한국통계학회논문집 제3권 1호
1996년 4월 pp. 229-234

선형결합을 통한 새로운 ψ-함수의 도출

박 노 진[1)]

요 약

로우버스트 추정에서 자주 사용되는 Huber의 ψ-함수와 재하강 Tanh-함수의 선형 결합을 통해 새로운 재하강 ψ-함수를 도출한다. 이 함수를 사용하면 적절한 조건하에서 앞의 두 함수를 사용할 때 보다 위치 모수(location parameter)에 대한 추정량의 점근분산(asymptotic variance)을 감소시킬 수 있음을 보였다.

1. 서론

최소제곱추정량이 이상치에 많은 영향을 받기 때문에 이상치에 영향을 적게 받거나 전혀 받지 않는 추정량을 구하기 위하여 로우버스트한 추정법들이 연구되어 왔다. 그 대표적인 방법을 'M-추정법' 이라 한다. M-추정법은 확률 표본 $X_1,\cdots,X_n$이 위치 모수 μ를 갖고 있는 어떤 확률 함수 $f(x-\mu)$에서 추출되었다면, 확률 변수들과 모수의 차이에 대한 어떤 함수를 정의하고, 그 합을 최소화함으로 추정량을 구하는 방법이다. 즉, 어떤 특정한 함수 $\rho(x-\mu)$를 정의하여

$$\sum_{i=1}^{n}\rho(X_i-\mu) \tag{1}$$

를 최소화하는 모수에 대한 추정량을 도출하는 방법이다. 여기서, $\rho(t)=t^2$ 일 때 위에 언급한 최소제곱추정량을 얻게 되고, $\rho(t)=|t|$ 일 때 중앙값(median)을 얻게 된다. 한편, 함수가 미분 가능하다면

$$\sum_{i=1}^{n}\psi(X_i-\mu)=0,\quad \psi(X_i-\mu)=-\rho'(X_i-\mu) \tag{2}$$

의 해가 원하는 추정량이 될 것이다. 대표적인 ψ-함수로서 Huber의 ψ-함수 (Huber, 1964)와 Tanh-함수 (Huber, 1980 그리고 Hampel 외 3인, 1986) 를 꼽을 수 있고, 그들은 다음과 같이 정의된다.

1) (300-716) 대전 광역시 동구 용운동 대전대학교 통계학과 전임강사.

Huber의 ψ-함수:

$$\psi_b(x) = \min\{b, \max\{x, -b\}\} = x \cdot \min\left(1, \frac{b}{|x|}\right), \quad 0 < b < \infty. \tag{3}$$

Tanh-함수: 주어진 상수 A, B, k, p, r 에 대하여

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

여기서, $0 < p < r$ 에 대하여

$$p = (A(k-1))^{1/2} \tanh[(1/2)((k-1)B^2/A)^{1/2}(r-p)]. \tag{4}$$

Huber의 함수에 의한 추정량 (Huber 추정량)과 Tanh-함수에 의한 추정량 (Tanh 추정량) 모두 로우버스트 하다. 그러나, 점근분산의 경우는 정규 분포를 가정하면 전자가 후자보다 작으나, 혼합된 분포(mixture distribution)나 코쉬 분포 같은 꼬리가 두꺼운 분포를 가정하면 반대로 Tanh 추정량의 점근분산이 Huber 추정량의 그것 보다 작아질 수 있다. 그렇다면, 로우버스트한 추정 법이 이상치가 존재할 때 그 진가를 발휘해야 한다면 Tanh-함수가 훨씬 유용하다고 할 수 있다. 실제로, Tanh-함수 같은 형태의 함수를 통틀어 재하강 ψ-함수(redescending ψ-function)라고 부르며 위에 언급한 Tanh-함수는 그들 중 효율성이 극대화된(optimal) 추정량을 도출한다. 본 논문에서는 재하강 하는 함수로서 정규 분포 하에서도 효율성이 뛰어난 함수를 찾기 위해, Huber 함수와 Tanh 함수의 선형 결합을 시도하였다. 그 결과 정규 분포 하에서 Tanh 추정량보다 효율성이 향상되고, 어떤 경우 코쉬 분포 하에서도 Tanh 추정량보다 점근분산이 감소한 약간 변형된 재하강 ψ-함수를 도출 할 수 있었다.

