Journal of the Korean Data & Information Science Society
2025, **36**(5), 1–13

http://dx.doi.org/10.7465/jkdi.2025.36.5.1
한국데이터정보과학회지

그래프 기반 신경망을 이용한 고차원 시계열 자료의 군집 탐지[†]

송민정[1] · 백창룡[2]

[12]성균관대학교 통계학과

접수 2025년 5월 29일, 수정 2025년 7월 11일, 게재확정 2025년 7월 22일

요 약

고차원 시계열 자료의 군집 분석에서 Guðmundsson과 Brownlees (2021)의 스펙트럴 클러스터링 기반 VAR-Blockbuster 알고리즘은 이론적 정당성과 계산 효율성으로 인해 널리 사용되고 있다. 그러나 해당 방법은 SB-VAR (Stochastic block-vector autoregression) 모형에 의존하므로 적용 범위에 제약이 존재한다. 본 연구는 이러한 제한을 보완하고자 모형 가정에 덜 민감한 딥러닝 기반 그래프 신경망인 Adaptive graph encoder (AGE)를 활용하여 고차원 시계열 군집 탐지에 대한 대안을 제시한다. 모의실험을 통해 표본 크기가 충분한 경우에는 VAR-Blockbuster가 안정적인 성능을 보였으나 표본 수가 제한적인 환경에서는 AGE가 상대적으로 우수한 군집 탐지 성능을 나타냈다. 또한, AGE 기반 방법을 S&P 500 지수 내 금융 및 소비재 섹터 종목 그리고 서울시 자치구별 주택 매매가격지수에 적용한 결과 구조적으로 해석 가능한 군집을 도출할 수 있음을 실증적으로 확인하였다.

주요용어: 고차원 시계열, 군집 탐지, SB-VAR, VAR-블록버스터, AGE.

1. 서론

시계열 데이터를 활용한 군집 탐지가 다양한 학문 분야에서 많은 관심을 받고 있다. 예를 들어, 환경 분야에서 Kim 등 (2021)은 전국 PM10 농도 자료 간의 시공간 상관관계를 기반으로 지역별 군집 구조를 도출하여 대기질 관리의 정책적 시사점을 제시하였다. 교통 분야에서는 Park과 Na (2022)가 광역철도역별 수송량 시계열에 동적 시간 워핑(Dynamic Time Warping; DTW) 알고리즘을 적용하여 고전적인 거리 기반 군집법보다 높은 응집도와 분리도를 확보한 군집 구조를 도출하였다. 한편, 에너지 분야에서는 Park과 Yoon (2017)과 Kim과 Kim (2021)이 전력 수요의 예측 가능성과 효율적 공급 관리를 위해 시계열 기반의 전력 수요량 군집분석을 수행하였다.

금융 분야에서도 군집 기반 분석은 고차원 시계열 내에서 구조적 관계를 도출하고자 하는 시도와 함께 활발히 연구되고 있다. Yoon과 Kim (2025)은 국내 주가의 비선형적 특성과 다양한 요인을 효과적으로 반영하기 위해 주식, 경제 지표, 뉴스 등의 데이터를 통합한 다이나믹 그래프 기반 주가 예측 모델을 설계하였다. Lee와 Baek (2023)은 암호화폐 시장의 시간에 따라 변화하는 변동성과 연결성 구조를 분석하기 위해 희박 VHAR-MGARCH 모형을 적용하여 코로나-19 시기 전후의 구조적 변화를 규명하였다. Baek과 Park (2021)은 주식시장의 실현 변동성 (Realized volatility; RV) 예측을 위해 희박 VHAR 모

[†] 이 성과는 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2025-00519717).

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

형을 활용하였으며 이 모형은 전 세계의 다양한 금융시장의 변동성 간 연결성을 파악하는 데 효과적임을 보였다.

이러한 중요성을 바탕으로, 최근 딥러닝 모델의 발달과 함께 군집 탐지를 위한 딥러닝 모델 연구가 활발히 진행되고 있다. 예를 들어, Tian 등 (2014)은 오토인코더 (Autoencoder; AE)의 손실 함수에 희소성 제약을 추가하여 군집 구조를 탐지하는 희소 그래프 인코더 (Sparse graph encoder; SGE) 모델을 제안하였다. 적대적 생성 네트워크 (Generative adversarial network; GAN)를 기반으로 Tao 등 (2019)은 적대적 정규화를 통해 딥 그래프 임베딩을 개선하며 성능 향상을 도모하였고, Yang 등 (2020)은 임베딩과 구조, 속성 정보를 판별 대상으로 삼아 보다 정교한 임베딩 학습을 수행하도록 하였다. 그래프 합성곱 네트워크 (Graph convolutional network; GCN)를 기반으론 Jin 등 (2019)은 JGE-CD (Joint GCN embedding for community detection)라는 비지도 군집 탐지 모델을 통해 성능을 향상시켰다. 이처럼 기존의 여러 딥러닝 모델에서 발전한 군집 탐지 모델이 제안되고 있다.

하지만, 현존하는 고차원 시계열에 기반한 탐지 방법들은 소위 탐색적인 방법들로 군집을 이루는 시계열 모형이나 방법론에 대한 이론적인 근거가 부족하였다. 최근 Guðmundsson과 Brownlees (2021)는 확률적 블록 모형 (Stochastic block model; SBM) 기반의 다변량 자기회귀 (Stochastic block-vector autoregression; SB-VAR) 모형을 소개하고 이에 기반하여 생성된 VAR 모형에 스펙트럴 클러스터링을 적용하는 VAR-블록버스터 (VAR-blockbuster) 알고리즘을 제안하였다. 또한 이 방법론에 대한 이론적인 결과들이 기존의 네트워크 자료의 군집분석과 같은 이론적인 성능임을 밝혔다. 이러한 중요한 이론적인 결과에도 불구하고 SB-VAR 모형에 의존하기 때문에 본 연구에서는 모형의 가정에서 조금은 더 자유로운 딥러닝에 기반한 방법이 VAR-블록버스터 알고리즘과 비교하여 견줄만한 방법론인지 살펴보고자 한다. 특히, 탐구한 딥러닝 방법 중에서 Cui 등 (2020)이 제안한 적응형 그래프 인코더 (Adaptive graph encoder; AGE)의 성능이 VAR-blockbuster와 견줄만하여 본 논문을 통해서 두 방법에 대해서 비교하고자 한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 Guðmundsson과 Brownlees (2021)이 제안한 SB-VAR 모형 및 VAR-블록버스터를 활용한 군집 탐지 방법을 소개하고 AGE에 기반한 군집 분석 방법이 SB-VAR 모형에 어떻게 적용될 수 있는지 소개한다. 제 3장에서는 SB-VAR 모형에 따라 생성된 모의실험 자료를 통해 두 방법의 성능을 비교한다. 제 4장에서는 S&P 500 종목에 해당하는 주가 데이터와 서울시 자치구별 매매가격지수 데이터에 군집 탐지 방법을 적용하여 그 유용성에 대해서 살펴보았다. 마지막으로 제 5장에서는 본 연구의 결론과 이에 관한 토의를 다룬다.

