The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.7.1429
29(7), 1429–1444

Monitoring mean change via penalized estimation

Okyoung Na[a] · Sunghoon Kwon[b,1]

[a]Department of Applied Information Statistics, Kyonggi University;
[b]Department of Applied Statistics, Konkuk University

(Received November 9, 2016; Revised December 7, 2016; Accepted December 7, 2016)

Abstract

We suggest a monitoring procedure to detect changes in the mean of the stochastic process. The monitoring procedure is based on penalized least squares estimates. Unlike the fluctuation (FL) monitoring, we use the numbers of nonzero estimates not the fluctuations of sequential parameter estimates. We investigate the behavior of the proposed monitoring procedure by means of a simulation study and compare its performance with CUSUM monitoring.

Keywords: CUSUM monitoring, FL monitoring, LASSO, penalized estimation, SCAD

1. 서론

자기회귀이동평균(ARMA) 모형이나 ARCH, GARCH 등의 대표적인 시계열모형들은 분석에 사용되는 시계열 자료가 정상성을 만족한다고 가정한다. 즉 관측기간 동안 시계열의 평균, 분산, 자기상관계수 등이 일정하고, 시간에 따라 모형의 모수 또한 변하지 않는다고 가정한다. 그러나 관측기간이 긴 경우 시계열의 특징을 나타내는 모수가 기간 내에 변하는 경우가 종종 있으며, 이 경우 전 기간 동안 관측한 시계열 자료를 모두 이용하여 모형을 적합하고 미래의 값을 예측하면 예측값이 잘 맞지 않는다. 그러므로 주어진 시계열 자료를 이용하여 모수의 안정성을 검토하고, 모수의 값이 변한 시점을 추정하는 것은 시계열 분석에서 매우 중요한 문제다. 보통 이 문제를 변화점 탐지 문제(change point problem)라고 부르며, 이와 관련된 연구에는 Brown 등 (1975), Ploberger와 Krämer (1986), Inclán과 Tiao (1994), Csörgő와 Horváth (1997), Lee 등 (2003) 등이 있다.

그러나 Chu 등 (1996)의 연구에 의하면 이들 연구 결과는 후향적 검정을 기반으로 하기 때문에 관심의 대상이 되는 전 기간 동안의 모든 자료가 주어지지 않고 자료가 하나씩 순차적으로 관측되어 추가되는 상황에서 모수의 변화 여부를 반복적으로 검토할 때에는 사용하기 어렵다. 따라서 Chu 등 (1996)은 후향적 검정이 아닌 순차적 검정 기법을 연구하였고, 선형회귀모형에서의 CUSUM 모니터링과 FL 모니터링을 제안하였다. CUSUM 모니터링은 자료가 하나씩 순차적으로 관측될 때마다 반복적으로 회귀 모형을 적합하고 얻은 잔차들의 누적합이 미리 정해진 경계함수를 벗어나는 순간 모니터링을 멈추고 회귀계수의 값이 변했다고 판단하는 방법이다. 그리고 FL 모니터링은 반복적으로 얻은 회귀계수의 추정값

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A1002995).
[1]Corresponding author: Department of Applied Statistics, Konkuk University, 120, Neungdong-ro, Gwangjin-gu, Seoul 05029. Korea. E-mail: shkwon0522@gmail.com

들의 변화량을 바탕으로 모니터링을 지속할지 여부를 판단하는 방법이다. Chu 등 (1996)이 CUSUM과 FL 모니터링 절차를 개발한 이후로 선형회귀모형 뿐만 아니라 여러 시계열모형에서의 모수의 변화를 검토하기 위한 다양한 모니터링 절차들이 연구 개발되었다. 이와 관련된 자세한 내용은 Leisch 등 (2000), Horváth 등 (2004), Berkes 등 (2004), Zeileis 등 (2005), Horváth 등 (2008), Na 등 (2011) 등을 참조하기 바란다.

본 연구에서는 least absolute shrinkage and selection operator(LASSO)나 smoothly clipped absolute deviation(SCAD) 등의 벌점함수를 이용한 벌점화 추정방법을 기반으로 평균에 대한 모니터링 절차를 개발하고자 한다. 평균에 대한 모니터링을 시행할 확률과정을 $\{y_i, i = 0, \pm 1, \ldots\}$이라고 하고, 편의상 모니터링을 시작한 시점을 $i = 1$이라고 표기한다. 즉 양의 시점과 0 이하의 시점은 각각 모니터링을 시작한 이후와 이전을 뜻한다. 그리고 Chu 등 (1996), Horváth 등 (2008), Na 등 (2011)처럼 확률과정 $\{y_i, i = 0, \pm 1, \ldots\}$는 모형식

$$y_i = \begin{cases} \mu + \epsilon_i, & i \in \{0, -1, -2, \ldots\}, \\ \mu + \theta_i + \epsilon_i, & i \in \{1, 2, 3, \ldots\} \end{cases} \tag{1.1}$$

을 만족한다고 가정한다. 여기서, $\mu(\in \mathbb{R})$는 모니터링 이전의 공통 평균을 의미하는 모수이고, $\theta_i(\in \mathbb{R})$은 모니터링을 시작한 이후 i 시점에서의 평균과 모니터링 이전의 공통 평균과의 차이를 나타내는 모수이다. 그리고 $\{\epsilon_i, i = 0, \pm 1, \ldots\}$는 평균이 0인 백색 잡음 확률과정으로 오차항을 나타낸다. 다시 말해 분산과 같은 평균 이외의 모수는 변하지 않으며, 평균만 모니터링을 시작한 이후에 변할 수 있다고 가정한다.

모니터링을 시작한 이후 평균이 μ에서 다른 값으로 변한 시점을 τ라고 표기하면,

$$\tau = \min\{i \in \mathbb{N} : \theta_i \neq 0\}$$

가 성립한다. 그러므로 본 논문에서는 벌점화 최소제곱법을 이용하여 0이 아닌 θ_i의 값을 식별해내고, 이 값을 이용하여 만든 모니터링 절차를 제안하고자 한다.

논문의 구성은 다음과 같다. 2절에서 모니터링의 대표적인 방법인 CUSUM 모니터링과 FL 모니터링을 평균 변화에 대한 문제에 적용하고, 정리하였다. 3절에서 θ_i에 대한 벌점화 최소제곱추정량을 정의하고, 이 추정량을 이용하여 만든 모니터링 절차를 소개하였고, 4절에 모의실험을 실시하여 얻은 결과를 보고하였다. 마지막으로 5절에 연구 결과에 대한 결론이 있다.

2. CUSUM 모니터링

CUSUM 모니터링은 Chu 등 (1996)이 처음 제안한 방법으로 모니터링 절차 중 가장 대표적인 방법이며, Na 등 (2011)에 의해 시계열에서 일반적인 모수에 대한 모니터링 문제로 확장 연구되었다. 2절에서는 평균 변화에 대한 CUSUM 모니터링을 소개하고, 그 성질을 간단히 살펴보고자 한다.

2.1. CUSUM 모니터링 절차

평균에 대한 모니터링을 시행하기 이전에 관측된 n개의 과거자료 $y_0, y_{-1}, \ldots, y_{1-n}$가 주어져 있고, 모니터링을 시작한 이후 $y_1, y_2, \ldots$이 순차적으로 하나씩 관측될 때, CUSUM 모니터링 절차는 다음과 같다.

(S1) 과거자료 $y_0, y_{-1}, \ldots, y_{1-n}$들의 표본평균 $\bar{y} = \sum_{i=1}^{n} y_{1-i}/n$과 표본분산 $s^2 = \sum_{i=1}^{n}(y_{1-i} - \bar{y})^2/(n-1)$을 계산한다.

