

# Permutation test for a post selection inference of the FLSA

Jieun Choi<sup>a</sup>, Won Son<sup>1, a</sup>

<sup>a</sup>Department of Applied Statistics, Dankook University

---

## Abstract

In this paper, we propose a post-selection inference procedure for the fused lasso signal approximator (FLSA). The FLSA finds underlying sparse piecewise constant mean structure by applying total variation (TV) semi-norm as a penalty term. However, it is widely known that this convex relaxation can cause asymptotic inconsistency in change points detection. As a result, there can remain false change points even though we try to find the best subset of change points via a tuning procedure. To remove these false change points, we propose a post-selection inference for the FLSA. The proposed procedure applies a permutation test based on CUSUM statistic. Our post-selection inference procedure is an extension of the permutation test of Antoch and Hušková (2001) which deals with single change point problems, to multiple change points detection problems in combination with the FLSA. Numerical study results show that the proposed procedure is better than naïve  $z$ -tests and tests based on the limiting distribution of CUSUM statistics.

Keywords: CUSUM statistic, fused lasso signal approximator, multiple change points, permutation test, post-selection inference

---

## 1. 서론

### 1.1. 변화점

시계열 자료와 같이 순차적으로 관측되는 데이터에서는 구조적 변화가 나타나는 경우가 흔히 있다. 이렇게 특정 시점 전후로 데이터 분포가 달라지는 점을 변화점(chage point)으로 정의할 수 있다. Pettitt (1979)는 순차 관측 데이터(sequential data)  $X_1, X_2, \dots, X_{T-1}, X_T, \dots, X_T$ 와  $F_1(x) \neq F_2(x)$ 인 분포 함수  $F_1(x), F_2(x)$ 에 대해

$$X_1, X_2, \dots, X_{T-1} \sim F_1(x),$$

$$X_T, X_{T+1}, \dots, X_T \sim F_2(x),$$

인 점  $X_T$ 를 변화점으로 정의한 바 있다. 순차 관측 데이터에서 변화가 일어난 횟수와 변화 시점, 변화의 폭 등을 정확히 식별하는 것은 중요한 과제 중 하나로 오랜 기간 동안 연구되어왔다.

분포의 구조적인 변화는 평균 수준, 기울기, 분산 등 다양한 측면에서 발생할 수 있는데 이 연구에서는 평균 수준이 구간별 상수(piecewise constant) 형태의 구조를 가지고 오차항들이 서로 독립인 데이터에서의 변화점 식별에 관해 살펴보려 한다. 즉, 평균 모형(mean model),

$$y_i = \mu_i + \epsilon_i, \quad i = 1, 2, \dots, n, \tag{1.1}$$

---

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1F1A1A01051039).

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

에서 대부분의 시점  $i$ 에서  $\mu_{i-1} = \mu_i$ 이고 일부 시점에서만  $\mu_{i-1} \neq \mu_i$ 인 경우에 대해 살펴보고자 한다. 각 시점에서 관측값들은 단 한 번만 관찰된다고 가정한다.

이 연구에서는 변화점이 하나 이상 존재하는 경우에 적용할 수 있는 다중변화점 모형(multiple change point model)을 고려한다. 즉, 아래와 같이 평균 수준이 달라지는 변화점  $j_1, j_2, \dots, j_k$ 들이  $k$ 개라 가정한다.

$$\mu_1 = \dots = \mu_{j_1-1} \neq \mu_{j_1} = \dots = \mu_{j_k-1} \neq \mu_{j_k} = \dots = \mu_n \quad (1.2)$$

관측값의 개수가  $n$ 일 때 단일변화점(single change point) 문제에서는 변화점이 나타날 수 있는 위치가 모두  $n-1$ 개로 비교적 단순한 문제로 볼 수 있지만 다중변화점 문제에서 변화점의 개수를 알 수 없을 때는 고려해야 하는 경우의 수가 급격히 늘어나게 되어 변화점의 정확한 식별이 어려워진다.

## 1.2. 선행 연구

변화점 식별 문제는 오랜 기간 동안 활발히 연구되어온 주제로 변화점의 식별 절차와 관련된 다양한 선행연구가 존재한다. 변화점의 위치를 변수로 본다면 변화점 식별 절차는 여러 개의 변수들 중에서 데이터를 가장 잘 설명하는 변수들을 선택하는 절차로 생각할 수 있다. 즉, 선형모형에서 변수를 선택하는 절차와 변화점 식별 절차는 많은 유사성을 가진다고 볼 수 있다. 실제로 변화점을 식별하기 위해 자주 사용되는 방법들로는 모든 가능한 경우를 확인(exhaustive search)하는 방법, 새로운 변화점의 추가(forward selection) 또는 변화점 제거(backward elimination) 등의 단계적 선택(stepwise selection)을 통해 변화점을 찾는 방법, 벌점(penalization) 모형을 이용하는 방법 등이 있다 (Niu 등, 2016).

이 연구에서 살펴보고자 하는 fused lasso signal approximator (FLSA; Tibshirani 등, 2005)는 총변동 벌점(total variation penalty)이 부여된 최소제곱법을 이용하여 변화점을 식별하는 방법이다. FLSA에서 사용되는 총변동(TV) 벌점  $\sum_{i=2}^n |\mu_i - \mu_{i-1}|$ 은 변화점의 개수에 대한 벌점  $\sum_{i=2}^n I(\mu_{i-1} \neq \mu_i)$ 을 볼록 완화(convex relaxation)한 형태이다. 여기서  $I(P)$ 는 표시함수(indicator function)로 가정  $P$ 가 참일 때는 1, 거짓일 때는 0의 값을 가진다. 변화점의 개수에 대한 벌점은 볼록함수가 아니므로 볼록 최적화(convex optimization)가 불가능한 데 비해 총변동 벌점은 볼록함수로 볼록 최적화가 가능하므로 계산 상의 이점을 기대할 수 있다. 하지만, 이와 같은 계산에 있어서의 장점 대신 총변동 벌점은 변화점 식별의 점근적 일치성(asymptotic consistency)은 보장하지 못한다는 문제점이 있다. 이런 문제점은 변화점 식별을 위해 FLSA를 사용할 때 큰 단점이 될 수 있으며 자세한 내용은 2절에서 다루기로 한다.

