The Korean Journal of Applied Statistics
2023, Vol. 36, No. 6, 529–545

DOI: https://doi.org/10.5351/KJAS.2023.36.6.529
Print ISSN1225-066X / Online ISSN 2383-5818

Banded vector heterogeneous autoregression models

Sangtae Kim[a], Changryong Baek[1,a]

[a]Department of Statistics, Sungkyunkwan University

Abstract

This paper introduces the Banded-VHAR model suitable for high-dimensional long-memory time series with band structure. The Banded-VHAR model has nonignorable correlations only with adjacent dimensions due to data features, for example, geographical information. Row-wise estimation method is adapted for fast computation. Also, two estimation methods, namely BIC and ratio methods, are proposed to estimate the width of band. We demonstrate asymptotic consistency of our proposed estimation methods through simulation study. Real data applications to pm2.5 and apartment trading volume substantiate that our Banded-VHAR model outperforms traditional sparse VHAR model in forecasting and easy to interpret model coefficients.

Keywords: banded coefficient matrices, vector heterogeneous autoregressive model, BIC, high dimensional time series, long memory property

1. 서론

장기 기억성(long memory property)을 가지는 고차원 시계열이 경제, 금융, 사회과학적 현상 등 많은 분야에서 널리 관측이 되고 있다. 특히 이러한 장기 기억을 가지는 고차원 시계열 데이터들을 동시에 분석하고 상호간의 동시적인 움직임(co-movement)을 파악하는 것이 나날이 중요해지고 있다. 대표적으로 장기 기억성 특성을 가지는 금융 분야에서의 변동성(volatility)의 경우, 실제 여러 개의 변동성 데이터들끼리 상호 간의 공통된 특성을 가지는 것으로 자주 확인된고 있어 이를 반영하는 모형의 개발이 활발하게 이루어 지고 있다 (Ray와 Tsay, 2000; Engle과 Marcucci, 2006; Baek과 Park, 2021).

한편으로는 대다수 데이터가 생성되는 현장을 살펴보면, 특정 시계열 데이터를 분석하는 데 있어 해당 데이터와 밀접한 데이터만을 수집하고 활용하는 것만으로 충분한 경우가 많다. 예컨대 자연 과학 데이터 중 특정 날씨 환경 변수(기온, 습도, 풍속 등)에 영향을 받는 경우, 해당 데이터 주변의 데이터들에만 영향을 받고 비교적 먼 데이터의 값들에는 영향을 받지 않을 수 있다. 대기 오염자료는 지리적으로 가까운 지역에서는 비슷한 오염농도를 보이다가 먼 지역으로 갈수록 0에 가까운 오염농도를 보이곤 한다. 또한, 여러 업종의 주식 데이터들을 모아볼 때, 특정 주식 데이터는 해당 회사와 비슷한 업계의 주식에만 의존하고 서로 다른 업종의 주식 데이터의 영향은 무시 가능한 수준일 수도 있다. 따라서 이렇게 지리적(spatial) 혹은 물리적(physical)으로 인접한 차원에서의 관측값끼리 상관관계가 크고 멀리 떨어진 곳에서는 0에 가까운 자료들을 종종 접하게 된다.

이러한 지리적 물리적 특수한 상관관계를 보통 밴드 구조화(banded structure)하여, 정해진 밴드의 띠너비 안에 있으면 상관관계가 존재하고 벗어나게 되면 0이 되는 특징을 가지는 밴드구조 VAR (vector autoregressive

This work was supported by the Basic Science Research Program from the National Research Foundation of Korea (NRF-2022R1F1A1066209).

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

Published 31 December 2023 / journal homepage: http://kjas.or.kr

model) 고차원 시계열 모형이 처음 Guo 등 (2016)에 의해서 제안되었고 연구되었다. 추후 Gao 등 (2019)는 공간시계열 모형에 보다 적합한 모형으로 계승시켰고, Zheng과 Cheng (2020)은 일반적인 선형의 제약 구조를 가지는 VAR 모형의 이론적인 성질을 탐구하였다.

본 논문은 밴드구조 VAR 모형을 고차원 장기억 시계열 모형에 적합하도록 확장하고자 한다. 보다 구체적으로 고차원 장기억 시계열 분석에 효과적인 VHAR (vector heterogeneous autoregressive model) 모형의 구조화된 모형으로 밴드구조를 가지는 VHAR 모형(밴드구조 VHAR; Banded-VHAR)을 제안하고자 한다. 또한 밴드구조 제약조건 하에서의 모형 추정 및 밴드의 크기를 결정하는 방법을 제안하고 제안한 Banded-VHAR 모형이 다른 모형과 비교하여 어떤 장단점들을 가지는지 탐구하고자 한다.

본 논문의 구성은 다음과 같다. 2장 방법론에서는 Banded-VHAR 모형의 기본적인 구조와 계수행렬 추정 방법 및 밴드의 크기를 결정하는 띠너비 모수 추정 방법 설명한다. 3장 모의실험에서는 제안된 계수행렬 및 밴드 너비 모수 추정 방법들의 점근적 일치성을 보인다. 이를 바탕으로 4장 실증 분석에서는 전국 초미세먼지 데이터와 수도권 아파트 거래량 데이터에 각각 제안한 모형을 적용하여, 예측 성능을 비교하고 추정 계수에 대해 살펴본다. 마지막으로 5장 결론 및 논의점에서는 본 연구에 대한 결론과 추가적인 논의 사항에 대해 다룬다.

2. 방법론

2.1. Banded-VHAR 모형

Corsi (2009)에 의해 제안된 HAR 모형의 다변량 확장인 VHAR 모형은 다음과 같이 표현할 수 있다.

$$y_t^{(d)} = A^{(d)} y_{t-1}^{(d)} + A^{(w)} y_{t-1}^{(w)} + A^{(m)} y_{t-1}^{(m)} + \epsilon_t, \quad t = 23, \ldots, T. \tag{2.1}$$

이 때 $y_t^{(d)} = (y_{1,t}^{(d)}, \ldots, y_{p,t}^{(d)})'$은 p차원을 가진 시계열 데이터의 t시점 일별 관측값을, $y_t^{(w)} = (y_{1,t}^{(w)}, \ldots, y_{p,t}^{(w)})'$과 $y_t^{(m)} = (y_{1,t}^{(m)}, \ldots, y_{p,t}^{(m)})'$은 각각 다음의 식을 만족하는 일별 시계열 데이터의 주차별, 월별 평균 벡터이다.

$$y_{t-1}^{w} = \frac{1}{5}\sum_{j=1}^{5} y_{t-j}^{d}, \quad y_{t-1}^{m} = \frac{1}{22}\sum_{j=1}^{22} y_{t-j}^{d}.$$

이노베이션(innovation) $\boldsymbol{\epsilon_t} = (\epsilon_{1,t}, \ldots, \epsilon_{p,t})'$는

$$\mathbb{E}(\boldsymbol{\epsilon}_t) = 0, \quad \mathrm{Var}(\boldsymbol{\epsilon}_t) = \Sigma_\epsilon \quad \text{and} \quad \mathrm{Cov}\left(y_{t-j}, \boldsymbol{\epsilon}_t\right) = 0 \quad \text{for all } j \geq 0$$

을 만족한다.

본 연구에서 $\{y_t\}$는 약정상성(weakly stationarity)을 만족한다고 가정한다. 또한, $A^{(d)}, A^{(w)}, A^{(m)}$은 각각 일별, 주차별, 월별 관계를 설명하는 $p \times p$ 계수행렬로 행렬 원소(element)는 모두 다음의 밴드구조를 만족한다.

$$a_{ij}^{(d)} = a_{ij}^{(w)} = a_{ij}^{(m)} = 0, \quad |i - j| > k_0. \tag{2.2}$$

즉 주어진 차원에서 인접한 차원의 시계열 데이터에 대해서만 상관구조를 가지는 선형 모형으로 구조적인 특성을 이용해서 성근성(sparsity)을 가지는 모형이다. 띠너비 모수 k_0는 고정된 양의 정수 값을 가진다. 모형의 단순함을 위해 본 연구에서는 $A^{(d)}, A^{(w)}, A^{(m)}$ 모두 동일한 띠너비를 가진다고 가정한다. 따라서 고정된 띠너비에 의해 각 계수행렬의 행들이 가질 수 있는 0이 아닌 성분 개수는 최대 $(2k_0 + 1)$개로 정해진다. 하지만, Guo 등 (2016)에서와 같이 $\mathrm{Var}(\boldsymbol{\epsilon}_t) = \Sigma_\epsilon$은 띠 행렬 구조를 따른 필요가 없으며 그 결과 $y_t^{(d)}$의 자기 공분산행렬들도 띠 행렬 구조를 따를 필요는 없다.