(S2) 모니터링을 시작한 이후 첫 번째로 얻은 관측값 y_1과 과거자료의 표본평균 $\bar{y}$와의 차이 $e_1 = y_1 - \bar{y}$를 구한다. 만약

$$|e_1| \geq s\sqrt{n}\left(1+\frac{1}{n}\right) b\left(1+\frac{1}{n}\right)$$

이면 모니터링을 멈추고, 평균이 모니터링을 시작한 이후 첫 번째 시점에서 변했다고 판단한다. 만약 그렇지 않으면 두 번째 시점에서 y_2을 관측하고 모니터링을 지속한다. 여기서 $b : \mathbb{R} \to \mathbb{R}$은 모니터링을 시작하기 전에 주어진 경계함수이다.

(S3) $(k-1)$ 시점까지 모니터링이 지속되어 멈추지 않았을 때, k 시점에서 y_k를 새롭게 관측하고 $e_k = y_k - \bar{y}$를 계산한다. 만약 이전에 계산한 차이 $e_i = y_i - \bar{y},\ i = 1, \ldots, k-1$들과 e_k의 합이

$$\left|\sum_{i=1}^{k} e_i\right| \geq s\sqrt{n}\left(1+\frac{k}{n}\right) b\left(1+\frac{k}{n}\right) \tag{2.1}$$

을 만족하면 모니터링을 멈추고, k 시점에서 평균이 변했다고 판단한다. 만약 그렇지 않으면 새로운 값 y_{k+1}을 관측하고 모니터링을 지속한다.

(S4) 모니터링을 멈출 때까지 단계 (S3)를 반복한다.

2.2. FL 모니터링과의 관계

표본평균의 정의에 의해 $\sum_{i=1}^{n}(y_{1-i} - \bar{y}) = 0$이므로, 식 (2.1)에서 차이들의 누적합은

$$\sum_{i=1}^{k} e_i = \sum_{i=1}^{k}(y_i - \bar{y}) = \sum_{i=1-n}^{k}(y_i - \bar{y}) = (n+k)(\bar{y}_k - \bar{y})$$

를 만족한다. 여기서 $\bar{y}_k = \sum_{i=1-n}^{k} y_i/(n+k)$이다. 그러므로 식 (2.1)은

$$|\bar{y}_k - \bar{y}| \geq sb\frac{1+k/n}{\sqrt{n}}$$

으로 다시 쓸 수 있고, 이는 Chu 등 (1996)과 Na 등 (2011)이 고려한 FL 모니터링과 같다. 일반적으로 순차적으로 반복해서 구한 모수의 추정값과 과거 자료만을 이용하여 구한 모수의 추정값의 차이를 이용하여 만든 FL 모니터링과 모형 적합 후 얻은 잔차들의 누적합을 이용하여 만든 CUSUM 모니터링은 같지 않다. 그러나 평균에 대한 모니터링에서는 2.1절에서 소개한 CUSUM 모니터링과 순차적으로 반복해서 구한 표본평균들의 변동을 이용해서 만든 FL 모니터링이 동일하다고 할 수 있다.

2.3. 변화시점 추정과 경계함수 결정

2.1절에 기술한 모니터링 절차에 의하면 정지시간

$$\hat{\tau}_b^C = \min\left\{k \in \mathbb{N} : \left|\sum_{i=1}^{k}(y_i - \bar{y})\right| \geq s\sqrt{n}\left(1+\frac{k}{n}\right) b\left(1+\frac{k}{n}\right)\right\} \tag{2.2}$$

를 정의할 수 있으며, 정지시간 $\hat{\tau}_b^C$는 평균 변화시점 τ에 대한 추정량으로 사용될 수 있다.

정지시간의 이론적 성질을 살펴보기 위해 모형식 (1.1)의 오차항 $\epsilon_i,\ i = 0, \pm 1, \ldots$들이 서로 독립이고, 동일한 분포를 갖는(iid) 확률변수들이며, $E(\epsilon_i) = 0$와 $E(|\epsilon_i|^\nu) < \infty$, $\nu > 2$를 만족한다고 가정하자.

만약 경계함수 $b(\cdot)$가 $\inf_{1<x<\infty} b(x) > 0$을 만족하는 연속함수이면, 오차항들의 누적합에 대한 불변성 원리에 의해

$$\lim_{n\to\infty} \mathbf{P}\left(\hat{\tau}_b^C < \infty | \theta_i = 0, \forall i \in \mathbb{N}\right) = \mathbf{P}\left(\sup_{0<x<1}\left\{\frac{|W(x)|}{b(1/(1-x))}\right\} \geq 1\right) \tag{2.3}$$

이 성립함을 보일 수 있다. 여기서 $W(\cdot)$은 표준 브라운 운동이고, 이와 관련된 자세한 내용은 Na 등 (2011)의 정리 1과 2를 참조하기 바란다.

식 (2.3)의 좌변에서 $\theta_i = 0$, $\forall i \in \mathbb{N}$은 평균이 모니터링 이후에도 변하지 않는다는 뜻이고, $\hat{\tau}_b^C < \infty$는 모니터링을 멈추었다, 즉 평균이 변화했다고 판단했다는 뜻이다. 그러므로 식 (2.3)은 모니터링 기간 동안 평균이 변하지 않았다는 귀무가설과 평균이 변했다는 대립가설에 대한 가설검정에서 제1종의 오류를 범할 확률에 대한 극한값을 나타낸다. 따라서 CUSUM 모니터링에 사용할 경계함수를 결정할 때, 유의수준 $\alpha(\in (0,1))$와 식 (2.3)을 활용할 수 있다. 예를 들어, 경계함수가 상수함수, $b(x) = c$라면 식 (2.3)은

$$\lim_{n\to\infty} \mathbf{P}\left(\hat{\tau}_b^C < \infty | \theta_i = 0, \forall i \in \mathbb{N}\right) = \mathbf{P}\left(\sup_{0<x<1} |W(x)| \geq c\right)$$

이 된다. 그러므로 $\sup_{0<x<1}|W(x)|$ 분포의 상위 α 분위수를 c_α라고 표시할 때, $b(x) = c_\alpha$가 유의수준 α에서의 CUSUM 모니터링 경계함수가 되고, 이 경우 변화시점 τ에 대한 추정량은 정지시간

$$\hat{\tau}_\alpha^C = \min\left\{k \in \mathbb{N} : \left|\sum_{i=1}^{k}(y_i - \bar{y})\right| \geq s\sqrt{n}\left(1 + \frac{k}{n}\right)c_\alpha\right\} \tag{2.4}$$

이다. 참고로 $\alpha = 0.01, 0.05, 0.10$에 대응되는 c_α는 각각 2.791, 2.214, 1.933이며, 좀 더 자세한 내용은 Na 등 (2011)을 참조하기 바란다.

3. 벌점화 최소제곱추정량을 이용한 모니터링

이제 평균 차이 θ_i, $i \in \mathbb{N}$에 대한 벌점화 최소제곱추정량을 정의하고, 순차적으로 구한 추정값을 이용하여 평균에 대한 모니터링을 시행하는 절차를 제안하고자 한다.

