응용통계연구 제 10권 1호
1997년 3월 pp. 137-149

푸리에 전개에 기초한 로그밀도추정[1)]

구 자 용[2)], 이 기 원[3)], 박 현 숙[4)]

요 약

본 논문에서는 푸리에 전개에 기초한 로그밀도추정법을 제안하였다. 삼각함수로 구성된 기저함수들은 베이즈정보 규준량에 근거하여 단계적 추가 및 삭제를 이용하여 결정하였고, 모수의 추정에는 최대가능성 방법을 이용하였다. 기존 자료의 분석 및 모의실험을 통하여 제안된 방법의 성능을 예시하였다.

1. 서 론

확률밀도함수의 추정방법으로는 크게 커널(kernel) 방법, 제약가능성(penalized likelihood) 방법 및 직교열(orthogonal series) 방법 등이 있다. 이러한 방법들은 Silverman(1986)에 잘 설명되어 있다. 이외에 Crain(1974), Stone과 Koo(1986), Stone(1990), Barron과 Sheu(1991), Kooperberg와 Stone(1991, 1992), Park(1994), Koo와 Chung(1995), Kooperberg(1995), Koo(1996), Koo와 Kim(1996) 그리고 Koo와 Park(1996a,b) 등에서 사용된 로그밀도추정법이 있다.

본 논문에서는 삼각함수(trigonometric function)를 이용한 로그-푸리에 밀도추정법을 제안하였다. 여기서 로그-푸리에 밀도추정법의 기본 생각은 추정대상인 밀도함수 f에 대하여 $\log(f)$를 푸리에 전개(Fourier expansion)했을 때 그 푸리에 계수(coefficient)를 최대가능성(maximum likelihood)방법으로 추정하여 f의 추정량을 구하는 것이다. 추정량을 결정하는 요소로는 삼각함수의 유형, 갯수 및 주파수(frequency)가 있다. 이들을 사용자의 주관이 개입되지 않은, 자동화된(automatic) 방법에 의해 결정하는 절차를 제시하는 것이 본 논문의 주된 내용이다.

이를 위해 Kooperberg, Stone과 Truong(1995), Koo와 Park(1996b) 등에서 사용된 접목점의 단계적 추가 및 제거 방법을 응용하였다. 제안된 방법이 보통의 직교열 방법에 비해 가지는 장점은 추정량이 하나의 밀도함수(양수값을 가지며 적분값이 1이 됨)가 된다는 것이다.

1) 첫 저자의 연구는 1996년도 교육부 기초과학 육성 연구비(BSRI-96-1418) 지원에 의한 것임.
2) 강원도 춘천시 옥천동 1번지 한림대학교 통계학과 부교수.
3) 강원도 춘천시 옥천동 1번지 한림대학교 통계학과 부교수.
4) 강원도 춘천시 옥천동 1번지 한림대학교 통계학과 박사과정.

2. 로그-푸리에 밀도추정량

구간 I=[0, 1]에서 정의된 밀도함수 f를 추정하는 문제를 먼저 고려해 보자. 기저함수(basis function)들의 집합 $\{\phi_j ; j=1,2,\cdots\}$를 $\{\sin(2\pi k_j x), \cos(2\pi k_j x) ; j=1,2,\cdots\}$이라 하자. 보통의 직교열 추정법에서는 주파수(frequency) k_j들이 정수값을 가지도록 하는데 비하여, 본 논문에서는 추정량의 유연성(flexibility: 여러 형태의 함수에 대해 좋은 적합도를 가짐)을 높이기 위하여 진동수 k_j가 정수가 아닌 경우도 고려했다.

Θ를 J-차원 열벡터 $\theta=(\theta_1,\cdots,\theta_J)'$의 집합이라 하고, Θ의 원소 θ에 대해

$$s(x;\theta)=\sum_{j=1}^{J}\theta_j\phi_j(x),$$

$$C(\theta)=\log\{\int_I \exp(s(x;\theta))dx\},$$

그리고

$$f(x;\theta)=\exp(s(x;\theta)-C(\theta)),\quad x\in I$$

라 하자. 이와 같이 정의한 함수 $f(\cdot;\theta)$는 구간 I에서 정의된 하나의 밀도함수가 된다. Θ에 속하는 θ에 대하여 $f(\cdot;\theta)$들을 J개의 모수를 가지는 로그-푸리에 모형(log-Fourier model)이라 정의한다. 여기서 상수함수 $\phi_0(x)=1$은 로그-푸리에 모형의 추정 가능성(identifiability)을 위하여 사용하지 않는다.

