The Korean Journal of Applied Statistics
2022, Vol. 35, No. 1, 63–75

DOI: https://doi.org/10.5351/KJAS.2022.35.1.063
Print ISSN1225-066X / Online ISSN 2383-5818

High-dimensional change point detection using MOSUM-based sparse projection

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

[a]Department of Statistics, Sungkyunkwan University

Abstract

This paper proposes the so-called MOSUM-based sparse projection method for change points detection in high-dimensional time series. Our method is inspired by Wang and Samworth (2018), however, our method improves their method in two ways. One is to find change points all at once, so it minimizes sequential error. The other is localized so that more robust to the mean changes offsetting each other. We also propose data-driven threshold selection using block wild bootstrap. A comprehensive simulation study shows that our method performs reasonably well in finite samples. We also illustrate our method to stock prices consisting of S&P 500 index, and found four change points in recent 6 years.

Keywords: change point, moving sum (MOSUM), sparse projection, high dimensional time series

1. 서론

주가지수, 기후 및 환경 데이터 등 우리가 일상 생활에서 접하는 대부분의 자료는 균질성(homogenity)을 가지기 힘들다. 이는 경제 정책의 변화, 환율 및 이자율의 변화, 기후 변화, 외부 요인 등의 수 많은 이유에 의해서 자료를 생성하는 메커니즘이 변화하기 때문이며, 따라서 변화점을 찾는 것은 매우 오랫동안 많은 연구자들의 주요 관심사였다. 본 논문은 빅데이터의 출현 이후 더욱 더 많은 변화를 보이는 고차원 시계열 자료에서의 변화점 추정을 연구하고자 한다.

이러한 변화점들을 정확하게 추정하는 것은 자료에 대한 올바른 이해를 도울 뿐만 아니라, 변화점을 통해 구분된(segmented) 부분 자료를 모델링 함으로써 보다 정확한 예측 모형을 세우는 등의 중요한 역할을 한다. 하지만, 고차원 자료의 경우 모든 차원에서 변화를 보이는 것이 아니라 일부 차원에서만 변화를 보일 수도 있어 이에 대한 특별한 고려가 필요하다. 예를 들어 주가의 경우 여러 개의 섹터로 나뉘어 있는데, 어떤 충격에 의해서 특별한 섹터의 주식들만 영향을 받고 나머지 섹터의 경우 변화를 보이지 않을 수 있다. 그 한 예로 은행에 대한 새로운 규제는 아마도 금융관련 회사 주식들의 주가에 민감하게 영향을 줄 것이다. 이렇듯 차원의 성근성(sparsity)을 고려한 변화점 탐지가 필요하다.

일변량 자료에서 변화점을 추정하는 전통적이면서 가장 널리 쓰이는 방법은 cumulative sum (CUSUM)이라 불리우는 자료의 변환이다. CUSUM 변환은 변화점 근처에서 뾰족함(peakedness)을 더욱 강조하기에 변화점을 쉽게 찾을 수 있도록 도와준다. 고차원 시계열 자료의 경우 CUSUM 변환을 각 차원에 대해서 적용한 뒤 합(sum)이나 최댓값(maximum)등을 사용하게 된다. 최근 Wang과 Samworth (2018)는 성근성을 가지는

This work was supported by the Basic Science Research Program from the National Research Foundation of Korea(NRF-2019R1F1A1057104).

[1] Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03023, Korea. E-mail: crbaek@skku.edu

Published 28 February 2022 / journal homepage: http://kjas.or.kr

고차원 시계열에 대해서 차원 축소(dimension reduction)을 통해 변화점을 찾는 새로운 방법을 제시하였다. 먼저 각 차원에 대해서 CUSUM 변환을 통해 변화점에 더 민감하게 반응하도록 자료를 변환한 다음에 이를 일차원의 자료를 압축한 뒤 얻어진 시그널을 통해 변화점을 찾는 것이다. 자료의 압축은 내적, 즉 프로젝션으로 이해할 수 있다. Wang과 Samworth (2018)는 첫 번째 가장 큰 평균 변화의 경우 최적의 프로젝션 방향은 CUSUM 변환 자료의 특이값분해(singular value decomposition)에서 첫 번째 좌특이벡터(left singular vector)가 됨을 밝혔다. 이 후 복수의 변화점을 추정하기 위하여 Fryzlewicz (2014)의 wild binary segmentation (WBS) 기법을 적용한다. 이는 프로젝션 된 CUSUM 벡터 중 최댓값을 갖는 변화점 위치를 찾은 후, 해당 변화점 기준으로 시퀀스를 분할하여 각 각각 하위시퀀스에서의 변화점을 동일한 방식으로 찾아내는 방법으로, 이를 반복하여 다중 변화점을 추정하는 방식이다.

Wang과 Samworth (2018)의 방법론은 특히 성근성이 있는 고차원 모델에서의 다중 변화점을 찾아낼 때 탁월한 성능을 보이나, WBS방식을 사용하여 복수의 가설검정을 수행함으로써 순차적 오류(sequential error)를 내포할 가능성이 있다. 이러한 순차적 오류를 최소화 하는 방법으로 단일 임계값으로 복수의 변화점을 한 번에 찾아내는 Eichinger와 Kirch (2018)가 제안한 moving sum (MOSUM)이 대안이 될 수 있다. 또한, CUSUM의 경우 소위 평균의 효과가 상쇄되는 경우에 변화점을 잘 찾지 못하나 MOSUM은 국지화된 값으로 이러한 평균 변화도 잘 찾아낼 수 있다.

본 논문에서는 Wang과 Samworth (2018)의 성근 프로젝션 기법을 기반으로, MOSUM 통계량과 block wild bootstrap (BWB)을 이용한 단일 임계값을 통하여 복수의 변화점을 추정하고자 한다. 이 추정의 과정을 2장에서 상세히 기술하였으며, 3장에서는 MOSUM 성근 프로젝션을 이용한 방법론을 모의 실험을 통해 그 성능을 확인하였다. 또한 4장에서는 실증자료분석을 시행하여 현실의 데이터에서도 우리가 제안한 방법이 잘 작동하는지 재확인 하였다. 마지막으로 5장에서 연구 결과를 요약하고 향후 연구에 대해 논의하였다.