2.2. Banded-VHAR 모형의 계수행렬 추정

본 장에서는 Banded-VHAR 모형의 추정 방법에 대해서 기술한다. 본질적으로 Baek과 Park (2021)에서 제안한 벡터화에 의한 추정 방법을 Banded-VHAR에서도 적용할 수 있다. 하지만, 벡터화에 의한 추정은 크로넥커 곱셈(Kronecker product;⊗)으로 인해서 행렬의 크기가 매우 커지는 단점이 있어 많은 양의 메모리와 계산 자원이 효율적이지 못하다. 이에 따라 계산 속도의 향상을 위해서 Guo 등 (2016)이 제안한 행별 추정 방법을 이용한 Banded-VHAR 모형의 추정도 소개한다.

2.2.1. 벡터화(vectorize)에 의한 추정 방법

VHAR 모형은

$$\mathbb{Y} = \mathbb{A}\mathbb{X} + \mathbb{Z}, \tag{2.3}$$

$\mathbb{Z} = (\epsilon_{23}, \ldots, \epsilon_T)$, $\mathbb{Y} = (y_{23}^{(d)}, \ldots, y_T^{(d)})$, $\mathbb{A} = (A^{(d)}, A^{(w)}, A^{(m)})$, $\mathbb{X} = (X_{22}, \ldots, X_{T-1})$,

$$X_t = \begin{pmatrix} y_t^{(d)} \\ y_t^{(w)} \\ y_t^{(m)} \end{pmatrix}$$

로 행렬을 이용하여 표현가능하다 (Baek과 Park, 2021). 이를 벡터화 하면, $\mathbf{Y} = \mathrm{vec}(\mathbb{Y})$, $\mathbf{X} = (\mathbb{X}' \otimes I_{p\times p})$, $\alpha = \mathrm{vec}(\mathbb{A})$ 그리고 $\mathbf{z} = \mathrm{vec}(\mathbf{Z})$에 대해서

$$\mathbf{Y} = \mathbf{X}\alpha + \mathbf{z} \tag{2.4}$$

로 쓸 수 있다. 띠너비 k_0에 대해서 0이 아닌 계수행렬을 벡터화한 γ에 대해서 $\alpha = R\gamma$로 표현이 가능하다. 예를 들어 3차원의 Banded-VHAR 모형에서 $k_0 = 1$인 경우 $A^{(d)}$의 0이 아닌 계수는

$$\begin{pmatrix} a_{11} \\ a_{21} \\ 0 \\ a_{12} \\ a_{22} \\ a_{32} \\ 0 \\ a_{23} \\ a_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{21} \\ a_{12} \\ a_{22} \\ a_{32} \\ a_{23} \\ a_{33} \end{pmatrix}$$

와 같이 나타낼 수 있으므로 Baek과 Park (2021)에 따라 α에 대한 일반화최소제곱(generalized least squares; GLS) 추정량은

$$\hat{\alpha} = R\left((\mathbf{X}R)'\left(I_{T-22}\otimes\hat{\Sigma}^{-1}\right)(\mathbf{X}R)\right)^{-1}(\mathbf{X}R)'\left(I_{T-22}\otimes\hat{\Sigma}^{-1}\right)\mathbf{Y}, \quad \hat{\Sigma} = \frac{1}{T-22}\left(\mathbb{Y}-\hat{\mathbb{A}}\mathbb{X}\right)\left(\mathbb{Y}-\hat{\mathbb{A}}\mathbb{X}\right)'$$

으로 $\hat{\mathbb{A}}$은 $\hat{\Sigma}$의 최소제곱(ordinary least squares; OLS) 추정량으로 $\hat{\Sigma} = I_p$일 경우 얻어지는 추정 행렬계수이다.

2.2.2. 행별(row-wise) 추정 방법

Guo 등 (2016)이 제안한 계수 추정 방법과 유사하게 수식 (2.1)의 우변을 행 별로 보면, 동일 띠너비(k_0) 가정과 수식 (2.2)에 의해 최대 $3\times(2k_0+1)$개의 0이 아닌 성분을 가진다. 이때 $\beta_i, i=1,\ldots,p$,를 $\mathbb{A}$ 행렬의 i 번째 행의 0이 아닌 성분들을 쌓아둔 열 벡터로 정의하고 τ_i를 β_i의 길이, 즉 i 번째 행의 0이 아닌 성분 개수라고 하자. 이때 τ_i값은 띠너비(k_0)에 대한 함수로

$$\tau_i := \tau_i(k_0) = \begin{cases} 3(2k_0+1), & i = k_0+1, k_0+2, \ldots, p-k_0, \\ 3(2k_0+1-j), & i = k_0+1-j \quad \text{or} \quad p-k_0+j, \quad j=1,\ldots,k_0. \end{cases} \tag{2.5}$$

으로 주어진다. 그러면 수식 (2.1)은 i 번째 행에 대해서 다음과 같이 재표현할 수 있다.

$$y_{i,t} = x_{i,t}'\beta_i + \epsilon_{i,t}, \quad i=1,\ldots,p. \tag{2.6}$$

$y_{i,t}$와 $\epsilon_{i,t}$는 각각 y_t와 ϵ_t의 i 번째 성분을 의미하고, $x_{i,t}$는 $y_{t-1}^{(d)}$, $y_{t-1}^{(w)}$, $y_{t-1}^{(m)}$의 값에 해당하는 $\tau_i\times 1$ 열 벡터이다. 따라서 위의 식(2.6)에 최소자승법을 적용해 구한 β_i의 추정량은

$$\hat{\beta}_i = (X_i'X_i)^{-1}X_i'y_i \tag{2.7}$$

으로 $y_i = (y_{i,23}^{(d)},\ldots,y_{i,T}^{(d)})'$, X_i는 $x_{i,22+j}'$를 j 번째 행으로 갖는 $(T-22)\times\tau_i$ 행렬이다. 따라서 모든 $i=1,\ldots,p$에 대해서 수식 (2.7)을 각각 적용하면, 행별 추정 방법에 기반한 식(2.1)의 계수 행렬들의 최소제곱추정량을 구할 수 있다. 위의 추정량을 토대로 구해지는 잔차의 제곱합(residual sum of squares; RSS)은 띠너비 모수(k_0)에 대한 함수로

$$\mathrm{RSS}_i := \mathrm{RSS}_i(k_0) = y_i'\left\{I_{T-22} - X_i\left(X_i'X_i\right)^{-1}X_i'\right\}y_i \tag{2.8}$$

이다.

2.3. 띠너비 모수(k_0) 결정

띠너비 모수(k_0)는 실제 값을 알 수 없으므로 추정해야 한다. 본 장에서는 베이지안정보기준(Bayesian information criterion; BIC)와 Lam과 Yao (2012) 및 Gao 등 (2019)가 제안한 비율에 기반한 방법 두 가지를 Banded-VHAR 모형에 적용한 방법을 제안한다.

2.3.1. BIC를 이용한 방법

먼저 Guo 등 (2016) 및 Wang 등 (2021)이 제안한 벌점함수에 대해서 Banded-VHAR 모형의 i-번째 차원에 대한 BIC 값은

$$\mathrm{BIC}_i(k) = \log\mathrm{RSS}_i(k) + \frac{\log\log T}{T}\tau_i(k)\log(p\vee T), \quad i=1,\ldots,p \tag{2.9}$$

으로 주어진다. 이때 $p\vee T = \max(p,T)$이고 $\tau_i(k)$는 식(2.5)로 정의된다. 이에 기반한 BIC를 활용한 최적의 띠너비는

$$\hat{k} = \max_{1\le i\le p}\left\{\operatorname*{argmin}_{1\le k\le K}\mathrm{BIC}_i(k)\right\} \tag{2.10}$$

로 주어진다. 여기서 K는 $K\ge k_0$를 만족하는 양의 정수이며, 사전에 정의되는 띠너비의 상한값으로 본 연구에서는 $K=[T^{1/2}]$로 두거나 $\mathrm{BIC}_i(k)$의 변화 곡선을 확인하여 상한을 정하였다.

