The Korean Journal of Applied Statistics
2022, Vol. 35, No. 3, 395–406

DOI: https://doi.org/10.5351/KJAS.2022.35.3.395
Print ISSN1225-066X / Online ISSN 2383-5818

Outlier detection for multivariate long memory processes

Kyunghee Kim[a], Seungyeon Yu[a] Changryong Baek[1,a]

[a]Department of Statistics, Sungkyunkwan University

Abstract

This paper studies the outlier detection method for multivariate long memory time series. The existing outlier detection methods are based on a short memory VARMA model, so they are not suitable for multivariate long memory time series. It is because higher order of autoregressive model is necessary to account for long memory, however, it can also induce estimation instability as the number of parameter increases. To resolve this issue, we propose outlier detection methods based on the VHAR structure. We also adapt the robust estimation method to estimate VHAR coefficients more efficiently. Our simulation results show that our proposed method performs well in detecting outliers in multivariate long memory time series. Empirical analysis with stock index shows RVHAR model finds additional outliers that existing model does not detect.

Keywords: outlier detection, multivariate long memory process, VHAR, robust regression

1. 서론

4차 산업의 발달에 따라 이전보다 많은 양의 데이터를 수집할 수 있게 되면서 데이터 내에 이상점(outlier)이 다양한 형태와 양으로 존재하게 됐다. 여기서 이상점이란 Hawkins (1980)가 정의한 바에 따르면 자료 내 다수의 관측값들의 패턴에서 크게 벗어나는 값이다. 이상점은 시계열 데이터 분석의 추정 및 검정에 큰 영향을 미치기 때문에 탐지의 중요성이 높아지고 있다. 반면 장기억 혹은 장기 종속 시계열은 Boubacar-Maïnassara 등 (2021)에서 언급된 바와 같이 수문학(hydrology), 기상학(climatology), 경제학(economics) 등 산업 전반에 걸쳐 다양한 응용 분야에서 관측되고 있다. 따라서 이러한 장기억성을 가지는 다변량 시계열에서의 이상점 탐지는 최근 많은 관심을 받고 있으나 그 방법론에 대한 연구는 많이 이루어지지 않고 있다.

보통 다변량 자료에 이상점 탐지를 할 때는 단변량 이상점 탐지 기법을 시계열마다 반복하는 방법을 사용한다. 그러나 이는 차원 간 존재하는 교차 상관(cross correlation) 구조를 효과적으로 반영하지 못하여 비효율적이고 다변량 구조(multivariate framework)에 기반하여 이상점 탐지를 하는 것과 결과의 차이가 있다 (Tsay 등, 2000). 이를 극복하고자 Tsay 등 (2000)은 다변량 구조를 고려하여 VARMA (vector autoregressive moving average) 모형을 활용한 이상점 탐지 방법을 제안하였다. 그러나 VARMA는 주로 단기 종속 시계열 (short range dependence; SRD)을 분석할 때 사용하는 통계 모형으로 장기 종속 성질을 고려하여 이상점을 탐지하는데에는 한계가 있다. 장기억성을 고려해주기 위해서는 높은 차수의 AR 계수들이 필요한데 자료의 차원 및 길이에 대한 제약으로 계수의 추정이 불안정해질 수 있기 때문이다. 이에 따라 본 논문에서는 Lee와 Baek (2021)의 방법을 계승하되 적은 수의 계수로 장기억성을 효과적으로 모델링할 수 있는 VHAR (vector heterogeneous autoregressive) 모형을 사용하여 다변량 시계열의 특징을 고려하면서도 장기억 시계열의 이상점 탐지에 효과적인 방법론을 제시하고자 한다. 또한 계수 추정에 있어서 이상점의 영향을 덜 받아 오차를 줄이는 추정 방법을 위해서 로버스트 회귀를 사용하여 이상점 탐지력을 높이고자 한다.

[1] Corresponding author: Department of Statistics, Sunkyunkwan University, Sunkyunkwan ro 25-2, Seoul 03063, Korea.
E-mail: crbaek@skku.edu

Published 30 June 2022 / journal homepage: http://kjas.or.kr

본 논문의 구성은 다음과 같다. 2장에서는 장기 종속 시계열의 정의와 RVHAR 모형 그리고 본 논문에서 사용할 이상점 탐지 방법에 관해 설명한다. 3장에서는 모의실험을 통해 여러 다변량 시뮬레이션 데이터를 가정하여 기존 Tsay 등 (2000)이 제안한 방법 및 RVAR 모형과 RVHAR 모형의 성능을 비교한다. 4장에서는 실증 데이터로 2년간의 주가지수의 실현 변동성에 대해 RVHAR 모형을 통해 이상점을 탐지하고 분석결과에 대해 설명한다. 마지막으로 5장에서는 본 논문의 결론과 논의점에 대해 다룬다.

2. 방법론

2.1. 벡터 장기종속 시계열 모형

장기 종속 시계열 혹은 장기 기억 시계열은 시차(lag)가 있는 관측값끼리 지속해서 의존성을 보이는 정상 시계열로 자기 상관 함수(autocorrelation function; ACF)가 멱함수(power law) 형태로 0을 향해 천천히 감소하는 형태를 지닌다. 좀 더 엄밀하게는 스펙트럴 밀도함수가

$$f(\lambda) \sim \lambda^{-D} G \lambda^{-D}, \quad \lambda \to 0^{+},\ D = \mathrm{diag}(d_1, \ldots, d_k),$$

여기서 $d_i \in (0, 1/2)$, $i = 1, \ldots, k$로 주어진 k차원의 정상 시계열이다. 이때 λ^{-D}는 $\mathrm{diag}(\lambda^{-d_1}, \ldots, \lambda^{-d_k})$로 나타낼 수 있으며 G는 비음정부호(non-negative definite)로 주어진 Hermitian 대칭 행렬이다.