## 1.3. 변화점 식별을 위한 가정

식 (1.2)와 같이 표현되는 다중변화점모형에서는 변화점 식별을 위해 필요한 추가적인 가정을 부여하는 경우가 많다. 일반적으로 평균 수준 변화폭과 구간별 상수 구간의 길이, 잡음의 세기(noise level) 등에 대한 가정을 하는데 이러한 가정의 구체적인 내용은 연구 목적이나 분석 대상이 되는 데이터의 성질에 따라 달라질 수 있다. 대체로 평균 수준의 변화폭이 클수록, 또 구간별 상수인 구간의 길이가 길수록 변화점 식별이 쉬워지며 반대로 잡음의 세기가 클수록, 상수 구간의 길이가 짧을수록 정확한 변화점 식별이 어려워진다. 다중변화점 모형에서 자주 고려되는 대표적인 가정들과 그에 따른 결과는 Niu 등 (2016)에 소개되어 있다.

이 연구에서는 다음과 같은 가정 하에 사후추론을 진행한다.

- (A1) 관측값들은 서로 독립이고 동일한 분산을 가지는 정규분포를 따르는 확률변수들이다.
- (A2) 관측값들의 기댓값은 구간별 상수 함수 형태의 구조를 가지고 있다.
- (A3) 각 변화점  $j$ 에서의 평균 수준 변화폭  $|\mu_{j-1} - \mu_j|$ 이 잡음의 세기( $\sigma$ )에 비해 충분히 크다.

FLSA의 점근적 성질 등 이론적인 측면을 연구하기 위해서는 (A3)에 비해 보다 더 구체적인 가정을 부여하기도 하지만 이 연구는 순열검정을 이용한 사후추론에 중점을 두고 있으므로 (A3)와 같이 일반적인 형태의 가정만을 고려하기로 한다.

## 2. Fused lasso signal approximator (FLSA)의 특징

### 2.1. FLSA 개요

FLSA는 다음과 같이 정의된다.

$$\hat{\boldsymbol{\mu}}^{\text{FL}}(\lambda_1, \lambda_2) = \arg \min_{\boldsymbol{\mu} \in \mathbb{R}^n} \left\{ \frac{1}{2} \sum_{i=1}^n (y_i - \mu_i)^2 + \lambda_1 \|\boldsymbol{\mu}\|_1 + \lambda_2 \|\boldsymbol{\mu}\|_{\text{TV}} \right\}. \quad (2.1)$$

여기에서  $\ell_1$ -노름(norm)  $\|\boldsymbol{\mu}\|_1$ 과 총변동 반노름(semi-norm)  $\|\boldsymbol{\mu}\|_{\text{TV}}$ 는

$$\|\boldsymbol{\mu}\|_1 = \sum_{i=1}^n |\mu_i|, \quad \|\boldsymbol{\mu}\|_{\text{TV}} = \sum_{i=2}^n |\mu_i - \mu_{i-1}|,$$

과 같이 표현된다. 식 (2.1)은 오차제곱합과  $\ell_1$  및 TV-벌점항이 결합된 형태로  $\ell_1$  벌점항은 0이 아닌 값이 드물게 나타나는 희소성(sparsity)을, 총변동 벌점항은 구간별 상수 구조를 구현하는 역할을 한다.

Friedman 등 (2007)은 FLSA의 최적화 과정을 총변동 벌점을 통해 구간별 상수 구조인 평균 수준을 식별하는 과정과  $\ell_1$  벌점을 통해 구간별 상수 구조의 평균 수준들 중 0에 가까운 값들을 0으로 만드는 소프트 임계화(soft-thresholding) 과정으로 나눌 수 있음을 보인 바 있다. 따라서 FLSA를 이용하여 다중변화점을 식별해내기 위해서는  $\ell_1$  벌점 없이 총변동 벌점만 부여되어 있는 형태의 식

$$\hat{\boldsymbol{\mu}}^{\text{F}}(\lambda_2) = \arg \min_{\boldsymbol{\mu} \in \mathbb{R}^n} \left\{ \frac{1}{2} \sum_{i=1}^n (y_i - \mu_i)^2 + \lambda_2 \|\boldsymbol{\mu}\|_{\text{TV}} \right\} \quad (2.2)$$

만으로도 충분하며 이 연구에서도 이와 같은 형태의 목적함수를 고려한다.

### 2.2. 변화점 탐색에 있어서의 특징

FLSA에 의해 식별되는 다중변화점의 개수는 식 (2.2)에 포함된 조절모수  $\lambda_2$ 의 값에 따라 결정된다.  $\lambda_2$ 가 0에 가까운 경우는 TV-반노름보다  $\text{RSS} = \sum_{i=1}^n (y_i - \mu_i)^2$  값을 작게 하는 것이 유리하므로 많은 점들을 변화점으로 식별하게 되며, 반대로  $\lambda_2$ 의 값이 큰 경우에는 RSS보다 TV-반노름을 작게 하는 것이 유리하므로 상대적으로 적은 점들만 변화점으로 판정하게 된다. 주어진 관측값  $\{y_i\}_{i=1}^n$ 에 대해 총변동벌점의 계수  $\lambda_2$ 를 크게 함에 따라 FLSA에 의해 선택되는 변화점의 개수가 점차 줄어들게 되며 구간별 상수 구간의 폭은 길어지게 된다.

Friedman 등 (2007)은 조절모수  $\lambda_2$ 의 값이 커짐에 따라 변화점이 점차 줄어드는 현상은 단조적으로 이루어지며  $\lambda_2$ 에서 변화점이 아닌 것으로 판정되었던 점은  $\lambda_2' > \lambda_2$ 인  $\lambda_2'$ 에 대해서도 항상 변화점이 아닌 것으로 판정됨을 보인 바 있다. 이와 같은 단조결합성질(monotone fusion property)을 고려할 때 FLSA는 단계적 선택과 같은 방식으로 변화점을 식별한다고 볼 수 있다. 즉,  $\lambda_2$ 에 의해 결정된 변화점의 집합을  $\widehat{\mathcal{J}}(\lambda_2)$ 라 하면  $\lambda_2^{(k)} < \lambda_2^{(m)}$ 을 만족하는 임의의  $\lambda_2^{(k)}, \lambda_2^{(m)}$ 에 대해 다음과 같은 포함관계

$$\widehat{\mathcal{J}}(\lambda_2^{(k)}) \supset \widehat{\mathcal{J}}(\lambda_2^{(m)}),$$

이 항상 성립한다.

FLSA의 이런 특성은 다중변화점 탐색에 유용하게 활용될 수 있다. 조절모수  $\lambda_2$ 의 값에 따라 변화점들의 집합이 결정되므로  $\lambda_2 = 0$ 에서 시작해서  $\lambda_2$  값을 증가시키거나 반대로 아주 큰  $\lambda_2$  값에서 시작해서 감소시키는 방향으로 진행하면서  $\lambda_2$  값의 변화에 따른 변화점 집합의 경로를 파악할 수 있다. 이렇게 경로 알고리즘(path algorithm)을 활용하면 수치적인 방법을 사용하지 않고도 어떤  $\lambda_2$  값에서 변화점들이 사라지게 되는지 정확히 알 수 있다 (Hoefling, 2010; Son과 Lim, 2019). 경로 알고리즘은 계산 상의 효율성도 높아 관측값의 개수가  $n$ 개일 때  $O(n \log n)$ 의 복잡도로 전체 변화점 집합들의 경로를 확인할 수 있다.

