The Korean Journal of Applied Statistics
2023, Vol. 36, No. 3, 197–209

DOI: https://doi.org/10.5351/KJAS.2023.36.3.197
Print ISSN1225-066X / Online ISSN 2383-5818

Detecting unusual observations in time series: An application of the normal-exp model

Ro Jin Pak[1,a]

[a]Department of Information Statistics, Dankook University

Abstract

A method to detect unusual or abnormal data in a time series or a discrete time signal is proposed. The idea of microarray background correction has been borrowed to detect abnormal data in a time series. Background correction to isolate anomalous signals or noise from the mircroarray's real signal is a very important step in tuning the data for ambient intensity surrounding each feature. The normal-exponential distribution was proposed to model background noise and signal by convoluting the exponential and the normal distribution. It is tried to model the error terms of a time series with the normal-exponential distribution. Once the residual components were well treated by the normal-exponential distribution, the observations with unexpectedly large residuals are detected as outliers or unusual observations. The marriage event data and the real estate price index data were considered for empirical studies and we were able to find several anomalous observations consistent with when the Korean economy or society actually underwent dramatic changes.

Keywords: microarray, normal-exponential distribution, time series, unusual data

1. 서론

시계열 혹은 이산시간 신호의 이상치를 발견하는 문제는 오래전부터 관심의 대상이었다. 본 연구에서도 시계열 속에 존재할지 모르는 이상치 혹은 비정상 데이터를 탐지하는 또 다른 한 가지 방법을 제시하려고 한다. 일반적으로 시계열 모형을 추정하고 잔차 분석을 통해 잔차가 오차의 정규성에서 크게 어긋나는 경우 이상치의 후보로 삼는다. 그런데 문제는 실제로 잔차들이 정규 분포를 따르지 않는 경우가 있고 이런 경우에는 이상치 탐색을 시도하는 것이 어려울 수밖에 없다는 것이다. 본 연구에서는 마이크로 어레이(microarray) 분석에서 사용되는 정규-지수 분포(normal-exponential distribution)를 기존에 사용하는 정규 분포 대신 차용하여 잔차 분석을 시도해보려 한다.

마이크로 어레이 분석 과정에서 형광 처리된 유전물질의 세기(intensity)를 스캔하여 관측값이

$$X\,(\text{측정된 세기; observed intensity}) = S\,(\text{세기의 참값; true intensity}) + B\,(\text{배경값; background})$$

와 같다고 하고 B는 정규 분포, S는 지수 분포의 확률 함수를 갖는다고 한다. 이제, X가 정규-지수 분포(normal-exponential distribution)라는 새로운 형태의 확률 분포를 갖는다고 하고 세기의 참값 S를 $E[S|X]$로서 추정하게 된다 (McGee와 Chen, 2006; Silver 등, 2009; Plancade 등, 2012).

The present research was supported by the research fund of Dankook university in 2022.
[1] Corresponding author: Department of Information Statistics, Dankook University, 152 Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do, 16890, Korea. E-mail: rjpak@dankook.ac.kr

Published 30 June 2023 / journal homepage: http://kjas.or.kr

현재까지 시계열에서 이상치를 찾는 방법으로는 먼저 Peña (1990)는 일변량 자기 회귀 통합이동평균(auto-regressive integrated moving average; ARIMA) 시계열 모형에서 이상치 혹은 영향력 있는 관측치를 판별하는 방법을 마할라노비스(Mahalanobis) 거리를 이용하여 제안했다. Lefrançois (1991)는 자기상관 함수에 대한 영향력을 측정하고 영향치를 판별하는 임계치를 도출하는 방법을 제안했다. Bruce와 Martin (1989)는 전체 데이터에서 관심이 있는 관측치를 제거하여 얻은 부분 시계열에 적합한 ARIMA 모형의 모수 추정치에 대한 변동량을 계산하여 영향치 혹은 이상치를 판단하는 방법을 제시하였다. 가장 최근에는 Shittu와 Shangodoyin (2008)과 Ren 등 (2019)이 스펙트럼 함수를 사용하여 주파수 영역에서 특이치를 식별하는 방법을 제안하였다. Gupta 등 (2013)이 여러 가지 방법론들을 포괄적이고 구조화된 형태로 정리한 서베이 논문을 발표했다. 앞서 언급한 방법들과 더불어 현재 가장 널리 인정받는 방법으로는 Chen과 Liu (1993)가 ARIMA 모형을 기반으로 제안한 방법으로 이상치 혹은 특이치를 '혁신적(IO; innovational)', '가산적(AO; additive)', '수준 이동(LS; level shifts)', '일시적 변동(TC; temporary changes)' 과 같은 유형으로 판단하는 방식이 있다.

본 연구의 구성은 다음과 같다. 1. 시계열 데이터의 오차가 정상적인 성분 또는 예상 가능한 성분과 비정상적인 성분 또는 예상 가능하지 않은 성분으로 구성되어 있다고 설정하고 정상적인 성분은 정규 분포를 따르는 B로 비정상적인 성분은 지수 분포를 따르는 S로 정의하고자 한다. 2. 마이크로 어레이 분석에서 배경 잡음과 신호를 정규 분포와 지수 분포를 이용하여 합성곱(convolution) 분포인 정규-지수 분포를 도출하는 방법을 차용한다. 3. 시계열 모형을 추정하고 잔차를 구한 후 잔차에 정규-지수 분포를 적합하여 일정한 한계를 벗어나는 잔차를 갖는 관측치를 이상치로 탐지하는 방법을 시도한다. 4. 비정상 오차 또는 비예상 오차를 추정하여 제거한 예상 가능 시계열을 도출한다. 5. 모의실험과 실제 자료 분석을 수행한다.

즉, 잔차분석에 있어서 정규 분포가 아닌 정규-지수 분포를 사용해보겠다는 것이 연구의 주제라고 하겠다. 다만, 마이크로 어레이 분석에서 배경값 (B)이 제거해야 할 성분이라면 본 연구에서는 B를 정상적인 성분으로 본다고 하겠다.