다변량 장기 종속 시계열에 대한 모수적 모형으로는 VARMA 모형에 분수 차분을 도입하여 확장한 모형이 널리 쓰인다. 하지만, 일반적으로 두 행렬의 곱은 교환 법칙이 성립하지 않으므로 벡터 AR 행렬과 분수 차분 행렬의 순서에 따라 두 모형이 존재한다. 먼저 VARFIMA(p, D, q) (vector autoregression fractional integrated moving average) 모형은

$$(I - B)^{D} \Phi(B) X_n = \Theta(B) \varepsilon_n \tag{2.1}$$

으로 주어지며 후행 연산자(backshift operator) $BX_t := X_{t-1}$에 대해서 $\Phi(z) = I_k - \Phi_1 z - \cdots - \Phi_p z^p$, $\Theta(z) = I_k - \Theta_1 z - \cdots - \Theta_q z^q$, ε_n는 백색잡음과정(white noise process)이며 분수 차분은 $(1-B)^D := \mathrm{diag}((1-B)^{d_1}, \ldots, (1-B)^{d_k})$에 대해서

$$(1 - B)^{d} := \sum_{j=0}^{\infty} (-1)^j \binom{d}{j} B^j = \sum_{j=0}^{\infty} (-1)^j \frac{\Gamma(d+1)}{\Gamma(j+1)\Gamma(d-j+1)} B^j$$

이다. 반면 FIVARMA(p, D, q) 모형은 AR 행렬과 차분의 순서를 바꾼 모형으로

$$\Phi(B)(I - B)^{D} X_n = \Theta(B) \varepsilon_n \tag{2.2}$$

으로 주어진다. 최대우도 추정량을 통해서 VARFIMA(p, D, q) 및 FIVARMA(p, D, q) 모형의 모수들을 추정할 수 있으며 자세한 내용은 Baek 등 (2017)을 참조하길 바란다. 본 논문에서는 두 모형의 혼동을 피하기 위해서 $p = 0$인 모형을 이용하여 시뮬레이션 연구를 진행하였다.

2.2. RVHAR 모형

장기 종속 모형인 FARIMA (fractionally integrated autoregressive moving average) 모형은 분수 차분을 이용한 모형이기에 추정에 있어서 계산이 복잡하고 해석이 어려운 단점이 있다. 이러한 단점을 극복하기 위해서 또 금융 시장의 특성을 반영하여 Corsi (2009)는 AR(22) 모형을 이용하여 장기 종속 모형을 근사시키는 모형인 HAR (heterogeneous autoregressive) 모형을 제안하였다. Baek과 Park (2021)은 이를 k차원 다변량 시계열로 확장한 VHAR 모형을 다음과 같이 제안하였다.

$$Y_t^{(d)} = \beta_0 + \Phi^{(d)} Y_{t-1}^{(d)} + \Phi^{(w)} Y_{t-1}^{(w)} + \Phi^{(m)} Y_{t-1}^{(m)} + \varepsilon_t, \qquad \varepsilon_t \sim WN(0, \Sigma),\ \ t = 23, \ldots, T. \tag{2.3}$$

여기에서 $Y_t^{(d)} \in \mathbb{R}^k$은 t 시점에서의 k차원 다변량 시계열 자료이고,

$$Y_{t-1}^{(w)} = \frac{1}{5}\sum_{j=1}^{5} Y_{t-j}^{(d)}, \qquad Y_{t-1}^{(m)} = \frac{1}{22}\sum_{j=1}^{22} Y_{t-j}^{(d)}, \tag{2.4}$$

여기서 $Y_{t-1}^{(w)}$, $Y_{t-1}^{(m)}$는 식 (2.4)로 정의한다. 이는 금융시장의 특성상 일주일을 5일로, 한 달을 22일로 표현한 집합체로서 각각 일별, 주별, 월별 영향을 의미하며 $\Phi^{(d)}, \Phi^{(w)}, \Phi^{(m)} \in \mathbb{R}^{k\times k}$은 각 일별, 주별, 월별 효과에 대한 모수 행렬이다. 따라서 VHAR 모형은 VAR(22) 모형의 특수한 형태로 장기억 속성을 잘 반영할 수 있도록 하되 제약 조건을 통해 모수의 개수를 $3k^2$개로 줄인 모형으로 생각할 수 있다.

본 논문에서는 VHAR 모형에 기반하여 다변량 장기억 시계열의 이상점을 탐지하고자 한다. 이를 위해서는 먼저 VHAR 모형의 모수를 추정하는 과정이 필요하나, 이상점의 영향을 최소화한 후 모수를 추정할 필요가 있다. 즉 시계열 자료에 이상점이 존재할 경우, 주변 이상점으로 인해 추정량이 편향될 수 있으며 백색잡음의 분산인 Σ 추정에도 영향을 줄 수 있다. 이를 효과적으로 줄이기 위해 본 논문은 후버의 로버스트 회귀(Huber's robust regression)를 적용하여 보다 신뢰성 있는 추정값을 사용하였다 (Kleiner 등, 1979). Fox와 Weisberg (2002)에서 계산한 대로 후버의 로버스트 M-계수를 얻는 방법은 다음과 같다.

우선 식 (2.3)을 회귀방정식 모형으로 재구조화하면

$$\mathbb{Y} = \mathbb{A}\mathbb{X} + \mathbb{Z},$$

$$\mathbb{Y} := \left(Y_{23}^{(d)}, \ldots, Y_T^{(d)}\right), \quad \mathbb{A} := \left(\Phi^{(d)}, \Phi^{(w)}, \Phi^{(m)}\right), \quad \mathbb{X} := (X_{22}, \ldots, X_{T-1}), \quad X := \begin{pmatrix} Y_t^{(d)} \\ Y_t^{(w)} \\ Y_t^{(m)} \end{pmatrix}, \quad \mathbb{Z} := (\varepsilon_{23}, \ldots, \varepsilon_T)$$

으로 표현이 가능하다. 이를 벡터화하면

$$\mathrm{vec}(\mathbb{Y}) = (\mathbb{X}^\intercal \otimes I_k)\mathrm{vec}(\mathbb{A}) + \mathrm{vec}(\mathbb{Z}) \tag{2.5}$$

으로 나타낼 수 있으므로 로버스트 회귀계수는

$$\mathrm{vec}\left(\hat{\mathbb{A}}\right) = \underset{\mathbb{A}}{\mathrm{argmin}}\, \rho(\mathrm{vec}(\mathbb{Y}) - (\mathbb{X}^\intercal \otimes \mathrm{I_k})\mathrm{vec}(\mathbb{A})) \tag{2.6}$$

을 통해 추정할 수 있다. 여기에서 쓰인 손실함수 ρ는 가장 널리 쓰이는 $\rho'(\cdot) = \psi(\cdot)$, $\psi(x) = \max\{-c, \min(c, x)\}$, $c = 1.345$을 사용하였다 (Lee와 Baek, 2021; Ripley 등, 2013).

본 논문에서 제안하는 장기억 시계열 다변량 이상점 탐지 방법은 로버스트한 추정 기법을 적용한 VHAR 모형을 기반으로 검정 통계량을 계산한다. 앞으로 위와 같이 로버스트한 추정 기법을 적용한 VHAR 모형을 RVHAR (robustly estimated VHAR) 모형이라 정의한다. 또한, VAR 모형에 동일한 방법을 적용한 경우는 RVAR (robustly estimated VAR) 모형이라 정의한다.

