응용통계연구 제 11권 2호, 1998년, pp. 363-376

이차 평활스플라인 *

김종태[1)]

요 약

선형 평활스플라인 추정은 경계 편의의 영향력을 제거 하기위해 수정된 것이다. 제시된 추정량은 적합된 값들과 관련있는 평활 모수 선택 기준의 계산을 개선시킨 $O(n)$ 알고리즘을 사용하여 효과적으로 계산할 수 있게 하였다. 추정량의 점근적 성질들이 균일 계획의 경우에 대하여 연구되었다. 이 경우에 경계수정 선형 평활스플라인들의 평균 제곱오차의 성질들은 표준 이차 커널 평활들에 대한 평균제곱오차들과 점근적 특성으로 비교하였다.

1. 서 론

평활스플라인 (Smoothing splines)은 비모수 회귀분석을 하는데 많이 사용되는 도구 중 하나이다. 그러나 평활스플라인은 경계 지역에서의 그들 고유의 추정을 조절하는 방법에 의해서 생기는 "경계 편의 (boundary bias)"의 문제점을 가지는 것으로 잘 알려져 있다. 이러한 경계 영향 (boundary effects)들을 제거시키는 일반적인 방법들이 Eubank와 Speckman (1991)과 Oehlert (1992)에 의해 제시되었다.

본 논문의 목적은 선형 평활스플라인의 경우에 있어서 Eubank와 Speckman이 제시한 경계수정 (boundary correction) 방법의 특성들을 개발하는데 있다. 본 논문에서 제시된 추정량은 평활 수준 (level of smoothing)의 정도를 자동적으로 선택 하는 방법을 사용함으로서 자료를 신속히 적합 시킬 수 있는 단순 평활들을 제공한다. 또한 추정량의 점근적 성질들을 균일 설계 (uniform design)인 경우에 대하여 조사하였다.

이제 비모수 회귀문제에 대하여 생각해 보자. 반응변수 $y_1, \ldots, y_n$이 다음의 모형 (1.1)로부터 동시에 발생하지 않는 설계점들 (non-coincident design points), $0 \le t_1 < \cdots < t_n \le 1$에서 구해진다고 하자.

$$y_i = \mu(t_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{1.1}$$

여기서 μ는 미지의 회귀함수이고, $\epsilon_1, \ldots, \epsilon_n$은 평균 0과 공통 분산 σ^2를 가지는 비상관 랜덤오차들 (uncorrelated random errors)이다.

회귀 모형 (1.1)에 있는 μ에 대해 적용될수 있는 추정량은 수 없이 많다. 본 논문에서의 관심의 대상이 되는 추정량은 다음의 식 (1.2)에서, 제곱 적분가능 도함수를 가지는 모든 절

