The Korean Journal of Applied Statistics
2022, Vol. 35, No. 3, 371–384

DOI: https://doi.org/10.5351/KJAS.2022.35.3.371
Print ISSN1225-066X / Online ISSN 2383-5818

An empirical evidence of inconsistency of the ℓ_1 trend filtering in change point detection

Donghyeon Yu[a], Johan Lim[b], Won Son[1,c]

[a]Department of Statistics, Inha University; [b]Department of Statistics, Seoul National University;
[c]Department of Information Statistics, Dankook University

Abstract

The fused LASSO signal approximator (FLSA) can be applied to find change points from the data having piecewise constant mean structure. It is well-known that the FLSA is inconsistent in change points detection. This inconsistency is due to a total-variation denoising penalty of the FLSA. ℓ_1 trend filter, one of the popular tools for finding an underlying trend from data, can be used to identify change points of piecewise linear trends. Since the ℓ_1 trend filter applies the sum of absolute values of slope differences, it can be inconsistent for change points recovery as the FLSA. However, there are few studies on the inconsistency of the ℓ_1 trend filtering. In this paper, we demonstrate the inconsistency of the ℓ_1 trend filtering with a numerical study.

Keywords: consistency, fused LASSO signal approximator (FLSA), ℓ_1 trend filtering, multiple chage points detection

1. 서론

특정 시점 이전과 이후의 관측값 분포가 달라지는 점을 변화점(change point)이라 한다 (Pettitt, 1979). 관측값에 변화점이 존재하는 경우 추정 및 검정방법, 모수 추정량 등이 달라져야하므로 변화점의 정확한 식별은 실제 데이터 분석에 있어서도 중요한 과제이다. 이러한 이유 때문에 변화점 식별 문제는 오랜 기간 많은 연구자들이 관심을 기울여왔으며 최근까지도 다양한 연구들이 이루어지고 있다.

변화점 식별 문제 중 흔히 다루어지고 있는 문제로는 구간별 상수인 평균 구조(piecewise constant mean structure)를 가지는 데이터에서 변화점을 식별하는 문제를 들 수 있다. Yao (1988)에서와 같이 가능한 모든 조합들을 대상으로 변화점 여부를 판별하는 방법도 고려할 수 있지만 이 경우 데이터의 크기가 n일 때 가능한 변화점의 조합은 모두 2^{n-1}개가 되어 n이 큰 경우에는 현실적으로 답을 찾기 어렵다는 문제점이 있다. 계산에 있어서의 효율성을 제고하기 위한 방법으로 CUSUM 통계량을 이용한 이진분할법(binary segmentation)과 변형된 방법들, 즉 WBS (wild binary segmentation) (Fryzlewicz, 2014), CBS (circular binary segmentation) (Olshen 등, 2004) 등이 제안된 바 있다. 한편, 구간별 상수 구조 데이터에서 변화점을 식별하는 또 다른 방법으로 벌점이 부여된 오차제곱합을 최소화하는 벌점화 회귀모형(penalized regression model)도 제안된 바 있다. 벌점화 회귀모형의 대표적인 방법으로 융합라쏘신호근사기(fused LASSO signal approximator; FLSA)를 들

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1F1A1A01048127 (Donghyeon Yu), No. 2021R1A2C1010786 (Johan Lim), and No. 2020R1F1A1A01051039 (Won Son)).

[1] Corresponding author: Department of Information Statistics, Dankook University, 152, Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do 16890, Repulic of Korea E-mail: son.won@dankook.ac.kr

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

수 있다. FLSA는 효율적으로 변화점을 식별할 수 있다는 장점이 있지만 변화점 식별에 있어서의 일치성이 보장되지 않는다는 문제점도 있다.

이 연구에서는 모형

$$y_i = \theta_i + \epsilon_i \tag{1.1}$$

에서 시점 i 전후의 기울기 $\theta_i - \theta_{i-1}$과 $\theta_{i+1} - \theta_i$의 차이

$$(\theta_{i+1} - \theta_i) - (\theta_i - \theta_{i-1}) \tag{1.2}$$

이 대부분 0이고 일부 시점에만 0이 아닌 값을 가지는 구간별 선형모형(piecewise linear model)을 가정한다. 이렇게 $\theta_{i+1} - \theta_i \neq \theta_i - \theta_{i-1}$인 점 i를 구간별 선형모형에서의 변화점이라 한다. 모형 (1.1)에서 $\theta_i = \mathbb{E}[y_i]$이고 오차항 ϵ_i는 정규분포 $N(0, \sigma^2)$를 따르는 것으로 가정한다.

이와 같은 가정 아래에서 구간별 선형모형 (1.1)을 따르는 관측값 $y_1, y_2, \ldots, y_n$으로부터 변화점을 식별하기 위해 ℓ_1 추세필터(trend filtering) (Kim 등, 2009)가 널리 사용되고 있다. ℓ_1 추세필터의 목적함수(objective function)는 잔차제곱합(residual sum of square, RSS)에 기울기의 차이 (1.2)와 관련된 벌점항이 더해진 형태라는 점에서 HP 필터(Hodrick-Prescott filtering) (Hodrick과 Prescott, 1997)와 유사한 점이 있다. 다만, HP 필터의 경우 기울기 차이의 제곱합을 벌점으로 사용하고 있어 능형회귀(ridge regression) 모형에 대응된다면, ℓ_1 추세필터는 기울기 차이의 절대값을 벌점으로 사용하므로 라쏘회귀(LASSO regression) 모형에 대응된다고 볼 수 있다. 따라서 HP 필터는 기울기의 변화점 식별에 이용하기에 부적합한 반면, ℓ_1 추세필터는 기울기의 변화점을 찾는 데 유용하게 사용될 수 있다. ℓ_1 추세필터의 목적함수 등에 대한 더 자세한 사항은 2절과 3절에서 다루기로 한다.

이 논문은 다음과 같이 구성된다. 먼저 2절에서는 FLSA와 ℓ_1 추세필터의 기본 개념을 살펴보고 3절에서는 두 방법의 변화점 식별에 있어서의 비일치성에 대해 생각해본다. 4절에서는 모의실험을 통해 변화점 식별에 있어서의 비일치성을 확인해보고 구간별 선형 모형을 따르는 데이터의 변화점을 FLSA를 이용하여 식별하는 방법을 제안한다. 5절에서는 결과를 정리하고 앞으로의 과제를 검토해본다.

2. 벌점회귀모형을 이용한 변화점 식별

2.1. FLSA

융합라쏘(fused LASSO) 회귀모형(Tibshirani 등, 2005)은 $n \times p$ 행렬의 설명변수 $\mathbf{X}$와 p차원 벡터인 반응변수 $\mathbf{y}$가 선형모형 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$로 표현될 수 있을 때 회귀계수 $\boldsymbol{\beta}$의 희소성을 구현하기 위한 방법으로 자주 사용되고 있다. 즉, 융합라쏘 회귀모형은 회귀계수 $\boldsymbol{\beta}$의 p개 원소 중 많은 원소의 값이 0이고 인접한 두 회귀계수의 값이 일치하는 경우가 많을 때 적합한 모형으로 볼 수 있다. 융합라쏘 회귀모형에서 회귀계수 $\boldsymbol{\beta}$의 추정량은

$$\hat{\boldsymbol{\beta}}^{\mathrm{FL}}(\lambda_1, \lambda_2) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_{\mathrm{TV}} \right\}$$

와 같이 라쏘 회귀모형에 총변동벌점(total variation penalty) $\|\boldsymbol{\beta}\|_{\mathrm{TV}}$가 더해진 형태의 목적함수를 최소화하여 구할 수 있다. 여기서 행렬 $\mathbf{X}$의 i번째 행 $\mathbf{x}_i$에 대해 $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^n (y_i - \mathbf{x}_i\boldsymbol{\beta})^2$로 정의되고 $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^n |\beta_i|$이며 조절모수는 부등식 $\lambda_1, \lambda_2 \geq 0$을 만족시키는 것으로 가정한다. 총변동벌점은 $\|\boldsymbol{\beta}\|_{\mathrm{TV}} = \sum_{i=2}^p |\beta_i - \beta_{i-1}|$로 표현할 수 있으며 인접한 회귀계수들의 값에 차이가 없으면 작은 값을, 차이가 클수록 큰 값을 가지게 된다.

