Journal of the Korean Data & Information Science Society
2016, **27**(1), 111–120

http://dx.doi.org/10.7465/jkdi.2016.27.1.111
한국데이터정보과학회지

잔차 수정을 이용한 불연속 분산함수의 비모수적 추정[†]

허집[1]

[1]덕성여자대학교 정보통계학과

접수 2015년 12월 15일, 수정 2016년 1월 4일, 게재확정 2016년 1월 6일

요 약

대부분의 불연속 회귀함수의 커널추정량은 알고 있거나 추정된 불연속점을 기준으로 자료를 분리하여 각각을 독립적으로 회귀함수를 적합하고 있다. 회귀모형에서 분산함수가 불연속점을 가지고 있을 때에도 잔차제곱들을 이용하여 위와 같은 불연속 회귀함수의 커널추정법을 활용하고 있다. Kang 등 (2000)은 Müller (1992)의 불연속점과 점프크기 커널추정량을 이용하여 반응변수의 표본을 연속인 회귀함수로부터 표본인 것처럼 수정하여 불연속 회귀함수를 추정하였다. 본 연구에서는 불연속 분산함수를 추정하기 위하여 Kang 등 (2000)의 방법을 이용한다. Kang과 Huh (2006)의 분산함수의 불연속점과 점프크기 추정량으로 잔차제곱들을 수정하고, 수정된 잔차제곱들을 이용하여 불연속 분산함수 커널추정량을 제안할 것이다. 제안된 추정량의 적분제곱오차의 수렴속도를 보여주고 모의실험을 통하여 기존의 추정량과 제안된 추정량을 비교하고자 한다.

주요용어: 분산함수, 불연속점, 잔차, 점프크기, 커널추정량.

1. 서론

분산함수는 회귀함수의 신뢰구간 혹은 가설검정 등의 추론에 이용되어 회귀함수의 추정 정도 (precision)에 영향을 주는 함수이다. 또한 커널함수를 이용한 비모수적 함수 추정의 띠폭 (bandwidth)의 선택에서도 추정되어져야 하는 함수이다. 그 외의 통계적 방법론이 필요한 영역에서도 분산 혹은 분산함수는 평균이나 회귀함수와 더불어 추론에 중요한 역할을 하고 있다. Rice (1984), Gasser 등 (1986), Müller와 Stadtmüller (1987), Hall과 Carroll (1989), Hall 등 (1990), Ruppert 등 (1997)과 Yu와 Jones (2004) 등이 회귀모형에서 분산함수의 비모수적 추정을 연구하였다. 이들의 연구는 분산함수가 연속인 경우에 이루어진 것이다.

회귀함수 등, 함수의 커널추정량의 편의 (bias)의 수렴속도 (rate of convergence)는 일반적으로 그 함수들의 연속의 정도와 관련이 있다는 것은 익히 알려져 있다. 흔히 쓰이고 있는 국소 p차다항적합 (local pth polynomial fit)을 이용한 회귀함수의 커널추정량은 그 함수가 최소 $p+2$미분된 도함수가 연속인 경우에 원하는 커널추정량의 편의를 얻을 수 있다. Müller (1992)은 불연속점을 고려하지 않은 불연속 회귀함수의 커널추정량은 일치추정량 (consistent estimator)이 되지 않으며, 불연속인 분산함수도 불연속점이 고려되지 않은 분산함수의 커널추정량은 회귀함수의 커널추정량처럼 일치추정량이 되지 않는다. 그 외의 통계적 함수의 불연속점 혹은 변화점 추정 연구로는 위험률함수의 연구로 Lee 등 (2015)이 있으며, 시계열모형의 연구로는 Sohn과 Cho (2015)가 있다.

[†] 본 연구는 덕성여자대학교 2014년도 교내연구비 지원에 의해 수행되었음.
[1] (01369) 서울시 도봉구 삼양로 144길 33, 덕성여자대학교 정보통계학과, 교수.
E-mail: jhuh@duksung.ac.kr

분산함수가 불연속점을 가질 때, Kang과 Huh (2006)는 연속인 회귀함수의 커널추정량으로 만들어진 잔차제곱들을 이용한 Nadaraya-Watson 추정량으로 불연속점과 점프크기 및 분산함수의 추정을 연구하였다. Huh (2005)는 회귀함수가 연속인 경우에 분산함수의 불연속점의 추정을 이차적률함수의 불연속점의 추정으로 제안하였다. 한편, Yu와 Jones (2004)는 음이 아닌 분산함수를 로그 변환하여 국소선형적합 (local linear fit)에 의한 로그분산함수 (log-variance function)의 추정으로 분산함수의 추정을 연구하였다. Huh (2009, 2015)는 Yu와 Jones의 연구를 이용하여 로그분산함수의 불연속점 추정으로 분산함수의 불연속점 추정에 대한 연구를 하였다. 음이 아닌 값을 가지는 분산함수의 제약 조건을 없애기 위해 Chen 등 (2009)은 잔차제곱들의 로그변환을 이용한 국소선형적합으로 로그분산함수의 커널추정량을 제안하였다. Huh (2014)는 Chen 등이 제안한 방법인 로그 잔차제곱들을 이용하여 불연속인 분산함수의 커널추정량을 제안하여 Huh (2015)의 방법과 비교 연구하였다.

