Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.3 pp. 471~479

Moving Averages Based on Robust Statistical Analysis[1)]

Ro Jin Pak[2)]

Abstract

Moving averages are the most popular statistics in analyzing time-series data like stock indices. However, moving averages are quite sensitive to unusual observations. In other words, they are not robust against unusual observations. We introduce the moving averages in terms of an M-estimator, and show how we can take advantages of using the proposed moving averages in fitting the data more than usual moving averages.

Keywords: Moving average, Robust Estimator, Running Median

1. 서론

주가와 같이 시간에 따라 발생하는 자료를 다룰 때, 이동 평균이 유용하게 사용되고 있다. 하지만, 로버스트 통계학의 입장에서 볼 때 평균은 이상치에 영향을 많이 받는 통계량으로써 변화가 심한 시계열 자료를 다룰 때 다소 왜곡된 정보를 제공할 수 있다고 생각된다. 이를 보완하기 위해 중위수를 이용한 이동중위수(running median)가 한 가지 대안으로 사용되나, 분산이 커지는 약점을 갖고 있다. 또 다른 대안으로서 Huber의 통계량으로 대표되는 로버스트 통계량을 이용할 수 있겠으나, 어떤 관측치가 이상치인지, 아니면 구조적인 변화의 결과인지 판단하기 어려운 상황에서 무조건적인 사용은 정보의 왜곡을 초래할 수 있다. 본 논문에서는 Huber의 통계량과 일반적인 평균의 적절한 조합을 통해 이상치의 영향을 덜 받으면서 중위수를 사용할 때처럼 분산이 커지지 않은 로버스트 이동평균을 제시하고 그 효용성을 모의실험과 실례를 통해 규명하고자 한다.

---

1) This research was conducted by the research fund of Dankook University in 2003

2) Associate Professor, Dankook University, Division of Information and Computer Sciences
E-mail: rjpak@dankook.ac.kr

2. 본론

2.1 기존의 방법

주어진 구간에서 n개의 관측치($X_1, X_2, \ldots, X_n$)가 주어졌을 때, 일반적인 이동평균은 μ에 대한 아래식의 해로 정의되는데,

$$\sum_{i=1}^{n}(X_i-\mu)=0,$$

이미 잘 알고 있는 것처럼 이렇게 구한 이동평균은 이상치의 영향을 직접적으로 받는다. 그래서, 이상치의 영향을 제한하고자 아래와 같은 로버스트 평균을 구하여 사용하곤 한다. Huber(1981)에 의해 제안된 위치모수의 추정량은 아래와 같은 목적함수를

$$Q(T)=\sum_{i=1}^{n}\rho\left(\frac{X_i-\mu}{\sigma}\right)$$

μ에 대하여 최소로 하는 통계량, 또는 음의 방정식

$$\sum_{i=1}^{n}\psi\left(\frac{X_i-\mu}{\sigma}\right)=0 \quad \text{여기서 } \psi=\rho',$$

의 μ에 대한 해로서 정의된다.

대표적인 M-추정함수로서 Huber(1981)는 다음과 같은 ψ-함수

$$\psi(r)=\begin{cases} r, & |r|\le b \\ b\cdot sign(r), & |r|>b \end{cases}$$

를 이용하여 위치모수에 대한 로버스트 추정량을 구하는 방법을 제시했고, 널리 사용되고 있다. ψ-함수는 b로 표시된 조율상수에 의해 형태가 결정되는데, 정규분포 가정하에서 95%의 점근효율을 갖도록 b를 1.345로 정하는 방법이 가장 널리 사용되고 있다. 즉, 표준화된 관측치가 ± 1.345를 벗어난다면 그 자료의 값을 1.345로 고정하여 그 영향치를 제어하게된다. 만일 b가 ∞가 되면, Huber의 추정량은 일반적인 최소제곱추정법, 또는 정규분포 가정에서의 최우추정법의 결과인 산술평균이 된다. 물론 1.345외 다른 조율상수의 값이 사용되지만, 조율상수 선택에 대한 광범위한 논의는 본 논문의 주안점이 아닌 고로 다음 기회에 논의하기로 하고 생략하기를 원한다. 본 논문에서 이하 모든 계산에서 1.345를 조율상수로 사용하겠다.