3.1. 평균 차에 대한 추정

모니터링 이전에 $n(\in \mathbb{N})$개의 과거 자료 $y_0, y_{-1}, \ldots, y_{1-n}$가 주어져 있고, 모니터링을 시작한 이후로 $k(\in \mathbb{N})$ 시간이 흘러 현재 $y_1, \ldots, y_k$가 추가로 관측되었다고 가정하자. 주어진 $(n+k)$개의 자료를 모두 이용하여 평균 차이를 나타내는 모수 벡터 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$를 벌점화 최소제곱추정방법을 이용하여 추정하고자 한다. 일반적으로 μ의 값이 알려져 있지 않으므로 $\boldsymbol{\theta}$와 더불어 μ도 동시에 추정한다.

$\boldsymbol{\theta}$와 μ에 대한 벌점화 최소제곱추정량은

$$\left(\hat{\boldsymbol{\theta}}_{\lambda,n}, \hat{\mu}_{\lambda,n}\right) = \arg\min_{\boldsymbol{\vartheta}\in\mathbb{R}^k, u\in\mathbb{R}} \{Q_n(\boldsymbol{\vartheta}, u) + 2P_\lambda(\boldsymbol{\vartheta})\} \tag{3.1}$$

과 같이 정의된다. 여기서

$$Q_n(\boldsymbol{\vartheta}, u) = \sum_{i=1-n}^{0}(y_i - u)^2 + \sum_{i=1}^{k}(y_i - u - \vartheta_i)^2 \tag{3.2}$$

는 제곱합 손실함수를 나타내며, ϑ_i는 $\boldsymbol{\vartheta}$의 i번째 원소를 뜻한다. 그리고 λ는 조절모수로 0 이상의 값을 가지며, P_λ는 조절모수 λ에 대응되는 벌점함수이다. 벌점함수로 다양한 함수를 사용할 수 있으나, 본 논문에서는 LASSO의 L_1 벌점함수

$$P_\lambda(\boldsymbol{\vartheta}) = \lambda \sum_{i=1}^{k} |\vartheta_i|$$

와 SCAD 벌점함수

$$P_\lambda(\boldsymbol{\vartheta}) = \sum_{i=1}^{k} \min\left\{ \lambda|\vartheta_i| - \max \frac{(0, |\vartheta_i| - \lambda)^2}{2a-2},\ (a+1)\frac{\lambda^2}{2} \right\}, \quad a > 2$$

만을 고려하였다.

모니터링 이전의 과거 자료 $y_0, y_{-1}, \ldots, y_{1-n}$의 표본평균 $\bar{y}$과 표본분산 s^2을 이용하면, 제곱합 손실함수 $Q_n(\boldsymbol{\vartheta}, u)$는 $\boldsymbol{\vartheta}$와 u의 함수

$$Q_{n,1}(\boldsymbol{\vartheta}, u) = (n-1)s^2 + (n+k)\left\{ \bar{y} - \mu + \frac{1}{n+k}\sum_{i=1}^{k}(y_i - \bar{y} - \vartheta_i) \right\}^2$$

와 $\boldsymbol{\vartheta}$만의 함수

$$Q_{n,2}(\boldsymbol{\vartheta}) = \sum_{i=1}^{k}(y_i - \bar{y} - \vartheta_i)^2 - \frac{1}{n+k}\left\{ \sum_{i=1}^{k}(y_i - \bar{y} - \vartheta_i) \right\}^2$$

의 합으로 나누어진다. 그러므로

$$\hat{\boldsymbol{\theta}}_{\lambda,n} = \left(\hat{\theta}_{\lambda,n,1}, \ldots, \hat{\theta}_{\lambda,n,k}\right)^T = \arg\min_{\boldsymbol{\vartheta}\in\mathbb{R}^k} \{Q_{n,2}(\boldsymbol{\vartheta}) + 2P_\lambda(\boldsymbol{\vartheta})\} \tag{3.3}$$

이고, $\hat{\mu}_{\lambda,n} = \bar{y} + \sum_{i=1}^{k}(y_i - \bar{y} - \hat{\theta}_{\lambda,n,i})/(n+k)$이다. 참고로 식 $Q_{n,2}(\boldsymbol{\vartheta})$을 살펴보면, $Q_{n,2}(\boldsymbol{\vartheta})$는 $y_i - \bar{y}$, $i = 1, \ldots, k$의 값들과 모니터링 이전의 과거 자료의 개수 n만 알면 구할 수 있다. 그러므로 모니터링 이전의 과거 자료들 $y_0, y_{-1}, \ldots, y_{1-n}$ 대신 과거 자료의 표본평균 $\bar{y}$과 표본의 크기 n만 알려져 있어도 추정값 $\hat{\boldsymbol{\theta}}_{\lambda,n}$을 계산할 수 있다.

만약 모니터링 이전에 관측된 자료가 존재하지 않고 μ의 값이나 μ에 대한 사전 추정값이 μ_0라고 알려진 경우라면, 식 (3.3)에서 $Q_{n,2}(\boldsymbol{\vartheta})$ 대신 $Q_0(\boldsymbol{\vartheta}, \mu_0) = \sum_{i=1}^{k}(y_i - \mu_0 - \vartheta_i)^2$을 사용하여 $\boldsymbol{\theta}$만을 추정할 수 있으며, 이 추정량을 $\hat{\boldsymbol{\theta}}_{\lambda,0}$라고 표기한다. 즉,

$$\hat{\boldsymbol{\theta}}_{\lambda,0} = \left(\hat{\theta}_{\lambda,0,1}, \ldots, \hat{\theta}_{\lambda,0,k}\right)^T = \arg\min_{\boldsymbol{\vartheta}\in\mathbb{R}^k} \{Q_0(\boldsymbol{\vartheta}, \mu_0) + 2P_\lambda(\boldsymbol{\vartheta})\} \tag{3.4}$$

이다. 만약 L_1 벌점함수를 사용하면 추정량 $\hat{\boldsymbol{\theta}}_{\lambda,0}$은

$$\hat{\theta}_{\lambda,0,i} = \mathrm{sgn}(y_i - \mu_0)\max(|y_i - \mu_0| - \lambda, 0)$$

을 만족하고, SCAD 벌점함수를 사용하면

$$\hat{\theta}_{\lambda,0,i} = \begin{cases} \mathrm{sgn}(y_i - \mu_0)\max(|y_i - \mu_0| - \lambda, 0), & |y_i - \mu_0| \le 2\lambda, \\ \dfrac{\mathrm{sgn}(y_i - \mu_0)\{(a-1)|y_i - \mu_0| - a\lambda\}}{a-2}, & 2\lambda < |y_i - \mu_0| \le a\lambda, \\ y_i - \mu_0, & |y_i - \mu_0| > a\lambda \end{cases}$$

을 만족한다. 여기서 $\mathrm{sgn}(x)$은 실수 x의 부호를 나타내는 부호함수이다. 자세한 사항은 Fan과 Li (2001)을 참조하기 바란다.

3.2. 일반화 정보함수

이제 조절모수 λ를 선택할 때 사용할 일반화 정보함수(generalized information criterion; GIC)에 대해 살펴보자. 본 논문에서는 모형 적합도에 대한 측도로 잔차제곱합을 이용한 GIC를 고려하였으며, GIC의 정의는 다음과 같다:

$$\mathrm{GIC}(\lambda) = \mathrm{RSS}_\lambda + \gamma\sigma^2\|\hat{\boldsymbol{\theta}}_{\lambda,n}\|_0. \tag{3.5}$$

여기서 첫 번째 항은 $\mathrm{RSS}_\lambda = Q_n(\hat{\boldsymbol{\theta}}_{\lambda,n}, \hat{\mu}_{\lambda,n}) = \sum_{i=1-n}^{0}(y_i - \hat{\mu}_{\lambda,n})^2 + \sum_{i=1}^{k}(y_i - \hat{\mu}_{\lambda,n} - \hat{\theta}_{\lambda,n,i})^2$이며, 조절모수 λ에 대응되는 잔차제곱합을 나타낸다. n은 모니터링 이전에 관측하여 주어진 과거 자료의 개수를 말하는 것으로 0 이상의 정수 값을 취하며, $n = 0$는 과거 자료가 주어지지 않은 경우를 말한다. $n > 0$인 경우 추정량 $\hat{\boldsymbol{\theta}}_{\lambda,n}$과 $\hat{\mu}_{\lambda,n}$은 식 (3.1)에서 정의된 값이고, $n = 0$일 때의 추정량 $\hat{\boldsymbol{\theta}}_{\lambda,0}$는 식 (3.4)에서 정의된 값이다. 그리고 이 경우 $\hat{\mu}_{\lambda,0}$는 사전에 주어진 값 μ_0를 말한다.

