The Korean Communications
in Statistics Vol. 7, No. 1, 2000
pp. 13-22

Modified Local Density Estimation for the Log-Linear Density[1)]

Ro Jin Pak[2)]

Abstract

We consider local likelihood method with a smoothed version of the model density in stead of an original model density. For simplicity, a model is assumed as the log-linear density, then we were able to show that the proposed local density estimator is less affected by changes among observations, but its bias increases little bit more than that of the currently used local density estimator. Hence, if we use the existing method and the proposed method in a proper way, we would derive the local density estimator fitting the data in a better way.

1. 배경

자료 속에 존재하는 어떤 구조를 찾기 위한 국소 회귀 분석은 개념의 간결성과 유연성이라는 좋은 이론적 성질을 갖고 있는 비모수 회귀 분석의 한 형태로 회자되고 있다. 근간이 되는 연구는 Stone (1977) 그리고 Cleveland (1979) 에 이루어졌다. 자세한 요약을 원한다면 Hastie and Loader (1993)가 도움이 될 것이다. 국소 회귀 분석이 국소 우도 (local likelihood) 분석의 특별한 경우로써 Tibshirani and Hastie (1987)에 의해 연구되었고, 여러 가지 분석에 적용되었다 (Staniswalis, 1989). 그후, Hjort and Jones (1996)와 Loader (1996)에 의해 국소 우도 함수를 이용한 확률 함수 추정에 대한 연구가 독립적으로 이루어졌다.

$X_1, \ldots, X_n$를 밀도함수 $f(x)$ 를 갖는 독립적이고 동일하게 추출한 확률표본이라 하자. 전통적인 $f(x)$의 커널 추정량 $\hat{f}(x)$는 다음과 같이 정의된다.

즉, $\hat{f}(x) = n^{-1} \sum_{i=1}^{n} K_h(x_i - x)$, 여기서 $K_h(z) = h^{-1}K(h^{-1}z)$ 그리고 $K(\cdot)$는 커널 함수이다. Hjort와 Jones (1996)는 비모수 밀도 함수 추정량의 도출 과정에 모수적 추정법을 가미하여 $\hat{f}(x)$에 버금가는 때로는 보다 나은 일종의 반모수적인(semi-parametric) 밀도 함수 추정량 $\hat{f}(x)$를 다음과 같이 제안했다. $f(\cdot, \theta) = f(\cdot, \theta_1, \ldots, \theta_p)$라는 모수족이 주어지고, 확률변수 X에 대하여 $f(x)$를 국소적(locally)으로 가장 근사(approximate)하는 밀도 함수 추정량, $\hat{f}(x) = f(x, \hat{\theta}_1(x), \ldots, \hat{\theta}_p(x))$, 이라고 정의하는데, $\hat{\theta}_1(x), \ldots, \hat{\theta}_p(x)$는 뒤에서 자세히 소개할

1) 이 논문은 '99년 대전대학교 교내 학술 연구비 지원에 의한 결과 입니다.
2) (300-716) 대전광역시 동구 용운동 대전대학교 정보통계학과 조교수

방법으로 구하게되는 모수들의 추정량들이다. 이때, $\hat{f}(x)$를 국소 모수적 밀도 함수 추정량 (local parametric density estimator)라고 부른다.

먼저, 주어진 x에 대하여 Hjort와 Jones (1996)는 국소 우도 함수 (local likelihood function)을 다음과 같이 정의한다.

$$L_n(x,\theta) = \int K_h(t-x)\{\log f(t,\theta)\,dF_n(t) - f(t,\theta)dt\}$$

$$= n^{-1}\sum_{i=1}^{n} K_h(x_i-x)\log f(x_i,\theta) - \int K_h(t-x)f(t,\theta)dt, \qquad (1)$$

여기서, $F_n(x)$는 $f(x)$의 누적 분포 함수이다. 그리고, 추정량들의 벡터 $\hat{\theta}$는 식 (1)을 최대화하는 다음의 p개의 모수들에 대한 방정식들의 해들로 정의된다.

$$n^{-1}\sum_{i=1}^{n} K_h(x_i-x)v(x,x_i,\theta) - \int K_h(t-x)\,v(x,t,\theta)f(t,\theta)f(t,\theta)dt = 0,$$

여기서 $v(x,t,\theta)$는 $(\partial/\partial\theta)\log f(t,\theta)$로 정의되는 $p\times 1$스코어 함수이다.