2.2 새로운 방법

Huber (1981)가 제안한 위의 방법은 샘플의 수가 고정된 경우, 위치모수에 대한 로버스트 추정량을 구하기 위해 사용되는데, 우리가 지금 관심을 갖고 있는 시간에 따라 샘플의 내용이 변하는 경우, 새로이 관측된 자료가 이상치라면 큰 문제가 없지만 만일 새로운 과정의 초기 자료라면 단순히 위와 같은 방법으로 그 자료의 영향력을 배제할 수는 없다.

$$\sum_{i=1}^{n} w_i r_i + (1 - w_i)\psi(r_i) = 0, \quad (1)$$

제안 1. 가중치 w_i들이 주어졌을 때, 아래의 식(1)의 해로써 추정량을 삼을 것을 제안한다.

여기서 $r_i = (X_i - \mu)/\sigma$.

식(1)은 자료의 본질적인 특성(r_i)과 로버스트 특성($\psi(r_i)$)을 동시에 고려하기 위해 가중치를 이용하여 추정량을 구할 것을 제안한다. 즉, 일반적인 이동평균을 구하는 식과 로버스트 평균을 구하는 식을 절충하여 보다 로버스트 하면서 단순한 M-추정의 방법보다는 분산을 작게 갖는 이동평균을 구해 보고자 한다. 물론, 가중치 w_i가 모두 0으로 지정되면 일반적인 Huber의 M-추정식이 될 것이고, 모두 1로 지정하면 일반적인 평균을 구하는 최소제곱법에 기인한 추정식이 될 것이다. 또는, 모두 0으로 지정하고 마지막 가중치 w_n을 1로 지정한다면 마지막 자료의 영향력을 계산에 받아드리고 오래된 자료의 영향력은 제어하는 상황에서의 추정량을 계산할 수 있게된다. 본 논문에서는 이동중위수 계산 시 사용되는 방법 중에 이항정리를 이용한 아래와 같은 가중법을 사용하여 보았다(Tukey, 1977). 예를 들어, 평균 계산 구간 너비가 4라면 $w_1 = 1/8$, $w_2 = 3/8$, $w_3 = 3/8$, $w_4 = 1/8$ 와 같이 중간에 있는 자료의 영향력을 오래된 자료와 새로운 자료의 영향력보다 상대적으로 많이 인정하는 방법을 의미한다. 이를 수식으로 표현하면, 구간 너비가 n 이라고 할 때,

$$w_i = \binom{n-1}{i-1} / 2^{n-1}, \quad i = 1, \cdots, n \quad (2)$$

으로 된다.

실제로 추정량의 계산은 뉴튼-랍슨의 수치 해석적 방법에 따라 수행된다. 즉, 어떤 미분 가능하고 연속인 함수 $f(x)$의 $f(x) = 0$이 되는 해를 구하는 뉴튼-랍슨의 수치 해석 방법에 의하면, 주어진 초기치 x_0를 이용하여 그 첫 번째 해

$$x_1 = x_0 - f(x_0)/f'(x_0)$$

가 주어지는데, 식(1)에 위의 방법을 적용한다. 그 과정에서 Hampel외 3인(1986, p106)이 설명한 방법을 기본으로 변수 x를 표준화하여 아래의 방법을 이용하여 계산할 수 있다.

여기서, T_0와 S_0는 μ와 σ에 대한 초기치이다. 일반적으로, T_0에 대하여는 중위수를,

$$T_n = T_0 + \frac{\frac{1}{n} S_0 \sum w_i \left(\frac{x_i - T_0}{S_0}\right) + (1 - w_i)\psi\left(\frac{x_i - T_0}{S_0}\right)}{\frac{1}{n}\sum w_i + (1 - w_i)\psi'\left(\frac{x_i - T_0}{S_0}\right)}$$

S_0에 대하여는 MAD (median absolute deviance)를 0.6745로 나눈 값을 사용한다 (Hampel외 3인 (1986)).

2.3 모의실험

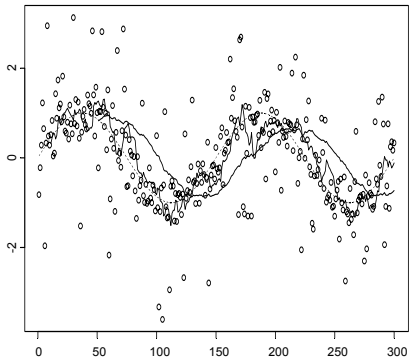
모의실험은 Sin함수를 기반으로

실험 1 (오염도 0%): 일양분포(-0.5,0.5)에서 생성된 오류를 추가한 크기가 300개의 자료를 300번 생성하여 구간너비가 10과 50인 경우에 대하여 일반적인 이동평균, 이동중위수, 로버스트 이동평균을 구하였다.