2. 방법론

본 장에서는 다변량 시계열의 군집을 탐지하기 위해 Guðmundsson과 Brownlees (2021)가 제안한 VAR-블록버스터 방법론과 Cui 등 (2020)이 제안한 그래프 기반 신경망 모델을 소개한다.

2.1. VAR-블록버스터에 기반한 군집 탐지

Guðmundsson과 Brownlees (2021)가 제안한 SB-VAR 모형은 시계열 데이터의 군집 구조를 반영하는 고차원 모형이다. 본 모형의 자기회귀 계수는 SBM에서 생성된 인접행렬을 기반으로 정의된다. SBM은 랜덤 그래프의 일종으로, 고정된 노드 집합 하에서 노드 간 엣지가 베르누이 시행에 따라 무작위적이고 독립적으로 형성되는 그래프를 생성한다. 이때 각 노드은 사전에 정의한 k개의 군집 중 하나에 속하며, 노드 간 엣지 확률은 그들이 속한 군집에 따라 결정된다. 따라서, 동일 군집 내 엣지 밀도가 이질 군집 간보다 상대적으로 높아지는 특성을 유도할 수 있어 군집 구조를 갖는 그래프를 생성하는 데 적합하며 군집 구조를 활용하여 미래 시점에 대한 예측을 가능하게 한다.

n차원의 다변량 시계열 $Y_t = (Y_{t,1}, Y_{t,2}, ..., Y_{t,n})^T$에 대해 SB-VAR($p$) 모형은

$$Y_t = \Phi_1 Y_{t-1} + \cdots + \Phi_p Y_{t-p} + \varepsilon_t \tag{2.1}$$

으로 정의된다. 여기서 Φ_l은 시차 l에 대응하는 $n \times n$ 자기회귀 계수 행렬이며, ε_t는 평균 $\mathbf{0} = (0, 0, ..., 0)^T$과 공분산 행렬 $\Omega = E[\varepsilon_t \varepsilon_t']$를 가지는 정상 확률 과정(stationary stochastic process)이다. SBM을 통해 생성된 k개의 군집을 가지는 그래프 $G_l \sim \text{SBM}(Z, B)$이 주어져 있다고 가정하자. 여기서 Z는 각 노드가 할당된 군집을 나타내며, B는 군집 간 연결 확률을 나타낸다. SB-VAR 모형은 네트워크 구조를 계수를 통해 반영하는 모형으로 SBM으로부터 생성된 그래프 G_l의 인접행렬 A_l을 기반으로

$$\Phi_l := \phi_l D_l^{-1/2} A_l' D_l^{-1/2} \tag{2.2}$$

로 정의된다. 여기에서 D_l 은 네트워크의 유입/유출 차수의 합으로 $[D_l]_{ii} = \sum_{j=1}^{n} [A_l]_{ij} + [A_l]_{ji}$ 이다.

잠재된 군집 구조를 탐지하기 위해 Guðmundsson과 Brownlees (2021)가 제안한 VAR-블록버스터 알고리즘은 다음과 같다.

1. 먼저, $X_t = \left(Y_{t-1}', ..., Y_{t-p}'\right)'$로 구성된 np 차원의 벡터를 이용하여 최소자승법(Ordinary Least Squares; OLS)을 통해 자기회귀 계수를 다음과 같이 추정한다.

$$[\hat{\Phi}_1 \ \hat{\Phi}_2 \ \cdots \ \hat{\Phi}_l] = \left(\sum_{t=p}^{T} X_t X_t'\right)^{-1} \left(\sum_{t=p}^{T} X_t Y_t'\right). \tag{2.3}$$

2. 추정된 계수 행렬을 대칭화하여 합산한 $\hat{\Phi}^S = \sum_{l=1}^{p} (\hat{\Phi}_l + \hat{\Phi}_l')$를 계산한다.
3. $\hat{\Phi}^S$의 상위 k개의 고유값에 대응하는 고유벡터들을 열로 갖는 $n \times k$ 차원의 행렬 $\hat{\boldsymbol{U}}$를 구한다. 각 행 벡터 $\hat{\boldsymbol{U}}_{i\bullet}$의 유클리드 노름의 역수를 대각 원소로 갖는 대각 행렬 $\hat{\boldsymbol{N}}$를 정의하여 정규화 행렬 $\hat{\boldsymbol{X}} = \hat{\boldsymbol{N}}\hat{\boldsymbol{U}}$를 계산한다.
4. $\hat{\boldsymbol{X}}$의 각 행에 대해 k-means 알고리즘으로 클러스터링을 수행하여 노드별 군집을 결정한다.

즉, VAR-블록버스터는 대칭화된 자기회귀 계수 행렬에 스펙트럴 클러스터링을 적용하여 시계열 간 군집 구조를 결정한다.