FLSA는 계획행렬 $\mathbf{X}$가 n차원 단위행렬이고 관측값 $\mathbf{y}$와 모수 $\boldsymbol{\mu}$가 n차원 벡터인 융합라쏘 회귀모형으로

$$\hat{\boldsymbol{\mu}}^{\mathrm{FL}}(\lambda_1, \lambda_2) = \underset{\boldsymbol{\mu} \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \|\mathbf{y} - \boldsymbol{\mu}\|_2^2 + \lambda_1 \|\boldsymbol{\mu}\|_1 + \lambda_2 \|\boldsymbol{\mu}\|_{\mathrm{TV}} \right\} \tag{2.1}$$

와 같이 표현할 수 있다. 식 (2.1)에서 ℓ_1 벌점항 $\|\boldsymbol{\mu}\|_1 = \sum_{i=1}^n |\mu_i|$은 0이 아닌 모수의 개수를 줄여 희소성을 구현하는 역할을 하고 총변동벌점 $\|\boldsymbol{\mu}\|_{\mathrm{TV}} = \sum_{i=2}^n |\mu_i - \mu_{i-1}|$은 인접한 두 모수가 동일한 값을 가지도록 한다. 즉, 총변동벌점은 모수들의 집합 $\{\mu_1, \mu_2, \ldots, \mu_n\}$에서 변화점을 찾기 위해 사용할 수 있다. 추정량 $\hat{\boldsymbol{\mu}}^{\mathrm{FL}}(\lambda_1, \lambda_2)$는 $\hat{\boldsymbol{\mu}}^{\mathrm{FL}}(0, \lambda_2)$를 소프트 임계화(soft-thresholding)하여 구할 수 있으므로 (Friedman 등, 2007) 모수들의 집합 $\{\mu_1, \mu_2, \ldots, \mu_n\}$에서 변화점을 찾기 위해서는 ℓ_1 벌점항을 제외한 총변동벌점항만을 포함한 모형

$$\hat{\boldsymbol{\mu}}^{\mathrm{FL}}(0, \lambda_2) = \underset{\boldsymbol{\mu}\in\mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{y} - \boldsymbol{\mu}\|_2^2 + \lambda_2 \|\boldsymbol{\mu}\|_{\mathrm{TV}} \right\} \tag{2.2}$$

를 사용하면 충분하다.

2.2. ℓ_1 추세필터

순차적으로 관찰되는 데이터는 여러가지 요인들로 분해할 수 있다. 예를 들어 시계열 데이터의 경우 추세(trend), 순환(cycle), 계절(season), 불규칙(irregular) 요인으로 분해할 수 있다. Hodrick과 Prescott (1997)은 미국의 국민총생산(GNP) 데이터로부터 경기순환에 기인한 변동만을 추출해내기 위해 GNP의 추세변동요인을 분리해내는 방법으로 HP 필터(Hodrick-Prescott filter)를 제안한 바 있다. 모형 (1.1)에서와 같이 관측값을 y_i, i 시점에서의 평균 수준을 θ_i라 할 때 HP 필터에 의해 구해지는 추세 추정량 $\hat{\boldsymbol{\theta}}^{\mathrm{HP}}$는 다음 식과 같이 표현된다.

$$\hat{\boldsymbol{\theta}}^{\mathrm{HP}}(\lambda) = \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2}\sum_{i=1}^n (y_i - \theta_i)^2 + \lambda \sum_{i=2}^{n-1} (\theta_{i-1} - 2\theta_i + \theta_{i+1})^2 \right\}. \tag{2.3}$$

식 (2.3)에서 조절모수 λ는 추세에 대한 추정량 $\hat{\boldsymbol{\theta}}^{\mathrm{HP}}$가 관측값 y_i에 가까운 값을 가지는 정도와 제약식 $\theta_{i-1} - \theta_i = \theta_i - \theta_{i+1}$을 만족시키는 정도를 결정해주는 역할을 한다. 즉, λ가 작을수록 $\hat{\boldsymbol{\theta}}^{\mathrm{HP}}$가 관측값 y_i에 가까운 값을 가지게 되고 반대로 λ가 클수록 제약조건 $\theta_{i-1} - \theta_i = \theta_i - \theta_{i+1}$을 잘 만족시키게 된다.

한편, 능형회귀(ridge regression) 모형

$$\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}$$

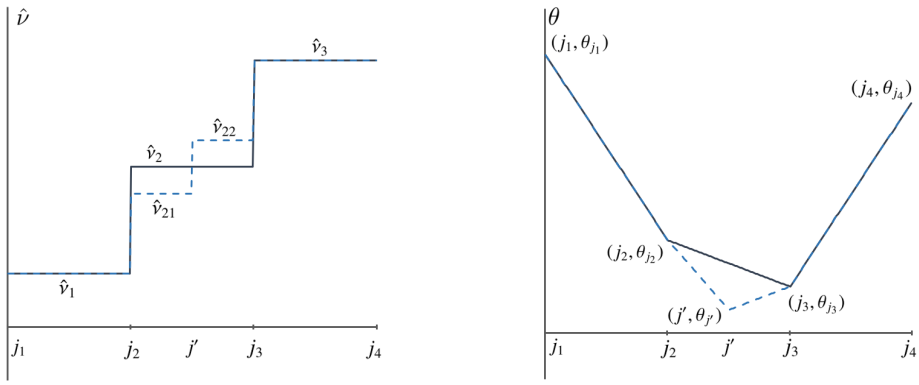
의 ℓ_2-벌점항 $\|\boldsymbol{\beta}\|_2^2$을 ℓ_1-벌점항 $\|\boldsymbol{\beta}\|_1$으로 대체한 라쏘(LASSO)회귀 모형

$$\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

에서와 같이 추세필터에서도 HP 필터의 ℓ_2-벌점항을 ℓ_1-벌점항으로 대체하는 방법을 고려할 수 있다 (Kim 등, 2009). ℓ_1 추세필터는 다음 식과 같이 표현된다.

$$\hat{\boldsymbol{\theta}}^{\ell_1}(\lambda) = \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2}\sum_{i=1}^n (y_i - \theta_i)^2 + \lambda \sum_{i=2}^{n-1} |\theta_{i-1} - 2\theta_i + \theta_{i+1}| \right\}. \tag{2.4}$$

라쏘회귀 모형에서 능형회귀 모형의 ℓ_2-노름(norm)을 ℓ_1-노름(norm)으로 대체하여 회귀계수 $\boldsymbol{\beta}$의 희소성(sparsity)을 구현한 것과 같이 식 (2.4)에서 HP 필터의 제약조건 $\theta_{i-1} - 2\theta_i + \theta_{i+1} = 0$에 대한 벌점항을 ℓ_1-노름으로 대체함에 따라 제약조건을 만족하지 않는 점, 즉 기울기에 차이가 있는 점의 희소성을 구현할 수 있게 된다. 다시 말해 HP 필터의 경우 정확히 $\theta_{i-1} - \theta_i = \theta_i - \theta_{i+1}$인 i가 존재하지 않기 때문에 모든 점에서 기울기가 변하지만 ℓ_1 추세필터의 경우 많은 점들이 정확히 $\theta_{i-1} - \theta_i = \theta_i - \theta_{i+1}$을 만족시킨다. 즉 ℓ_1 추세필터에 의해 구해진 추정량은 기울기에 변화가 없는 점들이 많다는 것을 의미한다.