한편, Basu와 Lindsay (1994)는 당시에 널리 연구되던 최소간격추정법 (minimum distance estimation)을 재조명하면서 밀도 함수 대신 평활 밀도 함수 (smoothed density function), 즉 $\int K_h(t-x)f(x)\,dt$ 를 사용하면, 점추정량의 점근적 극한 이론의 전개가 용이하고 로우버스트의 성질이 보장됨에 대하여 논하였다. 본 논문은 Basu와 Lindsay (1994)의 기법을 국소 밀도 함수 추정에 적용하면 어떤 효과를 갖게되는지 생각해 보고자 한다.

2. 국소 지수-선형 밀도 함수

국소 함수 추정에서 일반적으로 $\log f(t)$가 주어진 x 근방에서 낮은 차수의 다항식

$$\log f(t) \approx P(t-x) = a_0 + a_1(t-x) + \ldots + a_p(t-x)^p$$

에 어느 정도 잘 근사 한다고 가정한다. 그러나, 굴곡이 심한 다항식을 이용한 국소 밀도 함수 추정은 반모수 추정에서는 크게 매력적이지 못하다 (Hjort and Jones, 1996). 따라서, 본 논문에서는 다소 간단한 그러나 실용적인 모형을 가정하고자 한다. 주어진 모델 함수 $f(x)$에 대하여 국소 모델 $a\exp\{b(t-x)\}$을 가정하자. 우선 국소 우도 함수를 최대로 하는 a와 b를 구하기 위해 a와 b에 대하여 아래 방정식의 해를 구해야 한다.

$$n^{-1}\sum_{i=1}^{n} K_h(x_i-x)\begin{pmatrix} 1/a \\ x_i-x \end{pmatrix} = \int K_h(t-x)\begin{pmatrix} 1/a \\ t-x \end{pmatrix} ae^{(b(t-x))}dt.$$

Loader (1996)에 의해 국소 모수적 밀도 함수 추정량 $\hat{f}(x)$는 $\exp(\hat{a})$로 정의됨으로, 이에 따라 Hjort와 Jones (1996)는

$$\hat{f}_{HJ}(x) = \hat{f}(x)\exp[-\frac{1}{2}h^2\{\hat{f}'(x)/\hat{f}(x)\}^2]$$

가 됨을 보였다. 위 식에서 $\hat{f}(x)$, $\hat{f}'(x)$는 일반적인 밀도 함수와 그것의 일차 도함수에 대한 각각의 커넬 추정량들이다.

이제, 본 논문에서 제안하고자 하는 방법에 따라 밀도 함수 대신 평활 밀도 함수를 사용하는 경우의 추정량 $\hat{f}(x)$을 구해보자. 편이상, 커널함수를 가우시안으로 정의하자. 이러한 경우 평활 밀도 함수는 $f_s(t)=a\exp\{b^2h^2/2+b(t-x)\}$가 된다. $\hat{a}$, $\hat{b}$는 다음 식을 a,b에 대하여 최대로 하는 값들이 된다.

$$n^{-1}\sum_{i=1}^{n}K_h(x_i-x)\left(\frac{1/a}{bh^2+(x_i-x)}\right)$$

$$=\int K_h(t-x)\left(\frac{1/a}{bh^2+(t-x)}\right)a\exp\{\frac{1}{2}b^2h^2+b(t-x)\}dt.$$

위 식은 다시 a,b 에 대하여 다음과 같이 두 개의 식으로 쓸 수 있고,

$$\frac{1}{a}\hat{f}(x) = a\exp(\frac{1}{2}b^2h^2)\frac{1}{a}\int K_h(t-x)\exp\{b(t-x)\}dt$$

$$bh^2\hat{f}(x)+\tilde{g}(x) = a\exp(\frac{1}{2}b^2h^2)\Big[bh^2\int K_h(t-x)\exp\{b(t-x)\}dt$$

$$+\int K_h(t-x)(t-x)\exp\{b(t-x)\}\Big]dt,$$

이제, $K_h(t-x)$ 에 가우시안 커널 함수를 넣어 다시 정리하면,

$$\frac{1}{a}\hat{f}(x) = a\exp(\frac{1}{2}b^2h^2)\Big[\frac{1}{a}\exp(\frac{1}{2}b^2h^2)\Big]$$

$$bh^2\hat{f}(x)+\tilde{g}(x) = a\exp(\frac{1}{2}b^2h^2)\Big[2bh^2\exp(\frac{1}{2}b^2h^2)\Big]$$

