The Korean Journal of Applied Statistics (2019) DOI: http://dx.doi.org/10.5351/KJAS.2019.32.6.909
32(6), 909–922

Model selection for unstable AR process via the adaptive LASSO

Okyoung Na[a,1]

[a]Department of Applied Statistics, Kyonggi University

(Received October 7, 2019; Revised November 4, 2019; Accepted November 5, 2019)

Abstract

In this paper, we study the adaptive least absolute shrinkage and selection operator (LASSO) for the unstable autoregressive (AR) model. To identify the existence of the unit root, we apply the adaptive LASSO to the augmented Dickey-Fuller regression model, not the original AR model. We illustrate our method with simulations and a real data analysis. Simulation results show that the adaptive LASSO obtained by minimizing the Bayesian information criterion selects the order of the autoregressive model as well as the degree of differencing with high accuracy.

Keywords: AR model, unit-root, order selection, adaptive LASSO

1. 서론

자기회귀누적이동평균(autoregressive integrated moving average; ARIMA)모형은 지수평활법과 더불어 일변량 시계열 자료의 예측에 많이 사용되는 모형으로 최근 Hyndman과 Khandakar (2008)이 이들을 이용한 자동 예측 알고리즘을 개발하였다. 자기회귀모형은 자기회귀누적이동평균모형의 특수한 경우로 시계열의 현재 값을 과거 값들의 선형 결합으로 설명하려는 모형이며, 모형식이 선형회귀모형과 유사하다. 그리고 식 (1.1)에서 보듯이 모형에 오차항의 과거값으로 구성되는 이동평균 항이 포함되어 있지 않기 때문에 자기회귀누적이동평균모형에 비해 비교적 쉽게 선형회귀분석 기법을 적용하여 자기회귀모형의 모수를 추정할 수 있다. 또한 Brockwell과 Davis (2006)에 기술되어 있는 것처럼 실수값을 갖는 대부분의 정상시계열이 인과성을 만족하는 자기회귀모형으로 근사될 수 있고, 자기회귀 특성다항식의 형태에 따라 정상시계열 뿐만 아니라 비정상시계열도 자기회귀모형으로 모형화할 수 있다. 이와 같은 이유로 다양한 분야에서 자기회귀모형을 사용하고 있다.

q-차 자기회귀모형은 선형회귀모형과 유사하게 시계열 $\{y_t\}$가

$$y_t = \phi_1^* y_{t-1} + \phi_2^* y_{t-2} + \cdots + \phi_q^* y_{t-q} + \epsilon_t, \quad t = 0 \pm 1, \pm 2, \ldots \tag{1.1}$$

을 만족하는 경우를 말한다. 여기서 q는 자연수, $\phi_q \neq 0$, 오차항 $\{\epsilon_t\}$는 분산이 $\sigma^2 \in (0, \infty)$인 백색잡음 과정이다. 보통 q-차 다항식 $\phi^*(z) = 1 - \phi_1^* z - \phi_2^* z^2 - \cdots - \phi_p^* z^q$을 특성다항식이라고 부르며,

This work was supported by Kyonggi University Research Grant 2017-026-001.
[1]Department of Applied Statistics, Kyonggi University, 154-42, Gwanggyosan-Ro, Yeongtong-Gu, Suwon-Si, Gyeonggi-Do 16227, Korea. E-mail: okna@kyonggi.ac.kr

$\phi^*(z)$의 형태에 따라 다양한 시계열을 표현할 수 있다. 예를 들어, $\phi^*(z)$의 모든 근이 단위원 밖에 존재하면 식 (1.1)은 정상성을 만족하는 자기회귀모형으로 ARIMA$(q,0,0)$ 모형이 된다. 그리고 $\phi^*(z)$가

$$\phi^*(z) = (1-z)^d \left(1 - \psi_1^* z - \cdots - \psi_{q-d}^* z^{q-d}\right) \tag{1.2}$$

와 같이 인수분해되고, d는 자연수, 다항식 $\psi^*(z) = 1 - \psi_1^* z - \cdots - \psi_{q-d}^* z^{q-d}$은 모든 근이 단위원 밖에 존재하는 $(q-d)$-차 다항식이면, 식 (1.1)의 모형은 ARIMA$(q-d,d,0)$으로 비정상 시계열이 된다.

시계열 $\{y_t\}$를 가지고 자기회귀모형을 적합시킬 때, 특성다항식 $\phi^*(z)$의 형태, 즉 단위근의 개수와 더불어 다항식의 차수를 알면 Yule-Walker 추정법이나 조건부최소제곱법 등을 이용하여 모수 $\phi_1^*, \phi_2^*, \ldots, \phi_q^*$의 값을 어렵지 않게 추정할 수 있다. 최소한 단위근의 개수만 알려져 있어도 차분한 시계열을 이용하여 Box-Jenkins 방법에 따라 정상 자기회귀모형을 적합시킬 수 있다. 그러나 현실적으로 특성다항식 $\phi^*(z)$에 대한 사전정보가 없는 경우가 대부분이며, 모수를 추정하기 전에 단위근의 개수와 모형의 차수를 결정할 필요가 있다. 일반적으로 자기회귀모형을 적합시키는 절차는 Hyndman과 Khandakar (2008)이 제안한 것처럼 (1) 단위근의 개수 결정, (2) 자기회귀모형의 차수 선택, (3) 모수 추정의 3단계로 이루어진다.

(1)단계에서 단위근의 존재 여부를 판단하는 기본적인 방법은 시계열 그림과 표본자기상관그림을 활용하는 것이다. 이는 그림의 패턴을 파악하여 정상시계열의 그림의 특징과 유사한지 판단하는 방법으로 그림을 그리는 것은 어렵지 않으나 판단이 주관적이라는 단점을 가지고 있다. 이보다 좀 더 객관적으로 정상성을 판단하는 방법으로 단위근 검정(unit root test)이 있다. 단위근 검정 결과 정상시계열이라고 판단될 때까지 단위근 검정과 차분을 반복적으로 시행하여 단위근의 개수 d를 구한다. 대표적인 단위근 검정으로 augmented Dickey-Fuller (ADF) 검정 (Said와 Dickey, 1984), PP 검정 (Phillips와 Perron, 1988), KPSS 검정 (Kwiatkowski 등, 1992) 등이 있다.

(2)단계에서는 (1)단계에서 구한 단위근의 개수 d번만큼 차분한 시계열 $\{(1-B)^d y_t\}$를 이용하여 정상 자귀회귀모형의 차수를 선택한다. 여기서 B는 후향연산자를 의미하며, $By_t = y_{t-1}$을 만족한다. 정상성을 만족하는 자기회귀모형의 차수를 선택하기 위해서 표본부분자기상관그림을 이용할 수 있다. 또는 정상 AR(1) 모형부터 정상 AR$(q-d)$ 모형까지 $(q-d)$개의 모형을 적합시킨 후 계산한 Akaike information criterion (AIC) 값을 비교하여 AIC가 최소가 되는 모형을 선택할 수도 있다. 모형의 선택 기준으로는 AIC 대신 corrected AIC (AICc)나 Bayesian information criteria (BIC) 등의 다른 정보함수도 사용가능하다. 마지막으로 모수를 추정하고, 잔차시계열을 이용하여 모형의 적합 정도를 진단하는 것은 선형회귀분석과 크게 다르지 않다.

