Journal of the Korean Data & Information Science Society
2012, **23**(1), 79–87

한국데이터정보과학회지

확률밀도함수의 불연속점 추정을 위한 띠폭 선택[†]

허집[1]

[1]덕성여자대학교 정보통계학과

접수 2011년 11월 19일, 수정 2011년 12월 14일, 게재확정 2012년 1월 2일

요 약

Huh (2002)는 확률밀도함수가 하나의 불연속점을 가질 때, 한쪽방향커널함수를 이용하여 확률밀도함수의 오른쪽과 왼쪽 커널추정량을 제시하여 그 차를 최대로 하는 점을 불연속점의 위치추정량으로 제안하였다. 커널추정량의 평활모수인 띠폭의 선택의 중요함은 익히 알려져 있다. 최대가능도교차타당성은 확률밀도함수의 커널추정량에서 띠폭 선택의 기준으로 널리 쓰여지고 있다. 본 연구에서는 한쪽방향커널함수를 이용한 확률밀도함수의 오른쪽과 왼쪽 커널추정량들의 띠폭의 선택 방법을 Hart와 Yi (1998)의 한쪽방향교차타당성의 방법론을 최대가능도교차타당성에 적용하여 제안하고자 한다. 소표본 모의실험을 통하여 연구결과를 제시하고자 한다.

주요용어: 최대가능도교차타당성, 평활모수, 한쪽방향커널함수.

1. 서론

표본 $\{X_i : i = 1, \cdots, n\}$를 독립이고 동일한 확률밀도함수 f로부터의 표본이라고 하자. 확률밀도함수의 비모수적 추정을 위하여 일반적으로 f의 부드러움의 정도를 최소한 두 번 미분 가능하다고 가정한다. 하지만 실제 구현에서는 확률밀도함수가 불연속점 (discontinuity point)을 가질 수 있으며, 이러한 불연속점을 고려하지 않고 추정할 경우에는 그 불연속점 주변에서 비모수적 추정량의 정도 (precision)가 떨어지게 된다. 이는 확률밀도함수가 불연속점에서 미분이 존재하지 않아 추정량의 편의가 발생하게 되고 이로 인해 일치추정량 (consistent estimator)이 되지 않기 때문에 생기는 현상이다. Müller (1992)와 Huh와 Park (2004)이 이러한 문제점을 회귀모형에서 회귀함수가 불연속점을 가지는 경우에 언급하였고, Huh (2010b)는 일반화선형모형에서 회귀함수가 불연속점을 가지는 경우에 대해 설명하였다.

Cline과 Hart (1991)는 확률밀도함수 혹은 그 미분된 함수가 불연속점을 가질 때 그 함수의 커널추정량을 제시하였다. 그들의 연구는 Schuster (1985)의 대칭화한 데이터 (data symmetrized)를 이용한 확률밀도함수의 경계점 (boundary point) 연구 방법을 일반화 한 것이다. Cline과 Hart는 그들이 제시한 추정량의 평균적분제곱오차 (mean integrated squared error)를 구하였다. Huh (2002)는 확률밀도함수 혹은 그 미분된 함수가 알려져 있지 않은 한 점에서 불연속일 때, 그 불연속점의 위치 (location)와 점프크기 (jump size)를 한쪽방향 (one-sided)커널함수를 이용한 커널추정량으로 추정하였다. 또한 추정된 불연속점의 위치와 점프크기의 수렴속도와 극한분포를 구하였다. Otsu와 Xu (2010)는 알려져 있는 불연속점의 위치에서 점프크기 추정량을 회귀모형에서 흔히 쓰이는 비모수적 방법인 국소선형추정

[†] 본 연구는 덕성여자대학교 2010년도 교내연구비 지원에 의해 수행되었음.
[1] (132-714) 서울특별시 도봉구 쌍문동 419번지, 덕성여자대학교 정보통계학과, 부교수.
E-mail: jhuh@duksung.ac.kr

량을 이용하여 제시하였다. 확률밀도함수의 토대 (support)를 일정한 간격으로 구간 (bin)을 나누어 각 구간의 빈도를 이용하여 커널함수가중을 이용한 국소선형적합으로 불연속점의 위치의 오른쪽 추정량과 왼쪽 추정량의 차이를 점프크기 추정량으로 제시하였다. 생존분석에서 쓰이는 확률밀도함수 혹은 그 미분된 함수의 불연속점의 커널형 추정에 대해서는 Müller와 Wang (1990)에 의해 연구되었다. Lee 등 (2010)은 균등분포의 척도모수 (scale parameter)의 변화에 대한 추정을 연구하였다.

