Change-Points with Jump in Nonparametric Regression Functions

Jong Tae Kim[1)]

Abstract

A simple method is proposed to detect the number of change points with jump discontinuities in nonparamteric regression functions. The proposed estimators are based on a local linear regression fit by the comparison of left and right one-side kernel smoother. Also, the proposed methodology is suggested as the test statistic for detecting of change points and the direction of jump discontinuities.

keywords : Nonparametric regression; Jump discontinuities; Local linear regression fit.

1. 서론

본 연구는 국소선형회귀를 사용하여 다중점퍼 혹은 다중불연속에 의한 변화점들의 위치와 크기를 일괄적으로 찾는 추정과 검정방법을 제시하는데 그 목적이 있다. 또한 점퍼에 의한 불연속의 방향이 아래로 향하는지 위로 향하는가에 대한 해답을 제시하여 준다.

국소선형회귀 추정기법과 추정량들을 소개하고 다중점퍼 불연속점들에 의한 변화점들의 위치와 크기를 추정방법을 제시한다. 3절에서는 점퍼 불연속점들에 의하여 생기는 변화점들의 위치와 방향을 검정하는 방법을 소개하고, 모의실험 모형을 이용하여 그 타당성을 제시하였다.

2. 국소선형회귀적합에 의한 점퍼 크기의 추정

비모수 회귀 모형은 다음과 같이 제시된다.

$$Y_i = m(x_i) + \varepsilon_i, \quad i=1, \cdots n. \tag{2.1}$$

여기서 x_i들은 구간 [0,1]에서 정의된 순서화된 균일고정설계점(uniform fixed design point)들이고, Y_i들은 미지의 회귀함수 $m(x_i)$에 대한 반응 변수들이다. 또한, ε_i들은 평균 0과 분산 σ^2을 가지는 독립적이고 동일하게 분포되어진 확률오차들이다. 식(2.1)의 회귀함수 m은 미지의 불연속점들의 개수 q에 대하여 점퍼에 의한 불연속점들이 $t_j \in (0,1),\ j=1,\cdots,q$ 에서 존재한다고 가정하자. 그리고 $\triangle(t_j)$를 주어진 x에서의 점퍼의 크기라 하자. 그러면 t_j에서의 점퍼 불연속의 크기는 $\triangle(t_j)$이다. 위의 정의에 따라서 회귀함수 m은 스무드함수 $g \in C^2([0,1])$를 이용하여 다음과 같이 정의되어 진다.

---

1) Associate Professor, Dept of Statistics, Daegu University, 712-714, Korea.
E-mail: jtkim@daegu.ac.kr

$$m(x) = g(x) + \sum_{j=1}^{q} \triangle(t_j)\, 1_{[x \geq t_j]}\,. \tag{2.2}$$

임의의 점 x를 중심으로 오른쪽과 왼쪽의 함수극한값들을 각각 $m_+(x) = \lim_{y \downarrow x} m(y)$와 $m_-(x) = \lim_{y \uparrow x} m(y)$으로 정의하자. 그러면, 점퍼에 의한 불연속 점들 $t_j,\ j=1, \cdots, q,$ 에서 각 변화점들의 크기 $\triangle(t_j)$는 다음과 같다.

$$\triangle(t_j) = m_+(t_j) - m_-(t_j). \tag{2.3}$$

식(2.3)에서 점 t_0에서 $\triangle(t_0) = 0$ 이면 t_0에서는 점퍼에 의한 변화점이 발생하지 않는다고 간주할 수 있다. 어떤 불연속인 점 t_j는 고정설계점들 $x_i,\ i=1, \cdots, n$,에 대하여 $x_i \leq t_j < x_{i+1}$구간에 존재한다. 그러나 Loader(1996)의 언급처럼 불연속인 점을 그 구간에서 구별하기 어렵기 때문에 그 불연속인 변화점들은 고정설계점 상에서 발생한다고 가정하자.

임의의 설계점 x_k와 주어진 양의 정수 p에 대하여 다음 식(2.4)를 최소로 하는 커널가중 국소최소제곱(kernel weighted local least squares)법으로 구해진 회귀계수 $\hat{\beta}' = (\hat{\beta}_0, \cdots, \hat{\beta}_p)$를 이용하여 식(2.1)의 회귀함수 $m(x_k)$의 오른쪽 추정량을 $\hat{m}_+(x_k) \equiv \hat{\beta}_0$으로 제시할 수 있다.