Kwon 등 (2017), Nardi와 Rinaldo (2011), Chen과 Chan (2011) 등의 최근의 연구 결과를 보면, 정상 시계열의 경우 벌점화 추정기법을 이용하면 모수 추정과 더불어 차수까지 선택가능하다. 더 나아가 부분 모형까지 선택할 수 있으며, 이론적으로 참 모형을 잘 선택하기 위해 필요한 조건들이 기술되어 있다.

본 연구에서는 벌점화 기법 중 adaptive LASSO 방법을 정상 자기회귀모형뿐만 아니라 단위근이 존재하는 비정상 자기회귀모형까지 확대하여 적용하고자 한다. 현실적으로 2개 이상의 단위근을 갖는 경우는 많지 않으며, 상당수의 시계열 자료들은 한 번만 차분하여도 증가하거나 감소하는 경향을 제거하여 정상시계열로 만들 수 있다. 그러므로 본 연구에서는 정상 자기회귀모형이나 확률보행과정처럼 단위근의 개수가 기껏해야 1개인 자기회귀모형을 중점적으로 다룬다. 본 연구에서는 단위근의 개수 결정과 정상 자기회모형의 선택 및 모수 추정 추정을 각각 따로 수행하지 않고, 동시에 수행하기 위해 식 (1.1)과 같은 모형 대신 이를 변환한 모형에 adaptive LASSO 방법을 적용하였다. 이에 대해서는 2장에서 상세하게 기술할 것이다. 3장에서는 2장에서 설명한 adaptive LASSO 추정량의 성질을 모의실험을 통해 알아보고, 실제 자료분석에 적용해보고자 한다.

2. 모형 선택 및 모수 추정

2.1. 모형

시계열 자료 $y_1, y_2, \ldots, y_T$가 주어졌을 때 다음과 같은 자기회귀모형을 적합시키고자 한다:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t, \tag{2.1}$$

여기서 p는 자기회귀모형의 차수에 대한 허용한계이며, 일반적으로 $[10\log_{10}(T)]$ 또는 $[4(T/100)^{1/4}]$, $[12(T/100)^{1/4}]$, $[T^{1/3}]$처럼 자료의 개수 T에 대한 증가함수를 p의 값으로 많이 사용한다. 단, $[x]$는 실수 x 이하의 정수 중 최대값을 의미한다. 비록 주어진 시계열 $\{y_t\}$가 따르는 참 모형이 ARIMA$(q,0,0)$ 또는 ARIMA$(q-1,1,0)$라고 하더라도, 참 모형에 대한 사전 정보 없이 시계열 자료만 주어진 상태에서는 차수 q와 단위근의 존재 여부를 모른다. 그러므로 식 (1.1)에 주어진 모형 대신 식 (2.1)의 모형처럼 자료의 수가 충분히 많은 경우 참 모형을 내포할 수 있는 모형을 고려한 것이다.

참 모형이 식 (2.1)의 부분모형이라면, 참 모형의 차수 q는

$$q = \max\{j \in \mathcal{S}_F : \phi_j \neq 0\}, \quad \mathcal{S}_F = \{1, 2, \ldots, p\}$$

을 만족한다. 그리고 참 모형이 ARIMA$(q-1,1,0)$으로 단위근을 가지면

$$\phi_1 + \phi_2 + \cdots + \phi_p = 1$$

이 성립하고, 참 모형이 ARIMA$(q,0,0)$이면

$$\phi_1 + \phi_2 + \cdots + \phi_p \neq 1$$

이다. 그러므로 단위근 존재 여부에 대한 판단과 더불어 자기회귀모형의 차수까지 동시에 결정하기 위해서는 식 (2.1)의 계수들 $\phi_1, \phi_2, \ldots, \phi_p$ 중에서 0이 아닌 값들을 찾아내는 동시에 $\phi_1 + \phi_2 + \cdots + \phi_p = 1$이 성립하는지도 알아봐야 한다. 이를 위해 본 연구에서는 식 (2.1)의 모형 대신 변환된 모형을 고려하였으며, 변환된 모형식은 다음과 같다:

$$y_t - y_{t-1} = \beta_1 y_{t-1} + \beta_2(y_{t-1} - y_{t-2}) + \cdots + \beta_p(y_{t-p+1} - y_{t-p}) + \epsilon_t. \tag{2.2}$$

이 모형은 단위근 검정 중 Said와 Dickey (1984)의 ADF 검정에서 고려한 회귀모형으로 첫 번째 계수 β_1은

$$\beta_1 = \phi_1 + \phi_2 + \cdots + \phi_p - 1 \tag{2.3}$$

을 만족한다. 그러므로 $\beta_1 = 0$이면, 변환하기 전 모형의 특성다항식 $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$는 $\phi(1) = 1 - \phi_1 - \phi_2 - \cdots - \phi_p = 0$을 만족하며 단위근을 가진다. 이와 반대로 $\beta_1 \neq 0$이면 특성다항식 $\phi(z)$는 단위근을 갖지 않는다. 이와 같은 이유로 ADF 검정에서는 식 (2.2)의 모형을 적합시켜서 구한 β_1에 대한 추정값을 바탕으로 "$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 < 0$"을 검정하여 시계열 $\{y_t\}$의 정상성 여부를 판단한다. 본 연구에서도 시계열의 정상성, 즉 특성다항식 $\phi(z)$이 단위근을 갖는지 여부를 판단하는 것이 첫 번째 목적이므로 식 (2.1)의 모형이 아닌 식 (2.2)의 변형된 모형을 사용하였다.

식 (2.1)과 식 (2.2)를 비교하면, β_1을 제외한 나머지 계수들에 대하여 $\beta_2 = -(\phi_2 + \cdots + \phi_p)$, $\beta_3 = -(\phi_3 + \cdots + \phi_p), \ldots, \beta_{p-1} = -(\phi_{p-1} + \phi_p)$, $\beta_p = -\phi_p$의 관계식이 성립함을 쉽게 확인할 수 있다. 그리고 참 모형의 차수 q가 1 이상의 값일 때, 집합 $\mathcal{S}_\beta = \{j \in \mathcal{S}_F : \beta_j \neq 0\}$은

$$\max\{j \in \{1\} \cup \mathcal{S}_\beta\} = q$$

을 만족한다. 그러므로 우리는 식 (2.2)의 변환된 모형에서 $\beta_j \neq 0$인 계수들을 식별해냄으로써 원 모형 (2.1)의 단위근 존재 여부 판단과 더불어 자기회귀모형의 차수까지 선택할 수 있다.

특히 특성다항식 $\phi(z)$가 단위근을 1개 갖는 경우에는 $\beta_1 = 0$이므로 식 (2.2)와 계수 $\beta_2, \ldots, \beta_p$는 차분한 시계열 $\{(1-B)y_t\}$에 대한 정상 자기회귀모형과 그 계수를 나타내게 된다. 따라서 앞서 설명한 것처럼 3단계로 나누어 자기회귀모형을 적합시키지 않고, 변환된 모형을 적합시킴으로써 한번에 차분한 시계열의 정상 자기회귀모형까지 추정하는 것이 가능하다고 기대할 수 있다.

2.2. 모수 추정

우리는 식 (2.2)의 모형에 들어 있는 계수 $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$를 adaptive LASSO 방법을 이용하여 다음처럼 추정한다:

$$\hat{\boldsymbol{\beta}}^{\lambda} = \arg\min_{\mathbf{b}} \left\{ \frac{1}{2n} \sum_{t=p+1}^{T} (z_t - b_1 y_{t-1} - b_2 z_{t-1} - \cdots - b_p z_{t-p+1})^2 + \lambda \sum_{j=1}^{p} w_j |b_j| \right\}, \tag{2.4}$$

여기서 $\mathbf{b} = (b_1, b_2, \ldots, b_p)^T$는 n-차원 실벡터이고, $z_t = (1-B)y_t = y_t - y_{t-1}$, $n = T - p$이다. λ와 $w_1, w_2, \ldots, w_p$는 모두 0 이상의 실수이며, 각각 조절모수와 가중치를 나타낸다.

