한국통계학회논문집
2011, 18권, 2호, 245–255

DOI: 10.5351/CKSS.2011.18.2.245

분포변화 검정에서 경험확률과정과 커널밀도함수추정량의 검정력 비교

나성룡[a], 박현아[1,b]

[a]연세대학교 정보통계학과, [b]서울대학교 통계학과

요 약

자료의 분포변화를 검정하는 비모수적 방법으로 경험분포함수를 이용하거나 확률밀도함수 추정량을 이용하는 두 가지 방법을 고려할 수 있다. 이 논문에서는 분포변화 검정을 위한 두 가지 방법을 자세히 살펴보고 기존 연구의 결과를 정리한다. 여러 확률모형을 가정하고 분포변화 검정에 대한 모의실험을 실시하여 두 방법에 대한 이론적 극한 성질이 잘 성립하는가를 살펴본다. 검정력 비교를 통하여 모형에 따른 적절한 변화점 분석 방법을 알아본다.

주요용어: 분포변화 검정, 경험분포함수, 확률밀도함수 추정량, 비모수적 검정, 모의실험.

1. 서론

변화점 분석의 가장 일반적인 형태는 자료의 분포변화를 다루는 것이라 할 수 있으며 분포변화를 검정하기 위한 비모수적 방법으로 경험분포함수에 기초한 방법을 전통적으로 사용해왔다. Kolmogorov-Smirnov 검정, Anderson-Darling 검정 등의 비모수적 적합도 검정에서 중요한 역할을 담당하는 경험분포함수(empirical distribution function) 혹은 경험확률과정(empirical process)은 분포변화 검정을 위하여 순차적 경험확률과정(sequential empirical process)의 형태로 일반화된다. Bickel과 Wichura (1971), Shorack과 Wellner (1986)에는 독립 표본에서 정의된 순차적 경험확률과정이 Kiefer 확률과정으로 수렴하는 사실이 잘 증명되어 있으며 이는 변화점 분석의 이론적 기초를 제공한다.

정책 등의 외부 요인의 변화에 많은 영향을 받는 경제 및 사회 현상을 반영하는 자료에는 변화점이 있을 가능성이 크고 따라서 시계열 모형에서의 분포변화 검정은 중요한 문제라 하겠다. 다양한 시계열 모형에 대한 순차적 경험확률과정과 분포변화 검정 문제를 다룬 연구로 Picard (1985), Bai (1994), Koul (1996), Ling (1998) 등이 있다. Horváth 등 (2001)는 ARCH 모형에서의 순차적 경험확률과정을 다루면서 분포변화 검정에 대한 분포 무관 방법을 유도했으며 Berkes와 Horváth (2003)는 이 결과를 GARCH 모형으로 확장하였다. 또한 Na 등 (2006)은 측정 오차를 가지는 자기회귀모형에 대한 순차적 경험확률과정의 극한 성질을 유도하고 이에 기초한 분포 변화 검정 방법을 제안하였다. 시계열 모형에서 다루어지는 (순차적) 경험확률과정은 오차의 추정값인 잔차에 기초하는 것이 일반적이다. 시계열 자체에 기초하는 경험확률과정은 종속관계에 영향을 받는 극한분포를 가지기 때문에 변화점 분석에서 이론적인 기각역을 결정하기 어렵고 독립성을 가정하는 오차는 관측이 불가능하므로 직접 경험확률과정을 정의할 수 없기 때문이다.

적절한 종속관계를 만족하는 시계열 모형에서 시계열 자료를 직접 이용하는 방법이 Lee와 Na (2004)에서 제시되는데 경험분포함수 대신에 (순차적) 확률밀도함수 추정량이 변화점 검정에 이용된

[1] 교신저자: (151-742) 서울 관악구 신림동, 서울대학교 통계학과, 박사후연구원. E-mail: parkha03@yahoo.co.kr

다. 시계열 자료를 직접 이용하므로 잔차를 계산하기 위한 복잡한 모형 추정 과정을 거치지 않아도 되는 장점이 있다. 최근에 나성룡 (2009)에 의하여 이 방법은 종속 오차를 가지는 선형회귀모형의 변화점 분석 문제로 확장되었는데 다양한 시계열 모형을 포함한다.

지금까지 비모수적 변화점 분석을 위한 분포함수 방법과 확률밀도함수 방법을 직접 비교하는 연구는 이루어지지 않았다. 이 논문에서는 시계열 모형에서 정의되는 순차적 경험확률과정과 순차적 확률밀도함수 추정량의 비교를 직접 다루고자 한다. 특히 분포변화에 대한 변화점 검정에 적용했을 때 두 추정량에 기초한 검정 방법의 유한 표본 특성을 비교하는 것이 주요 목적이다. 이를 위하여 이론에 기초한 기각역을 계산하고 모의실험을 통하여 각각의 추정량에 기초한 검정의 경험 검정력을 산출한다. 설정된 각 시계열 모형에서 어떤 방법이 분포의 변화점 분석에 더 효율적인가를 산출된 검정력을 비교해서 알아본다.