비모수적 추정에서 평활모수의 역할은 익히 알려져 있듯이 매우 중요하다. 회귀함수가 불연속점을 가질 때, 커널함수를 이용한 불연속점의 추정에서의 평활모수인 띠폭 (bandwidth) 선택의 연구는 Gijbels와 Goderniaux (2004a, 2004b, 2005)에 의하여 연구되었다. Huh (2010b, 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 커널추정량의 띠폭 선택에 관한 연구는 현재까지 이루어지지 않았다.

이 논문에서는 확률밀도함수가 불연속점을 가질 때 그 불연속점의 위치와 점프크기의 커널추정량에 쓰이는 띠폭 선택에 대한 연구를 하고자 한다. 연속인 확률밀도함수의 커널추정량의 띠폭 선택에 흔히 쓰이는 최대가능도교차타당성 (maximum likelihood cross-validation)을 Hart와 Yi (1998)가 연구한 한쪽방향교차타당성의 방법에 적용하여 띠폭 선택 방법을 제시하고자 한다. 2절에서는 Huh (2002)가 제안한 확률밀도함수의 불연속점 추정에 대한 소개와 띠폭 선택 방법을 제시한다. 3절에서는 2절에서 소개한 띠폭 선택 방법에 대한 모의실험 결과를 소개하고자 한다.

2. 띠폭 선택

토대 [0,1]을 가지는 확률밀도함수 f가 미지의 점 $\tau \in Q$에서 다음과 같이 불연속점을 가진다고 가정하자. 여기서 Q는 토대 $[0,1]$에 포함되는 닫힌구간이다. 이때 불연속점을 가지는 f는 다음의 가정을 만족한다고 하자.

(A.1) 확률밀도함수 f는 $(x-\tau)(y-\tau) > 0$를 만족하는 모든 x, y에 대하여 다음을 만족하는 양의 상수 L이 존재한다.

$$|f(x) - f(y)| \le L|x-y|.$$

불연속점 τ에서 점프크기는 τ에서의 확률밀도함수의 우극한과 좌극한의 차이로 다음과 같이 표현된다.

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

여기서 $f_+(\tau) = \lim_{y\to\tau+} f(y)$와 $f_-(\tau) = \lim_{y\to\tau-} f(y)$이다. 불연속점이 존재한다면 식 (2.1)에서 $|\Delta| > 0$이고, 그렇지 않다면 $\Delta = 0$이다.

Huh (2002)는 위 가정 (A.1)하에 미지의 불연속점의 위치와 점프크기를 추정하기 위하여 한쪽방향커널함수를 이용하여 확률밀도함수의 오른쪽 추정량과 왼쪽 추정량을 정의한 후, 그 차이를 최대로 하는 점을 위치추정량이라고 제안하고, 그 위치추정량에서 추정된 점프크기를 불연속점의 점프크기로 추정량을 제시하였다. Huh (2002)가 제안한 임의의 점 x에서의 확률밀도함수의 오른쪽과 왼쪽 커널추정량은 각각

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

로 정의된다. 여기서 사용된 h는 평활모수인 띠폭이며, 한쪽방향커널함수 K는 다음의 성질을 만족한다.

(A.2) 토대 $[0,1]$을 가지는 커널함수 K는 $\int_0^1 K(u)du = 1$, $K(0) > 0$ 그리고 임의의 $0 < u \leq 1$에 대하여 $K(u) \geq 0$를 만족한다.

위 (2.2)식의 추정량들은 x를 기준으로 각각 오른쪽과 왼쪽의 데이터만을 사용하여 확률밀도함수의 커널추정량을 정의한 것이다. 이 두 추정량을 이용하여 임의의 점 x에서 점프크기추정량은 다음과 같이

$$\widehat{\Delta}(x) = \widehat{f}_+(x) - \widehat{f}_-(x)$$

로 정의할 수 있고, 이들 점프크기추정량들 중 그 절대값의 크기가 가장 큰 점프크기추정량의 위치를 불연속점의 위치추정량으로 다음과 같이 제시할 수 있다.

