Journal of Statistial
Theary & Methods
1994, Vol. 5, pp1 ~ 9

커널 확률밀도함수 추정량을 이용한 적합도 검정에 관한 연구

석 경 하[1) 김 대 학[2)

요약 확률밀도함수의 적합도 검정을 위한 새로운 검정 통계량을 소개하고 커널확률밀도함수 추정량을 이용한 제안된 검정 통계량의 점근 정규성을 규명하였다. 제안된 통계량과 콜모고르프-스미르노프 통계량과의 소표본 모의 실험비교를 통하여 제안된 통계량의 우수성을 입증하였다.

1. 서 론

최근에 밀도함수 추정에 관한 많은 연구가 이루어졌고 활발한 연구가 계속 되고 있다. 여러가지 확률밀도함수 추정량중에서도 커널확률밀도함수 추정량(kernel density eistimator)이 가장 많이 사용되고 있다. 이 커널 추정량에서 핵심적인 역할을 하는 평할계수(bandwidth)의 선택에 관한 연구가 아주 광범위하게 진행되고 있다. 평할계수의 추정법에 관한 좋은 참고문헌으로는 Silverman(1989)과 Marron(1989) 등이 있다.

수 많은 평할량(평할계수의 추정량)에 대한 연구중에서 근간이 될수있는것으로는 Rudemo(1982), Bowman(1984), Scott 와 Terrell(1987), Hall 과 Marron(1991), Jones, Marron 과 Park(1991), Hall,Marron 과 Park(1992) 그리고 Kim, Park 과 Marron(1992) 등을 꼽을수 있을것이다. 거의 대부분의 연구는 수행측도(Performance Measure)로서 누적제곱오차 (Intergrated Sqrared Error)혹은 평균누적제곱오차 (Mean Intergrated Squared error)을 채택하여 이를 최소화하는 최적평할계수를 추정하는 방법을 주제로 삼고 있다.

본 논문은 평균누적제곱오차의 최적평할계수에 대하여 최적수렴율을 가지는 평할량중의 하나인 Song, Seog 과 Cho (1991)의 평할량 $\hat{h}_s$을 이용한 누적제곱오차를 밀도함수의 적합도 검정을 위한 검정통계량으로 제안하였다. 그리고 제안된 검정통계

1) 인제대학교 통계학과
2) 대구효성가톨릭대학교 통계학과

량의 점근분포를 규명하였다. 또한 이 분야의 콜모고로프-스미르노프 검정법과의 소표본 모의실험비교를 통하여 제안된 방법의 우수성을 규명하였다.

2. 커널확률밀도함수 추정법

확률밀도함수 f를 가지는 크기 n인 랜덤표본을 $X_1,\cdots,X_n$이라 할 때 f의 커널추정량(Rosenblatt (1956), Parzen (1962) 은

$$\hat{f}_h = \sum_{i=1}^{n} K_h(x - X_i) \tag{1}$$

이다 . 여기에서 K는 커널함수, $K_h(t) = h^{-1}K(h^{-1}t)$이고 h는 평할계수이다. 커널추정량 $\hat{f}_h$에 미치는 영향을 비교하면 커널함수 K의 선택은 평할계수 h의 선택에 비해서 아주 미미한것으로 알려져있다. 그러므로 h의 추정에 관한 연구가 많이 이루어졌다. 일반적으로 커널함수는 대칭인 확률밀도함수를 이용한다.

h의 추정량 $\hat{h}$을 이용한 밀도함수 $\hat{f}_h$의 수렴율은 밀도함수 f의 매끄러운 정도와 커널함수 K의 차수에 의해서 대체로 결정된다. 그래서 이들에 대한 정의를 소개하기로 한다.

정의 1. 1) l은 정수이고 $\beta \in (0,1]$일때 $\nu = l + \beta$ 이고, f가 횔더연속이며 $\int f''^2 < \infty$를 만족할 때 모든 x와 y에 대해

$$\left| f^{(2+l)}(x) - f^{(2+l)}(y) \right| \le M \left| x - y \right|^{\beta}$$

를 만족하는 상수 M이 존재하면 f는 매끄러운 정도의 차수가 ν이다.