앞서 살펴본 것처럼 시계열 $\{y_t\}$가 따르는 참모형이 ARIMA$(q-d, d, 0)$ (단, d는 0 또는 1만을 가진다)일 때, q와 d는

$$q = \max\{j \in \{1\} \cup \mathcal{S}_\beta\}, \quad d = \begin{cases} 1, & \beta_1 = 0, \\ 0, & \beta_1 \neq 0 \end{cases}$$

을 만족한다. 그러므로 자기회귀모형의 차수 q와 차분 차수 d는 $\hat{\boldsymbol{\beta}}^{\lambda} = (\hat{\beta}_1^{\lambda}, \hat{\beta}_2^{\lambda}, \ldots, \hat{\beta}_p^{\lambda})^T$를 이용하여

$$\hat{q}^{\lambda} = \max\left\{ j \in \{1\} \cup \hat{\mathcal{S}}_\beta^{\lambda} \right\}, \quad \hat{d}^{\lambda} = \begin{cases} 1, & \hat{\beta}_1^{\lambda} = 0, \\ 0, & \hat{\beta}_1^{\lambda} \neq 0 \end{cases} \tag{2.5}$$

과 같이 추정할 수 있다. 여기서 $\hat{\mathcal{S}}_\beta^{\lambda} = \{j \in \mathcal{S}_F : \hat{\beta}_j^{\lambda} \neq 0\}$를 의미한다.

Zou (2006), Wang 등 (2009), Kwon 등 (2017)과 같은 기존의 연구 결과에서 보듯이 조절모수와 가중치의 값에 따라 추정량 $\hat{\boldsymbol{\beta}}^{\lambda}$의 성질이 결정되므로 두 값을 잘 선택해야한다. 이들은 선형회귀모형과 정상 자기회귀모형에서 adaptive LASSO 추정량이 "oracle property"를 만족하기 위한 조절모수와 가중치의 조건을 연구하였으며, 최소제곱추정량과 같은 일치 추정량의 역수나 거듭제곱을 가중치로 사용할 것을 제안하였다. 본 연구에서도 이들이 제안한 것처럼 가중치의 값으로

$$w_j = \frac{1}{\sqrt{n}|\tilde{\beta}_j|}, \quad j = 1, 2, \ldots, p \tag{2.6}$$

를 사용하였다. 여기서 $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_p)^T$는 변환된 모형 (2.2)의 모수 $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$에 대한 최소제곱추정값이다.

Adaptive LASSO 방법을 이용하여 모수를 추정하는 경우 가중치의 값 뿐만 아니라 조절모수 λ의 값도 정해주어야 한다. Zou (2006)나 Kwon 등 (2017)의 연구 결과에 기술된 것처럼 이론적으로 규명된 특정 조건이 있으면, 이 조건을 만족하는 λ를 조절모수의 값으로 사용할 수도 있다. 그러나 대부분의 경우에는 k-겹 교차검증(k-fold cross validation)이나 정보함수를 최소화하는 방법 등을 이용하여 조절모수

값을 선택하며, 본 연구에서는 10-겹 교차검증 방법과 두 개의 정보함수

$$\mathrm{AIC}(\lambda) = \log\left\{\frac{1}{n}\sum_{t=p+1}^{T}\left(z_t - \hat{\beta}_1^{\lambda} y_{t-1} - \hat{\beta}_2^{\lambda} z_{t-1} - \cdots - \hat{\beta}_p^{\lambda} z_{t-p+1}\right)^2\right\} + \frac{2}{n}N_\lambda,$$

$$\mathrm{BIC}(\lambda) = \log\left\{\frac{1}{n}\sum_{t=p+1}^{T}\left(z_t - \hat{\beta}_1^{\lambda} y_{t-1} - \hat{\beta}_2^{\lambda} z_{t-1} - \cdots - \hat{\beta}_p^{\lambda} z_{t-p+1}\right)^2\right\} + \frac{\log(n)}{n}N_\lambda$$

를 사용하였다. 여기서 N_λ는 조절모수가 λ일 때 구한 adaptive LASSO 추정값 $\hat{\beta}_1^{\lambda}, \hat{\beta}_2^{\lambda}, \ldots, \hat{\beta}_p^{\lambda}$ 중 0이 아닌 것의 개수를 나타낸다.

3. 모의실험 및 자료 분석

3.1. 모의실험

먼저 adaptive LASSO 추정량의 성질을 알아보기 위하여 8개의 자기회귀모형을 고려하였다:

- $\mathrm{ARIMA}(1,d,0) : (1-B)^d(1-0.3B)y_t = \epsilon_t,\ d = 0, 1;$
- $\mathrm{ARIMA}(2,d,0) : (1-B)^d(1-0.3B)(1-0.7B)y_t = \epsilon_t,\ d = 0, 1;$
- $\mathrm{ARIMA}(4,d,0) : (1-B)^d(1-0.3B^4)y_t = \epsilon_t,\ d = 0, 1;$
- $\mathrm{ARIMA}(5,d,0) : (1-B)^d(1-0.3B)(1-0.7B^4)y_t = \epsilon_t,\ d = 0, 1.$

여기서 오차항 $\{\epsilon_t\}$는 서로 독립적으로 표준정규분포를 따르는 확률변수들이다. 각각의 모형에서 T개의 시계열 자료 $y_1, y_2, \ldots, y_T$를 생성하여 식 (2.2)에 들어있는 모수 $\beta_1, \beta_2, \ldots, \beta_p$를 앞에서 설명한 adaptive LASSO 방법대로 추정하였다. 시계열 자료의 개수 T의 값으로 50, 100, 250, 500, 1000을 고려하였고, 자기회귀모형의 차수에 대한 허용한계는 $p = [12(T/100)^{1/4}]$을 사용하였다. 이 값은 Schwert (1989)가 PP 검정이나 ADF 검정 등의 단위근 검정에서 사용하도록 제안한 값으로, 비록 정상 자기회귀모형에서의 adaptive LASSO에 대한 연구이지만 Kwon 등 (2017)이 제안한 조건도 만족한다.

Adaptive LASSO 추정값을 구하기 위해서 식 (2.6)에 주어진 값을 가중치로 사용하였다. 그리고 조절모수 λ에 대한 값으로 $\lambda_1 > \lambda_2 > \cdots > \lambda_{100}$을 만족하는 100개의 실수를 고려하였다. 여기서 λ_1은 조절모수로 고려한 값 중 최대값으로 R-함수 중 glmnet에서 제공하는 lambda.max를 사용하였다. lambda.max는 주어진 자료로부터 계산되는 값으로, 이 값보다 큰 λ에 대해서는 모든 계수에 대한 adaptive LASSO 추정값이 0이다. 그리고 λ_{100}, 즉 고려한 조절모수 중 최소값은 $\min\{1, \lambda_1\}/10000$이며, $\log(\lambda_{99}), \log(\lambda_{98}), \ldots, \log(\lambda_2)$는 구간 $(\log(\lambda_{100}), \log(\lambda_1))$을 99등분하는 경계값이다.