2.3.2. 잔차 제곱합 비율(ratio)을 이용한 방법

두 번째 방법은 Lam과 Yao (2012)와 Gao 등 (2019)의 연구에서 제안한 잔차 제곱합 비율을 이용한 방법이다. 잔차 제곱합 비율은 참 값 k_0에 대해서

$$\frac{\mathrm{RSS}_i(k-1)}{\mathrm{RSS}_i(k)} = \begin{cases} \text{유한}, & 1 \le k < k_0, \\ \text{매우 큰 값}, & k = k_0, \\ 0/0, & k_0 < k < K \end{cases} \tag{2.11}$$

와 같은 성질을 가진다. 다만 $k_0 < k < K$일 때의 특이점 문제를 피하기 위해, 잔차 제곱합 비율에 $w_n = \log\log(T)$이라는 작은 상수 값을 추가하여 최적의 띠너비 추정량은

$$\hat{k} = \max_{1\le i\le p}\left\{\operatorname*{argmax}_{1\le k\le K} \frac{\mathrm{RSS}_i(k-1)+w_n}{\mathrm{RSS}_i(k)+w_n}\right\} \tag{2.12}$$

로 주어진다. 여기에서 $K \ge k_0$는 BIC와 같이 띠너비 상한값으로, 본 연구에서는 K를 $[T^{1/2}]$로 두거나 비율의 변화 곡선을 확인하여 정하였다.

3. 모의 실험

이번 장에서는 제안한 Banded-VHAR 모형의 유한 표본에서의 성능을 모의 실험을 통해 살펴본다. 정상성을 만족하는 Banded-VHAR 모형의 계수를 생성하기 위해서 다음의 알고리즘을 토대로 계수행렬을 생성하였다.

Banded-VHAR 프로세스는 제약 조건이 있는 22차수의 VAR(22)과 동일하므로, 정상성을 만족하는 VAR(22) 계수행렬들을 생성하여 이를 VHAR 계수행렬로 변환한다. 즉 Banded-VHAR 모형의 계수는

$$\begin{aligned} y_t^{(d)} &= \Phi_1 y_{t-1}^{(d)} + \Phi_2 y_{t-2}^{(d)} + \cdots + \Phi_{22} y_{t-22}^{(d)} + \epsilon_t, \\ \Phi_1 &= A^{(d)} + \frac{1}{5}A^{(w)} + \frac{1}{22}A^{(m)}, \quad \Phi_2 = \cdots = \Phi_5 = \frac{1}{5}A^{(w)} + \frac{1}{22}A^{(m)}, \quad \Phi_6 = \cdots = \Phi_{22} = \frac{1}{22}A^{(m)}. \end{aligned} \tag{3.1}$$

와 같이 나타낼 수 있다. 이 때 p는 자료의 차원이고 띠너비 모수(k)는 $1 \le k \le [(p-1)/2]$ 을 만족하는 양의 정수다. 또한 $p_d \in [0,1]$와 $p_o \in [0,1]$는 각각 계수행렬의 대각 성분과 비대각 성분의 크기를 통제하는 확률값이고, (l_d, u_d)와 (l_o, u_o)는 0과 1사이의 실수값으로 각각 계수행렬의 대각 성분과 비대각 성분값들을 수축(shrinkage)하기 위한 상수값을 추출하는 균등 분포(uniform distribution)의 하한, 상한값이다. 구체적인 알고리즘은 아래 Algorithm 1과 같다. 본 모의 실험에서는 이노베이션 ϵ_t은 독립 표준 정규 분포를 따른다고 가정하였다.

3.1. 계수행렬 추정 방법 비교

위의 방법론에서 소개한 두 가지 계수행렬 추정 방법, 벡터화를 이용한 방법 및 행별로 추정한 방법의 추정값을 비교한다. 또한 계산에 소요되는 시간을 비교하여 성능을 비교하고자 한다. 생성한 모형의 차원(p)은 10, 15, 30으로 데이터 크기(T)는 200, 500, 1000, 2000, 띠너비 모수(k_0)는 2를 사용하였다. 추정값의 성능은 각 세팅별로 100번의 반복을 통해서 각 방법으로 추정된 계수 행렬들의 프로베니우스 노름(Frobenius norm)을 계산하였으며 상자그림(box-plot)을 통하여 두 추정값에 대한 비교를 시각화하였다.

Figure 1은 벡터화 및 행별 추정 방법에 대한 비교를 상자그림으로 나타낸 그림이다. 먼저 모든 차원($p = 10, 15, 30$)에서 두 추정 방법을 통해 추정한 계수들 간의 차이가 없는 것을 확인할 수 있다. 특히 단순 상자그림만 보았을 때는 차원이 커짐에 따라 두 추정 방법간의 추정 계수들의 차이가 증가하는 것처럼 보이지만, y축의 범위를 보면 그 차이는 매우 미미한 것을 확인할 수 있다. 따라서 벡터화 및 행별 추정 방법 모두 비슷한 추정값을 줌을 알 수 있다.

Algorithm 1 : Coefficient matrices generating algorithm

Step 1. $\Phi_1 = [\phi_{ij}]_{p\times p}$, $i, j = 1, \ldots, p$ 생성

1. (Generating) 다음의 혼합 모형을 통해 계수 행렬의 대각 성분과 비대각 성분을 생성한다.
   (a) 대각 성분(ϕ_{ii}) :
   $\phi_{ii} \sim p_d \times (U[-0.2, -0.1] \cup U[0.1, 0.2]) + (1 - p_d) \times (U[-0.4, -0.3] \cup U[0.3, 0.4])$.
   (b) 비대각 성분(ϕ_{ij}, $i \neq j$) :
   $\phi_{ij} \sim p_o \times (U[-0.05, -0.01] \cup U[0.01, 0.05]) + (1 - p_o) \times (U[-0.25, -0.15] \cup U[0.15, 0.25])$
   $\phi_{ij} = 0 \quad \text{where} \quad |i - j| > k$.
2. (Threshold) 비대각 성분 중 임계값(0.05)보다 작은 값들은 정확히 0으로 설정한다.
   $\phi_{ij} = 0 \quad \text{if} \quad \phi_{ij} \leq 0.05$.
3. (Shrinkage) 대각 성분과 비대각 성분에 각각 수축 계수를 곱합으로써 값을 작게 만든다.
   (a) 대각 성분(ϕ_{ii}) : $\tilde{\phi}_{ii} = \phi_{ii} \times \eta_d \quad \eta_d, \sim U[l_d, u_d]$.
   (b) 비대각 성분(ϕ_{ij}, $i \neq j$) : $\tilde{\phi}_{ij} = \phi_{ij} \times \eta_o \quad \eta_o, \sim U[l_o, u_o]$.

Step 2. Step 1의 generating-threshold-shrinkage 과정을 사용하여 $\Phi_2 = \cdots = \Phi_5 = [\phi_{ij}]_{p\times p}$와 $\Phi_6 = \cdots = \Phi_{22} = [\phi_{ij}]_{p\times p}$, $i, j = 1, \ldots, p$를 생성한다.

Step 3. 정상성 검정 단계

특성 방정식(characteristic polynomial)의 고유값을 계산하여 VAR 프로세스의 정상성 검정을 한다 (Lütkepohl, 2005).

1. 만약 정상성을 만족하지 않는다면, Step 1으로 돌아가서 다시 생성한다.
2. 만약 정상성을 만족한다면, 다음 Step 4로 넘어간다.

Step 4. 생성된 정상 VAR 계수행렬들을 다음의 식을 통해 VHAR 계수행렬로 변환한다.

$$A^{(m)} = 22 \times \Phi_{22}, \quad A^{(w)} = 5 \times (\Phi_2 - \Phi_{22}), \quad A^{(d)} = \Phi_1 - \Phi_2.$$

두 방법의 결정적 차이는 계산 시간으로 Figure 2에 벡터화 및 행별 추정 방법에 추정 시간을 초(seconds)로 나타냈다. 행별 추정방법(row-wise) 방법이 벡터화(vectorized)에 비하여 추정 시간이 적음을 알 수 있다. 특히, 차원이 증가할 때 그리고 데이터가 증가함에 따라 두 방법의 소요 시간 차이는 현저하게 나타난다. 반면 차원이 작은 경우에는 데이터 크기가 커짐에 따라 소요 시간의 차이가 벌어지긴 하지만, 그 차이는 미미한 것을 확인할 수 있다. 이는 벡터화에 의한 추정 방법에서 적용되는 크로넥커 곱으로 인한 것으로, 소요 시간에 있어서 데이터의 차원보다는 자료의 크기가 계산 시간에 더 큰 영향을 끼침을 확인할 수 있다.