이 논문의 구성은 다음과 같다. 먼저 2절에서 경험분포함수, 커널 확률밀도함수 추정량을 정의하고 순차적 경험확률과정과 순차적 확률밀도함수 추정량에 대한 기존의 연구 결과를 정리한다. 분포변화의 검정에 필요한 통계량과 이론적인 기각역을 정리한다. 3절에는 분포변화 변화점 검정에 대한 모의실험 결과가 주어진다. 독립 자료를 비롯한 다양한 시계열 모형을 설정하고 각 모형에 가능한 여러 변화점 분석 방법을 시뮬레이션한 결과를 비교해서 의미있는 결과를 찾아 본다. 이 논문의 연구 결과를 4절에서 최종적으로 정리한다.

2. 순차적 경험확률과정과 확률밀도함수 추정량

이 절에서는 경험분포함수와 확률밀도함수 추정량을 정의하고 변화점 분석과 관련한 중요한 성질을 알아본다. 먼저 $X_1, X_2, \ldots, X_n$을 서로 독립이고 동일한 분포함수 $F(x)$를 가지는 확률변수들의 확률과정이라 하자. 분포함수 $F(x)$를 추정하기 위한 경험분포함수는 실수 x에 대하여

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$

식으로 정의되며, 여기에서 $I(\,)$는 지시함수를 의미한다. 이때 경험분포함수를 표준화한 확률과정

$$n^{\frac{1}{2}}(F_n(x) - F(x)) = n^{-\frac{1}{2}}\sum_{i=1}^{n}(I(X_i \le x) - F(x)), \quad -\infty < x < \infty \tag{2.1}$$

을 경험확률과정이라 한다. 독립 자료에 대한 경험확률과정이 브라운 다리(Brownian bridge)로 수렴하는 것은 잘 알려진 사실이며 Kolmogorov 검정 등의 비모수적 적합도 검정에서 중요한 역할을 한다.

경험확률과정 (2.1)을 확장한 순차적 경험확률과정은

$$K_n(s,x) = n^{-\frac{1}{2}}\sum_{i=1}^{[ns]}(I(X_i \le x) - F(x)), \quad 0 \le s \le 1, \ -\infty < x < \infty \tag{2.2}$$

식으로 정의되는 이차원 확률과정이다. 순차적 경험확률과정이 Kiefer 확률과정으로 약수렴(weak convergence)한다는 사실은 Shorack과 Wellner (1986)의 Theorem 3.5.1을 참고할 수 있으며 다음과 같이 정리할 수 있다. 그리고 약수렴에 대한 엄밀한 정의는 Billingsley (1999)를 참고할 수 있다.

정리 1. 식 (2.2)의 이차원 확률과정 $\{K_n(s,x) : 0 \le s \le 1, -\infty < x < \infty\}$에 대하여

$$K_n(s,x) \Rightarrow K(s, F(x))$$

의 약수렴이 성립한다. 여기에서 $\{K(s,t) : (s,t) \in [0,1]^2\}$은 $(s,t) \in [0,1]^2$, $(s',t') \in [0,1]^2$에 대하여

$$E[K(s,t)] = 0, \qquad E[K(s,t)K(s',t')] = \min(s,s')\left(\min(t,t') - tt'\right)$$

을 만족하는 이차원 정규확률과정, 즉 Kiefer 확률과정이다.

변화점 분석을 위해서 자료의 앞 부분과 뒷 부분에 기초한 분포함수 추정량의 차이를 비교한다. 즉

$$F_{[ns]}(x) = \frac{1}{[ns]} \sum_{i=1}^{[ns]} I(X_i \le x),$$

$$F^*_{n-[ns]}(x) = \frac{1}{(n-[ns])} \sum_{i=[ns]+1}^{n} I(X_i \le x)$$

의 차이를 비교하는데

$$\begin{aligned} D_n(s,x) &= n^{\frac{1}{2}} \frac{[ns]}{n} \frac{n-[ns]}{n} \left(F_{[ns]}(x) - F^*_{n-[ns]}(x)\right) \\ &= K_n(s,x) - \frac{[ns]}{n} K_n(1,x) \end{aligned}$$

에 기초한 분석을 수행한다. 여기에 정리 1의 결과를 적용하면

$$D_n(s,x) \Rightarrow K(s,F(x)) - sK(1,F(x)) \tag{2.3}$$

을 쉽게 얻으며, 여기에서 $\bar{K}(s,t) = K(s,t) - sK(1,t)$은 $(s,t) \in [0,1]^2$, $(s',t') \in [0,1]^2$에 대하여

$$E\left[\bar{K}(s,t)\right] = 0, \qquad E\left[\bar{K}(s,t)\bar{K}(s',t')\right] = \left(\min(s,s') - ss'\right)\left(\min(t,t') - tt'\right)$$

을 만족하는 정규확률과정이다.

참고 1. 자료에서 계산된 $D_n(s,x)$의 값이 크면 분포의 변화를 고려하게 된다. 검정통계량으로

$$\|D_n\|_\infty = \sup_{0\le s\le 1} \sup_{-\infty<x<\infty} |D_n(s,x)|$$

을 사용하는 것이 일반적이고 식 (2.3)에 기초해서 이론적으로 근사 기각역을 계산한다. 가령 유의수준 α의 검정을 위해서 $\|D_n\|_\infty \ge C_\alpha$일때 분포변화를 채택하는데, C_α는 $\sup_{(s,t)\in[0,1]^2} |\bar{K}(s,t)|$의 $1-\alpha$ 분위수이다. 이 분위수들은 Picard (1985)에 제공된다.

