Journal of the Korean Data & Information Science Society
2012, **23**(4), 765–775

http://dx.doi.org/10.7465/jkdi.2012.23.4.765
한국데이터정보과학회지

교차타당성을 이용한 확률밀도함수의 불연속점 추정의 띠폭 선택

허집[1]

[1] 덕성여자대학교 정보통계학과

접수 2012년 6월 27일, 수정 2012년 7월 14일, 게재확정 2012년 7월 19일

요약

교차타당성은 커널추정량의 평활모수인 띠폭의 선택 방법으로 흔히 활용되고 있다. 연속인 확률밀도함수의 커널추정량의 띠폭 선택으로 널리 쓰이는 교차타당성 방법으로는 최대가능도교차타당성과 더불어 최소제곱교차타당성과 편의교차타당성이 있다. 확률밀도함수가 하나의 불연속점을 가질 때, Huh (2012)는 불연속점 추정을 위한 커널추정량의 띠폭 선택으로 최대가능도교차타당성을 이용한 방법을 제시하였다. 본 연구에서는 Huh (2012)에 의해 최대가능도교차타당성으로 제안된 띠폭선택의 방법과 같이 한쪽방향커널함수를 이용한 최소제곱교차타당성과 편의교차타당성으로 띠폭 선택 방법을 제시하고, 이들 띠폭 선택 방법들과 Huh (2012)의 최대가능도교차타당성을 이용한 띠폭 선택 방법을 모의실험을 통하여 비교연구 하고자 한다.

주요용어: 최대가능도교차타당성, 최소제곱교차타당성, 편의교차타당성.

1. 서론

확률밀도함수가 혹은 그 미분된 함수가 불연속점 (discontinuity point)을 가질 때, Cline과 Hart (1991)는 Schuster (1985)의 대칭화한 데이터를 이용한 확률밀도함수의 경계점 (boundary point) 연구 방법을 일반화하여 커널함수를 이용한 불연속인 확률밀도함수의 추정방법을 제시하였다. Huh (2002)는 확률밀도함수 혹은 그 미분된 함수의 불연속점의 위치와 점프크기 (jump size) 추정량을 한쪽방향커널함수 (one-sided kernel function)에 의한 오른쪽과 왼쪽 커널추정량들의 차이를 이용하여 제시하였고, 추정된 불연속점의 위치와 점프크기의 수렴속도와 극한분포를 구하였다. Otsu와 Xu (2010)는 알려져 있는 불연속점의 위치에서 점프크기 추정량을 회귀모형에서 흔히 쓰이는 비모수적 방법인 국소선형추정량 (local linear estimator)을 이용하여 제시하였다. 확률밀도함수의 토대 (support)를 일정한 간격으로 구간을 나누어 각 구간의 빈도를 한쪽방향커널함수가중을 이용한 국소선형적합으로 불연속점 위치의 오른쪽 추정량과 왼쪽 추정량의 차이를 점프크기 추정량으로 제시하였다.

커널함수의 선택보다 평활모수인 띠폭의 선택이 커널추정량의 정도 (precision)에 더 큰 영향을 준다는 것이 익히 알려져 있다. 커널추정량의 띠폭 선택에서 널리 쓰이고 있는 방법은 교차타당성 (cross-validation)이다. 회귀함수가 불연속점을 가질 때, 커널함수를 이용한 불연속점의 추정에서의 띠폭 선택 연구는 Gijbels와 Goderniaux (2004a, 2004b, 2005)에 의하여 연구되었다. Huh (2010b,

[1] (132-714) 서울특별시 도봉구 쌍문동 419번지, 덕성여자대학교 정보통계학과, 부교수.
E-mail: jhuh@duksung.ac.kr

2011)는 회귀모형의 분산함수가 불연속점을 가질 때 로그분산함수를 이용한 불연속점의 추정과 일반화선형모형의 회귀함수의 불연속점 추정에 쓰이는 띠폭 선택에 대한 방법을 Hart와 Yi (1998)의 한쪽방향교차타당성 (one-sided cross-validation)을 이용하여 제시하였다. 확률밀도함수가 불연속점을 가질 때, Huh (2012)는 연속인 확률밀도함수의 커널추정량의 띠폭 선택에 쓰인 최대가능도교차타당성 (maximum likelihood cross-validation; MLCV)으로 Huh (2002)의 불연속점 추정량들의 띠폭 선택 방법을 제안하였다.

Härdle (1991)은 확률밀도함수의 커널추정량의 띠폭 선택을 위한 MLCV는 컴팩트토대 (compact support)를 가지는 커널을 사용하는 경우, 어떤 자료가 다른 자료들로부터 띠폭 이상 떨어져 있을 때 그 자료에서의 커널추정량이 0이 되어 로그가능도교차타당성의 값이 음으로 무한대가 될 수 있는 단점이 있다고 설명하였다. 교차타당성을 이용한 또 다른 방법으로는 최소제곱교차타당성 (least squares cross-validation; LSCV)과 편의교차타당성 (biased cross-validation; BCV)이 널리 쓰이고 있다. Scott과 Terrell (1987)에 의해 불편교차타당성 (unbiased cross-validation)이라 불리어진 LSCV는 확률밀도함수의 커널추정량의 적분제곱오차 (integrated squared error; ISE)를 최소로 하려는 목적으로 교차타당성을 이용한 점수함수를 고려하여 이 점수함수를 최소로 하는 띠폭을 선택하는 방법이다. Scott과 Terrell은 확률밀도함수의 커널추정량의 극한평균적분제곱오차 (asymptotic mean integrated squared error; AMISE)를 최소로 하려는 목적으로 교차타당성을 이용한 점수함수를 제안하여 띠폭의 선택 방법을 제시하였다. 이 방법은 최소제곱교차타당성 방법이 가지고 있는 큰 변동성의 단점을 보완한 것이지만, 띠폭의 과대평활 (oversmooth)을 만들어내는 단점을 가지게 된다.

