creative commons
COMMONS DEED

저작자표시 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

이학석사학위논문

A study on change point detection of time series data using topological data analysis

위상적데이터분석방법을 활용한 시계열데이터분석

BY

송원석

2022 년 8 월

A study on change point detection of time series data using topological data analysis

위상적데이터분석방법을 활용한 시계열데이터분석

지도교수 이 상 열

이 논문을 이학석사 학위논문으로 제출함

2022 년 8 월

서울대학교 대학원

통계학과

송 원 석

송원석의 이학석사 학위논문을 인준함

2022 년 8 월

위 원 장	임요한	(인)
부위원장	이상열	(인)
위 원	원중호	(인)

초록

이번 논문에서는 위상적데이터분석방법론(TDA)을 활용하여 GARCH(1,1) 데이터의 변화점 탐지에 적용하고자 한다. 그 전에 기존 변화점 탐지 방법에 대해서 설명하고자 한다. Lee (2020) 논문에서 제시한 기존의 브라우니안 브릿지를 활용한 통계량은 안정적인 GARCH(1,1) 데이터($\alpha + \beta$ 가 1보다 매우 작은 경우)에 대해 유용한 결과를 보여주었다.

반면에 위상적데이터방법과 지역선형임베딩을 통한 변화점 탐지 방법은 불안정적인 GARCH(1,1) 데이터($\alpha + \beta$ 가 1에 가까운 경우)에 대해서는 기존 방법보다 더 높은 power를 나타냈다. 하지만 기존방법과 신규방법 모두 불안정적인 데이터에서 높은 size를 나타내었다는 한계를 보여주었다.

신규방법은 Umar Islambekov(2019) 논문에서 제시한 방법과 같이 시계열 데이터로부터 기하학적 특성을 추출하고자 베티넘버를 활용하였다. 이어서 베터넘버 데이터의 차원을 지역선형임베딩(다양체학습)을 활용하여 축소시킨다. 마지막으로 Lee & Kim(2022) 벡터자기회귀모형에 대한 변화점 탐지방법을 차원축소된 데이터에 적용한다.

주요어: 시계열 자료, 위상적데이터방법론, 다양체학습, CUSUM.

학번: 2020-29046

목차

초록 **i**

1 소개 **1**

2 이론적 배경 및 적용방법 **3**

2.1 변화점 탐지이론 3

2.1.1 GARCH(1,1) 변화점 탐지 3

2.1.2 벡터자기회귀모형(VAR) 변화점 탐지 4

2.2 데이터 변환 방법론 7

2.2.1 위상적데이터방법론(Topological Data Analysis) 7

2.3 데이터 차원축소 방법론 11

2.3.1 지역선형임베딩(Locally Linear Embedding) 11

3 데이터 시뮬레이션 **14**

3.1 데이터 생성과 적용방법 14

3.1.1 데이터 생성 14

3.1.2 시뮬레이션 방법 17

3.1.3 시뮬레이션 결과 20

4 결론 **24**

참고문헌 **25**

Abstract **27**

표 목차

3.1 변화점탐지 비교방법 17
3.2 데이터1 적용방식별 Size 비교 20
3.3 데이터2 적용방식별 Power 비교 21
3.4 (데이터1) 적용방식2 모수 k값에 따른 Size 비교 22
3.5 (데이터2) 적용방식2 모수 k값에 따른 Power 비교 23

그림 목차

2.1 k-차원 simplex 및 단체복합체[1] 8
2.2 ϵ 증가시 비아토리스-립스 변화[2] 9
2.3 기하학적 공간에 따른 베티넘버[3] 10
2.4 지역선형임베딩[4] 13

3.1 데이터1 사진예시 15
3.2 데이터2 사진예시 16

제 1 장 소개

위상적데이터분석방법론(TDA)의 장점은 데이터 분석에 있어서 비모수적(nonparametric)이다는 점이 있으나 시계열 자료를 분석함에 있어 완전히 비모수적(nonparametric)하다고 보기는 어렵다. 특히 Umar Islambekov(2019)에서도 시계열자료를 베티넘버(Betti number)들로 변환하였고 이러한 변환된 데이터의 변화점을 탐지할때 기존에 연구되었던 변화점탐지 방법론이 활용되었다. 위상적데이터분석방법론을 활용함에 있어 현재까지는 기존 시계열 분석방법론들을 완전히 대체하기위해서는 보다 많은 연구가 필요하다고 판단된다. 다만 시계열 자료를 분석함에 있어서 위상적데이터분석방법론이 성능을 개선해줄 수 있는지를 확인해볼 필요가 있다고 판단였고 이를 주안점으로 삼아 논문을 작성하였다.

시계열 자료의 변화점탐지를 위해서 먼저 GARCH(1,1) 모형을 활용하여 데이터를 생성하였다. 기존연구를 적용하고자 생성된 데이터를 Lee (2020)에서 제시한 변화점탐지방법을 활용하여 size와 power를 계산하였다. 또한 생성된 데이터를 위상적데이터방법론을 활용하여 변형시키고 또한 다양체학습방법을 활용하여 3차원 축소시켰다. 변형된 데이터는 VAR모형의 병화점탐지방법론인 Lee & Kim (2022)의 VAR 변화점 탐지방법을 적용하여서 size와 power를 계산하였다.

이 논문의 순서는 다음과 같다. 제2창에서는 Lee (2022)를 기초로하여 이론적 배경과 적용방법을 다룰 예정이다. 또한, Lee (2020)을 통해 데이터 생성과 변화점탐지방법론을 다룰계획이고 Lee & Kim (2022)에서 제시한 VAR 변화점 탐지방법도 다룰 계획이다. Umar Islambekov(2019)에서 제시한 위상적데이터방법론에 대해

서도 설명할 계획이다. 제3장에서는 구체적인 시뮬레이션 방법과 결과를 제시할 계획이다. 마지막으로 제4장에서는 결론내용을 포함할 계획이고 관련증빙과 연구자료는 부록에 포함할 계획이다. 참고문헌에서는 관련된 논문목록을 추가하였다.