2) 커널함수 K가 다음을 만족할 때

$$\int x^j K(x)dx = \begin{cases} 1, & j = 0, \\ 0, & j = 1,\cdots\cdots,r-1 \\ c, & j = r,(c \ne 0). \end{cases}$$

커널 K의 차수는 r이다 라고 한다.

여러가지의 수행측도중 본 논문에서 언급할 것은 평균누적제곱오차

$$MISE(h) = E\int (\hat{f}_h(x) - f(x))^2 dx \tag{2}$$

이다. 기대되는 자료에 평균적으로 좋게 추정하고자 하는 이 측도, $MISE(h)$ 와 현재 존재하는 자료를 가장 잘 추정하고자 하는 측도, 누적제곱오차

$$ISE(h) = \int (\hat{f}_h(x) - f(x))^2 dx \tag{3}$$

와의 자세한 비교는 Hall and Marron (1991) 과 Jones (1991) 을 참조하기 바란다.

(2) 식의 $MISE(h)$ 를 최소화하는 h 를 h_{opt} (최적평할계수) 라 하자. h_{opt}와의 최적상대수렴율 $(\hat{h}/h_{opt} - 1)$이 $n^{-1/2}$가 되는 평할량 $\hat{h}$을 구하는 연구에는 Song, Seog

과 Cho(1991), Jones, Marron 과 Park(1991), Hall, Sheather, Jones 과 Marron(1991), 그리고 Kim, Park 과 Marron(1992)등이 있는데 Song, Seog 과 Cho(1991)의 방법을 응용한 적합도 검정이 본 논문의 주 내용이므로 그 내용을 간단히 소개하기로한다. 반복을 피하기 위해 자주 사용되는 기호를 다음과 같이 정의하자.

1) 어떤 함수 에 대해 $R(g)=\int\{g(x)\}^2dx$
2) $\mu_r=\int x^rK(x)dx$
3) $x_r=\int x^rK*K(x)dx$ 이때 *는 합성적(convolution)을 나타낸다.

위의 기호를 이용하여 (2)식을 근사적으로 표현하면

$$AMISE^*(h)=(nh)^{-1}R(K)+h^4R(f'')\mu_2^2/4-h^6\mu_2\mu_4R(f^{(3)})/24-n^{-1}R(f) \tag{4}$$

로 된다. 이 식을 최소화하는 평할계수의 추정량 $\hat{h}_s$을 식으로 나타내면

$$\hat{h}_s=\hat{h}_{AMISE}+\hat{h}^3{}_{AMISE}\tilde{R}_{a_2}(f^{(3)})\mu_4/20\mu_2\tilde{R}_{a_1}(f''), \tag{5}$$

이다. 그리고, 또 (5)식에서

$$\tilde{R}_{a_4}(f^{(3)})=n^{-1}a_2^{-7}R(U_2^{(3)})-2n^{-1}(n-1)^{-1}a_2^{-7}\sum_{i<j}\sum U_2{}^{(3)}*U_2^{(3)}\{(X_i-X_j)/a_2\}$$

이다. 여기에서 U_2는 차수가 2인 커널함수이고, a_2는 평할량으로서

$$a_2=\tilde{R}_{a_4}(f^{(4)})^{-1/9}\left\{R(U_2^{(3)})\Big/\int x^2U_2(x)dx\right\}^{1/9}n^{-1/9} \tag{6}$$

이다. 여기에서 $\hat{h}_{AMISE}$는 $(nh)^{-1}R(K)+h^4R(f')\mu_2^{2/4}$을 최소화하는 h의 추정량인데, 이는

$$\hat{h}_{AMISE}=\{R(K)/\mu_2^2\tilde{R}_{a_1}(f'')\}^{1/5}n^{-1/5}$$

이다. 그리고

$$\tilde{R}_{a_1}(f'')=n^{-1}a_1^{-5}R(U_1'')+2n^{-1}(n-1)^{-1}a_1^{-5}\sum_{i<j}\sum U_1''*U_1''\{(X_i-X_j)/a_1^*\} \tag{7}$$