3.2. 추정량의 일치성

제안한 Banded-VHAR 모형의 일치성을 확인하기 위해서 데이터가 증가할 때 실제 계수행렬과 추정행렬의 프로베니우스 노름이 감소하는지 확인하고자 한다. 또한 모형의 주요 모수인 띠너비 모수(k_0)를 추정하기 위해 제안한 두 가지 방법에 대한 일치성도 살펴보고자 한다. 시계열 데이터의 차원(p)은 10, 15, 30, 50으로, 데이터 크기(T)는 200, 500, 1000, 2000으로 띠너비 모수(k_0)는 2를 사용하였다. 일치성을 검증하기 위해서 5000번의 반복을 통해서 프로베니우스 노름을 계산하였으며 앞 절에서 살펴보았듯이 행별 추정 방법이 계산의 속도가

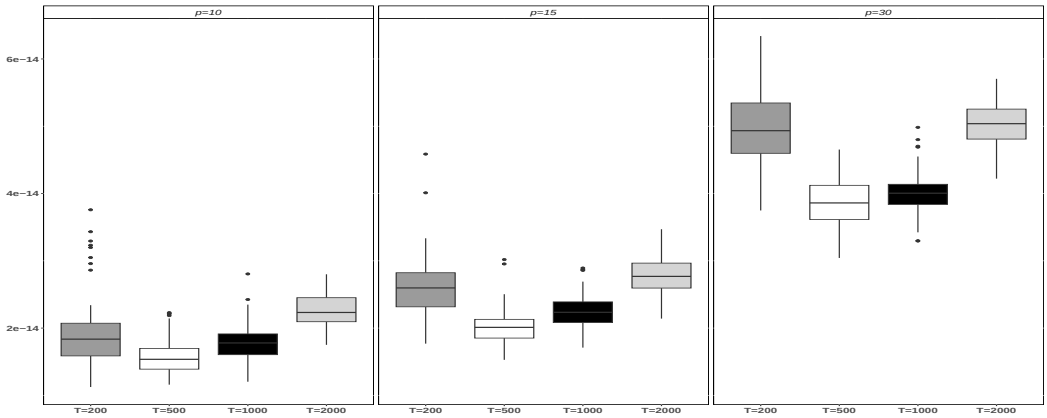

Figure 1: *Estimation errors (in Frobenius norm) of the coefficient matrices for row-wise and vectorized methods.*

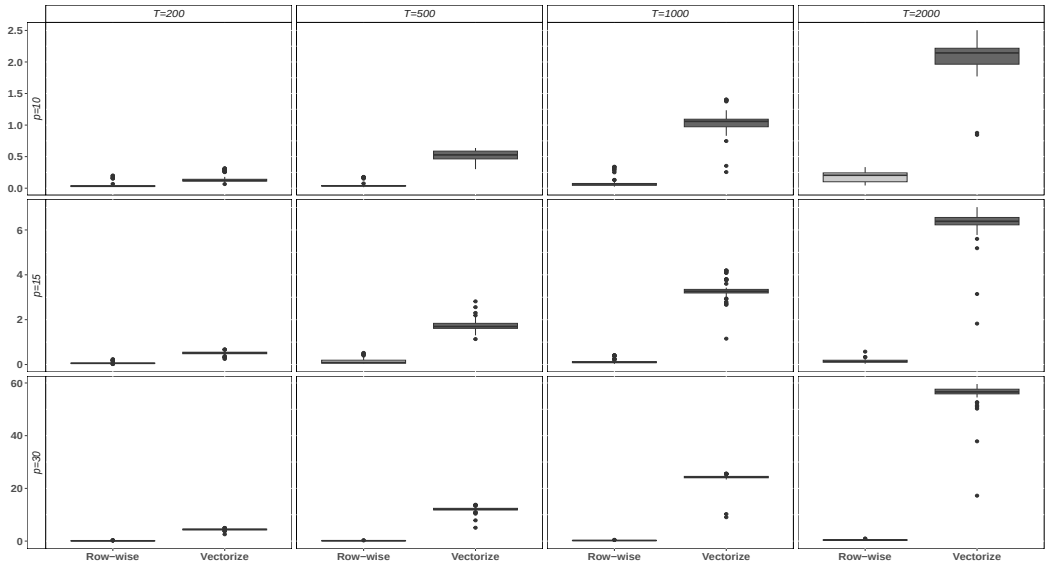

Figure 2: *The comparison of running time in seconds between row-wise and vectorized methods.*

빠르면서도 추정 성능은 동일하므로, 행별 추정 방법을 토대로 모의 실험을 진행하였다.

3.2.1. 계수행렬 추정의 일치성

Figure 3는 시계열 데이터의 차원과 데이터 크기 별로 일별, 주차별, 월별 계수행렬을 5000번 반복하여 추정했을 때, 추정된 계수행렬들과 실제 계수행렬들 간의 프로베니우스 노름값을 시각화 한 것이다. 상자그림을 보면 모든 차원($p = 10, 15, 30, 50$)에서 데이터 크기(T)가 증가함에 따라 오차가 감소하는 것을 확인할 수 있다. 이는 데이터가 증가함에 따라 차원에 상관없이 추정 계수행렬들이 실제 행렬들과 점근적으로 일치함을 의미한다. 차원이 큰 경우를 보면, 데이터의 차원이 작은 경우에 비해 데이터 크기가 작을 때의 추정 오차가 현저히 큰 것을 확인할 수 있다. 이는 시계열 데이터의 차원이 작은 경우 데이터 크기가 추정 계수행렬에 미치

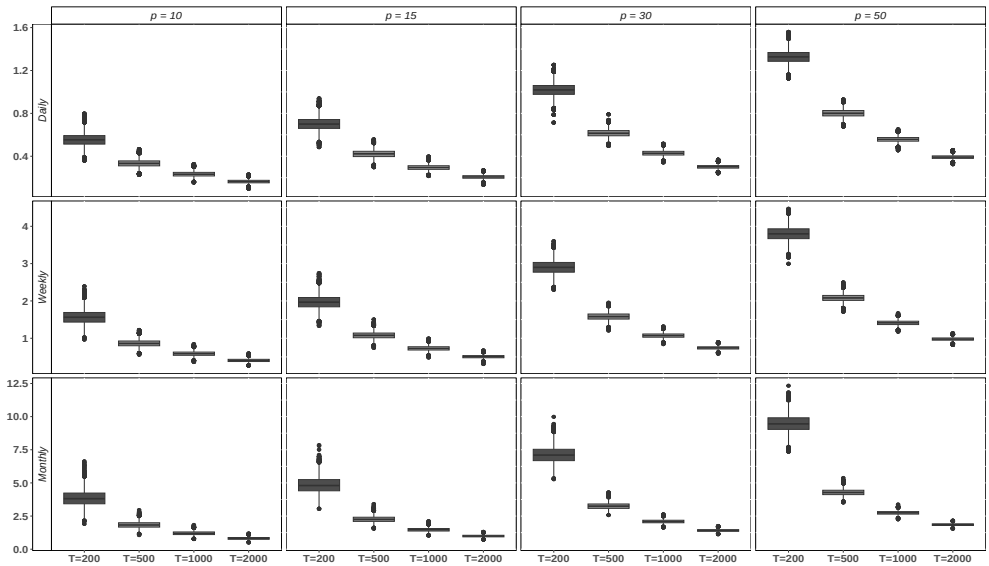

Figure 3: *Estimation errors in Frobenius norm for VHAR model coefficients.*

Table 1: Summary statistics of estimation errors, mean and standard deviation in parentheses, for VHAR model coefficients

Dimension	Sample size	Type		
		Daily	Weekly	Monthly
$p = 10$	200	0.5543(0.0594)	1.5690(0.1918)	3.8551(0.6198)
	500	0.3339(0.0349)	0.8631(0.0970)	1.8395(0.2445)
	1000	0.2331(0.0247)	0.5872(0.0650)	1.2063(0.1480)
	2000	0.1639(0.0176)	0.4063(0.0446)	0.8168(0.0962)
$p = 15$	200	0.7016(0.0610)	1.9719(0.1898)	4.8485(0.6329)
	500	0.4221(0.0363)	1.0792(0.0961)	2.2622(0.2393)
	1000	0.2950(0.0251)	0.7298(0.0654)	1.4669(0.1430)
	2000	0.2069(0.0175)	0.5063(0.0442)	0.993(0.0921)
$p = 30$	200	1.0192(0.0608)	2.9015(0.1927)	7.1281(0.6415)
	500	0.6151(0.0360)	1.5822(0.0993)	3.2529(0.2481)
	1000	0.4288(0.0252)	1.0722(0.0647)	2.0916(0.1425)
	2000	0.3011(0.0177)	0.7415(0.0445)	1.4142(0.0914)
$p = 50$	200	1.3269(0.0608)	3.8031(0.1947)	9.4851(0.6686)
	500	0.8016(0.0368)	2.0818(0.0993)	4.2857(0.2473)
	1000	0.5576(0.0251)	1.4117(0.0665)	2.7432(0.1414)
	2000	0.3919(0.0175)	0.9786(0.0446)	1.8509(0.0921)

는 영향력이 적지만, 차원이 커지는 경우 그 정도가 커지는 것을 의미한다. 더불어 데이터의 차원이나 크기에 상관없이 월별, 주차별, 일별 순으로 추정 오차가 높은 것을 확인할 수 있는데, 이는 VHAR 프로세스가 제약조건이 있는 VAR(22)와 동일하므로, VAR 모형의 관점에서 높은 차수의 계수 행렬의 추정 오차가 낮은 차수 추정 오차 대비 더 크다는 점에서 기인한다.

