The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.5.907
29(5), 907–920

Outlier detection in time series data

Jeong In Choi[a] · In Ok Um[a] · Hyung Jun Cho[a,1]

[a]Department of Statistics, Korea University

(Received May 25, 2016; Revised July 25, 2016; Accepted August 2, 2016)

Abstract

This study suggests an outlier detection algorithm that uses quantile autoregressive model in time series data, eventually applying it to actual stock manipulation cases by comparing its performance to existing methods. Studies on outlier detection have traditionally been conducted mostly in general data and those in time series data are insufficient. They have also been limited to a parametric model, which is not convenient as it is complicated with an analysis that takes a long time. Thus, we suggest a new algorithm of outlier detection in time series data and through various simulations, compare it to existing algorithms. Especially, the outlier detection algorithm in time series data can be useful in finding stock manipulation. If stock price which had a certain pattern goes out of flow and generates an outlier, it can be due to intentional intervention and manipulation. We examined how fast the model can detect stock manipulations by applying it to actual stock manipulation cases.

Keywords: outlier detection, quantile autoregressive model, time series data

1. 서론

특이치(outlier)란, 어떠한 분포를 가지는 변수(variable)의 값들 중에서 비정상적으로 본래의 분포를 벗어난 값이다. 자료에 특이치가 포함되었을 때, 추정치가 특이치로 인해 편향(bias)을 가지거나 그 타당도가 결여되는 문제를 가진다.

특이치를 발견하는 방법에 대한 연구는 다양한 분야에서 진행되어왔다. Hoaglin 등 (1986)은 정규분포, 지수분포 등을 따르는 간단한 일변량 자료에서의 특이치 발견방법을 제시하였고, 일반적인 데이터 형태 이외에도 Nardi과 Schemper (1999)는 생존 데이터에서의 특이치 발견 알고리즘을 제안하였다. Chen과 Liu (1993b)는 시계열 자료에서 모수적 모형을 이용한 특이치 발견 알고리즘을 제시하였다. 이는 시계열 모형의 모수와 특이치 효과 모수를 반복적으로(iteratively) 추정하는 방법이다. Fried (2004)는 시계열 자료에서의 비모수적인 특이치 발견 알고리즘을 제시하였다.

본 연구에서는 분위수 회귀(quantile regression)를 활용한 비모수적인 특이치 발견 알고리즘을 제안하고, 다양한 옵션의 모의실험을 통해 Fried (2004)의 알고리즘과 비교한다.

시계열 자료에서 특이치가 발생하는 실제 사례로 주가 데이터를 들 수 있다. 주가 데이터는 시간의 흐름

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1D1A1A09058602).
[1]Corresponding author: Department of Statistics, Korea University, 145 Anam-ro 145, Seongbuk-Gu, Seoul 02841, Korea. E-mail: hj4cho@korea.ac.kr

에 따라 관측되기 때문에 시계열 자료라 할 수 있으며, 보통의 흐름에서 벗어나는 주가가 발견되면 그것을 특이치라고 생각할 수 있다. 이를 통해 주가 조작(특이치)을 적발하는 데에 활용할 수 있을 것이다.

본 논문의 구성은 다음과 같다. 제 2절에서는 시계열 데이터의 특이치를 발견하는 비모수적인 방법 세 가지 알고리즘 Residual-based Algorithm, Boxplot Algorithm과 Fried (2004)가 제시한 Robust Filtering Algorithm을 살펴본다. 제 3절에서 각 알고리즘의 비교 모의실험을 수행하며, 모의실험에는 R 소프트웨어의 'quantreg'와 'robfilter' 패키지를 이용하였다. 제 4절에서는 동양시멘트의 주가 조작 사례에 적용 해본다. 마지막 제 5절에서 결론을 제시한다.

2. 시계열 자료에서의 특이치 발견 알고리즘

2.1. 분위수 회귀를 이용한 특이치 발견 알고리즘

분위수 회귀는 보통최소제곱추정(ordinary least squares estimation)에 비해 특이치에 민감하지 않고, 자료가 정규성을 만족하지 않더라도 회귀분석을 수행할 수 있다는 장점이 있다. 따라서 분위수 회귀에 자기상관성을 고려한다면 시계열 자료에서 특이치를 발견하는 데 적합할 것이다. Koenker과 Xiao (2006)은 분위수 자기회귀모형(quantile autoregressive model)을 제안하였다.

다음으로 소개될 두 개의 알고리즘, Residual-based Algorithm과 Boxplot Algorithm은 시계열 자료에서 분위수 회귀를 이용하여 특이치를 발견하는 방법이다. 이는 절단된 자료에서의 특이치 발견 알고리즘을 제시한 Eo 등 (2014)의 방법을 시계열 자료에 맞게 응용한 것이다.

2.1.1. Residual-based Algorithm 이 절에서 소개하는 잔차 기반 알고리즘의 전 과정은 SAS의 quantreg 프로시저 (SAS Institute Inc., 2008)의 특이치 발견 알고리즘을 응용한 것이다. 자료 Y_t는 p차 자기회귀과정 AR(p)를 따르는 시계열 자료라고 가정한다.

Step 1. 제 2사분위수($\tau = 0.5$) 회귀로 얻은 조건부 함수 $\hat{Q}_{y_t}(0.5|y_{t-1}, \ldots, y_{t-p})$를 통해 잔차 $r_t = y_t - \hat{Q}_{y_t}(0.5|y_{t-1}, \ldots, y_{t-p})$를 계산한다 ($t = 1, \ldots, T$).

Step 2. 각 시점 t에서의 스코어 s_t를 다음과 같이 정의한다.

$$s_t = \begin{cases} \dfrac{r_t}{\tilde{\sigma}_1}, & r_t \geq 0, \\ -\dfrac{r_t}{\tilde{\sigma}_2}, & r_t < 0. \end{cases}$$

이 때, $\tilde{\sigma}_1 = Q_3\{r_1, \ldots, r_T\}/\Phi^{-1}(0.75)$, $\tilde{\sigma}_2 = Q_1\{r_1, \ldots, r_T\}/\Phi^{-1}(0.25)$이다. 여기서, $Q_i\{r_1, \ldots, r_T\}$는 T개 잔차들의 i-사분위수를 의미한다.

Step 3. 특이치 지시변수 D_t를 아래와 같이 정의하고

$$D_t = \begin{cases} 1, & s_t > k, \\ 0, & \text{otherwise}. \end{cases}$$

$D_t = 1$인 시점의 관측치 y_t를 특이치라고 밝힌다 (기본값은 $k = 3$).

2.1.2. Boxplot Algorithm 상자수염그림(box-and-whisker plot) 혹은 상자그림은 Tukey (1977)가 제시한 내용으로, 일변량 자료에서의 특이치 발견에 널리 쓰여 오던 방법이다. 하지만 이 방법은 비대칭적인 자료에는 적절하지 않을 가능성이 존재하기 때문에 본 논문에서는 준 사분

위간 범위(semi interquartile range; SIQR)를 활용하여 보완하였다. 준 사분위간 범위의 정의는 $\text{SIQR}_U = Q_3 - Q_2$, $\text{SIQR}_L = Q_2 - Q_1$이며, $[Q_1 - 2k\text{SIQR}_L, Q_3 + 2k\text{SIQR}_U]$를 벗어나는 자료값을 특이치라고 할 수 있다. 아래에 제시된 알고리즘은 분위수 회귀모형으로 적합(fitted)된 시계열 자료에서 준 사분위간 범위를 기반으로 한 상자그림 방법으로 특이치를 발견하는 알고리즘이다. 여기서 자료 Y_t는 이전 알고리즘에서와 같이 p차 자기회귀과정 AR(p)를 따르는 시계열 자료이다.