이다. 여기서 U_1은 차수가 6인 커널함수이고, a_1는 평할량으로

$$a_1=\tilde{R}_{a_3}(f^{(5)})^{-1/11}\left\{6!R(U_1'')\Big/\left(2\int x^6U_1(x)dx\right)\right\}^{1/11}n^{-1/11} \tag{8}$$

이다. 그리고 또한 (7)식에서

$$\tilde{R}_{a_3}(f^{(5)})=n^{-1}a_3^{-11}R(U_3^{(5)})-2n^{-1}(n-1)^{-1}a_3^{-11}\sum_{i<j}\sum U_3^{(5)}*U_3^{(5)}\{(X_i-X_j)/a_3)\} \tag{9}$$

이고 (8)식에서

$$\tilde{R}_{a_4}(f^{(4)}) = n^{-1}a_4^{-9}R(U_4^{(4)}) + 2n^{-1}(n-1)^{-1}a_4^{-9}\sum_{i}\sum_{<j} U_4^{(4)} * U_4^{(4)}\{(X_i - X_j)/a_4\}$$

(10)

인데 (9) 식과 (10) 식의 U_3과 U_4는 차수가 2인 커널함수이고, a_3과 a_4는 평활량으로서

$$a_3 = \hat{\lambda}R(g_1^{(6)})^{-1/13}\left\{R(U_3^{(5)})\Big/\int x^2U_3(x)dx\right\}^{1/13} n^{-1/13}$$

$$a_4 = \hat{\lambda}R(g_1^{(5)})^{-1/11}\left\{R(U_4^{(4)})\Big/\int x^2U_4(x)dx\right\}^{1/11} n^{-1/11}$$

이다. 여기에서 $\hat{\lambda}$ 는 f의 척도(scale)에 대한 $\sqrt{n}$일치 추정량이다.

정리 1. (Song, seog 과 Cho(1992)) f의 매끄러운 정도의 차수 ν가 2.25보다 크고, 커널함수 U_1과 U_2가 $m=2,3$에 대하여 $U_i(i=1,2)$는 대칭이고 $2m$번 미분가능이고 $f^{(2m)}(\pm\infty)=0$를 만족하면

$$\sqrt{n}(\hat{h}_s / h_{opt} - 1) \to N\{0, 4(R(f')^{-1}\int(f^{(4)})^2 \cdot f - 1)/25\}$$

이다.

3. 확률밀도함수의 적합도 검정

확률밀도함수 f를 가지는 크기가 n 인 랜덤표본을 $X_1, X_2, \cdots, X_n$이라 할때 이 랜덤표본을 이용한 f 에 대한 적합도 검정 ($H_0 : f = f_0$ vs $H_1 : f \neq f_0$)방법들 중에서 가장 대표적인 검정법은 콜모고로프-스미노프 검정법이다. 이의 검정통계량은

$$\sup_x \left| \hat{F}(x) - F_0(x) \right|$$

이다. 여기에서 $\hat{F}(x)$는 랜덤표본의 경험적 분포함수이고 $F_0(x)$는 f_0의 분포함수이다.

3.1 검정통계량의 소개

본 논문에서는 적합도 검정을 위해서 통계량

$$ISE(\hat{h}_s) = \int(\hat{f}_{\hat{h}_s}(x) - f_0(x))^2 dx$$

을 제안한다. 이 검정통계량의 근사분포는 다음과 같다.

정리 2. 만약 f_0와 커널함수 U_1, U_2 가 정리 1의 가정을 만족한다면

$$ISE(\hat{h}_s) \to N(\mu, \sigma^2)$$

이다. 여기에서

$$\mu = n^{-1}h^{-1}R(K) + h^4\mu_2^2R(f'')/4 - R(f)/n$$

$$\sigma^2 = n^{-1}h^4\left\{(\mu_2 - x_2)^2\int ff''^2 - \mu_2^2R^2(f')\right\} + n^{-2}h^{-1}R(K*K)R(f)$$

이다.