2. MOSUM 기반의 다중 변화점 추정

고차원 시계열데이터에서의 다중 변화점을 추정하기 위하여, p차원의 관측값 벡터 X_t, $t = 1, \ldots, T$라 하자. 추정 대상 모델은 평균에 오차가 더해진 모형으로

$$X_t = \mu_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Sigma),$$

로 나타낼 수 있다. 이 때 평균의 변화점 위치인 k_j는 $k_0 = 0 < k_1 < \cdots < k_N < T =: k_{N+1}$로 정의하며 추정된 변화점의 위치는 $\hat{k}_j$, 추정된 변화점의 개수는 $\widehat{N}$이라 하자. 즉, 우리는 전체 시간 T에 대해서 평균의 경우 $N+1$개의 구간의 상수로 이루어 졌음을 다음과 같이 나타낼 수 있다.

$$\mu_{k_i+1} = \cdots = \mu_{k_{i+1}}, \quad 0 \le i \le N. \tag{2.1}$$

본 논문은 MOSUM 통계량을 이용하여 변화점의 위치 및 개수를 추정하는 방법을 제안한다.

Wang과 Samworth (2018)는 변화점 검정에 있어서 가장 널리 쓰이는 CUSUM 통계량에 근거하여 변화점을 추정하는 방법을 제시하였다. 좀 더 구체적으로 설명하기 위해 모형(2.1)을 행렬로 표현하면

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{W},$$

$\mathbf{X} = (X_1, \ldots, X_T)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$, $\mathbf{W} = (\epsilon_1, \ldots, \epsilon_T)$인 $p \times T$ 행렬로 나타낼 수 있다. 또한 성근성을 가정하기 위해서 모든 차원에서 변화하는 것이 아니라 $q \in \{1, \ldots, p\}$에서만 평균의 변화가 일어난다고 가정하자. 이 때의 q는 변화가 일어나는 차원의 개수를 나타내는 인덱스이다. Wang과 Samworth (2018)는 먼저 주어진

데이터에 대해서 CUSUM 변환을 한다.

$$\mathbf{T} = \mathcal{T}^{\text{CUSUM}}(\mathbf{X}),$$
$$\mathcal{T}^{\text{CUSUM}}(\mathbf{X}) := \sqrt{\frac{i(T-i)}{T}}\left(\frac{1}{T-i}\sum_{t=i+1}^{T} X_t - \frac{1}{i}\sum_{t=1}^{i} X_t\right). \tag{2.2}$$

이는 CUSUM 변환이 변화점에 대해서 뚜렷하게 뾰족함(peakedness)를 주어 변화점 탐지에 있어서 매우 유용하기 때문이다. 그 이후에는 고차원 자료를 일차원으로 프로젝션을 통해서 요약하는 단계를 거친다. 프로젝션 방향은 자료를 가장 잘 압축할 수 있는 방향으로 $\mathbf{T}$의 첫 번째 좌특이벡터(left singular vector)로 다음과 같이 표현할 수 있다.

$$\hat{z} \in \arg\max M \in \mathcal{M}\langle M, \mathbf{T}\rangle, \tag{2.3}$$

$\mathcal{M} = \{M \in R^{(p-1)\times(T-1)} : \|M\|_* \le 1,\ \text{rank}(M) = 1, M \text{ has at most } k \text{ non-zero rows}\}$, $\|M\|_*$는 핵노름(nuclear norm)이다. 하지만 이 프로젝션에 대한 수치해는 NP-hard에 해당하여 현실적으로 구하기 어려우므로 볼록화를 통해 완화한 공간에서의 해를 구한다 (Wang과 Samworth, 2018). 여기에서 NP-hard란 컴퓨터과학의 계산복잡도 이론에서 사용되는 용어로서, 다항시간 내에 답을 낼 수 없는 문제를 지칭한다.

$$\widehat{M} = \arg\max M \in \mathcal{S}_1\{\langle M, \mathbf{T}\rangle - \lambda\|M\|_1\}, \quad \lambda > 0, \tag{2.4}$$

$\mathcal{S}_1 = \{M \in R^{(p-1)\times(T-1)} : \|M\|_* \le 1\}$, $\|M\|_*$는 핵노름(nuclear norm)이다. 이와 같이 CUSUM 변환을 한 $\mathbf{T}$에서가 아니라, 이에 대한 볼록근사 행렬인 M에서의 첫 번째 좌특이벡터가 최적의 프로젝션의 방향이 되는데 이를 $\mathbf{z}$라고 부르자. 그렇다면 첫 번째 변화점은 사영된 벡터의 최댓값이 임계점을 넘을 때로

$$t_{\max} := \max_{1\le j<T} \left|\mathbf{z}'\mathbf{T}_j\right| \ge \eta,$$

이다. 이 때 $\mathbf{T}_j$는 $\mathbf{T}$의 j번 째 열이고, 변화점의 추정값은 최댓값을 가지게 되는 시점이다. 따라서 변화점 탐지의 오류를 줄이고 파워를 높이기 위해서는 임계점을 잘 설정하는 것이 중요하며 Wang과 Samworth (2018)는 이론적인 계산을 토대로 η를 $2\sigma\sqrt{\log(p\log T)}$로 제시하였는데, 이는 볼록화를 통한 완화과정에서 고려되는 페널티 λ와 같은 값이다. 두 개 이상의 변화점이 존재할 경우에는 Fryzlewicz (2014)의 WBS를 통해 위의 설명한 방법을 반복적(recursive)으로 적용함으로써 구할 수 있다.

WBS는 다수의 변화점 탐지에 대표적으로 쓰이는 binary segmentation (BS)을 개선한 방법이다. 먼저 첫 번째 변화점에 대해서 전후로 표본을 나눈 다음에 각 부분 표본에 대해서 임의의 크기의 구간을 여러 번 무작위로 뽑아서 생성된 $t_{\max}$ 값들의 최댓값을 임계점 η와 비교하여 임계점보다 클 경우 새로운 변화점으로 추가하는 알고리즘이다. 이 방법은 CUSUM 통계량이 가지는 전역적 성질에 대한 약점, 즉 평균이 상쇄되는 방향으로 존재할 경우 변화점을 탐지하지 못하는 단점을 보완한 방법이다. 예를 들어 $[(i/4)n]$, $i = 1, 2, 3$에 변화점이 있고 평균이 $\mu_1 = -1, \mu_2 = 1, \mu_3 = 0.5, \mu_4 = -0.5$과 같이 주어진다면 중간 $[n/2]$에서의 CUSUM 값은 0으로 평균의 변화가 상쇄되어 변화점을 찾기 어려워진다. WBS는 무작위로 구간을 뽑기때문에 평균이 상쇄되는 경우를 줄여 다중 변화점 탐지에 있어서 유용한 방법이다.