제2장에서 정규-지수 분포에 대해 간단히 서술하고 제3장에서는 새로운 방법론을 제시하려 한다. 제4장과 제5장에서는 경험적 분석으로 본 논문에서 제안하는 방법과 Chen과 Liu (1993)의 방법을 혼인 건수와 부동산 가격 지수 데이터에 적용하여 보았고 이상치나 특이치가 발생한 시점들을 특정할 수 있었다. 그 시점들이 경제적 혹은 사회적으로 변화가 일어나던 시점들 혹은 정부가 부동산 대책들을 발표하는 시점들과 매우 일치함을 확인할 수 있었다.

2. 정규-지수 분포

마이크로 어레이 데이터에 대한 RMA (robust multi-array average) 알고리즘을 구현하기 위한 확률적 도구가 McGee와 Chen (2006) 그리고 Silver 등 (2009)에 의해 소개되었다. 정규-지수 분포(normexp)로 불리는 정규 분포와 지수 분포의 합성곱 모형(convolution model)은 마이크로 어레이에서 시료의 광학적 세기를 분석하기 위한 확률적 모형이다. 모수가 세 개나 되는 정규-지수 분포의 최대우도추정을 위해 안장점 근사(saddle point approximation)라는 획기적인 방법도 제안되었다 (Ritchie 등, 2007).

본 단원에서는 정규-지수 분포에 대한 기본적인 내용을 간략하게 소개하겠다. 마이크로 어레이 분석에서 X를 발현값(혹은 관찰값), S를 참값 그리고 B는 배경값이라고 하면 $X = S + B$라고 정의한다. 확률 변수 S와 B를 각각 λ를 모수로 갖는 지수 분포와 μ와 σ^2을 모수로 갖는 정규 분포로 정의하고 모든 변수는 독립이라고 가정한다. B와 S의 결합확률함수는

$$f_{B,S}(b, s; \mu, \sigma, \lambda) = \frac{1}{\lambda} \exp(-s/\lambda)\phi\left(b; \mu, \sigma^2\right)$$

가 되고 $\phi(\cdot)$는 정규밀도함수이고 $S > 0$일 때 정의된다 (Bolstad, 2004; McGee와 Chen, 2006; Silver 등, 2009).

다음으로 $\mu_{S\cdot X} = x - \mu - \sigma^2/\lambda$라 하고 변수변환을 통해

$$f_{X,S}\,(x, s; \mu, \sigma, \lambda) = \frac{1}{\lambda} \exp\left(\frac{\sigma^2}{2\lambda^2} - \frac{x-\mu}{\lambda}\right) \phi\left(s; \mu_{S\cdot X}, \sigma^2\right)$$

를 얻을 수 있다. 변수변환을 하는 과정에서 음의 발현값을 제한하기 위해 $X \geq 0$으로 정의한다.

한편, $f_{X,S}(x, s)$를 s에 대해 적분하여 $f_X(x)$를 다음과 같이 구할 수 있다.

$$f_X\,(x; \mu, \sigma, \lambda) = \frac{1}{\lambda} \exp\left(\frac{\sigma^2}{2\lambda^2} - \frac{x-\mu}{\lambda}\right) \left[1 - \Phi\left(0; \mu_{S\cdot X}, \sigma^2\right)\right], \tag{2.1}$$

여기서 $\Phi(\cdot)$ 는 누적 정규 분포다. 이제, (2.1)의 f_X를 정규-지수확률 함수라고 부른다.

이제, S에 대한 X의 조건부 확률 함수

$$f_{S|X}\,(s \mid x; \mu, \sigma, \lambda) = \frac{\phi\left(s; \mu_{S\cdot X}, \sigma^2\right)}{1 - \Phi\,(0, \mu_{S\cdot X}, \sigma^2)}$$

를 구하여 참값 S에 대한 추정량을

$$E\,[S \mid X = x] = \mu_{S\cdot X} + \frac{\sigma^2 \phi\left(0; \mu_{S\cdot X}, \sigma^2\right)}{1 - \Phi\,(0, \mu_{S\cdot X}, \sigma^2)} \tag{2.2}$$

로 삼는다.

참고로 위의 내용은 Silver 등 (2009)을 주로 참조하였는데 Bolstad (2004)는 X_f를 발현값, X를 참값 그리고 X_b는 배경값이라고 하고 $X_f = X + X_b$라고 정의하였다. 저자에 따른 표기의 차이가 있음을 유의할 필요가 있다.

3. 이상치 판단을 위한 방법론

일반적으로 시계열 모형을 적합한 후 잔차의 확률 구조를 파악하고 어떤 관측치의 잔차가 확률적으로 가능한 한계를 벗어나면 그 관측치를 이상치 혹은 비정상치로 판단하곤 한다. 많은 경우 모형의 오차가 정규 분포를 따른다고 가정하게 되는데 추정 후 실제로 구한 잔차는 정규성 가정이 잘 맞지 않는 경우가 있다.

본 연구에서는 제1장에서 언급한 대로 오차가 정상적인 성분(B)와 비정상적인 성분(S)을 갖고 있다고 정의하고 오차의 확률적 표현을 위해 앞서 언급한 마이크로 어레이 분석의 정규-지수 확률모형을 차용하려 한다. 이때, B를 정규 분포, $N(\mu, \sigma)$, 를 따르는 정상 성분으로 정의하고 S를 예상외로 큰 값을 가질 수 있으나 그 확률은 낮은 비정상 성분으로 지수 분포, Exp(λ), 를 따르는 것으로 정의하여 관측치의 전차 성분이 설정한 한계치를 벗어나는 경우 이상치로 판별하려고 한다.

잔차 분석을 통한 이상치 탐색 방법을 다음과 같이 제안하고자 한다.