2.2. AGE 방법에 기반한 군집 탐지

Cui 등 (2020)이 제안한 적응형 그래프 인코더(Adaptive Graph Encoder; AGE)는 노드의 특성과 그래프 구조가 결합된 속성 그래프의 임베딩을 위한 비지도 학습 모델이다. 인코더를 통해 학습된 임베딩 벡터 간의 유사도는 그래프 구조 정보와 노드의 속성 정보를 함께 반영하므로 더 나은 그래프 구조 표현이 가능하다. 이러한 점을 고려하여, AGE는 적응형 인코더 구조로 설계되었으며 그래프 구조 학습을 위한 적응형 학습 방식을 채택하였다. 노드 특성 행렬 X는 인코더 f를 통해

$$Z = f(X; W) = X \cdot W \tag{2.4}$$

에 따라 저차원의 임베딩 벡터로 변환된다. 식 (2.4)에서 W는 인코더의 학습 가능한 가중치 행렬을 의미한다. 노드 간의 유사도를 측정하기 위해 코사인 유사도를 이용하여

$$S = \frac{ZZ^T}{||Z||_2^2} \tag{2.5}$$

와 같이 유사도 행렬 S를 구성한다.

인코더 학습에 사용할 샘플을 생성하기 위해 먼저 노드 특성 행렬 X로부터 유사도 행렬 S를 구성한다. 이후, 사전에 설정된 양성 및 음성 임계값을 바탕으로 유사도가 높은 노드 쌍은 양성 샘플로, 유사도가 낮은 노드 쌍은 음성 샘플로 간주하여 각각 라벨 1과 0을 할당한다. 이렇게 라벨이 부여된 샘플만을 활용하여 인코더를 학습하며, 이는 이진 분류 형태의 지도 학습 문제로 구성된다. 단, 양성 및 음성 샘플 간의 수적 불균형을 완화하기 위해 각 미니배치에서는 두 샘플이 동일한 크기로 포함되도록 구성한다. 구체적으로, 양성 샘플은 미리 정의된 배치 크기 bs만큼 순차적으로 선택하고 음성 샘플은 동일한 크기로 무작위로 추출하여 $2 \times bs$개의 균형있는 훈련 샘플 O를 구성한다. 이렇게 구성된 훈련 샘플에 대해 인코더를 통해 구한 노드 임베딩 벡터 간 유사도를 계산하고, 해당 유사도가 주어진 라벨에 근접하도록 인코더를 학습시킨다. 손실함수는 크로스-엔트로피 손실함수인

$$\mathcal{L} = \sum_{(v_i, v_j) \in O} -\ell_{ij} \log(s_{ij}) - (1 - \ell_{ij}) \log(1 - s_{ij}) \tag{2.6}$$

이며 이를 최소화하는 방향으로 학습이 이루어진다. 여기서 ℓ_{ij}는 이진 라벨, s_{ij}는 임베딩 벡터로 계산된 유사도를 의미한다.

또한, 사전에 설정한 임계값 갱신 횟수(upd)를 기준으로 주기적으로 임계값을 갱신하고 모델 성능을 평가한다. 즉, 전체 에폭($epoch$)을 갱신 횟수(upd)로 나누어 임계값 갱신이 이루어질 에폭(E_{upd})을 설정한다. 따라서, 특정 에폭마다 업데이트되는 양성 및 음성 임계값은

$$r'_{pos} = r_{pos} + \frac{r^{ed}_{pos} - r^{st}_{pos}}{upd} \tag{2.7}$$

$$r'_{neg} = r_{neg} + \frac{r^{ed}_{neg} - r^{st}_{neg}}{upd} \tag{2.8}$$

로 계산된다. 해당 에폭에 도달할 때마다, 현재까지 학습된 인코더를 기반으로 전체 노드 특성에 대한 임베딩 벡터를 생성하고, 이를 이용해 유사도 행렬을 계산한 뒤 k-means 클러스터링을 수행한다. 이렇게 구해진 모든 클러스터링 결과에 대해

$$\text{DBI} = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \tag{2.9}$$

와 같이 데이비드-볼딘 지수(Davies-Bouldin Index; DBI)를 계산하여 DBI가 가장 낮은 결과를 최종 클러스터링 결과로 채택한다. 여기서 K는 클러스터 수,

$$S_i = \frac{1}{|C_i|} \sum_{x \in C_i} \|x - c_i\|$$

는 클러스터 C_i의 중심 c_i로부터의 평균 거리로 정의되며, 클러스터의 내부 응집도를 나타낸다. $M_{ij} = \|c_i - c_j\|$는 클러스터 C_i와 C_j의 중심 간 거리로, 클러스터 간의 분리도를 의미한다. DBI는 각 클러스터 i에 대해 분리가 가장 잘되지 않은 클러스터 j를 찾아 그 유사도 $(S_i + S_j)/M_{ij}$를 계산하고, 이를 모든 클러스터에 대해 평균하여 전체 클러스터링 품질을 평가한다. DBI 값이 낮을수록 클러스터 간의 분리가 잘 되고, 내부 응집도가 높은 것으로 해석된다. 이후 양성 및 음성 임계값을 갱신한다. 이는 초기 학습 단계에서는 r_{pos}를 높게 설정하여 보다 많은 노드 쌍을 양성으로 간주함으로써 인코더가 전체적인 군집구조를 학습하도록 유도하고, 학습이 진행됨에 따라 r_{pos}를 점진적으로 감소시켜 신뢰도가 높은 양성 샘플에 집중함으로써 보다 정교한 구조 학습이 가능하게 한다. 이와 같은 과정을 통해 AGE는 탐색적이고 정밀한 군집 구분을 달성할 수 있으며, AGE의 학습 알고리즘은 다음과 같다:

1. 다변량 시계열 $\{Y_1, \ldots, Y_T\}$로부터 유사도 행렬 S를 계산한다.
2. S 의 원소 중 유사도 값이 높은 상위 r_{pos}인 쌍을 양성 샘플(라벨 1), 낮은 하위 r_{neg}인 쌍을 음성 샘플(라벨 0)로 정의한다.
3. 각 에폭(*epoch*)마다 다음을 반복한다:
   (a) 사전 정의된 배치 크기 bs만큼 양성 샘플을 순차적으로 추출하고, 동일한 크기만큼 음성 샘플을 무작위로 추출하여 총 $2 \times bs$개의 학습 샘플 O를 구성한다.
   (b) 임베딩 벡터로부터 계산된 유사도를 입력으로, 이진 라벨을 출력으로 사용하는 이진 분류 손실함수를 기반으로 인코더를 학습한다.
   (c) 설정된 임계값 갱신 주기 E_{upd}에 도달한 경우:
      i. 현재 임베딩 벡터를 기반으로 유사도 행렬을 다시 계산한 후, k-means 알고리즘을 적용하여 클러스터링을 수행한다.
      ii. 양성 및 음성 샘플의 임계값 r_{pos}, r_{neg}을 갱신하고, 이에 따라 새로운 이진 라벨을 할당한다.
4. 모든 클러스터링 결과 중 DBI가 가장 낮은 결과를 최종 클러스터링으로 선택한다.