을 얻게 된다. 이제,

$$\hat{a}=\hat{f}\exp(-\hat{b}^2h^2),\quad \hat{b}=\frac{1}{h^2}(\hat{g}(x)/\hat{f}(x))=\hat{f}'(x)/\hat{f}(x)$$

가 됨으로, 커널을 가우시안으로 갖는 평활밀도 함수를 이용한 국소 밀도 함수 추정량은

$$\hat{f}_s(x)=\hat{f}(x)\exp(-\hat{b}^2h^2)=\hat{f}(x)\exp[-h^2\{\hat{f}'(x)/\hat{f}(x)\}^2] \tag{2}$$

가 된다. 새로이 구한 추정량과 Hjort와 Jones (1996)의 추정량

$$\hat{f}_{HJ}(x)=\hat{f}(x)\exp(-\frac{1}{2}\hat{b}^2h^2)=\hat{f}(x)\exp[-\frac{1}{2}h^2\{\hat{f}'(x)/\hat{f}(x)\}^2] \tag{3}$$

을 비교해 보면, 밀도 함수의 기울기를 추정하는 $\hat{f}'(x)$를 $\hat{f}(x)$로 표준화한 함수를 이용한 $\hat{f}(x)$에 대한 가중치 함수의 값이 새로운 추정량 $\hat{f}_s(x)$ 에서는 기존의 추정량 $\hat{f}_{HJ}(x)$의 경우보다 지수적으로 감소되었음을 알 수 있다. 즉, 커널 밀도 함수 추정량 $\hat{f}(x)$ 의 변화가 x주변에서 급격히 감소 또는 증가할수록, 즉, $\hat{f}'(x)$의 절대값이 클수록, $\hat{f}(x)$에 대하여 기존의 추정량의 경우 보다 낮은 가중치를 부여하게 된다. 낮은 가중치를 곱하게 됨으로 새로운 추정량의 경우에는 추정 함수의 급격한 변화를 초래하는 자료의 영향력을 어느 정도 제어시킬 수 있게된다. Hjort와 Jones (1996)는 적당한 전제 조건 아래에서 국소 모수적 밀도 함수 추정량의 모델 하에서의 영향함수 (influence function)를 다음과 같이 구했다.

$$IF(F,t)=J_h^{-1}\{K_h(t-x)u(t,\theta_0)-\int K_h(t-x)u(t,\theta_0)f(t,\theta_0)dt\}$$

여기서, θ_0는 모델 하에서의 모수의 참값이고, $u(t,\theta_0)$는 모델의 스코어 함수, 그리고 $J_h = \int K_h(t-x)\,u(t,\theta_0)u(t,\theta_0)^T f(t,\theta_0)dt$ 이다.

이때, 커넬이 가우시안인 경우 Hjort와 Jones (1996)의 공식에 의해 영향함수를 구하면,

$$IF(F,t) = J_h^{-1}(F,t)\left\{\begin{pmatrix} 1/a \\ t-x \end{pmatrix}(2\pi)^{-1/2}e^{\{-\frac{1}{2h^2}(t-x)^2\}} - e^{(\frac{1}{2}b^2h^2)}\begin{pmatrix} 1 \\ 2bh^2 \end{pmatrix}\right\}$$

가 된다.

이제, 본 논문에서 제안한 방법대로 평활밀도 함수를 사용하여 영향 함수를 구해보자. 먼저, $f_s(t) = f(t)\exp(b^2h^2/2)$ 이므로 스코어 함수는 $f(t)$와 $f_s(t)$를 사용하는 경우 모두 동일한 형태를 갖는다. 따라서, 우리는 또한 $J_h(F_s,t) = J_h(F,t)\exp(b^2h^2/2)$임을 쉽게 알 수 있다. 결국, 평활 모델 함수에 대한 영향 함수는 아래와 같이 된다. 즉,