제 2 장 이론적 배경 및 적용방법

2.1 변화점 탐지이론

2.1.1 GARCH(1,1) 변화점 탐지

먼저 분석에 사용할 데이터는 GARCH(1,1)으로 아래와 같은 수식으로 정리할 수 있다.

$$\begin{cases} x_t = \sigma_t \epsilon_t, \\ \sigma_t^2 = \omega + \alpha x_{t-1}^2 + \beta \sigma_{t-1}^2, \end{cases} \tag{2.1}$$

상기 램덤분포 ϵ_t는 평균 0 및 i.i.d이다. 또한 $\omega > 0,\ E\epsilon_1^{4s} < \infty\ \ (s > 1),\ \alpha, \beta \geq 0,$ $\omega\alpha + \beta < 1$를 가정한다. GARCH 모형을 통해 길이 n의 데이터를 생성할 수 있다. 생성된 데이터가 변화점이 있는지 여부를 탐지하고자 Lee & Kim (2022)에서는 시계열 자료의 길이가 n이라고 가정할때 아래와 같은 가설에 대해서 CUSUM test를 실행할 수 있다고 안내하고있다.

$$\mathcal{H}_0 : (\omega_0, \alpha_0, \beta_0) \text{ 시간에 불변 } t = 1, \ldots, n. \text{ vs. } \mathcal{H}_1 : \mathcal{H}_0 \text{ 기각} \tag{2.2}$$

상기 귀무가설을 검정하기위해 아래와 같은 CUSUM test를 실행한다.

$$\tilde{T}_n = \frac{1}{\sqrt{n}\tilde{\tau}_1} \max_{1 \leq k \leq n} \left| \sum_{t=1}^{k} \hat{\epsilon}_t^2 - \frac{k}{n} \sum_{t=1}^{n} \hat{\epsilon}_t^2 \right| \tag{2.3}$$

$$\tilde{\tau}_1^2 = \frac{1}{n}\sum_{t=1}^{n}\hat{\epsilon}_t^4 - \left(\frac{1}{n}\sum_{t=1}^{n}\hat{\epsilon}_t^2\right)^2$$

상기 식에서 제시된 $\hat{\epsilon}_t$ 다음의 $\hat{\epsilon}_t = y_t/\hat{\sigma}_t$, $\hat{\sigma}_t^2 = 1 + \hat{\alpha}y_{t-1}^2 + \hat{\beta}\sigma_{t-1}^2$ 관계를 만족한다. 또한 GARCH(1,1) 모형 변화점 탐지를 위해 제시된 $\tilde{T}_n$는 GARCH(1,1)에서 생성된 데이터의 잔차로 계산된 통계량이다. 변화점이 존재하는 경우 해당 통계량이 임계점을 상회하며 이에 따라 길이 n의 데이터 내에 변화점이 있다고 판단할 수 있고 $\mathcal{H}_0$를 기각할 수 있다. Lee & Kim (2022)에서는 제시된 통계량 ($\tilde{T}_n$)이 특정 분포로 아래의 정리와 같이 수렴함을 밝혔다.

Theorem 1 *ϵ_1의 밀도함수인 $f := F'$가 미분가능하고 $\sup_x f(x) + \sup_x |f'(x)| < \infty$를 만족한다고 하자. 만일 $\mathcal{H}_0$을 만족한다고 가정할 경우, $n \to \infty$일때 $\hat{T}_n$는 아래와 같이 분포수렴한다.*

$$\hat{T}_n \xrightarrow{d} \sup_{0\le s\le 1}|B^\circ(s)|$$

Lee & Kim (2022) 논문에서 $\tilde{T}_n$이 특정분포로 수렴함을 증명하였고 시뮬레이션을 통해 기각역을 계산하였다. 금번 논문에서는 제시된 통계량과 기각역을 활용하여 GARCH(1,1) 모형의 변화점 탐지에 활용하였다.

2.1.2 벡터자기회귀모형(VAR) 변화점 탐지

다음 변화점 탐지를 위한 모형으로 벡터자기회기모형(VAR)을 생각할 수 있다.

$$Y_t = \sum_{i=1}^{p}\Phi_i Y_{t-i} + \epsilon_t \tag{2.4}$$

상기 Φ_i는 $i = 1, \ldots, p$ 에 대해 $d \times d$ 행렬이며, 랜덤벡터 ϵ_t 는 i.i.d 평균이 0이고 d-차원 랜덤벡터이다. 또한 ϵ_t 는양정치 공분산행렬 Σ를 가지고 $E\left\|\epsilon_t\right\|^4 < \infty$를 만족한다.

벡터자기회귀(VAR) 모형의 모수행렬로 이루어진 다음의 Λ은 $\Lambda = (\Phi_1, \ldots, \Phi_p)$로 정의되며 $\mathbb{R}^{pd^2}$의 컴팩트 집합 $\mathcal{S}$에서 내부해를 가진다고 가정한다. 모수행렬들은 $|z| \leq 1$ 일때 $\det\left(I_d - \sum_{i=1}^{p} \Phi_i z^i\right) \neq 0$ (I_d는 $d \times d$ 항등행렬)이라는 관계를 만족한다.

다음의 $\left(I_d - \sum_{i=1}^{p} \Phi_i z^i\right)^{-1} = \sum_{k=0}^{\infty} \Psi_k z^k$을 만족시키는 행렬 Ψ_k들은 $\Phi_1, \ldots, \Phi_p$의 함수로 생각할 수 있으며 이를 정리하여 VAR 모형을 아래와 같이 표현할 수 있다.

$$Y_t = \sum_{k=0}^{\infty} \Psi_k \epsilon_{t-k},$$

정리된 벡터자기회귀(VAR) 모형의 모수행렬인 Ψ_k 변수는 특정한 $C > 0$와 모든 $\rho \in (0, 1)$에 대해 $\|\Psi_k\| \leq C\rho^k$을 만족한다. 주어진 관측치 Y_n를 활용하여 구한 잔차는 다음의 $\hat{\epsilon}_t = Y_t - \sum_{i=1}^{p} \hat{\Phi}_i Y_{t-i}$와 같다.

이상열(2020)에서는 $\hat{\Phi}_i$은 $\sqrt{n}$-일치추정량이고 이는 Yule-Walker 또는 최소제곱 추정량(LSE)로 생각할 수 있다고 설명한다. 따라서 위와 같은 벡터자기회귀(VAR) 모형의 모수행렬(Φ_i와 Σ)의 변화가 있는지 검정하기위해서 아래와 같은 가설들을 세운다.

$\mathcal{H}_0 : \Phi_i, \Sigma, i = 1, \ldots, p$, 시간에 대하여 불변 t=1,$\cdots, n$

$\mathcal{H}_1 : \text{not}\, \mathcal{H}_0$

상기 귀무가설을 검정하기 위해 다음과 같은 통계량 $\hat{T}_n^{max}$을 계산하여 임계값을 넘어서는지 관찰한다.

$$\hat{T}_n^{\max} = \max_{1\le k\le n} \max\left\{ \frac{1}{\sqrt{n}\hat{\tau}_{1,n}} \left| \sum_{t=1}^{k} (Y_t - \hat{\epsilon}_t)^T \hat{\epsilon}_t - \left(\frac{k}{n}\right) \sum_{t=1}^{n} (Y_t - \hat{\epsilon}_t)^T \hat{\epsilon}_t \right|, \right.$$
$$\left. \frac{1}{\sqrt{n}\hat{\tau}_{2,n}} \left| \sum_{t=1}^{k} \|\hat{\epsilon}_t\|^2_{\hat{\Sigma}_n} - \left(\frac{k}{n}\right) \sum_{t=1}^{n} \|\hat{\epsilon}_t\|^2_{\hat{\Sigma}_n} \right| \right\}$$

$$\hat{\tau}^2_{1,n} = \frac{1}{n} \sum_{t=1}^{n} \left((Y_t - \hat{\epsilon}_t)^T \hat{\epsilon}_t \right)^2 - \left(\frac{1}{n} \sum_{t=1}^{n} (Y_t - \hat{\epsilon}_t)^T \hat{\epsilon}_t \right)^2$$
$$\hat{\tau}^2_{2,n} = \frac{1}{n} \sum_{t=1}^{n} \|\hat{\epsilon}_t\|^4_{\hat{\Sigma}_n} - \left(\frac{1}{n} \sum_{t=1}^{n} \|\hat{\epsilon}_t\|^2_{\hat{\Sigma}_n} \right)^2$$

