Journal of the Korean Data & Information Science Society
2014, **25**(1), 87–95

http://dx.doi.org/10.7465/jkdi.2014.25.1.87
한국데이터정보과학회지

불연속 로그분산함수의 커널추정량들의 비교 연구[†]

허집[1]

[1]덕성여자대학교 정보통계학과

접수 2013년 12월 2일, 수정 2013년 12월 22일, 게재확정 2014년 1월 2일

요 약

분산함수가 불연속인 경우 Kang과 Huh (2006)는 잔차제곱을 이용한 Nadaraya-Watson 추정량으로 분산함수를 추정하였다. 음의 실수 값도 가질 수 있는 로그분산함수를 추정 대상으로 하여, 오차제곱의 분포를 χ^2-분포로 가정하고 국소선형적합을 이용한 불연속 로그분산함수의 추정이 Huh (2013)에 의해 연구되었다. Chen 등 (2009)은 연속인 로그분산함수를 로그잔차제곱을 이용한 국소선형적합으로 추정하였다. 본 연구는 Chen 등의 추정법을 이용하여 불연속인 로그분산함수의 추정량을 제시하였다. 기존의 제안된 불연속인 로그분산함수의 추정량들과 제안된 추정량을 모의실험을 통하여 비교연구하고자 한다. 한편, 로그분산함수가 연속이지만 그 미분된 함수가 불연속일 경우, Huh (2013)의 방법과 제안된 방법으로 적합된 국소선형의 기울기를 이용하여 불연속인 미분된 로그분산함수의 추정량을 제시하고자 한다. 이들 추정량의 비교 연구 또한 모의실험을 통하여 제시하고자 한다.

주요용어: 국소선형적합, 로그분산함수, 로그잔차제곱, 불연속점, 커널함수.

1. 서론

이변량 표본 $\{(X_i, Y_i) : i = 1, \cdots, n\}$은 설명변수 X와 반응변수 Y의 확률벡터 (X, Y)로부터 추출된 임의의 표본이라 하자. 설명변수 X의 토대 (support)는 $[0, 1]$이며 확률밀도함수는 $f(x)$를 가진다고 할 때, 회귀함수와 분산함수를 각각 $m(x) = E(Y|X = x)$와 $v(x) = Var(Y|X = x)$라 두면 회귀모형은 다음과 같이 정의된다.

$$Y_i = m(X_i) + v(X_i)^{1/2}\varepsilon_i,\ i = 1, \cdots, n, \tag{1.1}$$

여기서 오차항 ε_i는 $X_1, X_2, \cdots, X_n$과 독립이며, 그들의 평균과 분산은 각각 0과 1이다.

회귀모형에서 중요한 추정 대상인 회귀함수와 분산함수의 비모수적 추정은 커널함수를 이용한 추정법이 이론적 추정의 우수성뿐만 아니라 실제 구현에서도 계산의 편리성으로 인해 널리 쓰이고 있다. 분산함수의 커널추정 연구로는 Rice (1984), Gasser 등 (1986), Müller와 Stadtmüller (1987), Hall과 Carroll (1989), Hall 등 (1990), Ruppert 등 (1997), Yu와 Jones (2004), Chen 등 (2009) 등이 있다. 이들의 연구는 분산함수가 연속인 경우이다.

분산함수가 불연속점인 경우는 두 가지이다. 첫 번째, 회귀함수가 불연속인 경우에 그 불연속점에서 분산함수도 불연속일 수 있다. 이러한 경우는 분산함수 추정법으로 불연속점을 추정하기 보다는 회귀

[†] 본 연구는 덕성여자대학교 2012년도 교내연구비 지원에 의해 수행되었음.
[1] (132-714) 서울특별시 도봉구 삼양로 144길 33, 덕성여자대학교 정보통계학과, 부교수.
E-mail: jhuh@duksung.ac.kr.