한편 $X_1, \ldots, X_n$이 종속성을 가정하는 시계열 자료인 경우에는 식 (2.1)과 (2.2)의 확률과정이 자료의 종속관계에 영향을 받는 극한을 가진다는 사실을 Berkes 등 (2009) 등에서 볼 수 있다. 결과적으로 검정통계량이 (2.3)보다는 훨씬 복잡한 극한분포를 가지게 되고 참고 1에서와 같이 종속관계와 무관하게 검정을 수행하기는 어렵다. 이러한 이유 등으로 기존 시계열 모형의 변화점 분석에서는 일반적으로 독립을 가정하는 오차항에 주목하였다. 그런데 오차는 관측이 불가능하기 때문에 자료에서 잔차를 계산하여 대신 사용한다. 예를 들면 서로 독립인 오차항 ϵ_i을 가정하는

$$X_i = \phi_1 X_{i-1} + \cdots + \phi_p X_{i-p} + \epsilon_i + \theta_1 \epsilon_{i-1} + \cdots + \theta_q \epsilon_{i-q}$$

의 ARMA 모형에서 모수에 대한 적절한 일치 추정량을 구하여 잔차를

$$\hat{\epsilon}_i = X_i - \hat{\phi}_1 X_{i-1} - \cdots - \hat{\phi}_p X_{i-p} - \hat{\theta}_1 \hat{\epsilon}_{i-1} - \cdots - \hat{\theta}_q \hat{\epsilon}_{i-q}$$

식으로 정의할 수 있다. 이때 잔차를 이용한 순차적 경험확률과정에 대하여 오차를 이용할 때와 같이 정리 1의 결과가 성립한다는 사실을 Bai (1994)가 증명하였다. 비슷한 연구가 다양한 시계열 모형에서 이루어졌는데 Bai와는 다르게 모수 추정량이 극한에 포함되는 결과가 많다. 하지만 분포변화를 위한 검정통계량에는 모수 추정량 부분이 상쇄되어 식 (2.3)의 결과가 다시 유도되고 분포무관 검정이 가능해진다. 서론에서 언급된 Koul (1996), Ling (1998), Horváth 등 (2001), Berkes와 Horváth (2003), Na 등 (2006) 등을 참고할 수 있다.

시계열을 직접 이용하지만 종속관계와 무관한 검정으로 Lee와 Na (2004)는 경험분포함수 대신 확률밀도함수 추정량을 이용하는 방법을 제안한다. 동일한 분포를 따르는 정상 시계열 자료 $X_1, \ldots, X_n$이 분포함수 $F(x)$와 확률밀도함수 $f(x)$를 가짐을 가정하자. 이때 $f(x)$의 비모수적 추정량으로

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

식으로 정의되는 커널밀도함수추정량(kernel density estimator)을 사용하는데, 여기에서 $K(x)$는 커널함수이고 $h = h_n$은 밴드폭(band width)이다. 커널함수, 밴드폭의 선택 및 $f_n(x)$의 성질에 대한 참고문헌으로는 Silverman (1986), Bosq (1998) 등이 있다.

변화점 분석을 위해서 순차적 커널밀도함수추정량

$$f_{[ns]}(x) = \frac{1}{[ns]h} \sum_{i=1}^{[ns]} K\left(\frac{x - X_i}{h}\right),$$

$$f^*_{n-[ns]}(x) = \frac{1}{(n-[ns])h} \sum_{i=[ns]+1}^{n} K\left(\frac{x - X_i}{h}\right)$$

의 차이에 기초한

$$d_n(s,x) = \left(\frac{nh}{f_n(x)\|K\|^2}\right)^{\frac{1}{2}} \frac{[ns]}{n} \frac{n-[ns]}{n} \left(f_{[ns]}(x) - f^*_{n-[ns]}(x)\right), \quad 0 \le s \le 1 \tag{2.4}$$

값을 이용한다. 여기에서 $\|K\|^2 = \int K^2(x)dx$이다. Lee와 Na (2004)는 정상 시계열 X_i들이 어떤 $\gamma \ge 3$에 대하여

$$\alpha(k) = \sup_{t \ge 1} \sup \left\{ |P(A \cap B) - P(A)P(B)| : A \in \mathcal{M}_1^t, B \in \mathcal{M}_{t+k}^{\infty} \right\} = O(k^{-\gamma}) \tag{2.5}$$

조건이 성립하는 강혼합(strong mixing) 성질을 만족할 때 다음의 사실을 증명하였다. 단, $\mathcal{M}_1^t$는 $X_1, \ldots, X_t$에 의하여 생성되는 시그마필드이고 $\mathcal{M}_{t+k}^{\infty}$는 $X_{t+k}, X_{t+k+1} \ldots$로 생성되는 시그마필드이다.

정리 2. $f(x_i) > 0$을 만족하는 서로 다른 $x_1, \ldots, x_m$에 대하여

$$(d_n(s,x_1), \ldots, d_n(s,x_m)) \Rightarrow \left(W_1^0(s), \ldots, W_m^0(s)\right), \quad 0 \le s \le 1$$

의 약수렴이 성립하며 $W_1^0(s), \ldots, W_m^0(s)$는 서로 독립인 브라운 다리들이다.

