응용통계연구 제 19권 2호, 2006년, pp. 305–317

로버스트 추정법을 이용한 자기상관회귀모형에서의 특이치 검출

이동희[1] 박유성[2] 김기환[3]

요 약

시계열 자료에서의 특이치, 특히 이 가운데 가법적 특이치가 모형의 식별, 모수의 추정 및 예측과 관련된 분석 전과정을 왜곡하는 것은 잘 알려져 있다. 그러나 특이치가 다수 발생하는 경우, 특히 연속적으로 집단을 이루어 발생할 때 대부분 특이치 검출방법은 가면화효과와 수렁화효과때문에 이들을 정확히 판별하지 못한다. 본 논문에서는 p차 자기상관회귀모형에 대한 고붕괴점 회귀추정량을 이용한 양방향 로버스트 필터방법을 제안했다. 실제 사례와 모의실험을 통해 제안한 방법이 매우 정확하게 시계열 자료에 포함된 특이치들을 검출하고 있음을 확인할 수 있다.

주요용어: 가법적 특이치, 고붕괴점추정량, 시간 가역성, 양방향 로버스트 필터, 자기상관회귀모형, 특이치군, 혁신적 특이치

1. 서론

다음과 같은 p차 자기상관회귀모형(autoregressive model of order p, AR(p))을 고려해보자.

$$z_t = \phi_0 + \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + a_t, \tag{1.1}$$

여기서 $t = 1, \cdots, T$이며, 오차항 a_t는 i.i.d 인 백색잡음 과정을 따르고, 이와 더불어 모수 ϕ_i $(i = 1, \cdots, p)$는 모든 $|\omega| < 1$인 복소수 ω에 대하여

$$\sum_{i=1}^{p} \phi_i \omega^i \neq 1$$

인 정상성(stationary)을 가정한다. Fox (1972)는 시계열 자료에서의 특이치를 두 가지 유형으로 구분하여 정의하였는데, 만일 관찰된 계열 y_t가 식 (1.2)를 만족한다면 이를 크기가 δ인 시점 t에서의 가법적 특이치 (additive outlier, AO)라 하며,

$$y_t = \delta I_t(S) + z_t, \; t = 1, \cdots, n \tag{1.2}$$

1) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 통계연구소, 연구조교수
E-mail: ld0351@korea.ac.kr
2) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 통계학과, 교수
E-mail: yspark@korea.ac.kr
3) (339-700) 충청남도 연기군 조치원읍 서창리 208, 고려대학교 자연과학대학 정보통계학과, 조교수
E-mail: korpen@korea.ac.kr

반면 시점 t에서 크기가 δ인 혁신적 특이치 (innovative outlier, IO)는 식 (1.3)과 같이 잡음 과정에 직접 영향을 주는 경우를 의미한다.

$$y_t = \phi(B)^{-1}\delta I_t(S) + z_t, \tag{1.3}$$

여기서 $I_t(S)$는 다음과 같은 지시함수(indicator function)이다.

$$I_t(S) = \begin{cases} 1, \text{ if } t = S, \\ 0, \text{ otherwise.} \end{cases}$$

Denby와 Martin(1979) 및 Chang 등(1988)의 연구에 의하면 AO는 모수 추정에 심각한 영향을 초래하지만, IO는 크게 영향을 주지 않는 것으로 알려져 있다. 본 연구에서 관심을 가지고 살펴보고자 하는 것은 추정과정에 많은 영향을 끼치는 AO들이 다수 발생하는 경우이다. 특히 이들 특이치들이 연속적으로 발생하게 되면 가면화효과(masking effect)와 수렁화효과(swamping effect)에 의하여 이제까지 알려진 시계열 자료에 대한 특이치 검출방법은 정확한 결과를 제공하지 못한다(Tsay 등, 2000; Justel 등, 2001). 이러한 왜곡 현상을 다음의 간단한 모의실험을 통하여 자세히 살펴보도록 하자.

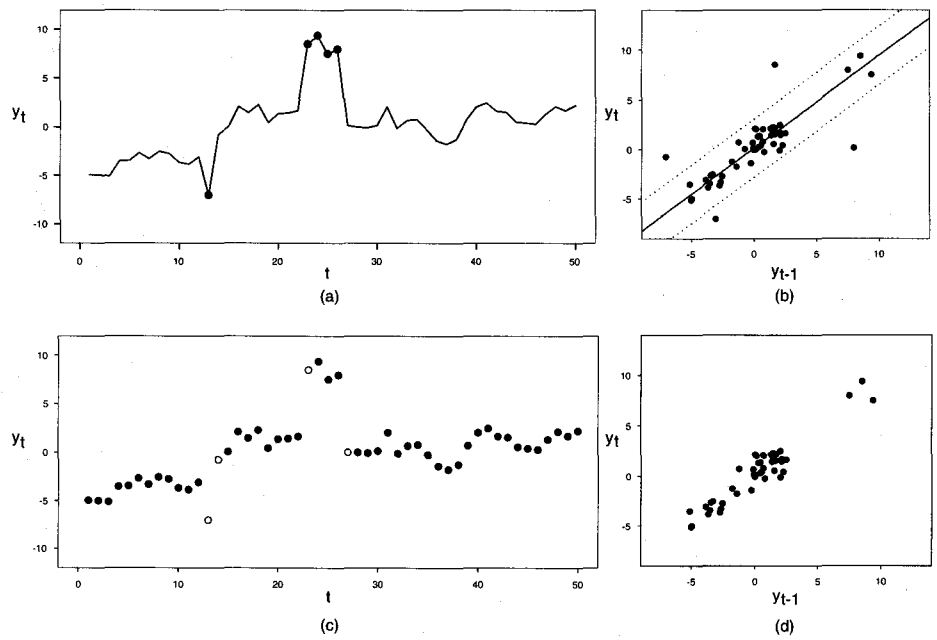

그림 1.1: (a) 특이치가 포함된 원계열, (b) S-추정량을 이용한 회귀적합선 (점선은 $\hat{y}_t \pm 2.5\hat{\sigma}$, (c) 검출된 특이치를 흰점으로 표현한 시도표, (d) 특이치가 제거된 자료를 회귀평면으로 표현