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{l=0}^{p} \beta_l (x_i - x_k) \right\}^2 K\left(\frac{x_i - x_k}{h}\right). \tag{2.4}$$

여기서 K는 토대(support)가 [0,1]인 커널함수이며 한쪽방향 커널(one-sided kernel)이라 불리어진다. 그리고 상수 h는 평활모수인 띠폭이다. 이러한 한쪽방향의 커널은 위의 식에서 x_k의 오른쪽에 있는 자료를 이용하여 회귀계수를 구하게 된다. 이러한 (2.4)식의 커널가중 국소최소제곱법에서 $K((x_i - x_k)/h)$ 대신 $K((x_k - x_i)/h)$를 사용하여 최소로 한 회귀계수 $\hat{\beta}' = (\hat{\beta}_0, \cdots, \hat{\beta}_p)$를 이용하여 회귀함수 $m(x_k)$의 왼쪽 추정량을 $\hat{m}_-(x_k) \equiv \hat{\beta}_0$로 제시할 수 있다. 위의 추정량들의 정의에서 $p=1$인 국소선형회귀적합(local linear regression fit)인 경우에, $m(x_k)$의 오른쪽과 왼쪽의 추정량은 각각 다음과 같이 표현된다.

$$\hat{m}_+(x_k) = \sum_{i=1}^{n} w_i^+(x_k)\, Y_i \,/\, \sum_{i=1}^{n} w_i^+(x_k),\quad \hat{m}_-(x_k) = \sum_{i=1}^{n} w_i^-(x_k)\, Y_i \,/\, \sum_{i=1}^{n} w_i^-(x_k) \tag{2.5}$$

여기서

$$\begin{aligned} w_i^+(x_k) &= K\left(\frac{x_i - x_k}{h}\right)\{S_2^+(x_k) - (x_i - x_k) S_1^+(x_k)\}, \\ w_i^-(x_k) &= K\left(\frac{x_k - x_k}{h}\right)\{S_2^-(x_k) - (x_i - x_k) S_1^-(x_k)\}, \\ S_j^{\pm}(x_k) &= \sum_{i=1}^{n} K\left(\pm\frac{x_i - x_k}{h}\right)(x_i - x_k)^j, \quad j = 1, 2 \end{aligned}$$

이다. 설계점들의 토대 $[0, 1]$에서 0과 1은 또 다른 형태의 불연속점으로 이해될 수 있고 이러한 이유로 많은 연구 논문(Loader (1996), Jose와 Ismail (1999))에서 불연속점들이 $[h, 1-h]$에 존재한다고 가정을 하고 있다. 위의 추정량들을 이용하여 이 구간 내에 있는

설계점들에서의 점퍼 크기 $\triangle(x_k)$에 대한 추정량 $\widehat{\triangle}(x_k)$는 다음같이 정의할 수 있다.

$$\widehat{\triangle}(x_k) = \widehat{m}_+(x_k) - \widehat{m}_-(x_k), \quad h \le x_k \le 1-h.$$

다음의 Lemma는 다음절에 제시되는 점퍼 불연속에 대한 검정통계량을 구하기 위한 목적으로 x_k에서 점퍼 크기의 추정통계량 $\widehat{\triangle}(x_k)$에 극한수렴분포를 보여주고 있다.

LEMMA2.1 식 (2.1)-(2.3)하에서 다음의 네 조건 (A.1)-(A.4)를 가정하자.

(A.1) 한쪽 방향 커널함수 K는 $\int_0^1 K(u)du = 1$ 그리고 $K(0) > 0$을 만족한다.

(A.2) 띠폭 h는 $n \to \infty$ 일 때 $h \to 0$, $nh \to \infty$ 와 $nh^3 \to 0$ 을 만족한다

(A.3) 임의의 두 불연속점간의 거리는 $2h$ 보다 크다.

