Permutation Test for a Post Selection Inference of the FLSA

Jieun Choi, Won Son

Introduction

■ 연구 목표

FLSA를 이용하여 식별된 다중 변화점에서 사후추론을 통해 거짓변화점을 탐색

■ 변화점의 정의

시계열 자료와 같이 순차적으로 관측되는 데이터에서 특정 시점 전후로 분포가 달라지는 지점.
본 연구에서는 아래와 같이 평균 수준이 구간별 상수 형태의 구조인 평균모형

$$y_i = \mu_i + \varepsilon_i,\ i = 1,2,\dots,n$$

으로 대부분의 시점 i에서 $\mu_{i-1} = \mu_i$이고, 일부 시점에서만 $\mu_{i-1} \neq \mu_i$인 경우를 가정함. 이 때 변화점이 여러 개인 다중 변화점을 고려하였음.

■ 변화점 식별을 위한 가정

다중변화점 모형에서는 변화점 식별을 위해 추가적인 가정을 부여하는데, 아래 세개의 가정 하에 사후추론을 진행하였음.

A1) 관측값들은 서로 독립이고 동일한 분산을 갖는 정규분포를 따르는 확률변수들이다.
A2) 관측값들의 기댓값은 구간별 상수함수 형태의 구조를 가지고 있다.
A3) 각 변화점 j에서의 변화폭 $|\mu_{j-1} - \mu_j|$이 잡음의 세기(σ)에 비해 충분히 크다.

Statistical Theory

FLSA의 특징

■ FLSA의 정의 및 특징

$$\hat{\mu}^{FL}(\lambda_1, \lambda_2) = \arg\min_{\mu_i} \left\{\frac{1}{2}\Sigma(y_i - \mu_i)^2 + \lambda_1\|\mu\|_1 + \lambda_2\|\mu\|_{TV}\right\}$$

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

λ_1은 작은 값들을 0으로 만드는 역할을 하기 때문에, 다중 변화점 식별을 위해서는 구간별 상수 구조를 구현하는 총변동 벌점항($\|\mu\|_{TV}$) 만을 고려해도 충분함. (Friedman et al. 2007)

장점: 다중변화점을 효율적으로 탐색 (Hoefling 2010)
단점: 관측값 개수 만큼의 변화점 집합만 고려하게 되어 점근적 일치성 보장 불가

따라서, 변화점 식별에 이용하기 위해서는 단점을 보완하기 위한 추가적인 검정방법이 필요.

변화점의 사후추론

■ 전통적인 검정방식의 문제점

전통적인 사후추론 방식: 식별된 변화점이 참인지 거짓인지 판별하는 가설검정.
하지만, 특정한 조건을 만족시키는 변화점을 기준으로 나뉘어진 각 구간에 대한 관측값은 더 이상 초기에 설정된 모형의 가정에 따르지 못하게 되어 가설검정 결과가 왜곡될 수 있음.

■ 전통적 사후추론 방식의 문제점을 확인하기 위하여 평균이 0이고 표준편차가 0.1인 정규분포를 따르는 21개의 난수를 생성하고 CUSUM 통계량에 따라 변화점을 식별하였음. 변화점으로 나뉘어진 두 구간의 표본평균이 같다고 볼 수 있는지 사후추론으로 z-검정을 이용.

* (그래프는 이 과정을 1000번 반복하여 계산한 z값과 유의확률임.)