위 식에서 사용한 기호 대해 설명하면 벡터 x와 양정치행렬 A에 대해 $\|x\|_A$는 $\|x\|_A = \left(x^T A^{-1} x\right)^{1/2}$을 의미하며 $\hat{\Sigma}_n$는 $\hat{\Sigma}_n = (1/n)\sum_{t=1}^{n} \hat{\epsilon}_t \hat{\epsilon}_t^T$를 의미한다.
상기 통계량 $\hat{T}_n^{max}$은 아래와 같이 특정분포로 수렴한다는 것을 Lee (2020)에서 설명하고 있다.

Theorem 2 *가설 $\mathcal{H}_0$ 가정하에, $n \to \infty$,*

$$\hat{T}_n^{\max} \xrightarrow{d} \sup_{0\le s\le 1} \max\left\{ \left|\mathrm{W}_{11}^{\circ}(s)\right|, \left|\mathrm{W}_{12}^{\circ}(s)\right| \right\}$$

상기 W_{11}°와 W_{12}°는 서로 독립인 브라우니안 브릿지(Brownian bridges)을 의미한다. Lee (2020) 논문에 따르면 $\hat{T}_n$이 브라우니안 브릿지로 구성된 분포로 수렴한다고 한다. 또한 시뮬레이션을 통해 기각역을 제시하고 있으며 $\hat{T}_n^{\max} > 1.4596$인 경우 유의수준 0.05에서 귀무가설을 기각한다고 안내하고 있다.

2.2 데이터 변환 방법론

2.2.1 위상적데이터방법론(Topological Data Analysis)

본 논문의 핵심주제인 위상적데이터방법론은 데이터로부터 기하학적 특성을 추출하여 본래 모집단의 분포를 추정해내는 방법이다. 연속적인 분포에서 추출된 유한개의 표본을 활용하여 본래의 모집단 분포를 추정하기에는 여러 한계점이 존재하나 많은 분야에서 위상적데이터방법론을 통해 데이터를 분석하여 유의미한 성과를 거두고 있다. 이번 장에서는 위상적데이터방법론에 대한 주요개념들을 소개하고 적용방식과 필수개념들에 대해 설명하고자 한다.

Definition 1 *(단체복합체(Simplicial Complex))*
어떤 거리공간 $\mathbf{X}$*의 부분집합* $\chi \subset \mathbf{X}$*에 대해 단체복합체(Simplicial Complex)* K*는* χ*의 유한한 부분집합들의 집합으로 구성되었고 다음을 만족한다.*

$$\alpha \in K,\ \beta \subset \alpha \ \Rightarrow\ \beta \in K$$

만일 $|\alpha| = p + 1$*이면,* α*는 p-simplex라고 한다.*

첫번째로 소개한 개념인 단체복합체는 고차원의 그래프라고 생각할 수 있다. 다만 단체복합체는 다소 추상적인 개념이므로 거리공간에서의 k-차원 simplex를 정의하여 단체복합체의 구성성분들에 대해서 알아보도록 한다.

Definition 2 *(k-차원 simplex)*
거리공간에서 다음과 같은 특징을 갖는 요소들의 집합을 k-차원 simplex라고 정의한다.

$$\Delta_k := \left\{ (x_0, \ldots, x_k) \in \mathbb{R}^{k+1} \mid \sum_{i=0}^{k} x_i = 1, x_i \geq 0 \right\}$$

앞서 소개한 k-차원 simplex 정의에 따르면 0-simplex는 한 점을 의미하고 1-simplex는 선의 일부, 2-simplex는 삼각형을 의미한다는 것을 알 수 있다.

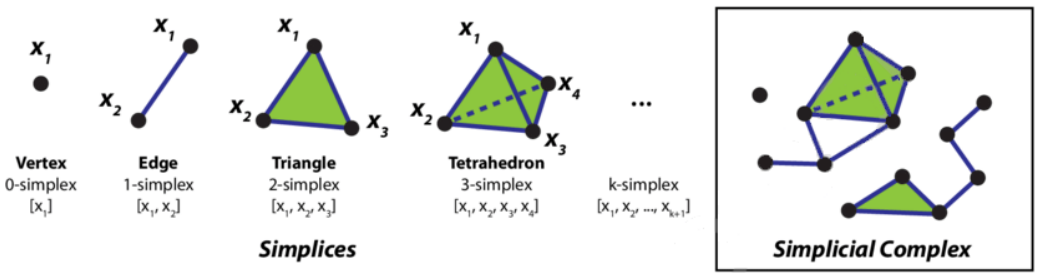


그림 2.1: k-차원 simplex 및 단체복합체[1]

위 그림 2.1에서 볼 수 있듯이 단체복합체(Simplicial Complex)는 k-차원 simplex들로 이루어진 집합으로 생각할 수 있고 특히 Definition1을 만족한다고 이해할 수 있다. 다음은 표본자료들로부터 기하학적 특성을 추출해내는 방법에 대해 설명하고자 한다. 유한개의 표본은 하나하나 별개의 점들이기에 그 자체로 기하학적 특성을 추출할 수는 없고 다음의 개념을 활용하여 유한한 점들의 기하학적 성격을 알아낼 수 있다.