3. 모의실험

본 장에서는 모의실험을 통해 VAR-블록버스터와 AGE의 군집 탐지 성능을 비교한다. 모의실험에 쓰인 자료 생성 과정(Data Generating Process; DGP)은 Guðmundsson과 Brownlees (2021)의 연구를 바탕으로 구성하였다. 그래프는 두 개의 군집을 가지는 SBM을 통해 군집 내 블록 확률이 0.5, 군집 간 블록 확률이 0.01인 구조를 갖도록 생성되었으며, 이를 통해 얻은 인접행렬을 기반으로

$$Y_t = \phi D^{-1/2} A' D^{-1/2} Y_{t-1} + \boldsymbol{\varepsilon_t} \tag{3.1}$$

와 같은 SB-VAR(1) 모형을 사용하였다. 여기서 ϕ는 0.99로 설정하였으며, ε_t는 평균이 $\mathbf{0} = (0, 0, ..., 0)^T$이고 공분산행렬이 $\Omega = I_n$인 n차원 다변량 정규분포

$$\boldsymbol{\varepsilon_t} \sim \mathcal{N}(\mathbf{0}, I_n) \tag{3.2}$$

로부터 독립적으로 추출하였다. 모든 실험은 100번의 반복 하에 수행하였으며 각 평가지표는 이들 결과의 평균값으로 계산되었다. 표본의 차원 $n \in \{30, 50, 100, 150\}$과 시계열 표본 크기 $T \in \{100, 150, 200, 300, 500, 1000\}$를 고려하여 다양한 조합에 대해 수행되었다.

두 방법론의 군집 탐지 성능을 비교하는 지표로는 정확도(Accuracy)와 F_1 스코어, 정규화된 상호정보량(Normalized Mutual Information; NMI)을 사용하였다. 정확도는 전체 노드(N) 중 올바른 군집(TP & TN)으로 분류된 노드의 비율을 의미한다. F_1 스코어는 정밀도(Precision)와 재현율(Recall)의 조화 평균으로, 두 지표 간의 균형을 측정하는 지표이다. 정밀도는 특정 군집으로 예측된 노드 중 실제 해당 군집에 속하는 비율을 의미하고, 재현율은 실제 해당 군집에 속하는 노드 중 모델이 정확히 해당 군집으로 예측한 비율을 의미한다. NMI는 예측된 군집 결과와 실제 군집 간의 유사도를 측정하는 클러스터링 평가 지표이다. 실제 군집 분포와 예측된 군집 분포 간의 상호정보량 $I(A; B) = H(A) + H(B) - H(A, B)$을 각 군집의 엔트로피 $H(A)$, $H(B)$의 기하평균으로 정규화하여 계산된다. 여기서 상호정보량은 값이 클수록 실제 군집과 예측된 군집이 공유하는 정보량을 의미하며,

값이 클수록 두 군집 간 일치도가 높음을 나타낸다. 정규화 과정을 통해 군집 수나 불균형한 군집 구조의 영향을 최소화할 수 있으며, 결과적으로 군집 간 일치 정도를 0에서 1 사이의 값으로 해석할 수 있다. 값이 1에 가까울수록 예측된 군집이 실제 클래스와 잘 일치함을 의미한다.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{FP}}{N}, \quad F_1 = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \quad \mathrm{NMI} = \frac{I(A;B)}{\sqrt{H(A) \cdot H(B)}}.$$

AGE 학습 시 필요한 초매개변수의 최적화 과정은 다음과 같다. 먼저, 모든 초매개변수 조합에 대해 최적화를 진행하는 것은 많은 시간이 요구되므로, 에폭 수($epoch$)와 후반 양성 임계값(r_{pos}^{ed}), 임베딩 차원 수(dim)를 중심으로 조합을 구성하여 최적화를 수행하였다. 각 조합에 대한 설정은 Table 3.1에 요약되어 있다. 다른 초매개변수 값으로 배치 크기(bs)를 8로 고정하였으며, 이에 따라 균형 있는 훈련 샘플 O에서 양성 및 음성 샘플의 크기는 각각 최대 8개씩, 총 16개의 노드 쌍을 기준으로 학습이 진행되었다. 양성 및 음성 샘플 라벨링을 위한 임계값 갱신(upd)은 총 5회 수행되었으며, 초기 양성 임계값(r_{pos}^{st})은 0.1, 음성 임계값(r_{neg}^{st})은 0.3으로 설정하였다. 후기 음성 임계값(r_{neg}^{ed})을 0.5로 설정하여 학습이 진행됨에 따라 양성 샘플 임계값을 감소시키고, 음성 샘플 임계값을 증가시켜 엄격한 기준을 적용하도록 설정하였다. 최적화 방법은 미니 배치(mini-batch) 학습 방식과 기울기 소실 문제를 완화할 수 있는 Adam(Adaptive Moment Estimation) 최적화 알고리즘을 사용하였다. 학습률(learning rate)은 0.01로 설정하였다. 전체 시계열 길이의 80%를 노드 특성 행렬로 사용한 격자 탐색을 통해 최적의 초매개변수 조합을 선택하였다. 이는 부록 Table A.1에 정리하였다. 이후 선택된 최적화된 초매개변수를 바탕으로 전체 데이터를 통한 학습 및 성능 평가를 실시하였다.

