The Pure and Applied Mathematics 2 (1995), No 1, pp. 67-82
J. Korea Soc. of Math. Edu. (Series B)

이상치가 존재하는 시계열모형 설정에 관한 연구

최 창 호 (강남대학교) · 박 천 건 (중앙대학교 대학원)

1. 서 론

경제학에서 분석하는 연도별 국민총생산액, 월별 소비자물가지수, 경영학에서 분석하는 어느 제품의 월별 판매량, 특정주식의 일별 종가 및 거래량, 기상학에서 관찰되는 일별 최고온도 및 최저온도, 태풍의 경로, 등등 여러 학문분야에서 접할 수 있는 통계자료들은 시간이 흐름에 따라 변하는 시계열자료(time series data)들이 많다. 따라서 대부분의 학문분야에서 시계열 분석이 필요하다.

우리가 이러한 시계열자료들을 분석하는 목적은 크게 두 가지이다. 첫째는 시계열자료를 관찰하고 분석함으로써 주어진 자료의 확률적 체계를 이해하고 모형화하는 것이고, 둘째는 과거의 자료를 가지고 미래의 값을 예측하는 것이다.

시계열자료는 어떤 확률적 구조를 가지고 있다. 그런데 파업, 전쟁의 발생, 갑작스런 정치·경제적 위기, 기록상의 오류와 같은 외부의 충격(interruptive events)으로 시계열자료의 확률적 구조와 불일치하는 가짜관찰치(spurious observations)들이 생긴다. 이러한 관찰치들을 이상치(outlier)라고 부른다. 본 논문에서는 이상치들(outliers)이 존재하는 시점을 탐색하고 그 효과를 측정하여 이상치가 존재하는 시계열자료를 모형화하는데 목적이 있다.

본 논문에서는 앞에서 언급한 연구목적을 얻기 위하여 이상치가 존재하는 시계열자료에 대해서 아래와 같은 반복적인 방법을 연구하였다.

다음은 이상치들을 탐색하고 그 효과를 측정하여 이상치가 존재하는 시계열자료를 모형화하는 절차이다.

단계 1: 1차자기상관 ρ_k에 대한 영향함수행렬(Influence Function Matrix)을 이용하여 이상치들이 발생한 시점을 찾는다.

단계 2: 이상치들이 존재하는 시점의 시계열자료들을 1차자기상관계수를 이용하여 그 효과를 잠정적으로 제거한다. .

단계 3: 이상치들의 효과가 잠정적으로 제거된 시계열자료를 확장된 표본자기상관함수(Extended Sample Autocorrelation Function)를 이용해서 혼합모형(ARMA)의 차수가 (p,q)인 잠정적인 모형을 설정한다.

단계 4: 단계 3에서 식별된 혼합모형(ARMA)의 모수를 최소제곱추정법(Gauss-Newton 알고리즘)을 이용해서 추정한다.

단계 5: 단계4에서 추정된 모수들을 이용하여 이상치들을 탐색하고 그 효과를 측정하여 이상치가 존재하는 시계열자료에서 측정된 효과를 제거한다.

이상치가 모두 제거될 때까지 단계 5의 절차를 반복하여 이상치가 존재하는 시계열자료를 모형화한다.

2 . 영향함수

다변량 자료에서 Devlin, Gnanadesikan 그리고 Kettenring(1975)은 이변량 상관에 대한 영향함수(Influence Function)가 이상치들을 탐색하기 위한 도구로써 유용하다는 것을 보여주었다.

시계열에서 차수를 결정하는 표본자기상관함수(Sample Autocorrelation Function: SACF), 표본편자기상관함수(Sample Partial Autocorrelation Function:SPACF), 그리고 확장된 표본자기상관함수(Extended Sample Autocorrelation Function:ESACF)와 같은 식별통계량(Identification Statistic)은 자기상관 ρ_k에 많이 의존한다. 그런데 이상치들(outliers)은 식별통계량에 많은 영향을 미친다. 2.1절에서는 자기상관 ρ_k에 관한 영향함수행렬(Influence Function Matrix)을 이용해서 시계열에서 이상치들의 영향력을 조사하는 방법을 제시하고, 2.2절에서는 2.1절에서 탐색된 이상치들의 효과를 잠정적으로 제거하는 방법을 제시한다.

