제주교육대학교 논문집(1995)
제24집 PP. 1~10

ELMS에서 SUBSET의 최적 크기

김 종 우*

The Optimal Size of Subset for ELMS

Jong - Woo Kim

요 약

Rousseeuw(1984)가 제시한 Least Median of Squares(LMS) 방법을 개선한 염준근, 박종구, 김종우(1995)의 ELMS는 효율적으로 이상치들을 인식한다. 그러나 이 방법에서 최적의 subset을 결정하기 위하여 사용하는 종유석 그림은 정확한 기준 점을 선정하는데 어려움을 나타내고 있다. 따라서 ELMS를 사용한 이상치 판별을 위한 적절한 크기의 subset 결정 방법을 제시하고자 한다.

1. 서 론

선형회귀분석(linear regression analysis)에서 다중 이상치의 존재시에는 이상치가 군집해 있는 방향으로 중심점을 끌어 당기는 영향에 의하여 이상치를 숨기려는 masking 효과와 정상적인 점들이 중심점에서 멀리 떨어져 있는 점으로 인식되는 swamping효과에 의해 이상치와 영향력 관찰점의 인식을 어렵게 하고 있다(Belsely, Kuh & Welch, 1980; Rousseeuw and Leroy, 1987). 이러한 문제점을 극복하고자 개발된 robust 추정량으로 Rousseeuw(1984)의 잔차제곱의 중위수를 최소화하는 least median of squares(LMS) 추정량은 매우 높은 breakdown point를 갖고 있으나, 효율성이 매우 나쁜 것으로 알려져 있다. 따라서 이 점을 극복하기 위하여 염준근, 박종구, 김종우(1995)의 ELMS는 높은 breakdown point를 유지하면서 빠른 추정을 할

* 제주교육대학교 수학교육과 부교수
Department of Mathematics Education, Cheju National University of Education, Cheju, 690-060, Korea.

수 있는 방법을 제시하고 있다. 그러나 ELMS에서 사용하고 있는 종유석 그림에 의한 이상치 판별은 개괄적인 이상치의 분포 모양을 제시하고 있으므로 보다 정확한 이상치 인식의 범위를 설정하는 데 부족함을 나타내고 있다.

본 연구에서는 선형회귀 구조를 갖는 모집단에서 다수 이상치 인식을 위해 ELMS를 재 분석하고 최적 subset의 크기를 얻기 위하여 염준근, 박종구, 김종우(1995)에서 제시한 ELMS algorithm을 사용하여 각 표본크기에 따른 이상치 인식 정도를 조사하고자 한다.

2. LMS와 ELMS

2.1 Ordinary Least Squares

고전적인 선형회귀모형(OLS)을 다음과 같이 설정하자.

$$Y = X\beta + \varepsilon \tag{1}$$

여기서 $Y = (y_1, y_2, \ldots, y_n)^t$는 종속변수인 $n \times 1$ 벡터.

$X = (x_1, x_2, \ldots, x_n)^t$는 독립변수인 $n \times p$ 행렬 $(p < n)$,

$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$는 $p \times 1$ 벡터.

$\beta = (\beta_1, \beta_2, \ldots, \beta_{p+1})$인 모수 $(p+1) \times 1$ 벡터.

$\varepsilon \sim N(o, \sigma^2)$인 $n \times 1$ 벡터.

β와 σ^2의 최소제곱 추정량은

$$\hat{\beta} = (X^tX)^{-1}X^tY, \tag{2}$$

$$\hat{\sigma^2} = e^te/(n-p), \text{ 여기서 } e = Y - X\hat{\beta}$$

으로 얻어진다. OLS에서 이상치의 여부를 파악하기 위하여 잔차를 이용한 다양한 함수들이 개발되어 사용되고 있다[염준근, 1993]. 그러나 이러한 최소제곱을 이용한 추정량은 이상치가 군집되어 있을 때 masking과 swamping 효과에 의하여 0%에 가까운 breakdown point를 갖게 된다[Rousseeuw and Leroy, 1987].

2.2 Least Median of Squares (LMS)

LMS의 주된 효과는 크기 $p+1$인 subset J를 다음과 같이 구하는데 있다.

$$\underset{\hat{\beta}}{\text{Minimize}} \ \underset{J}{\text{med}}\ e_i^2 \tag{3}$$

식(3)을 만족하는 subset J를 선택한다. 모수 $\hat{\beta}$을 얻기 위하여 수 subset J로 구성된 X_J와 Y_J를 이용하여

$$\hat{\beta} = (X_J^t X_J)^{-1} X_J^t Y_J \tag{4}$$

를 계산한다.