두 번째 항은 모형의 복잡도를 측정하는 부분으로 $\|\hat{\boldsymbol{\theta}}_{\lambda,n}\|_0$는 추정량 $\hat{\boldsymbol{\theta}}_{\lambda,n}$의 원소 중 0이 아닌 것의 개수를 의미하고, γ는 모형의 복잡도를 조절해주는 상수이며, $\sigma^2 = E(\epsilon_i^2)$이다. 일반적으로 σ^2는 미지의 값이므로 주어진 사전추정량이나 과거 자료의 표본분산 등으로 대체하여 GIC의 값을 구한다. γ의 값으로 Akaike information criterion(AIC)처럼 2를 사용할 수도 있고, Bayesian information criterion(BIC)처럼 $\log(n+k)$의 값을 사용할 수도 있다. 이 외에도 $2\log\log(n+k)$ 또는 3이나 4와 같은 상수도 γ의 값으로 사용할 수 있다.

3.3. 변화시점 추정

모니터링을 시작한 이후 $k(\in \mathbb{N})$번째 관측시점에서 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$에 대한 벌점화 최소제곱추정량 중 식 (3.5)에서 정의한 일반화 정보함수 GIC를 최소화하는 조절모수에 대응되는 추정량을 $\hat{\boldsymbol{\theta}}_n^k = (\hat{\theta}_{n,1}^k, \ldots, \hat{\theta}_{n,k}^k)^T$라고 표기하자. 그러면 모니터링을 시작한 이후 순차적으로 구한 추정량들 $\{\hat{\boldsymbol{\theta}}_n^k, k = 1, 2, \ldots\}$을 이용하여 모니터링 이후 평균이 변한 시점 τ를 추정할 수 있다. 본 논문에서는 $\|\hat{\boldsymbol{\theta}}_n^k\|_0$, 즉 $\hat{\theta}_{n,1}^k, \ldots, \hat{\theta}_{n,k}^k$ 중에서 0이 아닌 추정량의 개수를 이용하여 정지시간

$$\hat{\tau}_\kappa = \min\left\{k \in \mathbb{N} : \left\|\hat{\boldsymbol{\theta}}_n^k\right\|_0 \geq \kappa\right\} \tag{3.6}$$

을 정의하고, 정지시간 $\hat{\tau}_\kappa$를 평균 변화시점 τ에 대한 추정량으로 사용하고자 한다. 여기서 $\kappa(\in \mathbb{N})$는 평균에 대한 모니터링을 멈추고 평균이 변화했다고 판단할 때 사용되는 기준이 되는 개수로 미리 주어지는 상수이다.

4. 모의실험

여기에서는 2, 3절에서 다룬 CUSUM 모니터링과 벌점화 최소제곱추정량을 이용한 모니터링에 대한 모의실험을 실시하고, 그 결과를 분석해보기로 한다.

모의실험은 자료생성과정

$$y_i = \mu + \delta I(i \geq m) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0,1), \quad i = 1-n, 2-n, \ldots$$

을 이용하여 실시하였고, μ, δ, m, n 등 자료생성과정에 포함된 값들에 대해서는 다음 조합들을 고려하였다.

- 공통 평균: $\mu = 0$,

Table 4.1. The number of cases where the estimated stopping times are less than or qual to $N = 100$ for $\delta = 0$ among 100 simulations

CUSUM	n	α	0.10	0.05	0.01							
	10		7	5	1							
	100		1	0	0							
	300		0	0	0							
	1000		0	0	0							
SCAD	n	κ	1	2	3	4	5	6	7	8	9	10
	0		58	41	26	15	12	12	9	5	1	1
	10		98	88	71	60	53	47	39	36	31	26
	100		99	81	54	35	20	14	8	5	4	1
	300		97	75	40	16	9	3	1	1	0	0
	1000		49	30	10	2	0	0	0	0	0	0
LASSO	n	κ	1	2	3	4	5	6	7	8	9	10
	0		55	33	21	14	10	7	5	2	1	1
	10		95	65	40	30	27	22	18	16	15	12
	100		90	21	4	3	1	0	0	0	0	0
	300		86	11	1	0	0	0	0	0	0	0
	1000		15	1	0	0	0	0	0	0	0	0

- 평균의 변화크기: $\delta = 0, 1, 2, 3, 5$,
- 평균의 변화시점: $m = 10, 30$,
- 과거 자료의 개수: $n = 0, 10, 100, 300, 1000$.

각각의 경우에 3절에서 소개한 벌점화 최소제곱추정량을 기반으로 한 모니터링과 2절의 CUSUM 모니터링을 실시하였다. 다만 CUSUM 모니터링은 $n = 0$일 때 정의되지 않으므로 $n > 0$인 경우만 실시하였고, 편의상 모든 경우에 모니터링 종료 시점을 $N = 100$이라고 하였다.

벌점화 최소제곱추정량을 구하기 위해서 벌점함수로 LASSO의 L_1 벌점함수와 SCAD 벌점함수를 고려하였다. $n > 0$인 경우, 일반화정보함수 GIC에서 조절 상수 γ의 값으로 $\log(n + k)$를 사용하였고, σ^2의 값으로 과거 자료들의 표본분산 s^2을 이용하였다. 그리고 $n = 0$인 경우, $\gamma = \log(e^2 + k)$, $\mu_0 = 0$, $\sigma = 1$을 사용하여 GIC의 값을 계산하였다. 순차적으로 구한 벌점화 최소제곱추정값을 바탕으로 식 (3.6)에 정의된 $\kappa = 1, 2, \ldots, 10$에 대응되는 정지시간들 $\hat{\tau}_\kappa$을 구하였고, $\hat{\tau}_\kappa = 101$인 것은 모니터링을 종료할 때까지 $\sup_{k\in\mathbb{N}} \|\hat{\boldsymbol{\theta}}_n^k\|_0 < \kappa$임을 의미한다.

CUSUM 모니터링에서는 경계함수로 상수함수만을 고려하였다. 유의수준 α의 값으로 0.01, 0.05, 0.10 세 가지를 고려하였고, 이에 대응되는 정지시간 $\hat{\tau}_\alpha^C$을 구하였다. 여기서도 모니터링을 종료할 때까지 CUSUM 모니터링을 멈추지 못한 경우 정지시간을 $\hat{\tau}_\alpha^C = 101$이라고 표시하였다.