2.1 영향함수행렬

Hampel(1974)에 의해 정의된 영향함수는 우극한이 존재할 때 다음과 같이 정의한다.

$$I(F,T(F),x) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon\delta x) - T(F)}{\varepsilon} \qquad (2.1.1)$$

단, F : 분포함수
T(F) : 분포함수 F의 모함수
x : 관찰치들 중에 관심이 가는 관찰치의 위치
δx : 위치 x의 분포함수

ρ_k를 정상시계열 $\{X_t\}$, $t=1,2,\cdots$의 자기상관이라하고 $\mu = E(X_t)$이고 $\sigma^2 = Var(X_t)$라 하자. 그러면 X_t를 $Y_t = (X_t - \mu)/\sigma$, $t=1,2,\cdots$으로 변환해도 ρ_k에 아무런 영향을 미치지 않으므로 $I(F,\rho_k,x) = I(H,\rho_k,y)$이다.

<보조정리 2.1>

$$I(H,\rho_k,(z_t, z_{t+k})) = z_t z_{t+k} - \rho_k(z_t^2 + z_{t+k}^2)/2,$$

단, H는 $E(H)=0$, $Var(H)=1$과 공분산 ρ_k을 갖는 (z_t, z_{t+k})의 이변량분포이다.

<보조정리 2.1>를 다음과 같이 정리할 수 있다.

$$I(H,\rho_k,(z_t, z_{t+k})) = z_t z_{t+k} - \rho_k(z_t^2 + z_{t+k}^2)/2$$

$$= (1-\rho_k^2)U_{i,k,1}U_{i,k,2}$$

단, $$U_{i,k,1} = \{\frac{(y_i + y_{i+k})}{\sqrt{1+\rho_k}} + \frac{(y_i - y_{i+k})}{\sqrt{1-\rho_k}}\}/2$$

$$U_{i,k,2} = \{\frac{(y_i + y_{i+k})}{\sqrt{1+\rho_k}} - \frac{(y_i - y_{i+k})}{\sqrt{1-\rho_k}}\}/2.$$

정상가우시안과정(Stationay Gaussian Process) 즉, 정상시계열 $\{X_t\}$의 평균 μ, 표준편차 σ, 자기상관 ρ_k을 알면 $U_{i,k,1}$과 $U_{i,k,2}$는 각 시차 k에 대해 독립이고 표준정

규분포를 따른다. 따라서 영향함수 $I(H,\rho_k,(y_t,y_{t+k}))$는 표준정규분포에 상수배한 분포이다. 영향함수의 분포는 정상시계열 $\{X_t\}$의 관찰치에 대한 영향함수의 값가운데 영향력이 특히 큰 값을 찾는데 사용된다.

만약 시점 k에서 이상치가 있다면 영향함수행렬의 구조는 <그림2.1>과 같다.

<그림2.1> 영향함수행렬의 구조

시점 \ 시차	1	2	3	4	5	6	7	···	k-1	···	m
1								···	X		
2							X				
⋮						X					
⋮					X						
⋮				X							
k-3			X								
k-2		X									
k-1	X										
k	X	X	X	X	X	X	X	X	X	X	X
k+1											
⋮											
n											

$$\text{단, } X=\begin{cases} + & , \ I(H,\rho_k,(y_{t,y_{t-k}})) > 1 \\ - & , \ I(H,\rho_k,(y_{t,y_{t-k}})) < -1 \\ \text{공란} & , \ |I(H,\rho_k,(y_{t,y_{t-k}}))| < 1 \end{cases}$$

임계값(critical value) 1은 영향함수분포 즉, 상수배한 표준정규분포를 근거로 한 것이다. <그림2.1>은 옷핀모양을 하고있고 이상치는 시점k번째의 관찰치이다.