Table 3.2는 두 모델에 대한 군집 탐지 성능에 대한 요약이다. 두 모델 모두 전반적으로 높은 정확도와 F_1 스코어를 유지하였으나, 차원 수(n)와 시계열 길이(T)에 따라 성능에서 뚜렷한 차이를 보였다. 먼저, 차원이 낮고 시계열 길이가 길어질수록 두 모델 모두 성능이 향상되는 경향을 보여 두 방법 모두 일치성을 가지고 있음을 확인할 수 있었다. 하지만, $n = 30$에서 $T = 150 \sim 500$인 구간과, $n = 50$에서 $T = 200$일 경우 VAR-블록버스터가 AGE보다 대부분의 지표에서 더 우수한 성능을 보였다. 반면, $n = 100$, $T = 100 \sim 200$이나 $n = 150$, $T = 100 \sim 300$일 경우 AGE가 VAR-블록버스터에 비해 모든 지표에서 더 우수한 성능을 보였다. 예를 들어, $n = 100$, $T = 150$에서는 AGE의 ACC가 98.58%, NMI가 92.46였으며, 이는 VAR-블록버스터의 55.18%, 1.35와 비교하여 현저히 높은 수치이다. 즉 차원에 비하여 표본의 크기가 충분할 경우에는 VAR-블록버스터의 성능이 우수했고, 표본의 크기에 비해서 차원이 높은 경우에는 AGE 방법론이 더 좋은 성능을 보였다. 한편, 시간 측면에서는 AGE가 딥러닝 기반 구조로 인해 전반적으로 더 많은 연산 시간을 소요하였다. 이는 AGE가 초매개변수 최적화를 포함하는 복잡한 구조를 가지고 있기 때문이다. 결과적으로, AGE는 특히 고차원 및 단기 시계열에서 성능 우위를 보였으며, VAR-블록버스터는 저차원 및 상대적으로 긴 시계열 구간에서 계산 효율성과 높은 정확도를 모두 확보하는 데 강점을 보였다. 따라서, 데이터의 특성에 따라 두 모델을 선택적으로 활용하는 전략이 필요함을 알 수 있었다.

Table 3.1 Hyperparameter values used in grid search

Hyperparameter	**Candidates**
$epoch$	30, 35, 40, 45, 50
r_{ed}^{pos}	0.05, 0.03, 0.01
dim	32, 64, 128

Table 3.2 Summary of performance measures

		VARblockbuster				AGE			
n	**T**	ACC(%)	F_1(%)	NMI($\times 10^2$)	Time	ACC(%)	F_1(%)	NMI($\times 10^2$)	Time
30	100	96.83	96.82	85.21	1.38	98.08	98.06	92.52	158.03
30	150	99.60	99.60	97.93	1.27	99.16	99.16	96.04	133.05
30	200	99.90	99.90	99.46	1.49	99.49	99.49	97.49	235.19
30	300	100.00	100.00	100.00	3.21	99.76	99.76	98.96	283.65
30	400	100.00	100.00	100.00	3.74	99.87	99.86	99.43	380.49
30	500	100.00	100.00	100.00	3.77	99.90	99.90	99.62	272.66
30	1000	100.00	100.00	100.00	4.20	100.00	100.00	100.00	462.94
50	100	66.72	66.43	12.39	10.34	96.51	96.48	85.72	1060.36
50	150	96.12	96.11	80.37	6.27	99.15	99.15	95.75	684.74
50	200	98.98	98.98	94.00	10.40	99.86	99.86	99.18	943.92
50	300	99.96	99.96	99.76	15.71	99.98	99.98	99.88	1084.48
50	400	100.00	100.00	100.00	12.84	100.00	100.00	100.00	1201.24
50	500	100.00	100.00	100.00	12.09	100.00	100.00	100.00	1054.91
50	1000	100.00	100.00	100.00	13.40	100.00	100.00	100.00	1346.05
100	100	53.85	53.76	0.66	30.04	95.07	95.04	78.88	2638.14
100	150	55.18	55.05	1.31	41.71	98.58	98.57	92.46	3680.19
100	200	64.12	63.90	8.25	35.29	99.48	99.48	97.15	5531.03
100	300	95.68	95.68	76.53	48.60	99.87	99.87	99.11	4498.84
100	400	99.37	99.37	95.68	47.35	99.93	99.93	99.62	3600.21
100	500	99.86	99.86	99.01	38.01	99.99	99.99	99.93	4258.09
100	1000	100.00	100.00	100.00	66.13	100.00	100.00	100.00	4466.90
150	100	53.95	53.85	0.69	30.15	90.68	90.65	62.94	10033.99
150	150	52.73	52.69	0.39	35.78	96.34	96.32	81.91	7275.69
150	200	53.07	52.99	0.43	56.84	98.55	98.54	91.82	8828.72
150	300	61.05	60.87	5.37	67.34	99.67	99.67	97.86	7975.81
150	400	91.45	91.45	59.37	43.67	99.86	99.86	99.00	11218.66
150	500	98.05	98.05	87.25	41.83	99.97	99.97	99.74	11936.88
150	1000	99.99	99.99	99.95	69.81	100.00	100.00	100.00	11609.61

4. 실증자료분석

본 장에서는 기존의 실증자료를 바탕으로 두 방법의 성능을 비교하고자 S&P 500 주가와 서울시 자치구별 매매가격지수 데이터를 활용하였다.