본 연구에서는 확률밀도함수가 하나의 불연속점을 가지는 경우에, Huh (2002)가 제안한 불연속점의 위치와 점프크기의 커널추정량의 띠폭 선택 방법으로 LSCV와 BCV를 이용하여 제안하고자 한다. 이들 방법들과 Huh (2012)의 MLCV 방법을 모의실험을 통하여 비교하고자 한다. 2절에서는 Huh (2002)가 제안한 확률밀도함수의 불연속점 추정에 대한 소개와 LSCV와 BCV를 이용한 띠폭 선택 방법을 제시한다. 3절에서는 2절에서 소개한 띠폭 선택 방법들과 MLCV를 이용한 Huh (2012)의 띠폭 선택 방법을 모의실험 결과를 통하여 비교한다.

2. 교차타당성을 이용한 띠폭 선택들

확률밀도함수 f로부터의 랜덤표본을 $\{X_i : i = 1, \cdots, n\}$이라 하자. 확률밀도함수의 비모수적 추정에서 일반적으로 f의 부드러움 (smoothness)의 정도를 최소한 두 번 미분 가능하다고 가정한다. 하지만 실제 구현에서는 확률밀도함수가 불연속점을 가질 수 있으며, 이러한 불연속점을 고려하지 않고 추정할 경우에는 그 불연속점 주변에서 비모수적 추정량의 정도가 떨어지게 된다. 이는 확률밀도함수가 불연속점에서 미분이 존재하지 않아 추정량의 편의가 발생하게 되고 이로 인해 일치추정량 (consistent estimator)이 되지 않기 때문에 생기는 현상이다. 본 연구에서는 토대 [0,1]을 가지는 확률밀도함수 f가 미지의 한 점 $\tau \in Q$에서 불연속점을 가진다고 가정하자. 여기서 Q는 토대 $[0, 1]$에 포함되는 닫힌구간이다. 즉, 확률밀도함수 f는 $(x-\tau)(y-\tau) > 0$을 만족하는 모든 $x, y \in Q$에 대하여

$$|f(x) - f(y)| \leq L|x - y|$$

을 만족하는 양의 상수 L이 존재한다. 불연속점 τ에서 확률밀도함수의 우극한과 좌극한을 각각 $f_+(\tau) = \lim_{y\to\tau+} f(y)$와 $f_-(\tau) = \lim_{y\to\tau-} f(y)$라 하자. 불연속점의 점프크기 Δ는 불연속점 τ에서 확률밀도함수의 우극한과 좌극한의 차이이며 다음

$$\Delta = f_+(\tau) - f_-(\tau) \tag{2.1}$$

과 같이 표현된다. 불연속점이 존재한다면 식 (2.1)에서 $|\Delta| > 0$이고, 그렇지 않다면 $\Delta = 0$이다.

확률밀도함수가 불연속점을 가질 때 Huh (2002)는 미지의 불연속점의 위치와 점프크기를 추정하기 위하여 한쪽방향커널함수를 이용하여 어떤 점 x에서 확률밀도함수의 오른쪽 추정량과 왼쪽 추정량을 다음

$$\widehat{f}_+(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),\ \widehat{f}_-(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{2.2}$$

과 같이 각각 정의하였다. 여기서 사용된 h는 평활모수인 띠폭이며, K는 한쪽방향커널함수이다. 한쪽방향커널함수 K의 토대는 $[0,1]$이며 $\int_0^1 K(u)du = 1$, $K(0) > 0$이고 임의의 $0 < u \leq 1$에 대하여 $K(u) \geq 0$을 만족한다.

Huh (2002)는 식 (2.2)의 두 추정량을 이용하여 어떤 점 x에서 점프크기추정량은 $\widehat{\Delta}(x) = \widehat{f}_+(x) - \widehat{f}_-(x)$과 같이 정의하였고, 이들 점프크기추정량들 중 그 절대값의 크기가 가장 큰 점프크기추정량의 위치를 불연속점의 위치추정량으로 다음

$$\widehat{\tau} = \inf\{z \in Q : \widehat{\Delta}(z) = \sup_{x\in Q}\widehat{\Delta}(x)\} \tag{2.3}$$

과 같이 제시하였다. 한편, 추정된 위치추정량 $\widehat{\tau}$에서 점프크기추정량인

$$\widehat{\Delta}(\widehat{\tau}) = \widehat{f}_+(\widehat{\tau}) - \widehat{f}_-(\widehat{\tau}) \tag{2.4}$$

을 점프크기 Δ의 추정량으로 정의하였다. 또한, Huh (2002)는 제안한 위치추정량과 점프크기추정량인 식 (2.3)과 (2.4)의 수렴속도와 극한분포를 보였다.

Huh (2012)는 불연속점의 추정량들 (2.3)과 (2.4)를 위한 띠폭 선택을 연속인 확률밀도함수의 커널추정량의 띠폭 선택에 쓰인 MLCV를 이용하여 제시하였다. 먼저, 한쪽방향커널함수를 이용하여 X_i를 제거한 $f(X_i)$의 오른쪽과 왼쪽 커널추정량을 각각 다음과 같이

$$\widehat{f}_{h,i,+}(X_i) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{X_j - X_i}{h}\right),\ \widehat{f}_{h,i,-}(X_i) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{X_i - X_j}{h}\right) \tag{2.5}$$

정의하고, 다음의 교차타당성

$$CV_{ML}(h) = CV^{+}_{ML}(h) + CV^{-}_{ML}(h) \tag{2.6}$$

을 생각하였다. 여기서 어떤 주어진 양수 δ에 대하여

$$CV^{\pm}_{ML}(h) = \frac{1}{n^{\delta}}\sum_{i\in I_\delta} \log \widehat{f}_{h,i,\pm}(X_i) \tag{2.7}$$