* 이 논문은 1998년도 대구대학교 학술연구비 지원에 의한 연구임

1) (712-714) 경북 경산시 진량면, 대구대학교 자연과학대학 통계학과, 조교수

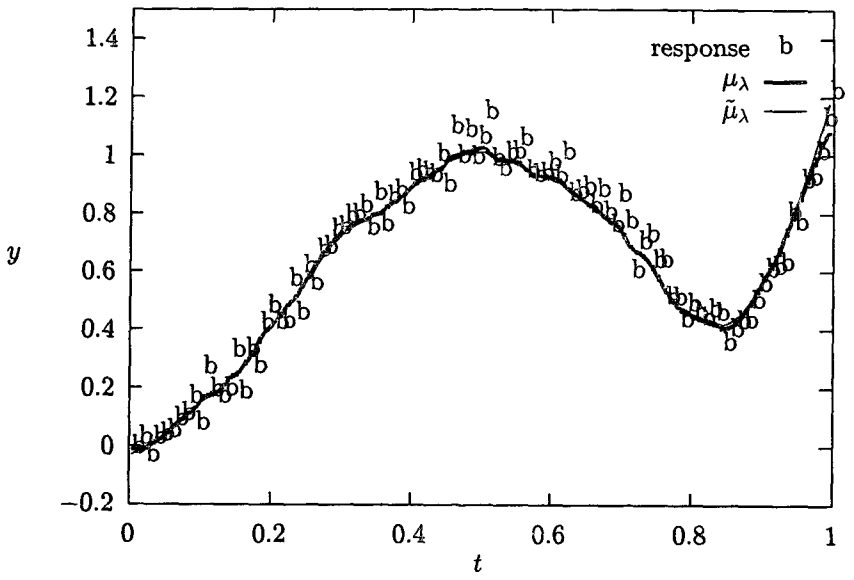

그림 1.1: 모의실험 데이타에 대한 평활스플라인의 적합.

대 연속함수 f에 대하여, 판단기준 식 (1.2)를 최소화 시킴으로서 구해지는 선형 평활스플라인 추정량 μ_λ이다.

$$n^{-1}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda\int_{t_1}^{t_n} f'(t)^2 dt, \quad \lambda > 0. \tag{1.2}$$

판단기준 식 (1.2)는 유일한 최소값을 가지며, 자신의 도함수가 0과 1에서 소멸되는 특성을 가지고 있는데, 위의 판정기준 식 (1.2)의 유일한 최소값은 각 설계점들 (design points)에서 추정값들을 가지는 선형 스플라인 추정이다. 식 (1.2)에 있는 λ의 값은 평활 모수 (smoothing parameter)이고, 데이타에 대한 추정량의 적용에 있어서의 평활 수준의 정도를 조절한다. λ의 값은 전형적으로 자료 조정 방법 (data-driven methods)에 의해 선택된다. 예를 들어, 그림 1.1은 λ를 선택하기 위해 일반적 교차 타당성 기준 (generalized cross validation)을 이용하여 구한 선형 평활스플라인 추정량을 적합시킨 것이다. 그림 1.1에 표현된 데이타는 (1.1)에 있는 모형에서 $\sigma = .05$를 가지는 정규 확률오차와 집합 A에 대한 지수함수 I_A를 가지는 $\mu(t) = 16(t(1-t))^2 + 48\mathrm{I}_{\{t>.7\}}(t-.7)^3$ 에 대하여 모의실험을 이용하여 구하였다.

μ_λ의 실행에 대하여, 다음 식 (1.3)의 평균제곱오차의 평균, 즉 위험 함수 (risk function), $\mathrm{R}_n(\lambda)$를 이용하여 μ_λ의 적합을 평가한다.

$$\mathrm{R}_n(\lambda) = n^{-1}\sum_{i=1}^{n}\mathrm{E}\,(\mu_\lambda(t_i) - \mu(t_i))^2. \tag{1.3}$$

그러면 식 (1.3)은 Eubank (1988, 6장)으로 부터 다음의 결과를 얻는다.

$$\inf_{\lambda} \mathrm{R}_n(\lambda) = O(n^{-2/3}). \tag{1.4}$$

즉, μ_λ는 μ의 일차 추정량이다. 만약 설계점이 균일하다면, 이결과는 Eubank (1997)가 지적하였듯이 $\sqrt{\lambda} \asymp n^{-1/4}$ 조건을 가지고 다음의 식 (1.5)에 대해 톱니모양이 될것이다.

$$\mathrm{R}_n(\lambda) \sim \frac{\sigma^2}{4n\sqrt{\lambda}} + \frac{\lambda^{3/2}}{2}[\mu'(0)^2 + \mu'(1)^2]. \tag{1.5}$$

선형 평활스플라인 추정량의 위험함수가 0과 1에서의 μ의 도함수 값에 의하여 표현되는 경계효과에 의해 조절되어진다는 것은 명백한 사실이다. 균등 설계의 경우에 있어서, 만약 $\sqrt{\lambda} \asymp n^{-1/5}$이라면, 다음의 식 (1.6)이 성립한다.