그림 1.1은 식 (1.4)로부터 생성된 오염된 시계열 자료에 대해서 로버스트 추정방법에 기반한 특이치 검출 과정이 어떻게 왜곡되는지 살펴보기 위하여 고붕괴점 로버스트 회귀 추정량(high breakdown robust regression estimator) 가운데 대표적인 S-추정방법을 이용한 특이치의 판별 과정과 그 결과를 나타낸 것이다.

$$y_t = .9\, y_{t-1} + a_t,\ t = 1, \cdots, 50, \tag{1.4}$$

여기서 a_t는 $N(0,1)$을 따르도록 하며, 계열 가운데 시점 13에서는 크기가 -6인 AO를, 시점 23부터 26까지 연속적인 AO들로 구성된 특이치군을 크기가 각각 7, 7, 5, 6이 되도록 설정하였다 (그림 1.1의 (a)). 그림 1.1의 (b)는 (y_{t-1}, y_t)의 쌍으로 회귀평면에 나타낸 것이며, 그림 1.1의 (c)는 특이치 검출 결과를 시도표로 다시 환원하여 표현한 것이다. 여기서 (y_{12}, y_{13}), (y_{13}, y_{14})는 y_{13}만이 특이치임에도 불구하고 수렁화효과에 의하여 모두 왼쪽 하단에 특이치로 나타나 결국 y_{14}까지 특이치로 분류되는 오류가 나타나며, (y_{23}, y_{24}), (y_{24}, y_{25}), (y_{25}, y_{26})은 모두 특이치임에도 불구하고 회귀평면 상에서 우측 상단에 좋은 지렛값 (good leverage point)으로 나타나는 가면화효과에 의해 결국 y_{24}, y_{25}, y_{26}은 특이치로 검출되지 못하고 있다. y_{27}의 경우 역시 특이치인 y_{26}에 의한 수렁화효과에 의해 (y_{26}, y_{27})은 우측 하단의 특이치로 분류된다. 결국 그림 1.1의 (d)에서 보듯이 시점 23에서 26까지 발생한 특이치군 가운데 최초 특이치를 제외하고는 가면화효과에 의해 나머지 특이치를 검출하지 못하고 있다.

본 논문의 목적은 기존의 시계열 자료에 대한 특이치 검출방법이 직면한 가면화효과와 수렁화효과에 의한 왜곡을 보완함으로써 정확한 특이치 검출이 가능한 방법을 제안하는 것이다. 우리가 제안한 방법의 평가를 위하여 얼마만큼 정확하게 자료에 포함되어 있는 특이치들을 검출하고 있는지 모의실험과 실제자료를 이용하여 살펴보고자 하며, 특히 다수의 특이치가 포함된 자료에 대해서 얼마나 정확한 결과를 우리에게 제시할 수 있는지 살펴볼 것이다.

2장에서는 본 연구에서 새롭게 제안한 시계열 자료에서의 특이치 검출방법의 내용을 살펴보고, 3장에서는 실제 사례와 모의실험을 통하여 제안한 방법을 평가해보고자 한다. 그리고 마지막 4장은 연구의 결론이다.

2. 양방향 로버스트 필터를 이용한 특이치 검출

자기상관회귀모형에서의 일반적인 특이치 검출 방법은 회귀모형에서와 같이 잔차(residual)에 기반한 방법을 사용한다. 추정된 모수 $\hat{\phi}$들을 이용하여 이에 대응하는 AR-잔차

$$r_t = r_t(\hat{\boldsymbol{\phi}}) = y_t - \hat{\phi}_0 - \sum_{i=1}^{p} \hat{\phi}_i y_{t-i},\ \ t = p+1, \cdots, T$$

를 얻은 후, 모형 (1.1)의 오차의 표준편차에 대한 추정량 $\hat{\sigma}_a$를 이용하여 특이치 기각규칙(rejection rule)을 적용한다. 여기서 $\hat{\sigma}_a$은 로버스트한 방법을 통해 얻어지며, 대표적인 추정치로써 중위절대편차(median absolute deviation, MAD)를 이용한다. 이로부터 표준화 잔

차 $r_t/\hat{\sigma}$가 어떤 상수 c보다 큰 경우만 해당 관찰치는 특이치로 구분된다. 그러나 이러한 방법은 다음의 두 가지를 고려해야 한다(Bianco 등, 2001).

- 잔차를 얻기위한 적합값은 모수의 추정값에 의존하므로, 이들 모수에 대한 추정 방법이 특이치에 의해 영향을 받는다면 특이치에 대한 판별 결과 역시 영향을 받게 된다.
- 시계열 자료의 특성에 따라 예측 및 적합값은 이전 시점의 관찰치로 구성되기때문에, 모수에 대한 추정값 뿐 아니라 이를 구성하는 관찰값의 특이치 존재 역시 특이치 판별에 큰 영향을 끼칠 수 있다.