2.2 1차자기상관계수로 이상치 효과의 잠정적인 제거

자기상관계수는 동일한 변수를 시점을 달리하여 관찰했을 때 시점이 서로 다른 관찰치 사이에 존재하는 상호작용 관계를 나타내고 시계열의 확률적 구조를 분석하는데 사용된다. 이때, 시차 k를 가진 동일 변수 사이의 자기상관계수를 다음과 같이 정의한다.

$$\rho_1 = \frac{E(Z_t-\mu)(Z_{t+k}-\mu)}{\sqrt{Var(Z_t)Var(Z_t)}}, \qquad (2.2.1)$$

단, $E(Z_t)=\mu,\ t=1,2,3,\cdots$

시간이 흐름에 따라 관찰되는 시계열 자료는 다른 시점 사이의 관찰치가 서로 독립이 아니고 종속이다. 그래서 관찰치 Z_k는 시점k이전의 관찰치 $Z_{k-1}, Z_{k-2}, Z_{k-3}, \cdots$에 영향을 받는다. 따라서 시점k번째에서 이상치가 탐색되었다면 다음과 같이 Z_k를 수정할 수 있다.

$$(Z_t-\overline{Z}) = \rho_1(Z_{t-1}-\overline{Z}) \qquad (2.2.2)$$

단, $\overline{Z} = \frac{1}{n}\sum_{t=1}^{n} Z_t$

$$\rho_1 = \frac{\sum_{t=1}^{n-1}(Z_t-\overline{Z})(Z_{t+1}-\overline{Z})}{\sum_{t=1}^{n}(Z_t-\overline{Z})}.$$

즉, 시계열 자료는 자기상관이 있으므로 시점 k번째의 관찰치 Z_k는 과거의 관찰치 Z_{k-1} , Z_{k-2} , Z_{k-3} , …로 나타낼 수 있다. 특히 시점 k번째의 관찰치 Z_k와 전시점 k-1번째 관찰치 Z_{k-1}사이에 1차자기상관계가 존재하므로 Z_k를 Z_{k-1}에 1차자기상관계의 곱으로 잠정적으로 나타낼 수 있다. 따라서 식(2.2.2)로 이상치의 효과를 잠정적으로 제거할 수 있다.

3. 시계열의 이상치

3.1 가법모형과 혁신모형

$\{X_t\}$가 다음과 같은 ARMA(p,q)모형을 따른다고 가정한다.

$$\phi(B)X_t=\theta(B)a_t \tag{3.1.1}$$

단, $\phi(B)=1-\phi_1B-\cdots-\phi_pB^p$, $\theta(B)=1-\theta_1B-\cdots-\theta_qB^q$, $\{a_t\}$는 백색잡음.

가법모형은 다음과 같이 정의한다.

$$Z_t = \begin{cases} X_t & ,t\neq T \\ X_t+w & ,t=T \end{cases} \tag{3.1.2a}$$

$$= X_t + wI_t^{(T)} \tag{3.1.2b}$$

$$= \frac{\theta(B)}{\phi(B)}a_t + wI_t^{(T)} \tag{3.1.2c}$$

단, $I_t^{(T)} = \begin{cases} 1 & ,t=T \\ 0 & ,t\neq T \end{cases}$

$I_t^{(T)}$는 T시점에서 이상치의 존재여부를 나타내는 지시변수(indicator variable)이다.

혁신모형은 다음과 같이 정의한다.

$$Z_t = X_t + \frac{\theta(B)}{\phi(B)}wI_t^{(T)} \tag{3.1.3a}$$

$$= \frac{\theta(B)}{\phi(B)}(a_t + wI_t^{(T)}) \tag{3.1.3b}$$

결국 가법이상치는 시점 T번째 관찰치에만 영향을 미치는 반면 혁신이상치는

$\frac{\theta(B)}{\phi(B)}$로 설명되는 가중치로 시점 T번째 이후의 관찰치 $Z_T, Z_{T+1}, \cdots$ 에 영향을 미친다.

일반적으로 몇개의 이상치를 포함하는 시계열 모형은 다음과 같이 정의된다.

$$Z_t = \sum_{j=1}^{k} W_j V_j(B) I_t^{(t_j)} + X_t \qquad (3.1.4)$$

단, $X_t = \frac{\theta(B)}{\phi(B)} a_t$

$V_j(B) = 1$, AO 시점 $t = T_j$

$V_j(B) = \frac{\theta(B)}{\phi(B)}$, IO 시점 $t = T_j$

3.2 이상치들의 시점을 알 때 이상치들의 효과 추정

가법이상치(Additive Outliers : AO)와 혁신이상치(Innovation Outliers : IO)를 탐색하는 절차를 확인하기 위하여 식(3.1.1)의 모수와 시점을 안다고 하고 $\frac{\phi(B)}{\theta(B)}$와 e_t를 다음과 같이 정의한다.