$$\begin{aligned} IF(F_s,t) &= e^{(-\frac{1}{2}b^2h^2)}J_h^{-1}(F,t)\left[\begin{pmatrix} 1/a \\ t-x \end{pmatrix}(2\pi)^{-1/2}e^{\{-\frac{1}{2h^2}(t-x)^2\}}\right. \\ &- \left. e^{(\frac{1}{2}b^2h^2)}\int(2\pi)^{-1/2}e^{\{-\frac{1}{2h^2}(t-x)^2\}}\begin{pmatrix} 1/a \\ t-x \end{pmatrix}f(t)dt\right] \\ &= J_h^{-1}(F,t)\left[e^{(-\frac{1}{2}b^2h^2)}\begin{pmatrix} 1/a \\ t-x \end{pmatrix}(2\pi)^{-1/2}e^{\{-\frac{1}{2h^2}(t-x)^2\}} - e^{(\frac{1}{2}b^2h^2)}\begin{pmatrix} 1 \\ 2bh^2 \end{pmatrix}\right]. \end{aligned}$$

이때, $IF(F_s,t)$의 전체적인 모양(shape)은 $IF(F,t)$와 동일하나, 그 수직폭이 $IF(F,t)$의 $\exp(-b^2h^2/2)$배로 좁아진 형태를 지님을 알 수가 있다. 즉, $f_s(t)$를 이용하는 경우에 $f(t)$를 이용할 때 보다 x의 주변에서의 영향력이 $\exp(-b^2h^2/2)$배로 감소한다고 하겠다.

그런데 편의는 어떤가? Hjort와 Jones (1996)에 의하면, 가우시안을 커넬 함수로 사용하는 경우, 주어진 상수 c에 대하여 다음과 같은 국소 밀도 함수 추정량

$$\hat{f}(x) = \tilde{f}(x)\exp[-ch^2(x)\{\tilde{f}'(x)/\tilde{f}(x)\}^2]$$

이 구해지고, 편의는 $(1/2)h^2[f''(x) - 2c\{f'(x)/f(x)\}^2] + O(h^4)$으로 구해진다. 식 (2)와 (3)에 의하면, 평활 밀도 함수를 사용하는 추정량은 c가 1인 경우가 되고, Hjort와 Jones (1996)의 경우는 c가 1/2로 보면 된다. 따라서, 평활한 함수를 사용하는 경우 그렇지 않은 경우 보다 편의가 절대값으로 $(1/2)h^2c\{f'(x)/f(x)\}^2$만큼 증가한다. 만일 $f(t)$가 실제 모델인 $a\exp\{b(t-x)\}$와 일치한다면, 지수 선형 밀도 함수를 가정하는 경우의 편의가 $O(h^4)$로 되어 편의가 거의 없으나, 평활화 시킨 지수-선형 함수를 사용하는 경우에는 $-ab^2\exp\{b(t-x)\} + O(h^4)$가 되어 바이어스가 어느 정도 존재함을 알 수 있다. 편의는 영향함수의 경우와는 반대의 현상을 보여주고 있다. 결국, 영향력을 감소시키기 위해 평활한 함수를 사용하는 경우 본래의 밀도 함수가 아닌 약간 변형된 밀도 함수를 사용하게 됨으로 편의 상승을 막을 수가 없다고 보여진다.

3. 모의 실험

모의 실험에서는 다음의 세 가지 분포를 가정했다.

(1) beta(2,2);

(2) beta(2,2)와 beta(20,20)의 1 대 1의 혼합분포

$$f(x)=\frac{1}{2}\frac{3!}{1!^2}x(1-x)+\frac{1}{2}\frac{39!}{19!^2}x^{19}(1-x)^{19},\quad x\in[0,1];$$

(3) beta(2,2)과 beta(40,40)의 1대 1의 혼합분포

$$f(x)=\frac{1}{2}\frac{3!}{1!^2}x(1-x)+\frac{1}{2}\frac{79!}{39!^2}x^{39}(1-x)^{39},\quad x\in[0,1].$$