$X_1,\cdots,X_n$을 구간 I에서 정의된, 확률밀도함수 f로부터의 확률표본이라 하자. 이 때 $\bar{\phi}_j=n^{-1}\sum_i\phi_j(X_i)$라 하고, $l(\theta)$를

$$l(\theta)=\frac{1}{n}\sum_i\log f(X_i;\theta)=\sum_j\theta_j\,\bar{\phi}_j-C(\theta)$$

로 정의하자. 여기서 $l(\theta)$를 로그가능성함수(log-likelihood function)라 부르도록 한다. $\hat{\theta}$을 Θ의 원소로서 $l(\theta)$를 최대로 하는 값으로 정의하고, $\hat{\theta}$을 최대가능성추정량(maximum likelihood estimator; MLE)이라 부르며, f의 추정량으로는 $\hat{f}=f(\cdot;\hat{\theta})$을 사용하는 데, $\hat{f}$를 로그-푸리에 추정량이라고 한다.

3. 최대가능성 추정량의 계산

최대가능성추정량 $\hat{\theta}$의 계산 방법을 설명하기 위하여 몇 가지 기호를 정의하자. 먼저 $\theta\in\Theta$에 대하여 스코어 함수(score function) $S(\theta)$를 j번째 원소가

$$\frac{\partial l(\theta)}{\partial\theta_j}=\bar{\phi}_j-\mu_j(\theta)=\bar{\phi}_j-\int_I\phi_j f(\cdot;\theta)$$

인 J-차원 벡터로 정의한다. 또한, $\theta\in\Theta$에 대하여 $C(\theta)$의 헤시안(Hessian)행렬 $H(\theta)$를 (j,k) 원소가

$$\frac{\partial^2 C}{\partial\theta_j\partial\theta_k}(\theta)=\int_I\phi_j\phi_k f(\cdot;\theta)-\mu_j(\theta)\mu_k(\theta)$$

인 $J\times J$ 행렬로 정의한다. 이 때, MLE $\hat{\theta}$은 식 $S(\hat{\theta})=0$을 만족해야 한다. 본 논문에서는 $\hat{\theta}$을 계산하기 위하여 다음과 같은 뉴톤-랩슨(Newton-Raphson)방법을 이용하였다.

INITIALIZE $\quad\hat{\theta}^{(0)}=\theta^{(0)}$

ITERATE FOR $\quad r=0,1,\cdots,M_1.$

SOLVE $\quad H(\hat{\theta}^{(r)})\eta=S(\hat{\theta}^{(r)})\quad$ for η

IF $\quad l(\hat{\theta}^{(r)}+\eta)>l(\hat{\theta}^{(r)})\ $ then

$\hat{\theta}^{(r+1)}=\hat{\theta}^{(r)}+\eta$

ELSE

$\hat{\theta}^{(r+1)}=\hat{\theta}^{(r)}+2^{-m}\eta\ ,$

for $m=\min\{k:\ l(\hat{\theta}^{(r)}+2^{-k}\eta)>l(\hat{\theta}^{(r)})\},\quad 0\le k\le M_2$

END IF

UNTIL

$l(\hat{\theta}^{(r+1)})-l(\hat{\theta}^{(r)})<\varepsilon.$

우리가 사용한 수렴기준은 반복수 $M_1=50$, $M_2=30$이며 $\varepsilon=10^{-6}$이다. 그리고 뉴톤-랩슨 반복(iteration) 중에 나타나는 여러 가지 적분값들은 가우스-르장드르(Gauss-Legendre) 수치적분법[Press 외 3인(1992)에 있는 gauleg]을 이용하여 계산한다.

이제 우리가 추정하고자 하는 밀도함수 f_X의 정의구역이 실직선(real line) R인 경우를 고

려해 보자. $X_{(1)}$ 및 $X_{(n)}$을 밀도함수 f_X로 부터 나온 확률표본 $X_1, \cdots, X_n$의 첫 번째 및 n 번째 순서통계량이라 하자. 적절한 상수 L 및 U 에 대해 일차변환

$$y = ax + b = \frac{(U-L)x + LX_{(n)} - UX_{(1)}}{X_{(n)} - X_{(1)}}$$

에 의해 재조정된(rescale)된 자료를 $Y_i = aX_i + b$ 라 하면 Y_i들은 구간 $[L, U]$ 사이의 값을 가지게 된다. L 및 U 를 $0 < L < U < 1$을 만족하도록 잡은 후에 재조정된 자료 $Y_1, \cdots, Y_n$을 구간 I 에서 정의된 밀도함수 f_Y로 부터 얻은 확률표본으로 간주하고 위에서 설명한 로그-푸리에 방법을 적용하여 로그-푸리에 추정량 $\hat{f_Y}$를 구한다. 최종적으로 밀도함수 f_X의 로그-푸리에 추정량 $\hat{f_X}$은, X_L과 X_U 를 각각 $x_L = (UX_{(1)} - LX_{(n)})/(U-L)$, $x_U = \{(1-L)X_{(n)} + (U-1)X_{(1)}\}/(U-L)$이라 할 때 변수변환법에 의해

$$\hat{f_X}(x) = \frac{1}{a}\hat{f_X}\left(\frac{x-b}{a}\right), \quad x \in [x_L, x_U]$$

로 구한다. 한편, 구간 $[x_L, x_U]$밖에서는 $\hat{f_X} = 0$ 으로 정의한다.