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

한편 추정된 위치 $\widehat{\tau}$에서의 점프크기추정량인

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

을 점프크기 Δ의 추정량으로 정의할 수 있다. Huh (2002)는 제안한 위치추정량과 점프크기추정량의 수렴속도와 극한분포를 보였고, 점프크기추정량의 극한분포를 이용하여 불연속점의 존재유무에 대한 가설검정의 검정통계량을 제시하였다.

본 연구에서는 위 Huh (2002)의 불연속점 추정량들 (2.3)과 (2.4)를 구하기 위한 띠폭 h의 선택에 대해 연구하고자 한다. 먼저 연속 가정 하의 확률밀도함수의 커널추정량에 쓰이는 대칭함수이고 토대가 $[-1,1]$인 커널함수를 L이라 하고 띠폭을 h_c라고 하자. 커널확률밀도함수추정량

$$\widehat{f}_{h_c}(x) = \frac{1}{nh_c}\sum_{i=1}^{n} L\left(\frac{X_i - x}{h_c}\right)$$

에서 띠폭 h_c의 선택에 흔히 쓰이는 최대가능도교차타당성은

$$CV_{ML}(h_c) = \frac{1}{n}\sum_{i=1}^{n} \log \widehat{f}_{h_c,i}(X_i) \tag{2.5}$$

로 정의되고, 이를 최대로 하는 h_c를 선택하게 된다. 여기서

$$\widehat{f}_{h_c,i}(X_i) = \frac{1}{(n-1)h_c}\sum_{j \neq i} L\left(\frac{X_j - X_i}{h}{}_c\right)$$

이고, X_i를 제거한 $f(X_i)$의 커널추정량이다. 식 (2.5)는 비모수적 함수추정에 흔히 쓰이는 거리의 척도인 쿨백-라이블러 정보 (Kullback Leibler information) $d_{KL}(f, \widehat{f}_{h_c})$와 다음과 같은

$$E[CV_{ML}(h_c)] \approx -E[d_{KL}(f, \widehat{f}_{h_c})] + \int (\log f(x)) f(x) dx$$

관계를 가진다. 여기서

$$d_{KL}(f, \widehat{f}_{h_c}) = \int \log(f/\widehat{f}_{h_c})(x) f(x) dx$$

이다. Hart와 Yi (1998)는 연속인 회귀함수의 커널추정량의 띠폭 선택을 위해 한쪽방향교차타당성을 제안하였다. Huh (2010b)는 Hart와 Yi의 방법을 사용하여 회귀함수의 불연속점의 추정량들을 위한 띠폭을 오른쪽과 왼쪽 한쪽방향교차타당성들의 합을 최소로 하는 것으로 선택하였다. 본 연구에서는 확률밀도함수의 불연속점의 추정량들을 위한 띠폭을 회귀함수의 불연속점 추정을 위한 Huh의 띠폭 선택 방법과 같이 (2.5)의 최대가능도교차타당성에 한쪽방향교차타당성 방법을 적용하여 선택하고자 한다.

먼저 X_i를 제거한 $f(X_i)$의 오른쪽과 왼쪽 한쪽방향커널추정량을 각각 다음과 같이

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

정의하고, 다음의 교차타당성

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

을 생각하자. 여기서

$$CV^{\pm}_{ML}(h) = \frac{1}{n_h}\sum_{i\in I_h} \log \widehat{f}_{h,i,\pm}(X_i)$$

이고 $I_h = \{i : X_i \in [h, 1-h]\}$이며 n_h는 I_h의 원소의 수이다. Müller (1992)는 회귀함수의 불연속점 추정을 위한 띠폭은 연속인 회귀함수의 추정을 위한 띠폭보다 상대적으로 작은 것을 추천하였다. Huh (2010b)는 Müller의 주장처럼 일반화선형모형의 회귀함수가 불연속점을 가질 때 불연속점 추정을 위한 띠폭을 오른쪽과 왼쪽 한쪽방향교차타당성의 합의 국소최소점 (local minimizer) 중 최소의 값으로 제안하였다. Müller (1992)와 Huh (2010b)가 제안한 방법처럼 확률밀도함수의 불연속점 추정을 위한 띠폭을 식 (2.6)의 국소최대점 (local maximizer) 중 가장 작은 값으로 제안하고자 한다.