(A.4) 모든 $i = 1, \cdots, n$에 대해 $E(|\varepsilon_i|^{2+s}) < \infty$를 만족하는 어떤 양의 실수 s가 존재를 만족하면 점 x_k 에 대하여 식(2.6)에서 정의된 $\widehat{\triangle}(x_k)$의 극한수렴분포는 다음과 같다.

$$\sqrt{nh}\left(\widehat{\triangle}(x_k) - \triangle(x_k)\right) \xrightarrow{d} N\left(0,\ 2\sigma^2 \int_0^1 \{K(u)\}^2 du\right).$$

증명은 Huh와 Carrière(2002)의 Corollary 1의 증명과정과 유사함으로 생략하였다. 조건 (A.3)는 인접해 있는 설계점들에서 각각의 점퍼 크기의 추정치들이 서로서로 영향을 주기 때문에 하나의 불연속점 주변의 설계점들에서도 불연속점으로 판정될 가능성이 있다. 따라서, Jose와 Ismail(1999)이 언급하였던 것처럼 근접해 있는 두 개의 변화점들 사이의 거리들은 $2h$ 보다 크다는 가정이 필요하게 된다. 조건 (A.4)는 중심극한정리를 보일 때 필요한 것이다.

3. 다중 점퍼 불연속의 존재에 대한 검정

2절에서 제안된 점퍼 크기의 추정량을 이용하여 다중 불연속인 변화점들의 위치추정과 미지의 점퍼 불연속인 변화점들의 개수를 가설검정을 이용하여 추정하여 보자. 설명의 편의를 위하여 점퍼의 크기 $\triangle(t_j)$, $j = 1, \cdots, q$, 들이 다음의 조건을 만족한다고 가정하자.

$$\triangle(t_1) \ge \triangle(t_2) \ge \cdots \ge \triangle(t_q).$$

직관적인 관점에서, 점퍼 크기의 추정값 $\widehat{\triangle}(x_k)$ 의 절대값이 가장 크다면, 설계점 x_k에서 가장 큰 점퍼 불연속점을 가질 것이라고 추정할 수 있다. 즉, 집합 $I_1 = \{x_k : h \le x_k \le 1-h\}$에서 가장 큰 점퍼 불연속인 변화점 위치 t_1의 추정량 l_1은 다음과 같이 정의할 수 있다.

$$l_1 = \arg\max_{I_1} |\widehat{\triangle}(x_k)|.$$

그리고, 두 번째로 큰 점퍼 불연속점 위치 t_2의 추정량 l_2는

$$l_2 = \arg\max_{I_2} |\widehat{\triangle}(x_k)|$$

으로 정의할 수 있다. 여기서 $I_2 = \{x_k : h \le x_k \le 1-h,\ |x_k - l_1| \ge 2h\}$이다. 이 집합은 Lemma2.1의 조건 (A.3)를 기초로 하여 인접한 두 불연속점의 거리를 최소 $2h$로 만들기

위한 것이다. 이와 같은 방법으로 계속해서 점퍼 불연속점 위치 t_3, t_4 등의 추정량들을 순차적으로 다음과 같이 제시 할 수 있다.

$$l_3 = \arg\max_{I_3} |\widehat{\triangle}(x_k)|,$$

$$l_4 = \arg\max_{I_4} |\widehat{\triangle}(x_k)|,$$

$$\vdots .$$

여기서 임의의 $j=1,2,\cdots$에 대해 $I_j = \{x_k : h \le x_k \le 1-h,\ |x_k - l_i| \ge 2h,\ i=1,\cdots,j-1\}$ 이다. 이 추정량 $l_j,\ j=1,2,\cdots$들을 기초로 하여 미지의 점퍼 불연속점의 개수 q를 다음의 통계적 가설검정의 결과에 의해 추정할 수 있다. 먼저, 어떤 불연속점 t_j에서 점퍼 불연속의 존재여부에 대한 가설검정을 하여 보자. 다음의 가설

$$H_0 : \triangle(t_j) = 0, \quad H_1 : \triangle(t_j) \neq 0$$

에 대해 t_j의 추정량 l_j을 이용하여 표준정규분포로 극한 수렴하는 아래의 검정통계량을 제시하고자 한다.

$$KC_h(l_j) = \frac{\widehat{\triangle}(l_j)}{\sqrt{\mathrm{Var}(\widehat{\triangle}(l_j))}} \xrightarrow{d} N(0,1). \tag{3.1}$$