우선 adaptive LASSO 추정량의 성질을 알아보기 위해서 다음의 두 가지 측도를 고려하였으며, 총 100번 반복실험을 수행하였다.

- 모형식별정확도(model identification accuracy; MIA_0):

  r번째 반복실험에서 사용한 100개의 조절모수들의 집합을 Λ_r이라고 할 때, Λ_r의 부분집합 A_r을 다음처럼 정의하자:

  $$A_r = \left\{\lambda \in \Lambda_r : \hat{\mathcal{S}}_\beta^{\lambda} = \mathcal{S}_\beta\right\}.$$

  이 집합은 참 모형을 올바르게 식별한 조절모수들의 집합으로, 공집합이 아닌 경우는 adaptive LASSO 방법을 이용하여 참 모형을 올바르게 식별할 수 있음을 의미한다. 그러므로 모형식별정

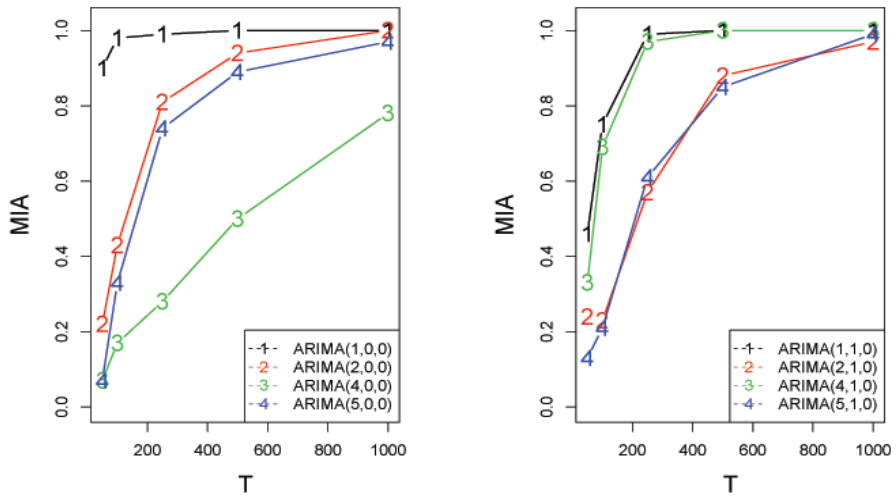

Figure 3.1. Model identification accuracy (MIA_0: $d = 0$ (left), $d = 1$ (right)).

확도에 대한 측도로

$$\text{MIA}_0 = \frac{1}{100}\sum_{r=1}^{100} I(A_r \neq \emptyset)$$

을 사용하였다. 여기서 $I(\ )$는 지시함수를 뜻한다.

- 추정량의 평균제곱오차(mean squared error of estimator; MSE_0):
  집합 A_r이 공집합이 아닐 때, A_r의 원소 중 최소값을 λ_r^0라고 하고 이에 대응되는 adaptive LASSO 추정값을 $\hat{\boldsymbol{\beta}}_r^0$라고 표시한다. 그리고 이 추정값들을 이용하여 추정량의 평균제곱오차와 유사한 측도인

$$\text{MSE}_0 = \frac{1}{|R_0|}\sum_{r\in R_0}\left\|\hat{\boldsymbol{\beta}}_r^0 - \boldsymbol{\beta}\right\|^2$$

  를 정의하였다. 여기서 $R_0 = \{r : A_r \neq \emptyset, r = 1, 2, \ldots, 100\}$이고, $|R_0|$는 집합 R_0의 원소의 개수로 100MIA_0와 일치한다. 그리고 $\|\hat{\boldsymbol{\beta}}_r^0 - \boldsymbol{\beta}\|$는 두 벡터 $\hat{\boldsymbol{\beta}}_r^0$, $\boldsymbol{\beta}$의 유클리드 거리를 나타낸다.

Figure 3.1과 Figure 3.2는 8개의 모형에 대한 MIA_0와 MSE_0를 자료의 개수별로 계산하여 그린 것이다. 먼저 Figure 3.1을 보면 모든 경우에 자료의 개수가 증가할수록 MIA_0의 값이 증가한다. 또한 대부분의 경우에 자료의 개수가 1,000개 정도만 되어도 1에 가까운 MIA_0 값을 가지는 것도 확인할 수 있다. 그러므로 우리는 시계열 자료의 개수가 충분히 많은 경우 adaptive LASSO 방법이 참 모형을 올바르게 식별해 낼 수 있음을 알 수 있다. 그리고 Figure 3.2를 보면 자료의 개수 T가 증가할수록 MSE_0의 값이 대부분 0으로 감소함을 알 수 있다. 이는 Figure 3.1의 결과와 연동해서 봤을 때, adaptive LASSO 방법이 시계열 자료의 개수가 충분히 많은 경우 참 모형을 올바르게 식별할 뿐만 아니라 추정량도 일치성을 만족하게 됨을 의미한다. 본 연구에서 제안한 방법에 따르면, 참 모형을 올바르게 식별해낸다는 것은 단위근의 존재 여부와 자기회귀모형의 차수를 동시에 정확하게 판단한다는 뜻이다. 그러므로 식 (2.2)의 ADF 단위검정에 사용되는 회귀모형에 adaptive LASSO 방법을 적용하면 정상 자기회귀모형 뿐만 아니라 단위근이 존재하는 비정상 자기회귀모형도 단위근 검정, 차수 선택, 모수 추정의 3단계를 거치치 않고도 잘 추정할 수 있다.

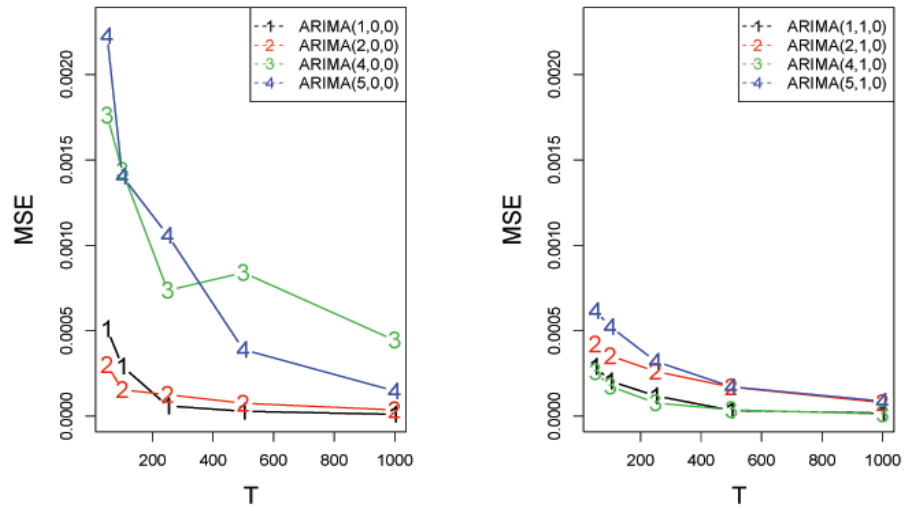

Figure 3.2. Mean squared error of estimator (MSE_0: $d = 0$ (left), $d = 1$ (right)).