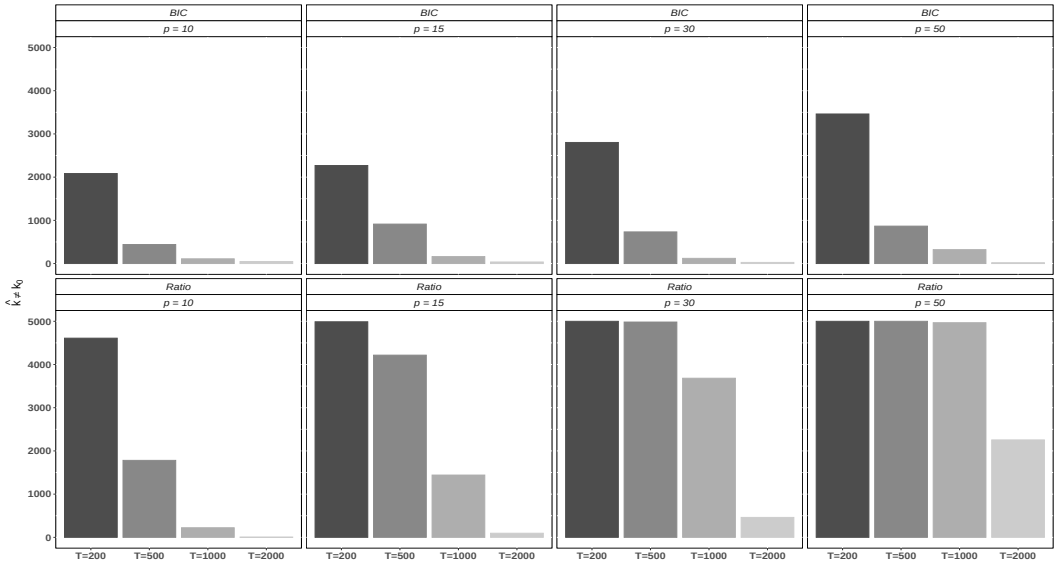

Figure 4: *The empirical frequency of misspecified bandwidth parameter k_0 for BIC and ratio methods.*

Table 2: Empirical frequency table for the estimation of bandwidth parameter k_0 for BIC and ratio methods

Dimension	**Sample size**	**Relative frequency**					
		$\hat{k} = k_0$		$\hat{k} > k_0$		$\hat{k} < k_0$	
		BIC	**Ratio**	**BIC**	**Ratio**	**BIC**	**Ratio**
$p = 10$	200	0.5832	0.0780	0.3364	0.9220	0.0804	0.0000
	500	0.9118	0.6436	0.0490	0.3546	0.0392	0.0018
	1000	0.9780	0.9556	0.0214	0.0428	0.0006	0.0016
	2000	0.9906	0.9988	0.0094	0.0004	0.0000	0.0008
$p = 15$	200	0.5460	0.0022	0.3816	0.9978	0.0724	0.0000
	500	0.8172	0.1566	0.0350	0.8434	0.1478	0.0000
	1000	0.9678	0.7116	0.0140	0.2884	0.0182	0.0000
	2000	0.9926	0.9808	0.0074	0.0192	0.0000	0.0000
$p = 30$	200	0.4394	0.0000	0.5440	1.0000	0.0166	0.0000
	500	0.8534	0.0034	0.0350	0.9966	0.1116	0.0000
	1000	0.9758	0.2634	0.0076	0.7366	0.0166	0.0000
	2000	0.9946	0.9076	0.0054	0.0924	0.0000	0.0000
$p = 50$	200	0.3076	0.0000	0.6898	1.0000	0.0026	0.0000
	500	0.8266	0.0000	0.0318	1.0000	0.1416	0.0000
	1000	0.9356	0.0066	0.0082	0.9934	0.0562	0.0000
	2000	0.9960	0.5490	0.0040	0.4510	0.0000	0.0000

Table 1은 Figure 3와 동일한 모의 실험에 대한 요약 통계량으로 프로베니우스 노름의 평균과 표준 편차를 기록한 표이다. 그림상과 마찬가지로 차원이나 일별, 주차별, 월별 계수 행렬의 상관없이 데이터 크기가 증가함에 따라 오차의 평균 및 표준편차가 감소함을 살펴 볼 수 있어 일치성을 확인할 수 있다.

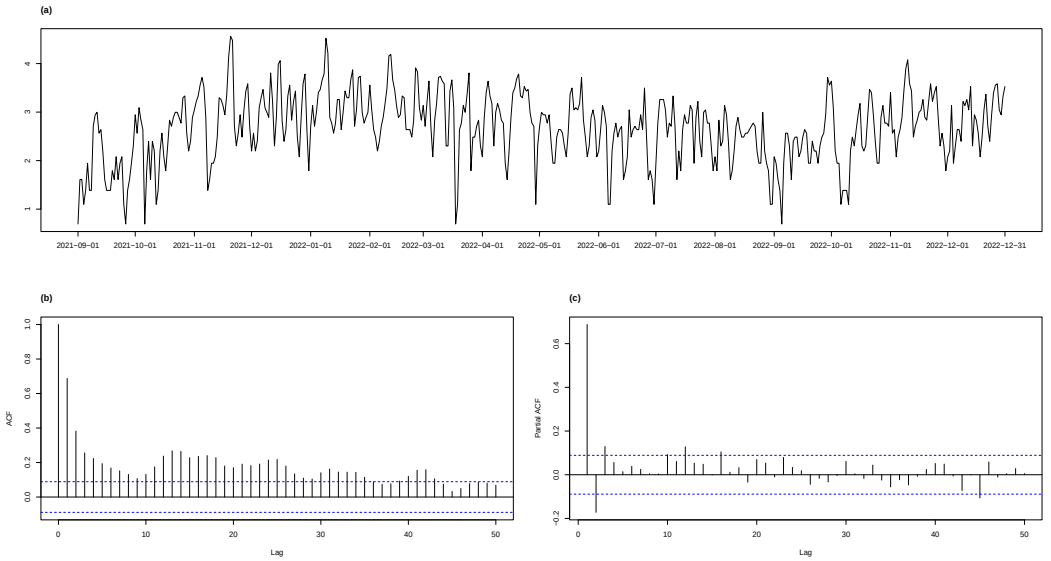

Figure 5: *The time plot (top), sample actocorrelation plot (bottom left) and sample partial autocorrelation plot (bottom right) for log-transformed daily pm2.5.*

3.2.2. 띠너비 모수 추정의 일치성

Figure 4는 시계열 데이터의 차원과 데이터 크기 별로 띠너비 모수($k_0 = 2$)를 앞선 방법론에서 제안한 BIC를 이용한 방법과 잔차 제곱합 비율을 이용한 방법으로 5000번 반복하여 추정($\hat{k}$)했을 때, 각 추정 방법 별로 $\{\hat{k} \neq k_0\}$에 대한 빈도를 시각화한 것이다. 차원이 작은 경우 데이터 크기에 상관없이 두 추정 방법 모두 실제 띠너비 모수 값과 동일하게 추정한다. 하지만 차원이 증가함에 따라, 데이터 크기가 작은 경우 모수 추정에 있어 오차 정도가 증가하는 것을 확인할 수 있다. 특히 고차원의 경우 잔차 제곱합 비율을 이용한 방법으로 띠너비 모수를 추정하기 위해서는 상당히 많은 양의 데이터 크기가 필요하다. 반면, BIC를 이용한 방법의 경우 일정 수준의 데이터 크기($T \geq 1000$) 이상에서는 차원에 상관없이 모수 추정의 오차가 작다. 그럼에도 불구하고 두 방법 모두 차원에 상관없이 데이터 크기가 증가함에 따라 모수 추정의 오차가 작아지는 것을 보이며, 이는 띠너비 모수 추정의 점근적 일치성을 내포한다.