함수 추정법으로 불연속점을 추정하는 것이 편리할 것이다. 이러한 회귀함수가 불연속인 경우의 연구로는 Müller (1992), Loader (1996), Gréoire와 Hamrouni (2002), Huh와 Carrière (2002), Huh와 Park (2004), Huh (2010) 등이 있다. 두 번째, 회귀함수는 연속이지만 분산함수가 불연속일 수 있다. 이 경우에 분산함수의 불연속점의 커널추정은 Huh (2005), Kang과 Huh (2006), Huh (2009), Huh (2013)에 의해 이루어졌다. Huh (2005)는 회귀함수가 연속이라는 점을 이용하여 커널추정법으로 이차적률 (second moment)함수를 추정하여 분산함수의 불연속점을 추정하였다. Kang과 Huh (2006)는 회귀함수의 커널추정량으로 만들어진 잔차제곱을 국소상수항적합 (local constant fit)에 의한 Nadaraya-Watson 커널추정량으로 분산함수의 불연속점을 추정하였다. 분산함수는 음의 값을 갖지 않는 함수이다. 고차커널함수 (higher-order kernel function)를 사용한 분산함수의 추정이나, 국소다항적합 (local polynomial fit)을 이용한 분산함수의 추정은 추정된 분산함수가 음의 값을 가질 수도 있는 단점이 있다. 분산함수가 연속일 때, Yu와 Jones (2004)는 오차제곱이 카이제곱분포를 따른다고 가정하고 로그분산함수 (log-variance function)을 국소선형적합으로 추정하여 분산함수의 국소선형적합이 음의 값을 가질 수 있는 단점을 보완하였다. Huh (2013)는 Yu와 Jones (2004)의 국소선형적합을 이용한 로그분산함수의 커널추정법을 이용하여 불연속점 추정량을 제시하였다.

한편, Kang과 Huh (2006)와 Huh (2013)는 추정된 불연속점을 기준으로 표본을 양분하여 양분된 각 영역의 표본을 독립적인 표본으로 간주하고 각 영역에서 분산함수 혹은 로그분산함수를 추정하였다. Kang과 Huh (2006)는 추정된 불연속점으로 양분된 두 영역의 잔차제곱을 Nadaraya-Watson 커널추정량을 이용하여 불연속인 분산함수의 커널추정량을 제시하였다. Huh (2013)는 추정된 불연속점에 의해 양분된 두 영역의 잔차제곱의 분포를 카이제곱으로 가정하고 로그분산함수를 국소다항적합으로 추정하여 Kang과 Huh (2006)가 제안한 불연속인 분산함수의 추정량의 정도 (precision)를 개선하였다. Fan과 Gijbels (1996) 등에 의해 국소상수항적합은 경계점 (boundary point)에서 이론적으로 추정의 정도가 떨어진다고 익히 알려져 있다. 추정된 불연속점을 기준으로 표본을 양분하여 독립적으로 분산함수 혹은 로그분산함수를 추정하므로, 추정된 불연속점은 경계점과 같은 역할을 하게 된다. 따라서, 경계점 문제가 없는 국소다항적합을 이용한 Huh (2013)의 로그분산함수 추정은 국소상수항적합으로 분산함수를 추정한 Kang과 Huh (2006)의 추정법을 개선한 것이 된다.

Huh (2013)의 잔차제곱으로 국소다항적합을 이용한 로그분산함수의 추정은 커널함수를 가중한 카이제곱분포의 가능도함수 (likelihood function)를 최대로 하는 과정을 거친다. 가능도함수를 최대로 하는 해는 명시적 형태 (explicit form)로 표현되지 않기 때문에 초기치 (initial value)를 주고 반복적으로 계산하여 최대로 하는 해를 구하게 된다. 이러한 내재적 형태 (implicit form)의 추정치를 구하는 방법은, 잘 알려져 있듯이, 적절하지 않은 초기치 설정으로 인해 최대로 하는 해를 구하지 못하거나, 비록 최대로 하는 해를 구하였더라도 방대한 계산량의 문제를 야기할 수도 있다.

본 연구에서는 명시적 형태의 추정량을 제시하면서도 국소선형적합의 장점을 활용할 수 있는 불연속인 로그분산함수의 추정량을 제안하고자 한다. 제안할 불연속 로그분산함수의 추정법은 Chen 등 (2009)이 로그잔차제곱의 국소선형적합으로 제안한 연속인 로그분산함수의 비모수적 추정법을 이용한 것이다. Huh (2013)와 제안된 불연속 로그분산함수의 추정법은 국소선형적합을 이용한 것이다. 로그분산함수는 연속이지만 미분된 로그분산함수가 불연속점을 가질 때, 적합된 국소선형의 기울기를 이용하여 불연속인 미분된 로그분산함수의 추정량을 제안할 것이다.