현실적으로 앞서 설명한 λ_r^0와 $\hat{\boldsymbol{\beta}}_r^0$는 구할 수 없으며, MIA_0와 MSE_0의 성질은 adaptive LASSO 방법이 참 모형을 잘 추정할 가능성이 있다는 것만 제시한다. 실제 자료를 분석하는 경우에는 100개의 조절모수 중 적절한 값을 선택해야 하며, 본 연구에서는 이를 위해 10-겹 교차검증, AIC, BIC를 사용하였다. 100번의 반복실험 중 r번째 반복실험에서 사용한 조절모수들의 집합 Λ_r에서 10-겹 교차검증으로 선택한 조절모수 값을 λ_r^C, 이에 대응하는 adaptive LASSO 추정값을 $\hat{\boldsymbol{\beta}}_r^C$라고 표시한다. 마찬가지로 Λ_r 중에서 AIC와 BIC가 최소가 되도록 선택한 조절모수의 값을 각각 λ_r^A와 λ_r^B라고 하고, 각각에 대응하는 adaptive LASSO 추정값을 $\hat{\boldsymbol{\beta}}_r^A$와 $\hat{\boldsymbol{\beta}}_r^B$라고 표기한다. 세 가지 추정량의 성질을 비교 검토하기 위해 다음에 주어진 4가지 측도를 고려하였다.

- 모형식별정확도(MIA_s):
  앞에서 모형식별을 측정하기 위해 사용했던 측도 MIA_0와 유사한 측도이며, 정의는 다음과 같다:

$$\text{MIA}_s = \frac{1}{100}\sum_{r=1}^{100} I\left(\hat{\mathcal{S}}_\beta^\lambda = \mathcal{S}_\beta, \lambda = \lambda_r^s\right), \quad s = C, A, B,$$

  여기서 첨자 s는 조절모수를 선택할 때 사용된 세 가지 방법을 나타내는 것으로 $s = C$는 10-겹 교차검증, $s = A$와 $s = B$는 각각 AIC와 BIC를 의미한다.

- 추정량의 평균제곱오차(MSE_s):
  조절모수를 선택하는 방법별로 추정량의 평균제곱오차의 성질을 파악할 수 있는 측도 MSE_s를 MSE_0와 유사하게 다음과 같이 정의하였다:

$$\text{MSE}_s = \frac{1}{100}\sum_{r=1}^{100} \left\|\hat{\boldsymbol{\beta}}_r^s - \boldsymbol{\beta}\right\|^2.$$

- 단위근 식별 정확도(unit-root identification accuracy; UIA_s)와 차수 선택 정확도(order selection accuracy; OSA_s):

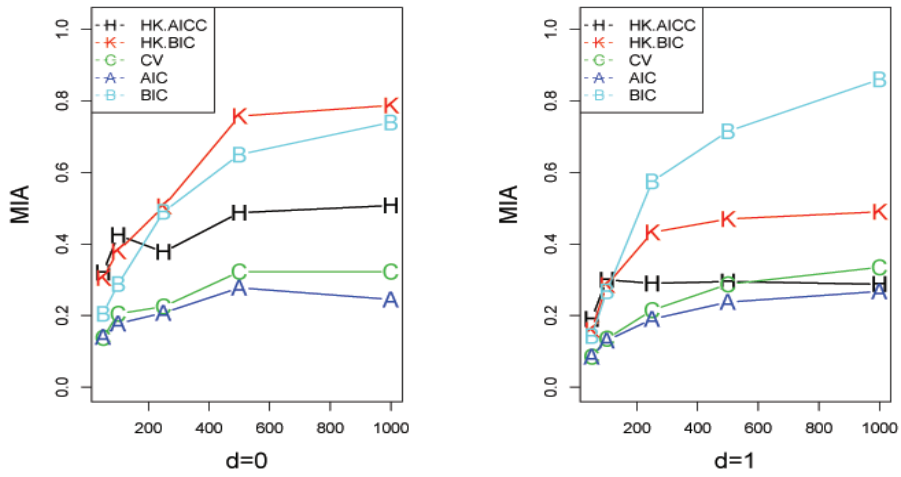

Figure 3.3. Model identification accuracy (MIA_s, $s = C, A, B, H, K$).

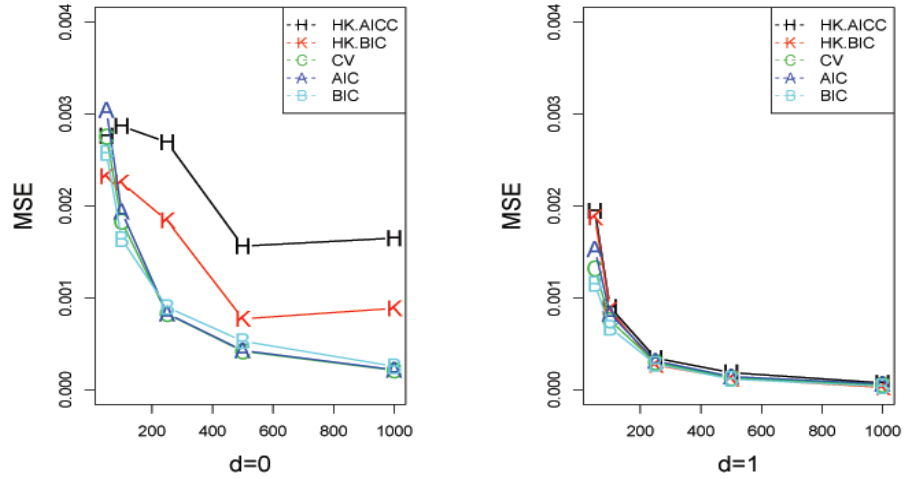

Figure 3.4. Mean squared error of estimator (MSE_s, $s = C, A, B, H, K$).

본 연구에서는 단위근의 존재 여부 판단과 자기회귀모형의 차수 선택을 동시에 하는 것이 중요한 문제이므로 다음과 같은 두 가지 측도를 더 고려하였다:

$$\text{UIA}_s = \frac{1}{100}\sum_{r=1}^{100} I\left(\hat{d}^{\lambda} = d, \lambda = \lambda_r^s\right);$$

$$\text{OSA}_s = \frac{1}{100}\sum_{r=1}^{100} I\left(\hat{q}^{\lambda} = q, \lambda = \lambda_r^s\right).$$

우선 모형별, 자료의 개수별로 이 측도값들을 계산하였다. 그리고 8개의 모형을 크게 $d = 0$인 경우와 $d = 1$인 경우로 분리한 후, 각 경우에 속하는 4개의 모형에 대한 측도값들을 산술평균하였다. Figures 3.3–3.5에서 $s = C, A, B$인 경우는 이렇게 구한 측도들의 평균값을 자료의 개수별로 정리하여 그린 것이