위의 검정통계량은 H_0 하에서 표준정규분포로 수렴함을 쉽게 보일 수 있다. 식(3.1)에서 분산은 여전히 미지이므로 $\mathrm{Var}(\widehat{\triangle}(l_j))$에 대신 그 추정량

$$\widehat{\mathrm{Var}}(\widehat{\triangle}(l_j)) = \widehat{\mathrm{Var}}(\widehat{\triangle}_+(l_j)) + \widehat{\mathrm{Var}}(\widehat{\triangle}_-(l_j))$$

$$= \hat{\sigma}^2 \left[\sum_{i=1}^{n} (w_i^+(l_j))^2 / \left(\sum_{i=1}^{n} w_i^+(l_j) \right)^2 + \sum_{i=1}^{n} (w_i^-(l_j))^2 / \left(\sum_{i=1}^{n} w_i^-(l_j) \right)^2 \right]$$

를 사용하여 불연속점의 위치를 찾는 통계량 $\widehat{KC}_h(l_j)$를 다음과 같이 제시한다.

$$\widehat{KC}_h(l_j) = \frac{\widehat{\triangle}(l_j)}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\triangle}(l_j))}}. \tag{3.2}$$

위에 식에 있는 $\hat{\sigma}^2$는 Gasser, Sroka와 Jennen-Steinmetz(1986)의 분산 추정량 (GSJS 분산 추정량)을 사용하였다. GSJS의 분산 추정량은 $(d_{-1}, d_0, d_1) = (-6^{-1/2}, (2/3)^{1/2}, -6^{-1/2})$일 때 다음과 같이 정의된다.

$$\hat{\sigma}^2 = (n-2)^{-1} \sum_{k=2}^{n-1} \left(\sum_{j=-1}^{1} d_j Y_{j+k} \right)^2 . \tag{3.3}$$

Kim(1998)의 연구에 의하면 GSJS 분산 추정량은 Rice(1984)의 추정량이나 Hall et al.(1990)의 추정량보다 평균제곱오차의 값에서 우수한 성질을 가지며 $\hat{\sigma}^2$은 σ^2에 확률적으로 수렴한다. 그러므로 H_0하에서 통계량 $\widehat{KC}_h(l_j)$는 다음과 같은 분포를 따른다.

$$\widehat{KC}_h(l_j) \xrightarrow{d} N(0,1) . \tag{3.4}$$

그러므로 실험점 t_j에서 점퍼 불연속에 대한 존재여부의 기각역은 유의수준 α에서 다음과 같이 주어진다.

$$|\widehat{KC}_h(l_j)| > \Phi^{-1}(1-\alpha/2). \tag{3.5}$$

여기서 Φ는 표준정규분포함수이다. 한편, t_j에서 점퍼크기 $\triangle(t_j)$에 대한 $100(1-\alpha)\%$신

뢰구간은 다음과 같이 정의된다.

$$\widehat{\triangle}(l_j) \pm \Phi^{-1}(1-\alpha/2)\left(\widehat{\mathrm{Var}}\left(\widehat{\triangle}(l_j)\right)\right)^{1/2}. \tag{3.6}$$

식 (3.5)의 기각역을 이용하여 점퍼 불연속점으로 추정된 l_j, $j=1,2,\cdots$ 에서 불연속점의 존재여부에 대한 가설검정을 할 수 있다 . 이 때 주어지는 유의수준 α 에서 기각역은

$$\left|\widehat{\mathrm{KC}}_h(l_j)\right| > \Phi^{-1}(1-\alpha/2), \quad j=1,2,\cdots,$$

이 된다. j에 대해 순차적으로 ($\widehat{\triangle}(l_j)$의 크기순으로) 가설검정을 해 나가면서 귀무가설이 기각되지 않은 단계가 $j=r$이라면 미지의 점퍼 불연속점의 개수 q의 추정량을 $\hat{q}=r-1$로 정의할 수 있다.

다중 불연속점의 개수 추정에 대한 타당성을 조사하기 위해, $0.25, 0.5, 0.75$의 3개의 점에서 변화점을 가지도록 설계하였고, 특히 변화점 0.5에서는 0.25와 0.75에서 보다 점퍼의 크기를 더 크게 주었으며, 변화점 0.75는 변화점 0.25보다 약간 더 큰 점퍼의 크기를 주었다. 데이터의 개수는 400으로, 오차항의 표준편차의 값은 0.1로 두고, 다음의 모형을 조사하였다.