이는 상대 빈도로 나타낸 Table 2를 통해 보다 객관적으로 확인할 수 있다. 차원이 고정되어 있을 경우 데이터 크기가 증가할수록 올바르게 추정한 빈도가 증가함을 관찰할 수 있다. 다만, 차원이 높아질수록 일치성을 위해서는 더 많은 자료가 필요함을 볼 수 있다. 특히 잔차 제곱합 비율을 이용한 추정 방법은 고차원에서의 성능이 BIC보다는 떨어지며, 실제 띠너비 모수보다 더 큰 띠너비 값($\hat{k} > k_0$)으로 과대 추정하는 경향이 있음을 살펴볼 수 있다. 따라서 BIC를 이용한 방법과 잔차 제곱합 비율을 이용한 방법 모두 점근적 일치성은 가지고 있지만, 본 실험에서 사용한 모형의 경우 BIC를 이용한 방법의 수렴 속도가 잔차 제곱합 비율을 이용한 방법보다 더 빠름을 알 수 있다.

4. 실증 분석

Banded-VHAR 모형을 지역에 따른 초미세먼지 및 아파트 거래량 자료에 적용하여 예측력을 비교하여 제안한 모형의 유용성에 대해서 살펴보았다.

Table 3: Out-of-sample forecasting errors of the estimated Banded-VHAR models for pm2.5 according to ordering directions

Ordering	Estimation method	$\hat{k}$	Non-zero coef	One-step ahead	Two-step ahead
Northwest to Southeast	BIC	4	399	0.1462	0.1911
	Ratio	6	537	0.1462	0.2109
Southwest to Northeast	BIC	2	237	**0.1396**	**0.1803**
	Ratio	6	537	0.1453	0.2112
North to South	BIC	5	471	0.1467	0.2086
	Ratio	7	597	0.1454	0.2055
West to East	BIC	5	471	0.1466	0.2306
	Ratio	5	471	0.1466	0.2306
VHAR	-	-	867	0.1538	0.2520
sVHAR	-	-	198	0.1407	0.2023
Seasonal ARMA	-	-	-	0.1576	0.1897

4.1. 초미세먼지(pm2.5)

사용한 자료는 2021년 9월 1일부터 2022년 12월 31일까지의 전국 17개의 광역단체 별로 관측한 일 별 초미세먼지(pm2.5) 데이터($p = 17$, $T = 487$)이다. 본 자료는 한국환경공단의 전국 실시간 대기오염도 공개 홈페이지 에어코리아(`https://www.airkorea.or.kr/web/`)에서 확인 가능하다. 일 별 초미세먼지 데이터의 경우 봄과 겨울에 초미세먼지 양이 증가하고, 여름과 가을에 감소하는 경향이 있다. 이에 본 분석에서는 17차원의 초미세먼지 자료가 정상성을 만족하기 위해 로그 변환을 진행하여 $\{y_t;\ t = 1, \ldots, 487\}$로 정의하였다. 로그 변환한 데이터에서 각 차원 별로 정상성을 만족하는지 검정하기 위해 ADF 검정(augmented Dickey–Fuller test)을 진행하였고, 모든 차원에서 정상성을 만족하였다.

Figure 5는 서울 특별시에 대한 시계열 데이터 플롯과 50 차수에 대한 표본 자기상관 함수(sample auto-correlation function; SACF)와 표본 편자기상관 함수(sample partial auto-correlation function; SPACF) 플롯이다. 특히 SACF 플롯에서 비교적 높은 차수에서도 상관성을 가지는 것을 보았을 때 장기 기억성 특성을 확인할 수 있어 VHAR 모형을 사용함이 타당해 보인다. 또한 초미세먼지의 경우 지리적으로 가까울수록 비슷한 농도를 보일 경우가 많으므로 Guo 등 (2016), Gao 등 (2019), Wang 등 (2021)와 유사하게 지리적 위치를 기반으로 17개 광역 단체에 임의의 순서를 정하였다. 본 분석에서는 지리적 정보를 토대로 북-남, 서-동, 북서-남동, 남서-북동 네 가지 방향을 기준으로 광역 단체들을 정렬한다. 각 방향 별로 BIC 이용한 방법과 잔차 제곱합 비율을 이용한 방법을 적용하여 각각 최적의 띠너비 모수를 추정한 후 모형을 적합한다. 그 후 마지막 30일에 대한 1-step, 2-step 표본외예측 오차를 구하여 모형간의 성능을 비교하였다. 1-step 표본외 예측값은 1부터 t까지 사용한 자료에 대해서 추정한 계수행렬 $\hat{A}^{(d)}, \hat{A}^{(w)}, \hat{A}^{(m)}$에 대해서

$$\hat{y}_{t+1}^{(d)} = \hat{A}^{(d)} y_t^{(d)} + \hat{A}^{(w)} y_t^{(w)} + \hat{A}^{(m)} y_t^{(m)}$$

으로 주어진다. 이를 활용하여, 즉 예측값 $\hat{y}_{t+1}^{(d)}$을 사용하여 $\hat{y}_{t+1}^{(w)}$와 $\hat{y}_{t+1}^{(m)}$을 구한 뒤, 2-step 표본외 예측값은

$$\hat{y}_{t+2}^{(d)} = \hat{A}^{(d)} \hat{y}_{t+1}^{(d)} + \hat{A}^{(w)} \hat{y}_{t+1}^{(w)} + \hat{A}^{(m)} \hat{y}_{t+1}^{(m)}$$

으로 계산할 수 있다. 따라서 마지막 30일에 대한 h-step ($h = 1, 2$) 표본외예측 평균제곱오차(mean squared prediction error; MSPE)는

$$\mathrm{MSPE}^{(h)} = \frac{1}{(30 - h + 1) \times p} \sum_{t=T-30}^{T-h} \left(y_{t+h}^{(d)} - \hat{y}_{t+h}^{(d)}\right)' \left(y_{t+h}^{(d)} - \hat{y}_{t+h}^{(d)}\right) \tag{4.1}$$

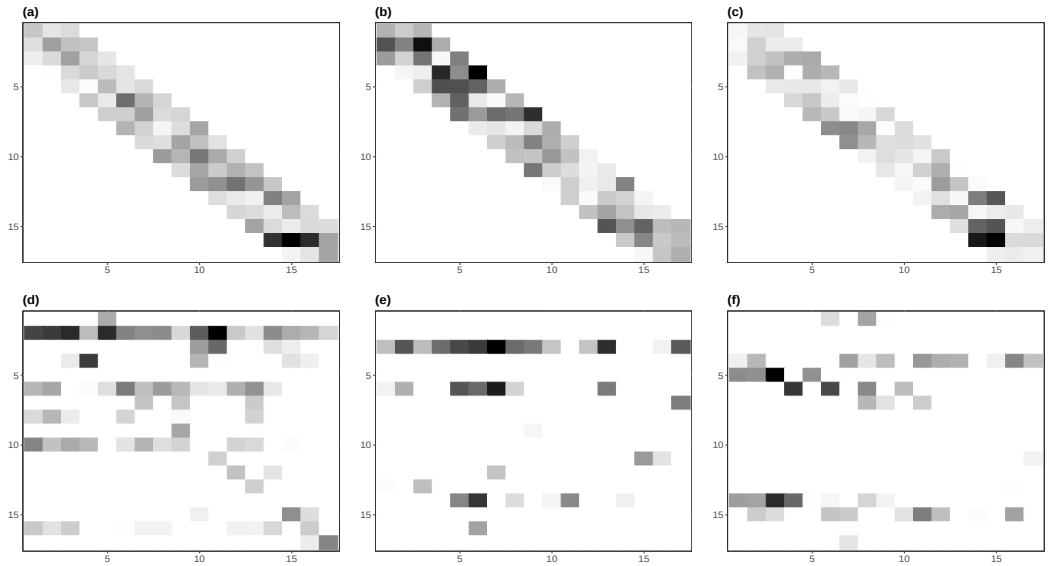

Figure 6: *Estimated daily, weekly and monthly coefficient matrices for Banded-VHAR coefficients (top) with southwest to northeast ordering; estimated daily, weekly and monthly coefficient matrices for sVHAR mode (bottom). Dark shade represents larger absolute value.*

으로 주어진다.

이 때 Banded-VHAR 모형 이외에도 Cubadda 등 (2017)에서의 일반적인 VHAR 모형과 Baek과 Park (2021)에서의 adaptive lasso를 이용한 sVHAR (sparse VHAR) 모형도 더불어 비교하였다. 또한 초미세먼지 데이터가 계절성(seasonality) 특징을 가지고 있으므로, 각 차원 별로 개별 seasonal ARMA (autoregressive moving-average) 모형도 적합하여 결과를 비교하였다. 이때 각 차원 별로 적용한 모형의 차수는 Hyndman과 Khandakar (2008)에서의 auto.arima 함수를 사용하여 결정하였다.