$$\pi(B) = \frac{\phi(B)}{\theta(B)} = (1 - \pi_1 B - \pi_2 B^2 - \cdots) \qquad (3.2.1)$$

$$e_t = \pi(B) Z_t \qquad (3.2.2)$$

식(3.1.2c)와 (3.1.3a)를 다음과 같이 나타낼 수 있다.

$$\text{AO} : \quad e_t = w\,\pi(B) I_t^{(T)} + a_t \qquad (3.2.3)$$

$$\text{IO} : \quad e_t = w\, I_t^{(T)} + a_t \qquad (3.2.4)$$

그러면 n개 관찰치에 대해 식(3.2.3)의 가법모형을 다음과 같이 쓸 수 있다.

$$\begin{pmatrix} e_1 \\ \vdots \\ e_{T-1} \\ e_T \\ e_{T+1} \\ e_{T+2} \\ \vdots \\ e_n \end{pmatrix} = w \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ -\pi_1 \\ -\pi_2 \\ \vdots \\ -\pi_{n-T} \end{pmatrix} + \begin{pmatrix} a_1 \\ \vdots \\ a_{T-1} \\ a_T \\ a_{T+1} \\ a_{T+2} \\ \vdots \\ a_n \end{pmatrix} \tag{3.2.5}$$

가법모형에서 w의 최소제곱추정치를 $\hat{w}_{AT}$라 하고 식(3.2.5)로부터 이를 구하면 다음과 같다.

$$\text{AO}: \quad \hat{w}_{AT} = \frac{e_T - \sum_{j=1}^{n-T} \pi_j e_{T+j}}{\sum_{j=0}^{n-T} \pi_j^2}$$

$$= \frac{\pi^*(F) e_T}{\gamma^2} \tag{3.2.6}$$

단, $\pi^*(F) = (1 - \pi_1 F - \pi_2 F^2 - \cdots - \pi_{n-T} F^{n-T})$,

$F : Fe_t = e_{t+1}$, $\gamma^2 = \sum_{j=0}^{n-T} \pi_j^2$.

추정치의 분산은 다음과 같다.

$$\text{Var}(\hat{w}_{AT}) = \text{Var}\left(\frac{\pi^*(F) e_T}{\gamma^2}\right)$$

$$= \frac{1}{\gamma^4} \text{Var}(\pi^*(F) a_T)$$

$$= \frac{\sigma_a^2}{\gamma^2} \tag{3.2.7}$$

위와 마찬가지로, 혁신 모형에서 w의 최소제곱추정치 $\hat{w}_{IT}$을 구하면 다음과 같다.

$$IO : \quad \hat{w}_{IT} = e_T \tag{3.2.8}$$

$$Var(\hat{w}_{IT}) = Var(e_T) = Var(w\, I_t^{(T)} + a_T) = \sigma_a^2 \tag{3.2.9}$$

위의 식으로부터 시점 T에서 IO효과의 최량추정치(best estimate)는 잔차 e_T인 반면 AO효과의 최량추정치는 $e_T, e_{T+1}, e_{T+2}, \cdots, e_n$의 선형결합이다.

따라서 $Var(\hat{w}_{AT}) \leq Var(\hat{w}_{IT}) = \sigma_a^2$되고 어떤 경우에는 $Var(\hat{w}_{AT})$가 σ_a^2보다 매우 작은 값이 될 수 있다.

Z_T에 대한 다음의 가설

H_0 : Z_T 는 AO 도 IO 도 아니다.

H_1 : Z_T 는 AO 이다.

H_2 : Z_T 는 IO 이다.

은 가법이상치와 혁신이상치의 우도비검정통계량

$$H_1 \text{ vs } H_0 : \lambda_{1,T} = \frac{\gamma\, \hat{w}_{AT}}{\sigma_a} \tag{3.2.10}$$

$$H_2 \text{ vs } H_0 : \lambda_{2,T} = \frac{\hat{w}_{IT}}{\sigma_a} \tag{3.2.11}$$

에 의하여 검정이 가능하여지며 여기서 H_0가 사실이면 $\lambda_{1,T}$와 $\lambda_{2,T}$는 N(0,1)분포를 따른다.

3.3 반복적인 절차를 통한 이상치 탐색과 효과 추정

영향함수, 1차자기상관계수, Chang과 Tiao(1983)가 제시한 가법이상치와 혁신이상치의 개수를 모를 때 이상치를 탐색하는 반복적인 절차를 근거로 하여 다음과 같이

이상치를 탐색하는 반복적인 절차를 제시한다.