4.1. S&P 500 주가 데이터

먼저 고려한 데이터는 Yahoo finance에서 제공하는 S&P 500 지수에 포함된 종목의 주가 데이터이다. 파이썬의 yfinance 패키지를 통해 수집하였고 미국 S&P 500 지수에 속하는 종목 중 금융 섹터와 경기 소비재 섹터에 해당하는 124개 종목의 2024년 3월 1일부터 2025년 3월 1일까지 일별로 측정된 250개의 주가 데이터를 사용하였다. 각 종목의 변동성을 구하기 위해 Guðmundsson과 Brownlees (2021)에서 진행한 방법과 동일하게 일별 고점과 저점의 차이를 이용하여

$$\tilde{\sigma}_{it}^2 = 0.361(p_{it}^{high} - p_{it}^{low})^2$$

의 식을 통해 각 종목의 일별 변동성을 계산하였다. 또한, 각 종목에 대한 S&P 500 지수의 공통적인 영향을 제거하기 위해 이전 시점의 S&P 500 지수의 변동성($\tilde{\sigma}_{s\&P500,t-1}$)를 독립변수로 하는 회귀모형

$$\log(\tilde{\sigma}_{it}^2) = \beta_0 + \beta_1 \log(\tilde{\sigma}_{s\&P500,t-1}) + Y_{it}$$

Table 4.1 Summary of model performance on S&P 500 stocks

	VARblockbuster				**AGE**			
	ACC(%)	F_1(%)	NMI($\times 10^2$)	Time	ACC(%)	F_1(%)	NMI($\times 10^2$)	Time
S&P 500	50.81	50.08	0.00	0.30	74.19	73.01	16.23	1994.37

을 설정하고 이로부터 구한 잔차 Y_{it}를 개별 종목의 최종 변동성을 사용하였다. 이때, AGE 학습 시 격자 탐색 방법을 통해 초매개 변수를 선택하였고 그 결과는 부록 Table A.2에 요약하였다. 이를 통해 선택된 초매개변수 조합에 기반하여 전체 데이터에 대한 AGE의 학습을 진행하였으며, 최종적으로 VAR-블록버스터와 AGE의 성능 비교 결과를 Table 4.1에 요약하였다.

두 방법 모두 S&P 500의 두 섹터에 해당하는 종목에 대한 군집 탐지에서 우수한 성능을 보였으나, AGE가 상대적으로 74%의 정확도로 S&P 500 섹터에 보다 일치한 결과를 주었다. F_1 및 NMI에서도 VAR-블록버스터보다 더 나은 성능을 보였다. 하지만, 그러나 AGE는 VAR-블록버스터 모델에 비해 학습 시간이 상당히 길어지는 단점이 있어 향후 더 최적화된 학습 전략을 통해 시간에 대한 개선이 필요해 보인다.

Table 4.2는 AGE의 노드별 임베딩 결과를 이용하여 추정한 군집 결과를 나타낸다. 첫 번째 군집은 S&P가 소비재 섹터로 구별한 종목들이 다수를 이루는 군집으로 일부 금융 섹터 종목들이 함께 군집된 결과를 보인다. 이들 금융 종목은 전통적인 상업은행이나 보험사가 아닌, 금융 인프라와 데이터 기반 서비스에 특화된 기업군으로 구성되어 있다는 점에서 구조적으로 차별화된다. 예를 들어, CBOE (Cboe global markets), CME (Chicago mercantile exchange), MKTX (Marketaxess holdings)는 주요 거래소 및 전자거래 플랫폼 운영사이며, FDS (FactSet), MSCI (Morgan stanley capital international)는 지수 및 금융 데이터 제공사, FIS (Fidelity national information services), GPN (Global payments), PYPL (PayPal)은 결제 솔루션 및 핀테크 기업이다. 두 번째 군집은 S&P가 금융 섹터로 구분한 종목들이 다수를 이루는 군집으로 AGE에 의해서 묶인 소비재 섹터는 AMZN (Amazon), DASH (DoorDash), GM (General Motors), LULU (Lululemon), TPR (Tapestry), TSLA (Tesla) 등 소비재 및 기술 기업과 BKNG (Booking holdings), CCL (Carnival), EXPE (Expedia), HLT (Hilton), MAR (Marriott) 등 여행·관광 업종이다. 첫 번째 군집은 소비자 지출, 주택시장, 실물경제 활동에 의존하는 기업들인 반면에 두 번째 군집은 신용 주기, 금리, 레버리지, 시장 유동성에 크게 의존하는 기업들로 파악된다. 따라서, AGE가 찾은 두 개의 군집은 소비와 실물경제 변수에 민감한 소비 중심-핀테크 융합형 구조를 가지는 군집과 전통 금융기관을 중심으로 금리나 시장 불안 같은 경제적 충격에 대한 공통 반응을 보이는 구조로 구별하는 것으로 파악되어 S&P 500 지수의 종목 구별과는 미묘한 차이가 있었다.

4.2. 서울시 자치구별 매매가격지수

이번 장에서 고려한 데이터는 서울시 자치구별 매매가격지수 데이터로 한국부동산원의 부동산통계정

Table 4.2 Summary of cluster of S&P 500 stocks

군집	종목 리스트
소비재 중심 군집	ABNB, APTV, AZO, BBY, CMG, CZR, DHI, DPZ, DRI, F, GPC, HAS, HD, KMX, LEN, LKQ, LOW, LVS, MCD, MGM, MHK, NKE, NVR, ORLY, PHM, POOL, ROST, SBUX, TSCO, ULTA, WSM, WYNN, YUM, *BEN, CBOE, CME, EG, ERIE, FDS, FIS, GL, GPN, JKHY, MKTX, MSCI, PYPL, TROW*
금융 중심 군집	ACGL, AFL, AIG, AIZ, AJG, ALL, AMP, AON, APO, AXP, BAC, BK, BLK, BRK-B, BRO, BX, C, CB, CFG, CINF, COF, CPAY, DFS, FI, FITB, GS, HBAN, HIG, ICE, IVZ, JPM, KEY, KKR, L, MA, MCO, MET, MMC, MS, MTB, NDAQ, NTRS, PFG, PGR, PNC, PRU, RF, RJF, SCHW, SPGI, STT, SYF, TFC, TRV, USB, V, WFC, WRB, WTW, *AMZN, BKNG, CCL, DASH, DECK, EBAY, EXPE, GM, GRMN, HLT, LULU, MAR, NCLH, RCL, RL, TJX, TPR, TSLA*

Table 4.3 Summary of model performance on sales price indices

	VARblockbuster				AGE			
	ACC(%)	F_1(%)	NMI($\times 10^2$)	Time	ACC(%)	F_1(%)	NMI($\times 10^2$)	Time
Sales price index	52.00	51.30	0.05	0.03	72.00	71.82	14.44	108.64