$$\mathrm{R}_n(\lambda) \sim \frac{\sigma^2}{4n\sqrt{\lambda}} + \lambda^2 \int_0^1 \mu''(t)^2 dt. \tag{1.6}$$

여기서 μ는 추정량으로서의 같은 경계 조건들, 즉

$$\mu'(0) = \mu'(1) = 0 \tag{1.7}$$

을 만족한다고 가정한다. 그러므로 선형 평활스플라인은 만약 μ가 식 (1.7)을 만족한다면, 실제로 μ에 대한 이차 추정량을 제공한다. 위와 같은 결론들은 Nychka (1995)의 연구에서 사용된 비균등 설계에 대해서도 적용할수 있다.

선형 평활스플라인들의 경계 편의에 대한 식 (1.5)와 같은 점근적인 결과는 비교적 매우 작은 표본의 크기에서도 적용되는 경향을 가진다. 예를들어, 그림 1.1에서 적합된 선형 평활스플라인의 경계 제한들은 시각적으로 잘 적합 되었음을 볼 수 있고, 추정량들은 1에 가까운 데이타들의 가파른 기울기를 해결하기 위해 매우 노력하였음을 볼 수 있다.

자연 선형 스플라인에 기저한 함수들에 대한 적절한 선택을 함으로서 $O(n)$ 작용들을 가지고 데이타에 잘 적합 되는 선형 평활스플라인들을 찾는 매우 효과적인 알고리즘을 개발할 수 있다. 그러나 추정량의 경계에서의 적합은 다소 제한적인 결론을 만들어 내는데 그 이유는 이차, 혹은 그 이상의 평활들이 일반적으로 선호되어 지며 데이타의 실제적인 적용에 있어서 식 (1.7)의 조건을 완정히 만족하는 것을 기대할 수 없다.

다음 절에서 μ_λ로 부터 경계효과들을 제거하기 위하여 Eubank와 Speckman (1991)의 경계수정 방법 (boundary correction method)을 사용할 것이다. 경계수정 방법으로 수정된 평활스플라인은 수정되지 않은 것보다 계산면에 있어서 보다 쉽고 그 계산을 위해 사용가능한 $O(n)$ 알고리즘을 제안한다. 경계수정 추정량의 점근적 성질은 3절에 연구된다. 여기서 이 점근적 성질이 점근적 평균 제곱오차 관점으로 부터 다른 표준 이차 평활 추정량들과 경쟁력을 가지는 것을 알수 있다. 모든 결과들에 대한 증명은 4절에서 있다.

2. 추정량의 계산

이절에서는 선형 평활스플라인의 경계수정 추정량 계산을 위한 $O(n)$ 알고리즘을 개발할 것이다. 먼저 식 (1.2)로부터 본래의 추정량을 계산하기 위한 효과적인 방법을 설명하고 난 뒤 경계수정이 어떻게 만들어지는가를 볼 것이다.

평활스플라인들이 선형추정량들이므로, 선형 평활스플라인에 대하여 적합된 값들의 벡타에 대해 반응 벡터 $\mathbf{y} = (y_1, \ldots, y_n)^T$를 변환한 $n \times n$ 행렬 $\mathbf{S}_\lambda$가 존재하고, 선형 평활스플라인은 다음과 같이 표현한다.

$$\mu_\lambda = (\mu_\lambda(t_1), \ldots, \mu(t_n))^T = \mathbf{S}_\lambda \mathbf{y}. \tag{2.1}$$

우리는 μ_λ에 대한 방정식의 구조를 해결하기 위한 효과적인 방법이 필요하다. 이러한 효과적인 방법을 얻기 위하여 본래의 선형 스플라인에 기저한 함수에 대한 편리한 사용을 위하여 다음 식 (1.9)와 같은 자연 선형 B-스플라인을 사용할 것이다.