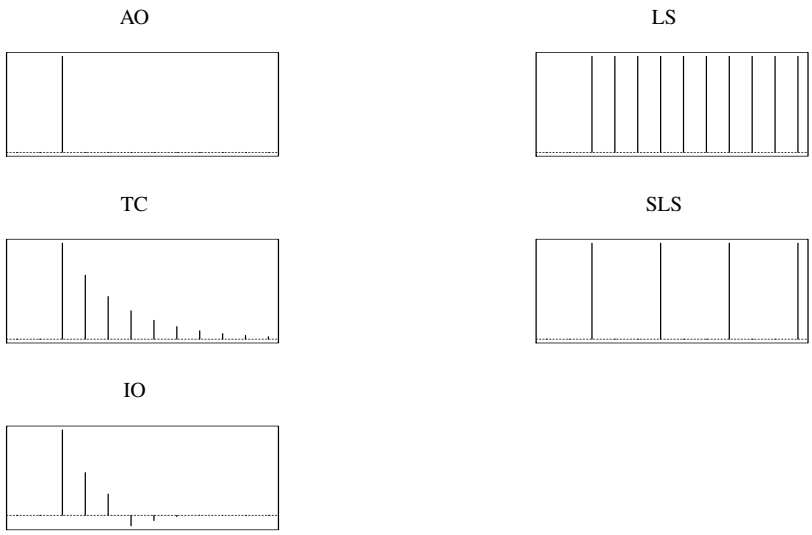
1. 분해법, 자기 회귀모형 추정법 등을 통해 시계열 모형을 적합 시키고 잔차를 구한다.
2. 정규-지수 분포 모형에서 $X \geq 0$이므로 잔차의 절대값을 X로 하여 $X = S + B$라고 정의한다.
3. 잔차의 절대값에 정규-지수 모형을 추정하고 추정값 들을 대입한 $f_X(x; \hat{\mu}, \hat{\sigma}, \hat{\lambda})$에 근거하여 주어진 유의수준 α에 따라 $(1-\alpha)$-분위수 $q_{1-\alpha}$,

   $$1 - \alpha = P\,[X \leq q_{1-\alpha}] = \int_0^{q_{1-\alpha}} f_X\left(x; \hat{\mu}, \hat{\sigma}, \hat{\lambda}\right) dx$$

   를 구한다.
4. 잔차의 절대값이 $q_{1-\alpha}$를 초과한다면 가정된 정상 오차 범위를 벗어났다고 보고 주어진 유의수준에서 해당 관측치가 이상치 혹은 비정상치일 가능성이 크다고 판단한다.

Table 1: Types of outliers

Type	Polynomial of lag operator B
Additive outlier (AO)	$L(B) = 1$
Level shifts (LS)	$L(B) = 1/(1 - B)$
Temporary change (TC)	$L(B) = 1/(1 - \delta B)$
Seasonal level shift (SLS)	$L(B) = 1/(1 - B^2)$
Innovative outliers (IO)	$L(B) = \theta(B)/(\phi(B)(1 - B)^{\alpha})$



4. 모의 실험

정규-지수 분포가 현재까지는 유전자 관련 데이터에서는 효과적으로 사용되고 있는데 시계열에서 이상치나 비정상치를 판별하는 데도 활용될 수 있는지를 혼인 데이터와 부동산 데이터를 예로 들어 실증해보려 한다.

정규-지수 분포의 모형 추정은 limma라는 R-패키지를 사용하였다. 새롭게 제안하는 방법과의 비교를 위해 현재 널리 사용되는 Chen과 Liu (1993)의 방법을 함께 사용하였다. Chen과 Liu (1993)가 제안 방법은 López-de-Lacalle (2016)에 의해 개발된 R의 tsoutliers 패키지의 tso함수를 사용하여 분석하였다. R 함수 tso를 사용하기 위해 매개모수 cval의 값을 지정해 주어야 하는데, 매개모수 cval은 만약 $n \leq 50$이면 3.0, $n \geq 450$ 이면 4.0 그 외의 경우는 $3 + 0.0025(n - 50)$로 계산할 것이 제안되었다 (López-de-Lacalle, 2016).

본 논문에서 제안하는 방법과 기존에 널리 사용되는 Chen과 Liu (1993)의 방법에 대한 성능을 비교하기 위해 모의실험을 하여 보았다. 먼저, 주기가 12인 sin(t/12), $t = 1, 2, \ldots, 50$에 오차항으로 $N(0, 0.2)$를 추가하여 길이가 50인 시계열을 생성하였다. 다음으로 무작위로 선택된 5개, 10개 그리고 15개 시점에 대하여 Table 1에 있는 다섯 가지 이상치 형태를 포함시켰다. 계절성 수준 이동(SLS)의 주기는 3으로 하였고 혁신적 이상치(IO)는 ARIMA(0.5, 0, −0.1)을 사용하였다.

이상치가 각각 5개(10%), 10개(20%) 그리고 15개(30%)가 포함된 모의 시계열에 대하여 Chen과 Liu (1993)의 방법과 제안하는 방법으로 이상치와 정상치를 탐지해내는 정확도를 계산하였다. 이 과정을 모두 1,000회 실시하였다.

모의실험 결과를 Table 2와 Figure 1에 정리하였다. 결과를 정리하면 이상치와 정상치를 탐지해내는 정

Table 2: Summay statistics for accuracy of detecting outliers ($\alpha = 0.1$)

Method	Contamination	Min	Median	Mean	Max	Variance	Range
Proposed	10%	0.7800	0.8800	0.8865	0.9800	0.0010	0.2000
	20%	0.6800	0.8000	0.7902	0.9000	0.0010	0.2200
	30%	0.6000	0.7000	0.6953	0.7800	0.0007	0.1800
Chen & Liu	10%	0.5200	0.9000	0.8919	0.9800	0.0018	0.4600
	20%	0.5000	0.8000	0.7985	0.8800	0.0008	0.3800
	30%	0.5400	0.7000	0.7031	0.8000	0.0005	0.2600