(a) FLSA (b) ℓ_1 Trend Filtering

Figure 1: *Monotonely increasing mean levels or slopes.*

한편, FLSA는 단조결합성질(monotone fusion property)이 성립(Friedman 등, 2007)되는 반면, ℓ_1 추세필터에 있어서는 이러한 성질이 성립되지 않는다. FLSA에서의 단조결합성질은 조절모수가 부등식 $\lambda_2 < \lambda_2'$을 만족시킬 때 $\hat{\boldsymbol{\theta}}^{\mathrm{FL}}(0, \lambda_2')$에 의해 식별된 변화점 집합 $\widehat{\mathcal{J}}(\lambda_2')$이 $\hat{\boldsymbol{\theta}}^{\mathrm{FL}}(0, \lambda_2)$에 의해서 식별된 변화점 집합 $\widehat{\mathcal{J}}(\lambda_2)$에 포함됨을 의미한다. 다시 말해 조절모수 λ_2가 증가할 때 한번 변화점 집합에서 제거된 점은 다시 변화점이 되지 못한다. 반대로 ℓ_1 추세필터에서는 단조결합성질이 성립되지 않으므로 변화점 집합에서 제거되었던 점도 다시 변화점이 될 수 있다. 이러한 이유 때문에 ℓ_1 추세필터의 해경로(solution path)는 FLSA에 비해 더 복잡하게 주어진다.

3. 변화점 식별에 있어서의 비일치성

3.1. FLSA

FLSA는 신호(signal)의 크기가 잡음(noise)에 비해 충분히 큰 경우에도 일반적으로 참변화점 식별에 있어서 비일치성(inconsistency)을 보인다. 참변화점의 집합을 $\mathcal{J}^* = \{j_1^*, j_2^*, \ldots, j_J^*\}$라 하고 모든 참변화점 j^*에서의 평균 수준의 차이 $\alpha_{j^*}^* = |\nu_{j^*}^* - \nu_{j^*-1}^*|$가 오차항의 표준편차 σ에 비해 충분히 크다고 가정하자. 이때 조절모수 λ_2에 따라서 생성된 변화점 집합의 경로 $\widehat{\mathcal{J}}(\lambda_2)$가 참변화점 집합 $\mathcal{J}^*$를 포함할 확률이 1보다 작음이 알려져 있다 (Rojas와 Wahlberg, 2014; Son과 Lim, 2019). 여기서 ν는 구간을 가리키는 인덱스 j와 관측값의 순서를 가리키는 인덱스 i를 구분하기 위해 도입한 기호로 ν_j^*는 j번째 구간에서의 평균 수준에 해당되고 i가 j번째 구간에 포함된 점일 때 $\mu_i^* = \nu_j^*$이다.

다만, FLSA는 참변화점으로 나누어진 각 구간의 평균값들이 모두 극값인 경우에는 일치성을 보인다. 즉, 참변화점들에 의해 나누어진 구간 $[1, j_1^*), [j_1^*, j_2^*), \ldots, [j_J^*, n]$에서의 평균 수준을 $\nu_1^*, \nu_2^*, \ldots, \nu_{J+1}^*$이라 하면 모든 j^*에 대해 $(\nu_{j^*}^* - \nu_{j^*-1}^*)(\nu_{j^*}^* - \nu_{j^*+1}^*) > 0$이 항상 성립할 때는 $\widehat{\mathcal{J}}(\lambda_2) = \mathcal{J}^*$인 조절모수 λ_2를 찾을 수 있고 변화점 식별에 있어서의 일치성이 충족된다. 반대로 평균 값들이 계단형태로 단조증가하거나 단조감소하는 구간이 있는 경우에는 일치성이 충족되지 않음이 알려져 있다. 즉, $(\nu_{j^*}^* - \nu_{j^*-1}^*)(\nu_{j^*}^* - \nu_{j^*+1}^*) < 0$인 j^*가 적어도 하나 존재하는 경우에는 $\widehat{\mathcal{J}}(\lambda_2) = \mathcal{J}^*$인 조절모수 λ_2를 찾을 수 있는 확률이 1보다 작으며 일치성이 충족되지 않는다.

이렇게 평균값이 계단 형태로 분포되어 있을 때 일치성이 충족되지 않는 것은 총변동벌점의 특성에 기인한다. Figure 1의 (a)와 같이 전체 구간이 세 개의 하위구간 $\widehat{B}_1 = [1, j_1), \widehat{B}_2 = [j_1, j_2), \widehat{B}_3 = [j_2, n]$으로 나누어진

경우와 $\widehat{B}_2 = [j_1, j_2)$가 $\widehat{B}_{21} = [j_1, j')$, $\widehat{B}_{22} = [j', j_2)$로 한번 더 나누어져 모두 네 개의 하위구간 $\widehat{B}_1, \widehat{B}_{21}, \widehat{B}_{22}, \widehat{B}_3$으로 이루어져 있는 경우를 생각해보자. 각 구간에서의 FLSA 추정값 $\hat{\nu}_1, \hat{\nu}_2, \hat{\nu}_3$와 $\hat{\nu}_{21}, \hat{\nu}_{22}$에 대해 두 부등식

$$\hat{\nu}_1 < \hat{\nu}_2 < \hat{\nu}_3, \quad \hat{\nu}_1 < \hat{\nu}_{21} < \hat{\nu}_{22} < \hat{\nu}_3$$

이 성립한다고 가정하자. 이때 총변동벌점은

$$|\hat{\nu}_1 - \hat{\nu}_2| + |\hat{\nu}_2 - \hat{\nu}_3| = |\hat{\nu}_1 - \hat{\nu}_3| = |\hat{\nu}_1 - \hat{\nu}_{21}| + |\hat{\nu}_{21} - \hat{\nu}_{22}| + |\hat{\nu}_{22} - \hat{\nu}_3|$$

와 같이 주어지므로 두 추정값의 총변동벌점에 차이가 없다. 반면, 오차제곱항의 경우 더 잘게 나누어진 추정값에서 더 작은 값을 가질 수 있기 때문에 FLSA에 의해 구해진 추정량은 거짓변화점을 포함하게 될 가능성이 높음을 알 수 있다.

예를 들어 참변화점이 $i_1^* = 3, i_2^* = 5$, 즉 $\mu_3^* \neq \mu_2^*, \mu_5^* \neq \mu_4^*$이고 구간 $B_1^* = [1, 3),\ B_2^* = [3, 5),\ B_3^* = [5, 6]$에서 평균수준이 $\nu_1^* = 1,\ \nu_2^* = 2,\ \nu_2^* = 3$이라 가정해보자. 관측값 $y_3 = 1.9,\ y_4 = 2.1$일 때 두번째 구간이 $\widehat{B}_2 = [3, 5)$로 하나로 묶여 있으면 $\hat{\nu}_2 = 2$, 두 구간 $\widehat{B}_{21} = [3, 4),\ \widehat{B}_{22} = [4, 5)$로 나누어져 있으면 $\hat{\nu}_{21} = 1.9,\ \hat{\nu}_{22} = 2.1$이 된다. 따라서 총변동벌점은 두 추정값에서 동일하지만 오차제곱합은 후자가 더 작은 값을 가지게 되므로 FLSA를 적용하면 B_2^*가 두 구간으로 나누어진다.