반면, FLSA는 위에서 살펴본 바와 같이 단계적 선택 방식과 같은 방식으로 변화점을 탐색하기 때문에 변화점 탐색에 있어서의 점근적 일치성(consistency)을 보장할 수 없다는 단점이 있다.  $n$ 개의 관측값들이 있을 때 가능한 변화점 집합은 모두  $2^{n-1}$ 개이지만 FLSA는 단조결합성질에 따라 이 중에서  $n$ 개의 변화점 집합만 경로에 포함할 수 있다. 좀더 엄밀하게 표현하면 평균 수준이 계단(staircase)식으로 증가하거나 감소하는 형태로 분포되어 있을 때에는 FLSA가 변화점 선택에 있어서 불일치성(inconsistency)을 보인다는 사실이 많은 연구자들에 의해 밝혀져 있다 (Rinaldo, 2014; Rojas와 Wahlberg, 2014; Son과 Lim, 2019).

### 2.3. 거짓 변화점 검정의 필요성

식 (1.2)에서 참 변화점들의 집합을  $\mathcal{J}^*$ 라 하자. 평균 수준이 계단 형태로 분포되어 있을 때 FLSA의 불일치성은,

$$\mathbb{P}[\#\lambda_2 \text{ s.t. } \mathcal{J}^* = \widehat{\mathcal{J}}(\lambda_2)] > 0 \quad \text{or} \quad \mathbb{P}[\exists \lambda_2 \text{ s.t. } \mathcal{J}^* = \widehat{\mathcal{J}}(\lambda_2)] < 1 \quad \text{as } \sigma_n \rightarrow 0$$

과 같이 표현할 수 있다. 이 식은 FLSA를 이용해서 변화점들을 탐색할 때 오차항의 표준편차  $\sigma_n$ 이 0으로 수렴하더라도 참 변화점의 집합  $\mathcal{J}^*$ 만 정확하게 식별해낼 수 있는 총변동별점의 조절모수  $\lambda_2$  값을 항상 찾아내는 것은 불가능하다는 것을 의미한다. 즉, FLSA에 의해 선택된 변화점 집합  $\widehat{\mathcal{J}}(\lambda_2)$ 가 참 변화점을 모두 포함하도록 총변동 별점의 조절모수  $\lambda_2$ 를 선택하였을 때는 적어도 하나 이상의 거짓 변화점이 포함되어 있을 확률이 0보다 크게 된다. 반대로 FLSA에 의해 선택된 변화점 집합에 거짓 변화점이 포함되어 있지 않도록 하려면 참 변화점 중에서 식별하지 못한 것이 적어도 하나 이상 존재하게 될 확률이 0보다 크게 된다.

이렇게 FLSA는 본질적으로 점근적 불일치성을 보이므로 변화점 식별에 FLSA를 이용하기 위해서는 FLSA에 의해 선택된 변화점 집합에 포함되어 있는 점들이 참 변화점인지 거짓 변화점인지 판단하기 위한 추가적인 검정절차가 필요하다.

## 3. 사후추론

### 3.1. 일반적인 $z$ -검정 또는 $t$ -검정의 문제점

이제 FLSA를 이용하여 찾아낸 변화점들이 참 변화점인지 거짓 변화점인지 판별하는 가설검정 과정에 대해 생각해보자. 가설검정 단계에서는 처음 설정한 모형 (1.1)에 따라 오차항이 서로 독립인 정규분포를 따른다고 가정한 다음 귀무가설 아래에서 검정통계량의 분포를 구해 볼 수 있다. 하지만 이런 방식의 가설검정은 변화점 식별과 검정통계량 계산이 동일한 데이터를 기반으로 이루어지기 때문에 왜곡된 결과를 나타내기 쉽다. 변화점 식별 과정에서 선택된 변화점은 랜덤으로 선택된 점이 아니고 특정한 조건을 만족시키는 점이기에 때문에 변화점을 중심으로 나누어진 각 구간의 관측값들이 당초에 설정된 모형의 가정에 부합한다고 보기 어렵다.

예를 들어 변화점 식별을 위해 사용되고 있는 대표적 통계량 중 하나인 CUSUM 통계량을 이용하여 변화점을 탐색하는 과정을 살펴보자. 가정 (A1), (A2)가 성립하고 모분산이 주어지지 않는 경우에 변화점이 존재하지 않는다는 귀무가설과 변화점이 하나 존재한다는 대립가설 아래에서 가능도비(likelihood ratio) 검정통계량이 CUSUM 통계량 절댓값의 최댓값으로 주어진다 (Sen과 Srivastava, 1975; Olshen 등, 2004). CUSUM 통계량은

CBS (Olshen 등, 2004), WBS (Fryzlewicz, 2014), double CUSUM 통계량 (Cho, 2016) 등 최근까지도 많은 연구에 폭넓게 활용되고 있다.

CUSUM 통계량을 이용하여 변수 추가 방식으로 변화점을 탐색하는 과정은 다음과 같다. 관측값  $\mathbf{y} = \{y_i\}_{i=1}^n$ 에 대해  $j+1$ 번째 점이 변화점이라고 가정하면 두 구간  $[1, j+1]$ 과  $[j+1, n]$ 에서의 표본평균

$$\bar{y}_{[1, j+1]} = \frac{\sum_{i=1}^j y_i}{j}, \quad \bar{y}_{[j+1, n]} = \frac{\sum_{i=j+1}^n y_i}{(n-j)}$$

에 대해 CUSUM 통계량은,

$$u(\mathbf{y}, j+1) = \sqrt{\frac{n-j}{nj}} \sum_{i=1}^j y_i - \sqrt{\frac{j}{n(n-j)}} \sum_{i=j+1}^n y_i = \frac{\bar{y}_{[1, j+1]} - \bar{y}_{[j+1, n]}}{\sqrt{1/j + 1/(n-j)}} \quad (3.1)$$

과 같이 주어진다. 식 (3.1)에서 확인할 수 있듯이 CUSUM 통계량은 이표본  $z$ -검정 또는  $t$ -검정에서 개별 관측값의 모분산  $\sigma^2$  또는 분산 추정량  $\hat{\sigma}^2$ 이 생략된 형태로 볼 수 있다.