시계열 $\{Z_T\}$가 ARMA(p,q) 모형을 따른다고 가정한다.

단계 1 : 2.1절의 자기상관 ρ_k에 대한 영향함수행렬을 이용하여 이상치들의 시점을 탐색한다.

단계 2 : 탐색된 이상치들을 2.2절의 1차자기상관계수로 그 효과를 잠정적으로 제거한다.

단계 3 : 단계2에서 이상치들의 효과가 잠정적으로 제거된 시계열을 ESACF로 모형을 식별하고 최소제곱법(Gauss-Newton알고리즘)으로 모수를 추정한다.

단계 4 : 단계3으로부터 추정된 잔차는 다음과 같다.

$$\hat{e}_t = \hat{\pi}(B)Z_t = \frac{\hat{\phi}(B)}{\hat{\theta}(B)}Z_t \qquad (3.3.1)$$

단, $\hat{\phi}(B)=(1-\hat{\phi}_1B-\cdots-\hat{\phi}_pB^p)$

$\hat{\theta}(B)=(1-\hat{\theta}_1B-\cdots-\hat{\theta}_qB^q)$

그리고 $\hat{\sigma}_a^2=\frac{1}{n}\sum_{t=1}^{n}\hat{e}_t^2$은 σ_a^2의 초기 추정치이다.

단계 5 : 추정된 모형을 통해 $\hat{\lambda}_{1,t}$와 $\hat{\lambda}_{2,t}$를 $t=1,2,\cdots,n$에 대해 계산하고 $\hat{\lambda}_T$를 다음과 같이 정의한다.

$$\hat{\lambda}_T = \max_t \max_i \{|\hat{\lambda}_{i,t}|\} \qquad (3.3.2)$$

단, T는 이상치의 효과가 최대로 일어났을 때의 시점.

만약 $\hat{\lambda}_T = |\hat{\lambda}_{1,T}| > c$이면 Z_T는 $\hat{w}_{AT}$의 추정 효과를 가지는 시점 T에서의 가법이상치이다. C는 3과 4사이의 양수이다.

따라서 식(3.1.2b)을 수정하면 다음과 같이 되고

$$\tilde{Z}_t = Z_t - \hat{w}_{AT}I_t^{(T)} \qquad (3.3.3)$$

단, 식(3.2.3)의 새로운 잔차는 다음과 같이 된다.

$$\hat{e}_t = \hat{e}_t - \hat{w}_{AT}\hat{\pi}(B)I_t^{(T)} \qquad (3.3.4)$$

따라서 만약 $\hat{\lambda}_T = |\hat{\lambda}_{2,T}| > c$이면 Z_T는 $\hat{w}_{IT}$의 효과를 가지는 시점 T에서의 혁신이상치이다.

따라서 식(3.1.3a)을 수정하면 다음과 같이 되고

$$\tilde{Z}_t = Z_t - \frac{\hat{\theta}(B)}{\hat{\phi}(B)} \hat{w}_{IT} I_t^{(T)} \tag{3.3.5}$$

식(3.2.4)의 새로운 잔차는 다음과 같이 된다.

$$\hat{e}_t = \hat{e}_t - \hat{w}_{IT} I_t^{(T)} \tag{3.3.6}$$

단계 6 : 수정된 잔차로 다시 $\hat{\lambda}_{1,t}$와 $\hat{\lambda}_{2,t}$를 계산해서 모든 이상치를 식별할때까지 단계5를 반복 수행한다. 모수 $\pi(B)$은 단계3에서 추정한것을 그대로 사용한다.

단계 7 : 단계 6으로부터 k개의 이상치가 식별되었다고 가정한다.

그러면 이상치의 효과와 시계열 모수는 다음과 같은 모형을 사용하여 동시에 추정한다.

$$Z_t = \sum_{j=1}^{k} w_j V_j(B) I_j^{(T)} + \frac{\theta(B)}{\phi(B)} a_t \tag{3.3.7}$$

단, $V_j(B) = 1$는 $t = T_j$에서의 가법이상치.

$V_j(B) = \frac{\theta(B)}{\phi(B)}$는 $t = T_j$에서의 혁신이상치.