첫번째에 제기된 특이치의 영향은 로버스트 추정량을 사용함으로써 극복할 수 있다. Rousseeuw와 Leroy (1987, Chap.7) 및 Martin과 Yohai (1991)는 각각 고붕괴점(high breakdown point)을 갖는 로버스트 추정량인 최소중위제곱추정량(least median of squares estimator, LMSE)과 S-추정량(S-estimator, SE)의 사용을 제안했으며, 이들 추정량이 오염된 시계열 자료에 대해서 매우 효과적임을 보여준 바 있다. 특히 Terpstra 등 (2001) 및 Meintanis와 Donatos(1999)는 여러가지 AO와 IO들로 오염된 시계열 자료에 대한 추정방법들간의 비교를 위한 모의실험에서 고붕괴점을 갖는 로버스트 추정량들이 다른 추정방법들에 비하여 낮은 평균제곱오차와 편향을 유지함으로써 고붕괴점 회귀추정량이 자기상관회귀모형에서도 역시 로버스트함을 보여주었다. 여기서 붕괴점 (breakdown point)이란 간단히 말하여 주어진 자료로부터 오염된 부분이 추정량을 왜곡시키기 시작하는 비율을 의미한다(Donoho와 Huber, 1983). 이러한 의미에서 최소제곱추정량(least squares estimator, LSE)는, 하나의 관찰값이 오염되어 있을 경우에도 영향을 받기때문에 LSE의 붕괴점은 $1/T$, 즉 근사적으로 0이 된다. 특히 추정량의 붕괴점이 50%가 되는 경우, 즉 자료의 절반 가까이 오염되었을 경우에도 영향을 받지 않는 추정량들을 일컬어 고붕괴점 추정량(high breakdown estimator)이라 한다.

두번째로 고려해야 할 관찰치 가운데 존재하는 특이치의 영향력은 좀 더 복잡한 양상을 보인다. 특히 특이치가 일정 시점동안 연속적으로 존재하는 특이치군(outlying patches)의 경우에는 가면화효과와 수렁화효과가 시점에 따라 지속적으로 발생함으로써 정확한 특이치의 검출을 어렵게 한다.

우리가 제안하는 특이치 검출방법은 고붕괴점 회귀추정방법이 갖는 다수의 특이치에 대한 로버스트성과 정상 시계열이 갖는 시간 가역성(time reversibility) 특징을 이용한 것이다. 다수의 오염된 시계열 자료에 대해서 로버스트성을 유지하는 고붕괴점 추정량을 이용하여 특이치 검출 과정에서 발생하는 가면화 및 수렁화효과에 의한 왜곡을 방지하기 위한 로버스트 필터 방법을 전방 및 후방 예측과정에서 함께 적용하였다. 즉 고붕괴점 추정량에 기반한 예측값과 실제 관찰값간의 차이에 따라 로버스트 필터를 이용한 정화된 값으로 해당 시점의 관찰값을 대체함으로써, 이전 시점 혹은 이후 시점의 관찰값에 대한 특이치 판별 과정에서 나타나는 수렁화효과와 가면화효과에 의한 왜곡을 줄이고자 한다. 다음은 정상 시계열 자료에서의 특이치 검출을 위한 양방향 로버스트 필터 방법(dual robust filtering) 절차이다.

단계 1 AR(p)모형에 대하여 LMSE 혹은 S-추정량과 같은 로버스트 추정방법에 기반한 자기

상관회귀계수 $\phi_0, \phi_1, \cdots, \phi_p$들을 추정한 후, 이와 더불어 a_t의 표준편차에 대한 로버스트 추정량인 MAD를 사용하여 식 (2.1)과 같이 $\hat{\sigma}_a$를 추정한다.

$$\hat{\sigma}_a = (.6745)^{-1} \text{ median of } |r_t| \tag{2.1}$$

여기서 r_t는 자기상관회귀잔차이다.

단계 2 (전방 로버스트 필터 방법) $t = p+1$시점부터 각 관찰값에 대한 적합값을 다음과 같이 전방예측값에 기반한 로버스트 필터를 이용한 정화된 계열값을 이용하여 얻는다. 시점 t의 적합값 $\hat{y}_t^f$는 다음과 같이 정의된다.

$$\hat{y}_t^f = \hat{\phi}_0 + \hat{\phi}_1 \tilde{y}_{t-1} + \cdots \hat{\phi}_p \tilde{y}_{t-p},$$

여기서 $\tilde{y}_{t-k}, (k = 1, \cdots, p)$는 로버스트 필터 w_{t-k}에 의해 정화된 계열값으로써

$$\tilde{y}_{t-k} = w_{t-k} y_{t-k} + (1 - w_{t-k}) \hat{y}_{t-k}^f \tag{2.2}$$

이며, 로버스트 필터

$$w_{t-k} = \begin{cases} 1, & \text{if } \left| \dfrac{y_{t-k} - \hat{y}_{t-k}^f}{\hat{\sigma}_a} \right| < c, \\ 0, & \text{otherwise .} \end{cases} \tag{2.3}$$

와 같이 정의된다.