하나 이상의 불연속점을 가지는 경우에, 한 불연속점의 점프크기 추정치들이 인접해 있는 점의 점프크기에 영향을 주기 때문에 하나의 불연속점 주변의 점들에서도 불연속점으로 판정될 가능성이 있다. 따라서, Jose와 Ismail (1999)이 언급하였던 것처럼 근접해 있는 두 개의 불연속점들 사이의 거리들은 2h보다 크다는 가정이 필요하게 된다. 이러한 가정을 바탕으로 Kim 등 (2003)과 Huh (2007, 2010a)는 각각 선형모형, 분산함수와 일반화선형모형에서 불연속점 수의 추정 알고리즘을 제안하였다. 확률밀도함수가 하나 이상의 불연속점을 가지는 경우에 본 연구에서 제안한 방법으로 띠폭을 선택한 후, Kim 등과 Huh가 제안한 알고리즘을 이용하여 불연속점의 수와 그 위치를 추정할 수 있을 것이다.

3. 모의실험

확률밀도함수의 불연속점 추정을 위한 제안된 띠폭 선택에 대한 모의실험 결과를 알아보고자 한다. 모의실험을 위하여 토대 $[0, 1]$을 가지는 다음의 확률밀도함수

$$f(x) = p\left\{\frac{\lambda_1}{2}\exp(-\lambda_1|x-0.5|)1_{[0\le x<0.5]} + \frac{\lambda_2}{2}\exp(-\lambda_2|x-0.5|)1_{[0.5\le x\le 1]}\right\}$$

을 생각하자. 여기서 $\lambda_1 = 1$이고 λ_2는 각각 3, 6, 9인 세 가지 경우를 고려하자. 상수 p는 확률밀도함수의 조건을 만족하게 하는 값으로, λ_2가 각각 3, 6, 9인 경우에 p는 각각 1.708906, 1.488447, 1.446801이 된다. 이 세 가지 경우에 확률밀도함수 f는 모두 $\tau = 0.5$에서 불연속을 가지며, 이 점에서 각각의 불연

속점의 점프크기는

$$\Delta = p(\lambda_2 - \lambda_1)/2 = \begin{cases} 1.708906, & \lambda_2 = 3 \\ 3.721118, & \lambda_2 = 6 \\ 5.787203, & \lambda_2 = 9 \end{cases}$$

이 된다. 불연속점을 가지는 확률밀도함수의 형태를 알아보기 위해 $\Delta = 3.721118$ ($\lambda_2 = 6$)인 경우의 확률밀도함수 f를 아래의 그림 3.1에 제시하였다.

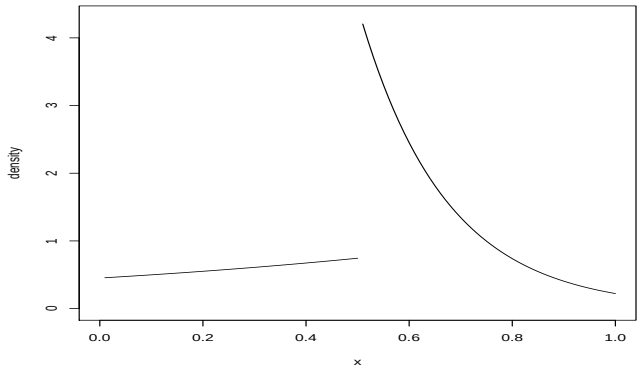


그림 3.1 점프크기 Δ=3.721118인 경우의 확률밀도함수 f

조건 (A.2)를 만족하는 한쪽방향커널함수는 토대 $[0,1]$을 기반으로 Epanechnikov 커널을 이용하여 만든

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

을 선택하였고, 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. Müller (1992)와 Huh (2010b)가 제안한 것처럼 구간 Q를 $[h, 1-h]$로 선택하였다. 세 가지 경우에 대한 불연속점의 추정의 정도를 보기 위하여 각 경우에서 표본의 수 n을 100, 200, 400으로 하여 각각 1000번 반복하였다.