정리 2에 필요한 밴드폭, 커널함수 및 모형에 대한 자세한 내용은 Lee와 Na (2004)를 참고할 수 있다. 한편 여기에서 유의해야 할 점은 정리 1 또는 식 (2.3)과는 달리 확률밀도함수 추정량에 대한 정리 2에서는 고정된 유한개의 x_i에 대한 수렴을 다루는 것이다. 현재로서는 확률밀도함수 추정량에 기초한 식 (2.4) 확률과정의 수렴결과는 고정된 유한 점에 대해서만 규명되어 있다. 순차적 경험확률과정의 Kiefer 확률과정으로의 수렴과 같은 보편적인 결과의 유도는 훨씬 더 진행되어야할 도전적인 과제라 하겠다.

참고 2. 분포에 대한 변화점 분석을 위해서 검정통계량

$$\|d_n\|_\infty = \max_{1\le i\le m} \sup_{0\le s\le 1} |d_n(s, x_i)|$$

값이 크면 정상성을 기각하고 분포변화를 채택한다. 정리 2의 결과를 이용해서 유의수준 α의 기각역을 $P(\sup_{0\le s\le 1} |W_1^0(s)| \ge C_\alpha) = 1 - (1-\alpha)^{1/m}$을 만족하는 C_α에 대하여 $\|d_n\|_\infty \ge C_\alpha$로 정한다. 정규분포 등의 주요 연속 확률분포를 고려해보면 m의 값은 2 내지 4로 하고 자료를 $m+1$개로 등분하는 값을 x_i로 정의하는 것이 적절하리라 판단된다.

참고 3. 실제 문제에서는 m과 x_i들의 선택이 중요하다. 확률밀도함수 값의 차이가 크거나 크리라 예상되는 1개의 점을 정해서 검정을 하는 것이 이론적으로는 효과적일 수 있다. 이것이 어려우므로 여러 개의 비교점을 사용하게 되는데 m의 값이 지나치게 크고 x_i들이 인접하게 되면 $d_n(s, x_i)$들의 종속성으로 인하여 유한표본 특성이 극한분포와 많이 달라진다. 한편 비교점을 m개 사용한다면 자료에 대한 누적확률을 $1/(m+1)$로 등분하는 점을 사용하는 것이 특별한 사전정보가 없는 상황에서는 타당하리라 판단된다. 참고로 정규분포, 혼합정규분포 등을 가정한 다양한 모의실험의 결과에서 m의 값이 5 이상되면 검정력이 더 이상 좋아지지 않거나 오히려 떨어지는 경우가 많았고 따라서 이 논문에서는 $m = 3$과 사분위수를 사용한 모의실험 결과를 제시한다. 확률모형에 따라 검정 결과는 많이 달라질 수 있으리라 보는데 최적의 m과 x를 정하는 문제는 계속 연구되어야 하겠다.

최근에 나성룡 (2009)은 정리 2의 결과를 오차에 종속성을 허용하는 선형회귀모형으로 확장하였다. 즉

$$X_i = \beta_1 z_{i1} + \cdots + \beta_p z_{ip} + \epsilon_i, \quad i = 1, \ldots, n$$

의 회귀모형에서 오차항 ϵ_i는 식 (2.5)의 약한 종속성을 가짐을 가정한다. $\beta_1, \ldots, \beta_p$는 미지의 회귀계수, $z_{i1}, \ldots, z_{ip}$는 고정 혹은 변동 모형의 독립변수이며 $X_1, \ldots, X_n$들이 관측된다. 회귀계수에 대한 $\sqrt{n}$-일치 추정량 $\hat\beta_1, \ldots, \hat\beta_p$을 이용하여 잔차를

$$\hat\epsilon_i = X_i - \hat\beta_1 z_{i1} - \cdots - \hat\beta_p z_{ip}$$

식으로 정의할 때 잔차에 기초한 $d_n(s, x_i)$에 대하여 정리 2가 성립한다. 한편 나성룡 (2009)에서 다루는 모형은 일반적인 중회귀모형, ARMA 모형, 측정오차 자기회귀모형 등의 다양한 모형을 포괄한다.

3. 분포변화 검정에 대한 모의실험

이 절에서는 앞에서 설명한 비모수적 변화점 분석 방법에 대한 유한 표본 확률적 특성을 모의실험을 통하여 살펴 본다. 검정통계량의 유한 표본 분포가 극한분포로 잘 수렴하는가를 확인하기 위하여 분포변화가 없다고 가정하는 귀무가설 하에서 이론적인 기각역에 대한 검정력을 모의실험으로 구해 본

표 1: 독립 모형에서의 경험 검정력

n	size		power	
	ED	KD	ED	KD
50	0.026	0.011	0.213	0.238
100	0.033	0.017	0.652	0.731
200	0.056	0.015	0.986	0.99
300	0.052	0.027	1	1
500	0.046	0.018	1	1

다. 한편 효율적인 검정법을 알아보기 위하여 모형에 분포 변화를 가정하고 각 검정법의 경험 검정력을 산출해서 비교한다.