Kang 등 (2000)은 회귀함수가 불연속인 경우에 Gasser-Müller 추정량으로 추정된 불연속점의 점프크기 추정량으로 반응변수의 표본을 연속인 회귀함수로부터 얻어진 표본인 것처럼 수정하여, 회귀함수의 커널추정량을 제시하고 다시 역으로 추정된 점프크기로 재수정한 불연속 회귀함수의 추정법을 제시하였다. 본 연구에서는 Kang 등 (2000)의 방법을 이용하여 불연속인 분산함수를 연속인 것처럼 수정하여 불연속 분산함수의 커널추정량을 제시하고자 한다.

2절에서는 분산함수의 불연속점 추정량과 점프크기 추정량을 이용한 잔차제곱의 수정 방법과 이를 이용한 불연속 분산함수의 추정법을 소개하고, 3절에서는 제안된 추정량의 적분제곱오차 (integrated squared error)의 수렴속도를 보여준다. 4절에서는 모의실험을 통하여 기존의 불연속 분산함수의 추정량과 비교연구하고자 한다.

2. 잔차 수정과 불연속 분산함수의 추정

이변량 표본 $\{(X_i, Y_i),\ i = 1, 2, \cdots, n\}$은 확률벡터 (X, Y)로부터 독립적으로 추출되었다고 하자. 회귀함수와 분산함수를 각각 $m(x) = E(Y|X = x)$와 $v(x) = Var(Y|X = x)$라 두면, 회귀모형은 다음과 같이

$$Y_i = m(X_i) + v^{1/2}(X_i)\varepsilon_i,\ i = 1, 2, \cdots, n \tag{2.1}$$

제시된다. 오차항 ε_i는 독립변수 $X_1, \cdots, X_n$과 독립이며 평균과 분산은 각각 0과 1이다. 공변량 X의 확률밀도함수는 $f(x)$이고 토대 (support)는 $[0, 1]$이라 하자. 어떤 함수 g에 대하여 임의의 점 x로 접근할 때 좌극한값과 우극한값을 각각 $g_-(x) = \lim_{y\to x-} g(y)$, $g_+(x) = \lim_{y\to x+} g(y)$라 두면, 분산함수가 한 점 τ에서 불연속인 경우 점프크기는

$$\Delta = v_+(\tau) - v_-(\tau) \tag{2.2}$$

가 된다. 분산함수가 불연속점을 가진다면 $|\Delta| > 0$이고, 그렇지 않다면 $\Delta = 0$이다. 분산함수가 불연속인 이유는 회귀함수가 불연속인 경우와 회귀함수는 연속이고 분산함수 자체가 불연속인 경우가 있다. 전자의 경우는 회귀함수의 불연속점의 추정으로 분산함수의 불연속점을 추정할 수 있다. 본 연구에서는 후자의 경우를 고려하고자 한다.

Kang과 Huh (2006)는 Nadaraya-Watson 추정량으로 분산함수의 불연속점과 그 점에서 점프크기를 추정하기 위하여, 먼저 회귀함수를 Nadaraya-Watson 추정량으로 추정한 후 잔차제곱들을 이용하여 토대가 $[0, 1]$인 한쪽방향커널함수 (one-sided kernel function)로 어떤 점 x에서 $v(x)$의 오른쪽 추정량

$\widehat{v}_+(x)$와 왼쪽 추정량 $\widehat{v}_-(x)$을 각각 다음과 같이 정의하였다.

$$\widehat{v}_\pm(x) = \frac{\sum_{i=1}^{n} K\left(\pm\frac{X_i - x}{h_1}\right)\widehat{R}_i^2}{\sum_{i=1}^{n} K\left(\pm\frac{X_i - x}{h_1}\right)}. \tag{2.3}$$

여기서 $\widehat{R}_i = Y_i - \widehat{m}(X_i)$는 회귀함수의 Nadaraya-Watson 추정량 $\widehat{m}$에 의한 잔차이고 h_1는 띠폭이며 K는 토대가 $[0,1]$인 한쪽방향커널함수이다. 추정량들 (2.3)을 이용하여 점 x의 점프크기 추정량을 $\widehat{\Delta}(x) = \widehat{v}_+(x) - \widehat{v}_-(x)$라 정의하여 $|\widehat{\Delta}(x)|$가 최대가 되는 점 x를 불연속점 추정량 $\widehat{\tau}$으로 제안하였고 추정된 불연속점 $\widehat{\tau}$에서 점프크기 $\widehat{\Delta}(\widehat{\tau})$를 식 (2.2)의 점프크기 Δ의 추정량으로 제안하였다.