$$\mathbf{S}_\lambda = (\mathbf{I} + n\lambda\mathbf{V})^{-1}. \tag{2.2}$$

여기서 $\mathbf{V}$는 한개의 비대각 띠 (off-diagonal band)를 가지는 대칭 행렬이다. $\mathbf{V} = \{v_{ij}\}$는 0이 아닌 요소들로서 다음과 같이 주어진다: $v_{11} = (t_2 - t_1)^{-1}, v_{nn} = (t_n - t_{n-1})^{-1}$,

$$v_{ii} = (t_{i+1} - t_i)^{-1} + (t_i - t_{i-1})^{-1}, i = 2, \ldots, n-1,$$

그리고

$$v_{i,(i+1)} = v_{(i+1),i} = -(t_{i+1} - t_i)^{-1}, i = 1, \ldots, n-1.$$

이러한 띠의 구성 때문에 우리는 다음과 같은 선형 방정식을 해결할 수 있다.

$$(\mathbf{I} + n\lambda\mathbf{V})\mathbf{c} = \mathbf{b}. \tag{2.3}$$

여기서 $\mathbf{b}$는 $\mathbf{I}+n\lambda\mathbf{V}$의 콜레스키 분해 (Cholesky decomposition)를 이용한 $O(n)$ 계산의 n-백터이다. $\mathbf{I} + n\lambda\mathbf{V}$의 콜레스키 분해는 다음과 같다.

$$\mathbf{I} + n\lambda\mathbf{V} = \mathbf{U}^T\mathbf{D}\mathbf{U}. \tag{2.4}$$

여기서 $\mathbf{D}$는 대각행렬이고, $\mathbf{U}$는 단위 대각 행렬들과 단지 하나의 비대각 띠를 가지는 하삼각 행렬 (lower triangular matrix) 이다. 특히, $\mathbf{b} = \mathbf{y}$이라면 이것은 데이타에 대한 선형 평활스플라인 적합을 만들어 낸다.

경계수정 추정량을 만들기 위해서 우리는 Eubank와 Speckman (1991)이 제시한 수정 방법을 이용할 것이다. 그들의 연구는 핵심적으로 같은 알고리즘을 사용하여 수정된 추정량을 계산한다. 이 알고리즘은 평활 다항식 (smoothed polynomials)를 사용한 일반적인 최소제곱 적합에 의해 얻어진 수정들만을 가지고도 본래의 적합에 유용하게 이용이 된다. 이 알고리즘을 실행하기 위하여 다음과 같이 두자.

$$q_0(t) = t - .5t^2, \tag{2.5}$$

$$q_1(t) = .5t^2. \tag{2.6}$$

그리고 $\mathbf{Q} = \{q_{ij}\}$는 $q_{ij} = q_{j-1}(t_i), j = 1, 2, i = 1, \ldots n$을 가지는 $n \times 2$ 행렬이다. 그러면 데이타에 대한 경계수정 추정은 다음과 같다.

$$\tilde{\mu}_\lambda = \mu_\lambda + \tilde{\mathbf{Q}}\mathbf{b}. \tag{2.7}$$

여기서

$$\tilde{\mathbf{Q}} = (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Q}. \tag{2.8}$$

그리고 $\mathbf{b}$는 다음 식 (2.9)의 해이다.

$$\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}}\mathbf{b} = \tilde{\mathbf{Q}}^T\tilde{\mathbf{y}}. \tag{2.9}$$

여기서 $\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}$ 이다. 이것은 $\tilde{\mu}_\lambda$를 계산하기 위하여, 차수 n 작용에 있어 $\tilde{\mathbf{Q}}$와 $\tilde{\mathbf{y}}$ 모두를 얻기위해 식 (2.3) - (2.4)을 사용하여 $\mathbf{Q}$의 두개의 열들과 $\mathbf{y}$를 변환시킬수 있다는 것을 의미한다. 그러면 식 (2.7)의 $\mathbf{b}$는 $\tilde{\mathbf{Q}}$의 열들에 있는 $\tilde{\mathbf{y}}$의 회귀함수에 대한 식 (2.9)의 2×2 정규방정식을 해결함으로 구한다. 그러므로 $\mathbf{b}$는 표준 통계 소프트웨어를 사용하여 계산하거나 혹은 2×2 행렬 $\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}}$의 직접적인 역전도에 의해 계산되어 진다.