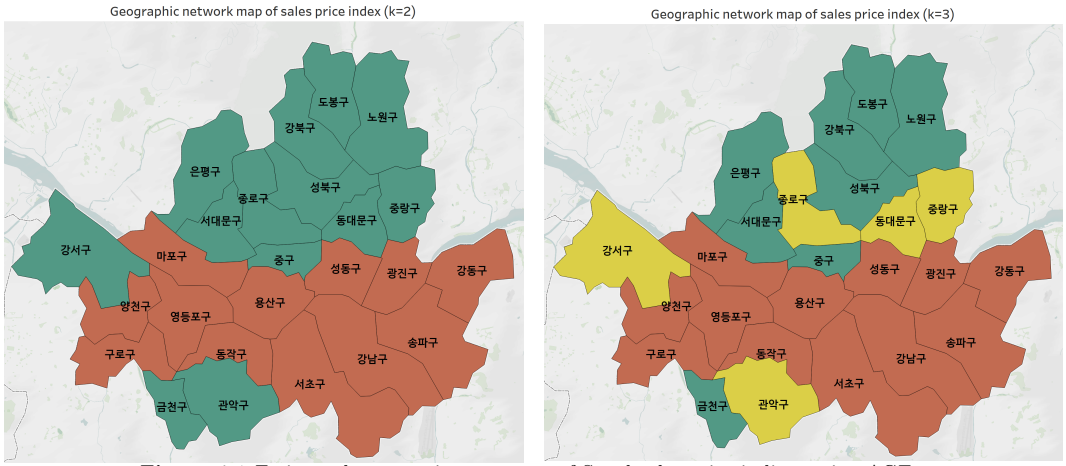

Figure 4.1 Estimated community structure of Seoul sales price indices using AGE

보시스템 (`https://www.reb.or.kr/r-one/portal/main/indexPage.do`)에서 제공하는 데이터이다. 2015년 9월 7일부터 2025년 4월 14일까지 주 단위로 측정된 500개의 데이터를 사용하였다. 서울시 자치구별 매매가격에 대한 서울시 전체의 매매가격의 공통적인 영향을 제거하기 위해

$$X_{it} = \beta_0 + \beta_1 X_{seoul,t-1} + Y_{it} \tag{4.1}$$

의 식에 따라 이전 시점의 서울시 전체의 매매가격지수(X_{seoul})를 독립변수로 하는 회귀모형을 통해 계산된 잔차를 이용하였다. 격자 탐색을 통해 선택된 AGE의 최적 초매개변수 조합은 부록 Table A.2에 요약하였다.

AGE의 노드별 임베딩과 군집화 결과를 바탕으로 서울의 자치구별 매매가격 간의 군집 구조를 Figure 4.1에 나타냈다. 먼저 왼쪽 그림은 군집을 두 개로 나누었을 때의 결과이다. 주황색 군집에 해당하는 지역은 강남을 포함한 한강변을 따라 위치한 대부분의 자치구가 해당되며, 강북 지역 중에서는 용산구, 성동구, 마포구 등이 함께 군집화된 것으로 확인된다. 즉, AGE 모형에 따르면 강남, 강북의 구별이 아니라, 한강변 지역과 그 외 지역으로 구별하는 것을 확인할 수 있었다. Figure 4.1의 오른쪽은 군집의 개수를 3개로 늘릴 경우 AGE에 기반한 군집 구조 결과이다. 이 경우 한강변 이외 지역을 추가적으로 2개로 더 구분하였는데, 종로구, 동대문구, 중랑구, 강서구, 관악구를 하나의 군집으로 구별하였다. 이들은 주황색 첫번째 군집에 인접한 구들로 재개발 호재 등으로 최근 평균 매매 가격이 상승추세에 있는 구들로 파악된다. 같은 자료에 대해 VAR-블록버스터로 군집 탐지를 진행하였으나 직관적으로 해석 가능한 군집 결과를 주지 못하여 이에 대한 비교는 생략하고자 한다.

5. 결론 및 논의

본 연구에서는 다변량 시계열 데이터의 군집 탐지에 딥러닝 방법인 AGE를 활용하였을 때 VAR-블록버스터와 어떤 성능의 차이가 있는지 살펴보았다. 두 모델은 전반적으로 높은 정확도와 F_1 스코어를 기

록했으며, 다양한 차원과 시계열 길이에 따라 성능이 변화하는 경향을 보였다. 특히, AGE는 고차원 데이터에서 더 효과적인 성능을 나타내었으며, VAR-블록버스터는 상대적으로 차원이 작은 데이터에서 더 우수한 성능을 보였다. 그러나 AGE는 딥러닝 기반 모델이므로 격자 탐색을 통한 초매개변수 최적화 및 학습으로 인해 상대적으로 더 많은 시간이 소요되었다.

실증 자료 분석에서는 S&P 500 주가 데이터과 서울시 자치구별 매매가격지수 데이터를 활용하여 두 모델의 성능을 평가하였다. S&P 500 주가 데이터를 대상으로 한 군집 탐지에서 AGE가 상대적으로 S&P 500의 섹터 분류에 대해 더 높은 정확도와 F_1 스코어를 기록한 반면, VAR-블록버스터 모델은 비교적 빠른 학습 시간과 우수한 성능을 보였다. 반면 S&P 500 자료의 경우 분석 기간을 $2 \sim 3$년으로 확장하였을 때 S&P 500의 섹터 분류와 각기 다른 ACC 일치도를 보였으며 군집된 종목들도 달라지는 양상을 확인할 수 있었다. 이는 군집 구조가 시간에 따라 변할 수 있음을 시사하며 군집의 동적 변화를 추적하기 위해 시간에 따라 변하는 다이나믹 모형이나 구조적 변화점을 반영한 모형으로의 확장 등 추가적인 모델링과 후속 연구가 요구된다. 서울시 자치구별 매매가격지수에 대한 군집 탐지에서는 AGE가 한강변을 기준으로 인접한 자치구와 떨어진 자치구로 군집을 형성함을 알 수 있어 소위 강남과 강북 지역의 구별이 아닌 한강의 접근성을 기준으로 군집이 형성됨을 살펴보았다. 전반적으로 AGE를 활용한 고차원 시계열 자료의 군집 탐지가 직관적이고 해석 가능한 군집을 도출함을 살펴볼수 있었다. 하지만, AGE의 경우 매우 긴 계산 시간에 따라 다양한 초매개변수 최적화 기법 등을 통하여 시간 효율성을 향상하는 연구가 필요해 보인다.