Step 1. 제 1사분위수($\tau = 0.25$), 제 2사분위수($\tau = 0.50$), 그리고 제 3사분위수($\tau = 0.75$) 회귀를 바탕으로 조건부 함수 $\hat{Q}_{y_t}(0.25|y_{t-1}, \ldots, y_{t-p})$와 $\hat{Q}_{y_t}(0.50|y_{t-1}, \ldots, y_{t-p})$, 그리고 $\hat{Q}_{y_t}(0.50|y_{t-1}, \ldots, y_{t-p})$을 얻는다. 편의상 각각 $\hat{Q}_{y_t}(0.25)$, $\hat{Q}_{y_t}(0.50)$, 그리고 $\hat{Q}_{y_t}(0.75)$이라 표기하기로 한다.

Step 2. 준 사분위간 범위를 아래와 같이 정의한다.

$$\text{SIQR}_U = \hat{Q}_{y_t}(0.75) - \hat{Q}_{y_t}(0.50),$$
$$\text{SIQR}_L = \hat{Q}_{y_t}(0.50) - \hat{Q}_{y_t}(0.25).$$

Step 3. 각 시점 t에서의 스코어 s_t를 다음과 같이 정의한다.

$$s_t = \begin{cases} \dfrac{y_t - \hat{Q}_{y_t}(0.75)}{2\text{SIQR}_U}, & y_t \geq \hat{Q}_{y_t}(0.50), \\ -\dfrac{y_t - \hat{Q}_{y_t}(0.25)}{2\text{SIQR}_L}, & y_t < \hat{Q}_{y_t}(0.50). \end{cases}$$

Step 4. 특이치 지시변수 D_t를 다음과 같이 정의하고

$$D_t = \begin{cases} 1, & s_t > k, \\ 0, & \text{otherwise.} \end{cases}$$

$D_t = 1$인 시점의 관측치 y_t를 특이치라고 밝힌다 (기본값은 $k = 1.5$).

2.2. Robust Filtering을 이용한 특이치 발견 알고리즘

Fried (2004)는 시간에 따라 움직이는 각각의 윈도우(time window) 안에서 중위수를 활용하여 선형추세적합(linear trend fitting)을 함으로써 시계열 데이터에서 비모수적으로 특이치를 발견하는 방법을 제시하였다.

2.2.1. 모형설정과 모수 추정방법 Fried (2004)가 설정하는 모형은 Y_t를 중앙값 μ_t와 분산 σ_t^2를 가지는 확률변수라고 할 때, 크기가 $n = 2m + 1$인 각 윈도우에 대해 고려하는 모형이다.

$$Y_{t+i} = \mu_t + i\beta_t + E_{t+i} + r_{t+i}, \quad i = -m, \ldots, m.$$

이 때 μ_t와 β_t는 각각 윈도우 안에서의 수준(level)과 기울기(slope)이며, E_{t+i}는 중앙값 0과 분산 σ_t^2를 가지는 랜덤잡음(random noise), r_{t+i}는 절대값이 $\varepsilon\sigma_t$보다 작은, 즉 $|r_{t+i}| \leq \varepsilon\sigma_t$인 근사오차(approximation error)이다.

μ_t와 β_t를 추정하는 방법에는 두 가지가 있다. 첫 번째 방법은 최소중위제곱(least median of squares; LMS)방법으로 오차항 제곱의 중위수를 최소로 만드는 μ_t와 β_t를 구하는 방법 (Rousseeuw, 1984)으로,

추정량의 식은 아래와 같다.

$$T_{LMS} = \arg\min \left[(\mu, \beta) : \text{median}(y_{t+i} - \mu - i\beta)^2\right].$$

두 번째로 제안된 방법은 반복된 중위수(repeated median; RM) 방법이다 (Siegel, 1982). Siegel (1982)는 RM 추정량이 적절한 조건 하에서 비편향(unbiased) 추정량이며 효율적이라고 말하고 있다.

$$T_{RM} = (\tilde{\mu}_t, \tilde{\beta}_t), \quad \text{단 } \tilde{\beta}_t = \text{med}_i \left(\text{med}_{j \neq i} \frac{y_{t+i} - y_{t+j}}{i - j} \right), \ \tilde{\mu}_t = \text{med}_i (y_{t+i} - i\tilde{\beta}_t).$$

다음으로 σ_t^2를 추정하는 네 가지 방법에 대해 알아보도록 하겠다. 위에서 추정한 μ_t와 β_t의 추정량을 각각 $\hat{\mu}_t$와 $\hat{\beta}_t$라 하면, 각 윈도우 안에서의 잔차 r_i는 다음과 같이 표현될 수 있다.

$$r_i = y_{t+i} - \hat{\mu}_t - \hat{\beta}_t i, \quad i = -m, \ldots, m.$$

위의 잔차를 이용하여 분산 σ_t^2를 추정할 수 있는데, 첫 번째 방법은 median absolute deviation(MAD) 방법으로 단순히 잔차 절댓값의 중앙값을 이용하는 전통적인 방법이다.

$$\tilde{\sigma}_{MAD} = c_{1,n} \text{med} \left\{ |r_{-m}|, \ldots, |r_m| \right\}.$$

이 때 $c_{1,n}$은 수정 인자(correction factor)로서 기본값으로 $c_{1,n} = 1/\Phi^{-1}(0.75) \approx 1.4826$이다. 두 번째 추정방법은 length of the shortest half(LSH)인데 이는 Rousseeuw (1988)에서 제안되었다. 순서에 따라 정렬된 잔차들을 $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$이라고 하면 LSH 추정값은 아래와 같다.

$$\tilde{\sigma}_{LSH} = c_{2,n} \min \left\{ |r_{(-m)} - r_{(m)}|; \ i = 1, \ldots, n - m \right\}.$$

$c_{2,n}$ 또한 수정 인자로서 $c_{2,n} = 1/(2\Phi^{-1}(0.75)) \approx 0.7413$을 기본값으로 가진다. 세 번째와 네 번째로 살펴볼 추정량은 각각 $\tilde{\sigma}_{QN}$과 $\tilde{\sigma}_{SN}$로, Rousseeuw과 Croux (1993)가 제안하였다.

$$\tilde{\sigma}_{QN} = c_{3,n} \left\{ |r_i - r_j| : -m \leq i \leq j \leq m \right\}_{(h)}, \quad h = \binom{m+1}{2},$$

$$\tilde{\sigma}_{SN} = c_{4,n} \text{med}_i \, \text{med}_{j \neq i} |r_i - r_j|.$$

이 때 각각의 수정 인자는 $c_{3,n} = 1/(\sqrt{2}\Phi^{-1}(5/8)) \approx 2.2219$, $c_{4,n} \approx 1.1926$으로 $c_{4,n}$은 $\Phi(\Phi^{-1}(3/4) + c^{-1}) - \Phi(\Phi^{-1}(3/4) - c^{-1}) = 1/2$를 만족하는 c값이다.