각 분포에서 표본의 크기가 200인 모의 자료를 1000번 생성하여, 함수를 추정한 후, 제 10백분위수와 제 90백분위수를 점선으로 이어 그렸다. 변화가 어느 정도 일정하게 유지되는 양쪽 끝과 중심부에서는 일반적으로 사용되는 커널 함수 추정량이 좋은 결과를 보여 주고 있다. <그림 1>에서 보면 단조로운 상승과 하강의 구조를 갖는 밀도 함수의 추정의 경우, 일반적인 커널 밀도 함수 추정량이 가장 적절한 것으로 여겨진다. <그림2>에서 보이듯이 어느 정도 기울기가 변하는, 즉 변곡이 일어나는 0.3와 0.7 근처에서 현재 사용되는 국소 우도 밀도 함수 추정량이 가장 균형있게 적합하고 있음을 알 수 있다. 한편, 기울기의 변화가 심한 경우 <그림 3>에서 보듯 급격한 변화에 영향을 덜 받는 평활 밀도 함수를 이용한 추정량이 변곡이 일어나는 0.4와 0.6 근처에서 가장 균형 있게 적합하고 있음을 알 수 있다. 기울기 변화의 정도에 따라 어떤 함수 추정량이 적당한 가를 결정하는 수학적 근거를 마련한다는 것은 사실 저자의 능력 밖인 것 같으나, 새로이 제한된 함수 추정량을 기존의 추정량들과 함께 사용하면 보다 유용한 정보를 얻을 수 있겠다고 사려된다.

실제 자료에 적용하기 위해 미국 엘로우스톤의 간헐샘인 Old Faithful의 분출(eruption) 자료 (Venables and Ripley, 1994)의 밀도 함수 추정량을 구하여 보았다. <그림 4>에서 보듯이, 본 논문에서 제시한 평활 밀도 함수를 이용한 국소 밀도 함수 추정량이 기존의 함수 추정량들 보다 자료가 두 개의 서로 다른 밀도 함수들의 혼합체에서 기인하지 않았는가 하는 추측이 더욱 가능하도록 하고 있다. 실제로 'waiting'이라는 변수를 'eruptions'와 함께 그려보면 우리의 추측이 옳다는 것을 시각적으로 확실히 보여 주고 있다.

4. 마무리

우리는 확률 밀도 함수를 평활 하여 국소 모수적 밀도 함수 추정을 실시할 때, 그렇지 않은 경우 보다 항상은 아니지만 변곡이 심한 특정한 함수 추정에 있어서 다소 효과가 있음을 간단하게 설명했다. 위 이론은 지수-선형 밀도 함수를 중심으로 한 아주 초보적인 결과이지만 보다 복잡한 모형에도 적용이 가능하리라 생각된다. 또한, 위 방법이 경계 (boundary)에서의 추정에 어떤 효과가 있는지도 한 가지 연구 주제가 될 것으로 생각된다. 위의 모의 실험에서는 평활 계수는 가장 일반적인 $1.06(MAD)n^{-1/5}$ (MAD는 Median Absolute Deviance)를 사용하고, Old Faithful 예에서는 Sheather & Jones (1991)가 제안한 데이터에 근거한 평활 계수를 사용하였으나, 국소 밀도 추정에 맞는 평활 계수에 대한 연구도 한 가지 과제로 남아 있다.

참고문헌

[1] Basu, A. and Lindsay, B. G. (1994). Minimum Dispartiy Estimation for Continuous Model: Efficiency, Distribution, and Robustness, *Annals of the Institute of Statistical Mathematics,* 46, 683-705.

[2] Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots, *Journal of the American Statistical Association*, 74, 829-836.

[3] Hastie, T. and Loader, C. (1993). Local Regression: Automatic Kernel Carpentry (with discussion), *Statistical Sciences*, 8, 120-143.

[4] Hjort, N. L. and Jones, M. C. (1996). Locally Parameter Nonparametric density estimation, *Annals of Statistics*, 24, 1619-1647.

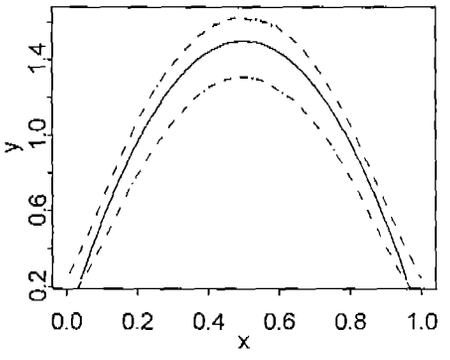
[5] Loader, C. R. (1996). Local Likelihood Density Estimation, *Annals of Statistics*, 24, 1602-1618.

[6] Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society* B, 53, 683-690.