2.3. 이상점 탐지 방법

다변량 시계열에서의 이상점은 모든 차원에 대해서 동일한 혹은 비슷한 시점에서 나타날 수도 있고, 특정 차원에 대해서만 나타날 수도 있다. 이러한 구조적인 특성을 무시한 채 단변량 이상점 탐지 방법을 모든 차원에 적용한다면 과도하게 이상점들을 찾을 수 있고 결합 변동성(joint dynamics)을 고려하지 못해 효율적이지 않은 추정 방법(inferior estimation)을 사용하게 된다. 본 연구에서는 다변량 구조를 직접 이상점 탐지에 적용해 단변량 이상점 탐지의 문제 를 극복하고자 하며, 특히 장기종속 시계열의 특징을 반영하기 위해서 VHAR 모형을 통해 탐지 능력을 키우고자 한다. 구체적인 방법론은 다음과 같다.

본 논문에서 제안한 방법은 VARMA를 사용하는 Tsay 등 (2000)의 다변량 이상점 탐지 방법을 계승하였다. VHAR 모형은 제약된 VAR(22) 모형의 표현에 따라

$$\widetilde{\Phi}(B)Y_t^{(d)} = c + \varepsilon_t, \quad t = 1, \ldots, T \tag{2.7}$$

로 표현할 수 있다. 여기서 c는 절편(intercept), T는 시계열 길이, $\Phi(B)$는 VHAR 모형의 특성 다항식으로

$$\begin{aligned}\widetilde{\Phi}(B) &:= I - \widetilde{\Phi}_1 B - \cdots - \widetilde{\Phi}_{22} B^{22} \\ &= I - \left(\Phi^{(d)} Y_{t-1}^{(d)} + \frac{\Phi^{(w)} Y_{t-1}^{(w)}}{5} + \frac{\Phi^{(m)} Y_{t-1}^{(m)}}{22}\right) B - \left(\frac{\Phi^{(w)} Y_{t-1}^{(w)}}{5} + \frac{\Phi^{(m)} Y_{t-1}^{(m)}}{22}\right) B^2 - \cdots - \left(\frac{\Phi^{(m)} Y_{t-1}^{(m)}}{22}\right) B^{22}\end{aligned} \tag{2.8}$$

이며 정상성을 위해 모든 해가 단위 원(unit circle) 밖에 있다고 가정한다. 이노베이션 $\{\varepsilon_t := (\varepsilon_{1t}, \ldots, \varepsilon_{kt})^\intercal\}$는 평균이 $\mathbf{0}$, 분산이 Σ인 i.i.d. k차원 다변량 정규벡터임을 가정한다.

시점 h에서 이상점의 유무를 나타내는 지시함수를 $\xi_t^{(h)}$라 정의하고, 이상점 크기를 ω로 표현할 때, 이상점이 있는 시계열 Y_t^*은

$$Y_t^* = Y_t^{(d)} + \omega\alpha(B)\xi_t^{(h)}, \quad \omega = (\omega_1, \ldots, \omega_k)^\intercal \tag{2.9}$$

으로 나타낼 수 있다. 수식 (2.9)의 $\alpha(B)$는 이상점의 효과를 나타내는 함수로 Fox (1972)가 제안한 이상점의 종류에 따라

$$\alpha(B) = \begin{cases} (1-\delta B)^{-1}, & \delta = 0 \quad \text{(AO)} \\ (1-\delta B)^{-1}, & 0 < \delta < 1 \quad \text{(TC)} \\ (1-\delta B)^{-1}, & \delta = 1 \quad \text{(LS)} \end{cases} \tag{2.10}$$

와 같이 주어진다. 여기에서 AO는 additive outlier, TC는 temporary change, LS는 level shift를 나타내는 약어로 이상점이 영향을 미치는 정도를 초모수(hyperparameter) δ를 통해 모형화 할 수 있다. 즉 δ가 1에 가까울수록 이상점이 발생한 이후에도 지속적인 이상점 효과를 주고 0에 가까울수록 특정 시점에만 영향을 미침을 뜻한다.

수식 (2.7)과 (2.9)를 이용하면 VHAR 모형의 오차가

$$a_t := \widetilde{\Phi}(B)Y_t^* - c = \varepsilon_t + \omega\widetilde{\Phi}(B)\alpha(B)\xi_t^{(h)}$$

임을 알 수 있다. 따라서 $\varepsilon_t \sim N(0, \Sigma)$와 이상점이 없다는 가정 하에서 추정한 VHAR 모형 계수 및 잔차 $\hat{a}_t$를 이용하면 이상치의 추정치 $\hat{\omega}$를 일반화 최소 제곱을 통해 다음과 같이 구할 수 있다.

$$\hat{\omega}_{i,h} = -\left(\sum_{j=0}^{n-h} \hat{\Phi}_j^{*\intercal} \Sigma^{-1} \hat{\Phi}_j^*\right)^{-1} \sum_{j=0}^{n-h} \hat{\Phi}^* \Sigma^{-1} \hat{a}_{h+j}, \quad \Phi^* = \widetilde{\Phi}(B)\alpha(B), \quad (i = \text{AO, LS, TC}). \tag{2.11}$$

단변량의 이상점 탐지는 t-검정 통계량와 임계값의 비교를 통해 이상점을 판단하지만, Tsay 등 (2000)에서는 다변량의 이상점 추정을 위한 새로운 검정 통계량으로 J 통계량(Joint Statistic)과 C 통계량(Component Statistic)을 제안했다. 여기서 주목해야 할 점은 단변량 시계열에서 오차 a_t는 $\Phi(B)\alpha(B)$에 대해서만 영향을 받지만, 다변량 시계열에서는 $\Phi(B)\alpha(B)$와 ω 사이의 상호작용에 의존한다는 것이다. 이를 고려하여 본 논문에서 사용되는 이상치 탐지 검정 통계량에는 ω의 추정량이 사용된다. 먼저 J 통계량은 타 시계열들과 정보를 결합해 ω 를 고려한다. 다음으로 C 통계량은 타 시계열을 고려하지 않고 개별 시점의 ω 정보만을 이용한다. 이 검정의 대립가설은 이상점 효과가 있음을 의미하는 $\omega \neq 0$으로 둔다.

$$J_{i,h} = \hat{\omega}_{i,h}' \Sigma_{i,h}^{-1} \hat{\omega}_{i,h}, \quad C_{i,h} = \frac{\max\limits_{1 \le j \le k} |\hat{\omega}_{j,i,k}|}{\sqrt{\sigma_{j,i,h}}} \tag{2.12}$$