또한 Kang과 Huh는 $\widehat{\tau}$과 잔차제곱들인 $\widehat{R}_i^2$들을 이용하여 Nadaraya-Watson 커널추정량으로 다음과 같이 불연속 분산함수를 추정하였다.

$$\widehat{v}_{KH}(x;\widehat{\tau}) = \frac{\sum_{i=1}^{n} L_h^*(X_i - x;\widehat{\tau})\widehat{R}_i^2}{\sum_{i=1}^{n} L_h^*(X_i - x;\widehat{\tau})}. \tag{2.4}$$

여기서 h는 식 (2.3)의 띠폭 h_1과는 다른 띠폭이며 $L_h^*(u-x;t)$는 다음과 같다.

$$L_h^*(u-x;t) = \begin{cases} \dfrac{1}{h}L\left(\dfrac{u-x}{h}\right)\mathrm{I}[x-h \le u < t], & t-h \le x < t \\ \dfrac{1}{h}L\left(\dfrac{u-x}{h}\right)\mathrm{I}[t \le u < x+h], & t \le x < t+h \\ \dfrac{1}{h}L\left(\dfrac{u-x}{h}\right), & \text{그 외.} \end{cases} \tag{2.5}$$

함수 L은 토대 $[-1,1]$를 가지는 커널함수이고, I는 표시함수 (indicator function)이다. 식 (2.5)의 커널함수 L_h^*는 불연속 점추정량 $\widehat{\tau}$를 기준으로 왼쪽 잔차제곱들과 오른쪽 잔차제곱들을 분리하는 기능을 하고 있다. 따라서 τ의 왼쪽 분산함수와 오른쪽 분산함수는 각각 $\widehat{\tau}$의 왼쪽 잔차제곱들과 오른쪽 잔차제곱들을 이용하여 Nadaraya-Watson 추정량으로 추정된다.

잘 알려져 있듯이 Nadaraya-Watson 추정량은 경계점에서 편의의 수렴속도 때문에 이론적으로 추정의 정도가 떨어진다. 추정량 (2.4)는 불연속점 추정량을 기준으로 잔차제곱들을 두 부분으로 나누어 각각 Nadaraya-Watson 추정량으로 분산함수를 추정하기 때문에 경계점 문제 (boundary problem)가 불연속점 추정량 주변에서도 만들어지게 된다. 불연속점 추정량 주변에서 일어나는 경계점 문제를 극복하기 위하여, 본 연구에서는 회귀함수가 불연속인 경우에 Gasser-Müller 추정량으로 추정된 불연속점에서 점프크기 추정량으로 반응변수의 표본을 연속인 회귀함수로부터 얻어진 표본인 것처럼 수정하여 회귀함수의 커널추정량을 제시하고 다시 역으로 추정된 점프크기로 재수정한 불연속 회귀함수의 추정법을 제시한 Kang 등 (2000)의 방법을 이용하여 불연속 분산함수의 커널추정량을 제시하고자 한다.

먼저, 식 (2.1)의 불연속 분산함수가 다음을 만족한다고 가정하자.

$$v(x) = w(x) + \Delta \times \mathrm{I}[\tau \le x \le 1]. \tag{2.6}$$

이때 함수 $w(x)$의 이계도함수는 연속으로 다음을 $w \in C^2([0,1])$만족한다. Kang과 Huh에 의해 추정된 $\widehat{m}$과 $\widehat{\tau}$을 이용하여 다음과 같은 수정된 잔차제곱들을 생각하자.

$$\widetilde{R}_i^2 = \widehat{R}_i^2 - \widehat{\Delta}(\widehat{\tau}) \times \mathrm{I}[\widehat{\tau} \le X_i \le 1],\ i = 1, \cdots, n. \tag{2.7}$$

위의 수정된 잔차제곱들을 이용한 Nadaraya-Watson 추정량으로 식 (2.6)의 연속인 $w(x)$의 추정량을 다음과 같이 제안한다.

$$\widehat{w}(x;\widehat{\tau}) = \frac{\sum_{i=1}^{n} L_h(X_i - x)\widetilde{R}_i^2}{\sum_{i=1}^{n} L_h(X_i - x)}. \tag{2.8}$$

여기서 $L_h(x) = h^{-1}L(x/h)$이다. 위 추정량은 표본을 두 부분으로 분리하지 않고 불연속점 추정량 부근에서 양쪽의 표본을 모두 이용하게 되는 장점을 가지게 된다. 불연속 분산함수 $v(x)$과 $w(x)$의 관계 (2.6)을 이용하여 다음과 같이 최종적으로 불연속 분산함수 $v(x)$의 커널추정량을 제안한다.

$$\widehat{v}_p(x;\widehat{\tau}) = \widehat{w}(x;\widehat{\tau}) + \widehat{\Delta}(\widehat{\tau}) \times \mathrm{I}[\widehat{\tau} \le x \le 1]. \tag{2.9}$$

3. 추정량의 점근 성질

이 절에서는 2절에서 제안한 식 (2.9)의 불연속 분산함수 추정량 $\widehat{v}_p(x;\widehat{\tau})$의 적분제곱오차의 수렴속도를 보여주고자 한다. Kang과 Huh (2006)와 Huh (2009)는 제안한 잔차제곱들의 수정 (2.7)에 쓰인 $\widehat{\tau}$과 $\widehat{\Delta}(\widehat{\tau})$의 점근성질을 다음과 같이 됨을 보였다.

$$|\widehat{\tau} - \tau| = O_P(n^{-1}),\ |\widehat{\Delta}(\widehat{\tau}) - \Delta| = O_P((nh_1)^{-1/2}). \tag{3.1}$$

여기서 $A_n = O_P(B_n)$은 임의의 양수 ε과 $n > N$에 대하여 $P(|A_n/B_n| > M) < \varepsilon$을 만족하는 상수 M과 N이 존재한다는 의미이다.