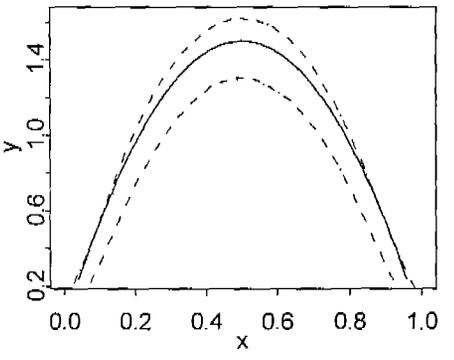
[7] Staniswalis, J. (1989). The Kernel Estimate of A Regression Function In Likelihood-Based Models, *Journal of the American Statistical Association*, 84, 276-283.

[8] Stone, C. J. (1977). Smoothing Bias In Density Derivative Estimation, *Annals of Statistics*, 5, 595-620.

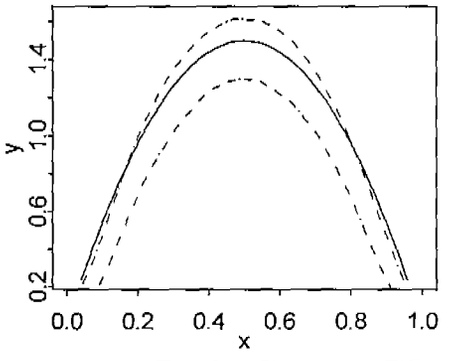
[9] Venables, W. N. and Ripley, B. D. (1994). *Modern Applied Statistics with S-Plus,* Springer, New York.