본 논문에서는 CUSUM이 가지는 다중 변화점의 문제뿐만 아니라 반복적인 방법의 적용으로 인한 검정의 크기 왜곡(size distortion)을 해결하고자 MOSUM 통계량을 사용하여 다중 변화점을 찾는 방법을 제안하고자 한다. 띠넓이(bandwidth) G에 대해서 k시점에서의 MOSUM 통계량은

$$\mathcal{T}^{\text{MOSUM}}(\mathbf{X}) := \frac{1}{\sqrt{2G}}\left(\sum_{t=k+1}^{k+G} X_t - \sum_{t=k-G+1}^{k} X_t\right) \tag{2.5}$$

과 같이 정의된다. 수식 (2.5)에서 살펴 보듯이 MOSUM 통계량은 사실 $2G$ 윈도우에 들어가 있는 데이터 $\{X_{k-G+1}, \ldots, X_{k+G}\}$에 대해서 CUSUM 통계량을 구한 것으로 이해할 수 있어, CUSUM 통계량이 가지는 좋은 성질을 국소적(localization)으로 살필 수 있는 장점이 존재한다.
MOSUM을 이용한 고차원 시계열의 변화점 탐지 알고리즘은 다음과 같다.

STEP 1 주어진 자료 $\mathbf{X}$에 대해서 띠넓이 G를 가지는 MOSUM 변환을 한다.

$$\mathbf{U} = \mathcal{T}^{\text{MOSUM}}(\mathbf{X})$$

STEP 2 Wang과 Samworth (2018)의 성근 프로젝션을 통해서 구한 사영된 벡터를 구한다. 즉, $\widetilde{M}$에서

$$\widetilde{M} = \arg\max M \in \mathcal{S}_1\{\langle M, \mathbf{U}\rangle - \lambda\|M\|_1\}, \quad \lambda > 0, \tag{2.6}$$

의 첫 번째 좌특이벡터 $\widetilde{\mathbf{z}}$ 을 구하면 된다. 검정통계량 값은

$$\mathbf{u}_j = |\widetilde{\mathbf{z}}'\mathbf{U}_j|, \quad j = G, \ldots, T-G,$$

으로 주어진다.

STEP 3 BWB를 통해서 임계값 $\widetilde{\eta}$를 구한다. 먼저 자료의 평균을 조정한 값에 대해서 평균이 0이고 분산이 1인 IID 확률변수(multiplier)를 곱하여 얻어진다. 다시 말해, 먼저 평균이 0이고 분산이 1인 IID 확률변수를 multiplier로 하고, 이 multiplier를 자료의 오차항에 곱하여 부트스트랩 표본이 얻어진다. 그리고 이 부트스트랩 표본이 임계값의 계산에 사용된다. 즉,

$$e_t = X_t - \frac{1}{T}\sum_{t=1}^{T} X_t,$$

$$X_t^* = \frac{1}{T}\sum_{t=1}^{T} X_t + w_t e_t.$$

곱해지는 확률변수 w_t 의 경우 블록크기 b_T에 대해서 같은 블록에 있으면 같은 확률변수 값을 공유하게 된다. 예를 들어 블록 크기가 5인 경우에 $w_1 = w_2 = \cdots = w_5$, $w_6 = \cdots = w_{10}$ 이다. BWB로 구해진 부트스트랩 표본 $\mathbf{X}^*$에 대해서 검정통계량 $\mathbf{u}_j^*$을 부트스트랩 반복 수만큼 얻어낸다. 임계값은 $\mathbf{u}_j^*$의 $100(1-\alpha)\%$ 분위수로 결정이 된다.

STEP 4 변화점은 임계값 $\widetilde{\eta}$을 넘은 $\mathbf{u}_j$의 국소 최댓값들이다.

위의 MOSUM을 이용한 변화점 탐지 과정을 Figure 1을 통해 예로 나타냈다. 검은색 실선은 검정통계량 $\mathbf{u}_j$를 나타내며 빨간색 선은 임계값을 나타낸다. 임계점을 넘은 값들 중에서 국소적 최댓값인 100, 250, 400을 MOSUM은 변화점으로 찾게 된다. 또한 Figure 2는 $t = 250$에서 평균이 상쇄되어 CUSUM 통계량의 경우 0을 주는 경우에 대해서도 MOSUM에 기반한 방법론은 문제 없이 변화점을 잘 찾음을 살펴볼 수 있다.

3. 모의 실험

본 장에서는 고차원 시계열의 다중 변화점 탐지에 있어 MOSUM을 이용한 방법론의 성능을 모의 실험을 통해서 살펴보고자 한다. MOSUM을 적용하기 위해서는 몇 가지 조율 모수(tuning parameter)가 등장한다. 가장 중요한 모수는 띠넓이 G로 너무 큰 띠넓이는 변화점을 탐지하는데 민첩하지 못하고, 또 너무 작은 띠넓이는