2.2.2. Robust Filtering Algorithm 자료 Y_t는 시계열 자료이며 윈도우의 크기는 $n = 2m+1$로, 즉 $j = -m, \ldots, m$이다.

Step 1. 첫 번째 윈도우 안에서 자료 y_{t+j}들을 이용하여 $\hat{\mu}_t$, $\hat{\beta}_t$, $\hat{\sigma}_t$를 추정한다.

Step 2. $r_j = y_{t+j} - \hat{\mu}_t - j\hat{\beta}_t$라 놓고 $|r_j| > d_0\hat{\sigma}_t$인 j번째 관측치 y_{t+j}을 특이치로 분류한 후, 해당 y_{t+j}를 ${y'}_{t+j} = \hat{\mu}_t + j\hat{\beta}_t + \text{sgn}(r_j) d_1 \hat{\sigma}_t$로 바꾼다. 이 때, sgn(•)은 -1, 0, 1의 값을 가지는 부호함수(sign function)이며 d_0과 d_1은 $0 \leq d_1 \leq d_0$인 상수이다. 여기서는 d_0과 d_1에 대한 몇 가지 선택방법을 제시하는데, 이는 다음과 같다.

T	$d_0 = 3$,	$d_1 = 0$	(trimming)
L	$d_0 = 3$,	$d_1 = 1$	(downsizing large values)
M	$d_0 = 2$,	$d_1 = 1$	(downsizing moderate values)
W	$d_0 = 2$,	$d_1 = 2$	(winsorization)

즉, $d_0\hat{\sigma}_t$ 이상으로 벗어나 있는 $d_1\hat{\sigma}_t$ 관측치를 만큼만 떨어져 있도록 임시 조정하는 것이다.

Step 3. 해당 윈도우에서 $n = 2m + 1$개의 관측치 중 양의 특이치로 밝혀진 값의 개수가 m개보다 많으면 ${y'}_{t+j}$을 y_{t+j}로 원상 복구하고 특이치로 분류되었던 관측치 모두를 비특이치(non-outlier)로 재분류한다. 음의 특이치의 경우도 마찬가지이다.

Step 4. 해당 윈도우에서 특이치로 밝혀지지 않는 값의 개수가 $\max\{\lfloor m/3 \rfloor, 5\}$개보다 적어도 ${y'}_{t+j}$을 y_{t+j}로 원상 복구하고 특이치로 분류되었던 관측치 모두를 비특이치(non-outlier)로 재분류한다.

Step 5. 다시 자료 y_{t+j}들을 이용하여 $\hat{\mu}_t$, $\hat{\beta}_t$, $\hat{\sigma}_t$를 추정한다.

Step 6. 추정된 $\hat{\mu}_t$, $\hat{\beta}_t$, $\hat{\sigma}_t$를 이용해 다음 윈도우의 첫 관측치인 $\hat{y}_{t+m+1}$가 특이치인지 판단한다. 즉, 만약 $|r_{m+1}| = |y_{t+m+1} - \hat{\mu}_t - \hat{\beta}_t(m+1)| > d_0\hat{\sigma}_t$라면 그 값을 특이치로 분류한 후, y_{t+m+1}를 다시 ${y'}_{t+m+1} = \hat{\mu}_t + \hat{\beta}_t(m+1) + \mathrm{sgn}(r_{m+1})d_1\hat{\sigma}_t$로 바꾼다.

Step 7. 윈도우의 중심 t를 $t+1$로(다음 윈도우로) 옮기고 Step 3으로 돌아간다.

3. 모의실험

특이치가 있는 시계열 데이터에서 앞서 설명한 알고리즘이 얼마나 특이치를 잘 발견할 수 있는지 모의실험을 통해 비교한다. 모의실험에서는 R 소프트웨어의 quantreg 패키지를 이용하여 Residual-based Algorithm과 Boxplot Algorithm의 분위수 회귀모형을 적합하고 알고리즘을 구현하였고, robfilter 패키지를 이용하여 Robust Filtering Algorithm을 적용하였다.

3.1. 모의실험 설계

Fox (1972)와 Chen과 Liu (1993a)에서는 시계열 데이터에서 발생할 수 있는 네 가지 종류의 특이치를 제시하였다. 먼저 Y_t가 정상(stationary) 시계열 ARMA모형을 따를 때, m번 $(t_1, \ldots, t_m)$의 특이치가 포함된 시계열모형식은 다음과 같다.

$$Y_t = \frac{\theta(B)}{\alpha(B)\phi(B)}a_t, \quad t = 1, \ldots, T,$$
$$Y_t^* = Y_t + \sum_{j=1}^{m} \omega_j L_j(B) I_t(t_j),$$

여기에서 ω_j는 특이치의 크기이고, $I_t(t_j)$는 지시함수로서 특이치가 $t = t_j$에서 발생하였을 때 1의 값을 가진다. 이 때 $L_j(B)$는 특이치의 종류에 따라 그 정의가 달라지는데, 네 가지의 특이치 종류는 각각 innovational outlier(IO), additive outlier(AO), temporary change(TC), level shift(LS)로 $t = t_j$시점에서의 영향력 모형은 다음과 같다.

$$\text{IO}: \quad L_j(B) = \frac{\theta(B)}{\alpha(B)\phi(B)},$$
$$\text{AO}: \quad L_j(B) = 1,$$
$$\text{TC}: \quad L_j(B) = \frac{1}{(1-\delta B)},$$
$$\text{LS}: \quad L_j(B) = \frac{1}{(1-B)}.$$

이처럼 네 가지 종류의 특이치를 기본으로, Chen과 Liu (1993b)의 연구를 바탕으로 하여 모의실험을 설계하였다. 먼저 정상 시계열의 모형을 설정하고 특이치의 개수와 위치, 종류를 바꾸어 가며 다양하

게 발생할 수 있는 경우를 고려하였다. 정상 시계열의 모수 또한 동 연구를 참고하여 정하였다 ($\phi = 0.6, \theta = 0.6$).

특이치가 한 개 발생할 때, 정상 시계열 모형이 AR(1), MA(1), AR(2), MA(2)인 경우들로, 각 모형마다 네 종류(IO, AO, LS, TC)의 특이치는 $\omega_1 = 5$의 크기로 $t_1 = 10$, $t_1 = 40$, $t_1 = 90$에서 발생하였다. 여기서 특이치 발생 시점을 $t_1 = 10$, $t_1 = 40$, $t_1 = 90$로 가정한 이유는 총 길이 $T = 100$에서 처음, 중간, 끝 부분에 특이치가 발생할 때의 경우를 알아보기 위함이다. MA(1), MA(2)의 경우에는, 분위수 자기회귀모형을 이동평균모형에 응용하여 적용하였다.

특이치가 두 개 발생할 때, 정상 시계열 모형이 AR(1), MA(1)인 경우, 각 모형마다 특이치는 $\omega_1 = 5$, $\omega_2 = 5$의 크기로 $t_1 = 90$, $t_2 = 91$ 또는 $t_1 = 90$, $t_2 = 95$에서 발생하였다. 특이치 발생으로부터 최대한 빠른 시일 내에 특이치를 발견하려면, 최근 시점($t_1 = 90$ 이후)에 발생하는 특이치를 잘 발견할 수 있어야 한다. 따라서 이 경우에서는 $t_1 = 90$ 이후에 발생하는 경우를 중점으로 결과를 살펴보도록 하겠다. 두 개의 특이치 종류는 서로 다른 종류(IO-TC, IO-AO, IO-LS), 혹은 같은 종류의 특이치들이 나타나는 경우로 고려하였다.