3.2 정리 2의 증명

$$ISE(h) = \int(\hat{f}_h(x) - f(x))^2$$
$$= n^{-1}h^{-1}R(K) + n^{-2}\sum_{i}\sum_{<j} K_h * K_h(X_i - X_j) + 2n^{-1}\sum_{i=1}^{n} K_h * f(X_i) + R(f)$$

로 되고 이때 $A_{ij} = K_h * K_h(X_i - X_j)$, $B_i = K_h * f(X_i)$로 두면

$$E(A_{12}) = \int\int\int K_h(x-y)K_h(x-z)f(y)f(z)dxdydz$$
$$= h^{-1}\int\int\int K(t)K(t+(y-z)/h)f(y)f(z)dxdydz$$
$$= \int\int\int K(t)K(t+u)f(y)\left(f(y) + h^2u^2f''(y)/2 + h^4u^4f^{(4)}(y)/4! + O(h^6)\right)$$
$$= R(f) - h^2\mu_2R(f') + (2\mu_4 + 6\mu_2^2)h^4R(f'')/4! + O(h^6)$$

$$E(A_{12}^2) = h^{-2}\int\int\left(K*K((x-y)/h)\right)^2f(x)f(y)dxdy$$
$$= h^{-1}R(K*K)R(f) + O(h)$$

$$E(B_1) = \int\int K_h(x-y)f(x)f(y)dxdy$$
$$= R(f) - h^2\mu_2R(f)/2 + h^4\mu_4/\{4!R(f')\} + O(h^6)$$

$$E(B_1^2) = \int\left(h^{-1}\int K((y-x)/h)f(x)dx\right)^2f(y)dy$$
$$= R(f^{3/2}) + h^2\mu_2\int f^2f'' + h^4\left(\mu_4\int f^2f^{(4)}/12 + \mu_2^2\int ff''^2/4\right)$$

$$E(A_{12}A_{13})$$
$$= h^{-2}\int\int\left(\int K(t)K(t+(x-y)/h)dt\right)\left(K(u)K(u+(x-z)/h)du\right)\times f(x)f(y)f(z)dxdydz$$
$$= \int\int\int K*K(t)K*K(u)f(x)f(x-ht)f(x-hu)dxdtdu$$
$$= R(f^{3/2}) + h^2x_2\int f^2f''^2 + h^4\left(x_4\int f^2f^{(4)}/12 + x_2^2\int ff''^2/4\right)$$

$$E(A_{12}^2) = h^{-2}\int\int\int K(t)K(t-(x-y)/h)dt\int K(u)K(u-(x-y)/h)du$$

$$E(A_{12}B_1)=\int\int\int K*K(t)K(z)f(x-ht)f(x)f(x-ht)dzdxdt$$
$$=R(f^{3/2})+h^2(\mu_2+x_2)\int f^2f''/2+h^4\left((\mu_4+x_4)\int f^2f^{(4)}/4!+6\mu_2x_2\int ff''^2\right)$$

로 되고 따라서

$$E\{ISE(h)\}=n^{-1}h^{-1}R(K)+h^4\mu_2^2R(f'')/4-R(f)/n$$

이다. 또

$$Var(A_{12})=h^{-1}R(K*K)R(f)-R^2(f)+2h^2\mu_2R(f)R(f')$$
$$Var(B_1)=R(f^{3/2})-R^2(f)+h^2\mu_2\left(\int f^2f''+R(f)R(f')\right)$$
$$+h^4\mu_2^2\left(\int ff''^2-R^2(f)\right)/4+h^4\mu_4\left(\int f^2f^{(4)}-R(f)R(f'')\right)/12$$
$$Cov(A_{12},A_{13})=R(f^{3/2})-R^2(f)+h^2\left(2\mu_2R(f)R(f')+x_2\int f^2f''\right)$$
$$+h^4\left(x_4\int f^2f^{(4)}/12+x_2^2\int ff''^2/4-\mu_2^2R^2(f')-(6\mu_4+2\mu_2^2)R(f)R(f'')\right)$$

이고

$$Cov(A_{12},B_1)=R(f^{3/2})-R^2(f)+h^2\left\{(x_2+\mu_2)\int f^2f''+3\mu_2R(f)R(f')\right\}/2$$