식(4)의 $\hat{\beta}$를 사용하여 잔차

$$e_i = Y_i - X_i \hat{\beta} \tag{5}$$

를 구한다. 다음에 scale factor σ^*를 구하기 위하여 식(3)에서 결정된 최소 중위수 잔차와 n과 p에 결정되는 유한표본수정계수를 사용하여 초기 수정계수 s^o와 가중치 w_i를 결정한다.

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i - p}} \tag{6}$$

여기서 $w_i = \begin{cases} 1 & |e_i/s^o| \le 2.5 \\ 0 & \text{otherwise.} \end{cases}$

$$s^o = 1.4826\,(1 + \frac{5}{n-p})\sqrt{\min \text{med}_J\, e_i^2}$$

이상치의 여부를 파악하기 위하여 식(5)에서 계산된 잔차 e_i와 식(6)에서 구한 σ^*를 사용하여

$$\left|\frac{e_i}{\sigma^*}\right| > 2.5 \tag{7}$$

이면 i번째 관찰치를 이상치로 간주한다.

OLS에서 문제시되는 효과를 제거하기 위한 방법으로 최근에 널리 사용되고 있는 LMS방법은 breakdown point가 거의 50%에 달하는 높은 안정성을 보이고 있다. 그러나 이 방법의 적용을 위한 random search algorithm은 크기 p인 subset을 구하기 위하여는 $n \times p$행렬에서 nCp회의 표본추출 즉, $n! / p!(n-p)!$회에 달하는 계산을 필요로 하는 n^{-3}의 접근 속도를 갖고 있다[송문섭, 1995]. 즉, LMS의 algorithm은 모수 $\hat{\beta}$를 결정하기 위해 행렬 X에서 크기 $p+1$인 subset을 random 추출하여 중위수를 구하고, 각각의 subset에서 얻어지는 중위수들 중에 최소인 subset을 택하여 전체 관측치의 잔차를 구하는 방법에 의한 이상치 식별을 하고 있다[염준근, 박종구, 김종우, 1995].

2.3 Extended Least Median of Squares (ELMS)

LMS에서 많은 계산 횟수를 줄이기 위한 방법으로 제시된 ELMS 방법은 다음과 같이 4단계로 구성되어 있다.

1단계. 초기 subset의 설정.

식(2)에서 제시된 LS의 잔차를 사용하여 잔차 e_i를 ascending order순으로 정렬하여 크기 순으로 크기가 $p+1$인 인 초기 subset J를 설정한다.

$$e_1 \leq e_2 \leq , \ldots , \leq e_n \text{ 일때,}$$

$$J = \{ y_1, y_2, \ldots , y_k \}, \quad \text{여기서 } k = p+1$$

2단계. 최적 subset J의 결정.

1단계에서 subset J에서 $k-1$개의 원을 취하고 나머지 subset에서 1개의 원을 취하여 새로운 subset J를 구성한다. 이 subset J를 총 $\binom{k}{k-1}$ 개의 subset로 구성되며, subset을 결정하는 방법인 $\min_\theta(\text{med}\gamma_i^2)_J$인 subset J를 찾는다.

3단계. 이상치 식별.

2단계에서 구한 subset J를 사용하여 LMS에서 사용한 이상치 식별 방법을 사용한다.

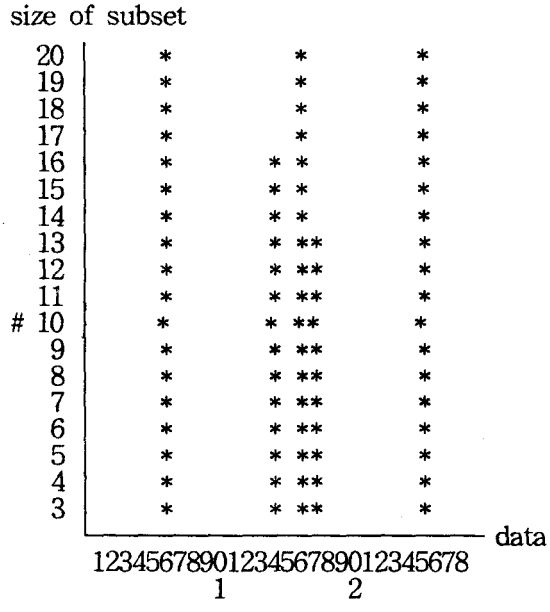
4단계. subset J의 크기 결정.