이고 $I_\delta = \{i : X_i \in [\delta, 1-\delta]\}$이며 n_δ는 I_δ의 원소의 수이다. Müller (1992)는 불연속점 추정을 위한 띠폭은 연속인 회귀함수의 추정을 위한 띠폭보다 상대적으로 작은 것을 추천하였기에 띠폭을 식 (2.6)의 국소최대점 (local maximizer) 중 가장 작은 값으로 제안하였다. 이때 Huh (2012)는 δ를 h로 제안하였다. Härdle (1991)은 MLCV가 컴팩트토대의 커널을 사용하고 한 자료가 다른 자료들로부터 띠폭 이상 떨어져 있는 경우에 그 자료에서의 커널추정량이 0이 되어 식 (2.7)이 음으로 무한대가 되는 단점이 있다고 설명하였다.가능도함수를 최대로 하는 목적으로 사용된 교차타당성이 MLCV라면, 다음의

$$ISE(\widehat{f}_h) = \int \left\{\widehat{f}_h(x) - f(x)\right\}^2 dx = \int \left\{\widehat{f}_h(x)\right\}^2 dx - 2\int \widehat{f}_h(x)f(x)dx + \int \{f(x)\}^2 dx \tag{2.8}$$

를 최소로 하는 목적으로 제안된 교차타당성은 LSCV이다. 여기서 $\widehat{f}_h$는 토대가 $[-1,1]$인 커널함수를 이용한 확률밀도함수의 커널추정량이다. 불연속점의 커널추정량을 위하여 LSCV를 이용하여 다음과 같이 띠폭의 선택을 제안하고자 한다. 먼저, Huh (2002)가 제안한 한쪽방향커널함수 K를 이용한 x의 오른쪽 커널추정량과 왼쪽 커널추정량을 각각 다음과 같이

$$\widehat{f}_{h,+}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \widehat{f}_{h,-}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

생각하고, 다음의 교차타당성

$$CV_{LS}(h) = CV_{LS}^{+}(h) + CV_{LS}^{-}(h) \tag{2.9}$$

을 제안한다. 여기서 $CV_{LS}^{\pm}(h)$는 어떤 주어진 양수 δ에 대하여

$$CV_{LS}^{\pm}(h) = \int_{\delta}^{1-\delta} \left\{\widehat{f}_{h,\pm}(x)\right\}^2 dx - \frac{2}{n_\delta}\sum_{i\in I_\delta} \widehat{f}_{h,i,\pm}(X_i) \tag{2.10}$$

이다. 식 (2.8)에서 띠폭 h와 무관한 마지막 항을 제외한 후 교차타당성 방법을 적용하여 식 (2.10)를 고려한 것이다. 식 (2.6)의 국소최대점 중 최소의 값을 띠폭으로 선택한 것처럼, 식 (2.9)의 국소최소점 (local minimizer) 중 가장 작은 값으로 불연속점 추정을 위한 띠폭으로 제안한다. 한편, LSCV는 ISE를 최소로 하는 목적으로 제안된 교차타당성이기에 LSCV의 변동성이 크다는 것이 익히 알려져 있다. 이러한 LSCV의 단점을 보완하기 위하여 Scott과 Terrell (1987)은 다음의

$$AMISE(\widehat{f}_h) = \frac{1}{nh}\int \{K(x)\}^2 dx + \frac{h^4}{4}\left\{\int x^2 K(x)dx\right\}^2 \int \{f''(x)\}^2 dx \tag{2.11}$$

를 최소로 하는 목적으로 교차타당성을 이용하여 BCV를 제안하였다. 두 번 미분한 한쪽방향커널함수를 이용하여 어떤 점 x의 두 번 미분된 확률밀도함수의 오른쪽 커널추정량과 왼쪽 커널추정량을 각각 다음과 같이

$$\widehat{f}''_{h,+}(x) = \frac{1}{nh^3}\sum_{i=1}^{n} K''\left(\frac{X_i - x}{h}\right), \ \widehat{f}''_{h,-}(x) = \frac{1}{nh^3}\sum_{i=1}^{n} K''\left(\frac{x - X_i}{h}\right)$$

생각하고, 다음의 교차타당성

$$CV_B(h) = CV_B^{+}(h) + CV_B^{-}(h) \tag{2.12}$$

을 제안한다. 여기서 $CV_B^{\pm}(h)$는 어떤 주어진 양수 δ에 대하여

$$CV_B^{\pm}(h) = \frac{1}{n_\delta h}\int \{K(x)\}^2 dx + \frac{h^4}{4}\left\{\int x^2 K(x)dx\right\}^2 \left[\int_{\delta}^{1-\delta} \{\widehat{f}''_{h,\pm}(x)\}^2 dx - \frac{1}{n_\delta h^5}\int \{K''(x)\}^2 dx\right] \tag{2.13}$$

이다. 두 번 이상 미분한 확률밀도함수가 연속일 때, 두 번 미분하여 연속인 커널함수로 추정한 f''의 커널추정량으로 식 (2.11)에서 $\int\{f''(x)\}^2 dx$를 추정한 경우, Härdle (1991)은 그 추정량이 $(1/nh^5)\int\{K''(x)\}^2 dx$의 편의가 있음을 설명하고 있다. 이러한 편의를 수정하여 BCV가 제안되어졌으며, 본 연구에서도 이러한 편의를 수정하는 방법을 이용하여 식 (2.13)을 고려한 것이다. 앞서 제안한 LSCV를 이용한 식 (2.9)의 국소최소점 중 가장 작은 값으로 띠폭을 선택한 것처럼, 식 (2.12)의 국소최소점 중 가장 작은 값으로 불연속점 추정을 위한 띠폭으로 제안한다. 연속인 확률밀도함수의 추정을 위한 띠폭 선택에서 BCV는 LSCV에 비해 일반적으로 큰 띠폭을 선택하여 과대평활하는 경향이 있는 것으로 알려져 있다.