모든 모의실험은 길이가 $T = 100$인 시계열 자료를 바탕으로 하였으며, 독립적으로 500번 반복하였다. 각 모의실험 경우에 대하여 세 가지 알고리즘(Residual-based Algorithm, Boxplot Algorithm, Robust Filtering Algorithm)을 비교하였고, 비교기준으로 민감도(sensitivity)와 특이도(specificity)를 사용하였다. 여기서 민감도는 특이치가 맞을 때 특이치라고 판단할 확률이고, 특이도는 특이치가 아닐 때 특이치가 아니라고 판단할 확률이다.

$$\text{민감도} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}},$$
$$\text{특이도} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}.$$

3.2. 모의실험 결과

우선, 특이치가 한 개인 경우를 살펴보면, Table 3.1과 같이 AR(1)기반 모형의 경우, 가장 큰 민감도와 특이도를 가지는 알고리즘은 대부분 Residual-based Algorithm인 것을 알 수 있다. 또한, 민감도와 특이도가 크게 차이나지 않는 것으로 보아 특이치의 종류와 위치에 따른 차이는 거의 없다고 할 수 있다. 한편, Robust Filtering Algorithm의 여덟 가지 방법 중 가장 높은 민감도와 특이도를 가지는 방법에는 각 Case별로 굵은 글씨로 나타내었다. 대부분의 경우에서 T_{RM}—$\tilde{\sigma}_{QN}$가 선택되었으며, 이는 Fried (2004)에서 주장한 바와 상통한다. 즉, μ_t와 β_t를 추정할 때 특이치가 적은 경우에는 RM 추정량이 LMS 추정량보다 효율적이며, σ_t^2의 추정방법 네 가지($\tilde{\sigma}_{MAD}, \tilde{\sigma}_{LSH}, \tilde{\sigma}_{QN}, \tilde{\sigma}_{SN}$) 중에서는 $\tilde{\sigma}_{QN}$ 추정량이 가장 좋은 성능을 보인다.

Table 3.2와 같이 MA(1) 기반 모형의 결과값도 Residual-based Algorithm과 Boxplot Algorithm이 비슷한 비율로 선택되는 것을 확인하였다. 두 알고리즘 사이에 큰 차이는 없고, 모두 분위수 회귀에 기초한 알고리즘이므로 이 경우에서도 역시 Robust Filtering Algorithm보다 분위수 회귀 알고리즘이 효과적인 방법임을 알 수 있다. 하지만 level shift(LS) 특이치에 있어서는 AR(1)을 바탕으로 한 모형에서보다 훨씬 낮은 민감도를 보이는데, 특히 $t_1 = 10$의 경우 민감도와 특이도 모두 Robust Filtering Algorithm보다도 좋지 않다. $t_1 = 40$ 혹은 $t_1 = 90$일 때에 비해서도 낮은 이유는 특이치가 초기에 발생할수록 LS 특이치의 특성상 발견이 어렵기 때문이다.

종합하자면, AR(1)을 바탕으로 MA(1)하는 모형에서 보다 좋은 민감도와 특이도를 보이고 있고, 대체

Table 3.1. Specificity and sensitivity at AR(1) based model ($\omega_1 = 5$)

			RES	BOX	RobFilter							
					T_{RM}				T_{LMS}			
					$\tilde{\sigma}_{MAD}$	$\tilde{\sigma}_{LSH}$	$\tilde{\sigma}_{QN}$	$\tilde{\sigma}_{SN}$	$\tilde{\sigma}_{MAD}$	$\tilde{\sigma}_{LSH}$	$\tilde{\sigma}_{QN}$	$\tilde{\sigma}_{SN}$
특이치없음		민감도	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		특이도	**0.992**	0.983	0.957	0.942	**0.965**	0.956	0.939	0.921	0.954	0.945
$t_1 = 10$	IO	민감도	**0.974**	0.970	0.548	0.470	0.520	**0.566**	0.540	0.454	0.552	0.548
		특이도	**0.993**	0.984	0.956	0.945	0.957	0.955	0.937	0.920	**0.961**	0.944
	AO	민감도	**0.970**	0.960	0.756	0.698	**0.822**	0.784	0.684	0.606	0.752	0.688
		특이도	**0.990**	0.980	0.957	0.940	**0.967**	0.956	0.941	0.914	0.963	0.947
	LS	민감도	**0.794**	0.480	0.074	0.074	0.046	0.064	0.090	**0.112**	0.076	0.106
		특이도	**0.993**	0.985	0.959	0.941	**0.968**	0.959	0.942	0.925	0.959	0.948
	TC	민감도	**0.968**	**0.968**	0.416	0.358	0.388	0.404	0.438	0.390	**0.484**	0.474
		특이도	**0.993**	0.984	0.958	0.944	**0.967**	0.957	0.939	0.919	0.952	0.946
$t_1 = 40$	IO	민감도	0.958	**0.964**	0.598	0.466	**0.692**	0.620	0.506	0.412	0.562	0.556
		특이도	**0.993**	0.984	0.954	0.942	**0.961**	0.954	0.939	0.914	0.959	0.942
	AO	민감도	**0.950**	0.948	0.712	0.614	**0.762**	0.710	0.596	0.484	0.652	0.622
		특이도	**0.989**	0.980	0.957	0.941	**0.966**	0.958	0.938	0.915	0.963	0.944
	LS	민감도	0.864	**0.866**	0.134	0.126	0.124	0.142	0.138	**0.176**	0.140	0.146
		특이도	**0.993**	0.984	0.961	0.945	**0.971**	0.964	0.945	0.925	0.960	0.949
	TC	민감도	0.956	**0.964**	0.542	0.402	**0.602**	0.540	0.432	0.388	0.472	0.496
		특이도	**0.993**	0.984	0.955	0.941	**0.963**	0.955	0.937	0.918	0.960	0.943
$t_1 = 90$	IO	민감도	0.968	**0.972**	0.614	0.556	**0.622**	0.626	0.580	0.522	0.568	0.572
		특이도	**0.993**	0.984	0.954	0.937	**0.963**	0.952	0.938	0.915	0.957	0.944
	AO	민감도	**0.970**	0.962	**0.734**	0650	0.732	0.724	0.630	0.604	0.688	0.660
		특이도	**0.990**	0.981	0.957	0.939	**0.966**	0.956	0.941	0.916	0.962	0.947
	LS	민감도	0.930	**0.948**	0.128	0.146	0.120	0.136	0.170	0.178	0.124	0.148
		특이도	**0.992**	0.983	0.962	0.947	**0.969**	0.960	0.944	0.923	0.959	0.949
	TC	민감도	**0.968**	0.966	0.550	0.496	**0.558**	0.566	0.564	0.518	0.522	0.554
		특이도	**0.993**	0.984	0.952	0.937	**0.961**	0.951	0.936	0.914	0.956	0.941

$*$ 특이치가 존재하지 않는 경우, 민감도를 계산할 수 없어 N/A로 표시하였다.
IO = innovational outlier; AO = additive outlier; LS = level shift; TC = temporary change.