분포변화 검정을 위한 비모수적인 방법으로 경험분포함수의 차이 즉 순차적 경험확률과정에 기초한 방법이 거의 유일했는데 최근에 Lee와 Na (2004), 나성룡 (2009)에 의하여 확률밀도함수추정량을 이용하는 방법이 새롭게 연구되었다. 각각의 경우에 $\|D_n\|_\infty$와 $\|d_n\|_\infty$ 통계량이 기존 연구에서 주로 다루어졌는데 직관적으로 이해하기 쉽기 때문이다. 이외에 적합도 검정에서의 Cramér-von Mises 혹은 Anderson-Darling 통계량과 같이 L^2-norm에 기초한 검정법 등을 고려할 수 있는데 검정통계량이 순차적 경험확률과정 혹은 확률밀도함수추정량의 함수 형태이므로 $\|D_n\|_\infty$ 혹은 $\|d_n\|_\infty$와 유사한 확률특성을 가질 것이 예상되므로 기존 문헌에서 다른 검정법은 거의 다루지 않고 있다. 이 논문에서는 $\|D_n\|_\infty$와 $\|d_n\|_\infty$의 비교를 중점적으로 살펴본다.

앞 2절의 참고 1에서 설명한 분포함수를 이용한 변화점 분석에서는 유의수준 5%에서 $\|D_n\|_\infty \geq 0.815$일 때 귀무가설을 기각한다. 여기에서 0.815는 $\sup_{(s,t)\in[0,1]^2} |\bar{K}(s,t)|$의 0.95-분위수에 해당하는 값인데 10,000번의 모의실험으로 산출한 값이다. 참고 2의 검정을 위한 커널밀도함수추정량을 구하기 위해 커널함수는 $N(0,1)$ 확률밀도함수 $K(x) = (2\pi)^{-1/2}e^{-x^2/2}$, 밴드폭은 $h = 0.2n^{-1/5}\log\log n$을 사용한다. 비교점의 수는 $m = 3$이고 비교점으로 3개의 사분위점을 사용한다. 이 경우 유의수준 5%의 기각역은 $\|d_n\|_\infty \geq 1.545$와 같다 (*cf.* Smirnov, 1948). 이 절에서는 유의수준 5%의 검정을 모의실험한다.

분포변화가 없다고 가정하는 귀무가설을 위하여 $N(0,1)$의 *i.i.d.* 자료 $Y_1, \ldots, Y_n$을 생성한다. 반면 분포변화를 위하여 $N(0,1)$의 독립 자료 $Y_1, \ldots, Y_{[n/2]}$과

$$Y_i = \frac{[(1-B_i)X_{1i} + B_iX_{2i}]}{[(1-\nu)+\sigma^2\nu]^{\frac{1}{2}}} \tag{3.1}$$

식을 만족하는 독립 자료 $Y_{[n/2]+1}, \ldots, Y_n$을 생성한다. 여기에서 $B_i \sim i.i.d.\, B(1,\nu)$, $X_{1i} \sim i.i.d.\, N(0,1)$, $X_{2i} \sim i.i.d.\, N(0,\sigma^2)$이며 $\sigma^2 = 10^2$, $\nu = 0.2$으로 설정한다. 위 식 (3.1)은 혼합정규분포의 확률변수이며 평균 0과 분산 1이 되도록 표준화되었다. $N(0,1)$과 식 (3.1)의 표준화된 혼합정규분포의 분포함수는 그림 1에 표현되었다. 그림 2는 이들 분포의 확률밀도함수를 나타낸다. 그래프의 형태에서 분포변화 검정의 효과가 분명히 나타날 것을 예상할 수 있다.

독립 모형에 대한 두 검정법의 경험 검정력은 표 1에 정리되어 있다. 표에서 size는 분포변화가 없는 귀무가설이 성립할 때 귀무가설을 기각한 비율이고 power는 분포변화가 있는 자료에서 귀무가설을 기각한 비율이다. ED는 경험분포함수를 이용한 방법을 나타내고 KD는 커널밀도함수추정량을 이용한 방법을 의미한다. 표본 크기는 $n = 50, 100, 200, 300, 500$이 고려되었고 모든 size와 power는 1,000번의 반복 횟수를 갖는 모의실험을 통하여 계산되었다.

표 1의 결과에서 ED 방법의 size가 명목 유의수준과 잘 일치하는 것을 볼 수 있다. 반면에 KD 방법에 대하여 어느 정도의 유의수준 왜곡(size distortion)이 있음을 보게 되는데 이는 극한으로의 수렴 속도가 ED 방법에 비하여 느림을 의미하는 것으로 해석할 수 있다. 그러나 그 정도가 심각하지는 않기