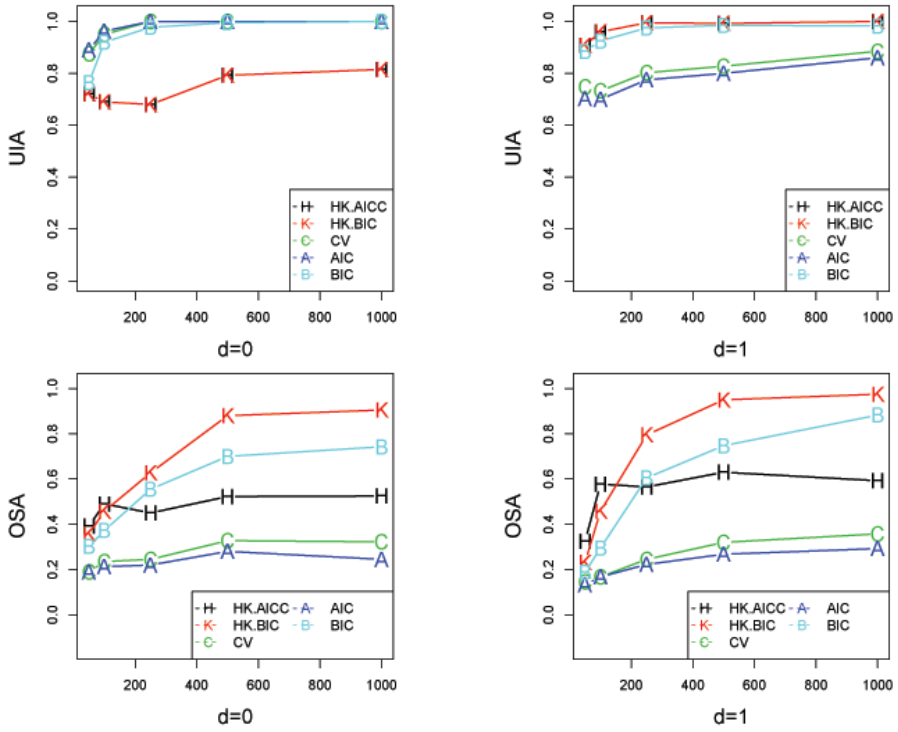

Figure 3.5. Unit-root identification accuracy (UIA_s, upper) and order selection accuracy (OSA_s, lower).

다. Figure 3.4를 보면 교차검증과 AIC, BIC 세 가지 방법에는 MSE_s의 차이가 별로 없으며 모두 자료의 개수가 증가할수록 그 값이 0으로 감소하는 것을 확인할 수 있다. 그러나 Figure 3.3과 Figure 3.5를 보면, 그 양상이 좀 다르다. 비록 모든 경우에 있어서 자료의 개수가 증가할수록 측도값이 증가하는 경향이 보이기는 하지만, 교차검증이나 AIC에 비해 BIC 방법이 증가하는 정도가 두드러지게 크게 나타난다. 다시 말해 교차검증이나 AIC보다 BIC를 이용하여 조절모수를 선택할 때 모형 식별과 더불어 단위근 존재 여부와 차수를 올바르게 잘 선택할 수 있다. 그러므로 본 연구에서는 조절모수를 선택할 때 BIC를 사용할 것을 추천한다.

Figure 3.6은 교차검증, AIC, BIC에 의해 선택된 조절모수들에 대한 상자그림으로 ARIMA$(2, d, 0)$에 대한 것이다. 모의실험에서 고려한 나머지 모형에 대한 상자그림들은 비록 본 논문에서는 생략하였지만, Figure 3.6과 모두 유사한 형태를 가짐을 확인할 수 있었다. 우리는 이 그림으로부터 조절모수를 선택할 때 사용한 세 가지 방법이 모두 λ_1이나 λ_{100}, 즉 후보로 고려된 조절모수들의 양 극단값이 아닌 실제 유효한 값을 선택함을 알 수 있다.

본 모의실험에서는 adaptive LASSO 방법과 기존의 방법을 비교하기 위하여 8개의 자기회귀모형에 대해 추가로 Hyndman과 Khandakar (2008)이 제안한 절차에 따라 각각 자기회귀모형을 적합시켰다. 본 연구에서는 차분 차수가 1 이하인 자기회귀모형만을 다루고 있으므로, Hyndman과 Khandakar (2008)이 개발한 auto.airma 함수를 이에 알맞게 조정하여 사용하였다. 그리고 차분 차수를 결정하기 위해 KPSS 검정을 사용하였고, 자기회귀모형의 차수를 선택하는 기준으로 AICc와 BIC 두 가지 정보함수를 고려하였다. r번째 반복실험에서 Hyndman과 Khandakar 방법을 이용하여 구한 차분 차수, 자

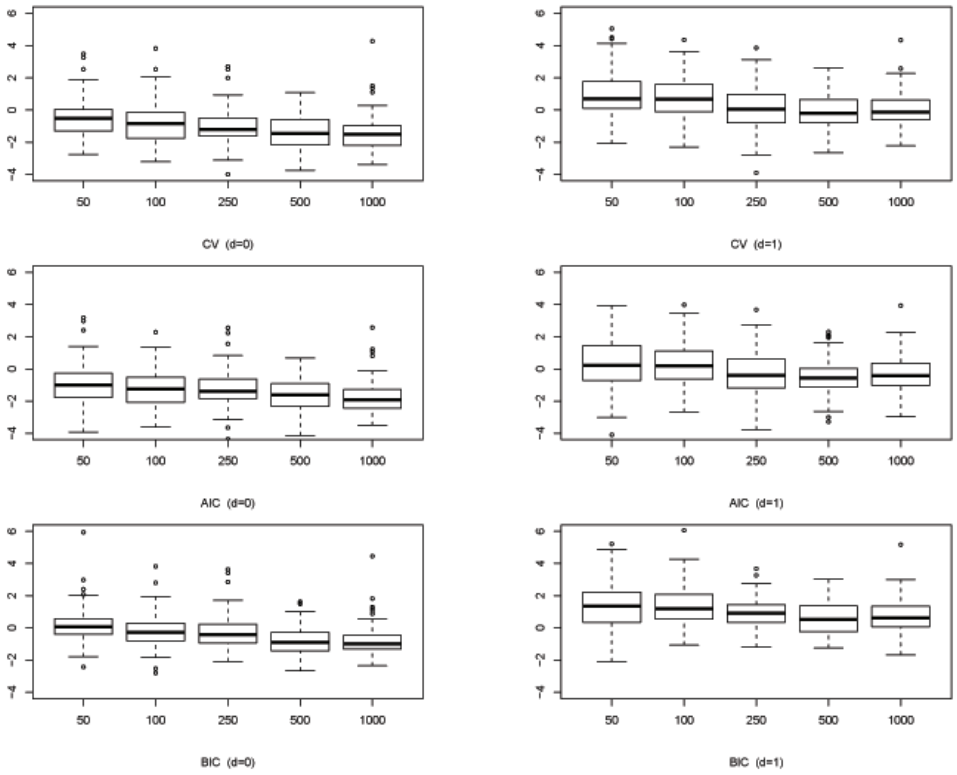

Figure 3.6. Box plots of λ_r^s $(s = C, A, B)$.

기회귀모형의 차수, 모수에 대한 추정값을 $\hat{d}_r^s, \hat{q}_r^s, \hat{\boldsymbol{\beta}}_r^s, s = H, K$라고 하자. 여기서 $s = H$와 $s = K$는 각각 AICc와 BIC를 이용한 경우를 나타낸다. 참고로 Hyndman과 Khandakar 방법에서는 식 (2.2)에 주어진 변환된 모형이 아닌 식 (2.1)의 원 자기회귀모형의 계수에 대한 추정값을 구하며, $\hat{\boldsymbol{\beta}}_r^H$과 $\hat{\boldsymbol{\beta}}_r^K$는 2.1절에서 설명한 계수들 사이의 관계식을 이용하여 원 자기회귀모형의 계수에 대한 추정값을 변환하여 구한 값이다. 우리는 $\hat{d}_r^s, \hat{q}_r^s, \hat{\boldsymbol{\beta}}_r^s, s = H, K$으로부터 앞서 정의한 4가지 측도들을 모두 계산할 수 있으며, Figures 3.3–3.5에서 $s = H, K$인 경우는 이 측도들을 자료의 개수별로 그린 것이다. 이 그림들을 보면 Hyndman과 Khandakar 방법에서는 자기회귀모형의 차수를 결정할 때 AICc보다는 BIC를 사용하는 것이 전반적으로 추정 결과가 좋음을 알 수 있다.