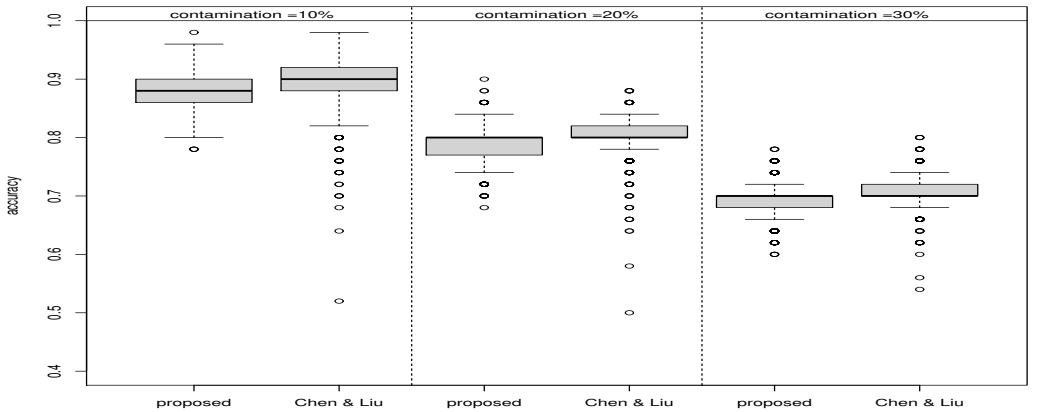


Figure 1: *The box plots for the accuracy of identifying outliers by tsoutliers and the proposed method according to levels of contaminations.*

확도 측면에서 Chen과 Liu (1993)의 방법이 평균을 근거로 하면 근소하게 우수하고 중간값을 근거로 하면 매우 유사한 수준이다. 모든 오염도 수준에 따른 두 방법 간 정확도의 평균에 대한 동일성 검정에서 p-값 < 0.00으로 두 방법 간 정확도의 평균 간에는 유의한 차이가 없다고 하겠다. 다만, Chen과 Liu (1993)의 방법에 따른 정확도의 범위가 제안된 방법보다 상대적으로 크고 상자 그림이 아래로 꼬리가 긴 모습을 보여서 제안된 방법이 조금 더 안정된(stable) 방법이라고 하겠다 (Table 2, Figure 1).

5. 자료분석

5.1. 결혼 건수

분석에 사용한 자료는 통계청 사이트에서 구한 월별 혼인 건수로써 2001년 1월부터 2020년 12월까지의 자료이다. Figure 2에서 보듯이 전체적으로 2015년 이후 혼인 건수가 감소하는 것이 뚜렷하고 계절적으로는 3, 4, 5월에 건수가 많고 여름을 지나 감소하다 11, 12월 증가하는 뚜렷한 주기성을 보인다. 분해법에 따른 시계열 모형화를 통해 얻은 잔차들의 히스토그램과 경험적 분포를 Figure 3에 그렸다. K-S 검정(Kolmogorov-Smirnov test)을 수행한 결과 (D = 0.52083, p-value < 2.2e-16) 잔차들은 정규 분포를 따르지 않는 것으로 판단된다.

본 연구에서 제안한 방법대로 잔차의 절댓값에 정규-지수 분포를 적합 시켜 보았다. 잔차의 절대값들이 정규-지수 분포에 잘 적합함을 히스토그램과 정규-지수 분포 그림(Figure 4)과 관련 통계량(Table 3)을 통해

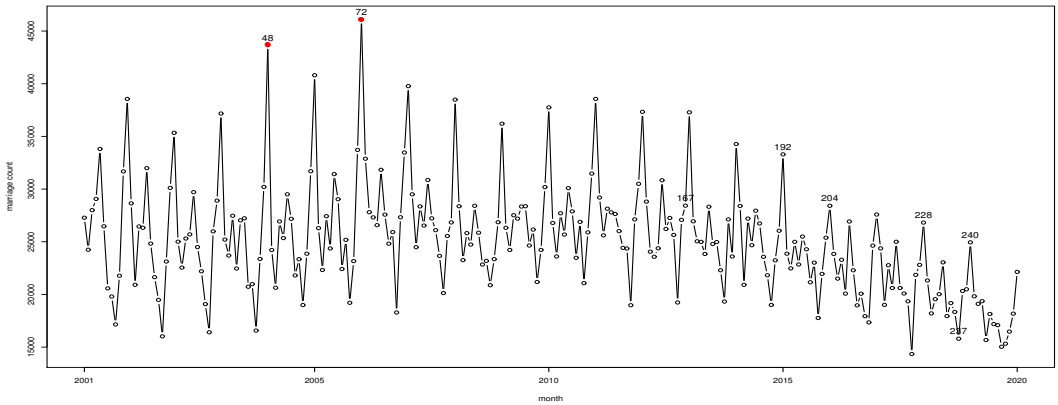


Figure 2: *Monthly marriage events: Extra-Ordinary months are numbered, the months marked with red circles are classified extra-ordinary by tsoutliers.*

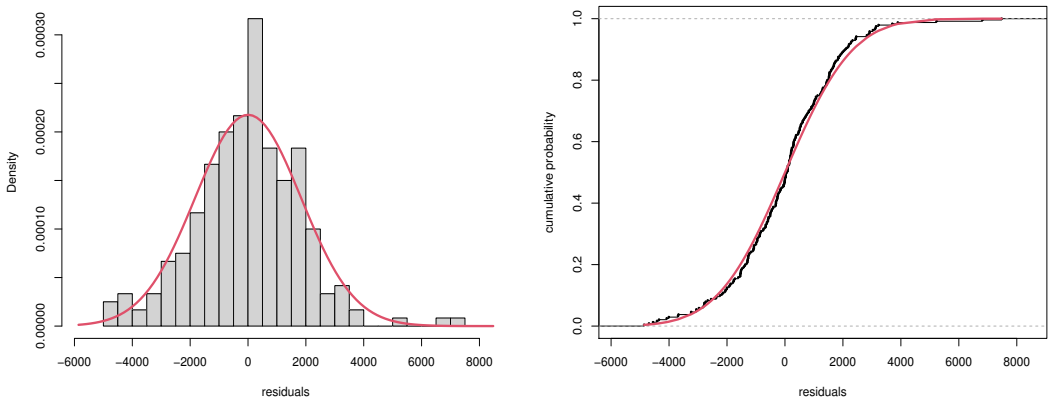


Figure 3: *Monthly marriage events: Noramlity check for residuals: Histogram & empirical cumulative distribution vs. normal distribution.*

확인할 수 있다.