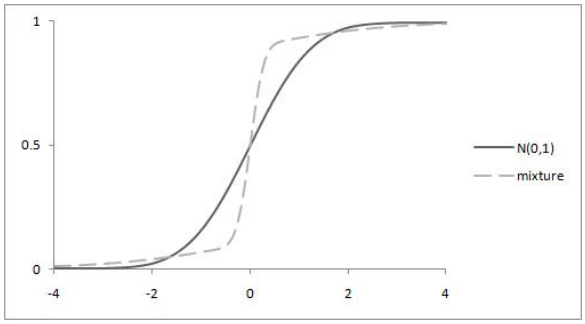

그림 1: *분포함수*

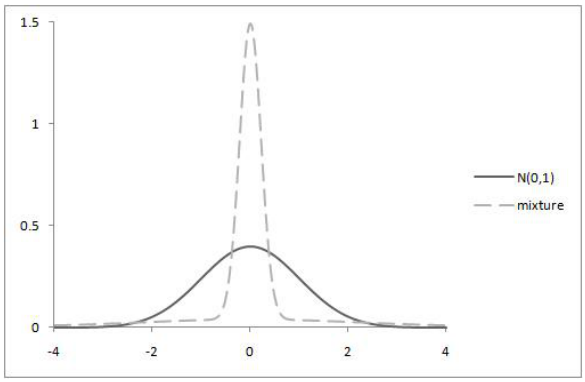

그림 2: *확률밀도함수*

때문에 이론적인 극한이 성립함을 확인하기에는 부족하지 않다. 한편 표본 크기가 작을 때 KD 방법의 power가 크게 나타나며, 300 이상에서는 두 방법 모두 거의 완벽한 검정을 수행함을 보게 된다. 이는 극한으로의 수렴속도 차이에 의하여 경험확률과정 방법이 독립 모형에서 더 효율적일 것이라는 일반적인 예상과는 반대되는 결과이다. 대립가설에서 설정한 식 (3.1)의 혼합 정규분포 특성에 기인한 것으로 해석가능한데 모형에 따라서는 KD 방법이 더 효율적일 수 있음을 알 수 있다.

변화점 분석을 위한 종속 모형의 예로 AR(1) 모형을 고려하였고, 특히 오차항의 분포 변화를 검정하는 문제를 다룬다. 즉 $|\phi| < 1$인 ϕ에 대하여 $Y_0 = 0, Y_i = \phi Y_{i-1} + \epsilon_i$ 식을 만족하는 시계열 $Y_1, \ldots, Y_n$을 생성한다. 이때 귀무가설로는 오차항 ϵ_i에 대하여 $\epsilon_1, \ldots, \epsilon_n \sim i.i.d.N(0,1)$을 가정한다. 반면 대립가설에 대하여 독립인 $N(0,1)$의 $\epsilon_1, \ldots, \epsilon_{[n/2]}$과 식 (3.1)을 만족하는 $\epsilon_{[n/2]+1}, \ldots, \epsilon_n$을 가정한다. 자기회귀계수의 추정량은 $\hat{\phi} = \sum_{i=1}^{n-1} Y_i Y_{i+1} / \sum_{i=1}^{n-1} Y_{i-1}^2$이고, 잔차는

$$\hat{\epsilon}_i = Y_i - \hat{\phi} Y_{i-1}, \quad i = 1, \ldots, n$$

식으로 계산한다. 잔차에 기초한 변화점 분석의 모의실험 결과는 표 2에 정리되어 있다. 시계열의 상관 정도를 표현하는 자기회귀계수 ϕ의 값으로 0.1, 0.3, 0.5, 0.8을 사용하였다.

표 2에서 경험분포함수 방법을 나타내는 ED_y와 ED_r은 각각 시계열 자체 혹은 잔차를 검정에 이용한다. 한편 KD_y와 KD_r은 각각 시계열과 잔차에 기초한 커널밀도함수추정량 방법을 의미한다. 일반적으로 알려진 바와 같이 ED_y 방법은 종속 자료에 직접 적용하기 어렵다는 사실을 size에서 확인할 수 있다. 시계열 자체의 확률밀도함수추정량을 이용하는 KD_y 방법은 어느 정도의 상관관계는 잘 견디는데 아주 높은 상관관계에서는 유의수준 왜곡과 검정력 손실을 보인다. 반면에 잔차를 이용하

표 2: AR(1) 모형에 대한 변화점 분석의 경험 검정력

n	size				power			
	ED_y	ED_r	KD_y	KD_r	ED_y	ED_r	KD_y	KD_r
				$\phi = 0.1$				
50	0.063	0.028	0.009	0.012	0.256	0.198	0.225	0.206
100	0.063	0.043	0.017	0.011	0.649	0.617	0.649	0.664
200	0.073	0.044	0.018	0.017	0.979	0.985	0.980	0.983
300	0.083	0.057	0.014	0.01	0.999	1	0.999	1
500	0.077	0.049	0.019	0.03	1	1	1	1
				$\phi = 0.3$				
50	0.124	0.018	0.014	0.012	0.342	0.220	0.198	0.206
100	0.146	0.032	0.007	0.016	0.653	0.624	0.469	0.688
200	0.204	0.036	0.024	0.013	0.920	0.975	0.859	0.986
300	0.184	0.047	0.027	0.016	0.988	0.999	0.974	1
500	0.209	0.061	0.035	0.018	1	1	1	1
				$\phi = 0.5$				
50	0.247	0.016	0.021	0.011	0.438	0.199	0.163	0.178
100	0.305	0.023	0.029	0.018	0.714	0.622	0.353	0.674
200	0.367	0.040	0.036	0.021	0.902	0.980	0.641	0.98
300	0.370	0.041	0.032	0.022	0.956	1	0.81	1
500	0.394	0.056	0.027	0.025	0.998	1	0.959	1
				$\phi = 0.8$				
50	0.663	0.021	0.059	0.006	0.790	0.236	0.214	0.223
100	0.798	0.028	0.072	0.012	0.888	0.612	0.296	0.664
200	0.891	0.040	0.103	0.022	0.952	0.977	0.342	0.978
300	0.913	0.060	0.095	0.010	0.967	0.998	0.415	1
500	0.928	0.056	0.106	0.019	0.990	1	0.481	1

는 ED_r, KD_r 방법은 시계열의 상관관계와 무관한 우수한 검정력을 보이며 전반적으로 독립 모형과 유사한 결과를 얻게 된다.