표본외예측 오차를 정리한 Table 3을 살펴 보면 일반적인 VHAR 모형보다는 Banded-VHAR 모형 및 sVHAR 모형의 예측 오차가 작음을 볼 수 있다. 즉 성근 모형을 적합함으로써 예측 오차를 줄일 수 있음을 내포한다. 또한 각 차원 별 적합한 seasonal ARMA 모형보다 모든 VHAR 모형의 예측 오차가 작은 것을 볼 수 있는데, 이는 각 지역 별 초미세먼지의 동시적인 움직임을 고려하는 것이 모형의 예측 오차를 줄일 수 있음을 의미한다. sVHAR 모형과 Banded-VHAR 모형을 비교하여 보면 두 모형 모두 비슷한 예측 오차를 주었으나 Banded-VHAR 모형의 방향에 따라 약간의 차이가 남을 볼 수 있다. 남서-북동쪽 방향으로 BIC를 통해서 추정한 모형이 1-step, 2-step 모두 가장 작은 예측 오차를 주었다. 하지만, Banded-VHAR 모형의 다른 큰 장점은 모형의 해석이 용이하다는 점이다. 예를 들어 Figure 6은 남서-북동쪽으로 정렬한 Banded-VHAR 모형의 계수 및 sVHAR 모형의 추정 계수들을 표현한 그림이다. 두 모형 모두 성근 모형을 지향하지만, 지리적인 정보를 이용한 해석에 있어서는 Banded-VHAR 모형이 훨씬 더 용이함을 살펴볼 수 있다.

4.2. 아파트 거래량

두 번째로 사용한 실증 자료는 국토교통부 실거래가 공개시스템(`https://rt.molit.go.kr/`)에서 제공하는 2006년 1월부터 2022년 12월까지의 수도권 76개의 시군구 별로 주차 별 아파트 거래량 데이터($p = 76, T = 901$) 자료이다. 주차 별 아파트 거래량 데이터의 경우 거래가 활발한 지역은 1년 52주 동안 고르게 거래가

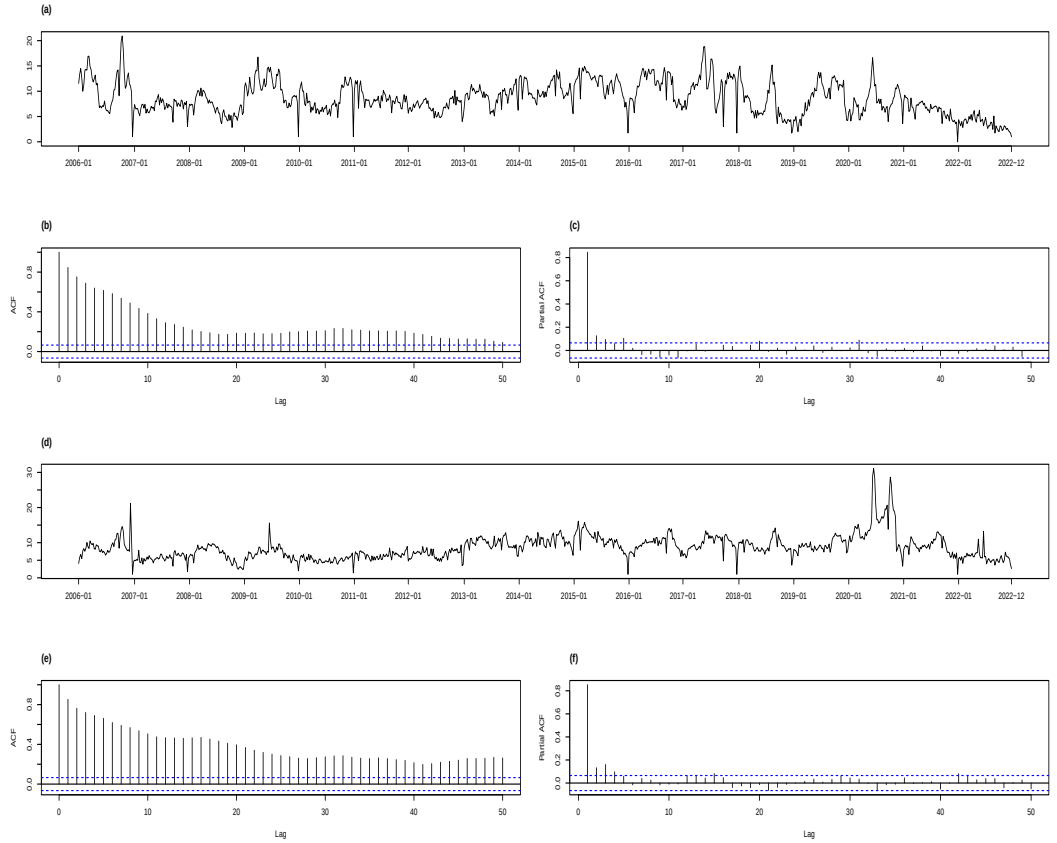

Figure 7: *Time, sample ACF and sample PACF plots for square-root transformed weekly apartment trading volume for Seoul Gangnam-gu ((a)~(c)) and Gyeonggi Gimpo ((d)~(f)).*

발생하지만, 비교적 비도시 시군구의 경우 드물게 거래가 발생하지 않는 경우가 있다. 따라서 본 분석에서는 정상성 만족과 더불어 거래량이 0건인 주가 존재한다는 점에서, 로그변환 대신 각 시계열 데이터에 제곱근 변환을 진행한 뒤 모형을 적합하였다. 제곱근 변환된 데이터에 대해서 초미세먼지 데이터와 마찬가지로 ADF 검정을 통해 모든 차원에서 정상성을 만족하는 것을 확인하였다. Figure 7는 서울특별시, 인천광역시, 경기도의 특정 시군구들의 시계열 데이터 플롯과 50차수의 SACF 플롯, SPACF 플롯이다. 장기억시계열의 대표적인 특징들이 뚜렷하여 VHAR 모형을 적합하는 것이 타당함을 확인할 수 있다.

앞선 초미세먼지 데이터 분석과 마찬가지로 아파트 거래량도 지리적 정보를 토대로 네 가지 방향으로 정렬한 후, 각 방향 별로 BIC를 이용한 방법과 잔차 제곱합 비율을 이용한 방법을 토대로 띠너비를 추정하고 모형을 적합하였다. 모형의 성능은 마지막 30일에 대한 1-step, 2-step 표본외예측 오차를 계산하였다. 이때 벡터화에 기반한 sVAR 모형의 추정은 본 자료의 차원처럼 큰 경우($p = 76$)에 매우 느려 Cavalcante 등 (2017)에서와 같이 행별 손실함수 추정방법을 사용하였다. 보다 구체적으로 Baek과 Park (2021)에서는 다음의 L_1-함수에 기반한 손실함수를 이용한다.

$$\hat{\alpha}^L = \underset{\alpha}{\operatorname{argmin}} \ \|\mathbf{Y} - \mathbf{X}\alpha\|^2 + \lambda \|\alpha\|_1. \tag{4.2}$$

Table 4: Out-of-sample forecasting errors of the estimated Banded-VHAR models for apartment trading volum e according to ordering directions

Ordering	Estimation method	$\hat{k}$	Non-zero coef	One-step ahead	Two-step ahead
Northwest to Southeast	BIC	4	1992	1.2263	1.8909
	Ratio	35	12408	2.1012	4.1601
Southwest to Northeast	BIC	8	3660	1.3773	2.2486
	Ratio	33	11910	2.1945	4.2463
North to South	BIC	6	2838	1.2819	1.9143
	Ratio	30	11118	2.1144	4.3916
West to East	BIC	4	1992	**1.1343**	**1.7585**
	Ratio	34	12162	2.0258	3.7409
VHAR	-	-	17328	2.3010	4.9141
sVHAR	-	-	1553	1.2859	1.9254