모의실험은 총 100번씩 반복하였으며, 각각의 경우에 구한 정지시간들을 가지고 그린 상자그림들을 Figure 4.1부터 4.5까지 정리하였다. Figure 4.1은 과거자료가 주어지지 않은 경우로 각각의 그림에서 처음 10개의 상자그림, $S1, \ldots, S10$이라고 표시된 상자그림은 SCAD 벌점함수를 이용하여 구한 정지시간 $\hat{\tau}_\kappa$, $\kappa = 1, 2, \ldots, 10$들에 대한 것이고, 뒤에 있는 $L1, \ldots, L10$이라고 표시된 10개의 그림은 LASSO의 L_1 벌점함수를 이용하여 구한 정지시간들의 상자그림이다. Figure 4.1을 보면 $n = 0$인 경우, 벌점함수에 따른 정지시간들의 차이는 크지 않음을 알 수 있다. 그리고 평균의 변화크기 δ의 값이 2 이상인 경우엔 3 또는 4 이상의 κ에 대응되는 정지시간들이 변화를 잘 탐지할 뿐만 아니라 변화시

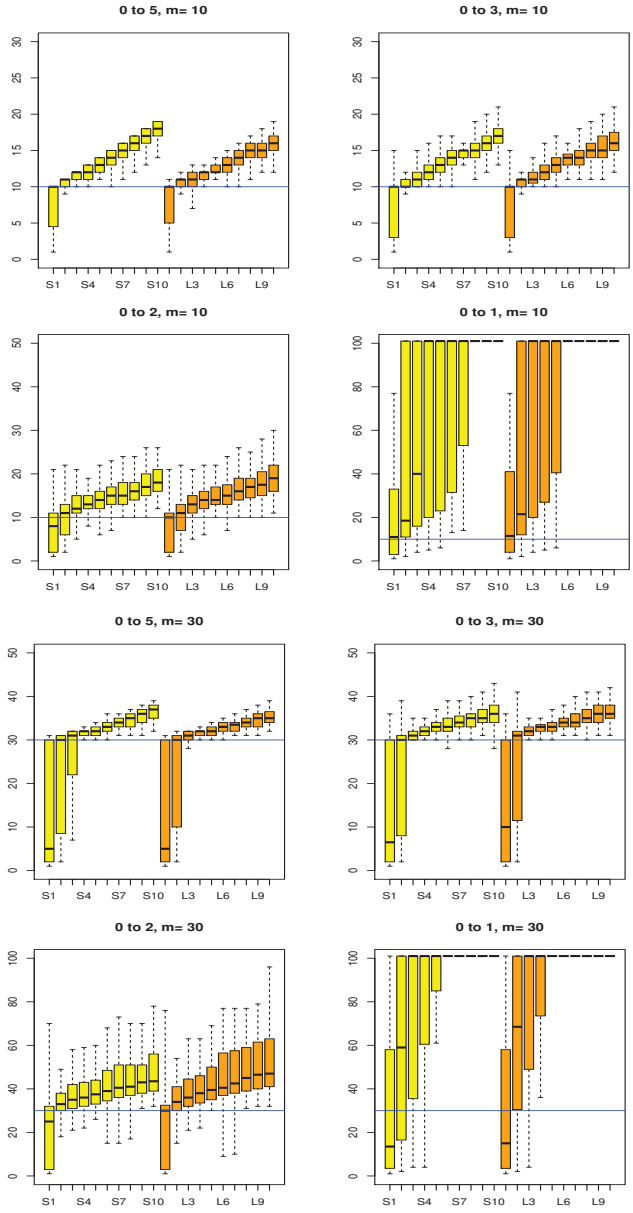

Figure 4.1. Boxplots of estimated stopping times obtained from 100 simulations with $n = 0$ when the mean changes from $\mu = 0$ to $\mu + \delta$ for $\delta \in \{5, 3, 2, 1\}$.

점과도 큰 차이가 나지 않는다. 그러나 $\delta = 1$로 크지 않은 경우엔 변화를 탐지하는 능력이 떨어지고, $\kappa \leq 2$인 경우는 변화가 없어도 평균에서 조금만 떨어진 값이 있으면 영향을 받는 것을 알 수 있다.

Figure 4.2부터 4.5까지는 과거 자료가 존재하는 경우에 해당하며, 각각의 그림에서 처음 3개 $C01$, $C05$, $C10$이라고 표시된 상자그림은 CUSUM 모니터링에서 구한 정지시간 $\hat{\tau}_\alpha^C$, $\alpha = 0.01, 0.05, 0.10$에

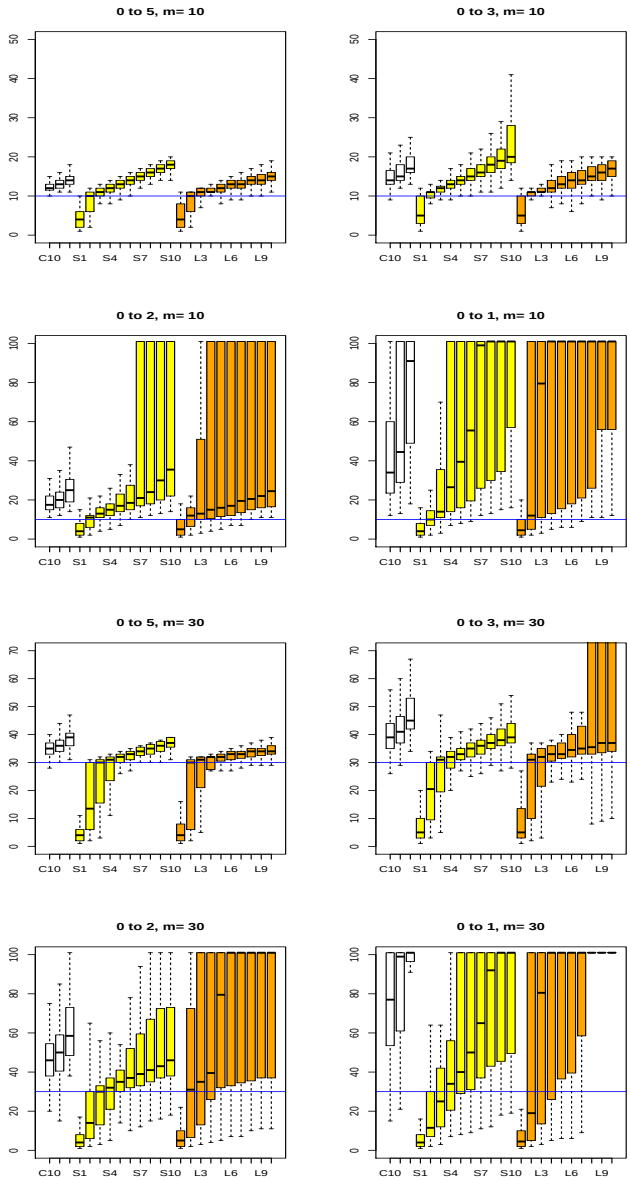

Figure 4.2. Boxplots of estimated stopping times obtained from 100 simulations with $n = 10$ when the mean changes from $\mu = 0$ to $\mu + \delta$ for $\delta \in \{5, 3, 2, 1\}$.