모형1:

$$y_1 = \sin(20x_i)1_{(x_{1\geq 5})} + 1(.25 \leq x_i \leq .5) - .5*(.5 \leq x_i \leq .75) - 1.5*1(x_i \geq .75) + \varepsilon_i,$$

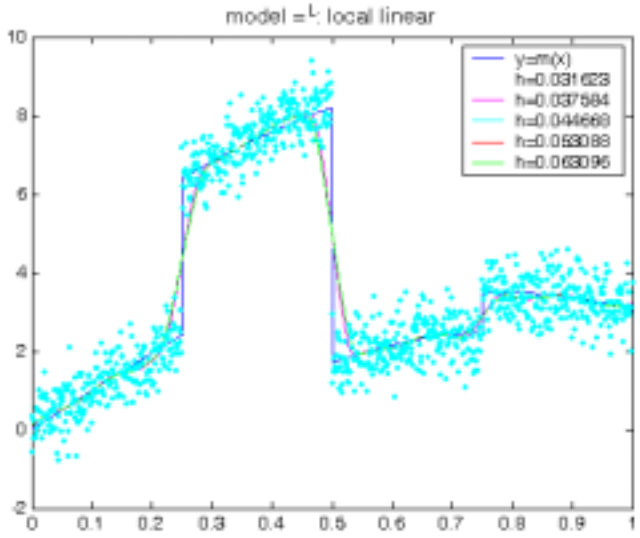

그림 3.1: 모형1의 함수와 띠폭들에 따른 국소선형적합

그림3.1은 위의 모형1에서의 본래의 데이터의 분포에 대하여, 띠폭 $h = 0.01, 0.0133, 0.0178, 0.0237, 0.0316$를 가지고, Epanechnikov커널을 사용한 커널함수 추정 그림이다. 식(3.4)에서 띠폭은 cross-validation의 추정된 h값을 사용하였다. Chaudhri와 Marron(1999)은 한 개의 적합한 h를 가지고서는 데이터가 실제로 보유하고 있는 변화점들의 정보를 모두 제공해 주지 않는다고 주장한다. Chaudhri와 Marron의 주장에 따라서 식(3.4)에서의 임의의 설계된 다양한 띠폭들을 적용시켜 보았다. 띠폭 h의 크기가 너무 작은 관계로 그림 3.1이나 그림 3.2에 그 차이점을 분간하기 어렵다. 그러나 앞서 설명한 것 같이 인접한 두 불연속점의 거리를 최소 $2h$로 두는 제약조건으로 인해 h의 크기가 두드러지게

커질 경우 변화점을 분석하는데 상당한 문제점을 가진다.

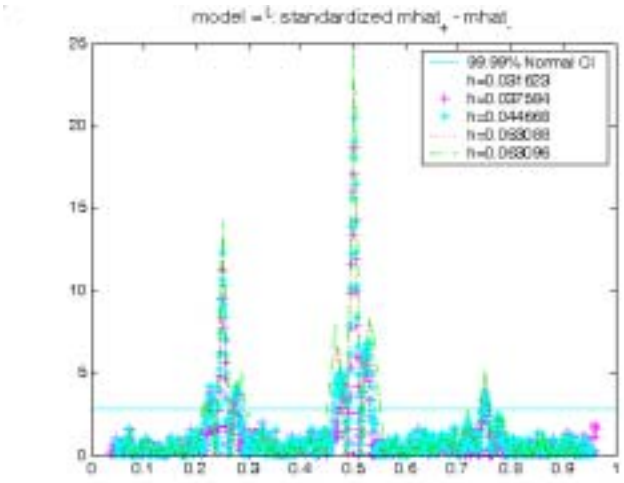

그림 3.2 모형1에서의 변화점 추정

그림3.2는 식(3.8)의 점퍼 불연속들의 변화점들에 대한 추정량 $|\widehat{KC}_h(t)|$가 $\Phi^{-1}(0.99)=2.33$보다 큰 값을 가지는 $t=0.25,0.5,0.75$ 위치의 국소적인 부분에서 발생하고 있음을 나타낸다. 또한 이들의 국소적인 부분들의 최대값들은 $t=0.25,0.5,0.75$에서 각각 위치하여 있다. 모형 1에서 설계한 것과 같이 $t=0.5$ 의 값에서 가장 큰 점퍼 불연속에 의한 변화점이 발생하고, 점퍼의 크기에 따라 0.75와 0.25에서 순차적으로 변화점이 발생함을 알 수 있다. 그러므로 식(3.8)의 추정량 $|\widehat{KC}_h(t)|$는 불연속점들을 가지는 다중변화점들의 위치 추정에 타당한 성질을 제공함을 알 수 있다.