(1). 커넬 밀도 함수 추정량

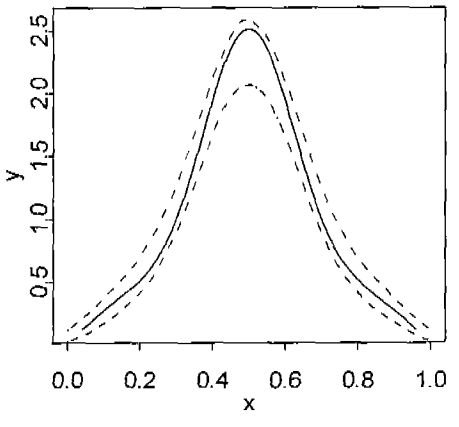


(2). 기존의 국소 우도 밀도 함수 추정량

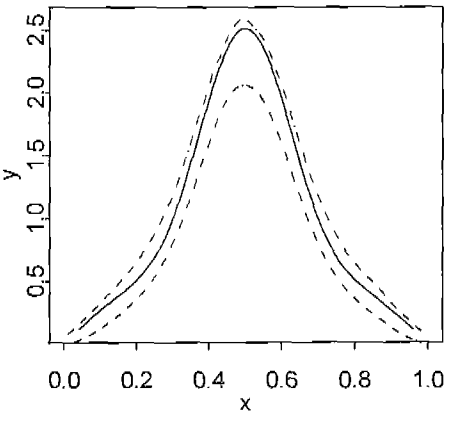


(3). 평활 국소 우도 밀도 함수 추정량

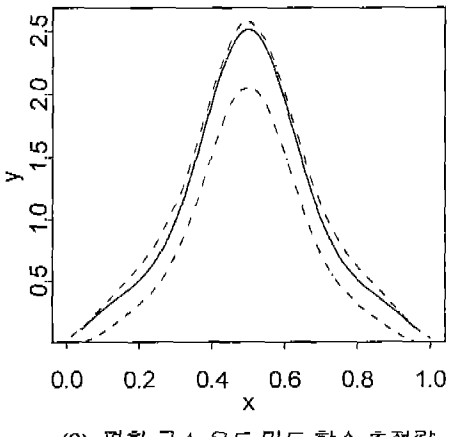
<그림1> beta(2,2)의 추정



(1). 커넬 밀도 함수 추정량

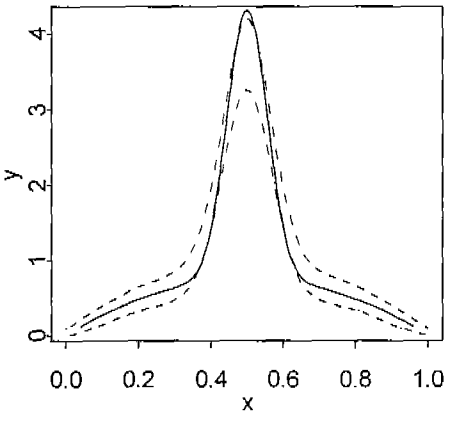


(2). 기존의 국소 우도 밀도 함수 추정량

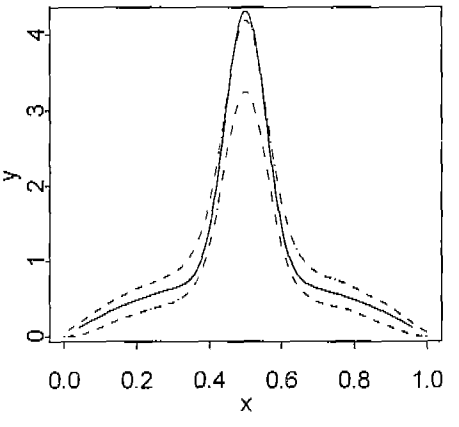


(3). 평활 국소 우도 밀도 함수 추정량

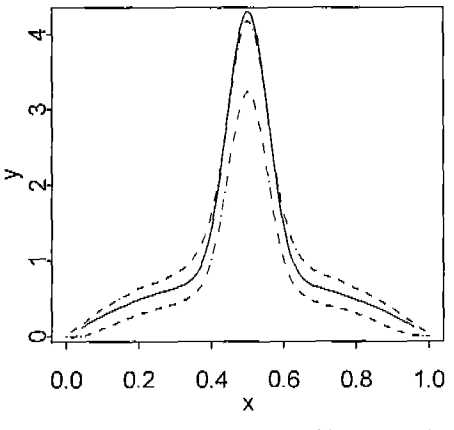
<그림 2> $\frac{1}{2}\,\mathrm{beta}(2,2)+\frac{1}{2}\,\mathrm{beta}(10,10)$의 추정



(1). 커넬 밀도 함수 추정량

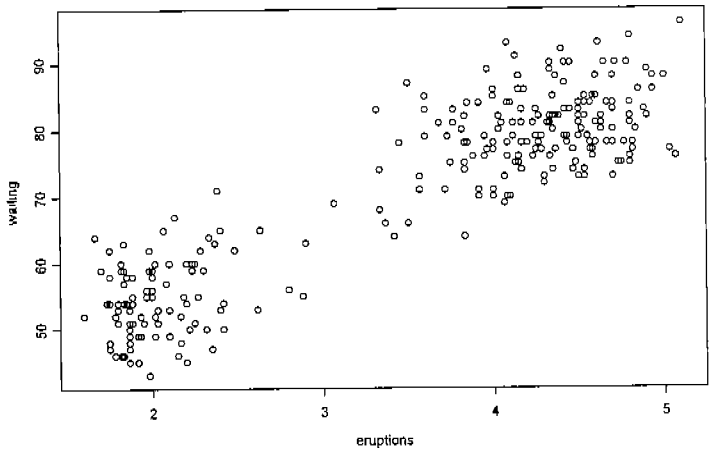


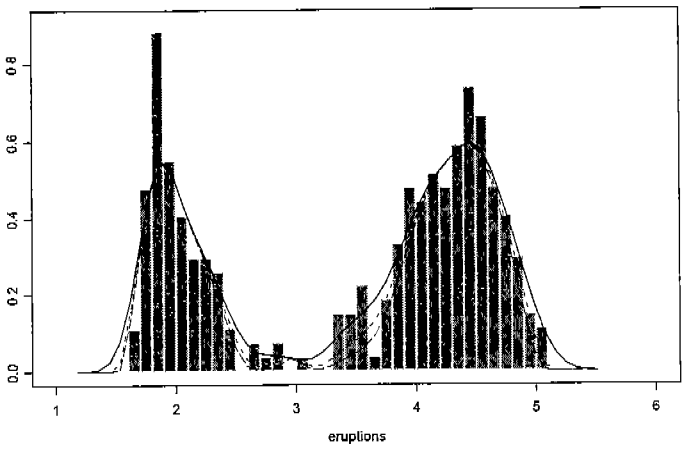
(2). 기존의 국소 우도 밀도 함수 추정량



(3). 평활 국소 우도 밀도 함수 추정량

<그림 3> $\frac{1}{2}\,\mathrm{beta}(2,2)+\frac{1}{2}\,\mathrm{beta}(40,40)$의 추정





<그림 4> The Old Faithful eruptions data에 적용한 추정법; 일반적인 밀도함수 추정량(실선), 기존의 국소밀도함수 추정량(얇은 점선), 제안된 국소밀도함수 추정량(굵은 점선). 그리고 eruptions와 waiting 의 2차원 산점도.