각 표본에서 선택된 $\widehat{h}$들을 이용하여 추정된 $\widehat{\tau}$와 $\widehat{\Delta}(\widehat{\tau})$의 평균제곱오차의 몬테카를로 추정치 (Monte Carlo estimates of the mean squared error, MSE)들이 표 3.1에 제시되었다. 불연속점 위치추정량의 수렴속도는 Müller (1992), Loader (1996), Huh와 Carrière (2002), Huh와 Park (2004), Huh (2002, 2010b)에 의해 n^{-1}이 됨이 밝혀졌고, 점프크기가 클수록 불연속점 추정량의 점근분산이 작아져서 추정의 정도가 좋음이 알려졌다. 이 결과에 따라 표 3.1의 위치추정량 $\widehat{\tau}$의 MSE는 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. 선택된 $\widehat{h}$과 추정된 $\widehat{\tau}$와 $\widehat{\Delta}(\widehat{\tau})$의 평균들이 표 3.1에 제시되어있고, 괄호 안은 추정치들의 평균 혹은 MSE들의 표준오차들이다. 선택된 띠폭이 $\Delta = 1.708960$ ($\lambda_2 = 3$)인 경우에 표본의 수가 증가함에 따라 선택된 띠폭의 평균이 커지고 있다. 이는 일반적으로 표본의 수가 증가함에 따라 교차타당성 등에 의해 선택된 띠폭이 작아지는 경향과 다른 현상이 일어나고 있다. 본 연구에서는 제안한 식 (2.6)의 국소최소점 중 최소로 하는 것을 띠폭으로 선택하였고, 이로 인해 교차타당성을 최소로 하는 것으로 선택

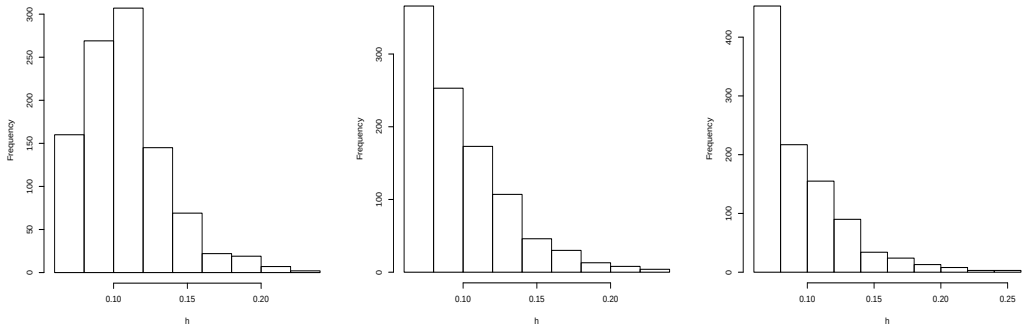
된 띠폭 값보다는 불안정한 값을 나타낼 수도 있어 모의실험의 모형에 따라서 표본의 수가 증가하더라도 선택된 띠폭의 평균이 커지는 경향이 있을 수 있다.

표 3.1 불연속점 추정치들의 모의실험 결과

Δ	n	$\widehat{h}$의 평균	$\widehat{\tau}$의 평균	$\widehat{\tau}$의 MSE	$\widehat{\Delta}(\widehat{\tau})$의 평균	$\widehat{\Delta}(\widehat{\tau})$의 MSE
1.708960	100	0.150870	0.536230	0.010886	1.005524	2.267489
			(0.003094)	(0.000778)	(0.042104)	(0.137054)
	200	0.154160	0.510950	0.004038	1.258034	0.828893
			(0.001979)	(0.000491)	(0.025012)	(0.072993)
	400	0.177150	0.497700	0.000445	1.340802	0.249216
			(0.000663)	(0.000128)	(0.010664)	(0.019992)
3.721118	100	0.112130	0.506460	0.001176	2.645269	3.090531
			(0.001065)	(0.000177)	(0.043967)	(0.274441)
	200	0.101500	0.500770	0.000194	2.875881	1.306066
			(0.000440)	(0.000071)	(0.024324)	(0.114815)
	400	0.097250	0.499880	0.000002	2.940343	0.851983
			(0.000047)	(0.0000005)	(0.015568)	(0.024360)
5.787203	100	0.092030	0.501200	0.000183	4.148902	4.112465
			(0.000426)	(0.000056)	(0.037795)	(0.282572)
	200	0.076650	0.500070	0.000014	4.401838	2.434765
			(0.000120)	(0.000012)	(0.022705)	(0.097462)
	400	0.069070	0.499990	0.0000003	4.535279	1.847470
			(0.000017)	(0.0000002)	(0.016738)	(0.043110)