위의 결과들을 만족시키는 조건들은 다음과 같다.

(A.1) X의 확률밀도함수 f는 $\inf_{x\in[0,1]} f(x) > 0$을 만족하고 Lipschitz 1차조건을 만족한다.
(A.2) 확률밀도함수인 한쪽방향커널함수 K는 Lipschitz 1차조건을 만족하고 $K(0) > 0$이고 $0 < u \le 1$에 대하여 $K(u) \ge 0$이다.
(A.3) $n \to \infty$일 때 $h_1 \to 0$, $h_1/\log n \to \infty$이고 $nh_1^3 \to 0$을 만족한다.

제안한 분산함수 추정량 $\widehat{v}_p(x;\widehat{\tau})$의 적분제곱오차의 수렴속도를 보여주기 위한 조건들과 결과는 다음과 같다.

(C.1) 식 (2.6)의 함수 w의 이계도함수 w''은 Lipschitz 1차조건을 만족한다.
(C.2) 확률밀도함수인 커널함수 L은 Lipschitz 1차조건을 만족한다.
(C.3) $n \to \infty$일 때 $h \to 0$, $nh/\log n \to \infty$을 만족한다.

정리. 조건 (A.1)~(A.3)와 (C.1)~(C.3)를 만족하면 분산함수 추정량 $\widehat{v}_p(x;\widehat{\tau})$의 적분제곱오차의 수렴속도는 다음과 같다.

$$\begin{aligned}&\int_0^1 (\widehat{v}_p(x;\widehat{\tau}) - v(x))^2 dx \\ =&O_P\left(h_1^4 + \left(\frac{\log n}{nh_1}\right)^2\right) + O_P\left(h_1^2\frac{\log n}{nh} + \frac{(\log n)^2}{n^2h_1h}\right) + O_P\left(h^2 + \frac{\log n}{nh}\right) + O_P\left(\frac{1}{n^2h_1^2}\right).\end{aligned}$$

증명 : 제안한 분산함수 추정량과 분산함수의 차는 다음과 같이 표현된다.

$$\widehat{v_p}(x;\widehat{\tau}) - v(x) = (\widehat{v_p}(x;\widehat{\tau}) - \widetilde{v}(x;\widehat{\tau})) + (\widetilde{v}(x;\widehat{\tau}) - v(x)) \,. \tag{3.2}$$

여기서 $\widetilde{v}(x;\widehat{\tau}) = \frac{1}{nh\widehat{f}(x)} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \left\{ R_i^2 - \widehat{\Delta}(\widehat{\tau}) \mathrm{I}[\widehat{\tau} \le X_i \le 1] \right\} + \widehat{\Delta}(\widehat{\tau}) \mathrm{I}[\widehat{\tau} \le x \le 1]$는 $\widehat{v}_p(x;\widehat{\tau})$에서 $\widehat{R}_i$대신 $R_i = Y_i - m(X_i)$를 이용한 분산함수의 추정량이며 $\widehat{f}(x)$는 확률밀도함수의 커널추정량인 $\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$이다.

$W_i(x) = \frac{1}{nh\widehat{f}(x)} K\left(\frac{X_i - x}{h}\right)$, $i = 1, \cdots, n$이라 두면, 식 (3.2)의 오른쪽 첫 번째 항은 다음과 같이 표현된다.

$$\begin{aligned}
\widehat{v}_p(x;\widehat{\tau}) - \widetilde{v}(x;\widehat{\tau}) &= \sum_{i=1}^{n} W_i(x) \left\{ \widehat{R}_i^2 - \widehat{\Delta}(\widehat{\tau}) \mathrm{I}[\widehat{\tau} \le X_i \le 1] - \left(R_i^2 - \widehat{\Delta}(\widehat{\tau}) \mathrm{I}[\widehat{\tau} \le X_i \le 1] \right) \right\} \\
&= \sum_{i=1}^{n} W_i(x) \left\{ (\widehat{m}(X_i) - m(X_i))^2 - 2R_i \left(\widehat{m}(X_i) - m(X_i) \right) \right\} .
\end{aligned}$$

위 결과에 의해서 다음이 성립하며

$$\begin{aligned}
\sup_{x\in[0,1]} |\widehat{v}_p(x;\widehat{\tau}) - \widetilde{v}(x;\widehat{\tau})| \le \sup_{x\in[0,1]} &\left| \sum_{i=1}^{n} W_i(x) \right| \left\{ \sup_{x\in[0,1]} |\widehat{m}(x) - m(x)| \right\}^2 \\
&+ 2\sup_{x\in[0,1]} \left| \sum_{i=1}^{n} W_i(x) R_i \right| \sup_{x\in[0,1]} |\widehat{m}(x) - m(x)| \,.
\end{aligned} \tag{3.3}$$

Mack과 Silverman (1982)에 의해서 $\sup_{x\in[0,1]} |\widehat{m}(x) - m(x)| = O_P(h_1 + \sqrt{\frac{\log n}{nh_1}})$이고 Kang과 Huh (2006)의 Lemma1에 의해서 $\sup_{x\in[0,1]} \left| \sum_{i=1}^{n} W_i(x) R_i \right| = O_P(\sqrt{\frac{\log n}{nh}})$이므로 식 (3.3)는 $\sup_{x\in[0,1]} |\widehat{v}_p(x;\widehat{\tau}) - \widetilde{v}(x;\widehat{\tau})| = O_P(h_1^2 + \frac{\log n}{nh_1}) + O_P(h_1\sqrt{\frac{\log n}{nh}} + \frac{\log n}{n\sqrt{h_1 h}})$를 만족한다.