여기에서 i는 이상점의 종류를 나타내는 첨자이고, $\hat{\omega}_{j,i,h}$는 $\hat{\omega}_{i,h}$의 j번째 요소를, $\sigma_{j,i,h}$는 행렬 $\Sigma_{i,h}$의 (j, j)에 위치하는 요소를 나타낸다. 위 식 (2.12)를 이용해 모든 시점에서 이상점 종류에 따라 J 통계량과 C 통계량을 구한 후, 식 (2.13)과 같이 전체 시점 중 최대 값을 구하면 이상점 탐지에 사용되는 통계량 $J_{\max}(i, h_i)$와 $C_{\max}(i, h_i^*)$를 얻을 수 있다.

$$J_{\max}(i, h_i) = \max_h J_{i,h}, \quad C_{\max}(i, h_i^*) = \max_h C_{i,h} \tag{2.13}$$

$J_{\max}$의 h_i는 J 통계량의 최댓값이 h시점에서 이상점 종류가 i일때 발생했음을 뜻한다. 유사하게 $C_{\max}$의 h_i^*는 C 통계량의 최댓값이 h^*시점에서 이상점 종류가 i 일때 발생했다는 의미이다. J 통계량의 임계값(critical value)을 C_J, C 통계량의 임계값을 C_c라면 다변량 이상점 탐지는 다음과 같은 절차로 진행된다.

Step 1: $J_{\max}(i, h_i) > C_J$인 경우를 찾아 이상점이라 판단한다.

Step 2: 이상점을 탐지한 경우, 해당 시점의 이상점 효과(ω)를 제거한다. 이 조정된 데이터(adjusted series)로 Step 1의 과정을 재시행한다.

Step 3: $J_{\max}(i, h_i) > C_J$를 만족하는 경우가 더 이상 없을 때, $C_{\max}(i, h_i^*) > C_C$인 경우를 이상점이라 판단한다.

Step 4: 이상점을 탐지한 경우, 해당 이상점 효과(ω)를 제거하며 이 조정된 데이터에 Step 3을 재시행한다.

Step 5: 더 이상의 이상점을 찾지 못하고 이상점 탐지를 종료한다.

임계값 C_J와 C_C는 일종의 Sieve 붓스트랩 방법을 통해서 추정하였다. 구체적인 기각역 추정 절차는 다음과 같다. 우선 RVHAR 모형을 자료에 적합하여 모형의 계수를 추정한다. 이 추정된 계수를 기반으로 시뮬레이션을 통해서 자료의 크기와 똑같은 붓스트랩 샘플을 생성하고, 잔차를 얻는다. 이때 검정 통계량 $J_{max}(i, h_i)$, $C_{max}(i, h_i)$은 잔차에 기반해 계산되기에 새로운 통계량들이 반복 수 만큼 생성된다. 따라서 이 과정을, 1000번 반복 시행하였으며, 구한 통계량들을 분위수에 맞춰 계산해 최종 경험적 기각역을 얻었다. 이 절차는 VAR 모형으로 시계열 이상 탐지를 제안한 Tsay 등 (2000)의 연구에서 기각역을 구한 방법과 동일하게 구현되었다.

3. 모의 실험

본 절에서는 제안한 RVHAR 모형을 이용한 장기억 다변량 시계열 이상점 탐지 성능을 알아보고자 모의실험을 시행하였다. 실험에는 시계열 길이가 300이고 차원의 수가 5와 10인 VFARIMA 모형을 사용했다. 모수 d는 상호 의존성의 정도에 따른 성능을 살펴보기 위해 세 개의 구간 [0.01, 0.15), [0.15, 0.3), [0.3, 0.45)에서 각각 랜덤하게 d를 추출해 모의 실험을 진행하였다. 차분의 순서에 영향을 받지 않도록 AR 모형의 차수 $p = 0$을 사용하였고 MA부분은 Tsay (1988)의 모의 실험과 같은 조건으로 $q = 1$을 사용하였다. 본 절에서 사용한 자료생성모형(data generating processes; DGPs)은 다음과 같다.

(DGP1)

$$\Theta_1 = \mathrm{diag}(0.5, 0.4, 0.1, 0.3, 0.7), \quad \Sigma_1 = \begin{bmatrix} 1 & 0.2 & \cdots & 0.2 \\ 0.2 & 1 & \cdots & 0.2 \\ \vdots & \vdots & \ddots & \vdots \\ 0.2 & 0.2 & \cdots & 1 \end{bmatrix}_{5\times 5},$$

(DGP2)

$$\Theta_2 = \mathrm{diag}(0.8, 0.6, 0.3, 0.47, 0.9, 0.1, 0.5, 0.6, 0.35, 0.2), \quad \Sigma_2 = \begin{bmatrix} 1 & 0.2 & \cdots & 0.2 \\ 0.2 & 1 & \cdots & 0.2 \\ \vdots & \vdots & \ddots & \vdots \\ 0.2 & 0.2 & \cdots & 1 \end{bmatrix}_{10\times 10}.$$

Table 1: Performance measures for DGP1 between RVHAR and RVAR(10)

		J statistics			*C* statistics		
	ω	[0.01, 0.15)	[0.15, 0.3)	[0.3, 0.45)	[0.01, 0.15)	[0.15, 0.3)	[0.3, 0.45)
AO	2	7.81	7.95	6.76	0.33	0.3	0.18
	3	12.76	11.61	12.79	0.41	0.39	0.3
	4	16.13	16.28	17.94	0.28	0.28	0.28
LS	2	7.49	4.49	2.12	0.79	0.48	0.3
	3	11.20	8.52	4.09	1.01	0.84	0.47
	4	12.10	10.85	9.13	0.89	0.88	0.78
TC	2	−0.18	0.53	1.85	0.02	0.05	0.22
	3	0.61	0.85	0.69	0.13	0.07	0.08
	4	1.01	2.50	4.22	0.15	0.21	0.30