ELMS의 사용은 염준근, 박종구, 김종우(1995)에서 지적하고 있는 바와 같이 효율성에서 매우 높음을 알 수 있다. 문제점으로는 지적하고 있는 Hadi & Simonoff(1994)

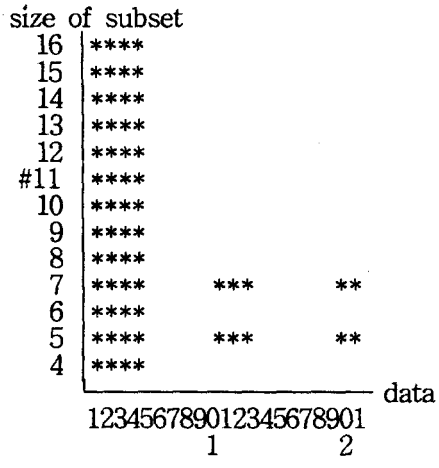
가 LMS 사용시에 발생할 수 있는 독립변수 X공간상에서의 상관성이 비교적 높게 발생하는 공선성의 경우에는 취약함을 보이고 있다.

3. 종유석 그림

ELMS의 실행이 subset의 크기에 따라 이상치를 적절히 인식할 수 있는 지를 알아보기 위하여 Atkinson & Mulira(1993)이 제안한 종유석 그림을 사용하여 분석한다.

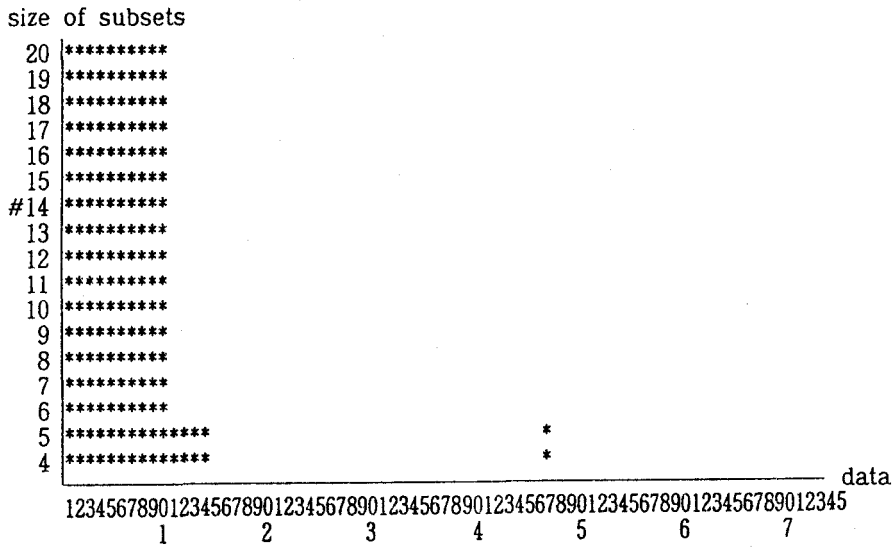


〈그림 1〉 The Stalactite Plot for Body and Brain Weight Data



〈그림 2〉 The Stalactite Plot for Stack Loss data

〈그림 1〉, 〈그림 2〉, 〈그림 3〉에서 사용하고 있는 자료는 Rousseeuw and Leroy (1987)의 해당 자료의 이상치 분석 결과와 비교하면, subset의 크기가 매우 폭 넓은 범위에서 이상치들을 잘 지적하고 있음을 알 수 있다. 더불어 염준근, 박종구, 김종우 (1995)에서 제시하고 있는 $[\log(n^3/p)]+p$개 근처([#]로 표시)에서 매우 안정성을 갖고 있는 것으로 나타난다. 한편 위 그림의 이상치 인식 정도에서 보여 주고 있듯이 과도한 이상치 인식 과정을 행하고 있음을 알 수 있다. 따라서 각 subset의 크기에 따른 단계별로 잔차의 정도를 분석하여 어느 정도의 subset크기가 적절하게 이상치를 인식하고 있는지를 알아보기로 한다.