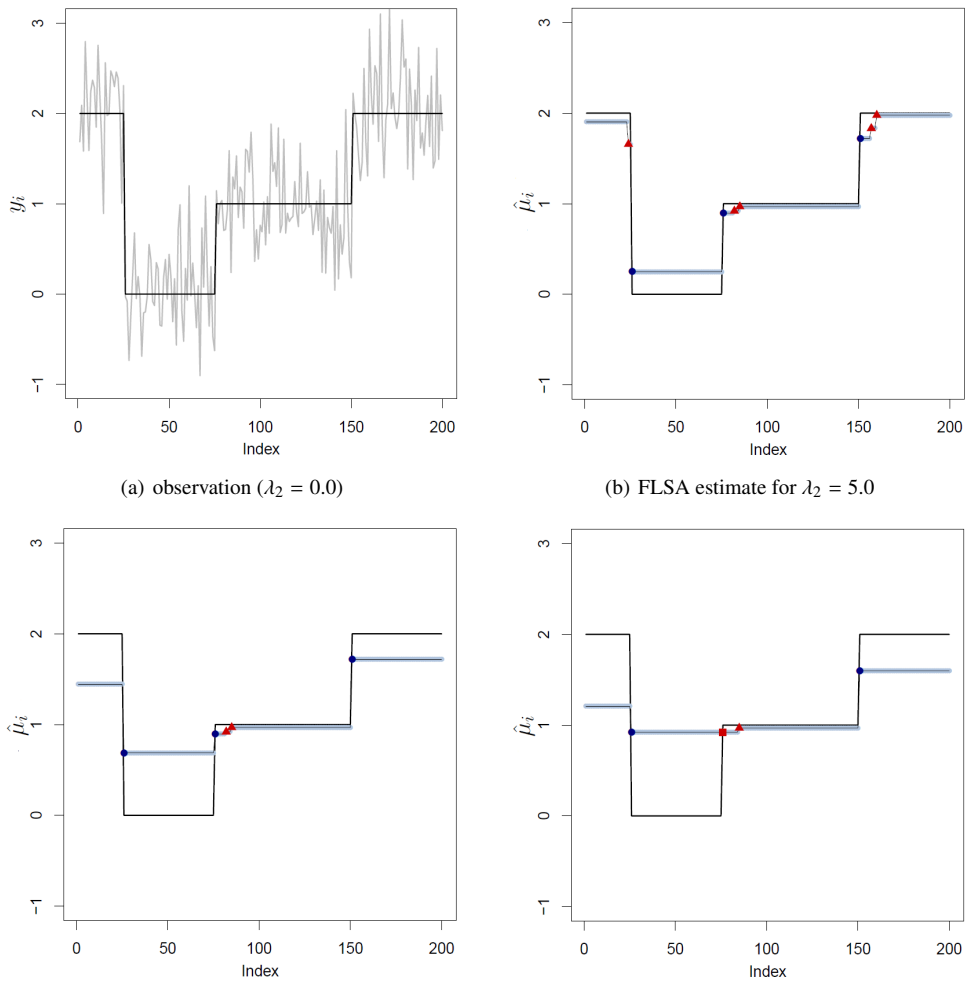
Figure 2는 구간별 상수구조를 가진 데이터에 대해 FLSA를 적용하여 구한 추정치를 조절모수 λ_2의 크기별로 나열한 그래프이다. Figure 2 (a)는 구간별 상수 구조 평균 수준과 관측값에 해당된다. 관측값은 $y_i = \mu_i + \epsilon_i (i = 1, 2, \ldots, 200)$로 오차항 ϵ_i는 서로 독립이고 표준편차 0.5인 정규분포($\epsilon_i \overset{iid}{\sim} N(0, 0.5^2)$)를 따른다. 기댓값 μ_i에는 계단 형태로 분포되어 있는 구간 $[76, 151)$이 존재하므로 FLSA가 비일치성을 보일 것으로 생각할 수 있다.

앞에서 살펴보았듯이 FLSA는 단조결합성질을 가지고 있으므로 n개의 관측값에 대해 서로 다른 변화점 집합은 모두 n개이고 조절모수 λ_2의 값이 커짐에 따라 변화점이 사라지게 된다. 따라서 λ_2의 값 $0 < \lambda_2^{(1)} < \lambda_2^{(2)} < \cdots < \lambda_2^{(n-1)}$에 따라 해경로(solution path), 즉 변화점 집합의 경로를

$$\widehat{\mathcal{J}}(0) \supsetneq \widehat{\mathcal{J}}(\lambda_2^{(1)}) \supsetneq \widehat{\mathcal{J}}(\lambda_2^{(2)}) \supsetneq \cdots \supsetneq \widehat{\mathcal{J}}(\lambda_2^{(n-1)})$$

과 같이 생각할 수 있다. 여기서 $\mathcal{J}(0)$은 $\lambda_2 = 0$으로 모든 점이 변화점인 집합을 의미하고 $\lambda_2^{(i)}$는 i개의 변화점이 사라진 변화점 집합을 의미한다. $i = n - 1$인 경우 모든 변화점이 사라져 모든 점들이 하나의 구간으로 묶인 상태가 된다. 관측값이 주어지면 이와 같이 변화점 집합이 달라지는 경계점(kink point) λ_2의 값 $0 < \lambda_2^{(1)} < \lambda_2^{(2)} < \cdots < \lambda_2^{(n-1)}$을 구하는 경로 알고리즘(Hoefling, 2010)이 알려져 있으므로 쉽게 해경로를 구할 수 있다.

Figure 2의 (b)~(d)는 조절모수 λ_2의 값 중 FLSA의 비일치성을 확인하는 데 유용한 3개의 값을 주관적으로 선택하여 작성한 그래프이다. 관측값에 FLSA를 적용하여 구한 추정치를 살펴보면 Figure 1의 (b)에서 알 수 있듯이 조절모수가 $\lambda_2 = 5.0$일 때 모든 참변화점이 정확하게 식별되었지만 참변화점이 아닌 다섯 개의 점도 변화점으로 식별된 것을 확인할 수 있다. 조절모수를 $\lambda_2 = 16.0$으로 증가시켰을 때는 Figure 1의 (c)와 같이 모든 참변화점이 정확하게 식별되었고 거짓변화점은 두 개만 남아 있다. 특히 남아 있는 두 개의 변화점은 평균 수준이 극값이 아닌 계단 형태로 증가하는 중간 부분에 해당되는 구간 $[76, 151)$에 포함되어 있다. 조절모수를 $\lambda_2^{(5)} = 21.9$로 증가시키면 Figure 1의 (d)에서와 같이 두 개의 거짓변화점이 여전히 유지된 상태에서 참변화점 세 개 중 하나가 사라지게 된다. 즉, 조절모수 $\lambda_2^{(5)} = 21.9$는 세 개의 모든 참변화점을 포함하고 있는 변화점 집합 $\widehat{\mathcal{J}}(\lambda_2^{(n-6)})$과 참변화점 중에 하나 이상의 변화점이 식별되지 못한 변화점 집합 $\widehat{\mathcal{J}}(\lambda_2^{(n-5)})$의 경계점에 해당된다. 이 예에서 관측값의 평균 수준이 계단 형태로 단조증가 또는 단조감소하는 데이터에 대해서는 FLSA를 이용하여 참변화점 집합을 정확하게 식별하는 것, 다시 말해 참변화점만 남겨두고 거짓변화점을 모두 제거하는 것은 쉽지 않다는 것을 알 수 있다.



(a) observation ($\lambda_2 = 0.0$)

(b) FLSA estimate for $\lambda_2 = 5.0$

(c) FLSA estimate for $\lambda_2 = 16.0$

(d) FLSA estimate for $\lambda_2 = 21.9$

* Blue circle : identified true change points, red triangle : identified false change points, red square : missed true change points.

Figure 2: *Inconsistency of the FLSA.*

3.2. ℓ_1 추세필터

식 (2.4)에서 볼 수 있는 바와 같이 ℓ_1 추세필터의 목적함수도 FLSA와 마찬가지로 오차제곱항과 ℓ_1 벌점항이 더해진 형태로 구성되어 있다. ℓ_1 추세필터의 벌점항 $\lambda \sum_{i=2}^{n-1} |\theta_{i-1} - 2\theta_i + \theta_{i+1}|$에서 $|\theta_{i-1} - 2\theta_i + \theta_{i+1}| = |(\theta_{i-1} - \theta_i) - (\theta_i - \theta_{i+1})|$은 점 $i-1$과 i, i와 $i+1$ 사이의 기울기의 차이로 해석할 수 있다.