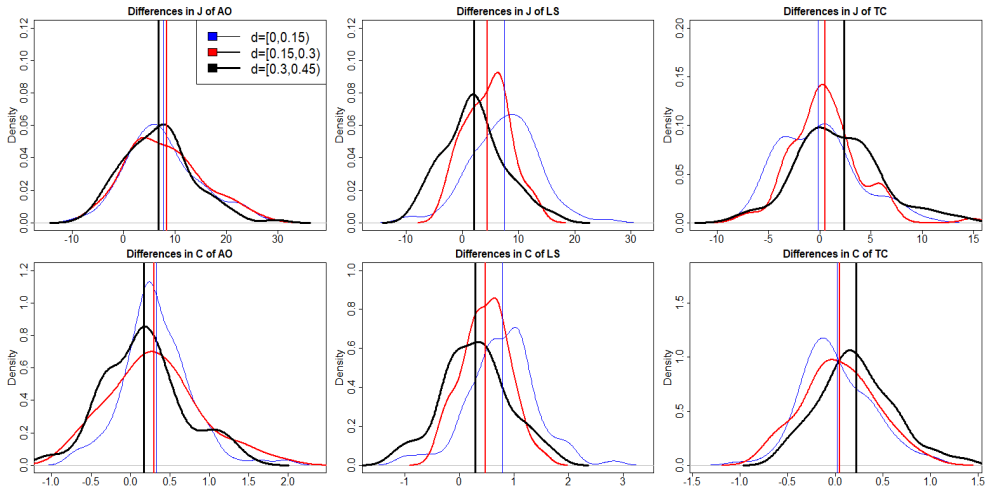

Figure 1: *Density estimation of performance measures on J test statistics (top) and C statistics (bottom) for DGP1 between RVHAR and RVAR(10).*

실험을 위해 데이터 중간 지점인 $t = 150$에 세 가지 이상점 형태 (AO, TC, LS)를 같은 시나리오 하에서 세가지 크기로 삽입하였다. 이때 이상점의 크기는 $\{2, 3, 4\}$를 사용하였고 이상점의 크기에 따른 성능도 비교하였다. 방법론에 따른 성능은 벤치마킹 모형인 RVAR(p)모형과의 검정 통계량 차이에 대한 경험적 평균

$$\mathbb{E}\left(\tau^H(150; i) - \tau^A(150; i)\right), \quad i = \text{AO, LS, TC} \tag{3.1}$$

을 사용하였다. 여기서 τ^H는 RVHAR의 검정 통계량 값, τ^A는 RVAR의 검정 통계량 값이며 i는 이상점 종류를 의미한다. 따라서, 경험적 평균이 양수가 나올 경우, 평균적으로 RVHAR의 검정 통계량이 RVAR의 검정 통계량보다 커서 이상점을 강하게 탐지할 가능성이 높은 것으로 간주한다. 반대로 경험적 평균이 음수가 나올 경우, RVAR이 RVHAR보다 이상점을 잘 탐지할 확률이 높은 것으로 간주하게 된다. 벤치마크 모형인 RVAR(p)의 차수 p는 5와 10을 사용하여 장기억성을 충분히 반영할 수 있도록 하였다. 두 차수에 대한 결과 차이는 미미하여 논문의 간결성을 위하여 $p = 10$인 경우에 대해서만 보고하였다. TC에서 사용되는 δ는 이상점 이후에도 일시적 영향을 줄 수 있도록 0.8을 부여했다.

Table 1은 DGP1에 대해서 RVAR(10) 모형과 RVHAR 모형의 비교를 요약한 표이다. 또한, Figure 1은 이

Table 2: Performance measures for DGP1 between RVHAR and RVAR(10)

	ω	J statistics			C statistics		
		[0.01, 0.15)	[0.15, 0.3)	[0.3, 0.45)	[0.01, 0.15)	[0.15, 0.3)	[0.3, 0.45)
AO	2	16.73	20.56	18.98	0.21	0.29	0.12
	3	28.74	29.97	30.94	0.25	0.36	0.27
	4	34.62	39.91	42.06	0.23	0.23	0.37
LS	2	3.36	0.49	−6.13	0.91	0.57	0.01
	3	13.28	9.21	2.16	1.46	1.25	0.72
	4	20.10	18.34	10.60	1.70	1.73	1.46
TC	2	−3.36	−3.84	−2.49	−0.17	−0.09	0.01
	3	−3.17	−3.50	−2.31	−0.02	0.04	0.01
	4	−0.50	−0.35	−0.01	0.20	0.16	0.16

표들을 그림으로 표현하였고 그림의 수직선은 d 에 따른 밀도 함수의 평균값이며 선이 두꺼워 질수록 d가 커짐을 의미한다. 수직선들을 참고하면 대부분의 통계량 평균값들이 0 이상임을 볼 수 있다. 즉, 통계량 차이 평균값이 RVHAR 모형이 RVAR 모형보다 더 크다는 결과를 시각적으로 확인 할 수 있다. 몇 가지 주요 결과는 다음과 같다. 먼저 LRD 모수 d와 관계없이 이상점 크기 ω가 커지면 RVHAR과 RVAR간의 격차도 커짐을 확인할 수 있다. 이는 Lee (2017)에서 충격량 즉, 이상치 크기가 커질수록 이상치를 정확하게 탐지한 비율이 높아졌던 연구 결과로 위 결과의 타당성을 뒷받침 할 수 있다. 다음으로 이상점 종류에 따른 결과 해석은 아래와 같다. 먼저 AO, LS의 경우 두 검정 통계량의 차이가 모두 양수로 본 논문에서 제안한 방법론이 우수함을 살펴볼 수 있다. 다만 C 검정 통계량에서의 차이는 J 검정 통계량과 비교하였을 때 상대적으로 작음을 살펴볼 수 있다. 이는 시뮬레이션 모형의 이상점이 모든 차원에서 한꺼번에 변화하기 때문에 각 차원의 합을 고려한 J 통계량이 한 차원의 최댓값을 사용하는 C 통계량보다 이상점 탐지에 더 유리한 영향인 것으로 보인다. 또한, TC에 대해서도 대부분은 양수 값을 가져 RVHAR에 기반한 이상점 탐지가 더 효과적임을 확인할 수 있다. 단 d의 값이 작고, $\omega = 2$인 경우엔 음수의 값을 가지는 것을 관찰할 수 있다. 이는 단변량인 경우에서도 Lee와 Baek (2021)에서도 보고했듯이 장기억 시계열과 평균변화 모형이 구별하기 어려움에 따라 관측된 것으로 추측한다. 이상점의 크기가 커지거나 장기억성이 강해지면 TC의 경우라 할지라도 본 논문에서 제안한 방법론의 성능이 좋아짐을 확인할 수 있다. 따라서 RVHAR모형이 장기억 시계열 이상점 탐지의 성질을 반영하여 탐지하고 있다고 볼 수 있다.