3. 모의실험

2절에서 LSCV와 BCV를 이용하여 제안한 띠폭 선택 두 가지 방법과 MLCV를 이용한 Huh (2012)의 띠폭 선택 방법을 모의실험을 통하여 비교해보고자 한다. 확률밀도함수의 부드러움의 정도를 고려하여 토대가 $[0,1]$인 두 가지 유형의 확률밀도함수를 모의실험 모형으로 생각하였다. BCV에서는 Härdle (1991)이 언급한 것처럼 두 번 미분하여 연속인 커널함수가 쓰인다. 이를 위하여 한쪽방향커널함수는 토대 $[0,1]$을 기반으로 이차커널 (quadratic kernel)을 이용하여 만든

$$K(x) = \frac{15}{8}(1-x^2)^2 1_{[0\le x\le 1]}$$

을 선택하고, 세 가지 띠폭 선택 방법에 사용하였다. 불연속점의 위치를 추정하기 위하여 $x_k = k/100,\ (k=1,\cdots,100)$에서 점프크기 $\widehat{\Delta}(x_k)$를 계산하여 구간 Q내에서 $|\widehat{\Delta}(x_k)|$를 최대로 하는 점을 불연속점의 위치로 추정하였다. Huh (2012)는 Müller (1992)와 Huh (2010b)가 제안한 것처럼 $\delta = h$로 하여 구간 Q를 $[h, 1-h]$로 선택하였다. 띠폭에 의존하는 구간 Q를 고려하면, 큰 띠폭의 경우에 Q의 구간의 폭이 줄어들게 되어 불연속점을 포함하지 않게 될 수도 있다. 본 모의실험에서는 δ=0.1로 하여 Q를 [0.1, 0.9]로 시행하였다. 불연속점의 추정의 정도를 보기 위하여 각 경우에서 표본의 수 n을 100, 200, 400으로 하여 각각 1000번 반복하였다.

먼저, 불연속점을 제외한 곳에서 두 번 이상 미분된 함수가 연속인 경우로서, 이중지수분포를 이용한 토대가 $[0,1]$이 되는 다음의 확률밀도함수

$$f_1(x) = p\left\{\frac{\lambda_1}{2}\exp(-\lambda_1|x-\tau|)1_{[0\le x<\tau]} + \frac{\lambda_2}{2}\exp(-\lambda_2|x-\tau|)1_{[\tau\le x\le 1]}\right\}$$

를 생각한다. 여기서 $\lambda_1 = 0.1$이고 λ_2는 1과 2인 두 가지 경우를 생각하고, 불연속점 위치 τ는 각각 0.5와 0.75를 고려하자. 양의 상수 p는 각 확률밀도함수의 조건을 만족하게 하는 값이다. 즉, $p = 1/\left(\int_0^\tau(\lambda_1/2)\exp(\lambda_1(x-\tau))dx + \int_\tau^1(\lambda_2/2)\exp(-\lambda_2(x-\tau))dx\right)$이다. 각 확률밀도함수의 점프크기는 $\Delta = p(\lambda_2-\lambda_1)/2$이다. 고려한 네 가지 확률밀도함수에 대한 상수 p와 점프크기 Δ는 아래의 Table 3.1과 같다. Table 3.1에 의하면 고려한 확률밀도함수 모형에서 불연속점이 0.5인 경우보다 0.75인 경우의 점프크기가 크다. Figure 3.1은 $\lambda_2 = 1$이며 불연속점이 0.5인 경우의 확률밀도함수를 보여주고 있다.

Table 3.1 Jump sizes of the density f_1

τ	λ_2	p	Δ
0.5	1	4.522432	2.035094
	2	2.937327	2.790461
0.75	1	6.815338	3.066902
	2	4.294372	4.079653

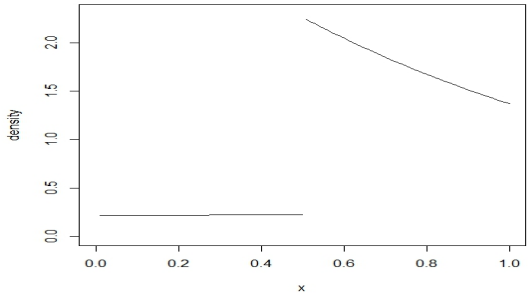

Figure 3.1 Truncated double exponential density in the case of $\lambda_1 = 0.1$ and $\lambda_2 = 1$ having a discontinuity point at 0.5

먼저, $\tau = 0.5$인 경우에 각 모형과 표본에서 식 (2.6), (2.9), (2.12)의 MLCV, LSCV, BCV에 의해 선택된 $\widehat{h}$들의 평균이 Table 3.2에 제시되었다. 또한 선택된 띠폭을 이용하여 추정된 $\widehat{\tau}$와 $\widehat{\Delta}(\widehat{\tau})$의 평균과 평균제곱오차의 몬테카를로 추정치 (Monte Carlo estimates of the mean squared error; MSE)들이 Table 3.2에 제시되었다. 각 결과의 괄호 안은 표준오차들이다. Table 3.3은 불연속점 $\tau = 0.75$인 경우에 각 모형과 표본으로부터의 불연속점의 추정에 대한 모의실험 결과를 보여준다.