CUSUM 통계량을 기반으로 추정된 변화점은  $n-1$ 개의 점  $j = 2, 3, \dots, n$  중에서,

$$u(\mathbf{y}, j) = \sqrt{\frac{n-j+1}{n(j-1)}} \sum_{i=1}^{j-1} y_i - \sqrt{\frac{j-1}{n(n-j+1)}} \sum_{i=j}^n y_i = \frac{\bar{y}_{[1, j]} - \bar{y}_{[j, n]}}{\sqrt{1/(j-1) + 1/(n-j+1)}}$$

의 절댓값을 최대로 만드는  $j$ 로 정의할 수 있다. 이 과정에서 선택된 변화점은 두 구간  $[1, j]$ 와  $[j, n]$ 에서의 표본평균  $\bar{y}_{[1, j]}$ 와  $\bar{y}_{[j, n]}$ 의 차이가 상대적으로 큰 점임에 유의할 필요가 있다. 변화점  $j$ 가 이런 과정을 통해 선택되므로 구간  $[1, n]$ 에서 모평균  $\mu_i$ 에 차이가 없다고 하더라도 단순한  $z$ -검정 또는  $t$ -검정으로는 두 구간의 평균 수준에 차이가 있는 것으로 판단하게 되는 경우가 있을 수 있다. 즉, 참 변화점이 아닌 점을 변화점으로 판단하는 검정 오류가 나타날 수 있다.

Figure 1은 정규분포  $N(0, 0.1^2)$ 에서 21개의 난수를 발생시켜 만든 데이터에 대해 CUSUM 통계량을 적용하여 변화점을 하나 찾고 식별된 변화점에 의해 나누어진 두 구간의 표본평균이 서로 같다고 볼 수 있는지  $z$ -검정하는 과정을 1,000번 반복하여 얻은 결과이다. 그림에서 살펴볼 수 있듯이 평균값이  $\mu_i = 0 (i = 1, 2, \dots, 21)$ 으로 모두 동일한 분포에서 난수를 발생시켰음에도 불구하고 1,000번의 반복실험 중에서  $p$ -값이 0.05보다 작은 경우가 모두 349번 나타났다. 즉, 변화점을 기준으로 두 구간의 표본평균이 서로 다르다고 판정되는 경우가 이론적으로 기대할 수 있는 것보다 상당히 많은 것을 알 수 있다. 이렇게 사후추론에  $z$ -검정을 사용하면 검정 결과에 오류가 생길 가능성이 높다는 것을 확인할 수 있다.

따라서 식별된 변화점이 참 변화점인지 확인하기 위한 가설검정을 위해서는 검정통계량의 표본분포를 구하는 것이 중요한 문제이다.  $z$ -검정 대신 검정통계량, 즉 CUSUM 통계량 절댓값의 최댓값 표본분포를 이용한 가설검정도 고려해 볼 수 있다. CUSUM 통계량 절댓값의 최댓값 분포는 극한분포(limiting distribution)를 기반으로 근사적으로 구할 수도 있다. 하지만 극한분포는 수렴속도가 느리기 때문에 관측값의 개수  $n$ 이 작은 경우에는 정확한 검정 결과를 기대하기 어렵다는 사실이 알려져 있고 이러한 문제점을 극복하기 위한 방법으로 순열검정이 제안된 바 있다 (Antoch와 Hušková, 2001).

### 3.2. FLSA의 사후추론에 대한 선행 연구

위에서 살펴본 것과 같이 FLSA를 이용한 변화점 식별 후에 변화점 식별에서와 동일한 관측값에 대해 가설검정이 이루어지는 경우 변화점 식별 과정에서 이루어진 오발견(false discovery)을 검정 과정에서 찾아내기 어려움을 알 수 있다. 이런 현상은 변화점 식별 문제에서뿐만 아니라 일반적인 모형 선택과 모형 선택 후의 추론과정에서 흔히 나타나며 Leeb과 Pötscher (2003, 2006) 이후 이러한 문제점을 해결하기 위한 사후추론

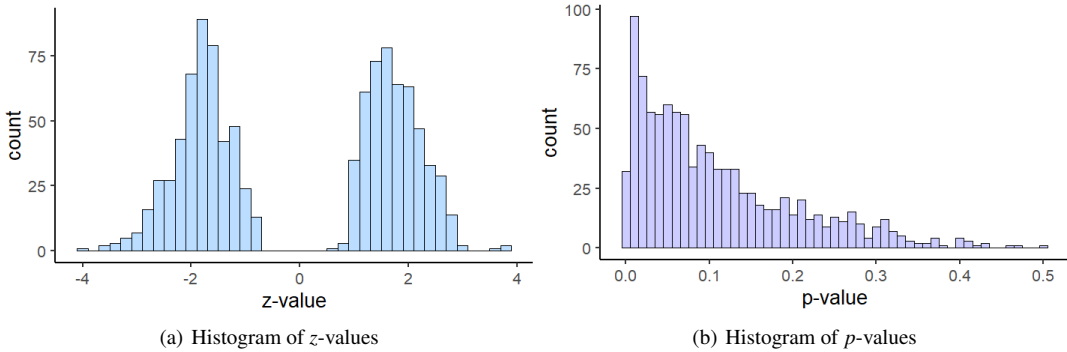


Figure 1: Results of Ordinary  $z$ -test for the Hypothesis  $H_0$  : no difference in mean values (1,000 Repetitions).

(post-selection inference) 절차를 개발하기 위한 많은 연구가 이루어지고 있다. 대표적 연구 결과 중 하나로 Lee 등 (2016)은 LASSO를 이용하여 모형을 선택한 상태에서 관측값들의 선형대비(linear contrasts)의 조건부 분포를 구하고 이 조건부 분포를 이용하여 사후검정을 진행하는 절차를 제안한 바 있다.

Hyun 등 (2018)은 일반화된 LASSO (generalized LASSO)의 사후추론을 위해 Lee 등 (2016)의 방식을 적용하여 구한 조건부 분포를 제시한 바 있다. FLSA는 일반화된 LASSO (generalized LASSO)의 특수한 형태 중 하나이므로 Hyun 등 (2018)에서 제안된 절차를 FLSA에서 발생하는 거짓 변화점을 제거하기 위한 사후검정 방법으로 사용할 수 있다. 한편, Hyun 등 (2018)에서 제안된 방법은 Lee 등 (2016)에서와 마찬가지로 귀무가설이 선형대비 형태로 주어진 경우에 적용될 수 있다. Hyun 등 (2018)에서는 변화점  $j$  주변의 모수만 이용하는 스파이크 대비(spike contrast),

$$H_0 : \mu_{j-1} = \mu_j \quad \text{vs} \quad H_1 : \mu_{j-1} \neq \mu_j$$

와 식별된 변화점으로 생성된 상수 구간에 해당되는 전체 모수의 평균을 이용하는 구간 대비(segment contrast)

$$H_0 : \bar{\mu}_{[1:j-1]} = \bar{\mu}_{[j:n]} \quad \text{vs} \quad H_1 : \bar{\mu}_{[1:j-1]} \neq \bar{\mu}_{[j:n]},$$

형태의 가설을 고려하고 있다.