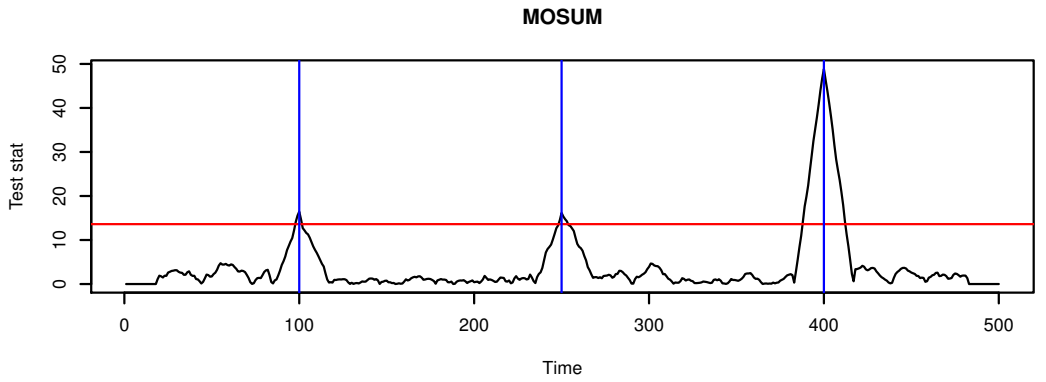

Figure 1: *MOSUM change point detection.*

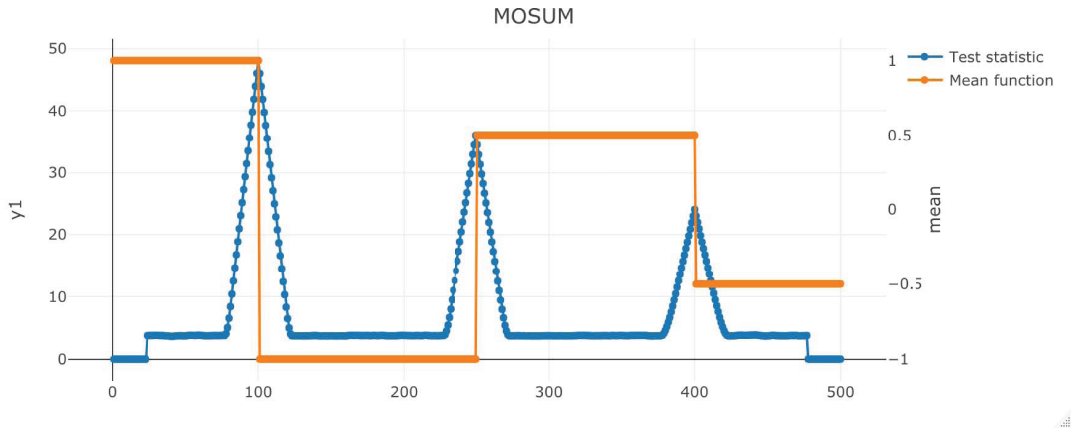

Figure 2: *MOSUM for cancelled mean at $t = 250$ case.*

변화점을 과대 추정하는 문제를 유발한다. 따라서 본 논문에서는 데이터 의존적인 Andrews (1991)의 방법을 이용하여 띠넓이를 추정하였다. 먼저 각 차원에 대해서 AR(1)모형을 적합 시켜 얻은 $\hat{\rho}_i$에 대해서

$$g_i = 1.147\left(\frac{4T\hat{\rho}_i^2}{\left(1-\hat{\rho}_i^2\right)^2}\right)^{\frac{1}{3}}$$

을 구한 다음 계산한 최종 띠넓이는

$$G = \left\lfloor \max\left(p^{-1}\sum_{i=1}^{p} g_i, K\right)\right\rfloor \tag{3.1}$$

이며 $K = 100$으로 설정하였다. K는 $\hat{\rho}_i \approx 1$인 경우에 발산하는 경우를 막기위해 설정한 값이다. 여기에서 $[x]$는 x와 가까운 정수를 찾는 함수이다. 또다른 조율 모수는 BWB를 적용할 때 필요한 블록 사이즈 b_T이다. Lee와 Baek (2019)의 연구 결과를 참조하여,

$$b_T = \sqrt{G}$$

Table 1: Size of MOSUM with $T = 500$ according to sparsity level

	Dimension	Sparsity				
		1	0.8	0.5	0.1	0.05
IID	$T = 500,\ p = 50$	0.05	0.02	0.03	0.04	0.03
	$T = 500,\ p = 100$	0.06	0.04	0.05	0.08	0.05
	$T = 500,\ p = 200$	0.05	0.04	0.05	0.06	0.03
AR(0.3)	$T = 500,\ p = 50$	0	0	0	0.02	0
	$T = 500,\ p = 100$	0	0	0	0	0
	$T = 500,\ p = 200$	0	0	0	0	0
AR(0.5)	$T = 500,\ p = 50$	0.01	0.02	0	0	0
	$T = 500,\ p = 100$	0	0	0	0	0
	$T = 500,\ p = 200$	0	0	0	0	0

를 사용하였다. 성근 프로젝션은 Wang과 Samworth (2018)가 제공한 R패키지 InspectChangepoint를 이용하여 구하였으며, 볼록화를 통한 완화(2.4)에 필요한 조율 모수는 패키지 기본값을 사용하였다.

시뮬레이션에 사용한 data generating process (DGP)는 다음과 같다. 샘플 크기는 $T = 500$으로 고정하였고, 고차원에서의 성능을 비교하기 위하여 각 DGP에서의 차원 변화로 $p = 50, 100, 200$ 로 세 경우를 고려하였으며, 변화점의 위치는 등간격으로 하였다. 먼저 평균 변화의 경우 변화점으로 구분된 각 구간에 대해서 다음의 평균값을 사용하였다.

DGP1 $m_1 = 0, m_2 = 0.8, m_3 = 1.6, m_4 = -0.8$

DGP2 $m_1 = 0, m_2 = 0.6, m_3 = 1.2, m_4 = -0.6$

DGP3 $m_1 = 0, m_2 = 0.8, m_3 = 1.6, m_4 = -0.8, m_5 = -1.6$

예를 들어 DGP1의 경우 변화점은 각 100, 250, 400에서 일어나고, 첫 번째 구간에서의 평균은 $\mu_1 = \cdots = \mu_{125} = 0$, 두 번째 구간에서의 평균은 $\mu_{126} = \cdots, \mu_{250} = 0.8$임을 의미한다. 오차항 ϵ_t의 경우 AR(1) 모형

$$\epsilon_t = \rho\epsilon_t + u_t, \quad u_t \sim i.i.d.\ \mathcal{N}\left(0, 0.25\left(1-\rho^2\right)\right)$$

을 사용하여 오차항의 평균은 0이고 분산이 0.5인 AR(1) 모형이다. 만약 $\rho = 0$ 이면, 이는 단순 IID 정규분포를 나타낸다. 또한 성근성이 있는 고차원시계열 데이터의 변화점 추정을 살펴보기 위해서 평균이 변화하는 차원 q를 p차원에서 랜덤 하게 뽑아 그 성능을 살펴보았다. 성근성은 q/p 로 sparsity = 1 인 경우는 모든 차원에서 평균이 변화하는 모형이며, 만약 sparsity = 0.1 이면 10%의 차원에서만 평균이 변화하고 나머지 90%의 차원에서는 평균 변화가 일어나지 않은 모형이다. 각 100번의 반복을 통해서 결과를 산출하였으며 BWB의 반복수도 101번을 사용하였다.