여기서 구간 $[L, U]$를 구하는 문제를 생각해 보자. 먼저 L 및 U 를 각각 0 및 1로 잡지 않은 이유는 f_X가 R에서 정의되려면 $|x| \to \infty$일 때 $f_X(x)$ 가 0으로 수렴해야 한다는 사실때문이다. 여기서 $X_{(1)} - x_L = L(X_{(n)} - X_{(1)})/(U-L)$, $x_U - X_{(n)} = (1-U)(X_{(n)} - X_{(1)})/(U-L)$ 이다. 만약 $L = .2$이고 $U = .8$이면 $X_{(1)} - x_L = x_U - X_{(n)} = .3(X_{(n)} - X_{(1)})$ 이므로 $\hat{f_X}$는 자료가 퍼져있는 구간 $[X_{(1)}, X_{(n)}]$ 밖에서 0에 가까운 값을 가진 것으로 기대되어 구간 $[x_L, x_U]$밖에서 $\hat{f_X} = 0$ 으로 정의했을 경우 적어도 본 논문의 예에서는 $\hat{f_X}$가 시각적으로 R 에서 연속으로 나타난다. 본 논문에서 $[L, U]$는 [.2, .8]로 정했는데, 이는 여러 가지 가능성, 예를 들어 [.1, .9], [.15, .85], [.25, .75], [.3, .7] 등과 비교했을 때, 본 논문에서 제시한 모의실험 결과에서 [.2, .8]이 전반적으로 좋았기 때문이다. 여기서 어떤 $[L, U]$이 좋은가 하는 판단 근거는 "모의실험에서는 f_X를 알 수 있으므로 어떤 $[L, U]$가 좋은지 그래픽(graphic)을 통하여 결정할 수 있다"는 사실이다.

4. 변수선택법

여기서는 로그-푸리에 모형을 생성하는 기저함수(basis function) ϕ_j들의 집합을 정의하는 데, 이는 최종적인 로그-푸리에 추정량을 결정한다. 기본적으로 Friedman(1991), Kooperberg,

Stone과 Truong(1995)과 Koo와 Park(1996b)등에서 사용된 단계적 기저(basis)의 추가 및 제거 방법을 이용한다. 아래에서 k_j는 $j/4$ 형태의 값을 가지도록 했는 데, 이는 시뮬레이션을 통하여 얻어진 결과이다. 현재로서는 k_j를 $j/4$로 정한 데에 대한 특별한 이론적 근거는 없고, k_j를 $j/3, j/4, j/5, j/6$ 등의 경우로 모의실험을 시행했을 때 k_j가 $j/4$일 경우 본 논문에서 제시한 모의실험 결과가 전반적으로 좋았다는 사실을 그 근거로 제시할 수 있을 뿐이다.

기저의 추가는 최소모형

$$f_0(x;\theta) = \exp\{\theta_1 \sin(2\pi jx/4) + \theta_2 \cos(2\pi jx/4) - c(\theta)\}$$

을 적합시키는 작업부터 시작한다.

각 단계에서는 $\{\phi_j\} = \{\sin(2\pi jx/4), \cos(2\pi jx/4), 1 \le j \le 30\}$에 속하면서 현재 모형에 사용되지 않은 기저함수들 중에서 다음의 Rao통계량을 최대로 하는 기저함수를 추가한다. 여기서 Rao통계량을 보다 구체적으로 설명하기 위해 F_0를 차원이 $J-1$인 현재모형이라 하고, F_0에 새로운 기저를 추가하여 결정되는 모형을 F라 하자. 현재 모형 F_0의 모수를 $\theta_0 = (\theta_1, \cdots, \theta_{J-1})'$, 모형 F의 모수를 $\theta = (\theta_1, \cdots, \theta_J)'$, θ_J를 추가대상인 기저에 대응하는 모수, $\hat{\theta_0}$을 F_0하에서 구한 MLE라 하자. S와 H를 각각 F에 대응하는 스코어함수 및 헤시안행렬이라 할 때, Rao통계량은

$$S(\hat{\theta_0})' H^{-1}(\hat{\theta_0}) S(\hat{\theta_0})$$

로 주어진다. 기저함수들의 추가는 모수의 숫자가 $J_{\max} = \min(3n^{1/5}, 30)$일 때까지 계속된다. 여기서 $n^{1/5}$이어야 한다는 것은 $\log(f)$가 두 번 미분 가능할 경우의 대표본이론(Barron과 Sheu, 1991)에 따른 것이고, 30이라는 숫자는 모형이 너무 커지지 않도록 하기 위해 정한 규칙이다.