이므로

$$Var(ISE(h))=Var(2n^{-2}\sum_{i<j}\sum A_{ij}-2n^{-1}\sum_{i=1}^{n}B_j)$$
$$=n^{-3}(n-1)Var(A_{12})+4n^{-1}Var(B_1)$$
$$+4n^{-3}(n-1)(n-2)Cov(A_{12},A_{13})-8n^{-2}(n-1)Cov(A_{12},B_1)$$
$$=n^{-1}h^4((\mu_2-x_2)^2\int ff''^2-\mu_2^2R^2(f'))+n^{-2}h^{-1}R(K*K)R(f)+O(n^{-2})$$

가 된다. 그리고 점근정규성은 Hall(1984)의 정리 (degenerate martingale 정리)에 의해서 증명된다. 또한 Taylor 정리에 의해서

$$ISE(h_{opt})=ISE(\hat{h}_s)+(h_{opt}-\hat{h}_s)ISE(h^*),$$

로 된다. 여기에서, h^*는 h_{opt}와 $\hat{h}_s$사이의 값이 되는데 $h_{opt}-\hat{h}_s=O(n^{-7/10})$이고 $ISE'(h^*)=O(n^{-3/5})$이므로 $ISE(h_{opt})$와 $ISE(\hat{h}_s)$는 동일한 점근분포를 가진다.

4. 모의실험

제안된 검정통계량의 소표본에서의 수행능력을 알아보기 위하여 모의실험을 수행하

였다. 제안된 통계량의 점근 귀무평균과 귀무분산의 계산을 용이하게 하기 위하여 f_0를 표준정규분포의 밀도함수로 제한하였다.

모의실험을 위한 n개의 난수(random number)의 생성은 *IMSL*의 *RNNOR*과 *RNCHI* 부프로그램을 사용하였다. 이 실험에서는 크기 n이 25, 50, 100, 200인 표본을 각각 1,000번 반복 생성하여 검정력을 비교하였다. 여러경우(긴 꼬리를 가지는 경우, 치우침이 있는 경우, 이봉인 경우)를 비교하기 위하여 사용된 분포는

표1. 유의수준이 0.1일때 두개의 검정법의 검정력 비교
(제안된 방법의 검정력이 콜모고르프-스미르노프의 검정력뒤에 있음)

f \ n	25	50	100	200
N(0, 1)	0.000 0.012	0.000 0.013	0.001 0.014	0.000 0.019
chi (1)	0.579 0.928	0.985 1.000	1.000 1.000	1.000 1.000
chi (2)	0.196 0.627	0.589 0.962	0.984 1.000	1.000 1.000
chi (3)	0.084 0.411	0.311 0.808	0.765 0.990	0.997 1.000
chi (4)	0.042 0.268	0.184 0.586	0.573 0.936	0.941 1.000
chi (5)	0.023 0.220	0.139 0.468	0.393 0.851	0.904 1.000
t (1)	0.725 0.846	0.551 0.988	1.000 1.000	1.000 1.000
t (2)	0.207 0.386	0.469 0.705	0.799 0.939	0.986 1.000
t (3)	0.060 0.180	0.160 0.364	0.362 0.586	0.705 0.863
t (4)	0.029 0.095	0.059 0.180	0.142 0.307	0.349 0.593
t (5)	0.014 0.068	0.023 0.091	0.059 0.179	0.157 0.358
Bimodal	0.002 0.037	0.000 0.005	0.005 0.107	0.004 0.540

표2. 유의수준이 0.05일때 두개의 검정법의 검정력 비교
(제안된 방법의 검정력이 콜모고르프-스미르노프의 검정력뒤에 있음)

f \ n	25	50	100	200
N (0, 1)	0.000 0.010	0.000 0.007	0.000 0.009	0.000 0.009
chi (1)	0.324 0.903	0.932 1.000	1.000 1.000	1.000 1.000
chi (2)	0.086 0.543	0.350 0.945	0.920 1.000	1.000 1.000
chi (3)	0.032 0.337	0.151 0.747	0.546 0.981	0.985 1.000
chi (4)	0.013 0.209	0.058 0.513	0.332 0.899	0.844 1.000
chi (5)	0.003 0.158	0.040 0.376	0.200 0.799	0.729 0.998
t (1)	0.632 0.849	0.922 0.983	0.999 1.000	1.000 1.000
t (2)	0.144 0.355	0.357 0.661	0.679 0.920	0.968 1.000
t (3)	0.030 0.150	0.102 0.317	0.236 0.538	0.550 0.833
t (4)	0.018 0.068	0.019 0.146	0.079 0.275	0.204 0.539
t (5)	0.002 0.049	0.001 0.062	0.025 0.132	0.071 0.304
Bimodal	0.000 0.019	0.000 0.032	0.000 0.063	0.000 0.157