2절은 Kang과 Huh (2006)가 제안한 불연속 분산함수 추정량과, Huh (2013)의 불연속 로그분산함수의 추정에 대해 소개하고, 로그잔차제곱의 국소선형적합을 이용한 불연속 로그분산함수의 추정법을 제안한다. 또한, Huh (2013)의 추정법과 로그잔차제곱을 이용한 제안된 추정법의 적합된 국소선형의 기울기를 이용하여 불연속인 미분된 로그분산함수의 추정량을 제안한다. 3절에서는 2절에서 소개된 추정법들을 모의실험을 통한 비교연구 결과를 설명하고, 본 연구에 대한 맺음말을 4절에 제시한다.

2. 불연속 로그분산함수의 추정

분산함수의 불연속점을 $\tau \in (0,1)$라 하고 알려져 있다고 하자. 즉, $v_+(\tau) - v_-(\tau) \neq 0$이다. 여기서 $v_+(\tau) = \lim_{x\to\tau+} v(x)$와 $v_-(\tau) = \lim_{x\to\tau-} v(x)$이다. 회귀함수의 어떤 커널형 추정량을 $\widehat{m}$이라 하자. 불연속점 τ를 알고 있을 때, Kang과 Huh (2006)는 국소상수항적합에 의한 커널추정량인 Nadaraya-Watson 커널추정량으로 다음과 같이 τ에서 불연속인 분산함수를 추정하였다.

$$\widehat{v}_{KH}(x;\tau) = \frac{\sum_{i=1}^{n} K_h^*(X_i - x;\tau)\widehat{R}_i}{\sum_{i=1}^{n} K_h^*(X_i - x;\tau)}, \tag{2.1}$$

여기서 h는 평활모수인 띠폭 (bandwidth)이며, $\widehat{R}_i = \{Y_i - \widehat{m}(X_i)\}^2$는 잔차제곱이고 $K_h^*(u-x;\tau)$는 다음과 같다.

$$K_h^*(u-x;\tau) = \begin{cases} \dfrac{1}{h}K\left(\dfrac{u-x}{h}\right) \mathrm{I}[x-h \le u < \tau], & \tau - h \le x < \tau \\ \dfrac{1}{h}K\left(\dfrac{u-x}{h}\right) \mathrm{I}[\tau \le u < x+h], & \tau \le x < \tau + h \\ \dfrac{1}{h}K\left(\dfrac{u-x}{h}\right), & \text{그 외,} \end{cases} \tag{2.2}$$

여기서 함수 K는 토대 $[-1,1]$를 가지는 커널함수이고, I는 표시함수 (indicator function)이다. 식 (2.2)의 커널함수 K_h^*를 사용함으로써 표본은 불연속점 τ를 기준으로 왼쪽 표본과 오른쪽 표본으로 분리되어진다. 따라서 τ의 왼쪽 분산함수와 오른쪽 분산함수는 각각 τ의 왼쪽 표본과 오른쪽 표본을 이용하여 Nadaraya-Watson 추정량으로 추정되는 것이다.

회귀모형 (1.1)에서 오차 ε_i가 정규분포를 가진다고 가정하면 오차제곱은 다음과 같이

$$\varepsilon_i^2 = \frac{\{Y_i - m(X_i)\}^2}{v(X_i)} = \frac{\{Y_i - m(X_i)\}^2}{\mathrm{e^s(X_i)}} \tag{2.3}$$

로 표현되고, X_i가 주어졌을 때 식 (2.3)은 $\chi^2(1)$분포를 가지게 된다. 여기서 $s(x) = \log v(x)$이다. 비록 ε_i가 정규분포를 가지지 않더라도, 그 분포가 평균 0을 중심으로 대칭이라고 가정하면 식 (2.3)의 분포는 근사적으로 $\chi^2(1)$이 될 수 있다. 이러한 오차 ε_i의 분포가 대칭이라는 가정은 논리적으로 충분히 타당성이 있다. 한편, 분산함수는 음이 아닌 실수 값을 가지는 함수이므로 국소선형적합이나 고차커널함수를 이용한 커널추정량으로 추정하는 것은 적절하지 않다. Yu와 Jones (2004)는 오차제곱의 분포를 $\chi^2(1)$이라 가정하고 이 분포의 가능도함수 (likelihood function)을 이용한 국소선형적합으로 연속인 로그분산함수를 추정하였다. 분산함수가 $v(x)$가 τ에서 불연속이면 로그분산함수 $s(x)$도 τ에서 불연속이다. Huh (2013)는 Yu와 Jones의 방법을 이용하여 식 (2.2)의 커널함수를 사용한 국소선형적합으로 불연속 로그분산함수를 다음과 같이 추정하였다.