불연속점이 0.5인 Table 3.2에서 Huh (2012)가 제안한 MLCV와 2절에서 제안된 LSCV에 의해 선택된 띠폭의 평균들이 유사한 결과를 보여주고 있지만, 불연속점이 0.75인 Table 3.3에서는 MLCV에 의해 선택된 띠폭의 평균들이 큰 경향을 띠고 있다. 한편, 모든 모의실험 모형에서 BCV에 의해 선택되어진 띠폭의 평균들은 MLCV와 LSCV에 의해 선택되어진 띠폭의 평균보다 큰 경향을 띠고 있음을 알 수 있다. 이는 연속인 확률밀도함수의 추정을 위한 띠폭 선택에서 BCV는 LSCV에 의해 선택된 띠폭들 보다 일반적으로 큰 띠폭을 선택하여 과대평활하는 경향이 있는 것과 동일한 결과이다. BCV에 의해 선택된 띠폭들을 이용한 불연속점의 위치 추정 $\widehat{\tau}$의 MSE가 MLCV와 LSCV에 의해 선택된 띠폭을 이용한 $\widehat{\tau}$의 MSE보다 작음이 Table 3.2와 3.3에 나타나고 있다. 또한, BCV에 의해 선택되어진 띠폭들을 사용하는 경우에 MLCV와 LSCV에 의해 선택되어진 띠폭들을 사용한 경우 보다 점프크기 추정 $\widehat{\Delta}(\widehat{\tau})$의 MSE도 모든 모형에서 작음을 알 수 있다. 불연속점 τ=0.5의 경우인 Table 3.2에서는 MLCV와 LSCV에 의해 선택된 띠폭들의 평균들이 유사하여, 두 방법에 의한 $\widehat{\tau}$의 MSE와 $\widehat{\Delta}(\widehat{\tau})$의 MSE도 유사한 결과를 나타내고 있다. 불연속점의 점프크기 Δ가 상대적으로 큰 τ=0.75인 경우에는 MLCV에 의해 선택되어진 띠폭들을 사용하는 경우에 LSCV에 의해 선택되어진 띠폭들을 사용하는 경우보다 $\widehat{\tau}$의 MSE와 $\widehat{\Delta}(\widehat{\tau})$의 MSE가 작은 경향을 보이고 있다.

회귀함수의 불연속점 위치추정량의 수렴속도는 Loader (1996), Huh와 Carrière (2002), Huh와 Park (2004), Huh (2010b)에 의해 n^{-1}이 됨이 밝혀졌다. Huh (2002)는 확률밀도함수의 불연속점의 위치추정량의 수렴속도가 n^{-1}이 됨이 보였고, 점프크기가 클수록 불연속점 추정량의 점근분산이 작아져서 추정의 정도가 좋음을 밝혔다. 이러한 현상이 Table 3.2과 3.3에서 나타나고 있다. 각 Table에서 위치추정량 $\widehat{\tau}$의 MSE가 표본의 수가 증가함으로써 급속히 작아지고 있다. 또한 점프크기가 커짐으로써 $\widehat{\tau}$의 MSE가 작아지는 경향이 있음을 알 수 있다. 점프크기추정량 $\widehat{\Delta}(\widehat{\tau})$의 MSE 또한 각 경우에서 표본의 수가 증가함에 따라 작아지고 있음을 알 수 있다.

Table 3.2 The Monte Carlo estimates of the MSEs and the averages for the discontinuity point estimators in the case of f_1 with τ=0.5

λ_2	n	CV	Average of $\widehat{h}$	Average of $\widehat{\tau}$	MSE of $\widehat{\tau}$	Average of $\widehat{\Delta}(\widehat{\tau})$	MSE of $\widehat{\Delta}(\widehat{\tau})$
		MLCV	0.112460	0.542880 (0.002831)	0.009855 (0.000874)	1.731917 (0.043871)	2.016620 (0.166966)
	100	LSCV	0.123800	0.535740 (0.002581)	(0.007938) (0.000796)	1.790413 (0.041177)	1.755435 (0.158737)
		BCV	0.315110	0.510700 (0.001494)	0.002348 (0.000548)	1.804126 (0.015737)	0.300991 (0.049219)
		MLCV	0.086620	0.515140 (0.001704)	0.003133 (0.000508)	1.897702 (0.026667)	0.729993 (0.093848)
1	200	LSCV	0.088800	0.515010 (0.001619)	0.002845 (0.000449)	1.888315 (0.027356)	0.769910 (0.099804)
		BCV	0.224910	0.502860 (0.000340)	0.000124 (0.000050)	1.876632 (0.008292)	0.093859 (0.012045)
		MLCV	0.060560	0.504470 (0.000890)	0.000812 (0.000252)	1.976765 (0.018171)	0.333583 (0.052963)
	400	LSCV	0.063680	0.504860 (0.000936)	0.000900 (0.000257)	1.951552 (0.019134)	0.373079 (0.057792)
		BCV	0.166600	0.500900 (0.000113)	0.000014 (0.000002)	1.912542 (0.006823)	0.061575 (0.002457)
		MLCV	0.097390	0.521990 (0.002000)	0.004484 (0.000591)	2.467622 (0.045631)	2.186391 (0.221180)
	100	LSCV	0.099450	0.521200 (0.001920)	0.004136 (0.000550)	2.428968 (0.048783)	2.510440 (0.252067)
		BCV	0.265570	0.503690 (0.000497)	0.000261 (0.000154)	2.369434 (0.013086)	0.348505 (0.036785)
		MLCV	0.080680	0.505470 (0.000889)	0.000821 (0.000224)	2.621245 (0.025068)	0.657015 (0.102310)
2	200	LSCV	0.072110	0.505800 (0.000871)	0.000792 (0.000195)	2.619183 (0.027905)	0.808024 (0.121016)
		BCV	0.190580	0.500990 (0.000117)	0.000015 (0.000002)	2.458482 (0.009417)	0.198885 (0.006774)
		MLCV	0.063610	0.500410 (0.000090)	0.000008 (0.000004)	2.712761 (0.013203)	0.180359 (0.021454)
	400	LSCV	0.051870	0.501150 (0.000336)	0.000115 (0.000057)	2.710744 (0.017517)	0.313195 (0.051730)
		BCV	0.135000	0.500210 (0.000047)	0.000002 (0.000001)	2.556770 (0.008086)	0.120000 (0.004381)

Table 3.3 The Monte Carlo estimates of the MSEs and the averages for the discontinuity point estimators in the case of f_1 with τ=0.75