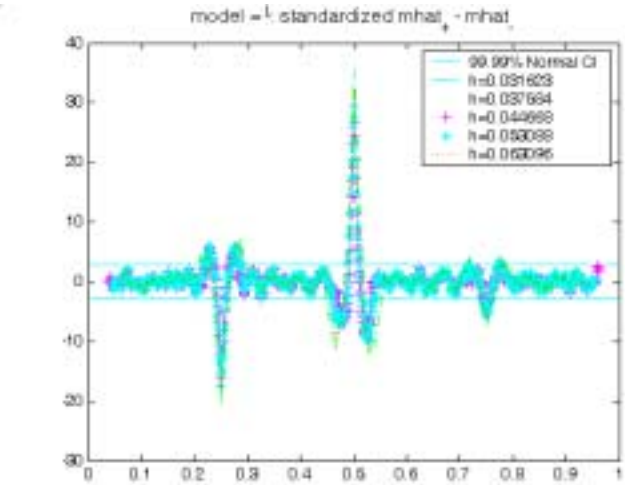

그림3.3는 식(3.8)의 점퍼 불연속들의 변화점들에 대한 추정량 $\widehat{KC}_h(\ )$가 $\Phi^{-1}(0.99)=2.33$ 보다 큰 값을 가지는 $t=0.25,0.5,0.75$ 위치의 국소적인 부분에서 발생하고 있음을 나타낸다. 또한 이들의 국소적인 부분들의 최대값들은 $t=0.25,0.5,0.75$

에서 각각 위치하여 있다. 모형 1에서 설계한 것과 같이 $t = 0.5$ 의 값에서 가장 큰 점퍼 불연속에 의한 변화점이 발생하고, 점퍼의 크기에 따라 0.75와 0.25에서 순차적으로 변화점이 발생함을 알 수 있다. 또한 $t = 0.25$ 와 $t = 0.75$ 에서 점퍼의 방향이 위로 향할 때는 검정통계량 $\widehat{KC}_h(t)$의 방향은 아래로 향하고, $t = 0.5$ 에서 점퍼의 방향이 아래로 향할 때는 검정통계량 $\widehat{KC}_h(t)$의 방향은 위로 향한다.

4. 결론

점퍼 불연속에 의한 다중 변화점들을 포함하고 있는 $\widehat{KC}_h$의 대립가설에서의 띠폭 h을 결정하는데 있어서 기존의 연구문헌들을 찾기가 힘들었다. cross-validation에 의해 구해진 띠폭을 이용하여 그 띠폭을 중심으로 하여 조금씩 변화를 주어 모의실험을 해보았다. 여기서 큰 문제점은 발견하지 못하였지만 변화점이 존재할 때, 띠폭 h를 추정하는 문제는 연구과제로 남아있다.

참고문헌

[1] Chaudhri, P. and Marron, J. S. (1999). SiZer for Exploration of Structures in Curve, Journal of American Statistical Association}, 94, 807-823.

[2] Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression, Biometrika}, 73, 625-633.

[3] Hall, P., Kay, J. W. and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. Biometrika 77, 521-528.

[4] Huh, J. and Carrière, K. C. (2002). Estimation of regression functions with a discontinuity in a derivative with local polynomial fits. Statistics and Probability Letters 56, 329-343.

[5] Jose, C. T. and Ismail, B. (1999). Change points in nonparametric regression functions, Communication in Statistics - Theory and Method, 28, 1883-1902.

[6] Kim, J. T. (1998). 비모수회귀모형의 차분에 기저한 분산추정에 대한고찰, 한국통계학회논문집, 5, 121-131.

[7] Loader, C. (1996). Change point Estimation using nonparametric regression, Annals of Statistics, 24, 1667-1678.

[8] Rice, J. (1984). Bandwidth choice for nonparametric kernel regression. Annals of Statististics, 12, 1215-1230.