성능을 알아보기 위해 사용한 통계량은 다음과 같다. 변화점의 개수에 대한 빈도표(frequency table)을 통해 변화점의 개수에 대한 분포를 살펴보았고, 각 변화점을 얼마나 잘 추정하는지 살펴보기 위해서 Cho (2016)에서 사용한 위치 정확도(location accuracy, %)를 계산하였다. 정확도는 전체 몬테카를로 반복 중에서, 추정된 변화점 k의 위치가 실제 변화점의 위치를 기준으로 일정 구간에 있는 경우의 비율로, 가까움의 척도는

$$\left|\hat{k}-k\right| < \log T$$

로 정의된다.

Table 1은 평균 변화가 없는 경우에 제안한 방법론이 변화점을 찾는지를 살펴보는 검정의 크기(size)를 나타낸다. 즉, 평균 변화가 없는 모형에서 MOSUM 방법이 적어도 하나 이상의 변화점을 찾는 경험적 빈도를 찾았다. 임계값은 부트스트랩 표본의 95% 분위 수를 사용하여 5%의 유의 수준을 기대한다. 오차항을 IID

Table 2: Performance of MOSUM for DGP1 and DGP2 with sparsity = 1

	Dimension	$\widehat{\mathbf{N}}$					Location accuracy		
		0	**1**	**2**	**3**	**> 3**	k_1	k_2	k_3
DGP 1									
IID	$T = 500, p = 50$	-	100	-	-	-	-	-	100
	$T = 500, p = 100$	-	100	-	-	-	-	-	100
	$T = 500, p = 200$	-	100	-	-	-	-	-	100
AR(.3)	$T = 500, p = 50$	-	-	-	100	-	100	100	100
	$T = 500, p = 100$	-	-	-	100	-	100	100	100
	$T = 500, p = 200$	-	-	-	100	-	100	100	100
AR(.5)	$T = 500, p = 50$	-	-	-	100	-	100	100	100
	$T = 500, p = 100$	-	-	-	100	-	100	100	100
	$T = 500, p = 200$	-	-	-	100	-	100	100	100
DGP 2									
IID	$T = 500, p = 50$	-	100	-	-	-	-	-	100
	$T = 500, p = 100$	-	100	-	-	-	-	-	100
	$T = 500, p = 200$	-	100	-	-	-	-	-	100
AR(.3)	$T = 500, p = 50$	-	-	9	91	-	95	96	100
	$T = 500, p = 100$	-	-	-	100	-	100	100	100
	$T = 500, p = 200$	-	-	-	100	-	100	100	100
AR(.5)	$T = 500, p = 50$	-	-	1	99	-	99	100	100
	$T = 500, p = 100$	-	-	-	100	-	100	100	100
	$T = 500, p = 200$	-	-	-	100	-	100	100	100

정규분포로 사용한 경우에 차원과 sparsity에 상관없이 대략 5% 유의 수준을 지킴을 확인할 수 있다. 하지만, 오차항이 상관관계를 가질 경우에는 차원이 높아질수록 보수적인 결과를 줌을 확인할 수 있다. 본 논문에서 결과를 보고하지는 않았지만 Wang과 Samworth (2018)의 경우 IID 오차항이 아닌 경우 사이즈 왜곡이 심하게 나타남을 관찰하였다. 하지만 Wang과 Samworth (2018) 역시 많은 조율 모수에 의존하여 이를 조절한다면 사이즈 왜곡을 줄일 수 있을 것이라 기대한다. 적어도 우리가 제안한 MOSUM에 기반한 방법 및 조율 모수의 선택이 변화점이 없는 경우에 적절한 사이즈를 유지함을 확인할 수 있었다.

Table 2는 sparsity = 1로 동일한 성근성 수준에서 DGP1과 DGP2에서 추정의 성능을 빈도표와 위치정확도로 요약하였다. DGP1과 DGP2는 모두 세 개의 변화점을 가지고 그 위치는 $k_1 = 100, k_2 = 250, k_3 = 400$ 이다. DGP1과 DGP2의 차이는 DGP2에서의 평균 변화(jump size)가 DGP1 보다 작다. 우선 AR모형의 오차를 가지는 경우에 DGP1은 변화점을 완벽하게 찾았으나, DGP2의 경우 또 차원이 작은 경우 약간 적게 추정하는 경향이 있었다. 이는 평균의 변화폭이 더 작기 때문에 당연한 결과로 이해할 수 있다. 또한 차원이 증가함에 따라 오히려 성능이 좋아지는 것은 sparsity가 1이기에, 즉 모든 차원에서 변화하기 때문에 얻어지는 일종의 차원의 축복(blessing of dimensionality)로 이해할 수 있을 것이다. 하지만 IID 평균정규분포를 사용한 경우에는 실제 변화점보다 더 적게 추정하는 것을 살펴볼 수 있다. 또 한 개의 변화점은 k_3로 평균의 변화가 가장 큰 경우를 찾음을 확인할 수 있다. 하지만 Figure 3에서 살펴 보듯이 MOSUM을 이용한 통계량이 참 변화점 근방에서는 주변과 다르게 뾰족함(peakedness)을 유지하지만, 임계값에 따라서 최종 선택되지 못함을 관찰할 수 있다. 보다 정확한 임계값에 대한 후속 연구가 필요하다고 본다.

Table 3은 DGP3에서의 MOSUM 성근 프로젝션 방법의 성능을 요약하였다. DGP3 에서는 각각 $k_1 = 100$, $k_2 = 200$, $k_3 = 300$, $k_4 = 400$에서 변화점이 위치하고 있다. DGP3의 결과를 살펴보면 오차항에 의존성이

