기 때문에 그 추정치들은 양의 값을 갖는 경향이 많게 된다. 반면, 새로운 함수와 Huber-함수를 이용한 경우 정상적인 분포, 비정상적인 분포 하에서 동일하게 추정치들이 0을 중심으로 분포되어 있음을 알 수 있다. 즉, 새로 제안된 함수가 보다 안정적인 추정치들을 산출해 낸다고 할 수 있겠다. <그림 2-(3)>과 <그림 2-(4)>는 위와 유사한 모의실험을 평균이 각각 1이고 4이고, 분산이 1로 동일한 정규분포에서 10개씩 택한 후, 표준 정규 분포에서 택한 30개와 합하여 크기가 40인 1000개의 표본 두 쌍을 갖고

수행하였다. 그림들을 통해 쉽게 알 수 있듯이, 분포가 비대칭인 경우 꼬리가 두터울수록 새로운 함수의 효과가 감소함을 알 수 있다.

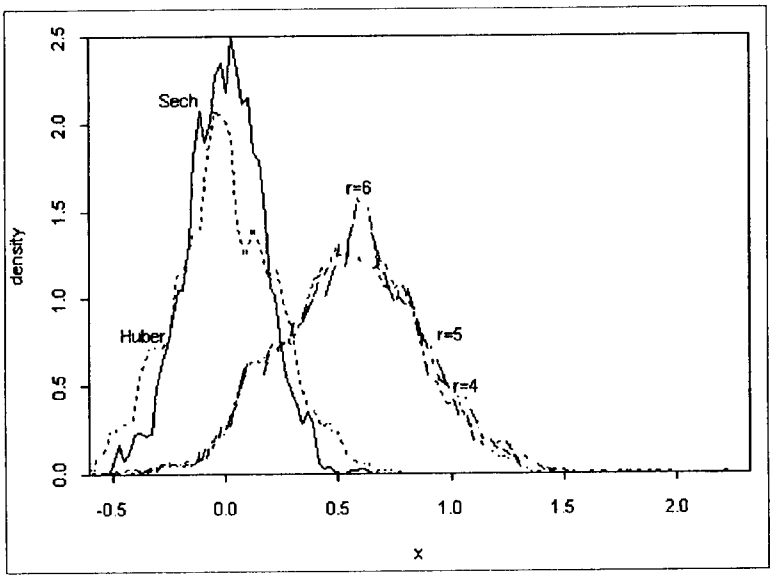
또한, 위의 함수들을 근거로 점근분산을 계산하여 <표 2>에 정리하였다. 새로운 함수의 점근분산들은 Huber-함수를 적용한 경우의 수치들과 그 크기나 여러 가지 분포 하에의 수치들의 변화하는 모습이 유사함을 알 수 있다. 또한, 정규 분포 하에서 효율의 경우 기존의 Tanh-함수를 적용한 경우 보다 좋아지는 경향이 있다. 물론, 새로운 함수와 근사한 r=6인 경우 실효성이 0.1%정도 떨어지나 추정치의 안정성을 고려한다면 큰 차이라고 생각되지는 않는다.

<그림 2> Tanh-함수 (r = 4, 5, 6)(일점 쇄선, 굵은 점선, 이점 쇄선), 새로운 함수(Sech)(실선), Huber-함수(가는 점선)에 의한 추정치들의 경험적 분포 함수; 세로축은 상대 빈도, 가로축은 추정치를 나타낸다.