기저함수의 추가가 끝나면 단계적 기저함수 제거를 시작한다. 각 단계마다 최소모형 f_0로 될 때까지 현재모형에 포함된 기저함수 중 아래에 정의된 Wald통계량을 최소로 하는 기저를 제거한다. 모형 F_0를 현재모형, $\hat{\theta} = (\hat{\theta_1}, \cdots, \hat{\theta_J})'$를 F_0에 대응하는 MLE, θ_J를 제거대상인 기저함수에 대응하는 모수라 하자. 이 때 $[H^{-1}(\hat{\theta})]_{JJ}$를 현재 모형 F_0하에서 구한 $H^{-1}(\hat{\theta})$의 J번째 대각원소라 하면, 제거 대상인 기저함수의 중요도를 측정하는 Wald통계량은

$$\frac{\hat{\theta_J}^2}{[H^{-1}(\hat{\theta})]_{JJ}}$$

로 주어진다.

이제 최종 모형의 선택 기준을 고려해 보자. 단계적 추가 및 제거로 적합된 일련의 모형들 중에서 적절한 모형선택기준에 의해 최종모형을 선택한다. 본 논문에서는 Kooperberg와 Stone(1992), Park(1994), Kooperberg(1995), Koo(1996), Koo와 Park(1996a,b) 등에서와 같이 Schwarz(1978)가 제안한 베이즈정보규준량 BIC(Bayesian Information Criterion)을 사용했는데, BIC는 특정모형의 MLE를 $\hat{\theta}$, 모수의 수를 J 라 할 때

$$\mathrm{BIC} = -2nl(\hat{\theta}) + (\log n)J$$

로 정의된다. 최종적으로 f의 추정량으로는 BIC를 최소로 하는 로그-푸리에 모형을 선택한다.

최종적인 로그-푸리에 밀도추정량을 구하는 과정을 간단하게 요약하면 다음과 같다. 아래에서 '적합'은 기저함수의 종류 및 숫자, MLE $\hat{\theta}$, BIC를 의미한다.

1. f영역이 I인지, R인지를 확인한다. 이 경우 R이면 자료를 I의 부분구간 $[L, U]$로 재조정한다.
2. 최소모형으로부터 단계적 기저함수 추가를 시작, 각 단계에서 '적합'을 구한다.
3. 모수의 숫자가 $J_{\max}$와 같아지면 추가를 종료한다.
4. 단계적 기저함수 제거를 시작한다.
5. 최소모형 f_0가 될 때까지 기저함수들을 제거한다.
6. 2-5단계에서 구한 모형 중 BIC를 최소로 하는 모형을 최종모형으로 선택한다.
7. f의 영역이 R인 경우 변수변환법으로 6단계에서 구한 최종모형을 변환하여 로그푸리에추정량을 구한다.

5. 모의 실험 및 자료 적합 예

앞에서 설명한 로그-푸리에 추정방법을 Old Faithful 간헐천(geyser)의 분출(eruption)자료에 적용한 결과를 <그림 1>에 나타내었다. 원시자료로는 Silverman(1986)의 8쪽에 있는 것을 이용하였다. 한 편, <그림 2(a)>는 같은 자료에 대해 Silverman(1986)의 그림 2.8에 설명된 대로 창폭(window width)를 0.25로 하고, 표준정규밀도함수를 커널함수로 했을 때의 커널밀도추정량을 나타낸 것이다. <그림 1(a)>의 로그-푸리에 추정량을 <그림 2(a)>에 제시된 커널추정량과 비교하면 다음과 같다. <그림 2(a)>의 커널 추정량과의 차이점으로는 왼쪽 정점이 커널 추정량의 왼쪽 정점보다 상대적으로 높고 그 폭이 좁으며, 오른쪽 정점의 경우는 높이와 폭이 커널 추정량의 오른쪽 정점과 비슷하다.