로그분산함수 $s(x)$를 국소선형으로 근사적으로 표현하고 $\chi^2(1)$분포의 커널가중국소로그가능도함수 (kernel weighted local log-likelihood function)

$$\sum_{i=1}^{n} \ell(\alpha_0 + \alpha_1(X_i - x), \widehat{R}_i) K_h^*(X_i - x;\tau) \tag{2.4}$$

를 생각하자. 이를 최대로 하는 해를 $\underline{\widehat{\alpha}} = (\widehat{\alpha}_0, \widehat{\alpha}_1)^T$라 하면 불연속점 τ를 가지는 $s(x)$와 $v(x)$의 추정으로 $\widehat{s}_{YJ}(x;\tau) = \widehat{\alpha}_0$과 $\widehat{v}_{YJ}(x;\tau) = \mathrm{e}^{\widehat{\alpha}_0}$을 각각 제안할 수 있다. 여기서

$$\ell(u,y) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\left(\log y + u + \frac{y}{\mathrm{e}^u}\right)$$

로 $\chi^2(1)$분포를 로그변환한 것이다. 즉, $X_i = x$가 주어졌을 때 $u = s(x)$라 두고 $\{Y_i - m(X_i)\}^2$대신 $\widehat{R}_i$을 사용한 $\widehat{R}_i/\mathrm{e}^u$가 근사적으로 가지게 되는 $\chi^2(1)$분포의 로그변환이다. 식 (2.4)에서 $\alpha_0 + \alpha_1(X_i - x)$대신 국소상수항 α_0를 사용하여 식 (2.4)를 최대로 하는 해 $\widehat{\alpha}_0$을 구하면, 분산함수의 추정량 $\widehat{v}_{YJ}(x;\tau) = \mathrm{e}^{\widehat{\alpha}_0}$은 Kang과 Huh (2006)의 추정량인 식 (2.1)의 $\widehat{v}_{KH}(x;\tau)$와 일치한다.

Huh (2013)가 제안한 추정량 $\widehat{s}_{YJ}(x;\tau)$는 실수 값을 취할 수 있는 로그분산함수를 추정 대상으로 하여 경계점에서 우수한 이론적 성질을 가지고 있는 국소선형적합을 이용하여 추정한 것이다. 경계점과 같은 역할을 하고 있는 불연속점을 가진 로그분산함수의 국소선형적합을 이용한 Huh (2013)의 추정 방법이 Kang과 Huh (2006)의 추정 방법보다 불연속점 근처에서 이론적으로 우수하다는 것을 Huh (2013)가 보였다. 식 (2.4)를 최대로 하는 해는 명시적 형태로 표현되어지지 않고 초기치를 이용한 반복적 계산으로 구해지는 내재적 형태를 가지고 있다. 따라서 적절한 초기치 설정과 많은 계산량이 단점이 될 수 있다.

로그분산함수가 연속일 때, Chen 등 (2009)은 $E(\log(\varepsilon_i^2/d)) = 0$가 되게 하는 상수 d를 이용하여 $\log(Y_i - m(X_i))^2 = \log(dv(x)) + \log(\varepsilon_i^2/d)$라 표현하고, 로그잔차제곱 $\log \widehat{R}_i$로 국소선형적합을 이용한 로그분산함수의 커널추정량을 제시하였다. Chen 등의 로그분산함수의 커널추정 방법과 식 (2.2)의 커널함수를 이용하여 한 점 τ에서 불연속인 로그분산함수의 추정을 다음과 같이 제안하고자 한다.