단계 3 (후방 로버스트 필터 방법) 단계 2의 과정을 후방에 대하여 동일하게 적용한도록 한다. 즉 앞서와는 반대로 $t = T - p - 1$ 시점부터 시차를 줄여나가며 고붕괴점 추정법에 의해서 단계 1에서 추정된 모수들을 이용한 시점 t의 정화된 계열값 적합값 $\hat{y}_t^b$를 다음과 같이 얻는다.

$$\hat{y}_t^b = \hat{\phi}_0 + \hat{\phi}_1 \tilde{y}_{t+1} + \cdots \hat{\phi}_p \tilde{y}_{t+p},$$

여기서 $\tilde{y}_{t+k}, (k = 1, \cdots, p)$는 로버스트 필터 w_{t+k}에 의해 정화된 계열값으로써

$$\tilde{y}_{t+k} = w_{t+k} y_{t+k} + (1 - w_{t+k}) \hat{y}_{t+k}^b \tag{2.4}$$

이며, 로버스트 필터

$$w_{t+k} = \begin{cases} 1, & \text{if } \left| \dfrac{y_{t+k} - \hat{y}_{t+k}^b}{\hat{\sigma}_a} \right| < c, \\ 0, & \text{otherwise .} \end{cases} \tag{2.5}$$

와 같이 정의된다.

단계 4 전방예측 및 후방예측에 기반한 로버스트 필터에 의해 정화된 관찰값에 기반한 예측값 $\hat{y}_t^f$, $\hat{y}_t^b$와 관찰값 y_t를 이용한

$$r_t^f = \frac{y_t - \hat{y}_t^f}{\hat{\sigma}_a},\ r_t^b = \frac{y_t - \hat{y}_t^b}{\hat{\sigma}_a}$$

를 이용하여 특이치 여부를 동시에 판별한다. 즉 r_t^f와 r_t^b의 절대값이 c보다 동시에 모두 크다면 해당 시점의 관찰치를 특이치로 규정하며, 반면 어느 한 값에서라도 기준값 c보다 작다면 정상 관찰치로 고려한다.

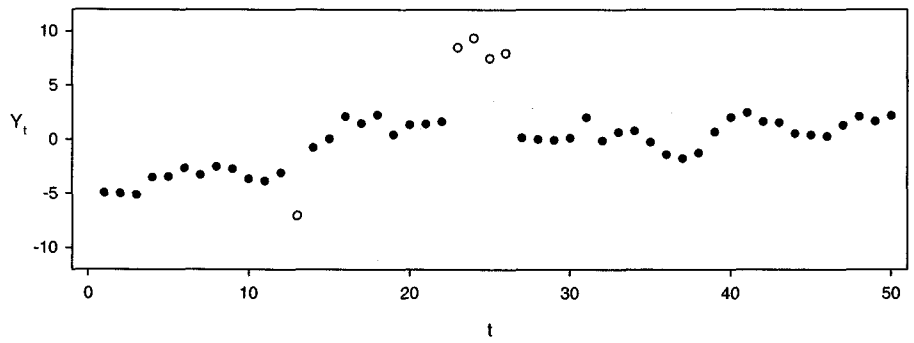

그림 2.1: 로버스트 필터 방법을 사용한 검출된 특이치를 흰점으로 표현한 시도표

이러한 과정에서 사용된 기준값 c는 Chang 등(1988)의 제안대로 다음과 같이 표본 크기 T 에 따라 결정하도록 한다.

$$c = \begin{cases} 3 & \text{if } T \leq 200 \\ 3.5 & \text{if } 200 < T \leq 500 \\ 4 & \text{if } T > 500 \end{cases}$$

그림 2.1은 앞의 그림 1.1 자료에 대하여 양방향 로버스트 필터 방법에 기반한 특이치 검출결과를 나타낸 것이다. 앞에서 나타난 가면화효과와 수렁화효과에 의한 특이치 검출 오류가 발생하지 않고 있음을 확인할 수 있다.

3. 사례분석과 모의실험 결과

로버스트 추정량으로써 LMSE와 S-추정량에 본 논문에서 제안한 양방향 로버스트 필터 방법에 의한 특이치 검출 방법을 실제 자료와 모의 실험을 통해 검토해 보기로 하자. 이 가운데 S-추정량은 앞서 밝힌 바와 같이 고붕괴점과 효율성간의 교섭이 발생하기 때문에 두 가지 경우를 고려하였다. 즉 정규성하에서의 효율성을 무시하고 LMSE와 같이 고붕괴점을

최대화한 경우 (SE(.5))와 이와는 달리 오염된 자료에 대하여 고붕괴점에 도달하지는 않지만 좀 더 높은 효율성을 유지할 수 있도록 하는 수준으로 붕괴점을 약 25% 정도가 되도록 조절한 경우 (SE(.75))를 함께 시도해 보았다.

3.1. 사례분석

표 3.1: 미국 라디오와 TV에 대한 월간 재고주문 자료에서의 특이치 검출 결과

Date	기존방법			양방향로버스트필터		
	LMSE	SE(.5)	SE(.75)	LMSE	SE(.5)	SE(.75)
6/68	o	o	o			
9/76	o			o	o	o
3/77	o	o	o			
8/77	o	o	o			
2/78	o			o	o	o
3/78				o	o	o
4/78	o	o	o			
9/78	o	o	o	o	o	o
5/79	o	o	o	o	o	o
8/79				o	o	o
10/79	o	o	o			

표 3.1은 1958년 1월부터 1980년 10월까지 미국의 TV와 라디오의 월간 재고주문 자료에 대하여 로그변환한 후 이를 다시 X11-ARIMA를 이용하여 계절조정한 다음, AR(3) 모형을 이용하여 고붕괴점 로버스트 회귀추정량에 의한 표준화잔차를 그대로 이용하는 기존방법과 양방향 로버스트 필터 방법에 의한 특이치 검출 결과이다. 분석을 위한 자료와 선택된 모형은 Justel 등(2001)의 연구결과를 이용하였다. 특이치 분류를 위한 임계값으로 관찰값의 수에 따라 두 방법 모두 앞서 밝힌 바대로 3.5를 사용하였으며, 표준화잔차의 절대값이 이보다 큰 경우를 특이치로 분류하였다. 그림 3.1은 분석에 사용된 자료의 시도표와 양방향 로버스트 필터 방법에 의한 특이치 검출결과를 나타낸 것이다.