다음으로 차원을 늘려 DGP2에 대한 이상점 탐지를 시행한 결과를 Table 2와 Figure 2에 요약하였다. 전반적으로 결과는 DGP1의 경우와 비슷하다. 이상점의 크기가 클수록 두 모형에 따른 검정 통계량의 차이가 벌어짐을 확인할 수 있으며, AO의 경우 모두 양수를 가져 RVHAR 모형에 기반한 이상점 탐지가 효과적임을 확인할 수 있다. LS의 경우 d의 값이 크고 $\omega = 2$일 때 J 검정 통계량이 음수의 값을 가지는 것을 관찰할 수 있다. 이 또한 $k = 5$의 경우와 마찬가지로 장기억 시계열과 평균변화 모형이 구별하기 어렵기 때문이라고 본다. TC의 경우 두 검정 통계량의 차이가 음수로, 특히 J통계량의 경우, RVHAR의 성능이 RVAR(10)보다 좋지 않음을 알 수 있다. 덧붙여 DGP1의 경우와 마찬가지로 d의 값이 커지거나 이상점의 크기가 커지면 두 검정 통계량의 차이가 줄어드는 것을 관찰할 수 있다. 일반적으로 평균 변화 모형과 장기억 시계열 모형의 구분이 어려움이 잘 알려져 이에 대한 영향으로 파악되나 더 자세한 원인에 대한 분석을 위한 추가 연구가 필요할 것으로 보인다.

본 절에서는 제안한 방법론의 성능을 살펴보기 위해서 RVAR(10) 모형을 벤치마킹하여 성능을 비교하였다. 그 결과 AO 및 LS의 경우에는 고려한 시뮬레이션 세팅 모두에서 RVHAR 모형에 기반한 방법이 우수함을 살펴볼 수 있었다. 또한 모든 차원에서 이상점이 같은 크기로 변화하기에 J 통계량이 C 통계량보다 더 좋은 성능을 보였다. TC의 경우 d의 값이 크거나 이상점의 크기가 클 때는 RVHAR 모형에 기반한 방법론이

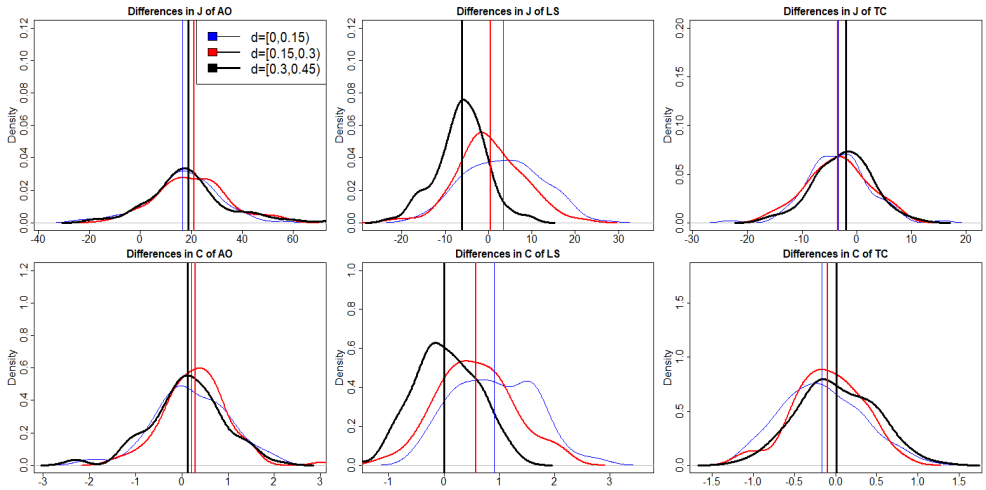

Figure 2: *Density estimation of performance measures on J test statistics (top) and C statistics (bottom) for DGP2 between RVHAR and RVAR(10).*

우수하였으나 일부 RVAR(10)의 성능이 좋은 점은 보완할 사항으로 파악된다.

4. 실증자료 분석

본 절에서는 실증자료에 대해 RVHAR 모형과 RVAR 모형의 다변량 이상점 탐지 성능을 3절과 동일하게 식 (3.1)에 근거해 비교한다. 실증자료는 2020년 1월 1일부터 2021년 12월 31일 기간에 대해 5가지 주요 주가지수 Dow Jones, Nikkei 225, S&P 500, Hang Seng, FTSE 100의 실현 변동성(realized volatility; RV)으로 (`https://realized.oxford-man.ox.ac.uk`)에서 수집하였다. 로그-수익률(log return)은 일별 자료 t시점의 자산 가격 P_t에 대해서

$$r_t = \log(P_t) - \log(P_{t-1})$$

로 널리 사용되는 지표이다. 이때 로그-수익률을 기반으로 자산 관리를 위해 다양한 변동성 추정 방법론들이 등장하게 되었는데, 그 중 대표적인 추정량이 Andersen 등 (2003)이 제안한 실현 변동성이다. 실현 변동성은 하루 중 M개의 등간격의 로그 수익률의 제곱 합이며 다음의 식으로 계산할 수 있다.

$$RV_t^{(d)} = \sqrt{\sum_{j=1}^{M} r_{t-j\cdot\Delta}^2}, \quad t = 1, 2, 3, \ldots$$

실증분석에는 10분 간격에 대한 실현 변동성 데이터에 로그 변환을 적용한 자료를 사용하였다. 최종 데이터는 Figure 3에 보여지는 것과 같이 시차가 증가함에 따라 자기상관성이 천천히 감소하는 것을 볼 수 있어 장기억 시계열 데이터라는 것을 확인할 수 있다. 또한 2.1절에서 언급한 d 값을 추정하기 위해 국소 휘틀(local Whittle; LW) 추정 값을 계산한 결과 d의 값이 대부분 0.4에서 0.5사이의 값을 가지고 있어 정상 장기억 시계열임을 확인하였다.