변화점 분석을 위한 또 다른 모형으로 측정오차 AR(1) 모형을 고려하였다. 이 모형은 나성룡 (2009)에서 다루어진 종속 오차를 가지는 선형회귀모형의 일종이다. 여기에서 $X_0 = 0$, $X_i = \phi X_{i-1} + \epsilon_i$ 식으로 AR 자료를 생성하고, $\delta_1, \ldots, \delta_n \sim i.i.d.\, N(0, 0.1^2)$의 측정오차에 대하여 $Y_i = X_i + \delta_i$ 식을 만족하는 관측값 $Y_1, \ldots, Y_n$을 생성한다. AR의 오차항 ϵ_i에 대한 분포 변화를 가정하는데 귀무가설과 대립가설은 표 2의 AR(1) 모형의 오차항에 대한 가정과 같다. 한편 측정오차 모형의 변화점 분석은 ϵ_i 대신에 $v_i = Y_i - \phi Y_{i-1} = \epsilon_i + \delta_i - \phi\delta_{i-1}$에 대하여 수행할 수 있다 (*cf.* Na 등, 2006). 자기회귀계수는 $\hat{\phi} = \sum_{i=1}^{n-2}(Y_i - \bar{Y})(Y_{i+2} - \bar{Y}) / \sum_{i=1}^{n-1}(Y_i - \bar{Y})(Y_{i+1} - \bar{Y})$ 식으로 추정하고, 잔차는

$$\hat{v}_i = Y_i - \hat{\phi} Y_{i-1}, \quad i = 1, \ldots, n$$

식으로 주어지며 $Y_0 = 0$을 가정한다. 측정오차 AR(1) 모형의 변화점 분석 모의실험 결과를 표 3에 정리하였다.

표 3에서 ED_r1은 잔차 전부를 이용한 경험분포함수 방법의 변화점 분석을 의미하고 ED_r2는 v_i의 종속성을 고려해 잔차의 일부를 사용하도록 Na 등 (2006)에서 제시된 경험분포함수 방법이다. 표 3의 size를 보면 AR(1) 모형의 ED_y보다는 정도가 많이 약하지만 ED_r1 역시 v_i의 종속성에 영향을 받는다는 것을 알 수 있다. ED_r2의 size는 안정적이지만 자료의 일부를 이용하는 관계로 power가 낮게 나온다. 반면 KD_y, KD_r은 power가 약간 낮아지는 점을 제외하면 AR(1)에서와 유사한 특성을 보인다.

표 3: 측정오차 AR(1) 모형에 대한 변화점 분석의 경험 검정력

n	size				power			
	ED_r1	ED_r2	KD_y	KD_r	ED_r1	ED_r2	KD_y	KD_r
				$\phi = 0.1$				
50	0.039	0.010	0.014	0.008	0.151	0.039	0.157	0.096
100	0.052	0.026	0.011	0.030	0.407	0.137	0.545	0.320
200	0.089	0.028	0.017	0.014	0.805	0.403	0.953	0.704
300	0.079	0.050	0.026	0.029	0.946	0.637	0.998	0.869
500	0.086	0.045	0.022	0.029	0.999	0.945	1	0.980
				$\phi = 0.3$				
50	0.058	0.013	0.014	0.014	0.187	0.039	0.151	0.118
100	0.055	0.028	0.014	0.021	0.461	0.142	0.421	0.433
200	0.066	0.048	0.018	0.013	0.852	0.462	0.782	0.862
300	0.071	0.037	0.031	0.026	0.975	0.769	0.959	0.978
500	0.059	0.058	0.028	0.025	0.999	0.977	0.997	0.999
				$\phi = 0.5$				
50	0.049	0.014	0.025	0.009	0.205	0.046	0.127	0.123
100	0.037	0.020	0.020	0.008	0.492	0.166	0.306	0.477
200	0.057	0.039	0.032	0.018	0.919	0.545	0.583	0.903
300	0.044	0.045	0.036	0.014	0.994	0.796	0.771	0.990
500	0.071	0.051	0.035	0.021	1	0.988	0.937	1
				$\phi = 0.8$				
50	0.042	0.019	0.057	0.006	0.262	0.065	0.173	0.125
100	0.050	0.031	0.071	0.009	0.561	0.199	0.283	0.478
200	0.064	0.047	0.115	0.019	0.939	0.537	0.350	0.916
300	0.045	0.045	0.109	0.016	0.997	0.815	0.412	0.990
500	0.065	0.059	0.112	0.021	1	0.995	0.466	0.999

4. 결론

본 연구에서는 시계열 자료의 분포변화를 검정하는 방법을 정리하였다. 대표적인 비모수적 방법으로 경험분포함수에 기초한 방법과 확률밀도함수추정량을 이용한 방법을 중점적으로 살펴보았다. 두 방법의 확률적 특성을 살펴보고 검정력을 비교하기 위하여 몇 가지 시계열 모형을 가정하고 모의실험을 수행하였다.

먼저 독립 모형에 대하여 정규분포 *i.i.d.* 자료와 정규분포에서 혼합정규분포로 변화하는 독립 자료를 시뮬레이션하여 경험 검정력을 살펴 보았다. 기존 예상과 달리 확률밀도함수추정량을 이용한 방법이 더 높은 검정력을 가지는 것을 보았다. 적절한 커널함수와 밴드폭의 선택이 이루어지면 자료의 확률분포에 따라서는 분포변화 검정에 경험분포함수보다 확률밀도함수추정량이 더 효율적일 수 있음을 알 수 있었다.