두 방법 모두 동일한 추정값을 사용하에도 불구하고 9/78와 5/79에서의 관찰값 (그림 3.1의 D와 E)을 제외하고는 서로 다른 특이치 검출결과를 보이고 있다. 특히 양방향 로버스트 필터 방법은 2/78과 3/78에서의 관찰값을 연속적으로 발생한 특이치로 분류하지만 (그림 3.1의 B와 C), 기존방법은 이들을 검출하는 대신 이후 시점인 4/78 관찰값을 특이치로 분류하고 있다. 즉 가면화효과와 수렁화효과에 의한 왜곡이 발생하고 있음을 볼 수 있다. 이들 외에 양방향 로버스트 필터 방법에 의하여 고립된 특이치로 판별된 9/76 (그림 3.1의 A)와 8/79 (그림 3.1의 F) 가운데 9/76은 그림 3.1을 통하여 특이치의 가능성을 의심해 볼 수 있으며, 기존방법 가운데 LMSE에 의한 결과는 이를 특이치로 분류하고 있다. 그러나 8/79는 시도표만으로 특이치 여부를 확인하기 어렵다. 그러나 사용된 모형이 AR(3)라는 점에서 이에 앞서 특이치로 분류된 5/79에 의한 가면화효과를 의심할 수 있다. 즉 5/79가 특

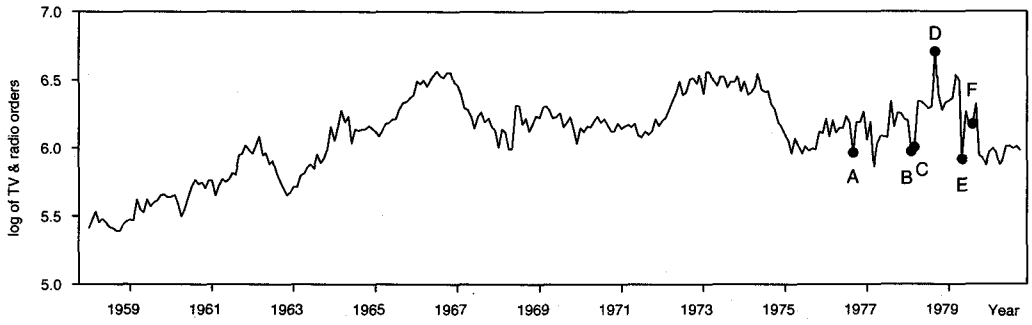

그림 3.1: 계절조정된 미국의 TV와 라디오 월간 재고주문 자료에 대한 시도표와 양방향 로버스트 필터 방법에 의한 특이치 검출결과 (점으로 표시).

이치이며 전계열이 AR(3)과정을 따른다면 5/79이후의 3개 시점까지 특이치에 의한 영향이 지속되기 때문이다.

3.2. 모의실험 결과

모의실험은 두 가지 과정을 나누어 실행하였다. 하나는 일련의 특이치군의 길이에 따른 변화를 표본크기를 늘여가며 로버스트 추정량의 효율성과 편향정도를 살펴보았으며, 다른 하나는 특이치 발생 형태를 여러가지로 변화시켜 가며 특이치 검출 능력을 평가해 보았다.

두 가지 모두 다음과 같은 정상시계열 과정으로부터 생성하였다.

$$y_t = 1.7y_{t-1} - .96y_{t-2} + .18y_{t-3} + 5\delta_t + a_t, \tag{3.1}$$

여기서 a_t는 $N(0,1)$을 따르도록 했으며, δ_t는 만약 시점 t에서 관찰값 y_t가 특이치라면 1의 값을 가지도록 하며, 그렇지 않다면 0의 값을 갖는 지시함수이다. 모형 (3.1)에서 다항식의 근은 각각 .6, .6, .5로써 정상 시계열이다. 첫번째 모의실험에서는 AO를 발생시킴에 있어 사전에 정한 비율만큼 연속적으로 특이치군을 발생시킨후 이렇게 오염된 자료에 대한 추정량의 성능을, 표본크기와 특이치군의 길이에 따른 변화에 따라 각 추정량의 전체평균제곱오차(total mean squared error, TMSE) 와 전체편향(total bias, TBIAS)을 비교하여 살펴보았다. 이들 전체평균제곱오차와 전체편향은 다음과 같이 계산하였다.

$$\text{TMSE} = \frac{1}{r}\sum_{i=1}^{p}\sum_{j=1}^{r}(\hat{\phi}_{ij} - \phi_i)^2, \quad \text{TBIAS} = \sum_{i=1}^{p}|\bar{\hat{\phi}}_i - \phi_i|$$

여기서 모의실험 반복횟수 r은 1000이며, $\bar{\hat{\phi}}_i = \sum_{j=1}^{r}\hat{\phi}_{ij}/r$이다. 표 3.2는 모의실험 가운데 특이치군을 계열의 중간에서 연속적으로 발생시킨 경우이다. 이 경우 외에도 특이치군의 위치를 계열의 전반 및 후반, 그리고 특이치의 크기를 변화시켜 가며 실험을 실시하였으나,

결과의 차이가 드러나지 않으므로 앞서 밝힌 바대로 특이치의 크기가 모두 5이며 계열의 중간에 이들 특이치군을 발생시킨 결과를 나타내었다.