λ_2	n	CV	Average of $\widehat{h}$	Average of $\widehat{\tau}$	MSE of $\widehat{\tau}$	Average of $\widehat{\Delta}(\widehat{\tau})$	MSE of $\widehat{\Delta}(\widehat{\tau})$
		MLCV	0.103730	0.757640 (0.000689)	0.000533 (0.000073)	2.978471 (0.030863)	0.960355 (0.148138)
	100	LSCV	0.077220	0.763870 (0.000978)	0.001149 (0.000110)	2.846047 (0.052164)	2.769811 (0.308279)
		BCV	0.202990	0.753430 (0.000364)	0.000144 (0.000022)	2.853746 (0.013048)	0.215687 (0.009090)
		MLCV	0.073840	0.753050 (0.000388)	0.000160 (0.000041)	2.996443 (0.020899)	0.441723 (0.064335)
1	200	LSCV	0.055730	0.755920 (0.000632)	0.000435 (0.000071)	2.957322 (0.032252)	1.052177 (0.148384)
		BCV	0.150070	0.751570 (0.000182)	0.000036 (0.000009)	2.886556 (0.011575)	0.166501 (0.006698)
		MLCV	0.049070	0.750840 (0.000197)	0.000039 (0.000021)	3.015570 (0.017095)	0.294865 (0.041882)
	400	LSCV	0.042380	0.751060 (0.000226)	0.000052 (0.000023)	3.013709 (0.019332)	0.376542 (0.058015)
		BCV	0.107950	0.750380 (0.000075)	0.000006 (0.000002)	2.941523 (0.010160)	0.118947 (0.005102)
		MLCV	0.087890	0.756250 (0.000631)	0.000437 (0.000069)	3.799385 (0.043030)	1.930173 (0.281906)
	100	LSCV	0.064230	0.759660 (0.000818)	0.000762 (0.000091)	3.747004 (0.059222)	3.617902 (0.426828)
		BCV	0.172470	0.751810 (0.000194)	0.000041 (0.000007)	3.634533 (0.015215)	0.429637 (0.015406)
		MLCV	0.076830	0.751070 (0.000196)	0.000040 (0.000015)	3.880958 (0.020621)	0.464719 (0.074877)
2	200	LSCV	0.047190	0.752950 (0.000436)	0.000199 (0.000044)	3.894685 (0.036403)	1.359366 (0.208006)
		BCV	0.124720	0.750560 (0.000091)	0.000009 (0.000002)	3.739849 (0.013827)	0.306641 (0.011669)
		MLCV	0.055470	0.750160 (0.000042)	0.000002 (0.000001)	3.940671 (0.015820)	0.269587 (0.011794)
	400	LSCV	0.040280	0.750170 (0.000046)	0.000002 (0.000001)	3.981041 (0.018806)	0.363380 (0.016038)
		BCV	0.089810	0.750110 (0.000033)	0.000001 (0.0000003)	3.826371 (0.012226)	0.213637 (0.008460)

두 번째는 미분된 함수가 0이 되는 균등분포 (uniform distribution)를 다음

$$f_2(x) = \begin{cases} c_1, & x \le 0.5 \\ c_2, & x > 0.5 \end{cases}$$

과 같이 생각한다. 여기서 $(c_1, c_2) = (1.5, 0.5)$와 $(1.75, 0.25)$두 가지를 고려하자. 이 경우에 각각 점프크기는 -1.0과 -1.5이다. 이 분포들은 두 번 미분된 함수가 0이므로 식 (2.13) 안의 $\int \{f(x)\}^2 dx$를 작은 값으로 추정할 것으로 추측된다. Figure 3.2는 $(c_1, c_2) = (1.5, 0.5)$이며 불연속점이 0.5인 경우의 확률밀도함수를 보여주고 있다.

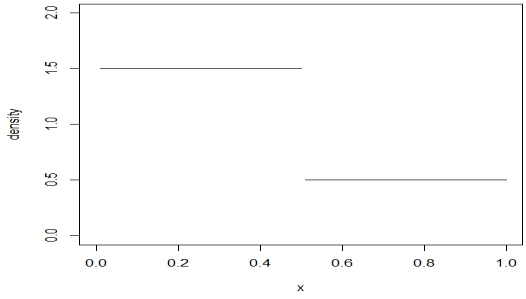

Figure 3.2 Uniform density having a discontinuity point at 0.5

Table 3.4에 의하면, 세 방법에 의해 선택되어진 띠폭들의 평균들이 Table 3.2와 3.3에 있는 선택되어진 띠폭들의 평균들보다 크게 나타나고 있다. 이는 연속인 확률밀도함수에서의 결과인 확률밀도함수의 변화가 작을수록 선택되어지는 띠폭이 일반적으로 크다는 것과 동일한 결과이다. 특히, BCV는 Table 3.2와 3.3과 같이 큰 띠폭을 선택하여 과대평활하고 있다. 표본의 수가 $n = 100$, 200이며 $(c_1, c_2) = (1.5, 0.5)$의 경우, Table 3.2와 3.3과는 다르게, BCV에 의해 선택된 띠폭의 $\hat{\tau}$의 MSE가 MLCV와 LSCV에 의해 선택된 띠폭들의 $\hat{\tau}$의 MSE 보다 조금 큼을 알 수 있다. 이는 선택된 띠폭이 상당히 커서 불연속점 추정을 위하여 넓은 영역의 표본들을 이용하여 확률밀도함수를 추정하기 때문이다. 띠폭을 상당히 크게 선택하는, 표본의 수가 $n = 100$, 200이며 $(c_1, c_2) = (1.5, 0.5)$의 경우를 제외하고는 BCV의 불연속점의 추정 정도는 Table 3.2와 3.3처럼 MLCV와 LSCV의 불연속점의 추정 정도에 비하여 좋음을 알 수 있다. 불연속점의 점프크기가 큰 $(c_1, c_2) = (1.75, 0.25)$일 때, 세 방법 모두 점프크기가 작은 $(c_1, c_2) = (1.5, 0.5)$경우 보다 불연속점의 추정 정도가 우수하다. 또한, Table 3.4는 모든 경우에 MLCV의 불연속점의 추정 정도가 LSCV의 불연속점의 추정 정도보다 우수함을 보여준다.