## 4. 순열검정을 이용한 FLSA의 사후추론

### 4.1. 단일변화점 문제에서의 순열검정

3.1절에서 살펴본 바와 같이 단일변화점 문제에서 순열검정을 사용하여 변화점을 검정하는 방법은 Antoch와 Hušková (2001)에 의해 제안된 바 있다. 식별된 변화점을  $j$ 라고 할 때 Antoch와 Hušková (2001)의 검정가설은,

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_{j-1} = \mu_j = \dots = \mu_n \quad \text{vs} \quad H_1 : \mu_1 = \mu_2 = \dots = \mu_{j-1} \neq \mu_j = \dots = \mu_n \quad (4.1)$$

로 표현할 수 있다. Antoch와 Hušková (2001)는 식 (3.1)에서 정의된 CUSUM 통계량의 분모에 분산 추정값  $\hat{\sigma}_n^2$ 이 포함된 통계량,

$$T_j = \frac{\bar{y}_{[1,j]} - \bar{y}_{[j,n]}}{\hat{\sigma}_n \sqrt{1/(j-1) + 1/(n-j+1)}} \quad (j = 2, 3, \dots, n) \quad (4.2)$$

을 검정통계량으로 사용하였는데 분산추정량을  $\hat{\sigma}_n^2 = 1/n \sum_{i=1}^n (y_i - \bar{y}_n)^2$ 으로 정의하였으므로 변화점의 위치가 분산추정량  $\hat{\sigma}_n^2$ 과 검정통계량  $T$ 에 영향을 미치지 않는다. Antoch와 Hušková (2001)의 검정절차는 다음과 같이 요약할 수 있다.

1. 식별된 변화점  $\hat{j}$ 을 기준으로 통계량  $T_j$ 의 절댓값  $t_0$ 를 구한다.
2. (4.1)의 귀무가설 아래에서는 관측값들이 모두 동일한 분포를 따른다는 점에 착안하여 관측값들을 랜덤으로 재배열한다.
3. 2에서 재배열된 관측값들에서 통계량  $T_j$  ( $j = 2, \dots, n$ )의 절댓값이 가장 큰 점을 하나 식별한다.
4.  $i$ 번째 재배열에서 과정 3을 통해 식별한 변화점에서의 통계량  $T_j$ 의 절댓값을  $t_i$ 라 한다.
5. 2~4의 과정을  $K$ 번 반복하여  $K$  개의 통계량  $t_1, t_2, \dots, t_K$ 의 분포를 구한다.
6. 5에서 구한  $t_1, t_2, \dots, t_K$ 의 분포와 1에서 구한  $t_0$ 를 비교하여 가설을 검정한다.

순열검정을 위해서는 많은 반복 계산이 필요하다는 단점이 있지만 이해하기 쉽고 분포 가정에 덜 민감하며 극한분포를 이용하는 경우에 비해 정확한 가설검정을 할 수 있다는 장점이 있다. 한편 식 (4.1)에서 주어진 검정가설은 대비 형태로 표현할 수 없다는 점에서 Hyun 등(2018)이 제시한 스파이크 또는 구간 형태의 검정가설과는 차이가 있다.

#### 4.2. 순열검정을 이용한 FLSA의 사후추론(post-selection inference)

이 연구에서는 Antoch와 Hušková (2001)의 순열검정절차를 확장하여 다중변화점 식별 방법인 FLSA의 사후검정절차를 제안하려 한다.

먼저 FLSA에 의해 식별된 변화점들이 참 변화점(true change points)인지 거짓 변화점(false change points)인지 판별하기 위한 가설을 설정해보자. FLSA의 특성상 나타나기 쉬운 거짓 변화점을 식별해서 제거하고 참 변화점은 유지하는 데 관심이 있으므로 귀무가설과 대립가설은,

$$H_0 : j \notin \mathcal{J}^* \quad \text{vs} \quad H_1 : j \in \mathcal{J}^*$$

와 같이 설정할 수 있다. 이 때  $j$ 는 FLSA에 의해 선택된 변화점의 집합  $\widehat{\mathcal{J}}$ 의 원소이며,  $\mathcal{J}^*$ 는 참 변화점들의 집합을 의미한다. 즉, 위 식에서 귀무가설은 FLSA에 의해 추정된 변화점  $j$ 가 거짓 변화점임을, 대립가설은 FLSA에 의해 추정된 변화점  $j$ 가 참 변화점임을 의미한다.

귀무가설  $H_0$  아래에서는  $j$ 를 기준으로 나누어진 두 구간  $\hat{B}_{j-}$ 와  $\hat{B}_{j+}$ 의 평균 수준이 서로 같다고 가정할 수 있다. 따라서 귀무가설은,

$$H_0 : i_1, i_2 \in \hat{B}_{j-} \cup \hat{B}_{j+} \text{인 모든 } i_1, i_2 \text{에 대해 } \mathbb{E}[y_{i_1}] = \mathbb{E}[y_{i_2}] \quad (4.3)$$

로 표현할 수 있다. 반대로 대립가설 아래에서는 두 구간  $\hat{B}_{j-}$ 와  $\hat{B}_{j+}$ 의 평균 수준이 다르고 각 구간 내에서만 평균 수준이 동일함을 가정한다. 대립가설은,

$$H_1 : i_1 \in \hat{B}_{j-}, i_2 \in \hat{B}_{j+} \text{인 모든 } i_1, i_2 \text{에 대해 } \mathbb{E}[y_{i_1}] \neq \mathbb{E}[y_{i_2}] \quad (4.4)$$

로 표현된다.