2.3절에 서술한 방법에 따라 Sieve 붓스트랩을 이용하여 얻은 기각역은 Table 3과 같다. RVHAR 모형과의 비교를 위해서 RVAR(5)모형을 사용하였고 기각역은 유의 수준 10%에 기반하여 이상점을 탐지하였다. Figure

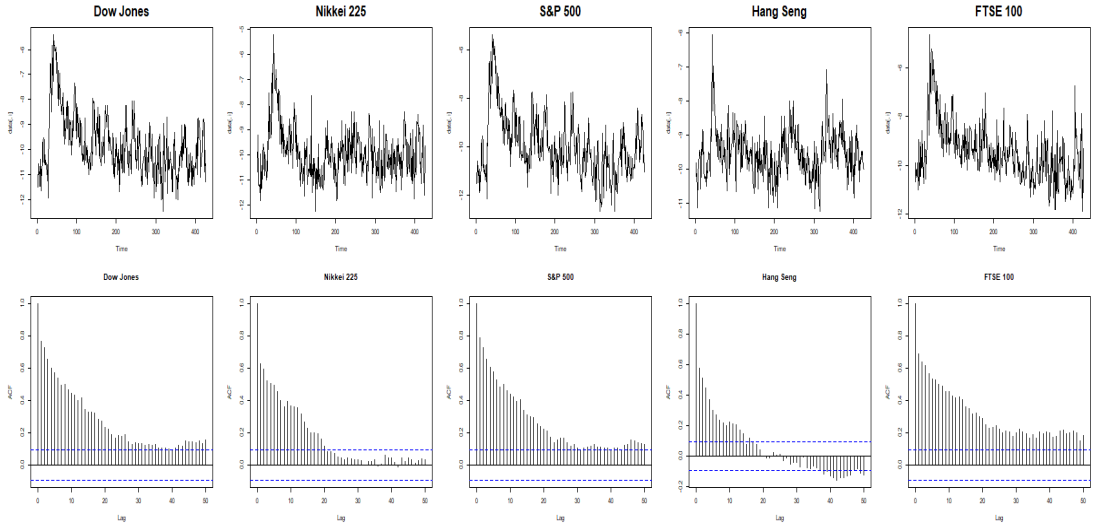

Figure 3: *Time plot and ACF plot of Realized Volatility Data for Index.*

Table 3: Critical region of Index data

Test	RVHAR					RVAR(5)				
	50%	90%	95%	97.5%	99%	50%	90%	95%	97.5%	99%
$J_{\max}(A, h_A)$	19.15	23.34	24.75	25.80	27.55	19.45	23.80	25.03	26.30	28.00
$J_{\max}(L, h_L)$	16.12	21.40	23.05	25.16	27.75	15.62	21.06	22.77	24.26	26.29
$J_{\max}(T, h_T)$	19.03	23.38	25.37	27.03	29.52	19.28	23.39	24.98	26.31	27.74
$C_{\max}(A, h_A)$	3.57	4.02	4.16	4.30	4.40	3.59	4.05	4.23	4.35	4.49
$C_{\max}(L, h_L)$	3.42	3.92	4.10	4.23	4.47	3.39	3.90	4.05	4.19	4.38
$C_{\max}(T, h_T)$	3.54	4.04	4.22	4.40	4.69	3.57	4.02	4.19	4.31	4.44

Table 4: Detected outliers by RVAR(5) model and RVHAR model (90%)

Iteration	Time	Date	RVHAR			RVAR(5)		
			$J_{max}(i, h_i)$	$C_{max}(i, h_i)$	Type	$J_{max}(i, h_i)$	$C_{max}(i, h_i)$	Type
1	38	2020-03-09	24.44	4.15	TC	22.83	4.41	LS
2	42	2020-03-13	25.03	4.96	TC	39.60	5.94	LS
3	138	2020-08-28	26.16	4.83	AO	24.29	4.76	AO
4	406	2021-11-26	20.98	4.12	AO	-	-	-

4는 실현 변동성 자료에 대한 RVAR(5)과 RVHAR 모형 기반 이상점 탐지 결과로 Table 4를 시각적으로 요약해 보여준다. 빨간색 선과 파란색 선은 각각 RVAR(5)과 RVHAR 이상점 탐지 시점을 의미한다.

Table 4는 유의 수준 10% 하에서 각 모형이 탐지한 이상점을 시간의 흐름 순으로 나타낸 결과이다. RVAR(5)은 총 3개, RVHAR은 총 4개의 이상점을 찾아냈고, RVAR이 찾은 시점들은 모두 RVHAR도 탐지했다. 구체적으로 시점 2020년 3월 9일, 3월 13일, 8월 28일은 공통으로 이상 탐지했고 시점 2021년 11월 26일은 RVHAR 모형만 탐지했다. 공통 탐지 시점 중 2020년 3월 9일, 8월 28일에서 본 연구에서 제안한 RVHAR 모형이 RVAR(5) 모형보다 큰 검정 통계량 값을 산출해 더 강력하게 이상점을 탐지하였다. 한편 RVAR(5) 모

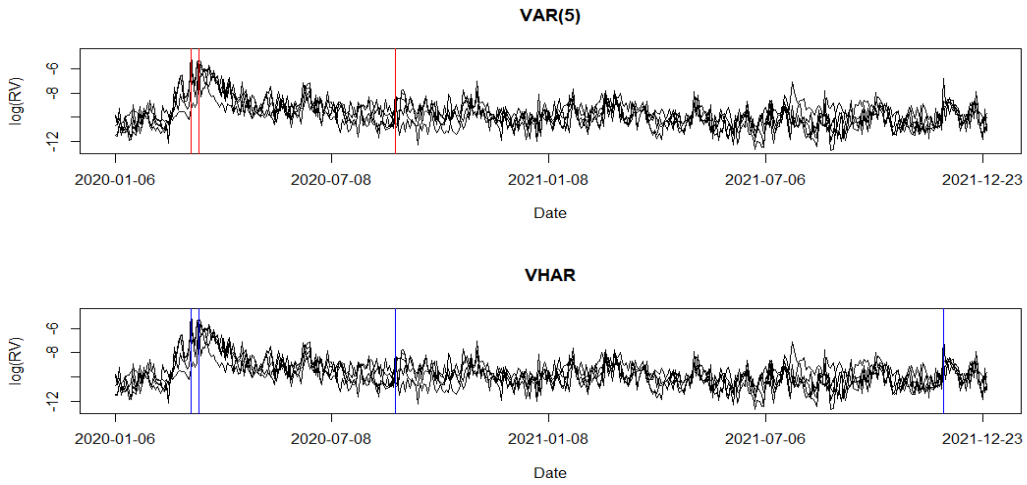

Figure 4: *Outlier detection result of realized volatility data (90%).*

형은 150 이후의 시점들에 대해 전혀 탐지하지 못했는데 Figure 4를 보면 해당 시점들에서 시계열들의 급격한 변화 형태를 확인할 수 있어 확연하게 이상점으로 판단하는 편이 나아 보인다. 반면 RVHAR 모형은 2021년 11월 26일 시점을 이상점으로 탐지해 RVAR(5)모형보다 적절하게 이상점을 탐지했다고 볼 수 있다. 특히, 2021년 11월 26일은 오미크론 변이 공포로 인해 S&P 500과 다우존스, FTSE 100 등의 지수들이 동반 하락한 시점으로 다변량 이상점으로 고려하기에 적절하다. 한가지 주목해야 할 점은 2020년 3월 9일과 3월 13일 시점에서는 RVHAR이 TC로 탐지하였는데 RVAR(5)이 LS로 탐지했다. 이를 통해 같은 시점을 이상점으로 탐지하더라도 TC와 LS의 구분이 어렵기 때문에 이상점의 타입을 구분하는데 유의해야 한다는 것을 확인할 수 있다.