표 3.2: 표본크기와 특이치군 비율의 변화에 따른 각 추정량의 전체평균제곱오차과 전체편향(괄호안)

표본크기	오염율(%)	LSE	LMSE	SE(0.5)	SE(0.75)
30	.00	0.187 (0.254)	0.879 (0.217)	0.367 (0.239)	0.241 (0.230)
	.05	1.823 (2.168)	0.839 (0.701)	0.604 (0.851)	0.631 (1.024)
	.10	1.141 (1.702)	0.976 (0.990)	0.786 (1.172)	0.901 (1.419)
	.15	1.078 (1.659)	0.941 (0.919)	0.777 (1.164)	0.850 (1.383)
	.20	1.034 (1.612)	0.925 (0.911)	0.747 (1.138)	0.833 (1.373)
	.25	0.999 (1.575)	0.887 (0.802)	0.712 (1.068)	0.793 (1.325)
	.30	1.001 (1.583)	0.817 (0.823)	0.699 (1.102)	0.800 (1.344)
60	.00	0.088 (0.135)	0.461 (0.101)	0.181 (0.132)	0.115 (0.130)
	.05	0.629 (1.243)	0.477 (0.459)	0.306 (0.583)	0.327 (0.730)
	.10	0.596 (1.219)	0.512 (0.486)	0.290 (0.550)	0.310 (0.708)
	.15	0.563 (1.174)	0.501 (0.443)	0.275 (0.526)	0.294 (0.677)
	.20	0.559 (1.167)	0.469 (0.427)	0.286 (0.499)	0.294 (0.669)
	.25	0.554 (1.159)	0.488 (0.351)	0.286 (0.468)	0.281 (0.628)
	.30	0.528 (1.130)	0.452 (0.370)	0.261 (0.474)	0.281 (0.649)
100	.00	0.050 (0.086)	0.272 (0.088)	0.092 (0.087)	0.064 (0.088)
	.05	0.342 (0.909)	0.306 (0.195)	0.145 (0.292)	0.126 (0.367)
	.10	0.332 (0.890)	0.302 (0.232)	0.139 (0.285)	0.127 (0.371)
	.15	0.329 (0.881)	0.308 (0.158)	0.144 (0.265)	0.126 (0.347)
	.20	0.331 (0.883)	0.300 (0.190)	0.135 (0.256)	0.121 (0.347)
	.25	0.311 (0.851)	0.294 (0.167)	0.127 (0.216)	0.113 (0.305)
	.30	0.310 (0.849)	0.320 (0.158)	0.134 (0.223)	0.113 (0.310)

비교한 추정방법들 가운데 S-추정량이 다른 추정량들에 비하여 낮은 전체평균제곱오차 및 편향을 갖는 것으로 보아 이들 가운데 가장 정확한 추정결과를 제공하는 것으로 보인다. 그러나 LMSE는 표본크기가 증가함에 따라 LSE와 비교하여 낮은 전체편향값을 갖지만 전체평균제곱오차에서는 그 차이가 점점 줄어들고 있음을 확인할 수 있다. 이것은 Rousseeuw와 Leroy (1987) 및 Meintanis와 Donatos (1999)가 고립된 AO 및 IO로만 이루어진 실험에서 LMSE가 보여준 결과와 비교하여 볼때 매우 불만족스러운 모습이다. 결국 AO가 연속적으로 발생하는 특이치군에 있어서 LMSE는 LSE에 비하여 크게 나은 모습을 보여주지 못하고 있음을 확인할 수 있다. S-추정량 역시 예상과는 다른 결과를 보여주고 있는데, 고붕괴점보다는 효율성에 맞춘 SE(.75)의 경우 표본크기가 그리 크지 않은 30 및 60에서는 고붕괴점을 유지하기 위한 SE(.5)와 비교할때 오염되지 않은 자료에 대해서는 우수한

결과를 제공하지만 오염된 자료에 대해서는 전체평균제곱오차 및 편향에서 보다 높은 값을 보이고 있다. 그러나 표본크기가 100인 경우에는 전체평균제곱오차는 줄어들고 있음을 확인할 수 있다.

표 3.3: 다양한 특이치 형태에 따른 각 추정량의 전체평균제곱오차 및 편향(괄호안)

오염률	구분	LSE	LMSE	SE(.5)	SE(.75)
.10	2OPs	1.319 (1.863)	0.308 (0.329)	0.278 (0.718)	0.219 (0.543)
	1OP+5IOs	2.670 (2.500)	0.425 (0.776)	1.235 (1.767)	0.591 (1.049)
.15	3OPs	1.433 (1.941)	0.366 (0.573)	0.717 (1.193)	0.335 (0.758)
	2OPs	1.262 (1.808)	0.320 (0.317)	0.269 (0.693)	0.208 (0.481)
	4OPs+3IOs	1.950 (2.153)	0.694 (1.099)	1.346 (1.826)	1.236 (1.706)
.20	4OPs	1.530 (2.008)	0.467 (0.762)	1.238 (1.730)	0.616 (1.066)
	1OP+10IOs	2.038 (2.269)	0.463 (0.787)	1.267 (1.766)	0.708 (1.162)

다음의 실험에서는 다양한 유형의 특이치로 오염된 자기상관회귀모형에 대하여 추정량의 효율성과 특이치 검출 능력을 살펴보았다. 이를 위하여 앞의 모형 (3.1)로부터 표본크기를 100으로 고정한 상태에서 다음과 같은 방법으로 특이치를 생성하여 각 반복수 1000번의 모의실험을 실시하였다.