한편, 식 (4.3)의 가정은 FLSA에 의해 식별된 변화점의 집합이 참 변화점을 모두 포함할 때 성립하는 것으로 참 변화점 중 식별되지 못한 변화점이 있는 경우에는 식 (4.3)이 성립되지 않아 가설검정절차에 문제가 발생할 수 있다. 이런 점을 고려하여 식별된 변화점 집합  $\widehat{\mathcal{J}}$ 이 참 변화점의 집합  $\mathcal{J}$ 를 포함( $\mathcal{J} \subset \widehat{\mathcal{J}}$ )할 수 있도록 FLSA의 조절모수  $\lambda_2$ 를 선택할 필요가 있다. 변화점 집합  $\widehat{\mathcal{J}}$ 을 찾기 위해서는 BIC 등의 정보기준(information criteria)을 고려할 수 있으며 Son 등 (2021)은 모든 참 변화점을 포함하고 거짓 변화점을 최소로 포함하는 변화점 집합을 찾기 위한 선택 기준으로,

$$\text{GIC} = \log \left( \frac{\text{RSS}_n(\widehat{\mathcal{J}})}{n} \right) + \frac{\kappa_n}{n} |\widehat{\mathcal{J}}| \quad (4.5)$$

과 같은 형태의 일반화 정보기준(generalized information criteria, GIC)을 제안한 바 있다. 식 (4.5)에서  $RSS_n$ 은 잔차제곱합을 의미하고  $\kappa_n = \kappa \log n$ 으로 정의된다.  $\kappa \geq 1$ 인 경우  $\sigma_n \rightarrow 0$ 일 때 참 변화점을 모두 포함하고 거짓 변화점을 최소로 포함하는 변화점 집합을 찾을 수 있는 확률이 1로 수렴함이 알려져 있다 (Son 등, 2021).

귀무가설 아래에서는 식 (4.3)과 같이 집합  $\hat{B}_{j-}$ 와  $\hat{B}_{j+}$ 에 포함된 관측값들이 동일한 평균모형(mean model)을 따른다고 가정하고 있으므로 순열검정을 적용할 수 있다. 이 연구에서는 CUSUM 통계량 (3.1)을 기반으로 하는 다음과 같은 순열검정절차를 제안한다.

1. FLSA에 의해 식별된 변화점  $j$ 에 의해 나누어진 두 구간  $\hat{B}_{j-}, \hat{B}_{j+}$ 에서의 통계량  $u_0 = |\bar{y}_{\hat{B}_{j-}} - \bar{y}_{\hat{B}_{j+}}| / \sqrt{\frac{1}{|\hat{B}_{j-}|} + \frac{1}{|\hat{B}_{j+}|}}$  값을 구한다.
2. 구간  $\hat{B}_{j-} \cup \hat{B}_{j+}$ 에 포함되어 있는 관측값들을 랜덤으로 재배열한다.
3. 2에 의해 재배열된 관측값들에 대해 각 점마다 CUSUM 통계량을 구하고 CUSUM 통계량의 절댓값이 가장 큰 점  $j'$ 을 찾아 새로운 변화점으로 한다.
4.  $i$ 번째 재배열에서 과정 3을 통해 식별된 변화점  $j'$ 에서  $u = |\bar{y}_{\hat{B}_{j'-}} - \bar{y}_{\hat{B}_{j'+}}| / \sqrt{\frac{1}{|\hat{B}_{j'-}|} + \frac{1}{|\hat{B}_{j'+}|}}$  값을 계산하여  $u_i$ 라 한다.
5. 2~4의 과정을  $K$ 번 반복하여  $K$  개의 통계량  $u_1, u_2, \dots, u_K$ 의 표본분포를 구한다.
6. 1에서 구한  $u_0$  값이 5에서 구한 표본분포에서 차지하는 위치를 이용하여 가설을 검정한다.

순열검정 절차에서 1과 4의 과정에서는  $t$ -통계량 대신 모분산 추정치  $\hat{\sigma}^2$ 이 소거된 형태의 CUSUM 통계량  $u$ 를 이용하여 검정을 실시하고 있다. 이렇게  $t$ -통계량 대신 CUSUM 통계량을 이용하는 것은 모분산 추정치  $\hat{\sigma}^2$ 이 동일하다는 가정 (A1)에 근거하는 것으로  $t$ -통계량의 분산을 추정하는 과정에서 발생하는 오차로 인해 검정 결과가 영향을 받는 것을 방지할 수 있다는 장점도 있다. 또, 순열검정 과정에서 식별된 변화점에 의해 나누어진 구간의 길이가 1인 경우에는  $t$ -통계량의 합동표본표준편차(pooled sample standard deviation)  $s_p$ 를 구할 수 없지만 이 경우에도 CUSUM 통계량 값을 구할 수 있다는 장점도 있다.

## 5. 모의실험

이 연구에서는 다음과 같은 평균모형을 가정하여 모의실험을 진행하려 한다.

- $\mu_1, \dots, \mu_{20} = 1, \mu_{21}, \dots, \mu_{40} = 0, \mu_{41}, \dots, \mu_{70} = 1, \mu_{71}, \dots, \mu_{100} = 2$
- $\sigma = 0.1, 0.2, 0.5$
- 순열검정에서 검정통계량의 표본분포를 구하기 위한 반복횟수  $K = 1000$
- 가설검정 횟수  $B = 1000$

기댓값이  $\mu_1, \dots, \mu_{20} = 1, \mu_{21}, \dots, \mu_{40} = 0, \mu_{41}, \dots, \mu_{70} = 1, \mu_{71}, \dots, \mu_{100} = 2$ 로 참 변화점은 모두 3개 ( $i = 21, 41, 71$ )이고 모표준편차  $\sigma = 0.1, 0.2, 0.5$ 인 평균모형을 고려한다. 이 평균모형에서 난수를 발생시켜 관찰한 관측값들에 대해 FLSA를 이용하여 다중 변화점을 식별한다. 이 때 FLSA의 조절모수는 변화점 집합 경로 상의 변화점 집합 중 모든 참 변화점들을 포함하고 거짓 변화점은 최소한으로 포함한 집합이 되도록 선택하였다. 이렇게 식 (4.5)에 제시된 GIC를 사용하지 않고 경로 알고리즘의 결과가 주어진 상태에서 사후적으로 최적의 조절모수를 선택한 이유는 GIC 형태의 정보기준을 사용하더라도 경우에 따라서, 특히 오차항의 분산이 큰 경우에는 참 변화점을 제대로 식별하지 못하여 참 변화점 중 일부가 누락된 변화점 집합이 선택될 수 있기 때문이다. 참 변화점이 누락된 경우에는 모의실험에서 관찰하고자 하는 각 가설검정 방법의 성능을 정확히 파악하기 어렵게 된다. 모의실험이 아닌 실제 데이터 분석 과정에서는 참 변화점과 거짓 변화점을 정확히 알 수 없으므로 앞서 언급한 바와 같이 식 (4.5)의 GIC를 이용하여 조절모수를 선택할 수 있다.