앞서 설명한 것처럼 adaptive LASSO와 Hyndman과 Khandakar 방법에서 모두 BIC를 사용한 경우가 추정 결과가 가장 좋으므로, 우리는 BIC를 사용한 경우에 한해서 두 방법을 비교하고자 한다. 먼저 Figure 3.3을 보면, $d = 0$인 경우에는 Hyndman과 Khandakar 방법이 adaptive LASSO 방법에 비해 모형식별도가 높긴 하지만 자료의 개수가 큰 경우에는 adaptive LASSO 방법과 큰 차이가 나지 않는다. $d = 1$인 경우를 보면 adaptive LASSO가 Hyndman과 Khandakar 방법에 비해 모형식별정확도가 더 높게 나타난다. 이는 Hyndman과 Khandakar 방법에서는 차분 차수와 자기회귀모형의 차수만을 선택하고, 부분모형을 고려하지 않았기 때문에 발생하는 현상이다. 실제 우리가 고려한 8개의 모형 중 ARIMA$(4, 1, 0)$와 ARIMA$(5, 1, 0)$는 나머지 6개의 모형과 다르게 $\beta_j = 0$, $1 < j < q$ (q는 자기회귀모형의 차수)을 만족하는 경우가 존재하는 경우로 두 모형에 대한 Hyndman과 Khandakar 방법의 모형

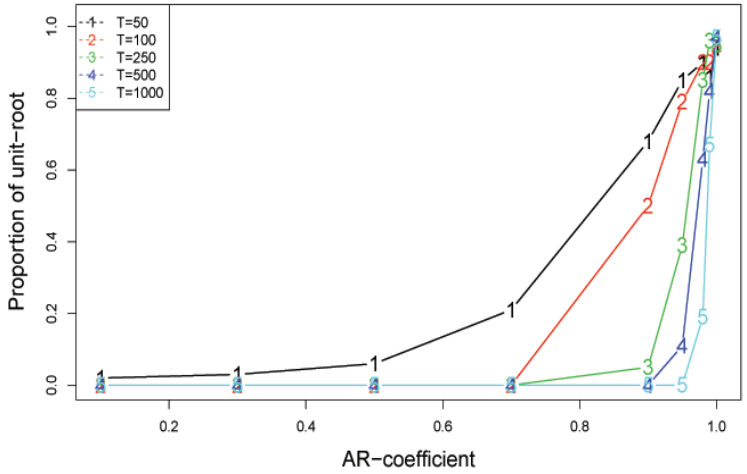

Figure 3.7. Proportion of unit-root.

식별정확도는 0이다. 실험 결과에서 보듯이 Hyndman과 Khandakar 방법에서는 추가로 계절형 자기회귀모형이나 부분모형을 선택하는 절차를 수행하지 않으면, 위와 같은 모형을 찾아낼 수 없다. 그러나 adaptive LASSO 방법을 사용하면 별도의 추가 분석 없이 최적의 부분모형을 선택할 수 있으며, 이는 기존 방법에 비해 큰 장점이라 할 수 있다.

Figure 3.5를 보면 단위근 식별 정확도는 adaptive LASSO가 Hyndman과 Khandakar 방법보다 우수하고, 차수 선택 정확도는 Hyndman과 Khandakar 방법이 adaptive LASSO보다 우수하다. 그리고 Figure 3.4에 주어진 MSE_s 측도를 비교해 보면 adaptive LASSO가 Hyndman과 Khandakar 방법보다 조금 더 우수하다. 종합적으로 두 방법을 비교하면, 측도에 따라 우수한 방법이 다르므로 어느 방법이 다른 방법에 비해 확실하게 우수하다고 판단하기는 어렵다. 그러나 일반적으로 자기회귀모형을 이용하여 분석할 때 단위근의 존재 여부에 따라 시계열의 성질과 예측값과 예측구간 등이 많이 달라지므로 차수보다는 단위근 존재 여부를 올바르게 판단하는 것이 더 중요하며, 이런 측면에서 본다면 단위근의 존재 여부를 좀 더 잘 구분해주는 adaptive LASSO 방법이 기존의 Hyndman과 Khandakar 방법보다 실제 자료분석에 적합할거라 생각한다.

마지막으로 단위근 존재 여부를 판단하는 정확성에 대해 좀 더 알아보기 위하여 1차 자기회귀모형의 계수를 다양하게 변화해가면서 모의실험을 수행하였다. 모의실험에서 고려한 모형의 구체적인 형태는 다음과 같다:

$$y_t = \phi^* y_{t-1} + \epsilon_t, \quad \phi^* = 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.98.0.99, 1.$$

오차항은 이전과 마찬가지로 표준정규분포를 따르는 백색잡음으로 하였다. 그리고 단위근 검정에서 ϕ^*의 값이 1에 가까울수록 단위근이 존재한다고 판단하는 경우가 많으므로, 본 연구에서도 0.9보다 큰 경우를 좀 더 세분하여 고려하였다.

본 연구에서 제안한 방법에 따르면, $\hat{\beta}_1^{\lambda}$의 값이 0이면 단위근이 존재한다고 판단하고 그렇지 않으면 단위근이 존재하지 않는다고 판단한다. 그러므로 총 100번의 반복 실험 중 $\hat{\beta}_1^{\lambda} = 0$을 만족하는 경우의 비율 P_{unit}을 구하여 단위근 식별에 대한 정확도를 판단하였다. 본 실험에서는 앞서 수행한 모의실험 결과에 따라 BIC를 이용하여 조절모수를 선택하였고, 이 때 구한 adaptive LASSO 추정값을 바탕으로 비율

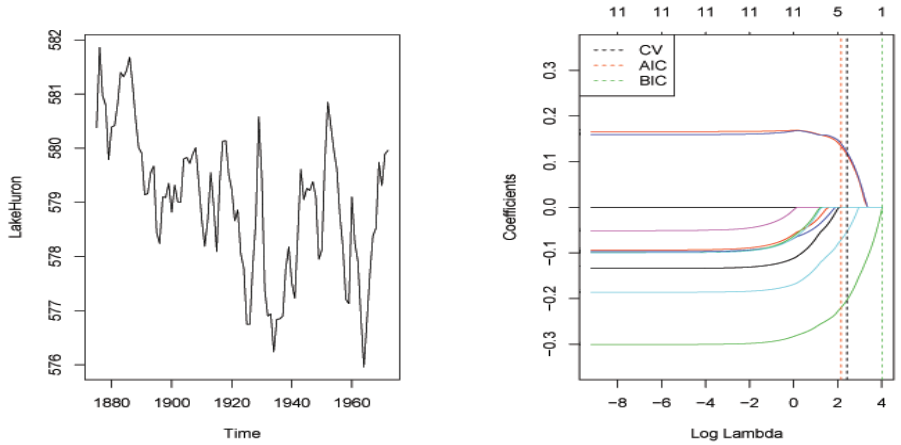

Figure 3.8. Time series plot and Path of adaptive LASSO coefficients.

P_{unit}을 계산하였다. Figure 3.7은 자기회귀모형의 계수 ϕ^*와 비율 P_{unit}에 대한 선형그래프를 자료의 개수별로 그린 것이다. 그림을 보면 모든 경우에 ϕ^*의 값이 커질수록, 특히 1에 가까운 값일수록 비율이 높아지는 것을 확인할 수 있다. 그리고 자료의 개수가 $T = 50$인 경우를 제외하고는 0.7 이하의 ϕ^*에 대해서 비율이 모두 0이다. 또한 비율이 0에서 양수로 바뀌는 지점에 해당하는 ϕ^*의 값이 자료의 개수가 증가할수록 1에 가까워지며, 같은 ϕ^*에 대한 비율 P_{unit}의 값이 자료의 개수가 증가할수록 작아진다. 이로부터 우리는 adaptive LASSO 방법이 자료가 많을수록 단위근의 존재 여부, 즉 시계열의 정상성 여부를 잘 판단할 수 있다고 기대할 수 있다.