FLSA의 경우 기댓값이 단조증가 또는 단조감소할 때 변화점 식별에서 비일치성을 보였던 것과 마찬가지로 ℓ_1 추세필터의 경우에는 기울기가 계단 형태로 단조증가 또는 단조감소할 때 변화점 식별에서의 비일치성이 나타날 것으로 추정된다 (Rojas와 Wahlberg, 2014).

예를 들어 Figure 1의 (b)와 같이 $1 < j_1 < j_2 < j_3$이고 $\theta_{j_1} - \theta_1 < \theta_{j_2} - \theta_{j_1} < \theta_{j_3} - \theta_{j_2}$라 가정하면

$$|(\theta_{j_2} - \theta_{j_1}) - (\theta_{j_1} - \theta_1)| + |(\theta_{j_3} - \theta_{j_2}) - (\theta_{j_2} - \theta_{j_1})| = (\theta_{j_3} - \theta_{j_2}) - (\theta_{j_1} - \theta_1)$$

임을 알 수 있다. 다음으로 $j_1 < j' < j_2$인 j'이 존재하여 부등식 $\theta_{j_1} - \theta_1 < \theta_{j'} - \theta_{j_1} < \theta_{j_2} - \theta_{j'} < \theta_{j_3} - \theta_{j_2}$가 성립한다고 가정하면 마찬가지로

$$|(\theta_{j'} - \theta_{j_1}) - (\theta_{j_1} - \theta_1)| + |(\theta_{j_2} - \theta_{j'}) - (\theta_{j'} - \theta_{j_1})| + |(\theta_{j_3} - \theta_{j_2}) - (\theta_{j_2} - \theta_{j'})| = (\theta_{j_3} - \theta_{j_2}) - (\theta_{j_1} - \theta_1)$$

임을 알 수 있다. 따라서

$$\begin{aligned}&|(\theta_{j'} - \theta_{j_1}) - (\theta_{j_1} - \theta_1)| + |(\theta_{j_2} - \theta_{j'}) - (\theta_{j'} - \theta_{j_1})| + |(\theta_{j_3} - \theta_{j_2}) - (\theta_{j_2} - \theta_{j'})| \\ &= |(\theta_{j_2} - \theta_{j_1}) - (\theta_{j_1} - \theta_1)| + |(\theta_{j_3} - \theta_{j_2}) - (\theta_{j_2} - \theta_{j_1})|\end{aligned}$$

이 성립한다. 즉, 기울기가 단조증가하거나 단조감소할 때 단조성이 훼손되지 않는 범위 내에서 구간이 나누어지더라도 벌점항은 변하지 않는다는 것을 알 수 있다. 이런 관점에서 볼 때 ℓ_1 추세필터도 변화점 식별에 있어서 불일치성을 가지게 될 것임을 짐작할 수 있다.

4. 모의실험

4.1. 모형

FLSA와 ℓ_1 추세필터의 변화점 식별에 있어서의 일치성을 확인하기 위해 모의실험을 진행하였다. FLSA와 ℓ_1 추세필터의 비교를 위해 구간별 선형 구조를 가지는 모형은 구간별 상수 구조 평균모형의 누적합 $\theta_i = \sum_{k=1}^{i} \mu_k$으로 생성하였다. 따라서 구간별 선형 구조 모형을 차분하면 구간별 상수 구조 평균모형을 구할 수 있게 된다. 구간별 상수 모형과 구간별 선형 모형의 관계는 Figure 3에서 확인할 수 있다.

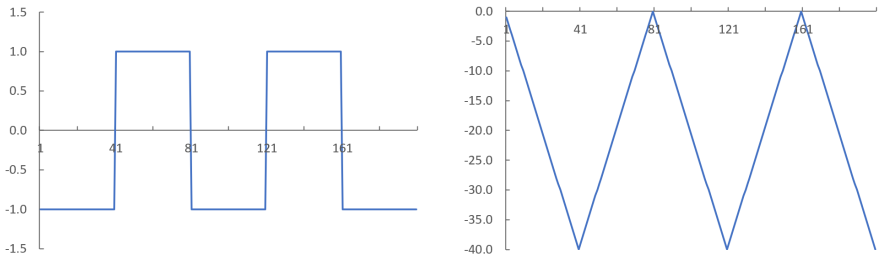
오차항 ϵ_i는 정규분포를 따르는 것으로 가정하고 관측값은 기댓값 μ_i, θ_i와 오차항 ϵ_i의 합 $y_i = \mu_i + \epsilon_i$와 $z_i = \theta_i + \epsilon_i$으로 정의하였다. 이렇게 정의된 관측값 y_i에 대해서는 FLSA를 적용하여 변화점을 식별하고, z_i에 대해서는 ℓ_1 추세필터를 적용하여 변화점을 식별하기로 한다. 즉, FLSA를 적용하여 변화점을 식별할 구간별 상수 모형과 ℓ_1 추세필터를 적용하여 변화점을 식별할 구간별 선형 모형의 i번째 관측값의 기댓값은 서로 다르지만 동일한 오차항을 가지도록 정의되었다. 이렇게 각 모형의 μ_i와 θ_i에 동일한 오차항 ϵ_i를 더한 것은 FLSA와 ℓ_1 추세필터를 비교할 때 오차항에 따른 차이를 가급적 축소하기 위해서이다.

모의실험에서는 아래와 같이 모두 네 가지 모형을 고려하였다.

• 모형1 : $\mu_i = \begin{cases} -1, & i = 1, 2, \ldots, 40,\ 81, 82, \ldots, 120,\ 161, 162, \ldots, 200, \\ 1, & i = 41, 42, \ldots, 80,\ 121, 122, \ldots, 160, \end{cases} \qquad \theta_i = \sum_{k=1}^{i} \mu_k$

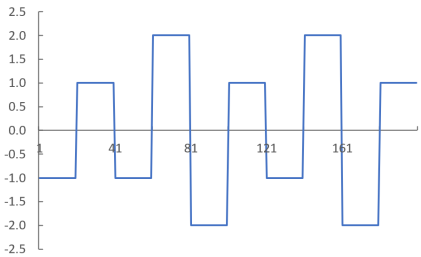
• 모형2 : $\mu_i = \begin{cases} -2, & i = 81, 82, \ldots, 100,\ 161, 162, \ldots, 180, \\ -1, & i = 1, 2, \ldots, 20,\ 41, 42, \ldots, 60,\ 121, 122, \ldots, 140, \\ 1, & i = 21, 22, \ldots, 40,\ 101, 102, \ldots, 120,\ 181, 182, \ldots, 200, \\ 2, & i = 61, 62, \ldots, 80,\ 141, 142, \ldots, 160 \end{cases} \qquad \theta_i = \sum_{k=1}^{i} \mu_k$

• 모형3 : $\mu_i = \begin{cases} -1, & i = 51, 52, \ldots, 100, \\ 1, & i = 101, 102, \ldots, 150, \\ 2, & i = 1, 2, \ldots, 50,\ 151, 152, \ldots, 200, \end{cases} \qquad \theta_i = \sum_{k=1}^{i} \mu_k$