로 Residual-based Algorithm이 가장 좋은 성능을 지닌다고 볼 수 있다. 특히 LS 특이치의 경우 초기에 발생할 경우 발견하기 어려운 경향을 보인다. 자세히 짚어보지 않은 AR(2), MR(2) 기반 모형의 결과는 각각 AR(1), MR(1) 기반 모형의 결과와 유사하기 때문에 생략하기로 한다.

다음으로 특이치 발견 알고리즘이 최근 시점에 발생한 특이치를 얼마나 잘 찾아내는지 확인하기 위하여, 특이치가 한 개 존재하는 경우에서 발생시점을 $t_1 = 85$에서 $t_1 = 100$까지 옮겨가면서 민감도와 특이도의 추이를 알아보았다.

Figure 3.1의 결과를 살펴보면, 열 가지의 알고리즘 모두 최근시차로 갈수록 민감도와 특이도가 높으므로, 최근 발생한 특이치를 가능한 한 빨리 찾아내는 데 유용하다. 하지만 그 중에서도 분위수 회귀 기반의 두 알고리즘이 특히 뛰어난 성능을 보인다. 또한, LS 특이치의 경우, 최근 시점에서 떨어져 발생할수록 발견하기 어렵다.

다음으로 특이치가 두 개 발생할 때의 모의실험 결과를 살펴보도록 하겠다. 특이치 한 개의 결과에서 확인하였듯이, AR(2)과 MA(2) 기반의 모형은 각각 AR(1), MA(1) 모형과 유사한 결과를 보이므로 여기

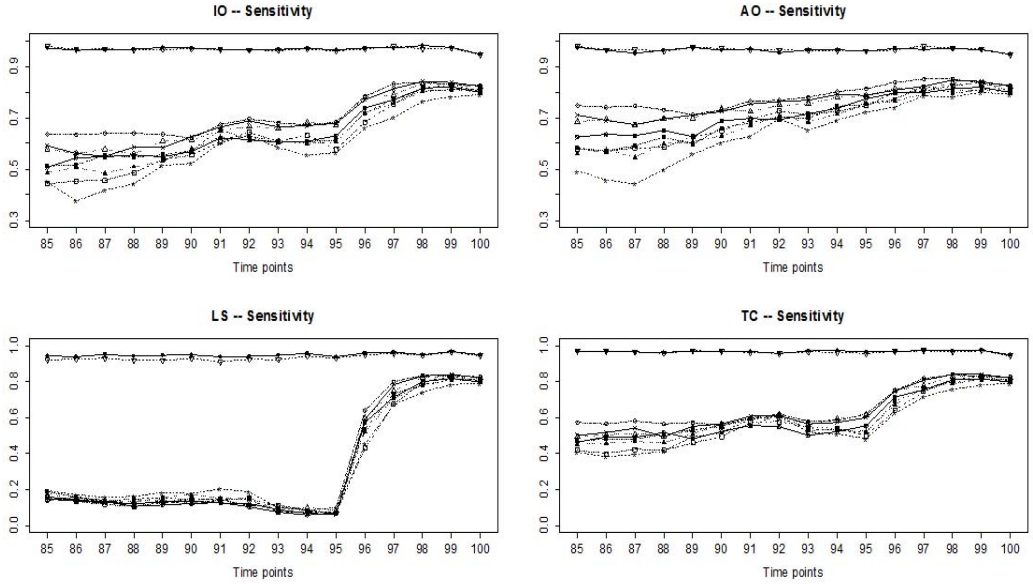
Table 3.2. Specificity and sensitivity at MA(1) based model ($\omega_1 = 5$)

					RobFilter							
			RES	BOX	T_{RM}				T_{LMS}			
					$\tilde{\sigma}_{MAD}$	$\tilde{\sigma}_{LSH}$	$\tilde{\sigma}_{QN}$	$\tilde{\sigma}_{SN}$	$\tilde{\sigma}_{MAD}$	$\tilde{\sigma}_{LSH}$	$\tilde{\sigma}_{QN}$	$\tilde{\sigma}_{SN}$
특이치없음		민감도	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		특이도	**0.992**	0.982	0.941	0.810	**0.984**	0.947	0.913	0.877	0.953	0.926
$t_1 = 10$	IO	민감도	0.958	**0.970**	0.742	0.742	0.476	**0.752**	0.674	0.552	0.652	0.692
		특이도	**0.992**	0.982	0.938	0.910	**0.983**	0.940	0.909	0.981	0.953	0.922
	AO	민감도	0.948	**0.966**	**0.802**	0.798	0.618	0.818	0.722	0.650	0.724	0.728
		특이도	**0.991**	0.982	0.940	0.917	**0.982**	0.944	0.911	0.873	0.954	0.924
	LS	민감도	0.010	0.002	0.010	0.024	0.004	0.008	0.024	**0.074**	0.016	0.020
		특이도	0.942	0.946	0.955	0.849	**0.990**	0.959	0.930	0.900	0.959	0.940
	TC	민감도	**0.918**	0.910	0.390	0.350	0.230	0.390	0.460	0.436	0.420	**0.480**
		특이도	**0.978**	0.973	0.950	0.855	**0.982**	0.954	0.926	0.781	0.956	0.930
$t_1 = 40$	IO	민감도	**0.918**	0.914	0.698	0.680	0.648	**0.708**	0.588	0.468	0.608	0.614
		특이도	**0.993**	0.982	0.932	0.794	0.979	0.935	0.906	0.870	0.951	0.919
	AO	민감도	0.916	**0.920**	0.702	0.694	0.660	**0.728**	0.600	0.496	0.622	0.624
		특이도	**0.992**	0.982	0.939	0.807	**0.982**	0.942	0.911	0.875	0.954	0.923
	LS	민감도	0.248	**0.268**	0.096	**0.148**	0.088	0.076	0.098	0.130	0.066	0.082
		특이도	**0.992**	0.987	0.612	0.438	0.614	**0.624**	0.450	0.396	0.514	0.494
	TC	민감도	**0.848**	0.836	0.612	0.438	0.614	**0.624**	0.450	0.396	0.514	0.494
		특이도	0.978	0.974	0.948	0.852	**0.980**	0.948	0.920	0.894	0.956	0.933
$t_1 = 90$	IO	민감도	0.946	**0.948**	0.742	0.768	0.770	**0.782**	0.646	0.536	0.666	0.662
		특이도	**0.993**	0.983	0.933	0.807	**0.966**	0.941	0.909	0.873	0.950	0.922
	AO	민감도	0.942	**0.944**	0.758	**0.812**	0.778	0.808	0.690	0.608	0.718	0.704
		특이도	**0.991**	0.982	0.937	0.812	**0.970**	0.946	0.914	0.878	0.952	0.926
	LS	민감도	0.518	**0.528**	0.098	**0.184**	0.128	0.094	0.092	0.170	0.068	0.108
		특이도	**0.973**	0.961	0.947	0.842	**0.974**	0.955	0.926	0.892	0.959	0.938
	TC	민감도	**0.892**	0.872	0.670	0.682	0.700	**0.716**	0.562	0.496	0.572	0.624
		특이도	**0.978**	0.973	0.934	0.821	**0.965**	0.942	0.914	0.882	0.949	0.924