Definition 3 *(비아토리스-립스 복합체(Vietoris-Rips complex))*
거리공간 $\mathbf{X}$*에서 그 유한한 부분집합* $\chi \subset \mathbf{X}$*과 임의의* ϵ*에 대해 비아토리스-립스 복합체* $VR(\chi, \epsilon)$*를 다음과 같이 정의할 수 있다.*

$$VR(\chi, \epsilon) = \{(x_1, \cdots, x_k) \in \chi \ ; \ d(x_i, x_j) < 2\epsilon \ for \ all \ 1 \leq i, j \leq k\}$$

앞서 정의한 $VR(\chi, \epsilon)$*은 단체복합체(Simplicial complex)의 성격을 가진다.*

[1]Mengsen Zhang, Topological portraits of multiscale coordination dynamics

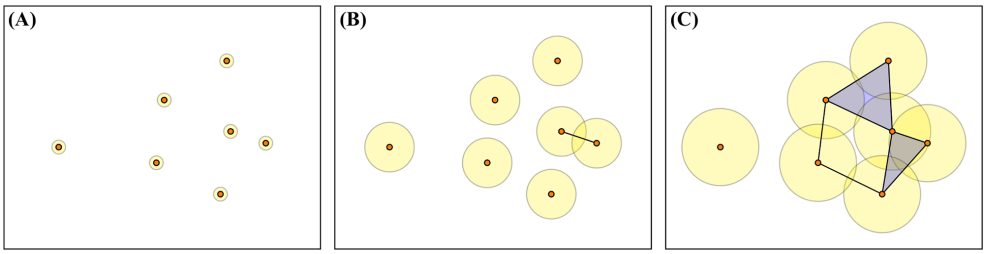


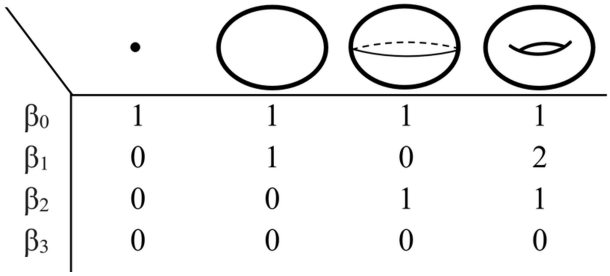
그림 2.2: ϵ 증가시 비아토리스-립스 변화[2]

위 그림 2.2에서 보았듯이 ϵ이 커지면서 연결되는 점들이 늘어남을 확인할 수 있다. ϵ이 매우 작은 (A)에서는 각 점들은 연결되지 않았기 때문에 0-차원 복합체만이 존재하고 (B)에서 점 2개가 이어지면서 연결체의 갯수가 하나 줄어들게된다. (C)에서 보면 점들이 연결됨에 따라 두개의 연결체가 구성되었으며 고리(루프)가 하나 생성된 것을 확인할 수 있다. 이렇듯 ϵ의 크기가 커지면서 기하학적 특성이 변하게 됨을 눈으로 확인할 수 있다. 특히 기하학적 특성의 변화를 계량적으로 나타내기 위해 베티넘버(Betti number)라는 개념을 사용하는데 이에 대해 설명하고자 한다.

Definition 4 *(베티넘버(Betti numbers, β_p))*
단체복합체(simplicial complex) K의 베티넘버(Betti number) β_p, $p \in \mathbb{Z}$이라함은 K의 p-차원 홀(hole)을 의미한다. p가 변함에 따라 베티넘버의 의미가 달라지는데 β_0는 연결된 요소의 개수를 의미한다. β_1은 루프(loop)의 개수를 의미하고 β_2는 빈공간(void)를 의미한다.

베티넘버에 대한 이해를 돕고자 그림 2.3에서 각 공간별 베티넘버들을 나열하였다. 각기 다른 기하학적 구조에 대한 정보를 베티넘버를 활용하여 추출할 수 있고 특성을 나타내는 대표값으로 사용할 수 있다.

[2]출처 : https://doi.org/10.1371/journal.pone.0213679.g003



β_0	1	1	1	1
β_1	0	1	0	2
β_2	0	0	1	1
β_3	0	0	0	0

그림 2.3: 기하학적 공간에 따른 베티넘버[3]

보다 자세하게 설명하면 한개의 점은 연결요소가 하나이기에 $\beta_0 = 1$이고 루프가 없으므로 $\beta_1 = 0$이며 빈공간이 없으므로 $\beta_2 = 0$이다. 다른 예로 원은 연결된 요소가 하나이므로 $\beta_0 = 1$이고 루프가 하나 있으므로 $\beta_1 = 1$이며 빈공간이 없으므로 $\beta_2 = 0$이다.

구의 경우를 생각하면 연결된 요소가 하나이므로 $\beta_0 = 1$이고 루프가 없으므로 $\beta_1 = 0$이며 구 내부에 빈공간이 있으므로 $\beta_2 = 1$이다. 토러스(torus)의 경우에는 연결된 요소가 하나이므로 $\beta_0 = 1$이고 루프가 두개가 있으므로 $\beta_1 = 2$이다. 루프가 두개있다고 함은 토러스 가운데를 통과하는 루프와 둘레를 아우르는 루프가 각각 존재한다. 마지막으로 토러스 내부에 빈공간이 있으므로 $\beta_2 = 1$이다. 다음은 3단계에 대해 나누어 설명하고자 하며 Benyamin Ghojogh(2020) 논문의 내용을 기반으로 작성하였다.

[3]Shafie Gholizadeh, ArminSeyeditabari, Topological Signature of 19th Century Novelists:Persistent Homology in Text Mining

2.3 데이터 차원축소 방법론

2.3.1 지역선형임베딩(Locally Linear Embedding)

다음은 다양체학습(Manifold learning) 방법 중에 하나인 지역선형임베딩(Locally Linear Embedding)에 대해서 알아보고자 한다. 지역선형임베딩은 3단계로 구성되는데 먼저 각 점들 주변에 가까운 순으로 k개 주변점들을 선정한다(k-Nearest Neighbors, kNN). 선형결합을 사용하여 각 점들과 k개 주변점들과의 관계를 나타내는 가중치들을 구한다. 마지막으로 계산한 가중치를 사용하여 모든점을 저차원으로 차원축소 시킨다. 지역선형임베딩의 주요개념은 고차원에서 구한 가중치를 사용하여 데이터의 차원을 축소시켜 저차원 데이터로 변환하는 것이다.

Step1 k개 주변점(kNN, k-Nearest Neighbors)
주변점들의 갯수를 나타내는 모수 k가 주어졌을때 이에 따른 주변점들은 유클리드 공간에서 거리가 가까운 순으로 지정하게 된다. $x_i \in \mathbb{R}^d$가 주어졌을때 $x_{ij} \in \mathbb{R}^d$는 x_i에게 j번째로 가까운 점이라고 할 수 있다. $\mathbf{X_i} \stackrel{\text{def}}{=} [x_{i1}, \cdots, x_{ik}] \in \mathbb{R}^{d \times k}$은 x_i의 k-주변점들로 이루어진 벡터라고 정의한다.