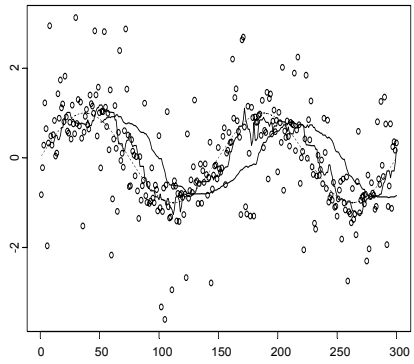
실험 2 (오염도 10%): 실험 1에서 생성한 데이터에 일양분포 (-3,3)에서 생성된 오류를 10% 섞어 얻은 오염된 자료 300세트에 대하여도 구간 너비가 10과 50일 때에 따라 이동평균, 이동중위수, 로버스트 이동평균을 구하였다.

실험 3 (오염도 20%): 실험 1에서 생성한 데이터에 일양분포 (-3,3)에서 생성된 오류를 20% 섞어 얻은 오염된 자료 300세트에 대하여도 구간 너비가 10과 50일 때에 따라 이동평균, 이동중위수, 로버스트 이동평균을 구하였다.

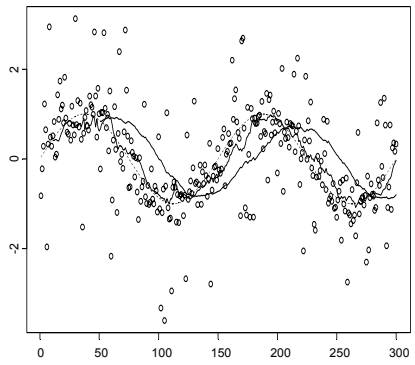
논문의 지면상 세 가지 주어진 상황에 따른 모의 실험 결과 중 오염도가 20%일 경우의 첫 번째 시뮬레이션 결과를 <그림 2-1, -2, -3>에 그려보았다. <그림 2-4>에서는 주어진 구간 너비와 오염도에 따라 이동평균, 이동중위수, 로버스트 이동평균의 300회의 모의 실험에 따른 평균제곱오차를 구하여 상자그래프로 그려보았다. 종합적으로 보면, 로버스트 이동평균의 평균제곱오차가 평균적으로 낮은 값을 갖고 그 분산도 다른 것들에 비해 상대적으로 적음을 알 수 있고, 즉, 본 논문에서 제안하고 있는 로버스트 이동평균이 다른 두 가지 추정량들에 비해 상대적으로 안정적이고 적합성도 뛰어남을 관찰할 수 있다. 그림들에서 보듯이 로버스트 이동평균은 특별히 구간 너비가 짧은 경우 (n=10), 이상치들의 영향을 덜 받아 가정된 모델인 Sin 함수(점선)에 대하여 가장 가깝게 부드럽게 움직임을 볼 수 있다.