해당한다. 그리고 중간 10개의 상자그림은 SCAD 벌점함수를 사용한 경우, 마지막 10개의 상자그림은 LASSO의 L_1 벌점함수를 사용하여 구한 정지시간들에 대한 것이다. 이 그림들을 보면 LASSO의 경우보다는 SCAD의 경우가 전반적으로 평균의 변화를 잘 탐지하고, 과거 자료의 개수 n이 크고 변화의 크기가 3 이상인 경우 SCAD나 LASSO 등 벌점함수를 사용하여 구한 정지시간들이 효과적으로 변화시점

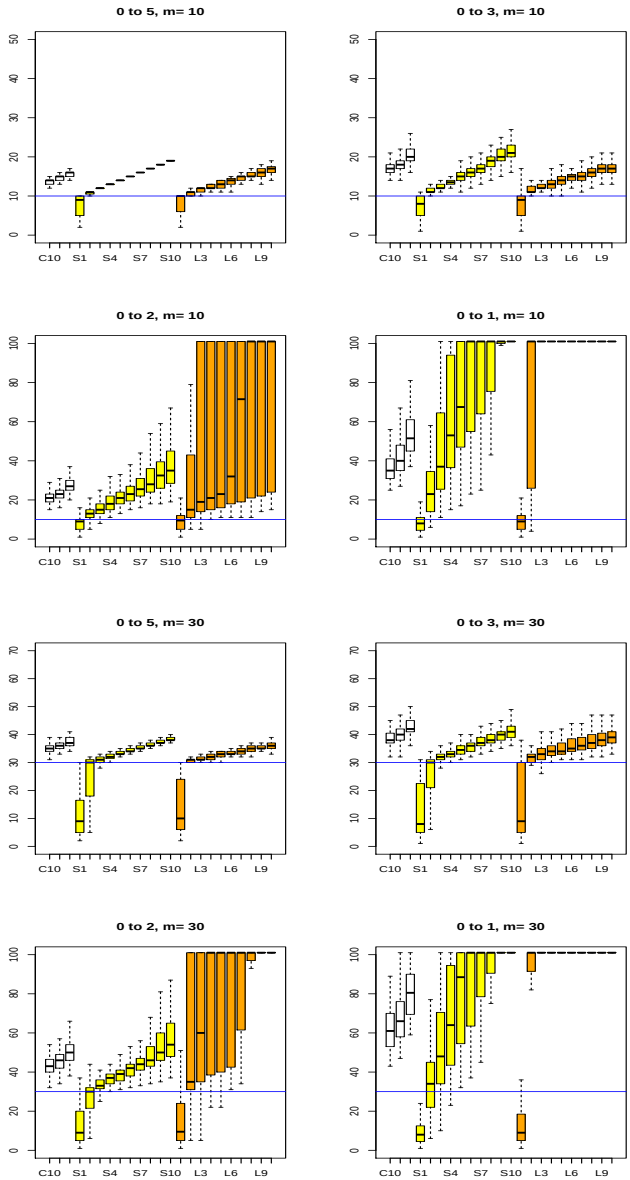

Figure 4.3. Boxplots of estimated stopping times obtained from 100 simulations with $n = 100$ when the mean changes from $\mu = 0$ to $\mu + \delta$ for $\delta \in \{5, 3, 2, 1\}$.

을 빠르게 잘 탐지할 수 있음을 확인할 수 있다. 물론 CUSUM 모니터링에 비해 벌점화 추정기법을 이용한 모니터링이 계산 속도가 훨씬 느리다는 단점이 있다. 또한 CUSUM 모니터링에서 구한 정지시간들이 변화가 일어난 뒤에 안정적으로 변화시점을 추정해주는 것에 비해 벌점화 추정방법을 이용하여 구한 정지시간들은 κ의 값에 따라 다른 현상들을 보이며 κ의 값을 선택하는 것이 문제이므로 이에 대한

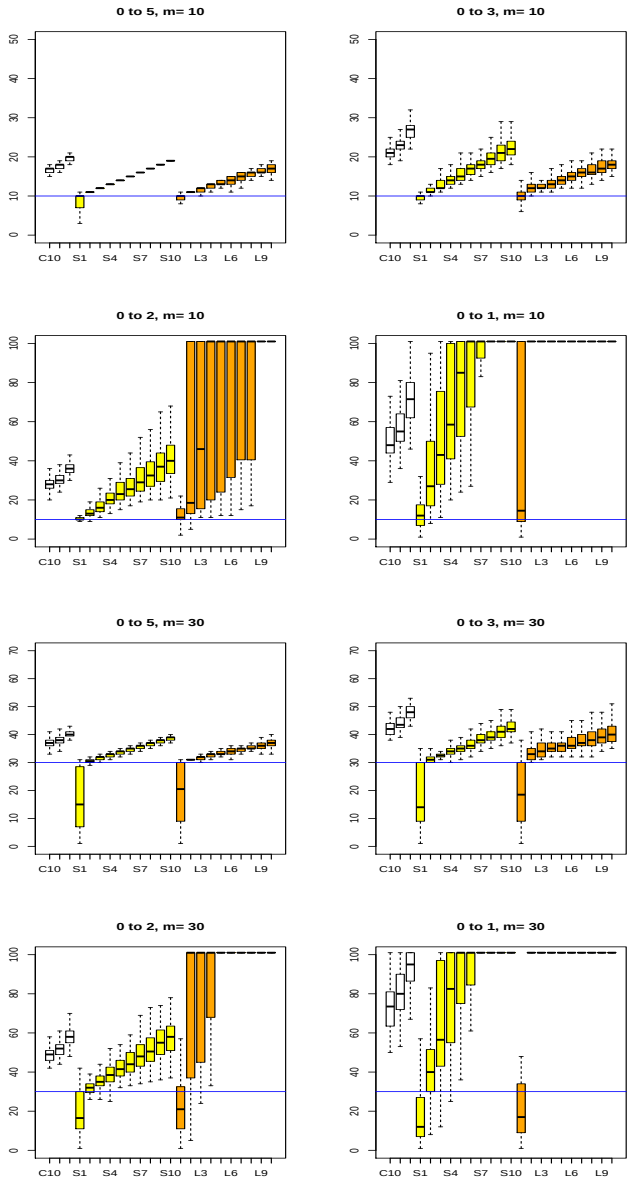

Figure 4.4. Boxplots of estimated stopping times obtained from 100 simulations with $n = 300$ when the mean changes from $\mu = 0$ to $\mu + \delta$ for $\delta \in \{5, 3, 2, 1\}$.

보완이 필요하다. 그러나 시간상의 문제로 본 연구에서는 이에 대한 연구는 생략하였다.

Table 4.1은 $\delta = 0$, 즉 평균이 변하지 않은 경우 100 이하의 값을 갖는 정지시간의 빈도를 정리해 놓은 것이다. 이 표를 보면 CUSUM 모니터링이 변화가 없는 경우 변화가 없다고 판단을 잘 하는 것에 비해 벌점화 추정방법을 이용한 모니터링의 경우는 κ의 선택에 세심한 주의가 필요함을 알 수 있다. κ의 값