한편, 식 (3.2)의 오른쪽 두 번째 항은 다음과 같이 표현된다.

$$\begin{aligned}
\widetilde{v}_p(x;\widehat{\tau}) - v(x) &= \sum_{i=1}^{n} W_i(x) R_i^2 - \Delta \mathrm{I}[X_i > \tau] - w(x) \\
&\quad + \sum_{i=1}^{n} W_i(x) \{ \Delta \mathrm{I}[X_i > \tau] - \widehat{\Delta}(\widehat{\tau}) \mathrm{I}[X_i > \widehat{\tau}] \} + \{ \widehat{\Delta}(\widehat{\tau}) \mathrm{I}[x > \widehat{\tau}] - \Delta \mathrm{I}[x > \tau] \}.
\end{aligned} \tag{3.4}$$

오차제곱항을 $\varepsilon_i^2 = \xi_i + 1$라 두면, 위 식의 첫 번째 항은 식 (2.6)과 $\sum_{i=1}^{n} W_i(x) = 1$에 의해 다음의 관계가 성립한다. 여기서 ξ_i는 평균이 0인 확률변수이다.

$$\begin{aligned}
&\sum_{i=1}^{n} W_i(x) \left\{ R_i^2 - \Delta \mathrm{I}[X_i > \tau] \right\} - w(x) \\
&= \sum_{i=1}^{n} W_i(x) \left\{ v(X_i)\varepsilon_i^2 - \Delta \mathrm{I}[X_i > \tau] \right\} - w(x) \\
&= \sum_{i=1}^{n} W_i(x) w(X_i) \xi_i + \left\{ \sum_{i=1}^{n} W_i(x) w(X_i) - w(x) \right\} + \sum_{i=1}^{n} W_i(x) \Delta \mathrm{I}[X_i > \tau] \xi_i .
\end{aligned} \tag{3.5}$$

Kang과 Huh (2006)의 Lemma1에 의해서 위 식 (3.5)의 첫 번째 항과 세 번째 항은 $O_P(\frac{\log n}{nh})$이며, Kang과 Huh (2006)의 식 (5.26)에 의해서 식 (3.5)의 두 번째 항은 $O_P(h + \sqrt{\frac{\log n}{nh}})$이다. 식

(3.4)의 오른쪽 두 번째 항과 세 번째 항은 식 (3.1)의 불연속점 추정량과 점프크기 추정량의 수렴속도와 $\sum_{i=1}^{n} W_i(x) = 1$에 의해 $O_P(n^{-1} + (nh_1)^{-1/2})$이 된다. 그러므로 식 (3.2)은

$$\begin{aligned}\sup_{x\in[0,1]} |\widehat{v}(x;\widehat{\tau}) - v(x)| =& O_P\left(h_1^2 + \frac{\log n}{nh_1}\right) + O_P\left(h_1\sqrt{\frac{\log n}{nh}} + \frac{\log n}{n\sqrt{h_1 h}}\right) \\ &+ O_P\left(h + \sqrt{\frac{\log n}{nh}}\right) + O_P\left(\frac{1}{n} + \frac{1}{n}h_1\right) \qquad (3.6)\end{aligned}$$

이 된다. 식 (3.6)에 의해서 정리의 결과를 얻을 수 있다. □

Kang과 Huh (2006)의 Therem 3.2의 결과와 비교해보자. $p = 2$인 적분제곱오차인 경우에 분산함수의 추정량 자체에 의한 수렴속도는 동일하다. 하지만, 본 연구는 점프크기 추정량을 이용한 수정된 잔차들을 이용하였기에 적분제곱오차는 점프크기 추정량의 수렴속도의 제곱인 $(nh_1)^{-2}$에 추가적으로 의존하게 된다. Kang과 Huh의 결과에서 추가적인 수렴속도는 불연속점 추정량의 수렴속도와 관련하여 $(nh)^{-2}$에 의존하며, 점프크기 추정량의 수렴속도에는 의존하지 않는다. 두 추정량의 각각의 추가적인 수렴속도는 그 외의 수렴속도 부분들에 비해 빠르기 때문에 적분제곱오차의 수렴속도에 영향을 주지 않는다. 비록 본 연구의 분산함수 추정량의 적분제곱오차의 수렴속도가 Kang과 Huh의 분산함수의 추정량의 적분제곱오차의 수렴속도와 동일하지만, 본 연구의 분산함수의 추정량은 불연속점 추정량 근처에서 점프크기 추정량에 의한 잔차들의 수정으로 경계점 문제를 보완하였기에 실제 구현에서는 Kang과 Huh의 추정량보다 추정의 정도가 우수할 것이다.