Figure 5를 보면 잔차의 절댓값이 정규-지수 분포의 95-분위수를 초과하는 경우는 모두 여덟 번으로 48 (2004/12), 72 (2006/12), 167 (2013/10), 192 (2016/12), 204 (2017/12), 228 (2019/12), 237 (2020/10), 240 (2020/12)로 파악되고 99-분위수를 초과하는 경우는 두 번으로 48 (2004/12), 72 (2006/12)와 같다. Figure 2에 의하면 시각적으로 보아도 48, 72번 관측치가 가장 큰 수치를 갖고 있고 167번 관측치가 전년도의 10월 데이터에 비해 상대적으로 낮게 상승하고 있음을 볼 수 있다. 더불어 이상치로 판단되는 192, 204, 228, 240번 관측치는 전년도의 12월 건수에 비해 낮은 값을 갖는 새로운 패턴을 보인다고 판단된다. 또한 237번 데이터도 이상치로 판단되는데 이는 2020년 10월 관측치로서 코로나 유행의 여파로 과거 10월 건수와는 크게 다름을 볼 수 있다.

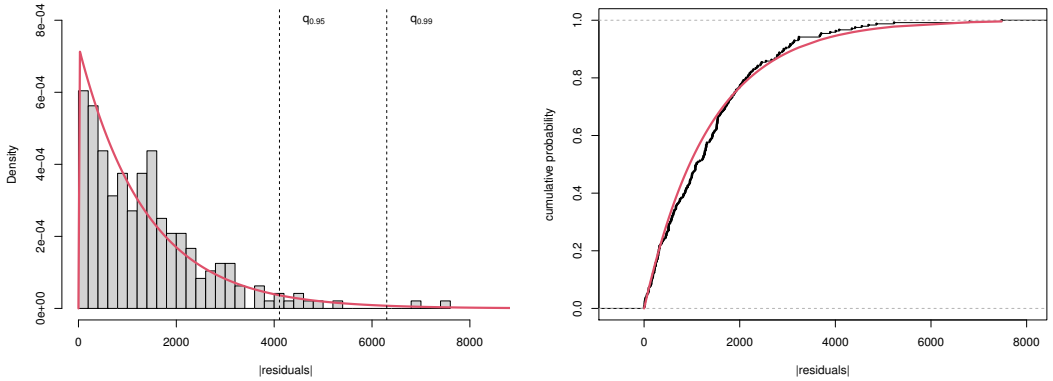


Figure 4: *Monthly marriage events: NormExp check for the absolute values of residuals: Histogram & empirical cumulative distribution vs. NormExp distribution.*

Table 3: Goodness-of-fit: Marriage event

Parameter	μ	$\log\sigma$	$\log\lambda$	KS-statistic (D)	p-value
Estimate	0.20	−10.71	7.22	0.0710	0.1684
Quantile	$q_{0.95}$ = 4108.18	$q_{0.99}$ = 6300.99			

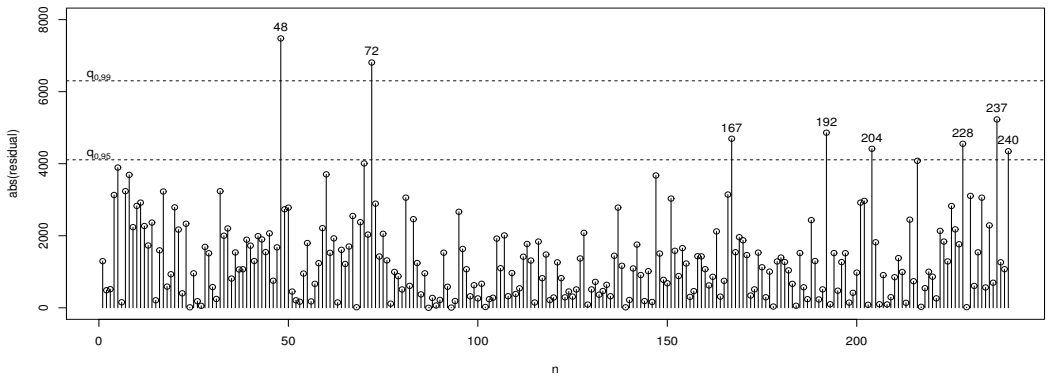


Figure 5: *Monthly marriage events: Absolute values of residuals: Critical lines.*

매개모수 cval를 3.47로 사용하여 tsoutliers를 수행하였을 때 Figure 10에 있는 tsoutliers의 결과에 따르면 48번째(AO; 2004/12)와 72번째(TC; 2006/12) 단 두 개의 관측치를 이상치로 판별하고 있다. 두 개의 관측치가 모두 추정값보다 큰 값을 갖고 있고 72번째 데이터는 일시적 변동에 해당하며 이후 큰 이상치가 발견되지 않고 그 효과도 오래 지속되지는 않는 것으로 볼 수 있다. Tsoutliers는 정규-지수분포를 이용하는 경우보다 이상치의 존재를 다소 보수적으로 판단하고 있고 2015년 이후 급격히 건수가 줄어드는 점은 파악하지 못하는 측면이 있다.

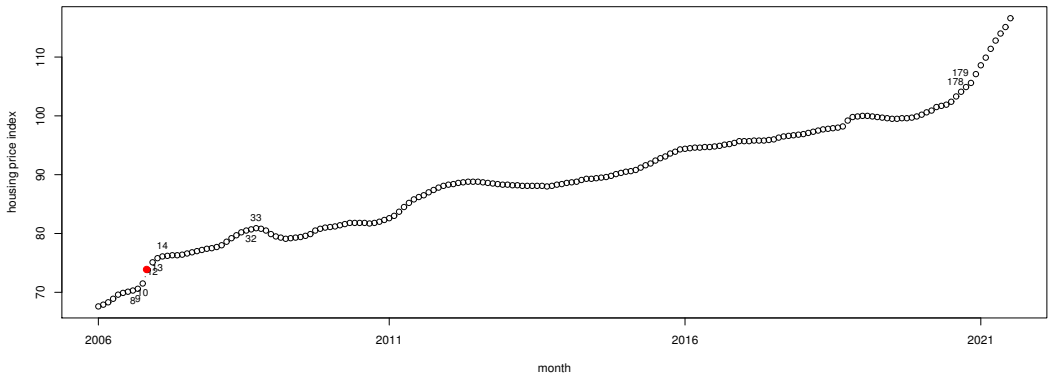


Figure 6: *Real estate data: Extra-Ordinary months are numbered, the months marked with a red circle is classified extra-ordinary by tsoutliers.*