일단 $\mathbf{b}$와 μ_λ가 계산되고 나면 $\tilde{\mu}_\lambda$와 잔차제곱합

$$\mathrm{RSS}(\lambda) = (\mathbf{y} - \tilde{\mu}_\lambda)^T(\mathbf{y} - \tilde{\mu}_\lambda)$$

이 식 (2.7)를 사용하여 $O(n)$ 작용에서 얻어질 수 있다. 그것은 또한 수정된 추정량의 평활 모수의 데이타 조정 (data-driven) 선택을 사용하는 것이 필요하다. 이것을 성취하는 하나의 방법은 다음의 일반화 교차 타당성기준을 최소화 시키는 λ의 값을 사용하는 것이다.

$$\mathrm{GCV}(\lambda) = \frac{n\mathrm{RSS}(\lambda)}{\mathrm{tr}(\mathbf{I} - \mathbf{H}_\lambda)^2}. \tag{2.10}$$

여기서

$$\mathbf{H}_\lambda = \mathbf{S}_\lambda + \tilde{\mathbf{Q}}(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T(\mathbf{I} - \mathbf{S}) \tag{2.11}$$

는 $\tilde{\mu}_\lambda$에 대한 해트행렬 (hat matrix) 혹은 평활 행렬이다.

이러한 판정기준을 사용하기 위하여서는 $\mathrm{tr}\mathbf{H}_\lambda$의 계산이 요구된다. 그 계산 절차는 다음과 같다.

먼저 Hutchinson과 de Hoog (1985)의 정리 3.1은 차수 n 계산들에 있어 $\mathrm{tr}\mathbf{S}_\lambda$를 구하기 위해 적용될 수 있다. 만약 식 (2.4)에 있는 행렬 $\mathbf{D}$가 대각 요소들을 $d_1, \ldots, d_n$ 가지고 $\mathbf{U}$의 비대각요소들이 $u_1, \ldots, u_{n-1}$이라면, 다음과 같은 $\mathbf{S}_\lambda$의 대각요소 s_{ii}들은 후향 귀납에 의해 계산되어 질수 있다.

$$s_{i(i+1)} = -u_i s_{(i+1)(i+1)}. \tag{2.12}$$

$$s_{ii} = d_i^{-1} - u_i s_{i(i+1)}. \tag{2.13}$$

여기서 초기값은 $s_{nn} = 1/d_n$이다. 이것은 식 (2.3)의 해를 위해 후향 대입 단계에서 구체화 되어지는데 이 이유는 $\mathbf{S}_\lambda$의 고유화 (trace)가 추정량들에 대한 적합되어진 값을 가지고 동시에 계산되어 질 수 있기 때문이다. $\mathrm{tr}\mathbf{H}_\lambda$의 나머지 부분은 다음과 같다.

$$\tau = \mathrm{tr}\ \tilde{\mathbf{Q}}(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T(\mathbf{I}-\mathbf{S}) = \mathrm{tr}\ (\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T\check{\mathbf{Q}}.$$

여기서 $\check{\mathbf{Q}} = (\mathbf{I}-\mathbf{S})\tilde{\mathbf{Q}}$이다. $\check{\mathbf{Q}} = [\check{\mathbf{q}}_1, \check{\mathbf{q}}_2]$라 하자. 그러면 우리는 정규방정식의 해에 의해 τ를 계산 할 수 있다.

$$\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}}\mathbf{a}_i = \tilde{\mathbf{Q}}^T\check{\mathbf{q}}_i, \quad i = 1, 2. \tag{2.14}$$

이것은 다시 표준 최소제곱 방법이나 직접적인 역도치법을 이용하여 계산되어 질 수 있다. 그 이유는 그 구조가 2×2이기 때문이다. 그 결과는 다음과 같다.

$$\tau = a_{11} + a_{22}. \tag{2.15}$$

여기서 a_{ii}는 (2.14)에 있는 $\mathbf{a}_i$의 i번째 요소이다.