제안된 방법에 의해 선택된 띠폭 $\widehat{h}$의 분포를 알아보고자 $\Delta = 3.721118$ ($\lambda_2 = 6$)인 경우의 $\widehat{h}$의 히스토그램을 그림 3.2에 나타내었다. 그림 3.2의 (a), (b), (c)는 각각 표본의 수가 100, 200, 400일 때의 선택된 띠폭 $\widehat{h}$의 히스토그램이다. 표본의 수가 증가함에 따라 선택된 띠폭 $\widehat{h}$는 작아지며 표준오차도 작아지는 경향을 보이고 있다.

표 3.2는 선택된 $\widehat{h}$를 이용하여 추정된 불연속점의 위치추정량 $\widehat{\tau}$의 분포를 보기 위하여, 점프크기 $\Delta = 3.721118$ ($\lambda_2 = 6$)인 경우에 추정된 불연속점의 위치 x_k의 지표 k와 해당되는 빈도를 표 3.2에 나열하였다. 불연속점 $\tau = 0.5$에 해당되는 k는 50이며, 표본의 수가 100, 200, 400에 따라서 $k = 50$의 경우 각각 722, 888, 978회로 나타났다. 또한 표본의 수가 증가함에 따라서 추정된 불연속점의 위치에 대한 산포가 현저히 줄고 있다는 것을 알 수 있다. 지금까지 확률밀도함수가 불연속점을 가질 때 불연속점의 추정을 위한 띠폭 선택에 대한 연구가 없었기에 모의실험의 결과를 비교할 수 없었지만, 선택된 띠폭을 사용한 불연속점 추정 결과들의 추정 정도는 좋다고 할 수 있다.



n=100 n=200 n=400

그림 3.2 Δ=3.721118인 경우 선택된 $\widehat{h}$의 히스토그램

표 3.2 Δ=3.721118인 경우 불연속점의 위치추정치 $\hat{\tau}$의 x_k에서의 빈도

$n \diagdown k$	45	46	47	48	49	50	51	52	53	54	55	56	58	59	60	61	62	63	64	65	66	>66
100	3	4	14	20	91	722	70	19	3	2	1	2	2	1	4	5	8	5	5	1	5	13
200	-	1	5	6	53	888	33	5	-	-	-	1	1	1	-	-	-	1	1	3	-	1
400	-	-	-	-	17	978	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

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. (2004a). 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.

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. 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.

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.

Lee, C. S., Chang, C. and Park, Y. W. (2010). Estimates for parameter changes in a uniform model with a generalized uniform outlier. *Journal of the Korean Data & Information Science Society*, **21**, 581-687.

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 Wang, J. L. (1990). Nonparametric analysis of changes in hazard rates for censored survival data: An alternative to change-point models. *Biometrika*, **77**, 305-314.

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.

Journal of the Korean Data & Information Science Society
2012, **23**(1), 79–87

한국데이터정보과학회지

Bandwidth selection for discontinuity point estimation in density[†]

Jib Huh[1]

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

Received 19 November 2011, revised 14 December 2011, accepted 2 January 2012

Abstract

In the case that the probability density function has a discontinuity point, Huh (2002) estimated the location and jump size of the discontinuity point based on the difference between the right and left kernel density estimators using the one-sided kernel function. In this paper, we consider the cross-validation, made by the right and left maximum likelihood cross-validations, for the bandwidth selection in order to estimate the location and jump size of the discontinuity point. This method is motivated by the one-sided cross-validation of Hart and Yi (1998). The finite sample performance is illustrated by simulated example.

Keywords: Maximum likelihood cross-validation, one-sided kernel function, smoothing parameter.

[†] This research was supported by the Duksung Women's University Research Grants 2010.
[1] Associate professor, Department of Information & Statistics, Duksung Women's University, Seoul 132-714, Korea. E-mail: jhuh@duksung.ac.kr