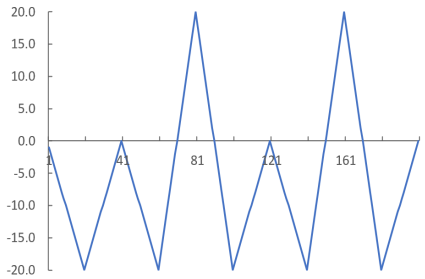


(a) Piecewise constant mean model 1

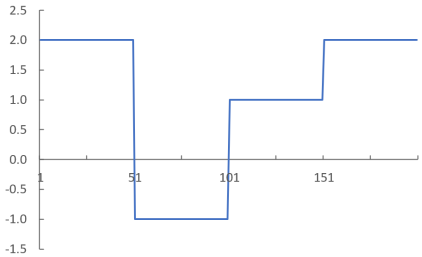
(b) Piecewise linear model 1



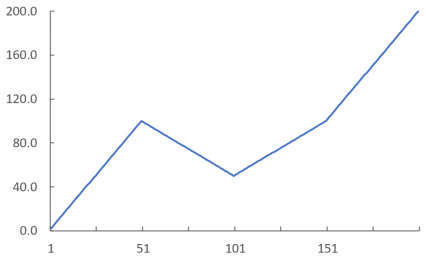
(c) Piecewise constant mean model 2



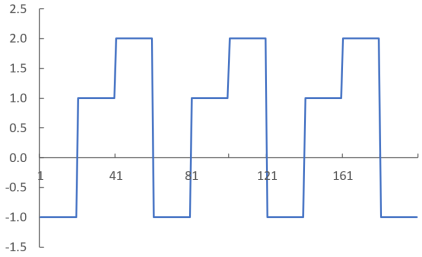
(d) Piecewise linear model 2



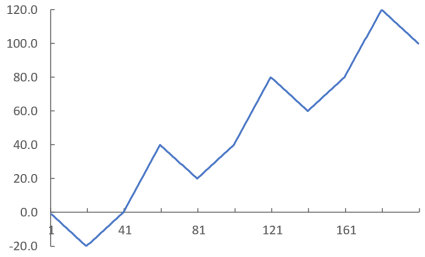
(e) Piecewise constant mean model 3



(f) Piecewise linear model 3



(g) Piecewise constant mean model 4



(h) Piecewise linear model 4

Figure 3: *Models for the numerical study.*

• 모형4 : $\mu_i = \begin{cases} -1, & i = 1, 2, \ldots, 20,\ 61, 62, \ldots, 80,\ 121, 122, \ldots, 140,\ 181, 182, \ldots, 200, \\ 1, & i = 21, 22, \ldots, 40,\ 81, 82, \ldots, 100,\ 141, 142, \ldots, 160, \\ 2, & i = 41, 42, \ldots, 60,\ 101, 102, \ldots, 120,\ 161, 162, \ldots, 180, \end{cases}$ $\theta_i = \sum_{k=1}^{i} \mu_k$

Table 1: Probability of detecting true change points set for the FLSA and ℓ_1 trend filtering

Model	FLSA					ℓ_1 Trend Filtering				
	$\sigma = 0.05$	$\sigma = 0.10$	$\sigma = 0.20$	$\sigma = 0.50$	$\sigma = 1.00$	$\sigma = 0.05$	$\sigma = 0.10$	$\sigma = 0.20$	$\sigma = 0.50$	$\sigma = 1.00$
1	1.00	1.00	0.99	0.57	0.08	0.00	0.00	0.00	0.00	0.00
2	1.00	1.00	0.99	0.42	0.01	0.00	0.00	0.00	0.00	0.00
3	0.02	0.02	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

먼저 모형 1과 2의 평균 수준 μ_i 값들은 모두 극값에 해당되어 FLSA에 의해 변화점 식별의 일치성이 보장되는 모형에 해당된다. 반면, 모형 3과 4의 평균 수준 중 일부, 즉 $\mu_i = 1$에 해당되는 부분은 극값이 아니므로 FLSA에 의해 변화점 식별의 일치성이 보장되지 않는다. Figure 3에서 확인할 수 있듯이 μ_i의 누적합 θ_i는 모형 1과 2의 구간별 선형 모형에서는 기울기가 증가하는 구간과 기울기가 감소하는 구간이 반복되는 반면, 모형 3과 4의 구간별 선형 모형에서는 기울기가 계단 형태로 점차 커지는 구간이 포함되어 있는 것을 알 수 있다.

한편, 오차항의 표준편차가 클수록 참변화점 집합을 정확하게 포착하기 어려운 반면, 오차항의 표준편차가 작은 경우에는 참변화점 집합을 정확하게 포착하기 쉬워지는 경향이 있으므로 다양한 잡음 수준에서 변화점 식별의 일치성을 확인해볼 필요가 있다. 이 모의실험에서는 오차항 σ의 표준편차로 0.05, 0.1, 0.2, 0.5, 1.0 등 다섯 가지 경우를 고려하였다.

4.2. 모의실험 결과

이렇게 각 모형마다 관측값 집합을 생성한 후 FLSA와 ℓ_1 추세필터를 적용하여 변화점을 구하는 과정을 1,000번 반복한 후 해경로가 참변화점 집합을 포함하는 횟수를 확인해보았다. 즉, 조절모수 λ_2에 의해 식별된 변화점 집합을 $\widehat{\mathcal{J}}(\lambda_2)$라 하고 참변화점 집합을 $\mathcal{J}$라 할 때 1,000번의 반복실험 중 $\widehat{\mathcal{J}}(\lambda_2) = \mathcal{J}$가 되게 하는 조절모수 λ_2를 찾을 수 있는 경우가 몇 번 있는지 구해보았다. Table 1에는 1,000번의 반복실험 중 $\widehat{\mathcal{J}}(\lambda_2) = \mathcal{J}$인 λ_2를 찾을 수 있었던 실험의 상대돗수가 기록되어있다.

표에서 확인할 수 있듯이 FLSA는 평균수준이 극값으로만 구성되어 있는 모형 1과 2에서 오차항의 표준편차 수준이 $\sigma = 0.05, 0.1, 0.2$ 등으로 낮으면 해경로에 참변화점 집합을 포함할 확률이 1 또는 1에 가까운 값으로 나타났다. 즉, FLSA는 평균수준이 극값으로 이루어져 있고 신호-잡음비(signal to noise ratio)가 클 때 일치성을 보이는 것을 모의실험을 통해 확인할 수 있다. 오차항의 표준편차가 $\sigma = 0.5, 1.0$ 등으로 커지면 해경로가 참변화점 집합을 포함하지 않는 경우가 점차 많아지는 것을 알 수 있다.

이와 반대로 모형 3과 4에서와 같이 평균수준이 극값이 아닌 계단형으로 증가 또는 감소하는 구간을 포함하고 있을 때는 오차항의 표준편차가 작아서 신호-잡음비가 클 경우에도 해경로가 참변화점 집합을 포함할 확률이 0에 가까운 매우 작은 값을 가지는 것을 확인할 수 있다. FLSA의 변화점 식별에서의 일치성과 관련된 이러한 결과는 Son과 Lim (2019)에 소개된 결과와 부합한다.

구간별 선형 모형에 ℓ_1 추세필터를 적용하는 경우에는 모형이나 오차항의 표준편차 σ에 관계없이 해경로가 참변화점 집합을 포함할 확률이 항상 0으로 나타났다. 즉, ℓ_1 추세필터는 FLSA와 달리 구간별 선형 모형의 구조나 오차항의 표준편차 수준에 관계없이 변화점 집합 식별에 있어서의 일치성이 보장되지 않음을 알 수 있다.