또한, <그림 1(b)>는 Geyser자료에 대하여 제안된 밀도추정량의 적합도를 그래픽으로 나타낸 것이다(이하 누적도표라 한다). <그림 1(b)>의 점들은

$$\left(\frac{i}{n}, \hat{F}(X_{(i)})\right), \quad i=1,\cdots,n$$

을 나타내는 데, $\hat{F}$은 로그-푸리에 추정량 $\hat{f}$에 대응하는 누적분포함수이며, $X_{(i)}$는 i번째 순서 통계량이다. $\hat{F}$의 계산은 역시 가우스-르장드르 수치적분을 이용하여 구하였다. 이 점들이 전체적으로 직선 $y=x$ 근처에 위치하며 규칙적인(systematic) 차이가 보이지 않으므로 제안된 로그-푸리에 추정량은 잘 적합되었다고 판단된다. <그림 2(b)>는 커널추정량에 대해 누적분포함수를 공식

$$\hat{F}_K(x)=\frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{x-X_i}{.25}\right)$$

에 의해 구하고 이를 누적도표로 나타낸 것이다. 여기서 Φ는 표준정규분포의 누적분포함수이다. 이 도표와 <그림 1(b)>를 비교해 보면 커널추정량의 양쪽꼬리에서의 적합도가 로그-푸리에 추정량의 적합도 보다 떨어짐을 알 수 있다.

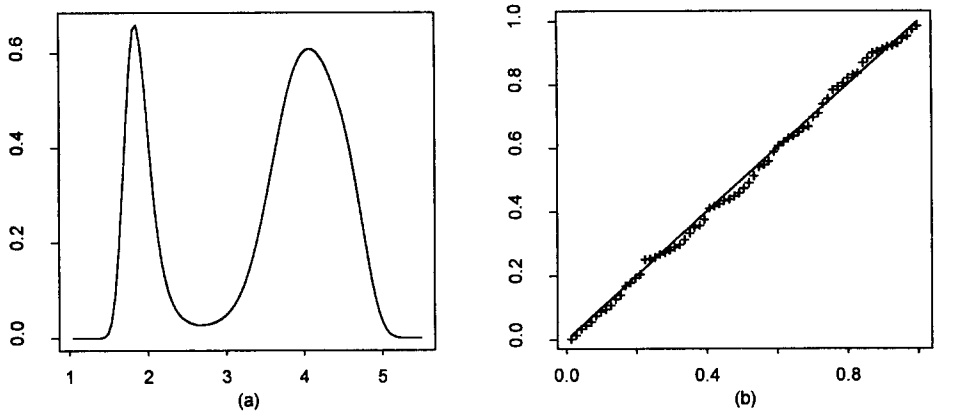
한편, Barron and Sheu(1991)는 4차 다항식에 의한 지수족(exponential family)밀도 추정량을 구하였다. 시각적으로 비교했을 때 우리의 추정량에서는 왼쪽 정점이 높게, 폭은 좁게 나타났으며, 전체적으로 높게 나타났다. 그 이유는 왼쪽정점의 폭이 좁고 2.6 근처에서의 추정량의 값이 작게 나타났기 때문이다.

이 분석 결과에 대한 보충으로 모의실험을 수행하였다. <그림 3>과 <그림 4>는 간단한 모형으로서 표준정규분포로부터 각각 $n=100$, $n=200$인 난수를 생성하여 얻은 모의자료에 대한 추정량이다. 난수는 Press외 3인(1991)의 gasdev를 이용하여 생성하였다. <그림 3(a)>의 로그-푸리에 추정량의 시각적인 적합도는 상당히 좋은데 비하여, 그의 누적도표 3(b)에 나타난 로그-푸리에 추정량의 꼬리 형태(behavior)는 상대적으로 떨어진다. 그러나 $n=200$인 경우 <그림 4>에는 밀도추정량 및 누적분포함수 추정량이 $n=100$에 비해 더 잘 적합되었음을 알 수 있다.

<그림 5>와 <그림 6>은 밀도함수

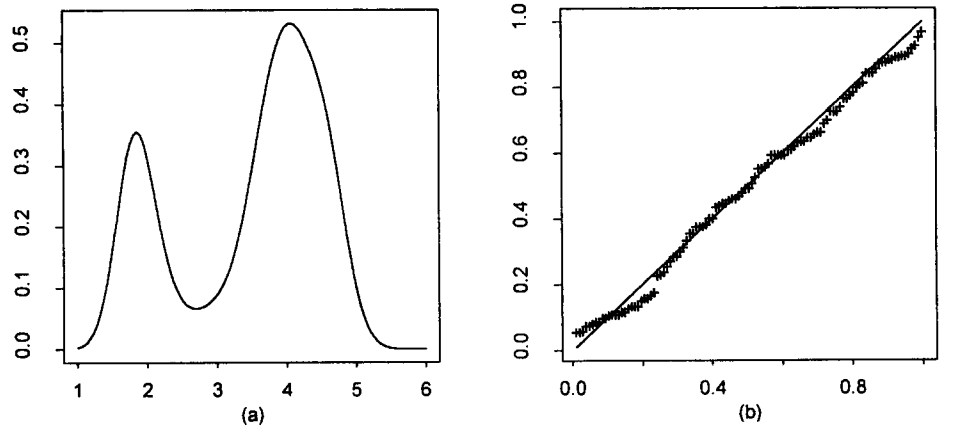
$$.3N(1.75,(\frac{1}{5})^2)+.7N(4,(\frac{1}{2})^2)$$

로부터 얻은 모의자료에 대한 실험결과로서 각각 $n=100$ 및 $n=200$인 경우들을 고려하였다. 이 밀도함수는 Geyser자료와 유사한 형태를 가지도록 고안되었다. 표준정규분포에 대한 결과분석과 비슷한 결론을 얻을 수 있다.