Table 3.4 The Monte Carlo estimates of the MSEs and the averages for the discontinuity point estimators in the case of f_2

(c_1, c_2)	n	CV	Average of $\widehat{h}$	Average of $\hat{\tau}$	MSE of $\hat{\tau}$	Average of $\widehat{\Delta}(\hat{\tau})$	MSE of $\widehat{\Delta}(\hat{\tau})$
$(1.5, 0.5)$	100	MLCV	0.181440	0.426300 (0.003833)	0.020124 (0.001201)	-1.005493 (0.026741)	0.715112 (0.055680)
		LSCV	0.173530	0.413970 (0.004107)	0.024265 (0.001351)	-0.957720 (0.036366)	1.324256 (0.088588)
		BCV	0.597450	0.381860 (0.005781)	0.047380 (0.002138)	-0.412557 (0.031817)	1.357406 (0.063582)
	200	MLCV	0.14696	0.450420 (0.003064)	0.011849 (0.000950)	-0.975172 (0.020152)	0.406734 (0.037511)
		LSCV	0.136270	0.440320 (0.003438)	0.015384 (0.001101)	-0.915046 (0.026900)	0.730836 (0.059787)
		BCV	0.393100	0.443830 (0.004048)	0.019537 (0.001559)	-0.794118 (0.021244)	0.493691 (0.039853)
	400	MLCV	0.117080	0.477990 (0.002068)	0.004761 (0.000629)	-0.993430 (0.012707)	0.161522 (0.021226)
		LSCV	0.103350	0.464800 (0.002621)	0.008111 (0.000810)	-0.922675 (0.019616)	0.390752 (0.038778)
		BCV	0.260180	0.492360 (0.001131)	0.001337 (0.000402)	-0.990203 (0.006858)	0.047135 (0.009785)
$(1.75, 0.25)$	100	MLCV	0.155960	0.447720 (0.003104)	0.012366 (0.000996)	-1.458339 (0.031609)	1.000859 (0.093227)
		LSCV	0.128790	0.428970 (0.003489)	0.017216 (0.001124)	-1.306472 (0.045792)	2.134354 (0.159516)
		BCV	0.423800	0.478790 (0.002041)	0.004615 (0.000736)	-1.453677 (0.016088)	0.260985 (0.044361)
	200	MLCV	0.122570	0.476310 (0.002168)	0.005261 (0.000686)	-1.485030 (0.020071)	0.403053 (0.052349)
		LSCV	0.096220	0.454110 (0.003026)	0.011261 (0.000965)	-1.376968 (0.032528)	1.073213 (0.097404)
		BCV	0.290730	0.493300 (0.000835)	0.000743 (0.000264)	-1.496396 (0.007847)	0.061593 (0.014205)
	400	MLCV	0.087630	0.490220 (0.001385)	0.002015 (0.000405)	-1.515496 (0.012145)	0.147737 (0.023620)
		LSCV	0.070070	0.477820 (0.002168)	0.005194 (0.000657)	-1.433313 (0.022262)	0.500034 (0.061094)
		BCV	0.197980	0.498160 (0.000211)	0.000048 (0.000011)	-1.499070 (0.005575)	0.031079 (0.001436)

세 방법에 의해 선택된 띠폭들로 추정된 불연속점의 위치추정량 $\hat{\tau}$의 분포를 보기 위하여, 첫 번째

모형인 f_1의 각 경우에서 표본의 수가 200일 때 추정된 불연속점의 위치 x_k의 지표 k와 해당되는 빈도를 τ=0.5, 0.75인 경우에 각각 Table 3.5와 3.6에 제시하였다. Table 3.2과 3.3의 결과에서 설명하였듯이, BCV에 의한 $\hat{\tau}$의 분포가 MLCV와 LSCV에 의한 $\hat{\tau}$의 분포들의 결과보다 우수하다는 것을 알 수 있다. 불연속점이 τ=0.5인 경우에 MLCV와 LSCV에 의한 $\hat{\tau}$의 분포는 유사하며, 점프크기가 큰 τ=0.75인 경우에는 MLCV에 의한 $\hat{\tau}$의 분포가 LSCV에 의한 $\hat{\tau}$의 분포의 결과보다 우수하다. Table 3.5와 3.6은 f_1모형에서는 상대적으로 큰 띠폭을 선택하는 방법들이 불연속점의 위치와 점프크기의 추정의 정도가 우수함을 보여주고 있다.

Table 3.5 Discontinuity point identification frequency in the case of f_1 with τ=0.5 and n=200

Δ	CV	k																		
		47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	>66
1	MLCV	-	1	12	739	111	42	23	8	6	2	4	-	2	3	3	4	1	2	2	37
	LSCV	-	1	12	723	126	41	21	9	4	4	5	-	1	6	3	5	1	1	2	35
	BCV	1	2	20	792	122	41	13	1	4	-	2	-	-	1	-	-	-	-	-	1
2	MLCV	-	-	3	859	82	22	5	3	2	2	2	-	2	3	1	2	-	1	-	11
	LSCV	-	-	3	850	85	23	7	4	-	3	3	-	1	4	1	1	1	2	1	11
	BCV	-	-	3	912	71	12	1	1	-	-	-	-	-	-	-	-	-	-	-	-

Table 3.6 Discontinuity point identification frequency in the case of f_1 with τ=0.75 and n=200

Δ	CV	k															
		74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89
1	MLCV	4	852	91	24	13	4	2	-	2	1	2	-	1	-	2	2
	LSCV	3	809	101	28	17	4	4	1	3	6	5	4	2	3	7	3
	BCV	6	881	83	19	7	3	-	-	-	-	1	-	-	-	-	-
2	MLCV	-	938	46	9	2	1	-	1	-	1	1	1	-	-	-	-
	LSCV	-	903	55	13	5	2	3	2	2	4	3	4	1	1	-	2
	BCV	-	955	37	6	1	1	-	-	-	-	-	-	-	-	-	-