위의 식으로부터 새로운 잔차는 다음과 같다.

$$\hat{e}_t^{(1)} = \hat{\pi}^{(1)}(B)[Z_t - \sum_{j=1}^{k} \hat{w}_j V_j(B) I_t^{(T_j)}] \tag{3.3.8}$$

모든 이상치가 탐색되고 그 효과가 추정될때까지 단계 5부터 단계 7까지의 절차를 반복수행한다. 따라서 최종적으로 이상치가 존재하는 시계열 모형은 다음과 같이 된다.

$$Z_t = \sum_{j=1}^{k} \hat{w}_j V_j(B) I_j^{(T)} + \frac{\hat{\theta}(B)}{\hat{\phi}(B)} a_t \tag{3.3.9}$$

단, $\hat{\phi}(B) = (1 - \hat{\phi}_1 B - \cdots - \hat{\phi}_p B^p)$,

$\hat{\theta}(B) = (1 - \hat{\theta}_1 B - \cdots - \hat{\theta}_q B^q)$.

4 . 적용사례

3.3절에서 제시한 반복적인 절차의 효율성을 검사하기 위하여, ϕ=0.6인 AR(1)모형을 따르는 정상시계열 $\{Z_t\}$가 아래와 같은 두가지 경우의 이상치들에 대한 분석을 실시하였다.

<예 1>

Z_{14}=1.208975, Z_{32}=0.3928414가 기록오차로 인하여 Z_{14}=9.208976 , Z_{32}=8.3928414로 시계열자료가 변경되었을 경우, 이상치들을 탐색하고 이상치가 존재하는 시계열을 모형화는 하는 절차는 다음과 같다.

단계 1: 이상치가 존재하는 시계열자료의 영향함수행렬은 다음과 같다.

<표1> 영향함수행렬

시점 \ 시차	1	2	3	4	5	6	7
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	2
8	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0
12	0	1	0	0	0	0	0
13	2	0	0	0	0	0	0
14	2	2	1	1	1	1	1
15	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0
25	1	1	0	0	0	0	2
26	1	0	0	0	0	2	0
27	0	0	0	0	2	0	0
28	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0
31	2	0	0	0	0	0	0
32	2	2	2	0	1	1	1
33	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0

<표1>을 분석한 결과 시점 14, 32에서 이상치가 존재한다고 할 수 있다.

단계 2 : 단계1에서 탐색한 시점14, 32의 이상치들의 시계열자료를 1차자기상관계수($\rho_1 = 0.2380044$)로 수정한 수치는 $Z_{14}=0.3803378$, $Z_{32}=0.3002427$이다.

3.3절의 나머지 절차를 수행하면 다음과 같은 최종적인 결과를 얻는다.

반복	이상치 평가			
	시점	형태	효과크기($\hat{w}$)	Var($\hat{w}$)
1	14	AO	8.547	1.89
2	32	AO	8.101	0.58

따라서 3.3절의 반복적인 절차로 정확하게 이상치의 시점과 형태를 찾아냈다.
그리고 이상치가 존재하는 시계열자료의 모형은 다음과 같다.

$$Z_t = \underset{(1.89)}{8.55\, I_t^{(14)}} + \underset{(0.58)}{8.10\, I_t^{(32)}} + \underset{(0.017)}{\frac{1}{1-0.594\,B}}\, a_t$$

<예 2>

$Z_{26}=-2.919735$가 기록오차로 인하여 $Z_{26}=2.919735$로 시계열자료가 변경되었을 경우, 이상치들을 탐색하고 이상치가 존재하는 시계열을 모형화는 하는 절차는 다음과 같다.

단계 1: 이상치가 존재하는 시계열자료의 영향함수행렬은 <표2>과 같다.