요약하면 본 실증데이터 분석에선 RVHAR 기반 알고리즘이 장기기억 종속 시계열 자료의 이상점을 더 많이 탐지했고 RVAR을 통해 찾아낸 이상점은 RVHAR도 모두 공통으로 검출하였음을 볼 수 있다. 그뿐만 아니라 대부분의 이상점에서 더 큰 검정 통계량 값을 가져 강력하게 이상점들을 탐지해냄을 확인할 수 있다. 더 나아가 RVHAR은 11월 26일과 같이 RVAR이 탐지하지 못했던 구간들에 대해서도 이상점을 잘 탐지하는 모습을 보인다.

5. 결론 및 논의

본 논문은 장기억성을 가진 고차원 자료를 효율적으로 모델링 할 수 있도록 RVHAR 모형에 기반한 이상점 탐지 방법론을 제안하였다. 3장 모의 연구를 통해 제안한 방법론이 벤치마킹 모형인 높은 차원의 RVAR(10) 모형보다 다변량 시계열의 차원 수, 이상점 크기, 종류 등에 대한 다양한 시나리오에서 더 큰 검정 통계량 값을 산출하여 이상점 탐지력이 높았음을 확인할 수 있었고 특히 AO와 LS는 대다수 양수 값으로 나타나 RVHAR 모형의 성능이 우수하였다고 설명할 수 있다. 제안한 방법론을 실증 자료인 주가지수의 실현 변동성 자료에 적용한 결과는 RVAR 모형이 탐지한 대부분의 이상점을 RVHAR도 동일하게 검출했다. 또한 기존 RVAR 모형에서 탐지하지 못한 일부 구간의 이상점들을 RVHAR에서는 찾아냈으며 동일하게 찾아낸 이상점들에 대해서도 대부분 RVHAR이 더 큰 검정 통계량을 가지고 있어 RVHAR의 성능의 우수성을 뒷받침할 수 있다. 더 나아가 RVHAR 모형은 RVAR(22) 모형의 계수를 3개로 축약해 계산 비용을 줄이고 따라서 계산 속도 또한

RVAR(p), $p > 3$보다 빠르다는 장점을 가진다. 물론 제안한 방법론에 대한 개선의 여지도 확인할 수 있었다. 이상점의 종류가 TC인 경우에 d가 작거나, 이상점의 크기 ω가 작으면 벤치마킹 모형과의 차이가 없거나 벤치마킹 모형이 더 뛰어난 예도 있었다. 이러한 경우는 장기억성이 뚜렷하지 않고, 평균변화 모형과 장기억성의 구별이 어려운 경우에 기인한 것으로 보이지만, 이에 대한 정확한 원인 분석과 개선 방향에 관해서는 추가 연구가 필요할 것으로 보인다.

References

Andersen TG, Bollerslev T, Diebold FX, and Labys P (2003). Modeling and forecasting realized volatility. *Econometrica*, **71**, 579–625.

Baek C, Kechagias S, and Pipiras V (2017). Semiparametric, parametric, and possibly sparse models for multivariate long-range dependence. *Wavelets and Sparsity XVII*, Vol. 10394, International Society for Optics and Photonics 103941S.

Baek C and Park M (2021). Sparse vector heterogeneous autoregressive modeling for realized volatility. *Journal of the Korean Statistical Society*, **50**, 495–510.

Boubacar-Maïnassara Y, Esstafa Y, and Saussereau B (2021). Estimating farima models with uncorrelated but non-independent error terms. *Statistical Inference for Stochastic Processes*, **24**, 549–608.

Corsi F (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, **7**, 174–196.

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

Fox J and Weisberg S (2002). Robust regression: appendix to An R and S-PLUS companion to applied regression.

Hawkins D (1980). Identification of outliers

Kleiner B, Martin RD, and Thomson DJ (1979). Robust estimation of power spectra. *Journal of the Royal Statistical Society: Series B (Methodological)*, **41**, 313–338.

Lee J (2017). Detecting outlier with exponential smoothing in time series. *Seoul National University.*

Lee K and Baek C (2021). Outlier detection for long memory processes. *The Korean Data & Information Science Society*, **32**, 1205–1218.

Ripley B, Venables B, Bates DM, Hornik K, Gebhardt A, and Firth D (2013). R package 'MASS'

Tsay RS (1988). Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, **7**, 1–20.

Tsay RS, Pena D, and Pankratz AE (2000). Outliers in multivariate time series. *Biometrika*, **87**, 789–804.

Received March 23, 2022; Revised May 3, 2022; Accepted May 9, 2022

다변량 장기 종속 시계열에서의 이상점 탐지

김경희[a], 유승연[a], 백창룡[1,a]

[a]성균관대학교 통계학과

요 약

본 논문에서는 장기 종속 다변량 시계열 자료에 대한 이상점 탐지 기법을 연구한다. 기존 다변량 시계열 이상점 탐지 방법은 단기 종속 시계열 모형인 VARMA에 기반한 방법으로, 장기억성을 띈 다변량 시계열 자료에는 적합하지 않다. 자기회귀 모형을 통해서 장기 종속성, 즉 장기억성을 고려하기 위해서는 높은 차수의 모형이 필요하고, 이는 곧 추정의 불안성으로 이어지기에 장기억성을 효율적으로 다룰 수 없기 때문이다. 따라서, 본 논문은 이러한 문제를 보완하고자 VHAR 구조에 기반한 이상점 탐지 방법을 제시하고자 한다. 또한 더욱 정확한 추론을 위해서 로버스트한 방법을 이용하여 VHAR 계수를 추정하였고 이를 활용하여 이상점을 탐지하였다. 모의실험 결과 우리가 제안한 방법론이 기존 VARMA에 기반한 방법론보다 이상점 탐지에 더 효과적임을 살펴볼 수 있었다. 주가지수에 대한 실증자료 분석에서도 기존의 방법론은 탐지하지 못하는 추가 이상점을 찾음을 확인할 수 있었다.

주요용어: 이상점 탐지, 다변량 장기 기억 시계열, VHAR, 로버스트 회귀

[1]교신저자: (03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: crbaek@skku.edu