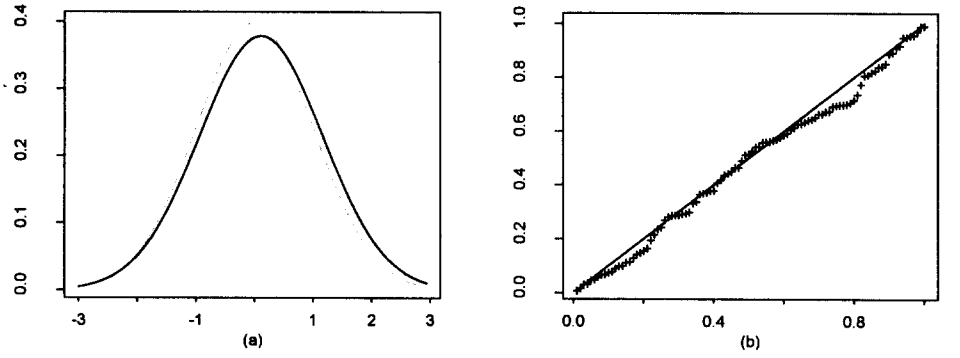
<그림 1> (a) 로그-푸리에 밀도추정량

(b) 누적분포표:+점은 $(\frac{i}{n}, \hat{F}(X_{1i}))$을 나타내며, 직선은 $y=x$ 임

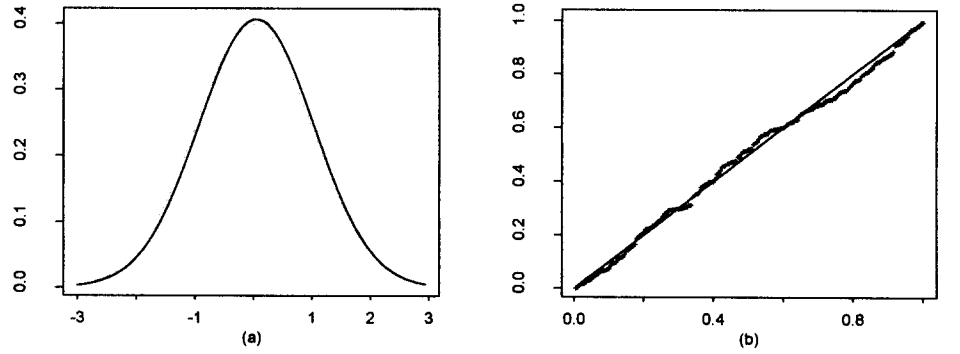


<그림 2> (a) 밴드길이 0.25이고 표준정규밀도함수를 이용한 커널밀도추정량

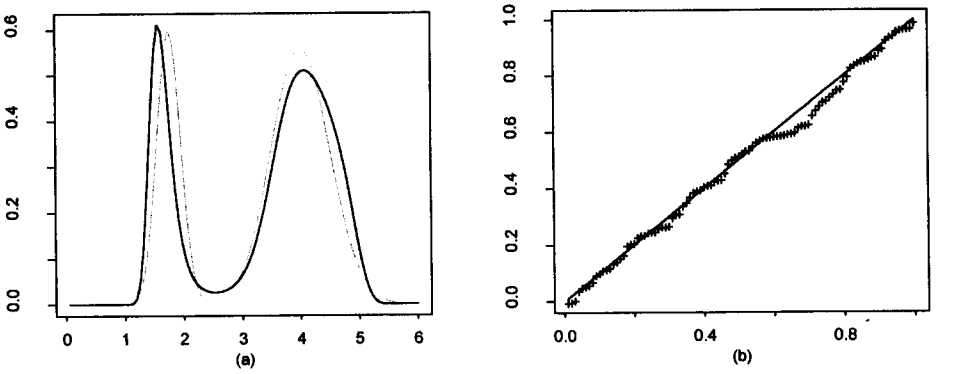
(b) 누적분포표:+점은 $(\frac{i}{n}, \hat{F}_K(X_{1i}))$을 나타내며, 직선은 $y=x$임



<그림 3> 모의실험($n=100$)
(a) 로그-푸리에 밀도추정량 : 실선은 로그-푸리에 추정량
점선은 추정 대상인 밀도함수(pdf)
(b) 누적분포표 : +점은 $(\frac{i}{n}, \hat{F}(X_{1i}))$을 나타내며, 직선은 $y=x$ 임