- 10% 오염 : (a) 길이가 같은 연속된 두 개의 특이치군(outlying patch,OP)을 시계열의 1/3과 2/3 지점에서 발생시킴 (2OPs); (b) 길이가 5인 하나의 특이치군은 2/3 지점에서 그리고 각 시점 11,21 ,31,41,51 에서 하나씩의 고립된 특이치(isolate outlier,IO)를 발생시킴 (1OP+5IOs).

- 15% 오염 : (a) 길이가 같은 연속된 세 개의 특이치군을 시계열의 1/2, 2/3, 9/10 지점에서 발생시킴 (3OPs); (b) 특이치 길이가 각각 10과 5인 두 개의 특이치 군을 1/3과 2/3 지점에서 발생시킴 (2OPs); (c) 특이치 길이가 각각 5,3,2,3인 네 개의 특이치군을 4/5,3/5,2/5,1/5 지점에서 발생시킴 (4OPs).

- 20% 오염 : (a) 길이가 같은 네 개의 특이치군을 시계열의 1/2, 3/5, 4/5, 9/10 지점에서 발생시킴 (4OPs); (b) 길이가 10인 하나의 특이치군을 2/3 지점에서 그리고 각 시점 10,15,17,27,31,39,50,54,56,62 에서 고립된 특이치를 발생시킴 (1OP+10IOs).

표 3.3은 추정량들에 대한 전체평균제곱오차와 편향을 계산한 것이다. 일률적으로 특이치군만을 생성하여 살펴봤던 표 3.2에서와는 달리 LSE에 비하여 LMSE가 안정적인 결과를 보여주고 있음을 확인할 수 있다. 반면 S-추정량은 특이치의 유형에 따라 상이한 결과를 보여주고 있다. 즉 고립된 특이치 및 특이치군의 수가 클수록 전체평균제곱오차와 편향은 커지고 있지만 LMSE는 이러한 차이에 대해서 다른 추정량들과는 달리 큰 변화가 나타나지 않는다.

표 3.4: 다양한 특이치 형태에 따른 로버스트 필터 방법을 적용한 특이치 검출 능력 비교 : 특이치 검출률 및 오분류율(괄호안)

오염률	구분	기존방법			양방향로버스트필터		
		LMSE	SE (.5)	SE (.75)	LMSE	SE (.5)	SE (.75)
.10	2OPs	45.1(5.3)	41.8(4.8)	37.9(4.4)	94.4(5.4)	96.8(1.7)	97.1(1.2)
	1OP+5IOs	68.5(12.1)	66.0(10.6)	60.2(6.9)	95.4(3.2)	96.9(0.9)	96.6(0.5)
.15	3OPs	40.3(7.3)	37.1(6.8)	29.4(5.4)	90.3(8.0)	93.2(3.1)	94.2(1.9)
	2OPs	30.1(5.6)	27.9(5.2)	25.3(4.6)	88.0(7.4)	90.8(2.4)	92.1(1.3)
	4OPs+3IOs	60.5(12.8)	49.0(9.1)	46.7(8.3)	93.9(4.1)	95.6(1.2)	96.1(0.7)
.20	4OPs	36.2(9.3)	31.9(8.2)	21.1(5.5)	87.1(11.7)	90.2(3.3)	90.8(1.7)
	1OP+10IOs	56.9(17.9)	55.0(13.8)	49.4(5.2)	86.6(9.8)	88.2(2.9)	88.0(1.5)

표 3.4는 표준화 잔차를 이용한 기존의 특이치 검출방법과 양방향 로버스트 필터를 이용한 특이치 검출 결과를 비교한 것이다. 특이치 검출률은 특이치로 생성된 관찰치를 특이치로 판별한 경우를 백분율로 나타낸 것이고, 오분류율은 특이치가 아닌 관찰치를 특이치로 판별한 경우를 백분율로 나타낸 것으로 1000번 반복실험에서의 평균을 계산한 것이다. 따라서 특이치 검출률은 100에 가까울수록 그리고 오분류율은 0에 가까울수록 정확한 특이치 검출능력을 나타내는 것으로 볼 수 있다. LSE의 경우 전혀 특이치를 판별하지 못하기 때문에 결과를 싣지 않았다. 고붕괴점 로버스트 추정량에서 전방 및 후방 예측의 과정을 모두 사용하는 양방향 로버스트 필터방법에 기반한 특이치 검출방법이 매우 효과적임을 확인할 수 있다. 그러나 로버스트 추정량을 사용했음에도 불구하고 횡단면 자료에 대한 회귀분석에서와는 달리 표준화 잔차만을 이용한 특이치 검출결과는 시계열 자료의 특성에 의한 가면화효과와 수렁화효과에 의하여 부정확함을 볼 수 있다. 이와는 달리 양방향 로버스트 필터를 이용한 특이치 검출방법은 동일한 추정값을 사용했음에도 불구하고 특이치에 대한 검출율과 오분류율 모두에서 정확한 결과를 제공하고 있다.