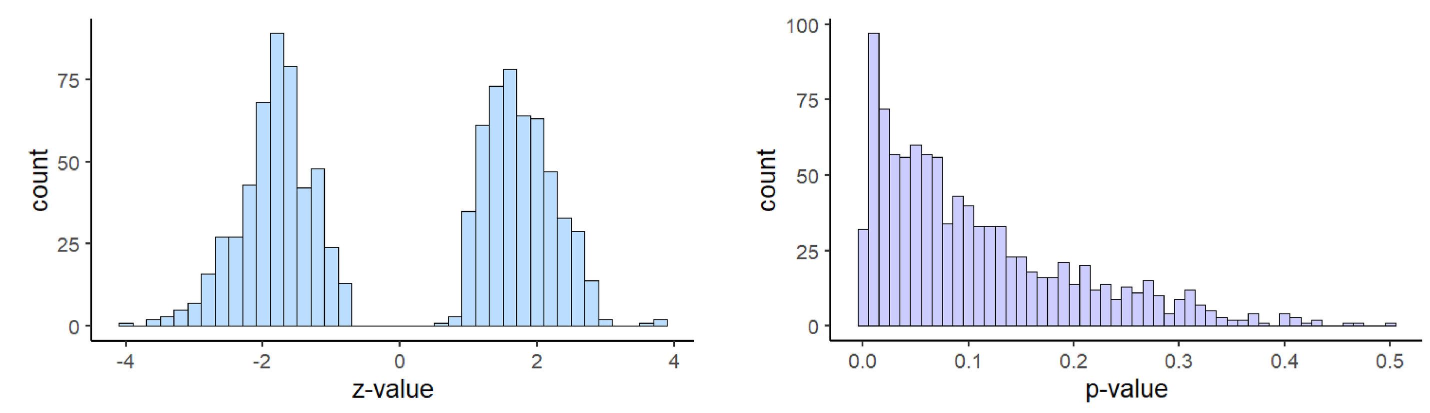

그림 1. '표본평균의 차이가 없다'는 귀무가설에 대한 z검정 결과

동일한 분포에서 난수를 발생시켰음에도 불구하고, z 검정 결과 표본평균의 차이가 있다고 나타난 경우가 1000번 중 349번 발생. 따라서, 각 구간은 처음 모집단에서 기대할 수 있었던 정규분포를 더 이상 만족하지 못한다는 것을 확인 할 수 있음.

■ 전통적인 검정방식의 대안

CUSUM 통계량을 이용하여 변화점을 탐색할 때는 통계량의 표본분포를 구하는 것이 중요하다고 할 수 있으며, 본 연구에서는 Antoch와 Huskova(2001)의 순열검정 절차를 응용하여 FLSA의 변화점의 사후 검정 절차를 제안하려 함.

Method

■ 거짓 변화점 식별을 위한 가설

$$H_0: \hat{j} \notin J \quad vs \quad H_1: \hat{j} \in J$$

($\hat{j}$는 FLSA에 의해 추정된 변화점 집합 $\hat{J}$의 원소이며, J는 참 변화점들의 집합)

귀무가설 하에서는 FLSA에 의해 추정된 변화점 $\hat{j}$가 거짓 변화점 임을 가정

■ CUSUM 통계량 기반의 순열검정 절차 제안 *

1. FLSA의 변화점 $\hat{j}$에 의해 나누어진 두 구간 $\hat{B}_{\hat{j}}, \hat{B}_{\hat{j}+1}$에서의 통계량 $u_0 = \dfrac{\bar{y}_{\hat{B}_{\hat{j}}} - \bar{y}_{\hat{B}_{\hat{j}+1}}}{\sqrt{\frac{1}{|\hat{B}_{\hat{j}}|} + \frac{1}{|\hat{B}_{\hat{j}+1}|}}}$ 을 구한다.
2. 구간 $\hat{B}_{\hat{j}} \cup \hat{B}_{\hat{j}+1}$에 포함되어 있는 관측값들을 랜덤으로 재배열한다.
3. 2에 의해 재배열된 관측값들에 대해 FLSA를 적용하여 새로운 하나의 변화점 $\hat{j}'$를 찾는다.
4. 3에서 식별된 변화점 $\hat{j}'$에 대하여 $u = \dfrac{\bar{y}_{\hat{B}_{\hat{j}'}} - \bar{y}_{\hat{B}_{\hat{j}'+1}}}{\sqrt{\frac{1}{|\hat{B}_{\hat{j}'}|} + \frac{1}{|\hat{B}_{\hat{j}'+1}|}}}$ 값을 계산한다.
5. 2~4의 과정을 K번 반복하여 K 개의 통계량 $u_1, u_2, \dots, u_K$의 표본분포를 구한다.
6. 1에서 구한 u_0값이 5에서 구한 표본분포에서 차지하는 위치를 이용하여 가설을 검정한다.

* (FLSA에 의해 식별된 변화점의 집합에 모든 참 변화점이 포함될 때 성립하기 때문에, FLSA 조절 모수를 적절하게 선정할 필요가 있다.)

Numerical Study

■ 모형 가정

$\mu_1, \dots, \mu_{20} = 1, \mu_{21}, \dots, \mu_{40} = 0, \mu_{41}, \dots, \mu_{70} = 1, \mu_{71}, \dots, \mu_{100} = 2$
$\sigma = 0.1\,, 0.3$
$K = 10{,}000$
순열검정과 가설검정 반복횟수 $B = 100$

■ 순열검정 vs z-검정

noise level	predicted signals	True Signals (FLSA): Permutation test: change point (P)	Permutation test: non-change (N)	Z-test: P	Z-test: N
$\sigma = 0.1$	change point(P)	300	3	300	12
	non-change point(N)	0	308	0	307
$\sigma = 0.3$	change point(P)	242	39	257	54
	non-change point(N)	54	374	43	359

표1. 예측된 변화점에 대한 혼동 행렬(confusion matrix)

먼저 잡음 수준이 작은 경우($\sigma = 0.1$) 순열검정 결과가 z-검정 결과에 비하여 모든 부분에서 우월함. z-검정 후 남아있는 거짓 변화점 중 그림 2의 구조와 같이 참 변화점으로부터 **멀리 떨어져 있는 경우의 비중이 높았으며,** 순열 검정은 참 변화점으로부터 가까운 곳에 위치한 경우가 많았음.
=> 구간이 긴 경우 순열검정이 거짓변화점을 탐색하는데 유용한 방법이라고 할 수 있음.