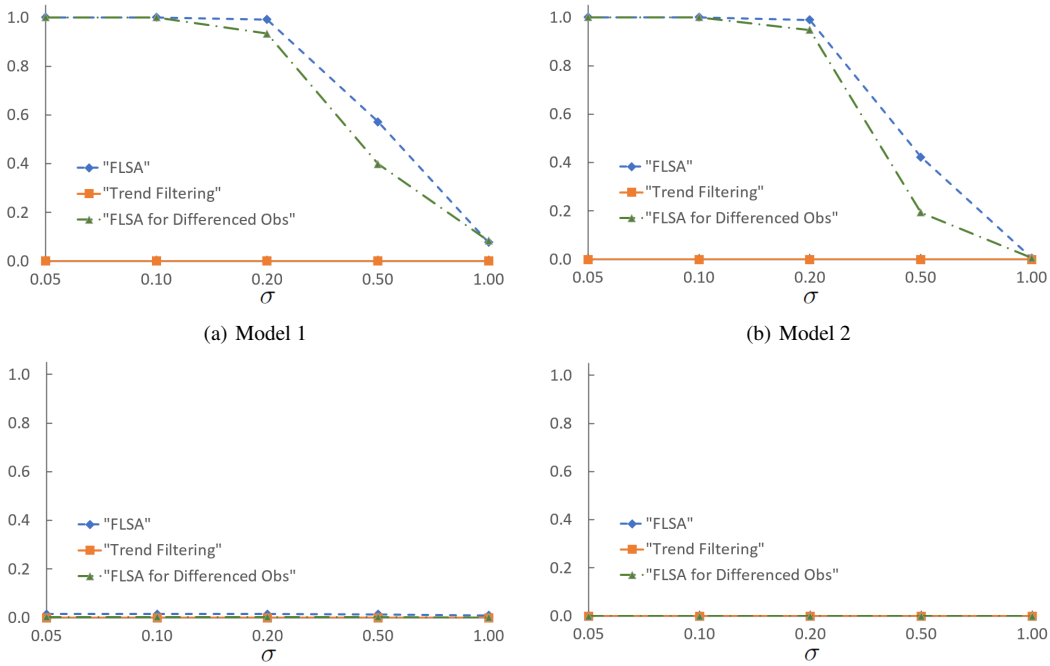
한편, $\theta_i = \sum_{k=1}^{i} \mu_k$에서 $\mu_i = \theta_i - \theta_{i-1}$이므로 z_i를 차분하면

$$w_i = \Delta z_i = z_i - z_{i-1} = (\theta_i + \epsilon_i) - (\theta_{i-1} + \epsilon_{i-1}) = \mu_i + \epsilon_i - \epsilon_{i-1} \tag{4.1}$$

임을 알 수 있다. 따라서 식 (4.1)에서 μ_i가 오차항의 표준편차에 비해 충분히 크다는 가정 하에 차분된 관측값

Table 2: Probability of detecting true change points set of the FLSA for the differenced observation

Model	FLSA for the Differenced Observations				
	$\sigma = 0.05$	$\sigma = 0.10$	$\sigma = 0.20$	$\sigma = 0.50$	$\sigma = 1.00$
1	1.00	1.00	0.93	0.40	0.08
2	1.00	1.00	0.95	0.19	0.01
3	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00



(a) Model 1 (b) Model 2

(c) Model 3 (d) Model 4

Figure 4: *Probability of detecting true change points set for the FLSA and ℓ_1 trend filtering.*

w_i에 대해 FLSA를 적용하면 $\mu_i = \theta_i - \theta_{i-1}$과 $\mu_{i+1} = \theta_{i+1} - \theta_i$가 다른 점, 즉 $\theta_{i+1} - \theta_i \neq \theta_i - \theta_{i-1}$인 점을 찾을 수 있게 된다. 따라서 차분된 관측값에 대해 FLSA를 적용하여 구간별 선형 모형의 변화점을 식별할 수 있을 것으로 기대할 수 있다. 이때 $\boldsymbol{\tau} = \Delta\boldsymbol{\theta}$의 추정량은 다음과 같이 표현할 수 있다.

$$\hat{\boldsymbol{\tau}}^{\mathrm{FL}}(\lambda) = \underset{\boldsymbol{\tau}\in\mathbb{R}^{n-1}}{\arg\min} \left\{ \frac{1}{2}\sum_{i=2}^{n}(w_i - \tau_i)^2 + \lambda \sum_{i=3}^{n} |\tau_i - \tau_{i-1}| \right\}.$$

다만, 시점 i에서의 w_i의 오차항 $\epsilon_i - \epsilon_{i-1}$과 시점 $i+1$에서의 w_{i+1}의 오차항 $\epsilon_{i+1} - \epsilon_i$ 의 공분산은

$$\mathrm{Cov}(w_i, w_{i+1}) = \mathbb{E}[w_i w_{i+1}] = \mathbb{E}[(\epsilon_i - \epsilon_{i-1})(\epsilon_{i+1} - \epsilon_i)] = -\sigma^2$$

이므로 오차항 간의 독립성을 가정하는 통상적인 FLSA와는 차이가 있다. 또, w_i의 표준편차도 $\sqrt{2}\sigma$로 y_i의 표준편차에 비해 큰 것을 알 수 있다.

Table 3: Probability of detecting true change points set in the ±1, ±2, ±3 windows

Model	Window		FLSA					ℓ_1 Trend Filtering					FLSA for the differenced obs.				
		σ	0.05	0.10	0.20	0.50	1.00	0.05	0.10	0.20	0.50	1.00	0.05	0.10	0.20	0.50	1.00
1	0		1.00	1.00	0.99	0.57	0.08	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.93	0.40	0.08
	±1		1.00	1.00	0.99	0.64	0.19	1.00	1.00	1.00	0.65	0.22	1.00	1.00	0.93	0.62	0.40
	±2		1.00	1.00	0.99	0.65	0.24	1.00	1.00	1.00	0.89	0.55	1.00	1.00	0.93	0.62	0.49
	±3		1.00	1.00	0.99	0.65	0.27	1.00	1.00	1.00	0.98	0.93	1.00	1.00	0.93	0.62	0.49
2	0		1.00	1.00	0.99	0.42	0.01	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.95	0.19	0.01
	±1		1.00	1.00	0.99	0.47	0.02	1.00	1.00	1.00	0.88	0.52	1.00	1.00	0.95	0.29	0.04
	±2		1.00	1.00	0.99	0.47	0.03	1.00	1.00	1.00	0.95	0.70	1.00	1.00	0.95	0.29	0.04
	±3		1.00	1.00	0.99	0.47	0.03	1.00	1.00	1.00	0.95	0.70	1.00	1.00	0.95	0.29	0.04
3	0		0.02	0.02	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	±1		0.04	0.04	0.04	0.04	0.03	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01
	±2		0.06	0.06	0.06	0.06	0.05	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01
	±3		0.08	0.08	0.08	0.08	0.07	0.00	0.00	0.00	0.00	0.01	0.02	0.02	0.03	0.03	0.03
4	0		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	±1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	±2		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	±3		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2에는 구간별 선형 모형에서 관찰된 관측값을 차분하여 구한 w_i에 대해 FLSA를 적용하여 구한 해경로 상에 참변화점 집합이 포함되어 있는 경우의 상대돗수가 기록되어 있다. w_i에 대해 FLSA를 적용한 결과는 관측값 y_i에 대해 FLSA를 적용했을 때와 비슷하게 모형 1과 2에서는 오차항의 표준편차가 작을 때 일치성을 보이는 것을 확인할 수 있다. 반면, 모형 3과 4에서는 변화점 식별에 있어서의 일치성이 충족되지 않았다.

Figure 4에는 Table 1과 2에서의 FLSA와 ℓ_1 추세필터의 변화점 식별 성능을 비교하여 제시하였다. 모형 1과 2의 경우 표준편차 σ 값이 작을 때 y_i에 대해 FLSA를 적용한 결과와 차분된 관측값 z_i에 대해 FLSA를 적용한 결과가 비슷한 것을 알 수 있다. 다만, w_i의 표준편차가 y_i에 비해 크기 때문에 표준편차 σ가 $0.2, 0.5$ 등으로 큰 경우에는 변화점 집합을 정확히 찾는 비율이 상대적으로 낮은 것을 볼 수 있다. 모형 3과 4의 경우에는 FLSA와 ℓ_1 추세필터와 마찬가지로 변화점 집합을 정확히 찾는 비율이 0에 가까운 것을 확인할 수 있다.