4. 모의실험

4절에서는 모의실험을 통하여 Kang과 Huh (2006)가 제안한 분산함수의 추정량과 본 연구에서 제안한 분산함수의 추정량을 비교해보고자 한다. 설명변수 X의 분포는 토대 $[0,1]$을 가지는 균등분포를 선택하였다. 분산함수가 한 점에서 불연속인 두 가지 경우를 다음과 같이 고려하였다.

$$v_1(x) = \begin{cases} \dfrac{x^2}{9}, & 0 \le x \le 0.6 \\ 9(1-x)^2, & 0.6 < x \le 1, \end{cases}$$

$$v_2(x) = \begin{cases} 9x^2, & 0 \le x \le 0.4 \\ \dfrac{(1-x)^2}{9}, & 0.4 < x \le 1. \end{cases}$$

두 분산함수에 공통으로 이용될 회귀함수는

$$m(x) = 4x + 4\mathrm{e}^{-100(x-0.5)^2},\ 0 \le x \le 1$$

이다. 식 (2.1)의 회귀모형의 오차 ε_i의 분포는 표준정규분포를 선택하였고, 표본의 수는 500이며 반복은 1000회 실시하였다. 분산함수 v_1과 v_2는 각각 0.6과 0.4에서 불연속이다. 점프크기를 양수인 경우와 음수인 경우를 각각 고려하기 위하여 v_1과 v_2의 점프크기를 각각 1.4와 -1.4를 선택하였다.

불연속점과 점프크기를 추정하는 식 (2.3)의 한쪽방향커널함수로는 Epanechnikov 커널을 토대가 $[0,1]$이며 확률밀도함수가 될 수 있도록 다음

$$K(x) = \frac{3}{2}(1-x^2) \times \mathrm{I}[0 \le x \le 1]$$

을 선택하였다. 회귀함수와 분산함수를 추정하기 위한 식 (2.5)와 (2.8)의 커널함수는 다음과 같이 Epanechnikov 커널을 선택하였다.

$$L(x) = \frac{3}{4}(1 - x^2) \times \mathrm{I}[-1 \le x \le 1].$$

Kang과 Huh의 분산함수 추정량과 제안한 추정량을 비교하기 위한 모의실험이므로 v_1과 v_2의 불연속점 τ들을 아는 것으로 하여 참값을 이용하였다. 잔차의 계산에 쓰이는 $\widehat{m}(x)$의 띠폭 h_m은 0.05, 0.10, 0.15를 선택하였고 불연속점 τ에서 점프크기 추정 $\widehat{\Delta}(\tau)$의 띠폭 h_1도 0.05, 0.10, 0.15를 선택하여 다양한 띠폭들에 대한 분산함수의 추정량의 추정 정도를 조사하였다. 분산함수의 추정량들의 계산에 쓰이는 띠폭 h는 다양하게 변화를 주면서 Kang과 Huh의 분산함수 추정량과 제안한 추정량의 적분제곱오차들을 계산하였다.

Table 4.1과 4.2는 각각 v_1과 v_2의 모의실험 결과이며, Kang과 Huh의 분산함수 추정량 $\widehat{v}_{KH}(x;\tau)$와 제안한 추정량 $\widehat{v}_p(x;\tau)$의 추정된 적분제곱오차를 최소로 하는 띠폭 h와 그 때의 추정된 적분제곱오차를 보여주고 있다. 괄호 안은 추정된 적분제곱오차들 각각의 표준오차들이다. 두 분산함수 모형에서 띠폭 h_m과 h_1이 커짐으로써 제안한 추정량 $\widehat{v}_p(x;\tau)$의 추정된 적분제곱오차들도 증가하는 경향을 보여주고 있다. 띠폭 h_m이 0.05인 경우에 $\widehat{v}_{KH}(x;\tau)$의 최소의 추정된 적분제곱오차가 $\widehat{v}_p(x;\tau)$의 최소의 추정된 적분제곱오차보다 작았으며 그 외의 경우에는 $\widehat{v}_p(x;\tau)$의 최소의 추정된 적분제곱오차가 작은 경향을 보여주고 있다. 비록 h_m이 0.05일 때 $\widehat{v}_{KH}(x;\tau)$보다 $\widehat{v}_p(x;\tau)$의 최소의 추정된 적분제곱오차가 크지만 그 차이는 미미하며 다양한 띠폭 h에 대해서 전반적으로 $\widehat{v}_p(x;\tau)$의 추정된 적분제곱오차들이 작은 결과를 얻었다.

Table 4.1 The minimum estimated ISE over $h$with the standard errors given in parentheses and the minimizing bandwidth $h$for the case of v_1

h_m	h_1	Kang and Huh		Proposed	
		h	ISE	h	ISE
0.05	0.05	0.09	0.006814 (0.000190)	0.10	0.006943 (0.000202)
	0.10			0.10	0.006827 (0.000190)
	0.15			0.10	0.007462 (0.000186)
0.10	0.05	0.11	0.011508 (0.000159)	0.15	0.010151 (0.000211)
	0.10			0.16	0.010998 (0.000163)
	0.15			0.15	0.013386 (0.000163)
0.15	0.05	0.24	0.064695 (0.000238)	0.26	0.042283 (0.000488)
	0.10			0.28	0.045339 (0.000306)
	0.15			0.28	0.060955 (0.000301)