우리는 $\log\lambda$에 있는 100 격자점들을 찾음으로서 GCV 판정기준의 최소화를 이용하여 자동적으로 선택되어진 λ의 값이 포함되는 알고리즘을 FORTRAN을 이용하여 구하였다. 그림 1.1에서는 FORTRAN을 이용하여 얻어진 적합의 예들로서 경계수정과 비경계수정 추정량들에 대하여 보였다. 예상하였던대로 경계수정의 효과는 거의1 로 가장 잘 표현되는데 실제 회귀함수의 도함수는 0과 대체로 다르다. 위의 제시된 알고리즘에 대한 검정과 경계수정의 실질적 효과에 대한 평가를 위하여 모의실험을 하였다. 실험 데이타들은 모형 (1.1)에서 $\sigma = .05$를 가지는 정규 확률오차를 사용하였다. 표본의 크기 $n = 100$, $t_i = (2i-1)/2n$, $i = 1, \ldots, n$,으로 하였고 회귀함수는 다음과 같이 주었다.

$$\mu(t) = \gamma g(t) + 16(t(1-t))^2. \tag{2.16}$$

여기서

$$g(t) = 64\mathrm{I}_{\{t>.7\}}(t-.7)^3 \tag{2.17}$$

이다. μ가 (1.7)에 있는 경계 조건을 만족하지않기 위해 $\mu'(0) = 0$와 $\mu'(1) = 17.28\gamma$로 주었다.

모의 실험은 각각의 테이타들에 대하여, 일반적인 선형 평활스플라인과 경계수정된 선형 평활스플라인들을 일반화 교차 타당성기준에 의해 선택된 그들 각각의 평활 모수를 가지고 계산 하였다. 이것은 0과 1 사이에 있는 γ의 10개의 동일한 간격 값들을 가지는 설계점들에 대하여 100개의 표본들을 100번 반복 시킨 것이다. 모든 표본들에 대하여 그 비 (*ratio*)를 계산 하였다.

$$\sum_{i=1}^{n}(\mu(t_i) - \tilde{\mu}_{\hat{\lambda}}(t_i))^2 / \sum_{i=1}^{n}(\mu(t_i) - \mu_{\hat{\lambda}}(t_i))^2.$$

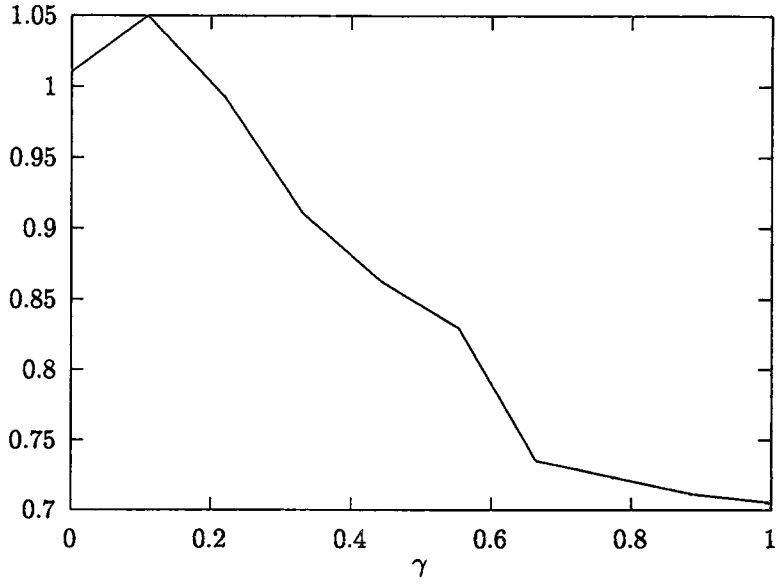

그림 2.1: 평균 위험비.