Table 3은 참변화점 j_i^* 바로 옆의 점들을 포함하는 구간 $[j_i^* - w, j_i^* + w]$ $(w = 0, 1, 2, 3)$에 추정된 변화점이 포함될 확률, 즉 근사적 일치성에 대한 확률이 기록되어 있다. 예를 들어 모형 1인 경우 $w = 1$일 때는 추정된 변화점들이 구간 $[40, 42]$, $[80, 82]$, $[120, 122]$, $[160, 162]$에 각각 하나씩 포함되어 있고 이 구간들 밖에는 추정된 변화점이 존재하지 않을 확률을 의미한다. 확률은 1,000번의 반복실험을 통해 계산하였다.

표에서 확인할 수 있듯이 모형 3과 4와 같이 평균수준 또는 기울기가 계단 형태로 단조증가 또는 감소하는 경우에는 여전히 참변화점 $\pm w$ $(w = 1, 2, 3)$ 구간에 추정된 변화점이 위치할 확률이 0에 가깝게 나타났다. 즉, FLSA와 ℓ_1 추세필터 모두 정확한 변화점 식별이 아닌 근사적인 변화점 식별에 있어서의 일치성도 기대하기 어려운 것을 알 수 있다.

한편, 모형 1과 2와 같이 변화점 식별에 있어서의 일치성을 기대할 수 있는 경우에 ℓ_1 추세필터는 정확한 변화점 집합 식별에 대한 일치성은 보장되지 않지만 참변화점들 근방에 추정된 변화점들이 위치할 확률이 상당히 높은 것으로 나타났다. 즉, ℓ_1 추세필터가 정확한 변화점 집합 식별에 있어서는 성능이 좋지 않지만 구간을 확대하여 참변화점에 가까운 점을 변화점으로 추정한 경우까지 고려하면 FLSA에 비해 변화점 식별

성능이 더 좋은 것으로 판단할 수 있다. 특히 오차항의 표준편차 σ가 큰 경우 ℓ_1 추세필터의 성능이 상대적으로 더 우수한 것으로 관찰되었다. 또 ℓ_1 추세필터의 경우 FLSA에 비해 구간의 폭이 넓어질수록 구간 내에 추정된 변화점이 포함될 확률의 증가폭이 큰 것을 확인할 수 있었다.

5. 결론 및 토의

이 연구에서는 구간별 선형 구조를 가지는 데이터에 ℓ_1 추세필터를 적용할 때의 변화점 식별의 일치성에 대해 살펴보았다. 모의실험 결과 기울기가 계단 형태로 단조증가 또는 단조감소하는 구간이 있는 경우 ℓ_1 추세필터로는 정확한 변화점 식별에 있어서의 일치성 확보가 어렵다는 것을 확인하였다. 이러한 ℓ_1 추세필터의 비일치성은 FLSA에서와 마찬가지로 기울기 차이의 절댓값을 벌점항으로 사용하기 때문인 것으로 판단된다.

또한, 구간별 선형 구조를 가지는 데이터에서 참변화점을 정확하게 식별하기 위해서는 ℓ_1 추세필터를 사용하는 것보다 관측값을 차분하여 FLSA를 적용하는 것이 더 좋다는 것을 모의실험을 통해 확인하였다. ℓ_1 추세필터에 의해 변화점이 정해지면 구간별 선형 모형의 기울기 등도 유일하게 결정되므로 변화점 식별에 있어서의 일치성이 보장되지 않으면 구간별 선형 모형이 정확하게 추정되지 않은 것으로 볼 수 있다. 따라서 구간별 선형 모형을 정확하게 추정하기 위해서는 관측값을 차분한 후 FLSA를 이용하여 변화점을 먼저 식별하고 이 변화점 집합을 이용하여 구간별 선형 모형을 추정하는 것이 바람직한 것을 알 수 있다. 한편, 일치성의 의미를 확장하여 추정된 변화점이 참변화점 주변에 위치하는 근사적인 일치성을 고려하면 ℓ_1 추세필터의 성능이 FLSA에 비해 좋은 경우도 있는 것으로 나타났다. 다만, 이 연구에서 확인한 결과는 모의실험에 근거한 것으로 ℓ_1 추세필터를 이용할 때 비일치성이 나타나는 원인을 정확하게 파악하기 위해서는 이론적 측면의 후속 연구가 필요하다.

또, 이 연구에서는 변화점 식별에 있어서의 일치성을 확인하기 위해 해경로에 참변화점이 포함되어 있는지에 대해서만 고려하였다. 실제 데이터 분석에 있어서는 최적의 변화점 집합을 찾을 필요가 있는데 최적 변화점 집합의 선택을 위해서는 벌점항의 조절모수 선택을 위한 절차가 필요하다. 특히 ℓ_1 추세필터의 경우 더 복잡한 해경로를 가지고 있으므로 이 해경로에 포함되어 있는 집합들 중에서 가장 좋은 변화점 집합을 찾기 위한 방법을 개발하는 것이 중요한 문제인 것으로 판단된다.

References

Friedman J, Hastie T, Höfling H, and Tibshirani R (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, **1** (2), 302–332.

Fryzlewicz P (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, **42** (6), 2243–2281.

Hoefling H (2010). A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, **19** (4), 984–1006.

Hodrick RJ and Prescott EC (1997). Postwar US business cycles: an empirical investigation. *Journal of Money, credit, and Banking*, **29** (1), 1–16.

Kim SJ, Koh K, Boyd S, and Gorinevsky D (2009). ℓ_1 trend filtering. *SIAM Review*, **51** (2), 339–360.

Olshen AB, Venkatraman ES, Lucito R, and Wigler M (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5** (4), 557–572.

Pettitt AN (1979). A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **28** (2), 126–135.

Rojas CR and Wahlberg B (2014). On change point detection using the fused lasso method. arXiv preprint

arXiv:1401.5408.

Son W and Lim J (2019). Modified path algorithm of fused lasso signal approximator for consistent recovery of change points. *Jounal of Statistical Planning and Inference*, **200**, 223–238.

Tibshirani R, Saunders M, Rosset S, Zhu J, and Knight K (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, **67** (1), 91–108.

Yao YC (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics and Probability Letters*, **6** (3), 181–189.

Received January 26, 2022; Revised February 16, 2022; Accepted February 19, 2022

ℓ_1 추세필터의 변화점 식별에 있어서의 비일치성

유동현[a], 임요한[b], 손원[1,c]

[a]인하대학교 통계학과, [b]서울대학교 통계학과, [c]단국대학교 정보통계학과

요 약

구간별 상수 구조를 가지는 관측값으로부터 변화점을 식별하기 위해 FLSA가 자주 사용되고 있다. FLSA는 총변동벌점을 이용하기 때문에 평균 수준이 단조성을 가지는 경우에는 변화점 식별에서의 일치성이 보장되지 않는다는 특징이 있다. ℓ_1 추세필터는 오차제곱합과 기울기 차이에 대한 ℓ_1 벌점의 합을 목적함수로 가지는 구간별 선형 구조 추정방법으로 구간별 선형 구조에서의 변화점을 식별하기 위해 활용할 수 있다. 한편, ℓ_1 추세필터의 경우에도 총변동벌점을 이용하므로 FLSA와 마찬가지로 변화점 식별에 있어서 비일치성을 보일 것으로 예상할 수 있는데 이와 관련된 연구는 아직까지 많이 이루어져 있지 않다. 이 연구에서는 모의실험을 통해 구간별 선형 모형에서 변화점을 식별하기 위해 사용되는 ℓ_1 추세필터의 비일치성에 대해 살펴본다.

주요용어: ℓ_1 추세필터, 다중변화점, 융합라쏘신호근사기, 일치성

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1F1A1A01048127(유동현), No. 2021R1A2C1010786(임요한), No. 2020R1F1A1A01051039(손원)).

[1]교신저자: (16890) 경기도 용인시 수지구 죽전로 152, 단국대학교 정보통계학과. E-mail: son.won@dankook.ac.kr