2. 변형된 재하강 Tanh 추정량

Huber-함수 와 Tanh-함수의 선형 결합을 다음과 같이 정의하자.

$$\psi_\alpha(x) = \alpha \cdot \psi_b(x) + (1-\alpha) \cdot \chi_{r,k}(x), \quad 0 \le \alpha \le 1. \tag{4}$$

(4)에 의하면, Huber-함수는 α=0인 경우이고, Tanh-함수는 α=1인 경우이다. 예를 들어 α=0.5 인 경우, <그림 1> 에서 보듯 $\psi_{\alpha=0.5}(x)$는 원점 근처에서 45도의 직선을 이루고, 한계점 밖에서 하강하는 형태를 보인다.

일반적으로, ψ-함수의 점근분산은 다음 같이 계산된다.

$$V(T,F) = \frac{\int \psi^2(x, T(F))dF(x)}{[\int (\partial/\partial\theta)[\psi(x,\theta(F))]|_{\{\theta(F)=T(F)\}} dF(x)]^2}, \tag{5}$$

여기서, F는 가정된 누적 확률 함수를 의미하고, $T(F)$는 가정된 F하에서의 모수에 대한 추정량을 의미한다. (5)에서 보듯이, 점근분산을 작게 하려면 분모를 크게 하거나 분자를 작게 하면 가능하다.

<그림 2>와 <그림 3>은 정규 분포와 모수가 0과 1인 코쉬 분포 하에서 (5)의 피적분 함수들을 보여주고 있다.

<그림 2-(a)>, <그림 3-(a)>는 정규 분포 하에서 (5)의 분자, 분모의 피적분 함수들의 상태를 보여주고 있다. <그림 2-(a)>를 보면 (5)의 분자에서 적분될 함수들의 차이가 거의 보이지 않는다. 반면, <그림 3-(a)>를 보면 (5)의 분모에서 적분될 피적분 함수들이 ψ-함수들의 기울기에 따라서 3 사분면과 4 사분면 위에서의 면적이 넓어지고, 결국 ψ_b, ψ_a, $\chi_{r,k}$ 순서로 피적분 함수들의 적분값들이 미세하지만 작아짐을 알 수 있다. 따라서, 점근분산은 Huber 추정량, 변형된 Tanh 추정량, Tanh 추정량 순서로 커지게 된다.

코쉬 분포 같은 꼬리가 두터운 분포를 가정할 경우, <그림 3-(b)>에서 보듯이 분모의 피적분 함수들의 적분값들은 그 크기의 순서에 있어서 정규 분포의 경우와 동일한 모습을 보여준다. 하지만, <그림 2-(b)>를 보면 정규 분포의 경우에 비해, 점근분산의 분자에 속한 세 피적분 함수들의 적분 값들이 분모의 적분값들 보다 상대적으로 크게 변화한다. 따라서, 점근분산은 정규 분포의 경우와는 반대로 Huber 추정량, 변형된 Tanh 추정량, Tanh 추정량 순서로 작아지게 된다. 그러나, 한가지 주지할 점은 코쉬 분포의 경우--아래에서 자세히 설명하겠지만-적당한 α에 대하여는 변형된 Tanh 추정량이 Tanh 추정량 보다 점근분산이 작아질 수 있다는 것이다.

위의 관측들을 정리해서 말하자면, <그림 2>와 <그림 3>은 한계점 밖에서의 함수들의 기울기 변화는 정규 분포같이 꼬리가 얇은 분포 하에서 점근분산의 분모에 영향을 끼치는 반면, 코쉬 분포 같이 꼬리가 두터운 분포 하에서 점근분산의 분자에 영향을 끼친다고 할 수 있다.

구체적인 결과를 보기 위해 표준 정규 분포 하에서 gross-error sensitivity (γ^*)를 1.6749로 하는 Huber-함수와 Tanh-함수를 택하자. γ^*=1.6749인 경우 (3)의 Huber-함수의 b 는 1.4088 로 계산되고, (4)의 Tanh-함수의 계수들은 각각 A=0.667, B=0.783, r=4, k=3.732, p=1.312 로 계산된다. (각 계수들의 계산 방식은 Hampel 외 3인의 2.6절을 참조하기 바란다. 두 함수의 모양은 <그림 1>을 참조하기 바란다.)