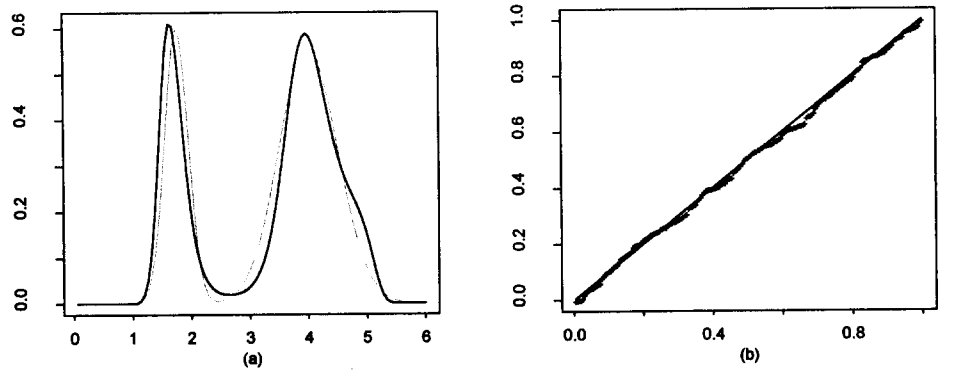
<그림 4> 모의실험($n=200$)
(a) 로그-푸리에 밀도추정량 : 실선은 로그-푸리에 추정량
점선은 추정 대상인 밀도함수(pdf)
(b) 누적분포표 : +점은 $(\frac{i}{n}, \hat{F}(X_{1i}))$을 나타내며, 직선은 $y=x$임



<그림 5> 모의실험($n=100$) $.3N(1.75,(\frac{1}{5})^2)+.7N(4,(\frac{1}{2})^2)$의 경우

(a) 로그-푸리에 밀도추정량 : 실선은 로그-푸리에 추정량
점선은 추정 대상인 밀도함수(pdf)

(b) 누적분포표 : +점은 $(\frac{i}{n}, \hat{F}(X_{1i}))$을 나타내며, 직선은 $y=x$ 임



<그림 6> 모의실험($n=200$) $.3N(1.75,(\frac{1}{5})^2)+.7N(4,(\frac{1}{2})^2)$의 경우

(a) 로그-푸리에 밀도추정량 : 실선은 로그-푸리에 추정량
점선은 추정 대상인 밀도함수(pdf)

(b) 누적분포표 : +점은 $(\frac{i}{n}, \hat{F}(X_{1i}))$을 나타내며, 직선은 $y=x$ 임

6. 토 의

본 논문에서는 푸리에 전개에 의한 밀도 추정량에 대해 연구하였다. 보통의 직교열 추정법에 비해 추정량이 양수값을 갖는다는 장점을 가지고 있다. 또한, 추정량을 결정하는 기저들은 자료에만 의존하는 방법에 의해 선택함으로써 탐색적 자료 분석에 이용될 수 있을 것으로 기대된다. 그러나 $n=100$정도에서 꼬리확률(tail probability) 계산 등과 같은 확증적 자료분석에의 이용에는 문제가 있을 수 있다.

Geyser자료에 대하여 본 논문에서 제시한 방법이 최근 여러 가지 형태로 개량된 커널방법들보다 성능이 우월하다는 주장을 할 수는 없다. 한 심사위원이 언급한 바와같이 Silverman(1986)에서는 Geyser자료에 대한 단순한 커널 추정치는 분포의 형태(즉 두 개의 모드(mode)를 가진 분포)를 파악하기 위한 커널추정방법의 예로서 제시된 것이지 정확히 그 자료의 밀도함수를 추정하기 위한 것이 아니다. 이 자료의 확률변수 X는 우선 $X>0$인 경계(boundary)가 존재하며, 또한 두 개의 모드(mode)가 존재함으로 두 모드(mode)근처에서의 X값들은 다른 부분에 비하여 조밀하게 관측되므로 동일한 창폭(window width)을 사용한 단순한 커널추정방법을 적용하기 곤란함을 알 수 있다. 따라서 개량된 커널방법(Local kernel method, Variable kernel method, 혹은 Boundary kernel method 등 : Wand와 Jones(1995) pp40-49, Silverman(1986) pp21-22, pp100-110)을 적용한 커널추정치와 로그-푸리에방법 추정치를 비교하는 것이 타당하다고 판단된다. 위의 개량된 커널방법을 적용하면 <그림 2(a)>의 두 봉우리는 폭은 좁게 그리고 높이는 조금 더 높게 추정되어 <그림 2(b)>의 경계(boundary)에서의 문제를 완화시키게 되어 <그림 1>의 로그-푸리에 추정결과의 성능과 비슷하리라 기대된다. 그러나 본 논문의 저자들은 커널방법에 대해 잘 알지 못하여 개량된 커널방법을 구현(implement)할 수 없어 <그림 2>와 같은 간단한 커널방법과의 비교만을 할 수 있었다. 개량된 커널방법과의 성능비교에 대한 연구는 향후의 좋은 연구과제가 될 것으로 기대된다.