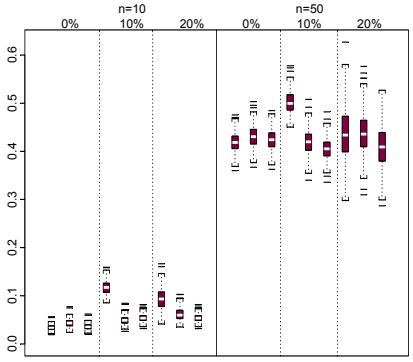
<그림 2-1> 이동평균, 오염도 20%
n=10과 50



<그림 2-2> 이동중위수;오염도 20%
n=10과 50



<그림 2-3> 로버스트 이동평균; 오염도 20%, n=10과 50

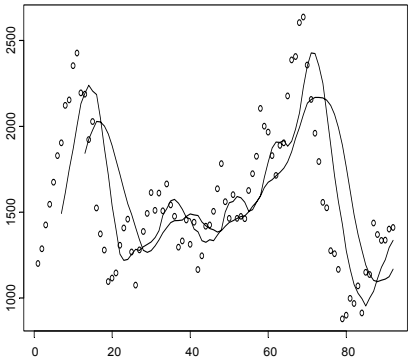


<그림 2-4> 평균제곱오차의 상자도표: 각 섹션에서 이동평균, 이동중위수, 로버스트 이동평균 순서로 그려짐

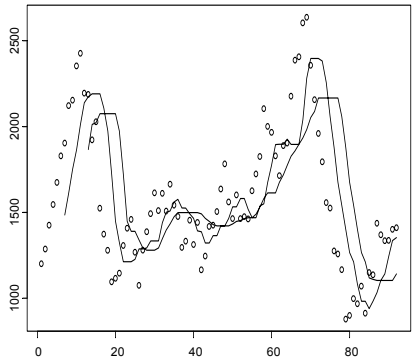
2.4 예를 이용한 분석: 반도체 판매량

1992년 1월부터 2002년 8월까지 한국무역협회가 제공하는 우리나라 반도체 수출액(단위 백만불)에 대하여 구간 너비 6개월과 12개월에 대하여 이동평균, 이동중위수, 로버스트 이동평균을 구하여 그림을 그려보았다. <그림 2-5,-6,-7>에서 보듯이 이동중위수는 변곡점 또는 고점, 저점에서 부드럽지 못함을 볼 수 있다. 이동평균과 로버스트 이동평균은 비슷한 모양을 갖고 있으나, 역시 변곡 또는 고점, 저점에서 로버스트 이동평균이 다소 부드럽게 움직이는 것을 볼 수 있다. 각 그림에 표기된 SSE의 크기나 실제로 잔차들의 상자그림 <그림 2-8>에서

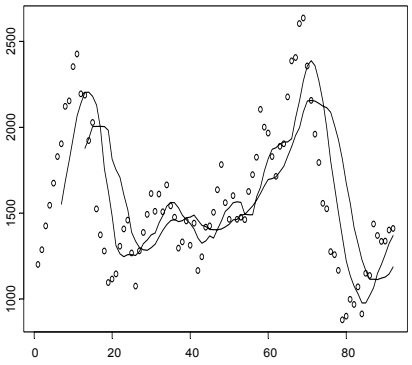
보듯이 로버스트 이동평균이 가장 데이터에 가장 잘 적합되고 있음을 알 수 있다.



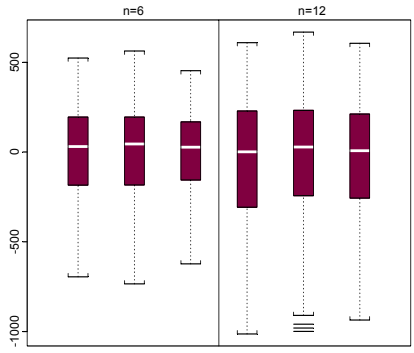
<그림 2-5>이동평균; 6개월,12개월
SSE=8320177(6개월),12528295(12개월)



<그림 2-6> 이동중위수
SSE=8474082(6개월),13717136(12개월)



<그림 2-7>로버스트 이동평균
SSE=6274331(6개월),10996107(12개월)

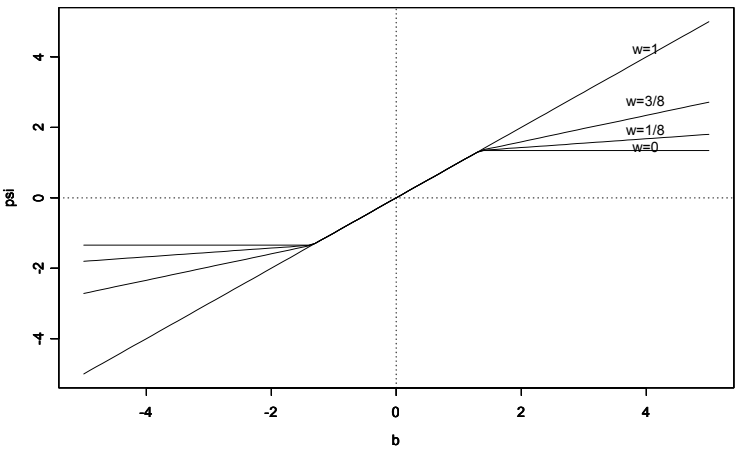


<그림 2-8> 반도체 수출 데이터의 잔차들의 상자그림; 구간너비 6개월, 12개월에 대하여 이동평균(첫 번째 상자그림), 이동중위수(두 번째), 로버스트 이동평균(세 번째).

2.5 관련 이론

<그림 2-9>에서 보듯이, 위에서 제시한 가중치를 사용한 추정함수들이 일반적인 평균을 주는 추정함수($w=1$)와 Huber형태의 추정함수($w=0$) 사이에 존재함으로써 알 수 있다. 따라서, 새로이 제안 추정함수들이 주어진 임계치를 초과하는 자료의 영

향력을 일반적인 평균의 추정함수와 달리 어느 정도 제어함으로 로버스트 성질이 보장되고, 또한 점근분산에 있어서는 Huber의 추정함수 보다 평균을 추정하는 함수에 가까움으로써 향상된 결과를 줄 것이다.