함수 $\log(dv(x))$를 국소선형으로 근사적으로 표현하고 다음의 식

$$\sum_{i=1}^{n}\left\{\log \widetilde{R}_i - \beta_0 - \beta_1(X_i - x)\right\}^2 K_h^*(X_i - x;\tau) \tag{2.5}$$

를 생각하자. 잔차제곱 $\widehat{R}_i$이 0에 매우 근접한 경우의 문제를 해결하기 위하여 식 (2.5)에는 다음과 같이 주어진 어떤 작은 실수 c에 대하여

$$\log \widetilde{R}_i = \begin{cases} \log \widehat{R}_i, & \widehat{R}_i > c \\ \log c, & \text{그 외}, \end{cases} \tag{2.6}$$

를 사용하였다. Chen 등 (2009)은 잔차제곱 $\widehat{R}_i$이 0에 매우 근접한 경우의 문제를 해결하기 위하여 $\log(\widehat{R}_i + n^{-1})$을 사용하였다. 여기서 n은 표본의 수이다. 식 (2.5)을 최소로 하는 해를 $\underline{\widehat{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)^T$라 하면 $dv(x)$의 추정으로 $\widehat{dv(x)} = \mathrm{e}^{\widehat{\beta}_0}$로 정의하면 분산함수를 $\widehat{v}_{CP}(x;\tau) = \mathrm{e}^{\widehat{\beta}_0}/\widehat{d}$로 추정할 수 있다. 여기서, d의 추정 $\widehat{d}$은 Chen 등 (2009)이 제안한 것으로

$$\widehat{d} = \left[n^{-1}\sum_{i=1}^{n}\widetilde{R}_i \exp(-\widehat{dv(X_i)})\right]^{-1}$$

이다. 불연속 로그분산함수의 추정으로는 분산함수의 추정량 $\widehat{v}_{CP}(x;\tau)$에 로그를 취한 형태로 $\widehat{s}_{CP}(x;\tau) = \widehat{\beta}_0 - \widehat{d}$로 정의할 수 있다. 식 (2.5)의 해 $\widehat{\beta}_0$는 명시적 형태로 다음과 같이 표현됨을 쉽게 알 수 있다.

$$\widehat{\beta}_0 = n^{-1}\sum_{i=1}^{n}\frac{\{u_2(x;\tau) - u_1(x;\tau)(X_i - x)\}\, K_h^*(X_i - x;\tau)\log \widetilde{R}_i}{u_2(x;\tau)u_0(x;\tau) - \{u_1(x;\tau)\}^2}, \tag{2.7}$$

여기서 $u_k(x;\tau) = n^{-1}\sum_{i=1}^{n}(X_i - x)^k K_h^*(X_i - x;\tau)$, $k = 0, 1, 2$이다. 제안된 추정량은 식 (2.7)과 같이 명시적 형태로 표현될 수 있는 장점이 있지만, 0 근처에서 로그값의 불안정성으로 수정된 $\log \widetilde{R}_i$을 사용하게 되어 왜곡된 추정량을 만들 수 있는 단점이 있다.

한편, 로그분산함수는 연속이지만 그 미분된 함수 $s'(x)$가 τ에서 불연속인 경우를 생각해보자. Huh (2013)는 식 (2.4)를 최대로 하는 $\widehat{\alpha}_1$을 이용하여 $s'(x)$를 $\widehat{s'}_{YJ}(x;\tau) = \widehat{\alpha}_1$으로 추정하였다. Chen 등 (2009)의 방법에 의한 식 (2.5)을 최소로 하는 해 $\widehat{\beta}_1$을 이용하여, $s'(x)$의 추정을 다음과 같이 $\widehat{s'}_{CP}(x;\tau) = \widehat{\beta}_1$으로 제안하고자 한다. 이 때 $\widehat{\beta}_1$의 명시적 형태는 다음과 같다.

$$\widehat{\beta}_1 = n^{-1}\sum_{i=1}^{n} \frac{\{u_0(x;\tau)(X_i - x) - u_1(x;\tau)\}\, K_h^*(X_i - x;\tau)\log \widetilde{R}_i}{u_2(x;\tau)u_0(x;\tau) - \{u_1(x;\tau)\}^2}.$$

3. 모의실험을 통한 비교

2절에서 소개한 추정량들을 모의실험을 통하여 비교해보고자 한다. 모의실험에 쓰이는 설명변수 X는 토대 $[0, 1]$을 가지는 균등분포 (uniform distribution)을 고려하였다. 로그분산함수가 불연속인 경우와 미분된 로그분산함수가 불연속인 두 모형을 다음과 같이

$$v_1(x) = \begin{cases} x^2, & 0 \le x \le 0.65 \\ 25(1 - x^2), & 0.65 < x \le 1, \end{cases}$$

$$v_2(x) = \begin{cases} 25x^2, & 0 \le x \le 0.5 \\ 25(1 - x^2), & 0.5 < x \le 1, \end{cases}$$

을 고려하였다. 이때 두 모형에 공통으로 이용될 회귀함수는 다음과 같다.