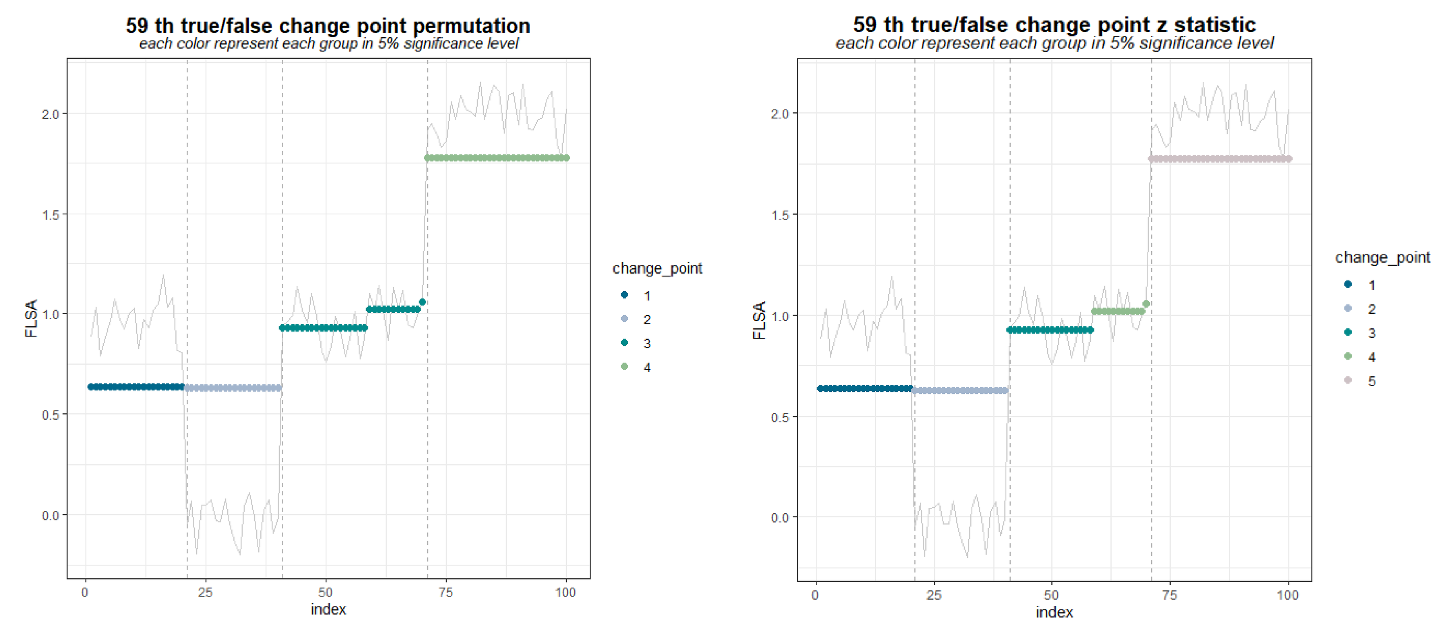

그림 2. 구간이 긴 관측치에 대한 순열검증(왼)과 z-검증(오) 결과 비교

(세 개의 점선은 참 변화점을 의미하며, 각 색상은 각 구간을 의미함. 관측치는 세 개의 참 변화점과 네 개의 구간을 가지고 있지만, z- 검정은 네 개를 참 변화점으로 식별하고 있음.)

하지만, 잡음 수준이 변화점에서의 평균 수준차이에 비해 무시할 수 없을 정도로 큰 경우 ($\sigma = 0.3$) 거짓 변화점 사이의 구간 폭이 짧아지면서 순열검정의 검정 성능이 저하됨. 거짓 변화점을 참 변화점으로 식별하거나, 참 변화점의 일부를 식별하지 못하는 상충현상 발생.
=> 변화점 식별을 위한 가정 중 A3)가 만족되는 관측치에 대해서 검증가능한 방법

Discussion

■ 결론

FLSA의 총변동벌점을 이용하여 다중변화점 탐색이 가능하지만, 점근적 일치성이 만족되지 않아 거짓 변화점이 남아 있을 수 있다는 단점 존재.

=> 본 연구는 단점을 보완하기 위하여 사후검정 방법으로 순열 검정 방법을 제안하였음. Antoch과 Huskova(2001)에 의해 제안된 단일변화점 모형과 관련된 순열검정 방법을 FLSA와 결합하여 다중 변화점 모형에 적용 가능하도록 확장하였음.

■ 토의

본 검정방법의 정확한 작동을 위해서는 FLSA로 식별된 변화점의 집합이 참 변화점을 포함할 수 있도록 조절모수 선택기법이 필요함. 또한, 각 구간의 길이가 짧을 때 순열검정으로 정확한 표본분포를 구하는 데 한계가 존재하여 검정결과에 오차 발생 가능.

=> 향후 이러한 문제를 해결하기 위한 추가적인 연구 필요