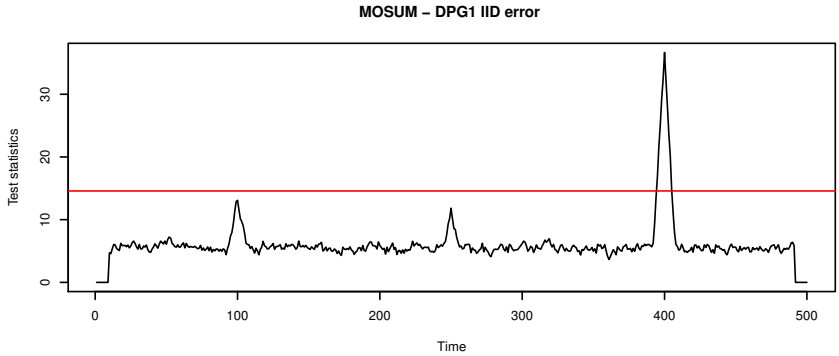

Figure 3: *MOSUM for DGP1 with IID errors.*

Table 3: Performance of MOSUM for DGP3 with sparsity = 1

DGP 3	**Dimension**	$\widehat{\mathbf{N}}$					**Location accuracy**			
		$<$ **2**	**2**	**3**	**4**	$>$ **4**	k_1	k_2	k_3	k_4
IID	$T = 500,\ p = 50$	-	-	-	97	3	100	100	100	100
	$T = 500,\ p = 100$	-	-	-	100	-	100	100	100	100
	$T = 500,\ p = 200$	-	-	-	100	-	100	100	100	100
AR(0.3)	$T = 500,\ p = 50$	-	-	11	89	-	96	96	100	97
	$T = 500,\ p = 100$	-	-	1	99	-	99	100	100	100
	$T = 500,\ p = 200$	-	-	-	100	-	100	100	100	100
AR(0.5)	$T = 500,\ p = 50$	-	6	64	30	-	81	71	100	72
	$T = 500,\ p = 100$	-	-	35	65	-	87	90	100	88
	$T = 500,\ p = 200$	-	-	3	97	-	99	99	100	99

강해질수록 성능이 점차 떨어짐을 볼 수 있다. 또 차원이 증가할수록 결과가 좋아짐도 살펴볼 수 있으며, 가장 좋은 위치정확성을 보이는 위치는 k_3로 평균의 변화가 가장 큰 경우이다. 또한 위치정확성을 살펴보면, 예를 들어 IID, $p = 50$의 경우 변화점의 개수는 약 세 번 정도 과대추정하지만, 위치정확도는 모두 100%임을 볼 수 있다. 이는 곧 MOSUM 기반 성근 프로젝션 방법이 과도한 변화점을 찾지만 실제 변화점 근처에서 찾는 것을 의미함을 볼 수 있어 정확도 면에서는 좋은 추정의 성능을 보인다고 말할 수 있다.

지금까지는 Table 2과 Table 3를 통해서 전반적으로 성근성이 존재하지 않는 환경, sparsity = 1 에서 변화점을 찾는 데 MOSUM 성근 프로젝션 방법이 좋은 성능을 나타내는 것을 확인하였다. 이 다음으로는 Table 4를 통하여 성근성이 존재하는 환경, 즉, sparsity = 0.8, sparsity = 0.5, sparsity = 0.1, sparsity = 0.05 일 때 MOSUM 성근 프로젝션의 성능을 확인하고자 한다. 사용한 시뮬레이션 모형은 DGP1에서 AR(0.5) 오차를 가지는 경우로 고정하였다.

Table 4 를 통해 성근도에 따른 성능을 살펴보면, sparsity = 0.8 과 sparsity = 0.5일 때 모든 변화점을 정확하게 찾는 것을 알 수 있다. 성근도가 높다고 할 수 있는 sparsity =0.1의 경우, 저차원인 $p = 50$ 조건에서는 많은 경우 변화점을 실제보다 적게 감지하였으나, 차원이 높아질수록 더 많은 수의 변화점을 찾아내었다. 실제보다도 더 많은 수의 변화점을 찾아낸 경우라 하더라도, 추정된 변화점 중에 정확히 추정된 변화점이 존재하므로 위치정확도가 모두 100%인 것으로 나타난다. 다만 sparsity = 0.05인 조건에서는 변화점을 탐지해내지 못하였으나, 이는 극단적으로 성근도가 높은 경우로, 현실의 데이터에서는 대개 잘 작동할 것으로 기대할 수 있다.

Table 4: Performance of MOSUM for DGP1 depending on sparsity levels

	Dimension	$\widehat{\mathbf{N}}$					Location accuracy		
		0	**1**	**2**	**3**	**> 3**	k_1	k_2	k_3
Sparsity=0.8	$T = 500, p = 50$	-	-	-	100	-	100	100	100
	$T = 500,\ p = 100$	-	-	-	100	-	100	100	100
	$T = 500,\ p = 200$	-	-	-	100	-	100	100	100
Sparsity=0.5	$T = 500,\ p = 50$	-	-	-	100	-	100	100	100
	$T = 500,\ p = 100$	-	-	-	100	-	100	100	100
	$T = 500,\ p = 200$	-	-	-	100	-	100	100	100
Sparsity=0.1	$T = 500,\ p = 50$	-	28	57	15	-	41	46	100
	$T = 500,\ p = 100$	-	-	-	94	6	100	100	100
	$T = 500,\ p = 200$	-	-	-	57	43	100	100	100
Sparsity=0.05	$T = 500,\ p = 50$	26	74	-	-	-	-	-	74
	$T = 500,\ p = 100$	70	30	-	-	-	-	-	30
	$T = 500,\ p = 200$	3	97	-	-	-	-	-	97

Table 5: Performance of MOSUM for DGP3 according to bandwidth (sparsity = 1)

	Dimension	Bandwidth	$\widehat{\mathbf{N}}$					Location accuracy			
			< 2	**2**	**3**	**4**	**> 4**	**k1**	**k2**	**k3**	**k4**
DGP 3											
Bandwidth=5	$T = 500, p = 50$	5	13	30	38	17	2	53	52	100	55
	$T = 500, p = 100$	5	89	10	1	-	-	2	4	100	6
	$T = 500, p = 200$	5	45	45	9	1	-	18	28	100	20
Bandwidth=15	$T = 500, p = 50$	15	-	-	-	100	-	100	100	100	100
	$T = 500, p = 100$	15	-	-	-	100	-	100	100	100	100
	$T = 500, p = 200$	15	-	-	-	100	-	100	100	100	100
Bandwidth=30	$T = 500, p = 50$	30	-	-	4	96	-	99	99	100	98
	$T = 500, p = 100$	30	-	-	1	99	-	100	100	100	99
	$T = 500, p = 200$	30	-	-	-	100	-	100	100	100	100
Bandwidth=50	$T = 500, p = 50$	50	65	35	-	-	-	14	10	100	11
	$T = 500, p = 100$	50	85	15	-	-	-	3	9	100	3
	$T = 500, p = 200$	50	96	4	-	-	-	2	1	100	1
Data-adaptive Bandwidth	$T = 500, p = 50$	12	-	-	-	99	1	100	100	100	100
	$T = 500, p = 100$	12	-	-	-	99	1	100	100	100	100
	$T = 500, p = 200$	12	-	-	-	100	-	100	100	100	100