Step2 주변점 활용한 가중치 계산
두변째 단계에서는 k개 주변점들을 활용하여 모든 점을 선형재구조화하기위한 가중치를 계산한다. 고차원 공간에서 주어진 샘플 데이터들의 선형관계를 나타내는 가중치를 계산한다고 볼 수 있으며 이를 위한 최적화식은 다음과 같다.

$$\underset{\widetilde{\boldsymbol{W}}}{minimize} \varepsilon(\widetilde{\boldsymbol{W}}) := \sum_{i=1}^{n} \left\| \boldsymbol{x}_i - \sum_{j=1}^{k} \widetilde{w}_{ij} \boldsymbol{x}_{ij} \right\|_2^2,$$

$$\text{subject to } \sum_{j=1}^{k} \widetilde{w}_{ij} = 1, \quad \forall i \in \{1, \ldots, n\}$$

$\widetilde{\boldsymbol{W}} := [\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_n]^\top \in \mathbb{R}^{n\times k}$과 $\widetilde{\boldsymbol{w}}_i := [\widetilde{w}_{i1}, \ldots, \widetilde{w}_{ik}]^\top \in \mathbb{R}^k$은 $\boldsymbol{x}_i$의 k-주변점들과의 관계를 나타내는 가중치들을 의미한다. 상기 최적화식에는 $\sum_{j=1}^k \widetilde{w}_{ij} = 1$라는 제약조건이 있는데 이는 k-주변점들의 가중치들의 합이 1임을 의미하고 제약조건 아래에서 샘플데이터와 주변점들의 선형결합과의 차이를 최소화하는 가중치를 계산한다고 생각할 수 있다.

Step3 선형임베딩

앞서 구한 가중치들은 샘플데이터와 주변점과의 선형관계를 의미한다. 선형임베딩은 해당 가중치를 활용해 데이터의 선형구조를 보존하면서 저차원으로 차원축소하는 것을 의미한다. 즉 지역적(국소적)인 선형구조를 유지하도록 가중치를 사용한다고 볼 수 있고 선형임베딩은 다음의 최적화식을 계산함으로써 구할 수 있다.

$$\underset{\boldsymbol{Y}}{\text{minimize}} \sum_{i=1}^n \left\| \boldsymbol{y}_i - \sum_{j=1}^n w_{ij}\boldsymbol{y}_j \right\|_2^2,$$

$$\text{subject to } \frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i\boldsymbol{y}_i^\top = \boldsymbol{I}, \quad \sum_{i=1}^n \boldsymbol{y}_i = \boldsymbol{0},$$

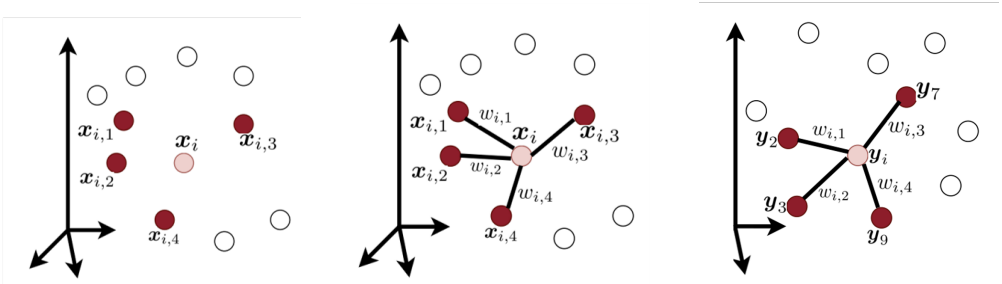
$\boldsymbol{I}$ 는 항등행렬이며 $\boldsymbol{Y} := [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]^\top \in \mathbb{R}^{n\times p}$로 정의했을때 $\boldsymbol{Y}$의 열벡터인 $\boldsymbol{y}_i \in \mathbb{R}^p$들은 차원축소된 데이터라고 볼 수 있다. 특히 w_{ij}는 가중치로서 만일 $\boldsymbol{x}_j$가 $\boldsymbol{x}_i$의 주변점 속하는지 여부에 따라 다음과 같은 가중치를 가진다.

$$w_{ij} := \begin{cases} \widetilde{w}_{ij} & \text{if } \boldsymbol{x}_j \in k\text{NN}(\boldsymbol{x}_i) \\ 0 & \text{otherwise} \end{cases}$$

앞서 설명한 지연선형임베딩의 세 단계에 대해 보다 간단한 설명을 위해 하단의

그림 2.2를 포함하였다. 차원축소된 데이터인 $\boldsymbol{y}_i$를 구하는 과정을 그림으로 나타내면 아래와 같이 나타낼 수 있다.

그림 2.2에서 k=4인 상황을 가정하고 있으며 $\boldsymbol{x}_{i,1} = \boldsymbol{x}_2,\ \boldsymbol{x}_{i,2} = \boldsymbol{x}_3,\ \boldsymbol{x}_{i,3} = \boldsymbol{x}_7,$ 이며 $\boldsymbol{x}_{i,4} = \boldsymbol{x}_9$인 경우를 가정하고 있다. 먼저 (a)에서는 $\boldsymbol{x}_i$로부터 가장 가까운 4개의 점들을 선정하고 있다. (b)에서는 $\boldsymbol{x}_i$ 주변의 4개 점들과 선형관계를 의미하는 가중치를 계산하고 있다. 마지막으로 가중치를 이용하여 고차원에서 저차원으로 선형임베딩을 실행하고있다.



(a) Step1. k개 주변점 (b) Step2. 가중치 계산 (c) Step3. 선형임베딩

그림 2.4: 지역선형임베딩[4]

[4]Benyamin Ghojogh, Ali Ghodsi, Locally Linear Embedding and its Variants

제 3 장 데이터 시뮬레이션

3.1 데이터 생성과 적용방법

3.1.1 데이터 생성

이번 장은 시뮬레이션을 통해 GARCH(1,1) 데이터에 대한 기존의 변화점 탐지 방법과 위상적데이터분석방법간의 성능을 비교분석해보고자 한다 이를 위해 앞서 설명한 GARCH(1,1) 데이터를 생성하였으며 구체적으로 길이(시간,t)가 600인 시계열자료 총 500개(N=500)를 생성하였다.

$$\begin{cases} x_t = \sigma_t \epsilon_t, \\ \sigma_t^2 = \omega + \alpha x_{t-1}^2 + \beta \sigma_{t-1}^2, \end{cases}$$

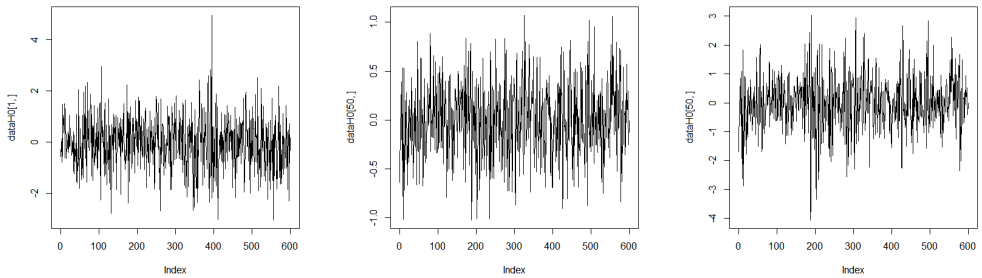
특히 데이터 분석을 위해서 상기 GARCH(1,1) 모형의 1 모수가 변하지 않을때와 2 모수가 변할때 각각에 대한 데이터를 생성하였고 각 경우를 데이터1과 데이터2로 명명하였다. 두 경우 모두 길이 600인 시계열자료 500개를 생성하였다. 데이터1은 $t = 1$ 600 시점동안 구조적 변화 또는 변화점이 없는 경우이고 데이터2는 600 시점동안 구조적 변화 또는 변화점이 존재하는 경우이다. 먼저 변화점이 없는 경우 세가지 Case에 대해 데이터를 생성하였고 변화점탐지를 분석하였다.