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

분산함수 $v_1(x)$는 $\tau = 0.65$에서 불연속이며, 연속인 $v_2(x)$는 그 미분된 함수가 $\tau = 0.5$에서 불연속이다. 오차 ε_i의 분포는 표준정규분포를 선택하였다. 표본의 수 n은 500으로 하고, 반복은 1000회를 실시하였다. 제안된 추정량 $\widehat{s}_{CP}(x;\tau)$와 $\widehat{s'}_{CP}(x;\tau)$를 계산하는 과정에서 식 (2.6)의 상수 c의 선택이 필요하다. 상수 c를 다양하게 변화를 주며 $\widehat{s}_{CP}(x;\tau)$와 $\widehat{s'}_{CP}(x;\tau)$를 계산하고, $s(x)$와 $s'(x)$의 추정이 우수한 적절한 상수 $c = 0.0001$을 선택하였다.

두 회귀모형에 사용될 잔차를 구하기 위하여, 회귀함수의 추정량 $\widehat{m}(x)$는 국소선형적합으로 추정하였으며, 이때 띠폭은 0.05로 하였고 사용된 커널함수는 흔히 사용되는 Epaneknikov 커널을 다음과 같이

$$K(x) = \frac{3}{4}(1 - x^2)I[-1 \le x \le 1] \tag{3.1}$$

을 선택하였다. 띠폭 h는 다양하게 변화를 주며 로그분산함수 혹은 미분된 로그분산함수를 추정하였고, 커널함수는 식 (3.1)의 Epaneknikov 커널함수를 사용하였다.

먼저, 불연속 분산함수 $v_1(x)$와 관련된 추정량들을 비교해 보자. 추정대상 함수는 로그분산함수로 하였다. Kang과 Huh (2006)가 제안한 불연속 분산함수의 추정량 $\widehat{v}_{KH}(x;\tau)$를 이용하여 불연속 로그분산함수를 다음과 같이 $\widehat{s}_{KH}(x;\tau) = \log \widehat{v}_{KH}(x;\tau)$로 추정할 수 있다. 띠폭 h를 다양하게 변화를 주며 불연속 로그분산함수의 추정량들 $\widehat{s}_{KH}(x;\tau)$, $\widehat{s}_{YJ}(x;\tau)$와 $\widehat{s}_{CP}(x;\tau)$를 계산하였다. 추정된 불연속 로그분산함수들의 추정 정도를 비교하기 위하여 적분제곱오차 (integrated squared error)의 몬테카를로 추정치를 구하였다. Figure 3.1은 분산함수 $v_1(x)$의 로그분산함수 추정량들인 $\widehat{s}_{KH}(x;\tau)$, $\widehat{s}_{YJ}(x;\tau)$와

$\widehat{s}_{CP}(x;\tau)$들의 $\log h$의 변화에 따른 적분제곱오차들을 그린 것이다. 가는점선, 실선, 굵은점선은 각각 $\widehat{s}_{KH}(x;\tau)$, $\widehat{s}_{YJ}(x;\tau)$, $\widehat{s}_{CP}(x;\tau)$의 적분제곱오차를 나타낸 것이다. $\widehat{s}_{YJ}(x;\tau)$의 최소 적분제곱오차가 다른 추정량들의 최소 적분제곱오차에 비해 작게 나타나고 있다. 한편, 국소선형적합을 이용한 두 추정량 $\widehat{s}_{YJ}(x;\tau)$, $\widehat{s}_{CP}(x;\tau)$의 최소 적분제곱오차가 국소상수항적합을 이용한 $\widehat{s}_{KH}(x;\tau)$의 최소 적분제곱오차보다 작다는 것을 알 수 있다.