마지막으로 띠넓이 선택의 중요성에 대해서 간단한 모의 연구를 통해 검토해 보고자 한다. DGP3의 IID 세팅을 기준으로 모의실험 하였으며, 그 결과는 Table 5 에 정리하였다. 각각 띠넓이 5, 15, 30, 50 으로 고정한 것과, 본 논문에서 제안한 자료 의존적(data adaptive)인 띠넓이 (3.1)를 사용한 조건일 때의 성능을 비교하였다. 띠넓이 값이 5 로 data-adaptive한 방식에 비해 작을 때는 추정의 성능이 떨어지지만, 띠넓이 15 및 30으로 증가할수록 추정의 성능이 개선됨을 확인할 수 있다. 특히 자료 의존적으로 구해진 띠넓이 12와 그와 유사한 띠넓이 15일 때가 가장 추정을 잘 해내며, 띠넓이 50으로 적절한 수준 이상으로 높아지면 다시 추정의 성능이 현저히 감소하는 것을 보았다. 이를 통해 변화점추정의 성능이 띠넓이의 선택에 따라 영향을 받을 수 있으므로 띠넓이의 선택이 추정에 있어 중요한 요소임을 확인하였으며, 본 논문이 제안한 자료 의존적 띠넓이 (3.1)

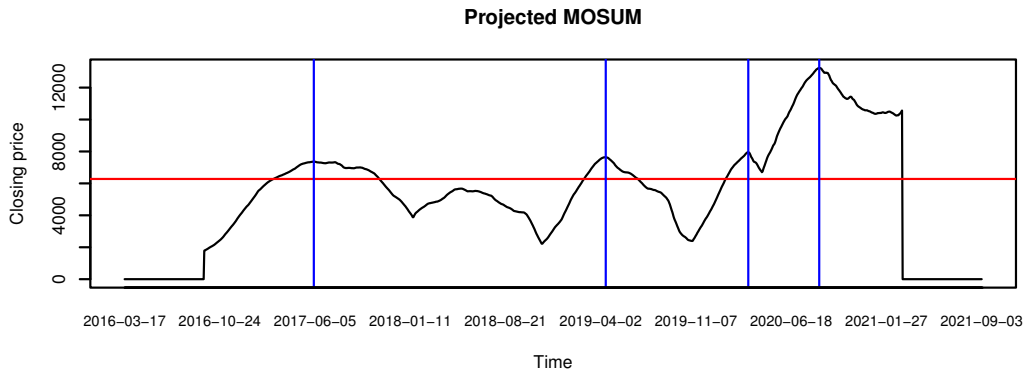

Figure 4: *Projected MOSUM of S&P 500 companies' stock price.*

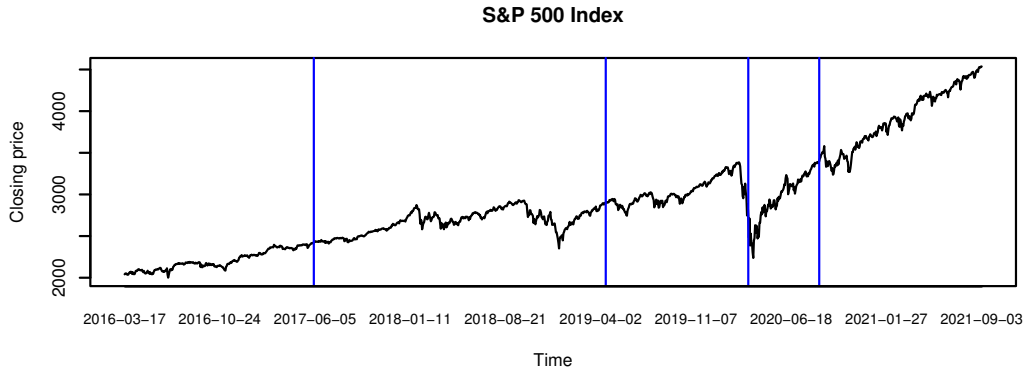

Figure 5: *S&P 500 index overlaied by estimated change points in vertical lines.*

가 적절함을 확인할 수 있다.

본 장에서는 모의 연구를 통해서 제안한 MOSUM 성근 프로젝션 방법의 성능을 살펴보았다. 먼저 사이즈 측면에서 오차항에 상관없이 기대한 유의 수준을 잘 지킴을 확인할 수 있었다. 또 다수의 변화점이 존재하더라도 성공적으로 변화점의 개수 및 위치를 정확히 찾아냈다. 또한 차원이 증가함에 따라 차원의 축복 효과로 오히려 성능이 좋아짐을 볼 수 있었다. 하지만, 일부 몇몇의 경우에는 변화점을 과소 추정 하였는데 통계량보다는 임계값이 적절하게 선택되지 못한 이유때문으로 파악되었다. 따라서 좀 더 정확한 임계값을 설정하는 연구가 요구된다.

4. 실증자료분석

실증자료분석을 위하여 먼저 미국 주식시장의 S&P 500 지수 산정에 포함되는 505개 기업의 일별 종가를 살펴보았다. 분석에는 2016년 3월 17일부터 2021년 9월 7일까지 총 2000일 중 주식가격이 산정된 영업일 총 1,378일 동안의 데이터를 사용하였다. 따라서 자료의 차원은 505×1378이다. Figure 4는 MOSUM 성근 프로젝션 방법을 사용한 결과 얻어진 통계량을 나타낸다. 빨간 수평선은 BWB 방법을 사용하여 구해진 임계값이다. 임계값을 뛰어넘는 국소 최댓값은 각각 4개로 변화점은 시간 순으로 2017년 6월 1일, 2019년 4월 12일, 2020

년 3월 11일, 2020년 08월 21일로 파란 수직선으로 나타냈다. Figure 5 는 S&P 500 지수 변화 추이를 나타내며 MOSUM 성근 프로젝션에 의해서 얻어진 네 변화점을 파란색 수직선으로 표현하였다. 추정된 변화점 근처의 주요 경제 이슈는 다음과 같다.