Table 4.2 The minimum estimated ISE over $h$with the standard errors given in parentheses and the minimizing bandwidth $h$for the case of v_2

h_m	h_1	Kang and Huh		Proposed	
		h	ISE	h	ISE
0.05	0.05	0.10	0.006924 (0.000211)	0.10	0.007344 (0.000263)
	0.10			0.11	0.006959 (0.000217)
	0.15			0.10	0.007577 (0.000206)
0.10	0.05	0.11	0.011698 (0.000191)	0.16	0.010645 (0.000297)
	0.10			0.17	0.010579 (0.000197)
	0.15			0.16	0.013479 (0.000181)
0.15	0.05	0.24	0.064500 (0.000254)	0.26	0.043140 (0.000607)
	0.10			0.28	0.045530 (0.000353)
	0.15			0.28	0.061035 (0.000323)

이를 보여주기 위하여, 모형 v_1에 대해서 Figure 4.1, 4.2와 4.3은 모의실험에 쓰인 띠폭 h_m인 0.05, 0.10과 0.15일 때 각각 다양한 띠폭 h에 대한 추정된 적분제곱오차들의 변화를 보여주고 있다. 이들 Figure에서 제안한 추정량 $\widehat{v}_p(x;\tau)$에 쓰인 점프크기 추정의 띠폭 h_1은 0.10인 경우만 고려하였다. 그 외의 h_1인 0.05와 0.15에 대해서도 추정된 적분제곱오차들의 변화는 비슷한 경향을 가지기에 그림을 생략하였다. 분산함수 모형 v_2도 모형 v_1의 결과와 유사하기에 추정된 적분제곱오차의 변화에 대한 그림을 생략하였다. Table 4.1과 4.2에서 설명한 것처럼 h_m이 0.05일 때 최소의 추정된 적분제곱오차는 $\widehat{v}_{KH}(x;\tau)$가 작은 값을 가지지만 h_m이 0.05와 0.10일 때 Figure 4.1과 4.2에서 대부분의 h에서는 $\widehat{v}_p(x;\tau)$가 추정된 적분제곱오차가 작은 값을 가짐을 알 수 있다. 띠폭 h_m이 0.15일 때 Figure 4.3에서는 모든 h에 대해서 $\widehat{v}_p(x;\tau)$의 추정된 적분제곱오차가 작음을 알 수 있다.

따라서, 분산함수 모형 v_1과 v_2의 모의실험 결과에 의하면 제안한 점프크기 추정량에 의해 수정된 잔차들을 이용한 분산함수 추정량이 Kang과 Huh가 제안한 분산함수 추정량보다 우수함을 알 수 있다. 이는 3절에서 언급하였듯이 제안한 분산함수 추정량은 불연속점 근처에서 점프크기 추정량에 의해 잔차들의 수정으로 경계점 문제를 보완하여 추정하였기 때문이다.

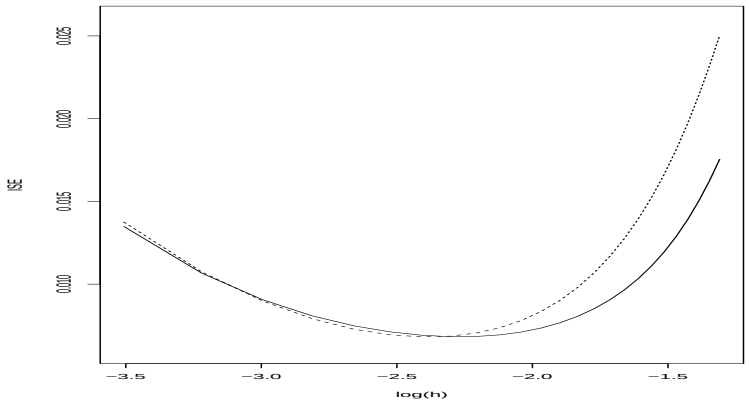

Figure 4.1 The estimated ISE as function of log-bandwidth for the case of v_1 and $h_m = 0.05$.
The estimated ISEs of $\widehat{v}_p(x;\tau)$ and $\widehat{v}_{KH}(x;\tau)$ represented by the solid and the dotted line respectively.

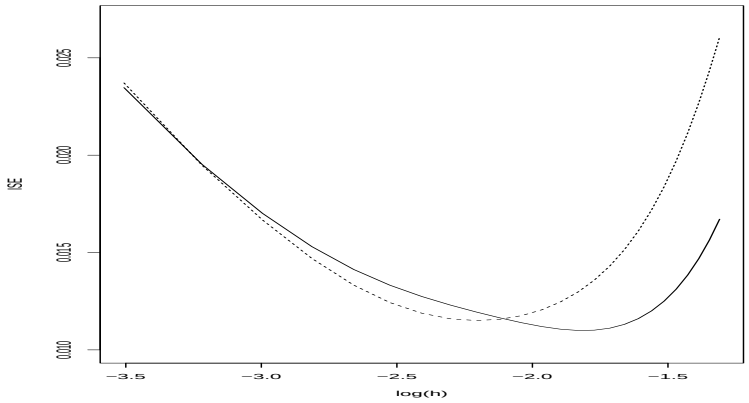

Figure 4.2 The estimated ISE as function of log-bandwidth for the case of v_1 and $h_m = 0.10$.
The estimated ISEs of $\widehat{v}_p(x;\tau)$ and $\widehat{v}_{KH}(x;\tau)$ represented by the solid and the dotted line respectively.