(데이터1) (ω, α, β)가 변하지 않는 길이(t) 600인 데이터 500개(N) 생성

Case1. $(\omega_0, \alpha_0, \beta_0) = (0.1, 0.1, 0.8)$

Case2. $(\omega_0, \alpha_0, \beta_0) = (0.1, 0.1, 0.2)$

Case3. $(\omega_0, \alpha_0, \beta_0) = (0.1, 0.3, 0.6)$



(a) Case1. (0.1, 0.1, 0.8) (b) Case2. (0.1, 0.1, 0.2) (c) Case3. (0.1, 0.3, 0.6)

그림 3.1: 데이터1 사진예시

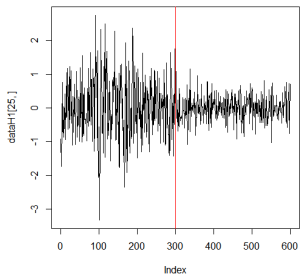
(데이터2) (ω, α, β)가 변하는 길이(t) 600인 데이터 500개(N) 생성

Case1. $(\omega_0, \alpha_0, \beta_0) = (0.1, 0.1, 0.8) \rightarrow (\omega_1, \alpha_1, \beta_1) = (0.1, 0.1, 0.2)$

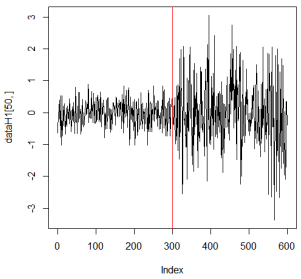
Case2. $(\omega_0, \alpha_0, \beta_0) = (0.1, 0.1, 0.2) \rightarrow (\omega_1, \alpha_1, \beta_1) = (0.1, 0.1, 0.8)$

Case3. $(\omega_0, \alpha_0, \beta_0) = (0.1, 0.3, 0.6) \rightarrow (\omega_1, \alpha_1, \beta_1) = (0.1, 0.1, 0.8)$

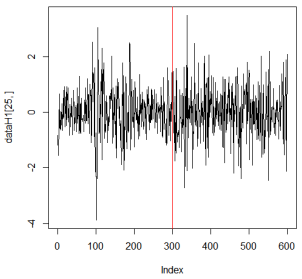
Case4. $(\omega_0, \alpha_0, \beta_0) = (0.1, 0.1, 0.8) \rightarrow (\omega_1, \alpha_1, \beta_1) = (0.1, 0.3, 0.6)$



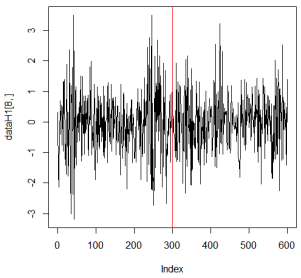
(a) Case1



(b) Case2



(c) Case3



(d) Case4

그림 3.2: 데이터2 사진예시

3.1.2 시뮬레이션 방법

GARCH(1,1) 데이터에 변화점의 존재여부를 확인하기위해 기존연구 방식의 일환으로 2.2.1절에서 통계량을 소개하였다. 해당 통계량을 사용한 변화점탐지방법은 아래 표에서 적용방식1로 표기하였다. 이어서 기존연구 방식이 아닌 새로운 방식의 변화점탐지방법을 만들어보고자 위상적데이터방법론(2.3.1절)과 지역선형임베딩(2.4.1절) 및 벡터자기회귀모형 변화점탐지방법(2.1.2)을 차례로 적용하였다. 이 방법을 적용방식2로 하단의 표메 표기하였다.

표 3.1: 변화점탐지 비교방법

구분	사용방법	순서	관련내용
적용방식1	GARCH 변화점탐지	-	2.2.1
적용방식2	위상적데이터방법	1	2.3.1절
	지역선형임베딩	2	2.4.1절
	VAR 변화점탐지	3	2.1.2절

다음은 적용방식2에 대해 자세히 설명하고자 한다. 적용방식2는 총 4단계(stage)에 나눌 수 있고 각 단계별 대한 설명은 아래와 같이 정리하였다.

Stage1. $GARCH\{x_t\}_{t=1}^{T} \subset \mathbb{R} \;\Rightarrow\; \{X_t\}_{t=1}^{T-2} = \{(x_t, x_{t+1}, x_{t+2})^t\}_{t=1}^{T-2} \subset \mathbb{R}^3$

Stage2. $\{X_t\}_{t=1}^{T-2} \subset \mathbb{R}^3 \Rightarrow TDA\ \{\boldsymbol{\beta}_p^{(s)}\}_{s=1}^{T-w-1} \subset \mathbb{R}^n$

Stage3. $\{\boldsymbol{\beta}_p^{(s)}\}_{s=1}^{T-w-2} \subset \mathbb{R}^m \Rightarrow LLE\ \{y_s\}_{t=1}^{T-w-1} \subset \mathbb{R}^3$

Stage4. $\{y_s\}_{s=1}^{T-w-1} \subset \mathbb{R}^3 \Rightarrow VAR\ change\ point\ detection$ (2.1.2)

▶ Stage1 설명 : GARCH(1,1) 모델을 따르는 데이터들을 1차원 데이터이기 때문에 기하학적 특성을 추출할 수 없다. 위상적데이터방법은 그자체로 고차원 데이터 분석을 위한 방법이기에 1차원 데이터를 다차원 데이터로 변형하는 것이 필요하다. 따라서 $X_1 = (x_1, x_2, x_3)$, $X_2 = (x_2, x_3, x_4)$ $\cdots$와 같이 연속적인 GARCH 데이터를 하나씩 좌표화하여 데이터를 변형한다.

▶ Stage2 설명 : 변수(w)를 지정하여 $\{\tilde{X}_t\}_{t=s}^{s+w-1} = \{X_s, X_{s+1}, \cdots, X_{s+w-1}\}$ $s = 1, \cdots, T-w-1$일때 각각 비아토리스-립스를 구한다. 본 논문에서 Umar Islambekov(2019) 참고하여 변수(w)는 5로 설정했다.
각 s시점에 비아토리스-립스 $VR(\tilde{X}_s, \epsilon)$는 ϵ가 커짐에 따라 연결되는 데이터가 많아지고 점차 커지게 된다. 이를 $0 < \epsilon_1 < \epsilon_2 < \ldots < \epsilon_n$일때 $VR(\tilde{X}_s, \epsilon_1) \subseteq VR(\tilde{X}_s, \epsilon_2) \subseteq \cdots \subseteq VR(\tilde{X}_s, \epsilon_n)$와 같이 표현할 수 있다. 각 $VR(\tilde{X}_s, \epsilon_i)$마다 베티넘버들을 구하여 다차원 베티넘버 벡터인 $\boldsymbol{\beta}_p^{(s)} = (\beta_p^{(s)}(\epsilon_1), \beta_p^{(s)}(\epsilon_2), \cdots, \beta_p^{(s)}(\epsilon_n))$를 계산한다. 베티넘버 $\beta_p^{(s)}$를 측정함에 있어 $p = 0, 1, \ldots, d$로 d-차원 홀을 측정할 수 있으나 본 논문에서는 $p = 0$인 경우만을 고려하였다. 데이터간 거리를 의미하는 ϵ_i은 다음의 $\epsilon_i = 0.01(i-1)$. $i = 1, \cdots n$와 같이 설정하였고 $n = 50$까지를 고려했다.