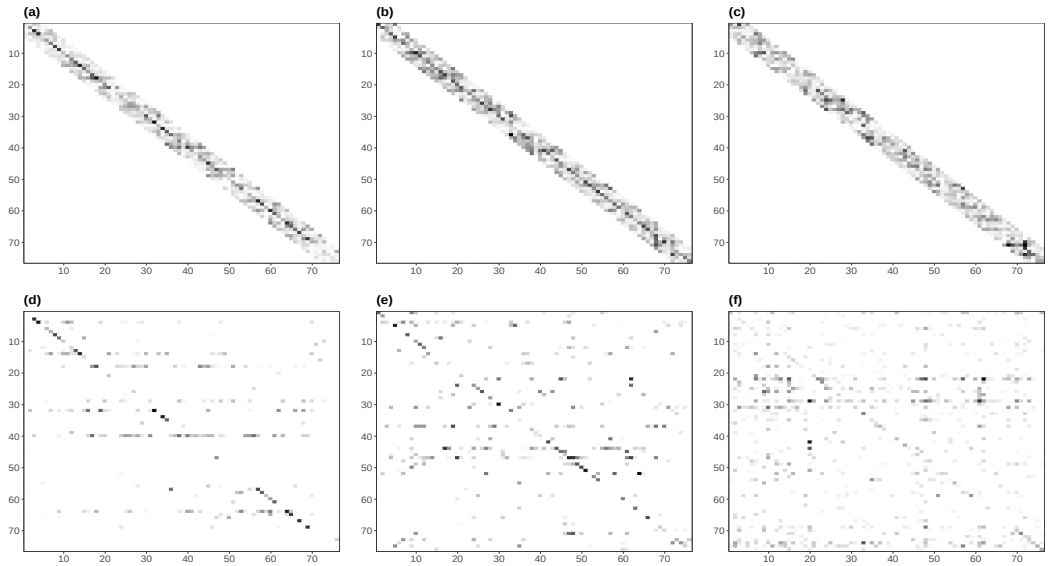

Figure 8: *Estimated daily, weekly and monthly coefficient matrices of Banded-VHAR coefficients (top) for apartment trading volume with west to east ordering; estimated daily, weekly and monthly coefficient matrices of sVHAR model (bottom). Dark shade represents larger absolute value.*

하지만 차원이 큰 경우에는 다음의 행별 손실 함수를 활용하는 것이 각 차원 별로 병렬화를 통해 추정이 가능하므로 컴퓨팅 계산 속도면에서 빠르다.

$$\hat{\alpha}_i^{row} = \underset{\alpha_i}{\operatorname{argmin}} \ \|\mathbf{Y} - \mathbf{X}\alpha_i\|^2 + \lambda \|\alpha_i\|_1, \quad i = 1, \ldots, p. \tag{4.3}$$

따라서 본 아파트 거래량 데이터의 경우 고차원 데이터라는 점에서 행별 손실 함수를 사용한 sVHAR 모형과, 일반적인 VHAR 모형, 네 가지 방향으로 정렬한 Badned-VHAR 모형을 비교하였다.

표본외예측 오차를 정리한 Table 4을 보면 앞선 결과와 비슷하게 일반적인 VHAR 모형보다는 Banded-

VHAR 모형과 sVHAR 모형의 성능이 뛰어나다. Banded-VHAR 모형만 보았을 때, 아파트 거래량 데이터가 차원이 크다는 점에서 모의 실험에서의 결과와 유사하게 잔차 제곱합 비율을 이용한 방법의 경우 BIC를 이용한 방법에 비해 띠너비 모수를 상당히 크게 추정하고 예측 오차 역시 sVHAR 모형보다 현저히 큰 것을 볼 수 있다. 반면 BIC를 이용한 방법으로 띠너비 모수를 추정한 경우 대다수의 방향에서 sVHAR 모형보다 뛰어난 성능을 가진다. 0이 아닌 계수의 개수를 보면 초미세먼지 데이터와 마찬가지로 sVHAR 모형이 가장 작지만, 예측 오차가 가장 작은 Banded-VHAR 모형(서–동, $\hat{k} = 4$)과 비슷하다. 하지만 Figure 8에서와 같이 Banded-VHAR 모형의 해석이 지리적인 특징을 이용하므로 sVHAR 모형보다 용이하다. 즉 sVHAR 모형의 경우 계수들이 무질서하게 퍼져 있어 지리적인 특징을 한눈에 파악하지 힘들다. 반면에, Banded-VHAR 모형의 경우 특정 시군구의 아파트 거래량은 인접한 시군구의 아파트 거래량에만 영향을 받는다는 식으로 해석이 가능하다는 점이 큰 장점으로 볼 수 있다.

5. 결론 및 논의점

본 연구에서 계수행렬이 밴드구조를 갖는 Banded-VHAR 모형을 제안하였다. 추정 방법에 대해서는 빠른 계산을 위해서 벡터화에 의한 추정 방법이 아닌 행별 추정방법을 제안하였다. 또 모형의 성근성을 결정하는 띠너비 모수 추정의 경우 BIC를 이용한 방법과 잔차 제곱합 비율을 이용한 방법을 제안하였다. 모의 실험 결과를 통해 계수행렬 추정 방법과 띠너비 모수의 추정 방법이 점근적 일치성을 가짐을 살펴볼 수 있었다. 제안한 모형의 실증례로 비교적 차원이 작은 초미세먼지 데이터와 고차원 상황인 아파트 거래량 데이터를 분석하여, 각각의 데이터에서 모두 Banded-VHAR 모형이 기존의 VHAR 모형들 보다 표본외예측 성능이 뛰어남을 보였다. 특히, 고차원에서의 띠너비 구조의 계수 행렬은 기존의 모형들보다 훨씬 뛰어난 해석을 제공하는 큰 장점이 있다는 것을 확인하였다. 이는 띠너비 모수만큼의 인접한 시계열 데이터에만 추정 계수만큼 영향을 받는다는 식으로 해석이 가능하기 때문에, 차원간의 계수들이 무질서하게 흩어져 있는 기존의 성근 모형과 큰 차이점을 가지고 있다. 하지만, Banded-VHAR 모형의 경우 차원을 정렬하는 순서에 따라 약간의 모형 성능 차이가 발생할 수 있는 한계점이 존재한다. 최적의 방향을 찾기 위한 방법에 대해서는 향후 연구를 통해 보완하고자 한다.

References

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

Cavalcante L, Bessa RJ, Reis M, and Browell J (2017). Lasso vector autoregression structures for very short-term wind power forecasting, *Wind Energy*, **20**, 657–675.

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

Cubadda G, Guardabascio B, and Hecq A (2017). A vector heterogeneous autoregressive index model for realized volatility measures, *International Journal of Forecasting*, **33**, 337–344.

Engle RF and Marcucci J (2006). A long-run pure variance common features model for the common volatilities of the Dow Jones, *Journal of Econometrics*, **132**, 7–42.

Gao Z, Ma Y, Wang H, and Yao Q (2019). Banded spatio-temporal autoregressions, *Journal of Econometrics*, **208**, 211–230.

Guo S, Wang Y, and Yao Q (2016). High-dimensional and banded vector autoregressions, *Biometrika*, **103**, 889–903.

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

Lam C and Yao Q (2012). Factor modeling for high-dimensional time series: Inference for the number of factors, *The Annals of Statistics*, **40**, 694–726.

Lütkepohl H (2005). *New Introduction to Multiple Time Series Analysis*, Springer Science & Business Media, Berlin.

Ray BK and Tsay RS (2000). Long-range dependence in daily stock volatilities, *Journal of Business & Economic Statistics*, **18**, 254–262.

Wang H, Luo X, and Ling L (2021). Semiparametric spatio-temporal models with unknown and banded autoregressive coefficient matrices, *Mathematical Methods in the Applied Sciences*, **30 December 2021**, 1–31.

Zheng Y and Cheng G (2020). Finite-time analysis of vector autoregressive models under linear restrictions, *Biometrika*, **108**, 469–489.

Received April 11, 2023; Revised May 23, 2023; Accepted June 21, 2023

밴드구조 VHAR 모형

김상태[a], 백창룡[1,a]

[a]성균관대학교 통계학과

요 약

본 논문에서는 장기 기억성을 가지는 고차원 시계열 데이터 분석에 유용한, 밴드 구조의 계수행렬들을 가지는 밴드구조 VHAR (Banded-VHAR) 모형을 제안한다. 밴드구조 VHAR 모형은 인접한 차원의 시계열에서만 상관구조를 가지는 성근 고차원 시계열 모형으로 밴드구조에 영향을 주는 요인으로는 대표적으로 지리적 특성이 있다. 밴드구조 VHAR 모형의 빠른 추정을 위해 본 논문은 행별추정방법을 사용하고 또 밴드의 크기를 추정하기 위해 BIC와 잔차제곱합의 비율을 이용한 추정 방법을 소개하였다. 더불어 모의 실험을 통해서 제안한 추정 방법의 점근적 일치성을 확인하였다. 실증자료 분석으로 지역별 초미세먼지 및 아파트 거래량 자료를 활용하여 모형을 적용한 결과 밴드구조 VHAR 모형이 표본외예측 능력의 우수하고, 지리적 정보에 기반하여 모형의 해석이 용이하다는 큰 장점이 있음을 살펴보았다.

주요용어: banded coefficients, vector heterogeneous autoregressive model, BIC, high dimensional time series, long memory

이 논문은 한국연구재단의 지원을 받아 수행된 기초연구 사업임 (NRF-2022R1F1A1066209).

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