다음은 미분된 로그분산함수가 불연속인 $v_2(x)$와 관련된 추정량인 $\widehat{s'}_{YJ}(x;\tau)$와 $\widehat{s'}_{CP}(x;\tau)$를 비교해 보자. Figure 3.2는 $\log h$의 변화에 따른 추정량 $\widehat{s'}_{YJ}(x;\tau)$와 $\widehat{s'}_{CP}(x;\tau)$들의 적분제곱오차들을 보여주고 있다. 실선과 점선은 각각 $\widehat{s'}_{YJ}(x;\tau)$와 $\widehat{s'}_{CP}(x;\tau)$의 적분제곱오차이다. 이 경우 또한 $\widehat{s'}_{YJ}(x;\tau)$의 최소 적분제곱오차가 $\widehat{s'}_{CP}(x;\tau)$의 최소 적분제곱오차보다 작게 나타나고 있다.

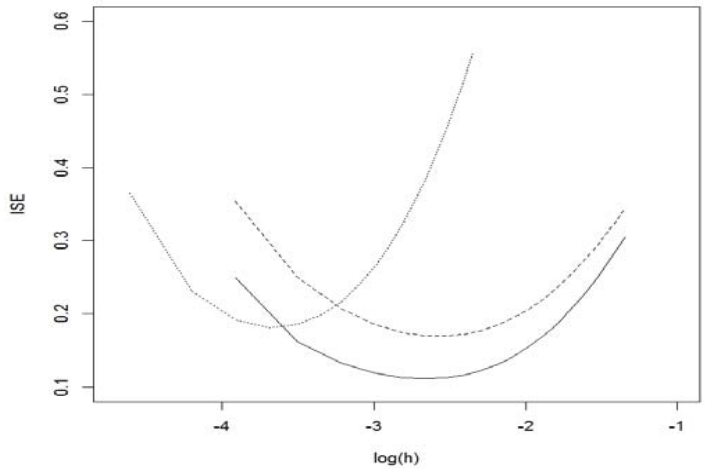

Figure 3.1 The ISEs as function of log-bandwidth for the case of v_1. The ISEs of $\widehat{s}_{KH}(x;\tau)$, $\widehat{s}_{YJ}(x;\tau)$ and $\widehat{s}_{CP}(x;\tau)$ represented by the dotted, the solid and the dashed line respectively.

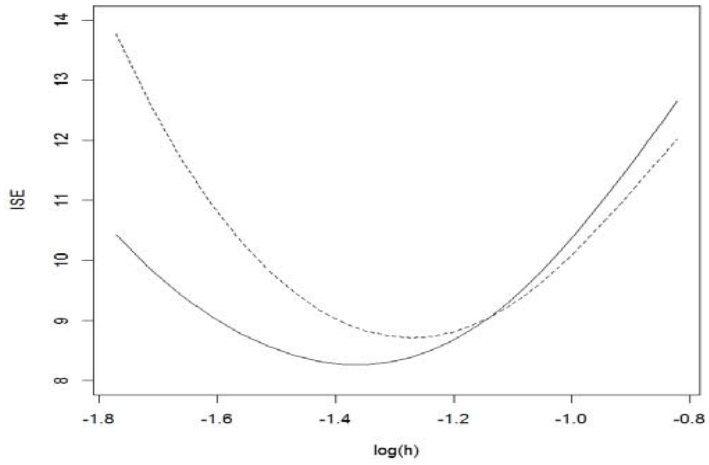

Figure 3.2 The ISEs as function of log-bandwidth for the case of v_2. The ISEs of $\widehat{s'}_{YJ}(x;\tau)$ and $\widehat{s'}_{CP}(x;\tau)$ represented by the solid and the dashed line respectively.

4. 맺음말

Kang과 Huh (2006)가 제안한 불연속 분산함수의 추정량은 잔차제곱을 이용한 국소상수항적합인 Nadaraya-Watson 커널추정량을 사용한 것이다. 이 추정량의 경우에 경계점과 경계점과 동일한 역할을 하고 있는 불연속점에서의 추정의 정도는 떨어지게 된다. 음이 값을 취하지 않는 분산함수이기에, 커널추정량에서 경계점의 문제가 없는 장점을 가진 국소선형적합을 사용하여 분산함수를 추정할 수 없어, Kang과 Huh (2006)의 추정량을 개선하고자 Huh (2013)는 로그분산함수를 추정 대상으로 삼아 오차제곱의 분산을 χ^2-분포로 가정하고 국소선형적합으로 불연속 로그분산함수를 추정하였다. Huh (2013)의 추정량은 Kang과 Huh (2006)의 추정량을 개선하였지만 명시적 형태를 가진 것이 아니기에, 커널가중 로그가능도함수에 초기치를 주어 반복적으로 계산을 해야 하는 단점이 있다.