Table 1: Confusion matrix of predicted true or false change points

Noise level	Predicted signals	True signals					
		z-test		CUSUM test		Permutation test	
		P	N	P	N	P	N
$\sigma = 0.1$	P	2391	25	2391	135	2391	4
	N	0	1667	0	1557	0	1688
$\sigma = 0.2$	P	2248	35	2201	130	2233	7
	N	0	1809	47	1714	15	1837
$\sigma = 0.5$	P	1028	393	879	1368	787	100
	N	330	3358	479	2383	571	3651

P: change point, N: non-change point

순열검정은 다음과 같은 방식으로 실시한다. 먼저 원래 관측값에 FLSA를 적용하여 구한 변화점들과 각 변화점마다의 CUSUM 통계량  $u_0$  값을 계산한다. 다음으로 4.2절의 2~5의 과정을 적용하여 재배열된 관측값들에 대해 CUSUM 통계량  $u$ 의 표본분포를 계산하고  $u_0$  값과 비교하여 순열검정을 실시한다. 이 때 표본분포를 구하기 위한 반복횟수( $K$ )는 1000으로 한다.

제한된 순열검정절차의 성능을 확인해보기 위해  $z$ -검정과 CUSUM 통계량 절댓값의 최댓값 극한분포를 이용한 가설검정 결과와 비교해 보았다. 표현 상의 편의를 위해 이하에서는 CUSUM 통계량 절댓값의 최댓값 극한분포를 이용한 검정절차를 간략하게 CUSUM-검정이라 부르기로 한다. 유의수준은 5%로 설정하였으며 CUSUM-검정을 위해서는 R 패키지 “CPAT”를 사용하였다.

Table 1에는 순열검정과 CUSUM-검정,  $z$ -검정을 각각  $B(1000)$ 번 반복한 결과가 혼동행렬(confusion matrix) 형태로 정리되어 있다.  $B$ 개의 평균모형에는 각각 3개씩의 참 변화점이 있으므로 참 변화점에 대한 검정은 모두 3,000번 이루어진다. 또한, FLSA에 의해 변화점으로 식별된 거짓 변화점에 대한 가설검정도 이루어지는데 거짓 변화점의 수는 관측할 때마다 달라지게 된다. 한편, FLSA에 의해 식별된 변화점으로 분할된 두 구간 중 길이가 1인 구간이 있는 경우 순열검정이 원활하게 이루어지지 않는다. 길이가  $m$ 인 구간과 길이가 1인 두 구간에 대해 순열검정을 하는 경우 구간 길이 1인 구간에 속한 점  $y$ 가 랜덤으로 재배열된 구간의 양 끝 점 중 하나에 위치할 확률이,

$$\frac{2 \cdot m!}{(m + 1)!} = \frac{2}{m + 1}$$

이다. 모의실험에서 가정한 평균모형의 구간 길이는  $m \leq 30$ 이므로 표본분포를 구하기 위한  $K = 1000$  번의 반복시행 중 구간 길이가 1인 구간에 속하는 점  $y$ 가 첫번째 또는 마지막에 위치하는 경우가  $1000 \cdot 2 \cdot m! / (m + 1)! = 2000 / (m + 1) \geq 64.5$ 번 이상 발생한다. 식별된 변화점으로 나누어진 구간 중에 길이가 1인 구간이 있는 경우에는 검정통계량과 같은 값을 가지는 CUSUM 통계량이 5% 이상 관찰되므로 유의수준 5%에서 순열검정이 이루어질 수 없다. 이런 점을 고려하여 Table 1에는 길이가 1인 구간이 포함된 경우를 제외한 가설검정 결과만 기록하였다. 예를 들어  $\sigma = 0.1$ 인 경우 참 변화점에 대한 검정은 모두 3,000번 이루어졌지만 그 중에서 한 쪽 구간 길이가 1인 609번의 가설검정을 제외한 2,391번의 검정 결과만 기록되어 있다.

Table 1에서  $z$ -검정과 CUSUM-검정에 비해 순열검정이 FLSA에서의 거짓 변화점을 잘 제거해내어 위양성(false positive)을 보이는 점들이 상대적으로 적음을 알 수 있다.  $\sigma = 0.1$ 인 경우 FLSA에 의해 변화점으로 식별된 4,083개의 점 중 2,391개는 참 변화점, 나머지 1,692개는 거짓 변화점에 해당된다.  $\sigma = 0.1$ 일 때는 참 변화점에서의 평균수준 변화폭 1에 비해 오차항의 표준편차가 작으므로 모든 검정방법들이 2,391개의 참 변화점을 정확하게 식별했지만, 잡음 수준( $\sigma$ )이 커짐에 따라 위양성(false positive)을 보이는 점들이 늘어나며 순열검정에 비해  $z$ -검정과 CUSUM-검정에서 위양성 점들이 상대적으로 더 많아지는 것을 확인할 수 있다.

Table 2: Performance comparison between post selection inference methods

Measures	True signals								
	z-test			CUSUM test			Permutation test		
	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.5$	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.5$	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.5$
TPR	1.000	1.000	0.757	1.000	0.970	0.647	1.000	0.993	0.580
TNR	0.985	0.981	0.895	0.920	0.930	0.635	0.998	0.996	0.973
precision	0.990	0.985	0.723	0.947	0.944	0.391	0.998	0.997	0.887
accuracy	0.994	0.991	0.858	0.967	0.957	0.638	0.999	0.995	0.869
F1 score	0.995	0.992	0.740	0.973	0.961	0.4887	0.999	0.995	0.701

\* TPR (sensitivity) =  $TP/TP+FN$ , TNR (specificity) =  $TN/TN+FP$ , precision =  $TP/(TP+FP)$   
accuracy =  $TP+TN/TP+FP+FN+TN$ , F1 score =  $2TP/2TP+FP+FN$  where true positives (TP),  
false negatives (FN), false positives (FP), true negatives (TN)

Table 1의 혼동행렬을 토대로 세 검정 방법의 성능을 비교하기 위해 구한 여러 가지 지표들이 Table 2에 정리되어 있다. 먼저 오차항의 분산 크기와 관계 없이 전반적으로 특이도(TNR, specificity)와 정밀도(precision)에서 순열검정절차가 다른 검정방법들에 비해 우수함을 확인할 수 있다. 특이도와 정밀도는 FLSA에 의해 변화점으로 식별되었지만 실제로는 변화점이 아닌 거짓 변화점(false positives)을 사후추론에서 변화점이 아닌 것으로 바르게 판단할 수 있는지 확인해 볼 수 있는 지표이므로 순열검정이 위양성 변화점을 제거하는 데 유용함을 이들 지표로부터 확인할 수 있다.

오차항의 분산 크기별로는  $\sigma = 0.1$ 로 분산의 크기가 작을 때는 순열검정 결과가 z-검정과 CUSUM-검정 결과에 비해 모든 지표에서 우월한 것을 확인할 수 있다. 모표준편차의 크기가 커지는 경우 거짓 변화점과 함께 참 변화점도 일부 식별하지 못하게 되므로 세 검정 방법의 지표들이 전반적으로 하락하는 것을 확인할 수 있다. 다만, 이 경우에도 순열검정이 z-검정과 CUSUM-검정에 비해 특이도, 정밀도, 정확도 등에서 우위를 보이고 민감도(TPR, sensitivity)는 상대적으로 낮게 나타났다.