먼저 2017년 6월의 변화점은 트럼프 대통령의 확대적 재정정책이 견인한 미국의 호황기를 반영한 것으로 생각된다. 당시 개인소득세/법인세 감면 등과 같은 재정지출 확대와 인프라 투자 확대로 인해 대내적 소비경제가 활성화 되었으며, 대외적으로 있었던 수출확대에 의한 무역적자 폭 감소 또한 주요 원인 중 하나라고 할 수 있겠다. 주식시장 내적으로는 Facebook, Amazon, Netflix, Google (FANG)으로 대변되는 기술주의 폭발적 성장 또한 전반적 주가 상승을 견인한 요인이었다. 또한, 미국의 외적 원인으로서, 2016년 까지 세계 경제에 존재했던 대형 리스크들, 즉, 2014년 국제유가 하락으로 촉발된 러시아 경제위기와 유럽의 Brexit로 인한 리스크들의 충격에서 벗어나기 시작한 것도 2017년 증시 호황의 주요 원인으로 생각해볼 수 있다. 두 번째로 2019년 4월 전후에는 소위 'R의 공포'라 불리우는 현상이 발생하였다. 이는 장단기 국채금리가 역전하여 장기금리가 단기금리를 하회하는 현상으로, 향후 시장에 대한 시장 참여자의 부정적 전망을 나타내며, 따라서 경기침체의 선행지표로 여겨진다. 특히 이 시기는 국제적으로 유사한 현상이 나타난 시기이기도 하였는데, 주요한 원인으로 미-중간 무역분쟁이 관련된 것으로 볼 수 있다. 이 영향으로 인해 주식시장의 변동성과 경제침체 리스크가 커짐에 따라 증시 불안이 있었던 시기이므로, 변화점이 탐지된 것은 당연한 결과라 할 수 있다. 세 번째 변화점인 2020년 3월 11일에는 Figure 5 상 증시 대폭락이 나타나는데, 이는 COVID-19의 영향으로 촉발된 세계적 경제불황이 반영된 것이다. 3월의 대폭락 이후 약 5개월 간 증시가 예전 수준을 회복하며 폭발적으로 성장하였는데, 이 시기에 연방준비위원회의 저금리기조가 유지됨에 따라, 개인의 투자규모가 대폭 확대된 것을 원인으로 꼽을 수 있다. 이후 9월에는 COVID-19 장기화로 인한 시장조정으로 증시가 급락하는데, 이 폭등과 급락의 변화를 2020년 8월 21일로 추정된 변화점을 통해서 확인할 수 있다. 이와 같이, S&P 500 기업 주가의 실증 데이터를 통하여 MOSUM기반 성근 프로젝션을 통한 변화점 추정 방법이 현실을 반영하여 변화점을 잘 추정해내는 것을 확인하였다.

5. 결론 및 논의점

본 논문은 MOSUM 성근 프로젝션을 이용하여 고차원 시계열에서 변화점을 추정하는 방법에 대해서 연구하였다. MOSUM 방법은 CUSUM 통계량을 작은 구간에서 구하는 것으로 이해할 수 있기에 CUSUM의 탁월한 변화점 탐지 능력을 기대할 수 있을 뿐만 아니라 다중 변화점을 임계값 이상의 국소 최댓값으로 찾아낼 수 있어 순차적 오류에서도 자유롭다. 모의 연구를 통해서 우리가 제안한 방법이 적절한 사이즈를 가지고 변화점 개수와 위치를 잘 찾음을 살펴볼 수 있었다. 또한 실증자료 분석을 통해 최근 6년 동안의 S&P 500 지수를 구성하는 500개 개별 종목을 동시에 분석하여 네 개의 변화점을 찾을 수 있었다.

몇몇 모의 실험에서 MOSUM 성근 프로젝션 방법론이 기대했던 성능을 보이지 못하였는데 이는 임계점이 적절히 찾지 못해서 생긴 결과였다. 따라서 좀 더 정확한 임계점을 찾는 방법에 대한 연구가 필요할 것으로 보인다. 또한 성근성에 의해서 변화점을 찾는 경우, 정확하게 어떠한 변수(차원)에서 변화가 일어나는지에 대해서는 현재 방법론의 경우 찾을 수 없다. 이러한 후속 연구를 통해 MOSUM 성근 프로젝션의 방법의 활용을 넓힐 수 있을 것이라 본다. 또한 국소 평균뿐만 아니라 국소다항추세(local polynomial trend)에 대하여 MOSUM 성근 프로젝션을 통해 변화점을 찾는 방법으로의 확장도 흥미로운 추가 연구임에 틀림없다.

References

Andrews DWK (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica*, **59**, 817–858.

Cho H (2016). Change-point detection in panel data via double cusum statistic, *Electronic Journal of Statistics*, **10**, 2000–2038.

Eichinger B and Kirch C (2018). A MOSUM procedure for the estimation of multiple random change points, *Bernoulli*, **24**, 526–564.

Fryzlewicz P (2014). Wild binary segmentation for multiple change-point detection, *The Annals of Statistics*, **42**, 2243–2281.

Lee T and Baek C (2020). Block wild bootstrap-based CUSUM tests robust to high persistence and misspecification, *Computational Statistics & Data Analysis*, **150**, 106996.

Wang T and Samworth RJ (2018). High dimensional change point estimation via sparse projection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**, 57–83.

Received September 13, 2021; Revised October 26, 2021; Accepted December 3, 2021

MOSUM 성근 프로젝션을 이용한 고차원 시계열의 변화점 추정

김문정[a], 백창룡[1,a]

[a]성균관대학교 통계학과

요 약

본 논문은 Wang과 Samworth (2018)가 제안한 성근 프로젝션 방법을 개선하여 MOSUM을 이용하여 고차원의 시계열데이터에 존재하는 다중 평균 변화점을 추정하는 방법에 대해서 제안한다. 제안한 방법은 국소방법으로 다중 변화점을 동시에 찾을 수 있어 순차적 오류를 최소화 할 뿐만 아니라 평균이 상쇄되는 경우에도 변화점을 추정하는 장점을 지니고 있다. 또한 데이터 의존적인 방법으로 블록 와일드 붓스트랩 방법을 활용하여 임계점을 찾는 방법을 제안한다. 모의 실험을 통해 제안한 방법이 좋은 성능을 보임을 확인하였으며 S&P 500 지수를 구성하는 개별 기업들의 금융 자료에 적용하여 최근 6년간 네 번의 변화점을 찾았다.

주요용어: 변화점, MOSUM, 성근 프로젝션, 고차원 시계열

이 논문은 한국연구재단의 지원을 받아 수행된 기초연구 사업임(NRF-2019R1F1A1057104).

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