$N(0,1), t(n), x^2(n)$ $n = 1, \cdots, 5$ 그리고 $0.5N(1, 4/9) + 0.5N(-1, 4/9)$이다. 생성된 난

수는 표준화시킨후에 가정에 사용 되었고 유의수준을 0.1, 0.05로 하였다. (6)식의 차수가 6인 커널함수 U_1은

$$U_1 = \phi(x) - \phi''(x)/2 + \phi^{(4)}(x)/8$$

를 이용하였다. 이때 $\phi(x)$는 표준정규분포의 밀도함수를 이용했고 차수가 2인 커널함수 U_2, U_3, U_4는 표준정규분포의 밀도함수를 사용하였다. 이 실험의 결과는 표1과 표2에 나타나 있다. 이 표에서 알수 있듯이 제안된 방법이 모든 경우에서 우수하다는 것을 알 수 있다. 특히 카이제곱 분포에서의 수행능력이 상당히 뛰어남을 알 수가 있다. 표본의 크기가 작을때의 수행성이 더욱 더 돋보인다.

표본의 크기가 25일때, 콜모고르프-스미르노프 검정법은 자유도가 3이상인 카이제곱분포와 t-분포에서는 검정력이 유의수준이하로 떨어진다. 표본의 크기가 커짐에 따라 검정력은 좋아지지만 높은 자유도를 가지는 분포에서의 검정력은 여전히 만족할만한 수준은 되지 못하는 것으로 평가된다. 그것에 비해 제안된 검정법의 검정력은 아주 좋은것으로 보인다. 특히 높은 자유도의 분포에서는 검정력이 최고 50배 정도에 이른다.

그리고 주목해야 될 것 중의 하나는 두 검정법 모두가 이봉성을 검정하는 능력이 상당히 뒤진다는 것이다. 여기에 대해서는 별도의 연구가 되어야 할 것으로 생각된다.

참고문헌

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density esimates. *Biometrika,* 71, 353, 360.

Chiu, S.T. (1991). Bandwidth selection for kernel density estimation. *Annals of Statistics* , 19, 1883-1905.

Hall, P. and Marron, J.S. (1987). Extents to which least- squares cross-validation minimises integrated square error in nonparametric density estimation. *Probability Theory and Related Fields.* 74, 567-581.

Hall, P. and Marron, J.S. (1991). Lower bounds for bandwidth selection in density estimation. *Probability Theory and Related Fields,* 90, 149-173

Hall, P., Marron, J.S. and Park, B.U. (1992). Smoothed cross-validation. *Probability Theory and Related Fields,* 92, 1-20.

Hall, P., Sheather, S.J., Jones, M.C. and Marron, J.S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika,* 78, 263-269.

Jone, M.C. (1991) The roles of ISE and MISE in density estimation. *Statistics and Probability Letters,* 12, 51-56.

Jones, M.C., Marron, J.S. and Park, B.U. (1991). A simple root n bandwidth selector. *Annals of Statistics,* 19, 1919-1932.

Kim, W.C., Park, B.U. and Marron, J.S. (1994). Asymptotically best bandwidth selectors in kernel density estimation. *Statistics and Probability letters,* *19*, 119-127.

Park, B.U. and Marron, J.S. (1994).Comparison of data-driven bandwith selectors. *Journal of the American Statistical Association,* 85, 66-72.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandanavian Journal of Statistics,* 9, 65-78.

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

Silverman, B.W.. (1986). Density estimation for Statistics and Data Analysis. *Chapman and Hall, London.*

Song, M.S., Seog, K.h., and Cho, S.S. (1991). On asymptotically optimal plug-in bandwidth selectors in kernel density estimation, *Journal of the Korean Statistical Society,* 20, 29-43