<표2> 영향함수행렬

	1 2 3 4 5 6 7
13	0 0 0 0 0 0 0
14	0 2 0 0 1 1 1
15	0 0 0 1 1 1 0
16	0 0 2 2 2 0 0
17	0 0 0 1 0 0 0
18	0 0 0 0 0 0 2
19	1 1 0 0 0 2 1
20	1 0 0 0 2 1 2
21	0 0 0 2 1 2 2
22	0 0 2 1 0 0 0
23	0 0 0 0 0 0 0
24	0 2 0 0 0 0 0
25	2 1 0 2 2 0 2
26	2 2 1 1 1 0 0
27	0 2 2 0 0 0 1
28	0 0 0 0 0 0 0

	1 2 3 4 5 6 7
29	0 0 0 0 0 0 0
31	0 0 0 0 0 0 0
30	0 0 0 0 0 0 0
32	0 0 0 0 0 0 0
33	0 0 0 0 0 0 0
34	1 0 2 2 2 2 1
35	0 2 0 2 2 1 0
36	0 0 0 0 0 0 0
37	0 0 0 2 0 2 2
38	0 0 2 0 0 2 2
39	1 2 2 2 2 2 2
40	2 2 2 2 0 0 0
41	1 1 1 1 1 0 0
42	1 1 0 1 0 0 0
43	1 1 1 0 0 0 0

<표2>을 분석한 결과 시점 26, 40, 41에서 이상치가 존재한다고 할 수 있다.

단계2 : 단계1에서 탐색한 시점 26,40,41번째 이상치들의 시계열자료를 1차자기상관계수(ρ_1 = 0.184205)로 수정한 수치는 Z_{26}=-0.41095, Z_{40} = 0.2984785(1.2562582), Z_{41} = 0.1010274 (-1.521501)이다.

3.3절의 나머지 절차를 수행하면 다음과 같은 최종적인 결과를 얻는다.

반복	이상치 평가			
	시점	형태	효과크기(w)	Var($\hat{w}$)
1	26	AO	4.734	0.5827

따라서 반복적인 절차로 정확하게 이상치의 시점과 형태를 찾아냈다.
그리고 이상치가 존재하는 시계열자료의 모형은 다음과 같다.

$$Z_t = \underset{(0.58)}{4.73\, I_t^{(26)1}} + \underset{(0.02)}{\frac{1}{1-0.553\,B}}\, a_t$$

5. 결 론

어떤 시계열자료가 주어졌을때, 여러 통계학자들은 이상치의 유무와 이상치가 있다면 그 시점과 효과 그리고 형태를 평가하는 방법을 많이 연구했으며 나름대로 타당성있는 방법을 제시하였다.
본 논문은 이상치가 존재하는 시계열을 모형화하기 위해 3.3절의 반복적인 절차를 제시하였고, 제4장의 적용사례를 통하여 다음과 같은 결론을 얻었다.

결과 1 : 적용사례의 예제1과 같이 이상치들을 정상시계열에 위반해서 주었을 경우에는 영향함수행렬을 이용하여 이상치들을 정확하게 탐색하였다. 그러나 예제2와 같이 정상시계열를 따르는 범위 내에서 이상치의 효과를 주었을 경우에는 영향함수행렬을 이용한 이상치들의 탐색이 부정확하였다.

결과 2 : 이상치들은 시계열자료의 모형를 식별하는데 중요한 역할을 하는 식별통계량과 모수추정에 영향을 미친다. 그래서 단계2에서 이상치들의 시점을 알면 1차자기상관계수로 잠정적으로 이상치들의 효과를 제거하였다. 그런데 자기상관계수는 이상치들이 많을수록 0에 가까워지므로 이상치들을 3.2절의 방법으로 수정를 하면 시계열 자료의 표본평균에 가까워지는 경향이 있다.

위의 결과로 다음과 같은 결론을 내릴수 있다. 결과1로부터 이상치들을 탐색하는데 있어 예제1과 예제2의 두 경우 모두에 효율적인 방법이 있어야 하겠다. 그리고 결과2로 부터 이상치의 시점을 알 때 그것을 본래의 자료와 근사하게 적합시킬 수 있는 효율적인 방법이 있으면 반복적인 절차를 통하여 이상치들을 정확하게 평가할 수 있을 것이다

< 참 고 문 헌 >

1. Box, G.E.P., and Jenkins, G.M.(1970), Time Series Analysis, Forecasting and Control, SanFrancisco, Holden-Day.
2. Gnanadesikan, R.(1977), Methods for Statistical Data Analysis of Multivariate Observations, New York: John Wiley.
3. Michael R. Chernick, Darryl I. Downing, and David H.Pike(1982), Detecting Outliers in time Series Data, Journal of American Statistical Association.
4. Ruey S. Tsay(1986), Time Series Model Specification in the Presence of outliers, Journal of American Statistical Association.
5. William, W.S.Wei(1990), Time Series Analysis, Addison-Wesley Publishing Company.