Huh (2013)의 추정량의 단점인 계산상의 문제점을 극복하기 위하여, 본 연구에서 명시적 형태의 추정량을 제시하면서도 Kang과 Huh (2006)의 추정량의 문제점인 경계점과 불연속점에서 추정의 정도를 개선할 수 있는 추정량을 제안하였다. 이는 Chen 등 (2009)이 로그잔차제곱을 이용하여 국소선형적합으로 연속인 로그분산함수의 추정량을 제안한 방법을 이용한 것이다. 제안된 추정법은 명시적 형태를 가진 추정량을 만들지만, 잔차제곱이 0 근처인 경우 로그잔차제곱의 불안정성으로 인해 수정된 로그잔차제곱을 사용하게 되어 왜곡된 추정량을 만들게 되는 단점이 있다. 3절의 모의실험에서 Huh (2013)의 추정량이 제안된 추정량 보다 조금 우수하게 나타나는 이유도 여기에 연유한 것이라 추측된다. 국소선형적합 기법을 사용한 Huh (2013)와 제안된 추정량은 적합된 국소선형의 기울기로 불연속인 미분된 로그분산함수의 추정량을 생각할 수 있는 장점이 있다.

Kang과 Huh (2006)와 Huh (2013)는 교차타당성 (cross-validation)을 이용한 띠폭 선택 방법을 제시하였다. 최대가능도교차타당성, 최소제곱교차타당성, 불편교차타당성, 편의교차타당성 등을 이용한 불연속점 추정의 띠폭 선택 방법의 연구는 Huh (2012a, 2012b)에 의해 이루어졌다. 본 연구에서 제안한 추정량 $\widehat{s}_{CP}(x;\tau)$와 $\widehat{s'}_{CP}(x;\tau)$의 띠폭 선택 방법에 대한 연구도 차후 필요하다고 본다.

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.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its application*, Chapman and Hall, London.

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

Grégoire, G. and Hamrouni, Z. (2002). Change point estimation by local linear smoothing. *Journal of Multivariate Analysis*, **83**, 56-83.

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. (2010). Detection of a change point based on local-likelihood. *Journal of Multivariate Analysis*, **101**, 1681-1700.

Huh, J. (2012a). Bandwidth selection for discontinuity point estimation in density. *Journal of the Korean Data & Information Science Society*, **23**, 79-87.

Huh, J. (2012b). Bandwidth selections based on cross-validation for estimation of a discontinuity point in density. *Journal of the Korean Data & Information Science Society*, **23**, 765-775.

Huh, J. (2013). Estimation of a change point in the variance function based on the χ^2-distribution. Preprint.

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.

Huh, J. and Park, B. U. (2004). Detection of change point with local polynomial fits for random design case. *Australian and New Zealand Journal of Statistics*, **46**, 425-441.

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

Loader, C. R. (1996). Change point estimation using nonparametric regression. *Annals of Statistics*, **24**, 1667-1678.

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

Müller, H. G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *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. *Technomtrics*, **39**, 262-273.

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
2014, **25**(1), 87–95

http://dx.doi.org/10.7465/jkdi.2014.25.1.87
한국데이터정보과학회지

Comparison study on kernel type estimators of discontinuous log-variance[†]

Jib Huh[1]

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

Received 2 December 2013, revised 22 December 2013, accepted 2 January 2014

Abstract

In the regression model, Kang and Huh (2006) studied the estimation of the discontinuous variance function using the Nadaraya-Watson estimator with the squared residuals. The local linear estimator of the log-variance function, which may have the whole real number, was proposed by Huh (2013) based on the kernel weighted local-likelihood of the χ^2-distribution. Chen *et al.* (2009) estimated the continuous variance function using the local linear fit with the log-squared residuals. In this paper, the estimator of the discontinuous log-variance function itself or its derivative using Chen *et al.* (2009)'s estimator. Numerical works investigate the performances of the estimators with simulated examples.

Keywords: Discontinuity point, kernel function, local linear fit, log-squared residual, log-variance.

[†] This research was supported by the Duksung Women's University Research Grants 2012.
[1] Associate professor, Department of Statistics, Duksung Women's University, Seoul 132-714, Korea. E-mail: jhuh@duksung.ac.kr