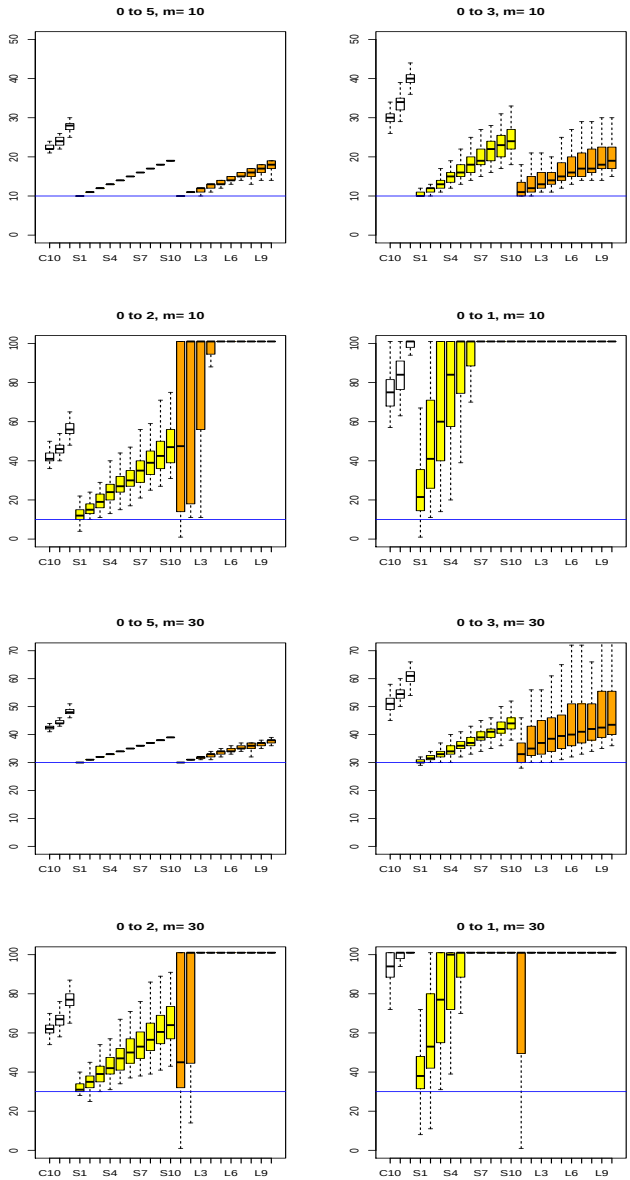

Figure 4.5. Boxplots of estimated stopping times obtained from 100 simulations with $n = 1000$ when the mean changes from $\mu = 0$ to $\mu + \delta$ for $\delta \in \{5, 3, 2, 1\}$.

이 작은 경우는 변화가 없음에도 불구하고 변화가 있다고 잘못 판단하는 비율이 매우 높다. 그러나 κ의 값이 증가하거나 과거 자료의 개수 n이 커지면 잘못 판단하는 비율이 빠르게 감소하는 것을 알 수 있다. 본 모의실험에서는 Table 4.1과 Figure 4.1부터 4.5를 근거로 각각의 n에 따라 잘못 판단한 비율이 약 10%에 가깝고 비교적 변화가 있을 때 변화를 잘 탐지하는 κ의 값들을 하나씩 선택하였다. $n \leq 10$인 경

Table 4.2. The sample quantiles of the estimator $\hat{\tau}_\kappa$ obtained from 100 simulations when $m = 10$ (The numbers in the parentheses are those of $\hat{\tau}^C_{0.10}$)

n	κ	δ	Q_0	$Q_{0.1}$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$Q_{0.9}$	Q_1
0	10	5	10.0	13.0	14.0	15.0	16.0	16.0	16.0
		3	10.0	12.9	14.0	15.0	15.0	16.1	18.0
		2	10.0	12.0	13.0	15.0	18.0	19.2	57.0
		1	14.0	19.0	53.5	101.0	101.0	101.0	101.0
10	10	5	14.0	18.0	18.0	19.0	20.0	20.0	21.0
			(10.0)	(11.0)	(11.8)	(12.0)	(13.0)	(14.0)	(16.0)
		3	15.0	18.0	19.8	22.0	100.0	100.0	100.0
			(9.0)	(12.0)	(13.0)	(14.0)	(16.3)	(19.0)	(22.0)
		2	14.0	19.0	25.0	100.0	100.0	100.0	100.0
			(11.0)	(12.9)	(15.0)	(17.5)	(22.0)	(26.0)	(88.0)
		1	19.0	30.0	100.0	100.0	100.0	100.0	100.0
			(12.0)	(15.9)	(23.8)	(34.0)	(60.0)	(101.0)	(101.0)
100	7	5	15.0	15.0	16.0	16.0	16.0	16.0	17.0
			(12.0)	(13.0)	(13.0)	(14.0)	(14.0)	(15.0)	(16.0)
		3	13.0	15.9	16.0	17.0	18.0	20.0	34.0
			(14.0)	(15.0)	(16.0)	(17.0)	(18.0)	(19.0)	(22.0)
		2	16.0	20.0	22.0	25.5	31.0	40.1	69.0
			(15.0)	(18.0)	(19.0)	(21.0)	(23.0)	(25.1)	(32.0)
		1	25.0	45.7	64.0	101.0	101.0	101.0	101.0
			(25.0)	(28.0)	(31.0)	(35.0)	(41.0)	(50.0)	(62.0)
300	5	5	13.0	13.0	14.0	14.0	14.0	14.0	15.0
			(14.0)	(15.0)	(16.0)	(17.0)	(17.0)	(18.0)	(20.0)
		3	13.0	14.0	14.0	15.0	17.0	19.0	28.0
			(18.0)	(19.0)	(20.0)	(21.0)	(22.0)	(23.0)	(26.0)
		2	15.0	17.0	20.0	23.0	29.0	35.0	45.0
			(20.0)	(24.0)	(26.0)	(28.0)	(30.0)	(31.0)	(36.0)
		1	24.0	40.9	52.8	85.0	101.0	101.0	101.0
			(29.0)	(39.0)	(44.0)	(48.0)	(57.0)	(64.1)	(81.0)
1000	3	5	11.0	12.0	12.0	12.0	12.0	12.0	12.0
			(20.0)	(21.0)	(22.0)	(22.0)	(23.0)	(23.0)	(24.0)
		3	11.0	12.0	12.0	13.0	14.0	16.0	20.0
			(26.0)	(28.0)	(29.0)	(30.0)	(31.0)	(33.0)	(37.0)
		2	11.0	13.9	16.0	19.0	23.0	27.0	48.0
			(36.0)	(37.0)	(40.0)	(41.0)	(44.0)	(46.0)	(50.0)
		1	14.0	31.0	40.0	60.0	101.0	101.0	101.0
			(57.0)	(63.9)	(68.0)	(75.0)	(81.3)	(87.1)	(101.0)

우엔 $\kappa = 10$이라고 하였고, n이 100, 300, 1000인 경우엔 각각 κ의 값으로 7, 5, 3을 선택한 후, SCAD 벌점함수를 이용하여 각각에 대응되는 정지시간들을 구하고, 이 값들을 CUSUM 모니터링 중 유의수준 $\alpha = 0.10$에 대응되는 정지시간과 비교하여 Table 4.2와 4.3에 정리하였다. 각각의 표에서 Q_p는 100번의 실험에서 얻은 정지시간의 $p \in \{0, 0.1, 0.25, 0.5, 0.75, 0.9, 1\}$ 분위수이다. 표 안에 정리된 결과들을 보면 과거 자료의 개수 n이 작거나 변화의 크기 δ가 작은 경우엔 CUSUM 모니터링이 변화시점을 좀 더 잘 탐지하지만, 반대로 과거 자료의 수 n이나 변화크기 δ의 값이 큰 경우엔 SCAD 벌점함수를 이용