〈그림 3〉 The Stalacitite Plot for Artificial Data Set

4. ELMS의 잔차 검토

ELMS algorithm을 사용한 이상치 인식 정도를 조사하기 위하여 앞에서 제시된 3가지 자료의 잔차가 subset의 크기에 따라 변화하는 정도를 측정하여 본다.

〈표 1〉, 〈표 2〉, 〈표 3〉의 이상치 선정은 Rousseeuw and Leroy (1987)에서 LMS에 근거를 둔 이상치 자료에 대하여 ELMS를 사용할 때 표본의 크기 변화에 따라 이상치의 잔차 변화를 측정한 것이다. 이 자료의 분석에 따르면 전체적으로 잔차의 크기는 표본의 크기가 초기 subset의 크기에서 보다 커져감에 따라 인식정도를 높이고 있으나 표본의 크기가 어느 일정한 정도에서 도달하면 잔차의 인식 정도도 서서히 낮아지고 있음을 보이고 있다. 계속해서 표본의 크기가 증가하면 결국에는 고전적인 LS와 매우 근사적인 잔차 값을 갖게 된다. 그러면 최적의 표본 크기를 현재 제시된 3가지 자료를 중심으로 조사하여 보자. 〈표 1〉에서 표본의 크기가 3 또는 4 정도에 달했을 때, 매우 높은 이상치 인식율을 보임을 알 수 있다. 〈표 2〉에서는 표본의 크기가 5 또는 7 정도의 크기를 가질 때, 〈표 3〉에서는 표본의 크기가 6 또는 7 정도에 달했을 때, 매우 높은 이상치 인식율을 보임을 알 수 있다. 따라서 이러한 표본의 크기는

$$[\frac{5}{59}n+\frac{96}{59}] \text{ 에서 } [\frac{5}{90}n+\frac{96}{90}]$$

범위의 표본 크기를 택하게 됨을 뜻한다.

이상치 / 표본크기	6	14	16	17	25
2	-7.334413	1.7653969	-6.717083	1.5907103	-7.990024
3	-8.787268	4.2084481	-7.879380	3.5147892	-9.410481
4	-9.633684	4.1473545	-8.676368	3.5036981	-10.35935
5	-9.474879	4.0304520	-8.537428	3.4117283	-10.19452
6	-9.093667	3.9322930	-8.189483	3.3354005	-9.787676
7	-8.965143	3.8126015	-8.078088	3.2253931	-9.644735
8	-8.685206	3.6990317	-7.825691	3.1336388	-9.346500
9	-8.341741	3.5882141	-7.513595	3.0408587	-8.976872
#10	-8.023014	3.4631404	-7.226208	2.9451679	-8.640877
11	-7.699505	3.3424023	-6.933046	2.8365066	-8.287862
12	-7.372493	3.2078598	-6.638414	2.7288787	-7.940322
13	-7.223645	3.0637734	-6.509768	2.5959872	-7.774296
14	-6.952520	2.9075391	-6.268814	2.4679126	-7.486447
15	-6.508456	2.7765970	-5.863705	2.3471077	-7.000321
16	-6.079249	2.6194013	-5.475614	2.2237357	-6.544772
17	-5.796572	2.4352076	-5.225920	2.0706569	-6.244073

〈표 1〉 Residual Index for Body and Brain Weight Data

이상치 / 표본크기	1	3	4	21
4	4.8426137	2.5569658	4.6822895	3.7820073
5	13.788582	6.7175144	13.376103	12.020815
6	6.6568669	4.0307464	6.1551137	5.0910501
7	14.082600	8.1666931	13.181366	10.540596
8	5.8918065	3.0357100	5.6176414	4.8751775
9	6.8676109	2.9529936	6.4341312	7.2048625
10	7.5416156	3.5259134	6.9460500	7.4767116
#11	6.5572270	2.8666331	6.1231896	6.8752214
12	4.6265773	1.9496624	4.4028185	4.8954311
13	3.7721034	1.5223300	3.5552962	4.0790489
14	3.1628202	1.2533430	3.0277994	3.4041626
15	3.2184665	1.4679233	3.0456010	3.1786542
16	1.9224840	0.3730920	1.9527006	2.5193086
17	0.8729293	-0.000522	0.9493794	1.2057568