여기서 $\hat{\lambda}$와 $\tilde{\lambda}$는 각 추정량에 대한 평활 모수들의 일반화 교차 타당성기준에 의해 선택된 것이다. 그 손실 비 (loss ratios)들은 각 γ 에 대한 100번 반복 자료들의 평균이고 이 평균들은 γ의 함수로서 그림 2.1에서 나타난다. 그 결과들은 경계수정이 심지어 상당히 작은 표본들에서도 매우 효과적이다는 사실을 보여 준다.

3. 경계수정 추정량의 점근적 성질

이 절에서는 균등설계 $t_i = (2i-1)/2n$, $i = 1, \ldots, n$의 경우에 있어서 경계수정 추정량의 대 표본에서의 점근성에 대해 연구할 것이다.

$q_{0\lambda}$ 와 $q_{1\lambda}$을 식 (2.5) - (2.6)에 있는 q_0 와 q_1에 대한 선형 평활스플라인 근사값들이라 하자. 보다 정확히 표현하면 $q_{0\lambda}$와 $q_{1\lambda}$는 다음의 판정 기준을 최소화 함으로 구한 함수이다.

$$n^{-1}\sum_{i=1}^{n}(g(t_i) - f(t_i))^2 + \lambda\int_0^1 f'(t)^2 dt. \tag{3.1}$$

여기서 f는 $g(t_i) = q_0(t_i)$, $i = 1, \ldots, n$와 $g(t_i) = q_1(t_i)$, $i = 1, \ldots, n$를 각각 사용한 것이다. 만약

$$\tilde{q}_i(t) = q_i(t) - q_{i\lambda}(t), \quad i = 0, 1, \tag{3.2}$$

정의된다면, 식 (2.7)에 있는 경계수정 선형 평활스플라인은 다음과 같다.

$$\tilde{\mu}_\lambda(t) = \mu_\lambda(t) + b_0\tilde{q}_0(t) + b_1\tilde{q}_1(t). \tag{3.3}$$

여기서

$$\tilde{q}_{ij} = \sum_{k=1}^{n} \tilde{q}_i(t_k)\tilde{q}_j(t_k), \quad i,j = 0,1, \tag{3.4}$$

에 대해서

$$b_0 = (\tilde{q}_{11}\sum_{i=1}^{n} \tilde{q}_0(t_i)y_i - \tilde{q}_{01}\sum_{i=1}^{n} \tilde{q}_1(t_i)y_i)/(\tilde{q}_{00}\tilde{q}_{11} - \tilde{q}_{01}^2) \tag{3.5}$$

이고

$$b_1 = (\tilde{q}_{00}\sum_{i=1}^{n} \tilde{q}_1(t_i)y_i - \tilde{q}_{01}\sum_{i=1}^{n} \tilde{q}_0(t_i)y_i)/(\tilde{q}_{00}\tilde{q}_{11} - \tilde{q}_{01}^2) \tag{3.6}$$

이다.

우리는 추정량의 편의와 위험함수에 대한 대표본에서의 성질들을 구하기 위해 4절에 있는 근사 lemma들을 가지고 식 (3.3) - (3.6)을 사용한다. 다음에 있어서, $\log n/n\lambda^2 \to 0$ 와 같은 방법으로 $n \to \infty$됨에 따라 $\lambda \to 0$ 되는 제한 조건을 부과할 것이다.