## 6. 결론 및 토의

FLSA는 총변동분해를 이용해 구간별 상수인 평균 구조를 구현할 수 있는 모형으로 다중변화점 탐색을 위해 활용되고 있다. 다만, FLSA는 점근적 일치성이 만족되지 않으므로 오차항의 분산이 매우 작은 경우에도 다수의 거짓 변화점이 여전히 남아 있을 수 있다는 단점이 있다. 이러한 FLSA의 문제점을 해결하기 위한 사후추론 방법으로 이 연구에서는 순열검정절차를 제안하였다. Antoch와 Hušková (2001)에 의해 소개된 단일변화점 모형에서의 순열검정 방법을 다중변화점 식별에 사용되는 FLSA에 적용하여 FLSA의 거짓 변화점들을 제거하는 검정절차를 제시하였다.

제안된 검정절차의 성능을 모의실험을 통해 살펴본 결과, 잡음의 세기가 평균수준의 차이에 비해 상대적으로 작을 때 순열검정이 z-검정과 CUSUM-검정에 비해 거짓 변화점 제거에 뛰어난 성능을 나타내는 것을 확인할 수 있었다. 다만, 잡음 크기가 커질 때에는 순열검정이 참 변화점도 많이 제거하는 오류를 보인다는 점에 유의할 필요가 있다. 변화점들로 나누어진 각 구간의 길이가 충분히 길 때는 CUSUM 통계량의 표본분포를 정확하게 계산할 수 있지만 구간 길이가 짧을 때는 정확한 표본분포를 구할 수 없게 되어 검정결과에 오류가 많이 발생하게 되기 때문이다. 향후 이런 문제점을 해결하기 위한 연구가 필요한 것으로 판단된다.

한편, 이 연구에서 제안된 검정방법이 잘 작동하기 위해서는 FLSA에 의해 식별된 변화점의 집합이 참 변화점을 모두 포함하고 있어야 한다. 따라서 실제 데이터 분석에서 FLSA를 적용하여 최적의 변화점 집합을 잘 찾아내기 위해서는 적절한 조절모수 선택기법이 필요함을 알 수 있다.

## References

- Antoch J and Hušková M (2001). Permutation tests in change point analysis, *Statistics and Probability Letters*, **53**, 37–46.
- Cho H (2016). Change-point detection in panel data via double cusum statistic, *Electronic Journal of Statistics*, **10**, 2000–2038.
- Friedman J, Hastie T, Höfling H, and Tibshirani R (2007). Pathwise coordinate optimization, *The Annals of Applied Statistics*, **1**, 302–332.
- Fryzlewicz P (2014). Wild binary segmentation for multiple change-point detection, *The Annals of Statistics*, **42**, 2243–2281.
- Hoefling H (2010). A path algorithm for the fused lasso signal approximator, *Journal of Computational and Graphical Statistics*, **19**, 984–1006.
- Hyun S, G'Sell M, Tibshirani RJ (2018). Exact post-selection inference for the generalized lasso path, *Electronic Journal of Statistics*, **12**, 1053–1097.
- Lee JD, Sun DL, Sun Y, and Taylor JE (2016). Exact post-selection inference, with application to the lasso, *Annals of Statistics*, **44**, 907–927.
- Leeb H and Pötscher BM (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations, *Econometric Theory*, **19**, 100–142.
- Leeb H and Pötscher BM (2006). Can one estimate the conditional distribution of post-model-selection estimators?, *Annals of Mathematical Statistics*, **34**, 2554–2591.
- Niu YS, Hao N, and Zhang H (2016). Multiple change-point detection: A selective overview, *Statistical Science*, **31**, 611–623.
- Olshen AB, Venkatraman E, Lucito R, and Wigler M (2004). Circular binary segmentation for the analysis of array-based dna copy number data, *Biostatistics*, **5**, 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**, 126–135.
- Rinaldo A (2014). Corrections to properties and refinements of the fused lasso, Retrieved December 8th from [https://www.stat.cmu.edu/arinaldo/Fused\\_Correction.pdf](https://www.stat.cmu.edu/arinaldo/Fused_Correction.pdf)
- Rojas CR and Wahlberg B (2014). On change point detection using the fused lasso method
- Sen A and Srivastava MS (1975). On tests for detecting change in mean, *Annals of Statistics*, **3**, 98–108.
- Son W and Lim J (2019). Modified path algorithm of fused lasso signal approximator for consistent recovery of change points, *Journal of Statistical Planning and Inference*, **200**, 223–238.
- Son W, Lim J, and Yu D (2021). Tuning parameter selection for the fused lasso signal approximator. Manuscript submitted for publication.
- 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**, 91–108.

Received June 1, 2021; Revised July 29, 2021; Accepted August 27, 2021

## 순열검정을 이용한 FLSA의 사후추론

최지은<sup>a</sup>, 손원<sup>1,a</sup>

<sup>a</sup>단국대학교 대학원 응용통계학과

---

### 요 약

FLSA는 총변동별점을 이용해 구간별상수인 평균 구조를 구현하는 별점모형으로 다중변화점 탐색을 위해 활용되고 있다. 한편, FLSA는 변화점 탐색에 있어서 점근적 일치성이 만족되지 않으므로 잡음의 크기가 0에 가깝게 수렴하는 경우에도 다수의 거짓 변화점이 식별될 수 있다는 단점이 있다. 이 연구에서는 이러한 FLSA의 문제점을 해결하기 위한 사후추론 방법으로 순열검정 방법을 제안한다. 단일변화점 모형과 관련된 순열검정 방법은 Antoch와 Hušková (2001)에 의해 제안된 바 있다. 이 연구에서는 Antoch와 Hušková (2001)의 검정절차를 확장하여 다중변화점 식별에 사용되는 FLSA와 결합함으로써 다중변화점 모형에 적용할 수 있는 순열검정절차를 제안한다. 모의실험 결과, 제안된 방법은  $z$ -검정과 CUSUM 통계량의 극한분포에 기반을 둔 검정방법에 비해 전반적으로 우수하였으며 거짓 변화점의 식별에 유용함을 확인할 수 있었다.

주요용어: CUSUM 통계량, 다중변화점, 순열검정, 사후추론, 융합라쏘신호근사기

---

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1F1A1A01051039).

<sup>1</sup>교신저자: (16890) 경기도 용인시 수지구 죽전로 152 단국대학교 정보통계학과, 대학원 응용통계학과.  
son.won@dankook.ac.kr