〈표 2〉 Residual Index for Stack Loss data

이상치 / 표본크기	1	2	3	4	5	6	7	8	9	10
4	5. 2028	5. 766	5. 8096	4. 7359	5. 2734	5. 0418	6. 0483	6. 1696	5. 1970	6. 1064
5	14. 801	15. 836	15. 973	15. 313	15. 708	14. 989	16. 303	15. 950	15. 477	16. 080
6	15. 609	16. 392	16. 710	15. 660	16. 285	15. 953	17. 217	16. 661	15. 837	16. 393
7	15. 746	16. 592	16. 895	15. 903	16. 490	16. 063	17. 355	16. 840	16. 082	16. 660
8	15. 461	16. 306	16. 581	15. 619	16. 192	15. 768	17. 049	16. 540	15. 795	16. 365
9	15. 275	16. 113	16. 376	15. 427	15. 994	15. 578	16. 848	16. 342	15. 600	16. 164
10	15. 309	16. 162	16. 376	15. 417	16. 004	15. 621	16. 925	16. 377	15. 589	16. 150
11	15. 182	15. 979	16. 222	15. 217	15. 831	15. 514	16. 786	16. 217	15. 388	15. 932
12	14. 111	14. 849	15. 179	14. 171	14. 745	14. 382	15. 560	15. 136	14. 374	14. 959
13	13. 723	14. 403	14. 744	13. 720	14. 304	14. 002	15. 133	14. 700	13. 919	14. 480
#14	13. 664	14. 342	14. 678	13. 660	14. 242	13. 944	15. 071	14. 636	13. 857	14. 413
15	14. 093	14. 835	15. 197	14. 209	14. 759	14. 348	15. 507	15. 128	14. 416	15. 011
16	14. 215	14. 964	15. 311	14. 329	14. 882	214. 478	15. 648	15. 247	14. 529	15. 115
17	13. 707	14. 434	14. 743	13. 808	14. 343	13. 968	15. 103	14. 694	13. 994	14. 546
18	13. 672	14. 407	14. 692	13. 775	14. 306	13. 934	15. 074	14. 653	13. 955	14. 499
19	13. 397	14. 106	14. 410	13. 493	14. 018	13. 653	14. 762	14. 361	13. 675	14. 215
20	13. 286	13. 985	14. 292	13. 376	13. 899	13. 542	14. 639	14. 241	13. 557	14. 092
21	13. 096	13. 781	14. 091	13. 183	13. 700	13. 347	14. 426	14. 038	13. 363	13. 894
22	13. 012	13. 690	13. 986	13. 077	13. 599	13. 267	14. 345	13. 943	13. 254	13. 777
23	12. 850	13. 533	13. 815	12. 932	13. 440	13. 096	14. 167	13. 776	13. 107	13. 626
24	12. 820	13. 501	13. 780	12. 907	13. 410	13. 068	14. 133	13. 739	13. 078	13. 589
25	12. 628	13. 303	13. 571	12. 719	13. 212	12. 870	13. 922	13. 534	12. 886	13. 389
26	12. 492	13. 153	13. 411	12. 550	13. 051	12. 739	13. 785	13. 383	12. 716	13. 211
27	12. 534	13. 212	13. 443	12. 613	13. 104	12. 782	13. 837	13. 423	12. 770	13. 255
28	12. 406	13. 085	13. 297	12. 488	12. 971	12. 652	13. 702	13. 285	12. 640	13. 115
29	12. 433	13. 109	13. 320	12. 517	12. 998	12. 685	13. 731	13. 304	12. 663	13. 129
30	12. 250	12. 925	13. 116	12. 336	12. 808	12. 497	13. 534	13. 109	12. 478	12. 935
31	12. 023	12. 689	12. 871	12. 110	12. 571	12. 264	13. 284	12. 866	12. 249	12. 697
32	11. 739	12. 383	12. 579	11. 819	12. 274	11. 972	12. 965	12. 569	11. 962	12. 408
33	11. 672	12. 318	12. 494	11. 752	12. 202	11. 907	12. 898	12. 490	11. 888	12. 323

〈표 3〉 Residual Index for Artificial Data Set

5. 결 론

ELMS algorithm을 사용하여 regression outlier를 파악하는 방법은 LMS를 사용하여 이상치 인식을 하는 방법에 비하여 매우 높은 효율성을 갖고 있음을 알 수 있다. ELMS에서 이상치를 인식하는 방법은 초기 subset의 안정성에 의존율이 매우 낮으며, 또한 LMS 방법에서의 주된 문제점으로 나타나고 있는 연산횟수에 비하여 상대적으로 매우 작은 연산 처리 시간을 필요로 하고 있다. 이러한 점은 subset의 크기를 증가시키는 방법을 취하고 있기 때문이다. 따라서 subset의 크기를 정확히 예측하는 것은 매우 중요한 요인이 된다.