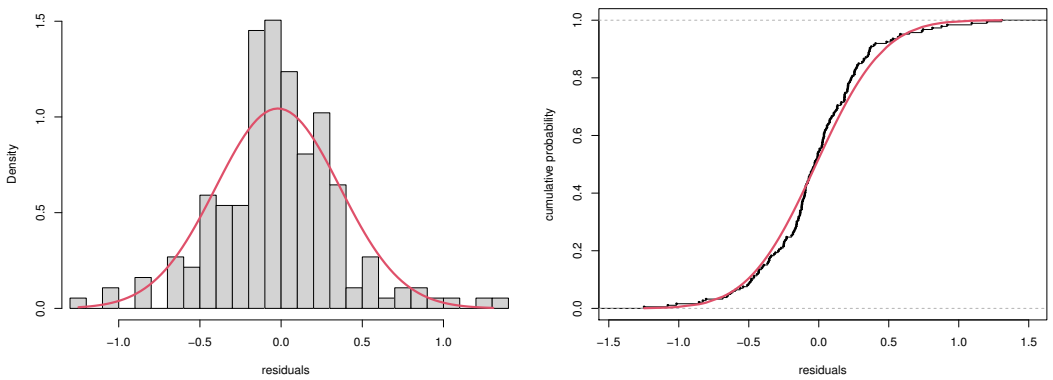


Figure 7: *Real estate data: Noramlity check for residuals: Histogram & empirical cumulative distribution vs. normal distribution.*

5.2. 부동산 가격 지수

국민 은행 (http://www.kbstar.com)의 주택 구매 가격 지수 자료를 사용하여 제안된 방법에 대한 실증 분석을 수행하였다. 사용한 국민 은행의 주택 매매 지수 데이터는 2006년 1월부터 2021년 2월까지의 월간 데이터이다 (Figure 6). 일정 시점(2019.01)의 주택의 매매 가격 수준을 100으로 환산하여 기준으로 정하고 시점별로 가격 수준의 상대적인 변동률을 작성한 수치이다.

분해법에 따른 시계열 모형화를 통해 얻은 잔차들의 히스토그램과 경험적 분포를 Figure 7에 그렸다. K-S 검정 결과(D = 0.25328, p-value = 1.444e-10) 정규 분포를 따르지 않는 것으로 판단된다.

반면에 잔차의 절댓값은 정규-지수 분포에 잘 적합한 것으로 판단된다 (Figure 8, Table 4).

Figure 9에서 잔차의 절댓값이 95-분위수를 초과하는 경우는 8 (2006/08), 9 (2006/09), 10 (2006/10), 12 (2006/12), 13 (2007/01), 14 (2007/02), 32 (2008/08), 33 (2008/09), 178 (2020/10), 179번 (2020/11)으로 모두

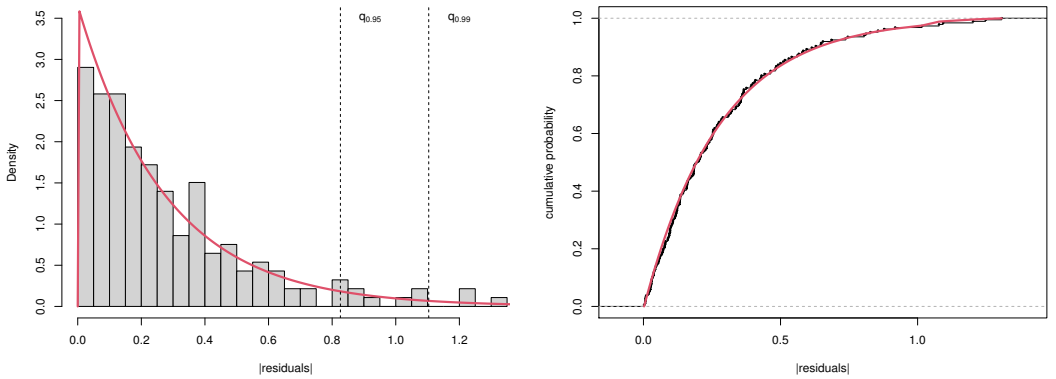


Figure 8: *Real estate data: NormExp check for the absolute values of residuals: Histogram & empirical cumulative distribution vs. Norm-Exponential distribution.*

Table 4: Goodness-of-fit: Real estate price index

Parameter	μ	$\log\sigma$	$\log\lambda$	KS-statistic (D)	p-value
Estimate	0.002	−22.88	−1.25	0.0367	0.9632
Quantile	$q_{0.95}$ = 0.8259	$q_{0.99}$ = 1.1036			

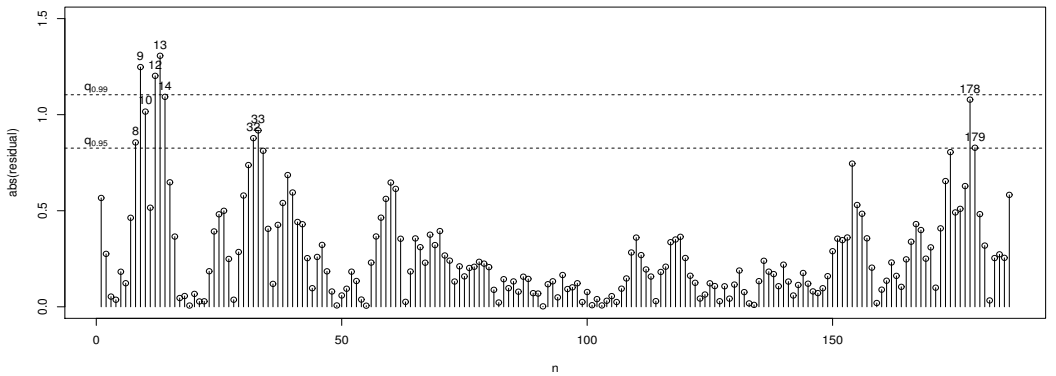


Figure 9: *Real estate data: Absolute values of residuals: Critical lines.*

열 번으로 판단되고 99-분위수를 초과하는 경우는 세 번으로 9 (2006/09), 10 (2006/10)과 13번 (2007/01)으로 나타났다. 즉, 2006년 4분기, 2008년 8, 9월 그리고 2020년 4분기에 부동산 가격 지수가 비정상적으로 변화한 것으로 판단된다. 최근 부동산 가격 상승이 국가적 문제로 대두되고 있는데 2020년 겨울 (178, 179번)이 이례적 상승의 시작이었다고 하겠다.