3.2. 자료 분석

예제 자료로 R에 내장되어 있는 LakeHuron를 사용하였다. 이 자료는 1985년부터 1972년까지 Huron 호수의 연 평균 수위를 피트 단위로 측정한 것으로 총 98개의 관측값이 존재한다. Figure 3.8의 왼쪽 그림은 LakeHuron 자료에 대한 시계열 그림으로, 이로부터 우리는 LakeHuron 자료가 비정상 시계열임을 알 수 있다. 또한 ADF 검정을 LakeHuron 자료와 이를 차분한 자료에 대해 실시하면, 유의확률이 각각 0.254와 0.01보다 작은 값이 나와 LakeHuron 자료는 한 개의 단위근이 존재함을 확인할 수 있다. 우리가 제안한 방법에서도 첫 번째 계수 β_1에 대한 adaptive LASSO 추정값이 0이 되어 동일한 결과가 갖는다.

본 연구에서 제안한대로 adaptive LASSO 방법을 이용하면 Figure 3.8의 오른쪽 그림과 같은 추정값들을 얻게 된다. 그림에서 세 개의 수직 참조선은 각각 10-겹 교차검증, AIC, BIC에 의해 선택된 조절모수의 위치를 나타내며, 세 방법에 의해 선택된 모형은 각각 다음과 같다:

$$\begin{aligned}
\text{CV} : y_t - y_{t-1} &= 0.1153(y_{t-1} - y_{t-2}) - 0.2032(y_{t-2} - y_{t-3}) \\
&\quad - 0.0535(y_{t-4} - y_{t-5}) + 0.1202(y_{t-9} - y_{y-10}) + \epsilon_t; \\
\text{AIC} : y_t - y_{t-1} &= 0.1335(y_{t-1} - y_{t-2}) - 0.2198(y_{t-2} - y_{t-3}) \\
&\quad - 0.0714(y_{t-4} - y_{t-5}) + 0.1394(y_{t-9} - y_{y-10}) + \epsilon_t; \\
\text{BIC} : y_t - y_{t-1} &= \epsilon_t.
\end{aligned}$$

Hyndman과 Khandakar (2008)의 방법을 사용하면 ARIMA$(0,1,0)$ 모형이 최적의 모형으로 선택되며, 이는 우리의 방법 중 BIC를 이용하여 얻은 결과와 동일하다.

4. 결론

본 논문은 기존의 연구 결과에서 보듯이 정상성을 만족하는 시계열 모형에 국한되어 개발 연구되었던 adaptive LASSO 방법을 비정상 자기회귀모형으로 확장 연구한 것이다. 자기회기모형을 적합시키는 전통적인 절차가 단위근 검정을 통한 시계열의 정상성 판단, 정상화 변환을 거친 시계열에 대한 모형의 차수 선택, 모수 추정 및 진단의 3단계로 이루어진 것에 비해 본 논문에서 제안한 방법은 ADF 검정에서 사용하는 회귀모형에 adaptive LASSO 기법을 적용하여 모수를 추정하는 것으로 단위근 존재 여부와 차수 선택, 모수 추정을 동시에 자동으로 할 수 있다는 장점을 가지고 있다. 그리고 모의실험 결과를 보면 BIC를 최소화하는 조절모수를 사용하여 얻은 adaptive LASSO 추정량은 일치성을 만족할 뿐만 아니라 모형 식별에 대한 정확도도 높음을 알 수 있다. 또한 우리의 관심의 초첨이었던 단위근 존재 여부와 자기회귀모형의 차수도 자료의 개수가 충분히 많은 경우 비교적 정확하게 판단함을 확인할 수 있었다.

그러나 아직 비정상 시계열 모형에서의 adaptive LASSO 추정량에 대한 oracle property나 조절모수 선택에 대한 이론적인 연구는 미흡한 상태이다. 그리고 단위근이 2개 이상 존재하는 일반적인 경우에 대한 연구 또한 미비하다. 이에 단위근의 개수를 결정하는 것과 이론적인 연구는 후속 연구로 계속 진행할 필요가 있다.

References

Brockwell, P. J. and Davis, R. A. (2006). *Time Series: Theory and Methods* (2nd ed), Springer.

Chen, K. and Chan, K. S. (2011). Subset ARMA selection via the adaptive lasso, *Statistics and Its Interface*, **4**, 197–205.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R, *Journal of Statistical Software*, **27**, 1–22.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root, *Journal of Econometrics*, **54**, 159–178.

Kwon, S., Lee, S., and Na, O. (2017). Tuning parameter selection for the adaptive LASSO in the autoregressive model, *Journal of the Korean Statistical Society*, **46**, 285–297.

Nardi, Y. and Rinaldo, A. (2011). Autoregressive process modeling via the Lasso procedure, *Journal of Multivariate Analysis*, **102**, 528–549.

Phillips, P. C. and Perron, P. (1988). Testing for a unit root in time series regression, *Biometrika*, **75**, 335–346.

Said, S. E. and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order, *Biometrika*, **71**, 599–607.

Schwert, G. W. (1989). Tests for unit roots: a Monte Carlo investigation, *Journal of Business and Economic Statistics*, **7**, 147–160.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **71**, 671–683.

Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418–1429.

비정상 자기회귀모형에서의 벌점화 추정 기법에 대한 연구

나옥경[a,1]

[a]경기대학교 응용통계학과

(2019년 10월 7일 접수, 2019년 11월 4일 수정, 2019년 11월 5일 채택)

요 약

벌점화 추정 기법 중 adaptive LASSO 방법은 모형 선택과 모수 추정을 동시에 할 수 있는 유명한 방법으로 이미 정상 자기회귀모형에서 연구된 적이 있다. 본 논문에서는 이를 확장하여 확률보행과정과 같은 비정상 자기회귀모형에서 adaptive LASSO 추정량이 갖는 성질을 모의실험을 통해 연구하였다. 다만 비정상 자기회귀모형에서는 단위근의 존재 여부를 판단하는 것과 모형의 차수를 선택하는 것이 가장 중요하므로, 이를 위해 원 자기회귀모형이 아닌 ADF 검정에서 고려하는 회귀모형으로 변환하여 adaptive LASSO를 적용하였다. 일반적으로 Adaptive LASSO를 적용할 때 조절모수의 선택이 가장 중요한 문제이며, 본 논문에서는 교차검증, AIC, BIC 세 가지 방법을 이용하여 조절모수를 선택하였다. 모의실험 결과를 보면, 이 중에서 BIC가 최소가 되도록 선택한 조절모수에 대응되는 adaptive LASSO 추정량이 단위근의 존재 여부를 잘 판단할 뿐만 아니라 자기회귀모형의 차수 또한 비교적 정확하게 선택함을 확인할 수 있다.

주요용어: 자기회귀누적이동평균 모형, 단위근, 차수 선택, 벌점화 추정방법

이 논문는 2017학년도 경기대학교 연구년 수혜로 연구되었음.

[1](16227) 경기도 수원시 영통구 광교산로 154-42, 경기대학교 응용통계학과. E-mail: okna@kyonggi.ac.kr