<그림 2-9> 주어진 가중치들에 따른 추정함수의 예

또한, 새로운 추정함수에 근거하여 점근분산을 수식적으로 전개한다면,

(1)에 근거하여 $\eta(w,r)=wr+(1-w)\psi(r)=w(r-\psi(r))+\psi(r)$라고 정의한 후, Maronna and Yohai(1981)와 Hample외 3인(1986)이 위치모수와 회귀계수 추정시 제시한 일반적인 조건들(regularity conditions)을 에 적용하여 만족한다면, 올바른(true) 모델 하에서의 점근분산은

$$V(\eta,F)=M^{-1}(\eta,F)\cdot Q(\eta,F)\cdot M^{-1}(\eta,F),$$

여기서,

$M(\eta,F) = E\eta' = E[1-\psi'(r)]\cdot Ew + E\psi'(r);$

$Q(\eta,F) = E\eta\eta^{T} = E[r-\psi(r)]^2\cdot Eww^{T} + E[(r-\psi(r))\psi(r)]\cdot 2Ew + E\psi(r)\psi(r)^{T}$

가 된다.

구간의 너비가 주어지고, 여러 가지 편이상 가중치들이 상수로 주어진다면, 표본점근분산은 아래의 M과 Q를 이용하여 추정된다.

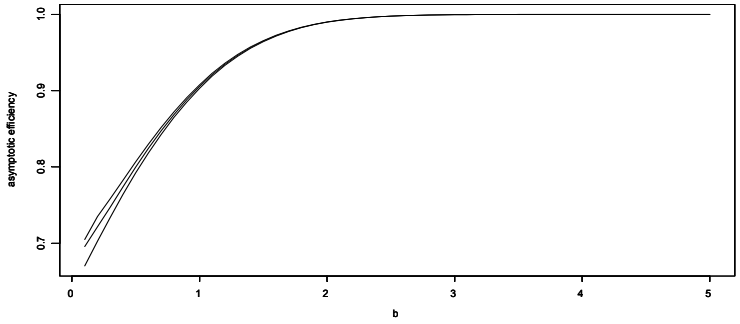
$$V(\psi,F)=M^{-1}(\psi,F)\cdot Q(\psi,F)\cdot M^{-1}(\psi,F),$$

여기서

$$M(\psi,F)=\frac{1}{n}\sum_i w_i\int(1-\psi'(r))dF(r)+\int\psi'(r)dF(r)$$

$$Q(\psi,F)=\frac{1}{n}\sum_i w^2{}_i\int(r-\psi(r))^2\,dF(r)+\frac{2}{n}\sum_i w_i(1-w_i)\int r\psi(r)dF(r)+\int\psi^2(r)\,dF(r)$$

만일 $\psi(\cdot)$는 Huber의 식을 사용하고, 확률함수는 표준정규분포로, 가중치는 위에서 정의된 식(2)대로 주어진 경우, 점근분산의 역수인 점근효율(asymptotic efficiency)을 조율상수 b에 대하여 <그림 2-10>과 같이 그릴 수 있다. 새로운 이동평균의 점근효율이 기존의 Huber에 기초한 이동평균 보다 큼을 알 수 있다. 모든 경우에 공히 조율상수가 증가할수록 점근효율이 증가하여 1에 접근함을 알 수 있다. 따라서, 조율상수 b는 원하는 점근효율을 만족하도록 정할 수 있겠다.



<그림 2-10> 제안된 이동평균의 점근효율의 변화; n=50(위), 100(중간), 기존의 Huber(아래)

3. 결론

기존의 이동평균에 비해 로버스트하고 분산도 적당하게 작은 새로운 형태의 이동평균을 제시했다. 간단한 이론 전개와 자료분석을 통해 기존의 이동평균, 이동중위수 보다 바람직한 성질들을 갖고 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*, John Wiley & Sons, New York.
2. Huber, P. J. (1981), *Robust Statistics,* John Wiley & Sons, New York.
3. Maronna, R. A. and Yohai, V. J. (1981), Asymptotic behavior of general M-estimates for regression and scale with random carriers, Zeit. Wahrsch. Verw. Gebiete, 58, 7-20.

4. Tukey, J. W. (1977), *Explorative Data Analysis,* Addison-Wesley, New York.
5. http://kotis.kita.net/, 한국무역협회, 종합무역정보, 품목별 수출입정보

[2003년 4월 접수, 2003년 7월 채택]