4. 결론

대부분의 회귀모형에 대한 추정방법은 자기상관회귀모형과 같은 시계열 자료에 대해서 확장하여 사용될 수 있으며, 실제로 로버스트 회귀추정량을 이용함으로써 자기상관회귀모형에 대하여 로버스트한 결과를 얻을 수 있는 것으로 알려져 왔다. 그러나 오염된 시계열 자료에 대해서는 로버스트 추정 방법조차 가면화효과와 수렁화효과때문에 추정량의 정확성은 물론 특이치 검출능력 또한 상당히 훼손되는 것이 사실이다. 특히 기존 연구가 AO와 IO로 구분된 시계열 자료의 특이치가 소수 발생한 경우에 대해서만 연구가 진행되어 왔으나 본 연구에서 살펴봤듯이 AO가 연속적으로 발생되는 특이치군의 경우는 로버스트 추정량이라 할지라도 LSE에 비하여 평균제곱오차 관점에서 항상 높은 효율성을 보여주지는 못하고 있다. 더구나 표준화 잔차만을 이용하는 기존의 특이치 검출 방법 역시 가면화효과와 수렁화효과에 의한 왜곡에 의하여, 횡단면 자료에 대한 회귀모형과는 달리 자기상관회귀

모형에서의 특이치 검출 능력은 상당히 떨어짐음을 모의실험을 통해 확인할 수 있었다.

본 논문에서는 고붕괴점을 갖는 로버스트 회귀추정방법에 근거한 자기상관회귀모형에서의 특이치 검출방법을 새롭게 제안하였다. 이 방법 역시 로버스트 추정량을 이용한 잔차를 특이치 검출의 기준으로 고려하지만, 이전 단계의 오염된 관찰치의 영향력을 제한하기 위한 로버스트 필터를 사용하고 이와 더불어 이를 전방 예측과 후방 예측 전 과정에 적용함으로써 정확한 특이치의 검출을 가능하도록 한다. 이러한 양방향 로버스트 필터 방법은 다양한 AO 발생형태에 대해서 정확한 특이치 검출력을 보이고 있으며, 오분류율 역시 상당히 줄일 수 있음을 모의실험을 통해 확인할 수 있었다.

참고문헌

Bianco, A. M., García B. M., Martínez, E. J. and Yohai, V. J. (1996). Robust procedures for regression models with ARIMA errors. *COMPSTAT 96, Proc. Computational Statistics.* Part A, 27-38, Physica-Velag, Berlin.

Chang, I., Tiao, G. C. and Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, **3**, 193-204.

Denby, L. and Martin, R. D. (1979). Robust estimation of the first-order autoregressive parameter. *Journal of the American Statistical Association*, **74**, 140-146.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*(Bickel,P. J., Doksum,K. A. and Hodges,J. L., eds.) 157-184. Wadsworth, Belmont, California.

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society Ser. B*, **34**, 350-363.

Justel, A., Peña, D. and Tsay, R. S. (2001). Detection of outlier patches in autoregressive time series. *Statistica Sinica*, **11**, 651-673.

Martin, R. D. and Yohai, V. J. (1991). Bias robust estimation of autoregressive parameter. *Directions in Robust Statistics and Diagnostics*, Part I. 233-246, Springer, New York.

Meintanis, S. G. and Donatos, G. S. (1999). Finite-sample performance of alternative estimators for autoregressive models in the presence of outliers. *Computational Statistics and Data Analysis*, **31**, 323-339.

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

Terpstra, J. T., McKean, J. W. and Naranjo, J. D. (2001), Weighted Wilcoxon Estimators for Autoregression. *Austrailia and New Jealand Journal of Statitistics*, **43**, 399-419.

Tsay, R. S., Peña, D. and Pankratz, A. E. (2000). Outliers in multivariate time series. *Biometrika*, **87**, 789-804.

[2006년 3월 접수, 2006년 4월 채택]

Outlier Detection of Autoregressive Models Using Robust Regression Estimators

Dong-Hee Lee[1] Yousung Park[2] Kee Whan Kim[3]

ABSTRACT

Outliers adversely affect model identification, parameter estimation, and forecast in time series data. In particular, when outliers consist of a patch of additive outliers, the current outlier detection procedures suffer from the masking and swamping effects which make them inefficient. In this paper, we propose new outlier detection procedure based on high breakdown estimators, called as the dual robust filtering. Empirical and simulation studies in the autoregressive model with orders p show that the proposed procedure is effective.

Keywords: additive outlier, autoregressive model, high breakdown estimator, innovative outlier, outlying patch, robust filter, time reversibility

1) Research-Assistant Professor, Institute of Statistics, Korea University, 5-1 Anam-Dong, Sungbuk-gu, Seoul, 136-701, Korea.
E-mail: ld0351@korea.ac.kr
2) Professor, Dept. of Statistics, Korea University, 5-1 Anam-Dong, Sungbuk-gu, Seoul,136-701, Korea.
E-mail: yspark@korea.ac.kr
3) Assistant Professor, Dept. of Informational Statistics, Korea University, 208 Seochang-Ri, Jochiwon-Eup, Yeonki-Gun, Chung-Nam, 339-700, Korea.
E-mail: korpen@korea.ac.kr