Table 4.3. The sample quantiles of the estimator $\hat{\tau}_\kappa$ obtained from 100 simulations when $m = 30$ (The numbers in the parentheses are those of $\hat{\tau}^C_{0.10}$)

n	κ	δ	Q_0	$Q_{0.1}$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$Q_{0.9}$	Q_1
0	10	5	17.0	32.0	33.0	34.0	35.0	36.0	36.0
		3	12.0	32.0	33.0	34.0	35.3	37.0	39.0
		2	10.0	32.9	36.0	40.5	51.0	64.4	101.0
		1	9.0	63.7	101.0	101.0	101.0	101.0	101.0
10	10	5	14.0	34.9	36.8	38.5	40.0	40.0	40.0
			(11.0)	(30.9)	(33.0)	(35.0)	(37.0)	(38.0)	(40.0)
		3	19.0	34.9	38.0	41.0	100.0	100.0	100.0
			(9.0)	(33.0)	(35.0)	(39.0)	(44.0)	(49.1)	(56.0)
		2	21.0	36.0	44.0	100.0	100.0	100.0	100.0
			(13.0)	(35.0)	(38.0)	(46.0)	(54.3)	(62.2)	(89.0)
		1	20.0	42.5	100.0	100.0	100.0	100.0	100.0
			(15.0)	(38.9)	(53.8)	(77.0)	(101.0)	(101.0)	(101.0)
100	7	5	32.0	34.0	35.0	35.0	36.0	36.0	37.0
			(31.0)	(33.0)	(34.0)	(35.0)	(36.0)	(37.0)	(39.0)
		3	33.0	34.0	36.0	37.0	39.0	40.0	46.0
			(31.0)	(35.0)	(37.0)	(38.0)	(40.3)	(43.0)	(45.0)
		2	33.0	37.0	41.0	44.0	47.0	54.1	101.0
			(32.0)	(38.0)	(40.0)	(43.0)	(46.3)	(50.0)	(54.0)
		1	38.0	58.0	79.3	101.0	101.0	101.0	101.0
			(43.0)	(50.9)	(53.0)	(61.0)	(70.0)	(76.1)	(101.0)
300	5	5	31.0	33.0	33.0	34.0	34.0	34.0	35.0
			(33.0)	(35.0)	(36.0)	(37.0)	(38.0)	(39.0)	(41.0)
		3	31.0	33.0	34.0	35.0	36.0	39.0	45.0
			(38.0)	(40.0)	(40.0)	(42.0)	(44.0)	(46.0)	(48.0)
		2	32.0	35.9	38.0	41.5	46.0	50.1	66.0
			(42.0)	(44.0)	(46.0)	(49.0)	(51.0)	(54.0)	(59.0)
		1	32.0	56.9	75.0	101.0	101.0	101.0	101.0
			(50.0)	(61.0)	(63.8)	(73.5)	(81.0)	(89.0)	(101.0)
1000	3	5	31.0	31.0	32.0	32.0	32.0	32.0	33.0
			(39.0)	(41.0)	(42.0)	(42.5)	(43.0)	(44.0)	(46.0)
		3	28.0	31.0	32.0	33.0	34.0	36.0	37.0
			(45.0)	(48.0)	(49.0)	(51.0)	(53.0)	(54.0)	(60.0)
		2	30.0	32.0	35.0	39.0	43.0	47.1	57.0
			(52.0)	(57.0)	(60.0)	(62.0)	(64.0)	(66.0)	(76.0)
		1	31.0	44.0	56.0	77.0	101.0	101.0	101.0
			(72.0)	(83.9)	(88.8)	(94.0)	(101.0)	(101.0)	(101.0)

한 모니터링이 변화시점을 더 잘 탐지하는 것을 알 수 있다. 특히 $n \geq 300$이고 $\delta \geq 3$인 경우엔 변화시점과 정지시간의 차이가 별로 나지 않고, 변화가 일어나자마자 빠르게 탐지하는 것을 확인할 수 있다.

5. 결론

본 연구에서는 벌점화 최소제곱추정량을 기반으로 한 모니터링 절차를 제안하고, 모의실험을 통해 대표적인 방법인 CUSUM 모니터링과 비교분석을 하였다. 비록 변화시점의 추정량에 대한 이론적인 연구는

이루어지지 못했지만, 모의실험 결과를 보면 변화의 폭이 크고 모니터링 이전에 관측된 과거자료의 개수가 많은 경우 CUSUM 모니터링에 비해 변화시점을 빠르게 잘 추정함을 알 수 있었다. 그리고 벌점화 최소제곱추정량을 이용한 모니터링에서는 벌점함수의 선택, 정보함수의 선택, 정시시간의 기준으로 사용되는 κ, 변화시점의 위치와 크기 등이 모니터링 절차와 변화시점의 추정에 영향을 주는 것을 확인할 수 있었다.

본 연구의 자연스러운 후속 연구로 평균 차에 대한 벌점화 최소제곱추정량과 정지시간에 대해 이론적으로 연구할 수 있으며, 벌점함수, 정보함수, κ 등의 선택에 대한 심도 있는 연구와 평균 이외의 다양한 모수와 모형의 변화에 대한 확장도 고려할 수 있다.

References

Berkes, I., Gombay, E., Horváth, L., and Kokoszka, P. (2004). Sequential change-point detection in GARCH(p, q) models, *Econometric Theory*, **20**, 1140–1167.

Brown, R. L., Durbin, J., and Evans, J. M. (1975). Techniques for testing the consistency of regression relationships over time, *Journal of the Royal Statistical Society Series B*, **37**, 149–192.

Chu, C.S.J., Stinchcombe, M., and White, H. (1996). Monitoring structural change, *Econometrica*, **64**, 1045–1065.

Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*, Wiley, New York.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348–1360.

Horváth, L., Hušková, M., Kokoszka, P., and Steinebach, J. (2004). Monitoring changes in linear models, *Journal of Statistical Planning and Inference*, **126**, 225–251.

Horváth, L., Kühn, M., and Steinebach, J. (2008). On the performance of the fluctuation test for structural change, *Sequential Analysis*, **27**, 126–140.

Inclán, C. and Tiao, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variances, *Journal of the American Statistical Association*, **89**, 913–923.

Lee, S., Ha, J., Na, O., and Na, S. (2003). The cusum test for parameter change in time series models, *Scandinavian Journal of Statistics*, **30**, 781–796.

Leisch, F., Hornik, K., and Kuan, C. H. (2000). Monitoring structural changes with the generalized fluctuation test, *Econometric Theory*, **16**, 835–854.

Na, O., Lee, Y., and Lee, S. (2011). Monitoring parameter change in time series models, *Statistical Methods and Applications*, **20**, 171–199.

Ploberger, W. and Krämer, W. (1986). On studentizing a test for structural change, *Economic Letters*, **20**, 341–344.

Zeileis, A., Leisch, F. Kleiber, C., and Hornik, K. (2005). Monitoring structural change in dynamic econometric models, *Journal of Applied Econometrics*, **20**, 99–121.

벌점화 추정기법을 이용한 평균에 대한 모니터링

나옥경[a] · 권성훈[b,1]

[a]경기대학교 응용정보통계학과, [b]건국대학교 응용통계학과

(2016년 11월 9일 접수, 2016년 12월 7일 수정, 2016년 12월 7일 채택)

요 약

본 연구에서는 벌점화 최소제곱추정방법을 이용하여 평균의 변화를 모니터링할 수 있는 방법에 대해 연구하였다. 모니터링 이전의 공통 평균과 모니터링을 시작한 이후 순차적으로 관측되는 관측값들의 평균의 차이를 벌점화 최소제곱추정방법을 이용하여 추정하였으며, 이 추정값들에서 0이 아닌 것의 개수를 바탕으로 모니터링 절차를 개발하였다. 이는 기존의 모니터링 절차들이 순차적으로 얻은 추정값들의 변동성을 기반으로 만들어진 것과 다른 점이다. 모의실험을 통해 본 연구에서 제안한 모니터링 절차가 가지고 있는 특징들을 살펴보았고, 대표적인 모니터링 절차인 CUSUM 모니터링과 비교 분석도 하였다.

주요용어: CUSUM 모니터링, FL 모니터링, LASSO, 벌점화 추정, SCAD

이 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2014R1A1A1002995).

[1]교신저자: (05029) 서울시 광진구 능동로 120, 건국대학교 응용통계학과. E-mail: shkwon0522@gmail.com