자료의 종속성이 변화점 분석에 미치는 영향을 보기 위하여 시계열 모형에서의 분포변화 검정을 모의실험하였다. 특히 AR(1) 모형과 측정오차 AR(1) 모형을 고려하면서 오차항의 분포변화를 시뮬레이션하였다. 종속 시계열에 기초한 경험분포함수 방법은 분포무관 검정을 찾기 어렵다는 사실이 잘 알려져 있는데 모의실험을 통하여 독립 자료에서의 기각역이 성립하지 않음을 확인하였다. 대안으로 잔차에 기초한 방법이 가능한데 모의실험 결과 경험분포함수 방법이 변화점 분석에 적당함을 볼 수 있었다. 반면에 확률밀도함수추정량 방법은 잔차뿐만 아니라 상관관계가 아주 높지 않다면 시계열을 직접 이용하는 것도 가능할 수 있음을 보았다.

종속적인 시계열 자료의 변화점 분석에서 순차적 경험확률과정의 사용은 잔차를 사용할 수 있는 경우로 제한이 필요함을 볼 수 있다. 모의실험 연구를 통하여 최근에 새롭게 제시된 확률밀도함수추정량 방법이 시계열 분포변화 검정에 효율적으로 쓰일 수 있다는 점을 확인하였다. 그런데 확률밀도함수추정량을 이용하기 위해서는 커널함수, 밴드폭, 비교점의 수와 위치 등을 사전에 결정해야 하는 복잡한 문제가 있다. 비교점은 본 연구와 같이 3개의 사분위수가 적절하리라 판단된다. 밴드폭이 분석 결과에 많은 영향을 주는 것은 비모수적 함수 추정에서 잘 알려진 사실인데, 실제로 밴드폭 선택에 따라 분포변화 검정 결과에 많은 변화가 있음을 모의실험으로 확인할 수 있다. 모의실험을 통한 경험으로 보면 변화점 분석에서는 기존의 함수추정에서 연구된 최적값보다는 작은 밴드폭을 사용하는 것이 효율적인 것으로 보인다. 자료의 변화점 분석에 확률밀도함수추정량을 효율적으로 활용하기 위해서는 향후 최적 밴드폭, 비교점 등을 포함한 다양한 연구가 이루어져야 하리라 판단된다.

참고 문헌

나성룡 (2009). 종속 오차에 대한 분포 변화 검정법, <한국통계학회논문집>, **16**, 587–594.

Bai, J. (1994). Weak convergence of the sequential empirical processes of residuals in ARMA models, *Annals of Statistics,* **22**, 2051–2061.

Berkes, I. and Horváth, L. (2003). Limit results for the empirical process of squared residuals in GARCH models, *Stochastic Processes and their Applications*, **105**, 271–298.

Berkes, I., Hörmann, S. and Schauer, J. (2009). Asymptotic results for the empirical process of stationary sequences, *Stochastic Processes and their Applications*, **119**, 1298–1324.

Bickel, P. J. and Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications, *Annals of Mathematical Statistics*, **42**, 1656–1670.

Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd edition, John Wiley & Sons, New York.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes*, 2nd edition, Springer, New York.

Horváth, L., Kokoszka, P. and Teyssière, G. (2001). Empirical process of the squared residuals of an ARCH sequence, *Annals of Statistics*, **29**, 445–469.

Koul, H. (1996). Asymptotics of some estimators and sequential residual empiricals in nonlinear time series, *Annals of Statistics*, **24**, 380–404.

Lee, S. and Na, S. (2004). A nonparametric test for the change of the density function in strong mixing processes, *Statistics and Probability Letters*, **66**, 25–34.

Ling, S. (1998). Weak convergence of the sequential empirical processes of residuals in nonstationary autoregressive models, *Annals of Statistics*, **26**, 741–754.

Na, S., Lee, S. and Park, H. (2006). Sequential empirical process in autoregressive models with measurement errors, *Journal of Statistical Planning and Inference*, **136**, 4204–4216.

Picard, D. (1985). Testing and estimating change-points in time series, *Advances in Applied Probability*, **17**, 841–867.

Shorack, G. and Wellner, J. (1986). *Empirical Processes with Applications in Statistics*, John Wiley & Sons, New York.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions, *Annals of Mathematical Statistics*, **19**, 279–281.

2010년 12월 접수; 2011년 1월 채택

Power Comparison between Methods of Empirical Process and a Kernel Density Estimator for the Test of Distribution Change

Seongryong Na[a], Hyeonah Park[1,b]

[a]Department of Information and Statistics, Yonsei University
[b]Department of Statistics, Seoul National University

Abstract

There are two nonparametric methods that use empirical distribution functions and probability density estimators for the test of the distribution change of data. In this paper we investigate the two methods precisely and summarize the results of previous research. We assume several probability models to make a simulation study of the change point analysis and to examine the finite sample behavior of the two methods. Empirical powers are compared to verify which is better for each model.

Keywords: Test for distribution change, empirical distribution function, probability density estimator, nonparametric test, simulation study.

[1] Corresponding author: Post-Doctorial, Department of Statistics, Seoul National University, Seoul 151-742, Korea.
E-mail: parkha03@yahoo.co.kr