Atkinson & Mulira (1993)에서 제시하고 있는 stalactite plot를 사용한 subset의 크기 평가에서 rank$(X)=p$일때, $p+2$ 이상에서 매우 뛰어난 이상치 인식을 하고 있으며, 염준근, 박종구, 김종우(1995)에서는 $[\mathrm{Log}(n^3/p)]+p$를 중심으로 안정성을 보이고 있다고 지적하고 있다. 본 논문의 잔차 검토에 의한 결과는 $[\mathrm{Log}(n^3/p)]+p$ 보다 상대적으로 작은 크기의 표본 추출 $[\frac{5}{59}n+\frac{96}{59}]$ 과 $[\frac{5}{90}n+\frac{96}{90}]$ 사이에서 이상치의 인식은 더 뛰어남을 찾을 수 있었다. median을 이용한 표본의 크기 증가는 앞에서 지적한 적정 수준에 도달하면서 뚜렷한 이상치 인식을 보이고, 표본의 크기가 커져감에 따라 점차 이상치 인식율은 서서히 낮아짐을 얻을 수 이었다.

최적 표본 크기의 유도는 ELMS algorithm의 이상치 인식에 대한 수렴성을 보다 안정적으로 유도하고, 빠른 자료 처리 시간을 갖게 한다.

6. 참 고 문 헌

[1] 송문섭(1995), "로버스트 추정법", Tutorial session, 춘계학술발표회, 한국통계학회.

[2] 염준근(1993), 선형회귀분석, 자유아카데미.

[3] 염준근, 박종구, 김종우(1995), "다변량 자료에서 다수 이상치 인식의 절차", 품질경영학회지, 출간중.

[4] Atkinson, A. C., and Mulira, H. -M. (1993), "The Stalactite Plot for the Detection of Multivariate Outliers," *Statistics and Computing* 3, 27-35.

[5] Atkinson, A. C. (1994), "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association*, 89, No 428, 1329-1339.

〔6〕 Hadi, A. (1992), "Identifying Multiple Outliers in Multivariate Data," *Journal of the Royal Statistical Society Series-B*, 54, No 3, 761-771.

〔7〕 Hadi, A., and Simonoff, J. S. (1993), "Procedures for the Identifying of Multiple Outliers in Linear Models," *Journal of the American Statistical Association*, 88, No 424, 1264-1272.

〔8〕 Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, E. (1986), *Robust Statistics : The Approach Based on Influence Functions*, John Wiley & Sons, New York.

〔9〕 Hawkins, D. M. (1993), "The Accuracy of Elemental Set Approximations for Regression," *Journal of the American Statistical Association*, 88, No 422, 580-589.

〔10〕 Huber, P. J. (1973), "Robust Regression : Asymptotics, conjectures and Monte Carlo," *The Annals of Statistics*, 1, 799-821.

〔11〕 Kalos, M. H., and Whitlock, P. A. (1989), *Monte Carlo Methods*, Vol I : Basics, New York : John Wiley & Sons.

〔12〕 Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, No 388, 871-880.

〔13〕 Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York : John Wiley.

〔14〕 Rousseeuw, P. J., and van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, No 411, 633-639.

〔15〕 Woodruff, D. L., and Rocke, D. M. (1993), "Heuristic Search Algorithms for the Minimum Volume Ellipsoid," *Journal of Computational and Graphical Statistics*, 2, 69-95.

〔16〕 Woodruff, D. L., and Rocke, D. M. (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimation," *Journal of the American Statistical Association*, 89, No 427, 888-896.

〔17〕 Yohai, V. J. (1987), "High Breakdown Points and High Efficient Robust Estimates for Regression," *The Annals of Statistics*, 15, 642-656.