위 두 함수를 이용하여 선형 결합을 시행하고 그 중 α=0.5인 경우가 <그림 1>에 나타나 있다. <그림 4>는 표준 정규 분포, 코쉬 분포, 표준 정규 분포와 분산이 9이고 평균이 0인 정규 분포의 혼합 분포 그리고 자유도가 3인 t 분포 하에서 α에 대한 점근분산들의 변화를 보여주고 있다. 이와 더불어 크기가 40인 1000개의 표본을 생성하여 모의 실험한 결과도 보여주고 있다. 예상대로 정규 분포 하에서 변형된 Tanh 추정량이 Tanh 추정량 보다 점근분산이 적으며, 코쉬 분포 하에서도 α 를 0.16 부근에서 정하면 변형된 Tanh 추정량이 Tanh 추정량 보다 점근분산이 적게 된다. 그 밖의 분포에서도 특정한 α 값들에서 본 논문이 제안하는 추정량의 점근분산이 현재 널리 받아들여지고 있는 Tanh 추정량 보다 적어짐을 알 수 있다. 모의 실험에의한 분산들은 이론에 의한 값들의 변화와 동일한 방향으로 변화하고 있으며, 이론치와 차이를 보이나 실제로 그 차이는 거의 동일하다고 볼만큼 수학적으로 크지는 않다.

3. 결론

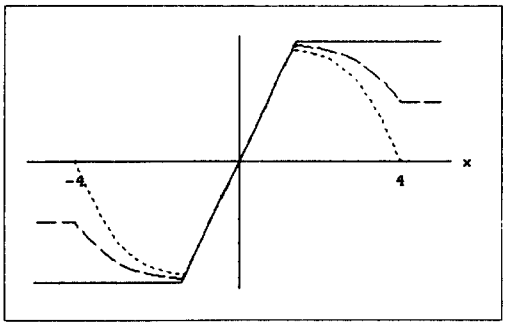
Huber의 ψ_b-함수와 Tanh-함수는 로우버스트 추정에서 중요한 가치를 지니며, 나름대로 장점을 지니고 있다. 두 함수의 선형 결합을 통해, Tanh 추정량보다 점근분산이 감소된 추정량을 산출 해내는 재하강 형태의 함수를 유도 할 수 있음을 보였다. 특히, 꼬리가 두터운 확률함수에 본 논문에서 제안된 함수를 적용할 경우 Huber-함수와 Tanh-함수에 근거한 추정량보다 점근분산이 작게될 수가 있음을 보였다.

참고문헌

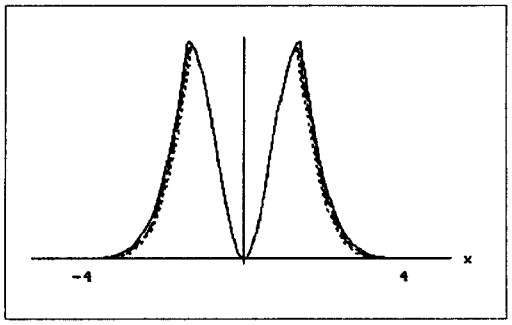
[1] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics, The approach based on influence functions*, Wiley, New York.

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

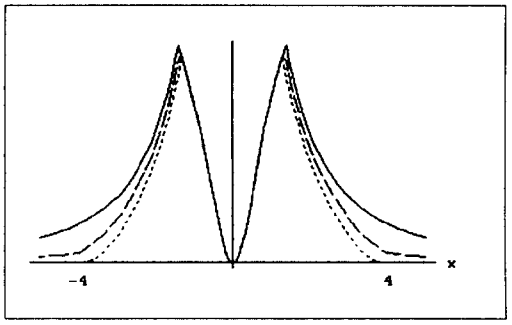
[3]. Huber, P. J. (1980). *Robust Statistics,* Wiley, New York.



<그림 1> $\psi_{b=1.4088}(x)$(실선), $\psi_{a=0.5}(x)$(굵은 점선), $\chi_{r=4,\ k=3.732}(x)$(점선)
; 원점 근처에서는 세 그래프가 겹쳐 있음.

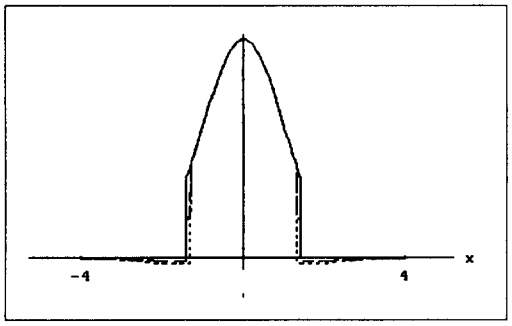


(a). $f(x)$가 표준 정규 분포인 경우.