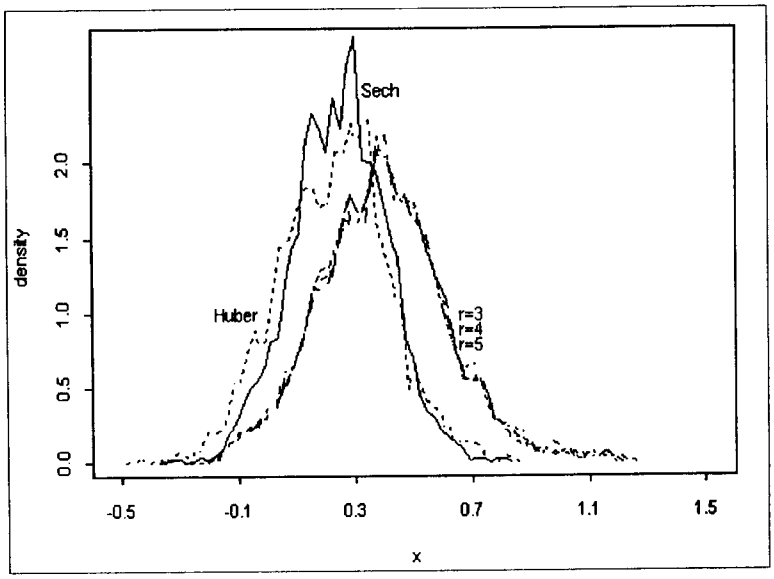
(1). N[0, 1]의 경우



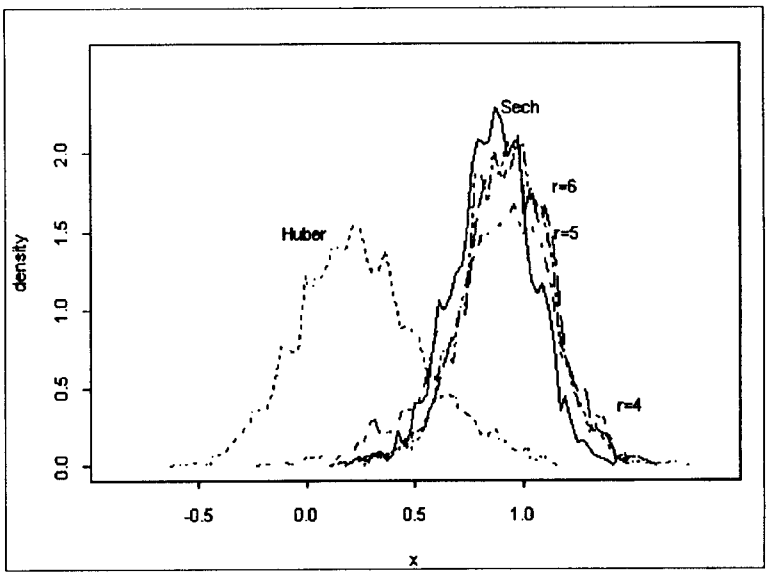
(2). (1/8)N[4, 1] + (1/8)N[-4, 1]+ (3/4)N[0, 1]의 혼합 분포의 경우



(3). (1/4)N[1, 1] + (3/4)N[0, 1]의 혼합 분포의 경우



(4). (1/4)N[4, 1] + (3/4)N[0, 1]의 혼합 분포의 경우



<표 2> 여러 가지 분포들에 대한 점근 분산

	r	효율 (efficiency)	점근 분산 5%3N	10%10N	t(3)	25%3N	Cauchy(0,1)
$\psi_{r,k}(x)$	4.0	1.032546	1.1434	1.21553	1.61644	1.77563	2.55528
	5.0	1.018197	1.14196	1.23019	1.61723	1.82723	2.66232
	6.0	1.015934	1.14856	1.25955	1.62559	1.86101	2.74855
$\psi^*_{r,k}(x)$	9.0	1.017454	1.16747	1.34886	1.68622	1.98268	3.2028
Huber b = 1.8506		1.015566	1.15623	1.51074	1.65323	1.87413	3.31104

여기서, 5%3N은 $0.95\Phi(x) + 0.05\Phi(x/3)$, t(3)는 자유도가 3인 t-분포를 의미한다. 위의 모든 함수들은 gross-error-sensitivity가 1.97762에 근사하도록 정하였다.

4. 결론

본 논문에서는 기존에 여러 가지 면에서 최적하다고 생각되어 사용되는 Tanh-함수의 계산상의 오류를 극복하기 위한 방편으로 Huber-함수와 유사하게 상승 후 잠시 수평을 유지하고 하강하는 재하강 추정함수를 도출했다. 새로 제안된 함수는 예상대로 추정치가 기존의 Tanh-함수를 사용할 때 보다 정상적인 분포, 비정상적인 분포 하에서 대체로 안정되게 계산되었다. 한편, 동일한 gross-error-sensitivity를 갖는 Tanh-함수를 적용한 경우 보다 정규 분포 하에서 대체로 효율이 좋아진다. 비록, 어떤 경우 효율이 Tanh-함수 보다 조금 떨어질 수도 있었으나, 추정치의 안정성을 고려할 때 그렇게 큰 손실은 아니라고 생각된다. 이 새로운 형태의 함수를 기존의 함수들과 함께 사용할 때 통계분석의 안정성을 조금이라도 향상시킬 수 있으리라 본다.

참 고 문 헌

[1] Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton University Press, Princeton, N.J.

[2] 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.

[3] Huber, P. J. (1964). Robust estimation of a location parameter *The Annals of Mathematical statistics,* Vol. **35**, 73-101.

[4] Huber, P. J. (1980). *Robust Statistics*, Wiley, New York.

[5] Staudte, R. G. and Sheather, S. J. (1990). *Robust estimation and testing*, Wiley, New York.