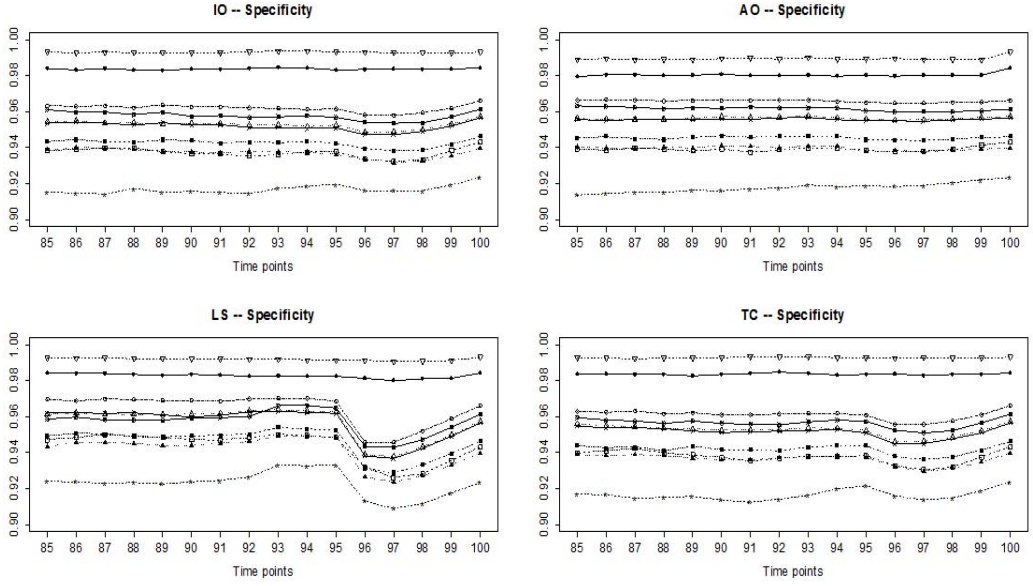
* 특이치가 존재하지 않는 경우, 민감도를 계산할 수 없어 N/A로 표시하였다.
IO = innovational outlier; AO = additive outlier; LS = level shift; TC = temporary change.

서는 AR(1) 모형과 MA(1) 모형만을 다루어보았다. 특이치가 두 개일 때에는 앞에서 살펴본 민감도와는 조금 다른 의미의 민감도로 세분화할 필요가 있다. 단순히 실제 특이치 두 개 중 몇 개를 발견했는가를 확인하기보다는 두 개의 특이치를 모두 발견했는가, 혹은 둘 중 하나만 발견했는가 등으로 나누어 살펴보아야하기 때문이다.

Table 3.3에는 AR(1) 기반 모형에서 특이치의 크기는 $\omega_1 = 5$와 $\omega_2 = 5$, 발생시점은 $t_1 = 90$와 $t_2 = 91$인 경우의 결과가 나타나 있다. 성능 판단 기준으로서 두 개의 특이치를 모두 발견한 비율을 우선적으로 생각해볼 때, Residual-based Algorithm이 뛰어난 민감도를 보이고 있다. 이는 특이치가 한 개 발생할 때의 모의실험 결과와 유사한 결과이다. 마찬가지로 특이치 발생 시점과 크기를 바꾸어가며 살펴본 결과, 대부분 Residual- based Algorithm이 좋은 민감도를 가지고 있으며, 특이도의 경우에는 0.9 이상으로 높은 특이도를 가지고 있다. LS 특이치가 포함된 경우들에서 상대적으로 낮은 민감도를 보이는 것 또한 특이치가 한 개 발생할 때의 이유와 유사하기 때문이라고 할 수 있다.



(a) 민감도



(b) 특이도

RES( ), BOX( ), T_{RM}-$\tilde{\sigma}_{MAD}$( ), T_{RM}-$\tilde{\sigma}_{LSH}$( ), T_{RM}-$\tilde{\sigma}_{QN}$( ),
T_{RM}-$\tilde{\sigma}_{SN}$( ), T_{LMS}-$\tilde{\sigma}_{MAD}$( ), T_{LMS}-$\tilde{\sigma}_{LSH}$( ), T_{LMS}-$\tilde{\sigma}_{QN}$( ), T_{LMS}-$\tilde{\sigma}_{SN}$( )

Figure 3.1. Specificity and sensitivity according to occurrence of outliers at AR(1) based model.

Table 3.3. AR(1) $\omega_1 = 5/\omega_2 = 5$, $t_1 = 90/t_2 = 91$

				RobFilter							
		RES	BOX	T_{RM}				T_{LMS}			
				$\tilde{\sigma}_{MAD}$	$\tilde{\sigma}_{LSH}$	$\tilde{\sigma}_{QN}$	$\tilde{\sigma}_{SN}$	$\tilde{\sigma}_{MAD}$	$\tilde{\sigma}_{LSH}$	$\tilde{\sigma}_{QN}$	$\tilde{\sigma}_{SN}$
IO	기준	**0.89**	0.78	0.32	0.30	0.31	0.35	0.35	0.39	0.30	0.38
TC	특이도	0.99	0.99	0.95	0.94	0.96	0.95	0.94	0.91	0.96	0.94
IO	기준	**0.92**	0.71	0.59	0.53	0.59	0.60	0.55	0.52	0.53	0.56
AO	특이도	0.99	0.98	0.96	0.94	0.96	0.95	0.94	0.92	0.96	0.95
IO	기준	**0.66**	0.60	0.06	0.08	0.04	0.05	0.08	0.11	0.04	0.08
LS	특이도	0.99	0.98	0.96	0.95	0.97	0.96	0.95	0.92	0.96	0.95
IO	기준	**0.90**	0.78	0.38	0.34	0.38	0.40	0.40	0.43	0.35	0.42
IO	특이도	0.99	0.99	0.95	0.94	0.96	0.95	0.94	0.91	0.96	0.94
AO	기준	0.24	0.23	**0.61**	0.55	**0.61**	**0.61**	0.56	0.55	0.56	0.58
AO	특이도	0.99	0.98	0.96	0.94	0.97	0.96	0.94	0.92	0.96	0.95
LS	기준	**0.85**	**0.85**	0.01	0.03	0.00	0.01	0.01	0.01	0.00	0.01
LS	특이도	0.99	0.99	0.97	0.95	0.97	0.96	0.95	0.93	0.95	0.96
TC	기준	**0.93**	0.83	0.28	0.24	0.23	0.27	0.33	0.36	0.26	0.34
TC	특이도	0.99	0.99	0.95	0.94	0.96	0.95	0.93	0.91	0.95	0.94

∗ 기준: 성능 판단 기준, 500회의 모의실험 중 두 개의 특이치를 모두 발견한 비율.