Tsoutliers의 경우 cval = 3.33을 사용하였고 부동산 지수 데이터에서는 11번째 데이터(2006/11) 한 개만을 이상치로 분류하고 있다. 이상치는 2006년 11월 자료로서 앞서 정규-지수 분포의 결과인 2006년 9월과 2007년

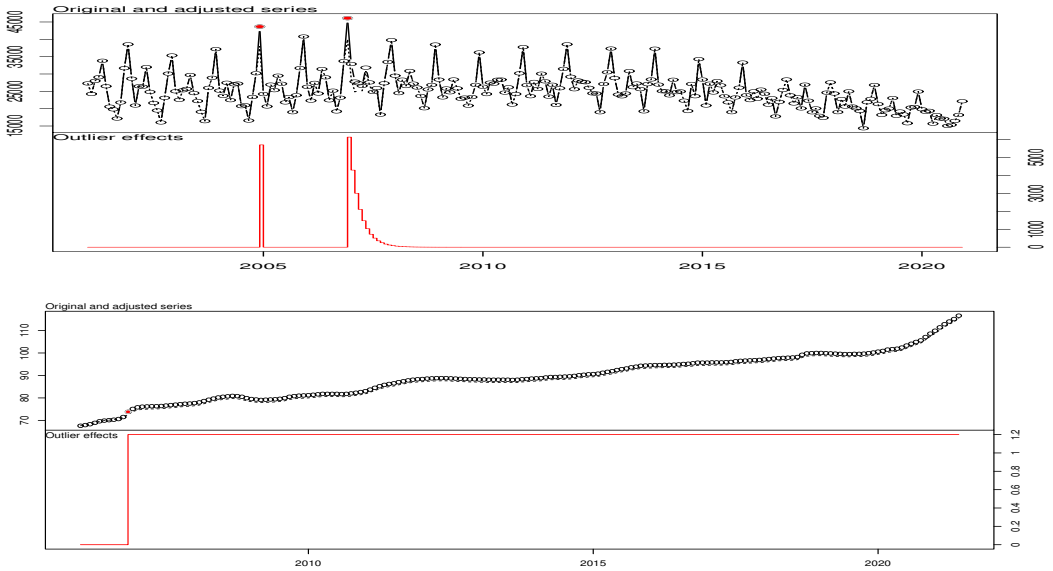


Figure 10: *Plots by tsoutliers. Marriage data (top) and real estate data (bottom).*

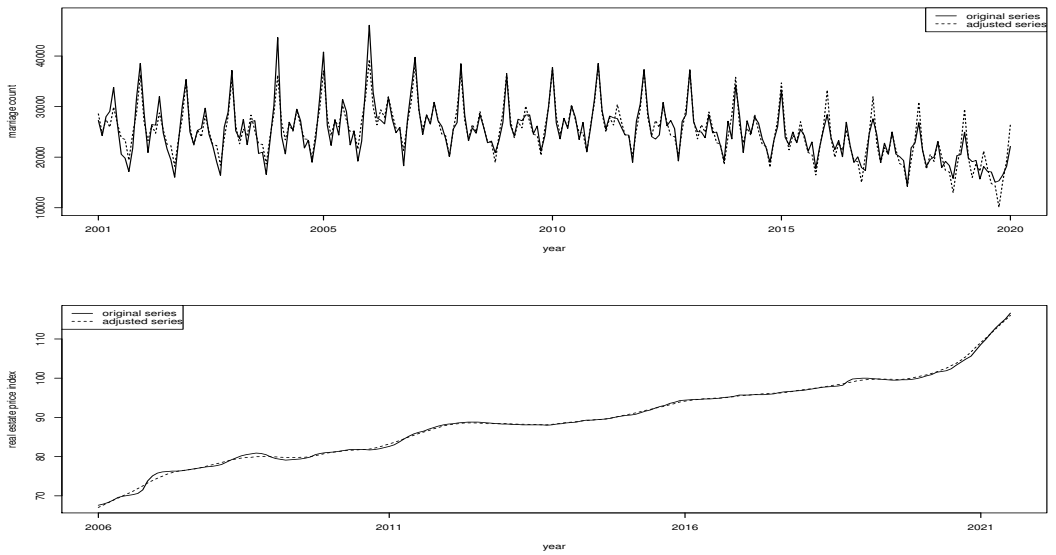


Figure 11: *Original (solid) and adjusted (dotted) plots. Marriage data (top) and real estate data (bottom).*

2월의 중간 시점으로 LS (수준 상승)으로 판단된다 (Figure 10). 즉, 2006년 11월을 기점으로 시계열의 수준에 새로운 변화가 시작되었다고 하겠다.

위 예제 자료들에 대하여 식 (2.2)를 이용하여 오차에 대한 비정상 성분의 조건부 기대값을 계산하고 비정상 성분을 제거한 시계열들을 Figure 11에 그려보았다. 결혼 건수의 경우 2015년 이후 결혼 건수는 앞선

연도들 보다 분산이 작아지면서 구조적으로 새로운 시계열의 형태를 보인다고 할 수 있겠다. 부동산지수의 경우 2008년 금융위기 전후와 2017년부터 시작하여 2022년에 이르기까지 부동산 시장은 국내·외부적 원인으로 정상적인 상황은 아니었던 것으로 보인다.