▶ Stage3 설명 : 지역선형임베딩(LLE)를 적용하는 단계로 Stage2에서 구한 $\boldsymbol{\beta}_p^{(s)} = (\beta_p^{(s)}(\epsilon_1), \beta_p^{(s)}(\epsilon_2), \cdots, \beta_p^{(s)}(\epsilon_n)) \in \mathbb{R}^n$이므로 지역선형임베딩을 적용하여 차원축소시키고자 한다. 다만 지역선형임베딩을 적용할때 k개 주변점 선정을 위해 k를 어떻게 설정할지 결정해야한다. 본 논문에서는 k는 3 또는 5인경우에 한해 분석을 진행했으며 앞서 구한 n차원 베티넘버벡터를 3차원 데이터로 차원축소하였고 이를 $\{y_s\}_{s=1}^{T-w-1}$로 표현했다.

▶ Stage4 설명 : 차원축소된 데이터 $\{y_s\}_{s=1}^{T-w-1} \subset \mathbb{R}^3$에 2.1.2절에서 소개한 벡터자기회귀 모형의 변화점 탐지를 위한 통계량($\hat{T}_n^{\max}$)을 적용하여 변화점 여부를 확인했다.

Remark1 비아토리스-립스의 데이터간 거리를 의미하는 ϵ이 0 0.5까지의 숫자를 가지나, GARCH(1,1)모형의 불안정성으로 데이터간 거리가 상당히 크게 나타날 수 있다. Umar Islambekov(2019)에서도 소개하고 있듯이 Stage2에서 위상적데이터방법(TDA) 적용하기 전 데이터 거리를 [-1, 1]이내로 정규화하였다. 또한 베티넘버 $\{\boldsymbol{\beta}_p^{(t)}\}$에서 p=0인 경우만을 고려했다.

Remark2 분석에 있어 size는 변화점이 없는 경우에 변화점이 있다고 판단하는 비율을 의미하며, power는 변화점이 있는 경우에 변화점이 없다고 판단하는 비율을 의미한다. 금번 연구에서 데이터1는 변화점이 없는 경우로 size는 데이터1에서 통계량이 기각역보다 높게 나타나는 경우의 비율이다. 또한 데이터2는 변화점이 있는 경우로 power는 데이터2에서 통계량이 기각역보다 낮은 경우의 비율이다.

3.1.3 시뮬레이션 결과

이번 절에서는 앞서 설명한 적용방식1, 2에 대한 시뮬레이션한 결과를 설명하고자한다. 변화점이 없는 경우(데이터1) 두 적용방식을 비교하면 표 3.2와 같다. 일반적인 GARCH 변화점탐지방법(적용방식1)은 Case 1에서 0.34, Case 2에서 0.028, Case 3에서 0.484의 Size를 나타내었다. 이는 α와 β의 합이 1에 가까울 수록 불안정해진다는GARCH 모델의 특징이 반영된 결과로 해석할 수 있다. 다만 α와 β의 합이 1보다 훨씬 작은 Case 2에서는 size가 0.05보다 작게 나타남을 확인할 수 있었다. 적용방식 2에서는 Case 1과 Case 2은 동일한 결과인 0.254의 size를 나타내었고 Case 3에서 0.446을 나타내었다. 적용방식 2에서는 GARCH 모델이 불안정하던 아니던 관계없이 유사한 size를 확인할 수 있었다. GARCH 모델이 불안정해질 수록 기존방식(적용방식1)보다 size가 개선되는 효과를 보였다.

표 3.2: 데이터1 적용방식별 Size 비교

구분	$(\omega_0, \alpha_0, \beta_0)$	**적용방식 1 GARCH**	**적용방식 2 TDA+LLE+VAR CPD**
Case1	(0.1, 0.1, 0.8)	0.34	0.254
Case2	(0.1, 0.1, 0.2)	0.028	0.254
Case3	(0.1, 0.3, 0.6)	0.484	0.446

이어서 변화점이 있는 경우(데이터2) 적용방식1, 2를 비교한 결과를 표 3.3에 나타내었다. 적용방식1의 경우에 Case 1와 Case 2에서 Power가 1로 나타났다. Case 3에서는 0.482, Case 4에서는 0.48의 Power를 기록하였다. 반면에 적용방식 2에서 Case 1은 0.99, Case 2에서는 0.988의 Power를 기록하였다. Case 3에서는 0.542, Case 4에서는 0.494를 나타내었다. 결과에서 확인할 수 있듯이 데이터에 구조적 변화가 있는 경우에 적용방식2은 우수한 Power를 나타내었다. 특히 GARCH 모델이 불안정한 상황에서 모수변화(Case3/Case4)가 발생하는 경우에 적용방식2가 적용방식1 보다 높은 power를 기록하였다.

표 3.3: 데이터2 적용방식별 Power 비교

구분	$(\omega_0, \alpha_0, \beta_0)$ $\rightarrow (\omega_1, \alpha_1, \beta_1)$	**적용방식 1** **GARCH**	**적용방식 2** **TDA+LLE+VAR CPD**
Case1	(0.1, 0.1, 0.8) $\rightarrow$ (0.1, 0.1, 0.2)	1	0.99
Case2	(0.1, 0.1, 0.2) $\rightarrow$ (0.1, 0.1, 0.8)	1	0.988
Case3	(0.1, 0.1, 0.8) $\rightarrow$ (0.1, 0.3, 0.6)	0.482	0.542
Case4	(0.1, 0.3, 0.6) $\rightarrow$ (0.1, 0.1, 0.8)	0.48	0.494

적용방식2를 사용함에 있어 여러 모수들을 미리 설정하는 것이 필요했다. 그 중 지역선형임베딩을 적용시 k-주변점을 선정해주어야했는데 k값에 따라 결과값이 상이하게 나타났다. k가 3, 5일때 결과를 표3.4와 표3.5에 포함했으며 주변점 갯수 변할때 지역선형임베딩을 위한 선형관계와 가중치값들이 변하기 때문으로 판단된

다. 구체적으로 살펴보면 먼저 변화점이 없는 경우(데이터1) 분석한 결과 Case1 이고 k=5일때 size는 0.254으로 나타났고 k=3일때 size인 0.178보다 크게 나타났고 Case 2에서도 유사한 결과가 확인할 수 있었다. Case3에서 k=5일때 size가 0.446이었으며 k=3일때 0.292보다 크게 나타났다.