4. 논의

본 연구는 확률밀도함수가 하나의 불연속점을 가지고 있을 때, 커널추정량을 이용한 불연속점의 위치와 점프크기를 추정하기 위한 띠폭의 선택 방법들을 교차타당성을 이용하여 제시하고 모의실험을 통하여 비교·연구하였다. 모의실험 결과에 의하면 BCV에 의해 선택된 띠폭은 MLCV와 LSCV에 의해 선택된 띠폭들보다 큰 경향이 있다는 것을 알 수 있다. 이는 연속인 확률밀도함수의 커널추정량을 위한 띠폭 선택의 경향과 동일한 것이다. 3절의 결과에서는 큰 띠폭은 불연속점의 추정 정도를 높이고 있다. 하지만, 확률밀도함수의 변화가 작은 경우에는 BCV는 매우 큰 띠폭을 선택하여 불연속점 추정을 위하여 넓은 영역의 표본들을 이용하여 확률밀도함수를 추정하기 때문에 불연속점의 추정 정도가 떨어짐을 알 수 있었다. 이 경우에는 MLCV가 LSCV와 BCV에 비해 불연속점의 추정 정도가 안정적이라 할 수 있다.

Jose와 Ismail (1999)이 언급하였던 것처럼 근접해 있는 두 개의 불연속점들 사이의 거리들은 2 h 보다 크다는 가정 하에, 확률밀도함수가 두 개 이상의 불연속점을 가지는 경우에도 선택된 띠폭을 이용하여 구간 Q내의 점프크기 추정량 $\widehat{\Delta}(x)$의 절대치를 이용하여 순차적으로 불연속점들을 추정할 수 있다. 만약 선택된 띠폭이 매우 크면 한 불연속점의 위치를 추정하고 두 번째 불연속점의 위치를 추정할 때, 두 개의 불연속점들 사이의 거리들은 2 h보다 크다는 조건을 만족하기 위해서는 두 번째 불연속점의 위치는 추정할 수 없을 수도 있다. 반대로, 선택된 띠폭이 매우 작다면 순차적으로 추정되어진 불연속점의 위치가 서로 이웃하는 결과를 초래할 수도 있다. 즉, 두 개 이상의 불연속점의 추정은

선택된 띠폭의 크기에 따라 추정된 불연속점의 수가 달라질 수도 있고, 하나의 불연속점 주변에서 여러 불연속점을 추정할 수도 있게 된다. 이러한 점을 고려하여 불연속점이 두 개 이상인 경우에 대한 띠폭 선택 방법의 추가 연구가 필요하다고 본다. 한편, Kim 등 (2003)과 Huh (2007, 2010a)는 각각 회귀함수, 분산함수와 일반화선형모형의 회귀함수가 두 개 이상의 불연속점을 가릴 때 Jose와 Ismail (1999)이 언급한 조건 하에서 임의로 주어진 띠폭을 사용하여 불연속점 수를 추정하는 알고리즘을 제안하였다.

참고문헌

Cline, D. B. H. and Hart, J. D. (1991). Kernel estimation of densities with discontinuities or discontinuous derivatives. *Statistics*, **22**, 69-84.

Gijbels, I. and Goderniaux, A. C. (2004a). Bandwidth selection for change point estimation in nonparametric regression. *Technometrics*, **46**, 76-86.

Gijbels, I. and Goderniaux, A. C. (2004b). Bootstrap test for change points in nonparametric regression. *Journal of Nonparametric Statistics*, 16, 591-611.

Gijbels, I. and Goderniaux, A. C. (2005). Data-driven discontinuity detection in derivatives of a regression function. *Communications in Statistics-Theory and Methods*, **33**, 851-871.

Härdle, W. (1991). *Smoothing techniques with implementation in S*, Springer-Verlag, New York.

Hart, J. D. and Yi, S. (1998). One-sided cross-validation. *Journal of the American Statistical Association*, **93**, 620-631.

Huh, J. (2002). Nonparametric discontinuity point estimation in density or density derivatives. *Journal of the Korean Statistical Society*, **31**, 261-276.

Huh, J. (2007). Nonparametric detection algorithm of discontinuity points in the variance function. *Journal of the Korean Data & Information Science Society*, **18**, 669-678.

Huh, J. (2010a). Estimation of the number of discontinuity points based on likelihood. *Journal of the Korean Data & Information Science Society*, **21**, 51-59.

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

Huh, J. (2011). Likelihood based estimation of the log-variance function with a change point. submitted to *Journal of Statistical Planning and Inference.*

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

Huh, J. andCarrière, 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.

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

Kim, J. T., Choi, H. and Huh, J. (2003). Detection of change-points by local linear regression fit. *The Korean Communications in Statistics*, **10**, 31-38.

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.

Otsu, T and Xu, K.-L. (2010). Estimation and inference of discontinuity in density. preprint.

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics-Theory and Methods*, **14**, 1123-1136.

Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, **82**, 1131-1146.

Journal of the Korean Data & Information Science Society
2012, **23**(4), 765–775

http://dx.doi.org/10.7465/jkdi.2012.23.4.765
한국데이터정보과학회지

Bandwidth selections based on cross-validation for estimation of a discontinuity point in density

Jib Huh[1]

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

Received 27 June 2012, revised 14 July 2012, accepted 19 July 2012

Abstract

The cross-validation is a popular method to select bandwidth in all types of kernel estimation. The maximum likelihood cross-validation, the least squares cross-validation and biased cross-validation have been proposed for bandwidth selection in kernel density estimation. In the case that the probability density function has a discontinuity point, Huh (2012) proposed a method of bandwidth selection using the maximum likelihood cross-validation. In this paper, two forms of cross-validation with the one-sided kernel function are proposed for bandwidth selection to estimate the location and jump size of the discontinuity point of density. These methods are motivated by the least squares cross-validation and the biased cross-validation. By simulated examples, the finite sample performances of two proposed methods with the one of Huh (2012) are compared.

Keywords: Biased cross-validation, least squares cross-validation, maximum likelihood cross-validation.

[1] Associate professor, Department of Statistics, Duksung Women's University, Seoul 132-714, Korea. E-mail: jhuh@duksung.ac.kr