부록. 모의실험 및 실증자료 초매개변수 선택 결과표

모의실험과 실증자료 분석에서 격차 탐색 방법을 통해 선택된 최적의 초매개변수는 다음과 같다.

Table A.1 Optimal hyper-parameters chosen by AGE and grid search

n	T	dim	$epoch$	r_{ed}^{pos}	DBI
30	100	32	35	0.05	0.0564
30	150	32	35	0.05	0.0268
30	200	64	40	0.05	0.0214
30	300	32	45	0.05	0.0255
30	400	32	45	0.05	0.0147
30	500	32	50	0.05	0.0086
30	1000	32	45	0.03	0.0093
50	100	64	40	0.05	0.0642
50	150	32	30	0.01	0.0202
50	200	64	30	0.03	0.0149
50	300	64	35	0.03	0.0159
50	400	64	45	0.03	0.0173
50	500	64	45	0.01	0.0126
50	1000	32	50	0.03	0.0091
100	100	64	30	0.01	0.0615
100	150	64	40	0.03	0.0352
100	200	128	50	0.05	0.0312
100	300	64	45	0.03	0.0151
100	400	64	35	0.01	0.0096
100	500	64	45	0.01	0.0083
100	1000	32	50	0.03	0.0070
150	100	32	50	0.03	0.6959
150	150	64	30	0.03	0.1718
150	200	128	30	0.03	0.0692
150	300	64	50	0.01	0.0429
150	400	64	45	0.03	0.0500
150	500	128	50	0.03	0.0465
150	1000	32	45	0.05	0.0066

Table A.2 Optimal hyper-parameters chosen by AGE and grid search

Dataset	n	T	dim	$epoch$	r_{ed}^{pos}	DBI
S&P 500	124	250	64	50	0.01	0.5778
Sales price index	25	500	32	40	0.03	0.1279

References

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

Cui, G., Zhou, J., Yang, C. and Liu, Z.(2020). Adaptive graph encoder for attributed graph embedding. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 976-985.

Guðmundsson, G. S. and Brownlees, C.(2021). Detecting groups in large vector autoregressions. *Journal of Econometrics*, **225**, 2-26.

Jin, D., Li, B., Jiao, P., He, D., and Shan, H.(2019). Community detection via joint graph convolutional network embedding in attribute network.*Artificial Neural Networks and Machine Learning - ICANN 2019: Workshop and Special Sessions*, 594-606.

Kim, I. and Kim, J.(2021). Multivariate time series clustering of electricity consumption data *Journal of the Korean Data & Information Science Society*, **32**, 569-584.

Kim, S., Jeon, Y. and Oh, H.(2021). Analysis of PM10 data using spatio-temporal correlation clustering and time series similarity. *Journal of the Korean Data & Information Science Society*, **32**, 1259-1279.

Lee, S. and Baek, C.(2023). Volatility changes in cryptocurrencies: evidence from sparse VHAR-MGARCH model. *Applied Economics Letters*, **30**, 1496-1504.

Park, D. and Yoon, S.(2017). Clustering and classification to characterize daily electricity demand *Journal of the Korean Data & Information Science Society*, **28**, 398-406.

Park, H. and Na, J.(2022). Time series clustering of metropolitan railway traffic using dynamic time warping. *Journal of the Korean Data & Information Science Society*, **33**, 775-783.

Tao, Z., Liu, H., Li, J., Wang, Z., and Fu, Y.(2019). Adversarial graph embedding for ensemble clustering. *In Proceedings of the International Joint Conferences on Artificial Intelligence Organization*.

Tian, F., Gao, B., Cui, Q., Chen, E. and Liu, T.-Y.(2014). Learning deep representations for graph clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, **28**.

Yang, L., Wang, Y., Gu, J., Wang, C., Cao, X., and Guo, Y.(2020). JANE: Jointly adversarial network embedding. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 1381-1387.

Yoon, M. and Kim, Y.(2025). Predicting Korean stock prices using heterogeneous dynamic graph neural networks(HDGNN). *Journal of the Korean Data & Information Science Society*, **36**, 13-22.

Journal of the Korean Data & Information Science Society
2025, **36**(5), 1–13

http://dx.doi.org/10.7465/jkdi.2025.36.5.1
한국데이터정보과학회지

Community detection in high dimensional time series data using graph-based neural networks[†]

Minjeong Song[1] · Changryong Baek[2]

[12]Department of Statistics, Sungkyunkwan University, Seoul, Korea.

Received 29 May 2025, revised 11 July 2025, accepted 22 July 2025

Abstract

In high-dimensional time series clustering, the VAR-Blockbuster algorithm proposed by Guðmundsson and Brownlees (Guðmundsson and Brownlees, 2021), which combines spectral clustering with vector autoregression (VAR), has been widely adopted due to its theoretical soundness and computational efficiency. However, its reliance on the Stochastic Block-Vector Autoregression (SB-VAR) model may limit its general applicability. To address this limitation, we investigate an alternative approach based on Adaptive graph encoder (AGE), a deep learning graph neural network that does not depend on model assumptions. Simulation studies show that while VAR-Blockbuster performs reliably when the sample size is sufficiently large relative to dimensionality, AGE exhibits superior clustering performance in small-sample, high-dimensional settings. Furthermore, applying the AGE-based method to financial and consumer sector stocks in the S&P 500 index, as well as to housing price indices across districts in Seoul, reveals that it can successfully identify interpretable and meaningful cluster structures.

Keywords: High-dimensional time-series, community detection, SB-VAR, VARblockbuster, AGE.

[†] This work was supported by the National Research Foundation of Korea grant funded by the Korea government(MSIT)(RS-2025-00519717).

[1] Graduate student, Department of Statistics, Sungkyunkwan University, Seoul 03063, Korea.
[2] Corresponding author: Professor, Department of Statistics, Sungkyunkwan University, Seoul 03063, Korea. E-mail: crbaek@skku.edu