Figure 4에서 결혼건수의 히스토그램 모양을 보면 지수 분포나 감마 분포로 적합하는 것이 가능할 것으로 보이는데 실제로는 지수 분포와 감마 분포를 적합하기에는 x축의 범위가 지수 분포나 감마 분포의 통상적인 정의역을 벗어나서 성공적이지 못하였다. 하지만 결혼건수를 10000 혹은 1000으로 나누어 일 단위 혹은 십 단위로 줄여서 지수 분포를 적합하여 K-S 검정을 수행하면 결혼 데이터의 경우 p-값이 0.1 보다 커서 지수 분포에 적합하다고 할 수 있다. 그런데 정규-지수 함수를 사용하면 단위를 줄이기 위해 어떤 변환이 적당할지를 고민하지 않고도 적합할 수 있었다.

6. 맺음말

시계열 모형을 추정함에 있어서 잔차의 정규성을 따지는 것은 모형의 적합성을 진단하는 최종 과정이라고 할 수 있는데 잔차가 정규성을 만족하지 않은 경우들이 있다. 본 논문에서는 생물정보학에서 사용되는 정규-지수 분포를 차용하여 잔차 분석에 사용해보았다. 만일 정규-지수 분포가 잔차들을 다루는 데 적합하다면 관측치의 잔차의 절대값이 주어진 임계치를 넘는 경우 그 관측치를 이상치로 판별할 것을 제안하였다.

결혼과 부동산 관련 데이터를 이용하여 실증 분석을 수행해 보았다. 두 경우 모두 모형의 적합성을 진단하는 과정에서 정규 분포가 유효하지 않았고 정규-지수 분포는 역할을 잘 수행하였다. 분석 결과 이미 알려진 경제적 혹은 사회적 변화에 따른 변동이 일어난 시점과 비슷한 시점에서 이상치가 발생함을 확인할 수 있었다. 또한 현재 공신력 있게 사용되는 방법과도 핵심 내용에서 유사한 결과를 보임을 확인할 수 있었다. 기존에 널리 사용되는 방법이 이상치들의 다양한 형태를 제공하는 것에 비해 본 논문의 방법이 이상치의 유무만을 판단한다는 면에서는 정교함이 부족하다고 하겠지만 기존의 방법이 매개모수의 값에 따라 그 결과가 달라짐을 생각한다면 제안된 방법이 이해하기 쉽고 사용하기 편리하다는 점에서 기존의 방법과 더불어 추가적인 분석을 위해 유용할 것으로 생각한다. 지수 분포 외에 감마 혹은 웨이블 분포의 효과를 점검해 볼 필요와 시계열 분석 외에 회귀 분석 등에 제안된 방법을 시도해 볼 필요가 있을 것 같다.

References

Bolstad BM (2004). Low level analysis of high-density oligonucleotide array data: Background, normalization and summarization (Dissertation), University of California-Berkeley, Berkeley, CA.

Bruce AG and Martin RD (1989). Leave-k-out diagnostics for time series, *Journal of the Royal Statistical Society B*, **51**, 363–401.

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

Gupta M, Gao J, Aggarwal CC, and Han J (2013). Outlier detection for temporal data: A survey, *IEEE Transactions on Knowledge and Data Engineering*, **26**, 2250–2267.

López-de-Lacalle J (2016). *Tsoutliers: Detection of Outliers in Time Series*, R package version 0.6-8.

Lefrançois B (1991). Detecting over-influential observations in time series, *Biometrika*, **78**, 91–99.

McGee M and Chen Z (2006). Parameter estimation for the exponential-normal convolution model for background correction of Affymetrix GeneChip data, *Statistical Applications in Genetics and Molecular Biology*, **5**, Article 24.

Peña D (1990). Influential observations in time series, *Journal of Business & Economic Statistics*, **8**, 235–241.

Plancade S, Rozenholc Y, and Lund E (2012). Generalization of the normal-exponential model: Exploration of a more accurate parametrisation for the signal distribution on Illumina BeadArrays, *BMC Bioinformatics*, **13**, 1–16.

Ren H, Xu B, Wang Y, Yi C, Huang C, Kou X, and Zhang Q (2019). Time-Series anomaly detection service at Microsoft, In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Anchorage, AK, USA, 3009–3017.

Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, and Smyth GK (2007). A comparison of background correction methods for two-colour microarrays, *Bioinformatics*, **23**, 2700–2707.

Shittu IO and Shangodoyin DK (2008). Detection of outliers in time series data: A frequency domain approach, *Asian Journal of Scientific Research*, **1**, 130–137.

Silver JD, Ritchie ME, and Smyth GK (2009). Microarray background correction: Maximum likelihood estimation for the normal–exponential convolution, *Biostatistics*, **10**, 352–363.

Received November 28, 2022; Revised January 4, 2023; Accepted January 4, 2023

시계열에서 비정상적인 관측치의 식별: 정규-지수 분포의 활용

박노진[1,a]

[a]단국대학교 정보통계학과

요 약

시계열 혹은 이산 시간 신호에서의 이상치 혹은 비정상적 관측치를 탐지하는 새로운 방법을 제안하였다. 마이크로 어레이 분석에서 시료의 발현값으로부터 배경 잡음을 제거하고 신호의 세기를 추정하기 위해 정규 분포와 지수 분포의 합성곱으로 이루어진 정규-지수 분포를 사용한다. 정규-지수 분포를 차용하여 시계열에서 이상치를 탐지하는 방법을 제안하여 보았다. 시계열 모형에서 오차항을 정규 분포로 정의하지만 실제로는 모형 적합 후 잔차들이 정규 분포에 적합하지 않은 경우가 있을 수 있다. 정규-지수 분포가 잔차 성분들을 잘 적합한다면 확률적으로 주어진 한계를 벗어나는 잔차를 갖는 관측치를 이상치로 판별하려 하였다. 실증적 검정을 위해 결혼 건수와 부동산 가격 지수를 예로 사용하였다. 제안된 방법을 통해 이상치 혹은 비정상적인 관측치를 판별해 내었고 판별된 시점들이 경제적으로 혹은 사회적으로 격변하던 시점들과 일치하는 측면이 있음을 확인할 수 있었다.

주요용어: 마이크로 어레이, 이상치, 시계열, 정규-지수 분포

이 연구는 2022년도 단국대학교 대학연구비의 지원으로 연구되었음.

[1]교신저자: (16890) 경기도 용인시 수지구 죽전로 152, 단국대학교 정보통계학과. E-mail: rjpak@dankook.ac.kr