Koo와 Chung(1995)에서는 선형역문제에서의 로그밀도추정법에 대한 이론적 고찰을 하였다. 그 한 예로서 X에 측정오차가 더해진 경우에 X의 밀도함수를 구하는 디컨볼루션(deconvolution)에서는 본 논문의 추정방법이 직접적으로 응용될 수 있을 것으로 기대된다.

감사의 글

이 글을 여러 가지로 도움이 되는 말씀을 해 주셔서 더욱 발전된 논문이 될 수 있게 심사해 주신 심사위원들에게 심심한 사의를 표합니다.

참 고 문 헌

[1] Barron, A. R.와 Sheu, C. -H. (1991). Approximation of density functions by sequences of exponential families, *The Annals of Statistics*, Vol. 19, 1347-1369.

[2] Crain, B. R. (1974). Estimation of distributions using othogonal expansions, *The Annals of Statistics*, Vol. 2, 454-463.

[3] Friedman, J. H. (1991). Multivariate adaptive regression splines(with discussion), *The Annals of Statistics*, Vol. 19, 1-141.

[4] Koo, J. -Y. (1996). Bivariate B-splines for tensor logspline density estimation, *Computational Statistics and Data Analysis*, Vol. 21, 31-42.

[5] Koo, J. -Y.와 Chung, H. -Y. (1995). Log-density estimation in linear inverse problems, manuscript.

[6] Koo, J. -Y.와 Kim, W. -C. (1996). Wavelet density estimation by approximation of log-densities, *Statistics and Probability Letters*, Vol. 26, 271-278.

[7] Koo, J. -Y.와 Park, B. U. (1996a). B-spline deconvolution based on the EM algorithm, *Journal of Statistical Computation and Simulation*, Vol. 54, 275-288.

[8] Koo, J. -Y.와 Park, J. (1996b). Logspline density estimation under censoring and truncation, manuscript.

[9] Kooperberg, C. (1995). Density estimation for bivariate survival data, Technical report No. 296, Department ofstatistics, University of Washington.

[10] Kooperberg, C.와 Stone, C. J. (1991). A study of logspline density estimation. *Computational Statistics and Data Analysis*, Vol. 12, 327-347.

[11] Kooperberg, C.와 Stone, C. J. (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, Vol. 1, 301-328.

[12] Kooperberg, C., Stone, C. J.과 Truong, Y. K. (1995a). Hazard regression, *Journal of the American Statistical Association*, Vol. 90, 78-94.

[13] Park, H. S. (1994). 텐서 로그스플라인 밀도추정에 관한 연구, 한림대학교 석사학위논문.

[14] Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992). *Numerical Recipes in C*, Cambridge University Press.

[15] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, Vol. 6, 461-464.

[16] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapmanand Hall, London.

[17] Stone, C. J. (1990). Large sample inference for logspline model, *The Annals of Statistics*, Vol. 18, 717-741.

[18] Stone, C. J.과 Koo, C. -Y. (1986). Logspline density estimation, *Contemporary Mathematics*, Vol. 59, 1-15, Amer. Math. Soc., Providence, R. I.

[19] Wand, M. P.와 Jones, M. C. (1995). *Kernel Smoothing*, Monographs on Statistics and Applied Probability 60, Chapman & Hall, London.

Log-density estimation based on a Fourier expansion[5)]

Ja-Yong Koo[6)], Kee-Won Lee[7)] and Hyun-Suk Park[8)]

Abstract

In this paper we propose a logdensity estimation based on a Fourierexpansion. The basis functions consisting of trigonometric functions are determinded by stepwise addition and deletion and the Bayes Information Criterion, where the maximum likelihood method is used to estimate the parameters. Numerical examples using real data and simulated data are provided to show the performance of proposed method.

5) Research of Ja-Yong Koo was supported by the Basic Science Research Institute Program, Ministry of Education, Korea, 1996, Project No. 1418.

6) Associate Professor, Department of Statistics, Hallym University, Chunchon, Kangwon-Do, 200-702, Korea.

7) Associate Professor, Department of Statistics, Hallym University, Chunchon, Kangwon-Do, 200-702, Korea.

8) Ph. D. Candidate, Department of Statistics, Hallym University, Chunchon, Kangwon-Do, 200-702, Korea.