Table 3.4. AR(1) $\omega_1 = 5/\omega_2 = 5$, $t_1 = 90/t_2 = 95$

				RobFilter							
		RES	BOX	T_{RM}				T_{LMS}			
				$\tilde{\sigma}_{MAD}$	$\tilde{\sigma}_{LSH}$	$\tilde{\sigma}_{QN}$	$\tilde{\sigma}_{SN}$	$\tilde{\sigma}_{MAD}$	$\tilde{\sigma}_{LSH}$	$\tilde{\sigma}_{QN}$	$\tilde{\sigma}_{SN}$
IO	기준	0.91	**0.92**	0.30	0.33	0.27	0.30	0.32	0.33	0.26	0.32
TC	특이도	0.99	0.98	0.95	0.94	0.96	0.96	0.94	0.92	0.96	0.94
IO	기준	**0.92**	**0.92**	0.41	0.41	0.37	0.44	0.42	0.46	0.37	0.44
AO	특이도	0.99	0.98	0.95	0.94	0.96	0.95	0.94	0.92	0.96	0.94
IO	기준	0.87	**0.88**	0.11	0.08	0.09	0.10	0.09	0.11	0.07	0.09
LS	특이도	0.99	0.98	0.96	0.95	0.97	0.93	0.94	0.93	0.96	0.95
IO	기준	0.91	**0.94**	0.34	0.37	0.32	0.36	0.35	0.37	0.30	0.37
IO	특이도	0.99	0.98	0.95	0.94	0.96	0.95	0.94	0.92	0.96	0.94
AO	기준	0.91	**0.93**	0.62	0.57	0.61	0.63	0.55	0.54	0.54	0.55
AO	특이도	0.99	0.98	0.96	0.94	0.96	0.96	0.94	0.92	0.96	0.95
LS	기준	**0.87**	0.84	0.02	0.04	0.02	0.03	0.02	0.03	0.01	0.02
LS	특이도	0.99	0.98	0.96	0.95	0.97	0.96	0.95	0.93	0.97	0.95
TC	기준	**0.92**	**0.92**	0.24	0.23	0.21	0.23	0.28	0.31	0.20	0.29
TC	특이도	0.99	0.98	0.95	0.94	0.97	0.96	0.94	0.92	0.96	0.94

∗ 기준: 성능 판단 기준, 500회의 모의실험 중 두 개의 특이치를 모두 발견한 비율.

4. 적용 사례

4.1. 자료 설명

여기에서 적용할 사례는 동양시멘트의 주가 조작 사례이다. 서울중앙 지방검찰청 보도 자료에 따르면 2011년 12월 5일부터 2012년 3월 16일까지, 그리고 2013년 6월 27일부터 2013년 9월 10일까지 총 2차에 걸쳐 주가 조작이 일어났다. 이에 주요 관련자들은 132억 원 상당의 부당이득과 277억 원 상당 경제

Table 4.1. Outlier detection points at the 1^{st} manipulation period

해당 시점	발견된 특이치 시점							
2011/11/30	10/7							
2011/12/01	10/7							
2011/12/02	10/7	11/25						
2011/12/05	10/7	11/25	**12/5**					
2011/12/06	10/7	11/25	**12/5**	**12/6**				
2011/12/07	**12/5**	**12/6**	**12/7**					
2011/12/08	**12/5**	**12/6**	**12/7**	**12/8**				
2011/12/09	**12/5**	**12/6**	**12/7**					
2011/12/12	**12/5**	**12/6**	**12/7**	**12/8**	**12/9**	**12/12**		
2011/12/13	**12/5**	**12/6**	**12/7**	**12/8**	**12/12**			
2011/12/14	**12/5**	**12/6**	**12/7**	**12/8**	**12/9**	**12/12**	**12/13**	**12/14**
2011/12/15	**12/5**	**12/6**	**12/7**	**12/8**	**12/9**	**12/12**	**12/13**	
2011/12/16	**12/5**	**12/6**	**12/7**	**12/8**	**12/9**	**12/12**	**12/13**	**12/15**
2011/12/19	**12/5**	**12/6**	**12/7**	**12/8**	**12/9**	**12/12**	**12/14**	**12/14**

Table 4.2. Outlier detection points at the 2^{nd} manipulation period

해당 시점	발견된 특이치 시점		
2013/06/24	N/A		
2013/06/25	N/A		
2013/06/26	N/A		
2013/06/27	N/A		
2013/07/01	N/A		
2013/07/04	N/A		
2013/07/05	**7/4**		
2013/07/08	**7/4**		
2013/08/23	**7/4**		
2013/08/26	**7/4**		
2013/08/27	**7/2**	**7/4**	**8/20**
2013/08/28	**7/4**	**8/16**	

적 이익을 취득한 혐의로 기소되었다. 따라서 주가 조작이 일어나기 약 두 달 전부터(2011년 10월 4일) 2차시기 종료 후 두 달 후까지(2013년 11월 29일)의 자료를 바탕으로 적용해 보기로 한다. 주가 자료는 한국거래소(KRX)의 일자별 주가자료를 사용하였다 (http://www.krx.co.kr).

4.2. 알고리즘 적용 결과

3절 모의실험 결과에서 가장 좋은 민감도와 특이도를 보인 알고리즘인 Residual-based Algorithm을 중심으로 위의 자료에 적용해보도록 하겠다. 정상 시계열로는 기본적인 모형인 AR(1)을 사용하였다.

Table 4.1과 Table 4.2는 동양시멘트의 주가자료에 Residual-based Algorithm을 적용한 결과, 특이치로 발견된 시점을 정리한 것이다. 표의 첫 번째 열인 해당 시점을 기준으로 45 영업일 전부터의 데이터(즉, 해당 시점은 사용된 45 영업일 자료 중 마지막 날)를 사용하여 분석하였다.

Table 4.1을 보면, 2011년 12월 5일부터는 시점을 옮겨감에 따라 12월 5일부터의 날짜가 계속해서 발견되는 것을 알 수 있다. 이처럼 상당 기간 동안 특정 날짜가 반복적으로 발견된다면 주가가 조작되

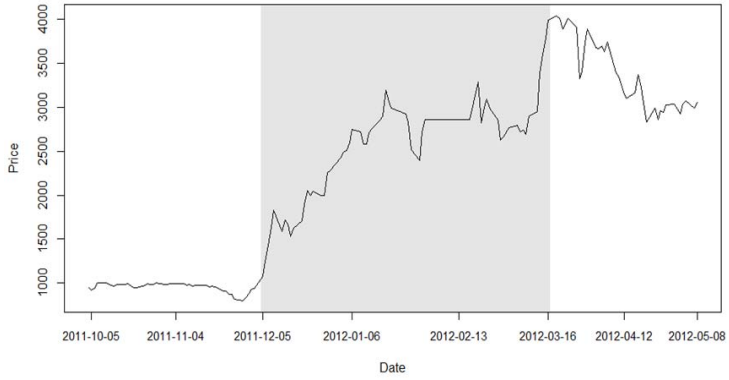


Figure 4.1. Time-series plot at the 1^{st} manipulation period.

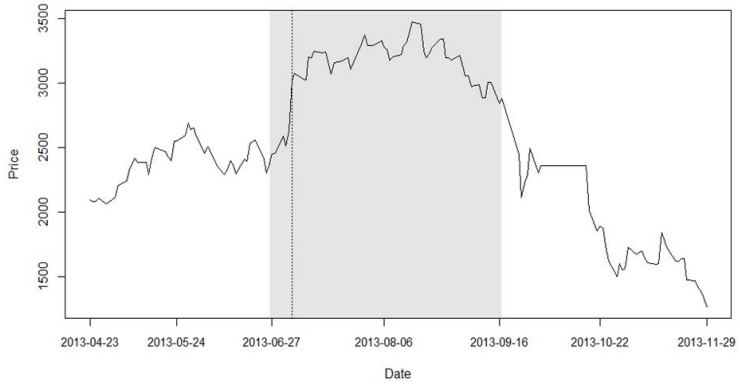


Figure 4.2. Time-series plot at the 2^{nd} manipulation period.