표 3.4: (데이터1) 적용방식2 모수 k값에 따른 Size 비교

구분 $(\omega_0, \alpha_0, \beta_0)$		**적용방식 2**	
		k=5	**k=3**
Case1	(0.1, 0.1, 0.8)	0.254	0.178
Case2	(0.1, 0.1, 0.2)	0.254	0.178
Case3	(0.1, 0.3, 0.6)	0.446	0.306

변화점이 없는 경우(데이터1) k값이 클때 더 큰 size를 갖는다는 경향을 확인할 수 있었다. 변화점이 있는 경우(데이터2) power에서도 비슷한 결과를 얻을 수 잇었다. 표 3.5의 결과에 따르면 Case 1인 경우 k=5일때 0.99로 높은 power를 보이나, k=3일때 0.878로 낮은 power를 나타내었다. Case 2 Case 4에서도 유사한 경향성을 보였다. 종합하면 적용방식 2의 모수 k값이 변할때 size와 power도 크게 변화했다. k값이 변화점 탐지에 영향을 미치기 때문에 지역선형임베딩외에도 다른종류의 차원축소방법도 추후 적용해보는 것이 필요하다고 생각된다.

표 3.5: (데이터2) 적용방식2 모수 k값에 따른 Power 비교

구분	$(\omega_0, \alpha_0, \beta_0)$ $\rightarrow (\omega_1, \alpha_1, \beta_1)$	**적용방식 2**	
		k=5	**k=3**
Case1	(0.1, 0.1, 0.8) $\rightarrow$ (0.1, 0.1, 0.2)	0.99	0.878
Case2	(0.1, 0.1, 0.2) $\rightarrow$ (0.1, 0.1, 0.8)	0.988	0.89
Case3	(0.1, 0.1, 0.8) $\rightarrow$ (0.1, 0.3, 0.6)	0.542	0.332
Case4	(0.1, 0.3, 0.6) $\rightarrow$ (0.1, 0.1, 0.8)	0.494	0.292

제 4 장 결론

금번연구에서는 위상적데이터방법과 다양체분석(지역선형임베딩)방법을 시계열모형에 적용해보고자 여러 선행연구들을 활용하였다. 특히 Umar Islambekov (2019)에서 비아토리스-립스와 베티넘버들을 활용하여 시계열데이터로부터 기하학적 정보를 추출 및 변환한 점에 착안하여 이를 GARCH모형에도 적용해보고자 연구를 진행했다.

베티넘버들을 활용해 고차원 벡터를 생성하였고 이에 따라 차원축소방식을 선택해야했으며 다양체분석(지역선형임베딩)방법을 활용했다. GARCH 모형의 변화점 탐지를 위해서는 기존 연구방법들을 활용했으며 특히 Lee & Kim (2022)에서 논의된 통계량을 적용했다. 그리고 차원축소된 베티넘버들의 변화점 탐지를 위해서 Lee (2020)에서 논의한 벡터자기회귀(VAR) 모형의 변화점 탐지를 위한 통계량을 활용했다.

다만 기존 변화점탐지 방법보다 위상적데이터방법이 우수한 측면도 있었으나 높은 size라는 방법론의 한계가 관측됨에 따라 이를 보완하여 발전시키는 것이 필요하다고 판단된다. 또한 모수들을 사전에 설정해야하는데 이에 대한 기준이 있는지 보완이 필요하다고 판단된다. 결론적으로는 모형의 구조적변이 또는 변화점탐지를 하는데 새로운 방법들을 적용해보고 결과를 도출해보았다는 점에 있어서 의미있는 연구였다고 판단한다.

참고문헌

Benyamin Ghojogh, and Ali Ghodsi, Locally Linear Embedding and its Variants: Tutorial and Survey. *2020, ArXiv Preprint arXiv:2011.10925*, 2-3.

Henri Riihimäki, and Wojciech Chachólski. A topological data analysis based classification method for multiple measurements. *2020 BMC Bioinformatics* **21**, 3-5.

Frédéric Chazal, Bertrand Michel, An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. *2021, ArXiv Preprint arXiv:1710.04019v2*, 2-21.

Kanika Mehta1, Gajendra Tyagi. Modified Locally Linear Embedding with Affine Transformation. *2017, National Academy Science Letters* **40**, 189–196.

Lee S, Change point test for time series via topological data analysis. *2022, Unpublished manuscript*

Mengsen Zhang,Scott Kelso. Topological portraits of multiscale coordination dynamics. *2020, Journal of Neuroscience Methods* **40**, 189–196.

Sangyeol Lee, Location and scale-based CUSUM test with application to autoregressive models. *2020, Journal of Statistical Computation and Simulation.*, 5–8.

Sangyeol Lee, and Chang Kyeom Kim. Test for conditional quantile change in GARCH models. *2022, Journal of the Korean Statistical Society*, 480–499.

Sangjo Lee, and Sangyeol Lee, Statistical Process Monitoring for Vector Autoregressive Time Series. *2022, Journal of Statistical Computation and Simulation.*, 5–8.

Shafie Gholizadeh, ArminSeyeditabari, Topological Signature of 19th Century Novelists:Persistent Homology in Text Mining *2018 Big Data and Cognitive Computing*

Umar Islambekov. and Monisha Yuvaraj. HARNESSING THE POWER OF TOPOLOGICAL DATA ANALYSIS TO DETECT CHANGE POINTS IN TIME SERIES. *2019, ArXiv Preprint arXiv:1910.12939*, 2-5.

Abstract

In this paper, the topological data analysis to detect change point in GARCH(1,1) data will be introduced. Conventional change point method for GARCH(1,1) will be discussed first. The conventional statistic using a brownian bridge distribution which is discussed in Lee (2020) is useful at detecting change point of stable GARCH data ($\alpha + \beta$ is much less than 1).

On the other hand, newly introduced change point detection method using topological data analysis and Manifold learning has higher power than conventional method for analyzing unstable data from GARCH(1,1).

However, both the classical method and the newly introduced method have limitations because they have high false alarm when there is no change point with unstable data ($\alpha + \beta$ is close to 1).

The newly introduced method is referred to Umar Islambekov (2019) by utilizing betti numbers to extract geometric features from time series data. In addition, the dimension of time series betti number data is reduced by Locally Linear Embedding(Manifold learning). Lastly, a change point method for VAR model from the papers Lee & Kim(2022) is applied to find a change point of the reduced time series data.

Keywords: Time series, Topological Data Analysis, Manifold learning, CUSUM

test.

Student Number: 2020-29046