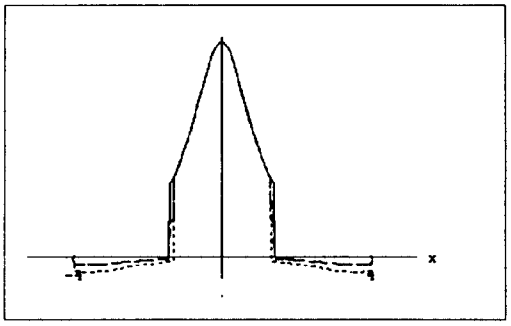


(b). $f(x)$가 코쉬 분포인 경우.

<그림 2> $\psi^2_{b=1.4088}(x)*f(x)$(실선), $\psi^2_{a=0.5}(x)*f(x)$(굵은 점선), $\chi^2_{r=4,k=3.732}(x)*f(x)$(점선)
; 원점 근처에서는 세 그래프가 겹쳐 있음.

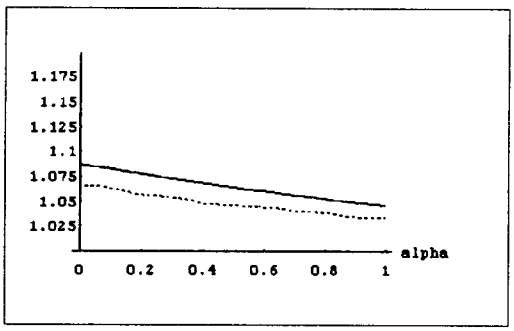


(a). $f(x)$가 표준 정규 분포인 경우.

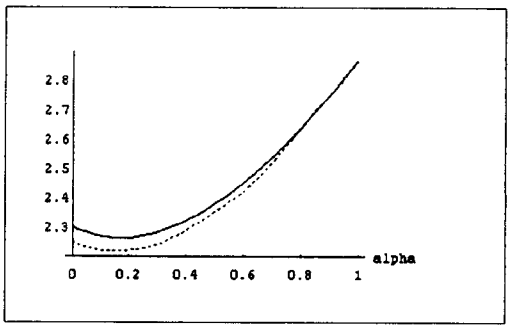


(b). $f(x)$가 코쉬 분포인 경우.

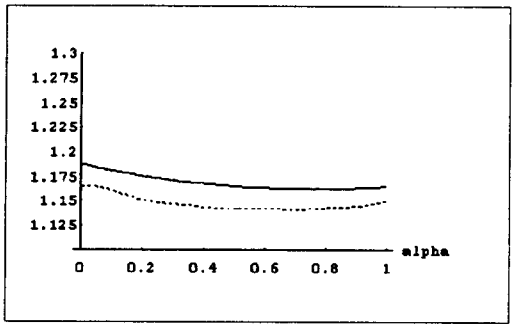
<그림 3> $\psi'_{b=1.4088}(x)*f(x)$(실선), $\phi'_{a=0.5}(x)*f(x)$(굵은 점선), $\chi'_{r=4,k=3.732}(x)*f(x)$(점선)
;원점 근처에서는 세 그래프가 겹쳐 있음.



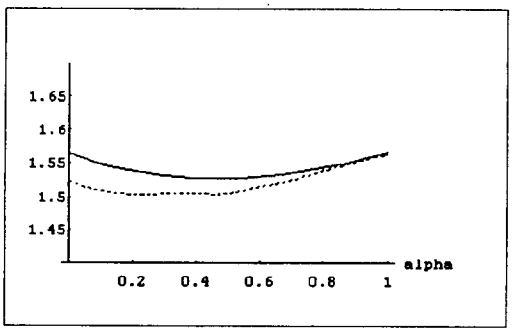
(a). 표준 정규 분포인 경우.



(b).코쉬 분포인 경우.



(c). 0.95*N(0, 1) + 0.05*N(0, 9)인 경우.



(d) 자유도가 3인 t 분포의 경우.

<그림 4> α에 따른 점근분산의 이론에 의한 변화(실선)와 모의 실험에 의한 변화(점선)