지는 않았는지 의심할 필요가 있다. Table 4.2을 보면, 주가 조작이 있기 약 3일 전부터 7월 4일까지는 어떤 특이치도 발견되지 않았고, 7월 5일부터 8월 26일까지는 7월 4일의 주가가 계속해서 특이치로 발견되었다. 2차 시기에서도 특정 날짜가 두 달 이상 반복해서 나타나는 것으로 보아, Residual-based Algorithm을 이용하여 주가 조작이 의심되는 징후를 포착할 수 있다고 결론지을 수 있다.

더불어, Figure 4.1는 2011년 10월 4일–2012년 5월 8일까지의 주가 흐름을 나타낸 것이며 회색 음영으로 처리된 기간이 주가 조작이 발생했던 기간이다. 해당 기간의 주가가 상당히 상승한 것을 알 수 있다. Figure 4.2는 2차 시기의 동양시멘트 주가흐름을 나타낸다. 회색 음영으로 표시된 기간은 2차 조작 기간이며, 점선으로 표시된 세로선은 7월 4일을 나타낸다. 조작 시기동안 주가가 상승하여 일정한 값을 유지하다가, 조작 시기가 끝난 뒤 다시 하락하는 모습을 보인다. 따라서 이는 LS 특이치에 해당한다고 할 수 있다.

5. 결론

본 연구에서는 시계열 데이터에서 특이치를 발견하는 알고리즘에 대해 알아보았다. 대표적인 비모수적 방법으로는 Fried (2004)가 제시한 Robust Filtering Algorithm을 들 수 있다. 이 알고리즘은 중위수를 활용한 국소적 선형적합을 기초로 한다. 시간에 따라 움직이는 윈도우 안에서 중위수를 이용하여 선

형추세적합을 함으로써 시계열 데이터에서 비모수적으로 특이치를 발견하는 방법을 제시하고 있다. 특히 선형추세를 적합하는 데 있어 총 여덟 가지의 옵션이 가능한데, 모의실험에서 모든 경우를 비교하여 살펴보았다.

더불어, 분위수 회귀와 시계열 분석을 접목한 분위수 자기회귀모형을 이용하여, 시계열 자료에서 특이치를 발견하는 새로운 알고리즘을 제시하였다. 분위수 회귀는 보통최소제곱추정에 비해 로버스트하다는 장점이 있다. 따라서 이에 자기상관성을 함께 고려한다면, 시계열 자료에서 특이치에 편향되지 않고 특이치를 발견해낼 수 있는 알고리즘이 가능할 것이다. Eo 등 (2014)가 제시한 절단된 자료에서의 특이치 발견 알고리즘을 시계열 자료에 적용한 두 가지 알고리즘을 새롭게 제시하였다.

본 연구의 모의실험을 통해, Fried (2004)의 여덟 가지 방법과 새롭게 제시한 두 가지 방법을 비교하여 여러 가지 특이치 상황에서 민감도와 특이도를 비교하였다. 발생한 특이치의 종류와 크기, 위치, 개수, 그리고 기본 모형의 종류에 따라 각각의 알고리즘이 얼마나 잘 특이치를 발견해내는지 살펴보았다. 그 결과 상당한 경우에서 분위수 자기회귀모형을 이용한 Residual-based Algorithm이 좋은 민감도와 특이도를 가지는 것을 확인하였다.

마지막으로 동양시멘트 주가 조작 사례를 통해 Residual-based Algorithm이 실제 상황에도 잘 적용될 수 있는지 알아보았다. 1차와 2차에 걸친 주가 조작 시기 동안 본 연구에서 제시한 알고리즘을 이용하여 최대한 빨리 조작을 적발할 수 있는지 살펴본 결과, 두 시기 모두에서 상당히 빠른 시일 내에 특이치 시점을 찾아낸 것을 알 수 있었다. 주가 조작 적발의 경우 의심될 만한 징후를 최대한 빠른 시일 내에 발견하는 것이 매우 중요하기 때문에 본 알고리즘이 유용하게 사용될 수 있으리라 기대한다.

References

Chen, C. and Liu, L. M. (1993a). Forecasting time series with outliers, *Journal of Forecasting*, **12**, 13–35.

Chen, C. and Liu, L. M. (1993b). Joint estimation of model parameters and outlier effects in time series, *Journal of the American Statistical Association*, **88**, 284–297.

Eo, S. H., Hong, S. M., and Cho, H. (2014). Identification of outlying observations with quantile regression for censored data. arXiv preprint arXiv:1404.7710.

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

Fried, R. (2004). Robust filtering of time series with trends, *Journal of Nonparametric Statistics*, **16**, 313–328.

Hoaglin, D. C., Iglewicz, B., and Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling, *Journal of the American Statistical Association*, **81**, 991–999.

Koenker, R. and Xiao, Z. (2006). Quantile autoregression, *Journal of the American Statistical Association*, **101**, 980–990.

Nardi, A. and Schemper, M. (1999). New residuals for cox regression and their application to outlier screening, *Biometrics*, **55**, 523–529.

Rousseeuw, P. J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871–880.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation, *Journal of the American Statistical Association*, **88**, 1273–1283.

Rousseeuw, P. J. and Leroy, A. M. (1988). A robust scale estimator based on the shortest half, *Statistica Neerlandica*, **42**, 103–116.

SAS Institute Inc (2008). *SAS/STAT 9.2 User's Guide: the QUANTREG Procedure*, Cary, SAS Institute Inc, NC.

Siegel, A. F. (1982). Robust regression using repeated medians, *Biometrika*, **69**, 242–244.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley, MA.

시계열 자료에서의 특이치 발견

최정인[a] · 엄인옥[a] · 조형준[a,1]

[a]고려대학교 통계학과

(2016년 5월 25일 접수, 2016년 7월 25일 수정, 2016년 8월 2일 채택)

요 약

본 논문의 목표는 분위수 자기회귀모형을 활용하여 시계열 자료에서 특이치를 발견하는 알고리즘을 제안하고, 기존의 방법들과 그 성능을 비교하여 실제 주가 조작 사례에 적용해 보는 것이다. 지금까지의 특이치 발견 연구는 대부분 일반적인 데이터 형태에서만 있어왔기 때문에 시계열 데이터에서의 연구는 미미한 편이다. 또한 모수적인 방법에만 제한되었는데, 모수적 모형은 복잡할 뿐만 아니라 소요되는 분석 시간도 길기 때문에 편리하지 않다. 따라서 본 연구에서는 분위수 자기회귀모형을 활용한 특이치 발견 알고리즘을 새롭게 제시하고, 다양한 경우의 모의실험을 통해 기존 알고리즘과 비교하도록 한다. 특히 시계열 자료에서의 특이치 발견은 주가 조작을 적발하는 데에 유용하게 활용될 수 있다. 시간에 따라 관측되던 주가가 갑자기 그 동안의 흐름에서 벗어나 특이치로 발견되었다면 혹시 인위적인 개입으로 조작된 것은 아닌지 의심해 볼 수 있기 때문이다. 따라서 실제 주가 조작 사례에 적용해 봄으로써 얼마나 빠른 시일 내에 주가 조작을 적발해 낼 수 있는지 살펴보았다.

주요용어: 특이치 발견, 분위수 자기회귀모형, 시계열 자료

이 논문은 제1저자 최정인의 석사학위논문의 일부를 발췌한 것임.

이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2015R1D1A1A09058602).

[1]교신저자: (02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과. E-mail: hj4cho@korea.ac.kr