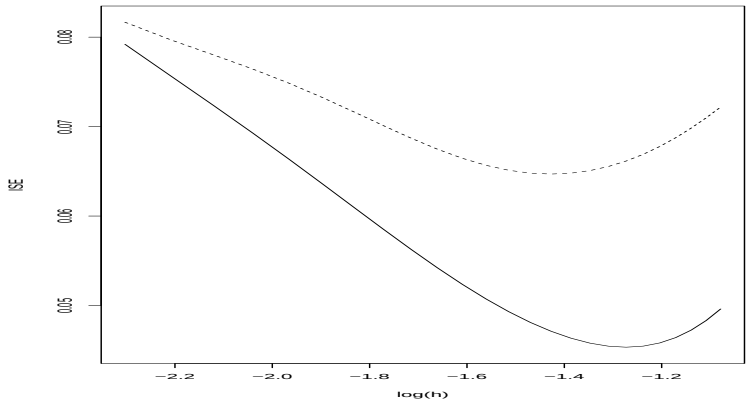

Figure 4.3 The estimated ISE as function of log-bandwidth for the case of v_1 and $h_m = 0.15$. The estimated ISEs of $\widehat{v}_p(x;\tau)$ and $\widehat{v}_{KH}(x;\tau)$ represented by the solid and the dotted line respectively.

References

Chen, L., Chen, M. and Peng, M. (2009). Conditional variance estimation in heteroscedastic regression models. *Journal of Statistical Planning and Inference*, **139**, 236-245.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625-634.

Hall, P. and Carroll, R. J. (1989). Variance function estimation in regression: The effect of estimating the mean. *Journal of the Royal Statistical Society B*, **51**, 3-14.

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

Huh, J. (2005). Nonparametric detection of a discontinuity point in the variance function with the second moment function. *Journal of the Korean Data & Information Science Society*, **16**, 591-601.

Huh, J. (2009). Testing a discontinuity point in the log-variance function based on likelihood. *Journal of the Korean Data & Information Science Society*, **20**, 1-9.

Huh, J. (2014). Comparison study on kernel type estimators of discontinuous log-variance. *Journal of the Korean Data & Information Science Society*, **25**, 87-95.

Huh, J. (2015). Estimation of a change point in the variance function based on the χ^2-distribution. *Communications in Statistics-Theory and Methods*, in press.

Kang, K. H. and Huh, J. (2006). Nonparametric estimation of the variance function with a change point. *Journal of the Korean Statistical Society*, **35**, 1-24.

Kang, K. H., Koo, J. Y. and Park, C. W. (2000). Kernel estimation of discontinuous regression functions. *Statistics and Probability Letters*, **47**, 277-285.

Lee, S., Shim, B. Y. and Kim, J. (2015). Estimation of hazard function and hazard change-point for the rectal cancer data. *Journal of the Korean Data & Information Science Society*, **26**, 1225-1238.

Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **61**, 405-415.

Müller, H G. (1992). Change-points in nonparametric regression analysis. *The Annals of Statistics*, **20**, 737-761.

Müller, H. G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics*, **15**, 610-625.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics,* **12**, 1215-1230.

Ruppert, D., Wand, M. P., Holst, U. and Hössjer, O. (1997). Local polynomial variance-function estimation. *Technometrics,* **39**, 262-273.

Sohn, S.-Y. and Cho, G.-Y. (2015). A change point estimator in monitoring the parameters of a multivariate IMA(1,1) model. *Journal of the Korean Data & Information Science Society*, **26**, 525-533.

Yu, K. and Jones, M. C. (2004). Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association*, **99**, 139-144.

Journal of the Korean Data & Information Science Society
2016, **27**(1), 111–120

http://dx.doi.org/10.7465/jkdi.2016.27.1.111
한국데이터정보과학회지

Nonparametric estimation of the discontinuous variance function using adjusted residuals[†]

Jib Huh[1]

[1]Department of Statistics, Duksung Women's University

Received 15 December 2015, revised 4 January 2016, accepted 6 January 2016

Abstract

In usual, the discontinuous variance function was estimated nonparametrically using a kernel type estimator with data sets split by an estimated location of the change point. Kang *et al.* (2000) proposed the Gasser-Müller type kernel estimator of the discontinuous regression function using the adjusted observations of response variable by the estimated jump size of the change point in Müller (1992). The adjusted observations might be a random sample coming from a continuous regression function. In this paper, we estimate the variance function using the Nadaraya-Watson kernel type estimator using the adjusted squared residuals by the estimated location of the change point in the discontinuous variance function like Kang *et al.* (2000) did. The rate of convergence of integrated squared error of the proposed variance estimator is derived and numerical work demonstrates the improved performance of the method over the exist one with simulated examples.

Keywords: Change point, jump size, kernel function, residual, variance function.

[†] This research was supported by the Duksung Women's University Research Grants 2014.
[1] Professor, Department of Statistics, Duksung Women's University, Seoul 01369, Korea.
E-mail: jhuh@duksung.ac.kr