이러한 조건하에서 lemma 4.1로 부터 다음을 가진다.

$$\tilde{q}_0(t) = \sqrt{\lambda}e^{-t/\sqrt{\lambda}} + O(\lambda)$$

그리고

$$\tilde{q}_1(t) = \sqrt{\lambda}e^{-(t-1)/\sqrt{\lambda}} + O(\lambda).$$

lemma4.3은 다음을 제공한다.

$$\mathrm{E}\, b_0 = \mu'(0) + O(\sqrt{\lambda}) \tag{3.7}$$

그리고

$$\mathrm{E}\, b_1 = \mu'(1) + O(\sqrt{\lambda}). \tag{3.8}$$

그러므로 $\tilde{\mu}_\lambda$의 편의는 다음과 같다.

$$\begin{aligned} \mu(t) - \mathrm{E}\,\tilde{\mu}_\lambda(t) &= \mu(t) - \mathrm{E}\,\mu_\lambda(t) \\ &\quad -\mu'(0)\sqrt{\lambda}e^{-t/\sqrt{\lambda}} - \mu'(1)\sqrt{\lambda}e^{(t-1)/\sqrt{\lambda}} \\ &\quad +O(\lambda(e^{-t/\sqrt{\lambda}} + e^{(t-1)/\sqrt{\lambda}}) + \lambda^{5/2}). \end{aligned} \tag{3.9}$$

식 (3.9)은 $\tilde{\mu}_\lambda$가 각각의 경계에 관계한 부분들을 제거함으로 본래의 추정량 μ_λ의 편의에 대하여 수정한다는 사실을 명백히 보여준다. 수정은 경계 지역에 대해 편중되어 있음을 주목하자. (3.9)과 조합된 lemma 4.1은 다음과 같이 주어진다.

$$\mu(t) - \mathrm{E}\,\tilde{\mu}_\lambda(t) = \lambda\mu''(t) + O(\lambda^{5/2}).$$

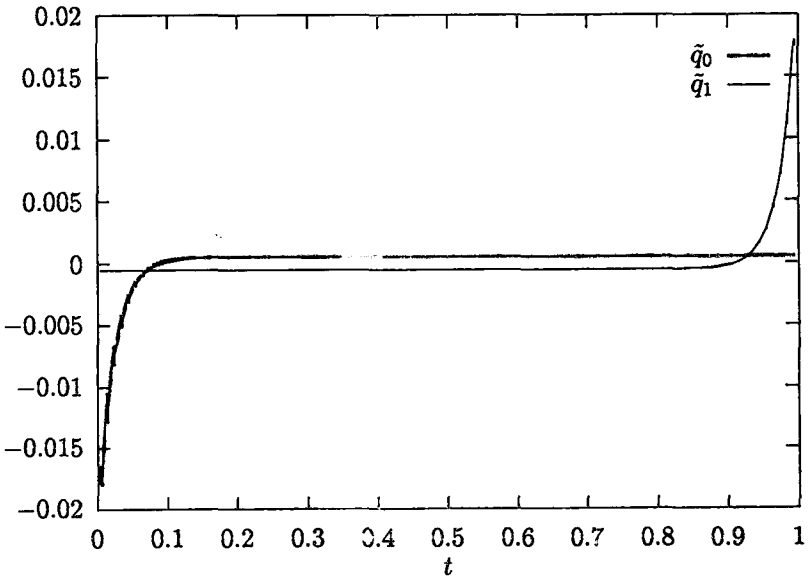

그림 3.1: $\tilde{q}_0$와 $\tilde{q}_1$의 함수.

여기서 $t \in (0,1)$. 그러므로 $\tilde{\mu}_\lambda$는 μ_λ와 같이 $[0,1]$에서 2차 추정량의 편의 성질을 가진다. 그러나 $t = \sqrt{\lambda}v$ 이나 $t = (1-\sqrt{\lambda})v$와 같은 경계점들에 대하여 $\mu(t) - \mathrm{E}\,\mu_\lambda(t)$가 차수 $\sqrt{\lambda}$ 를 가지는 반면에 $\mu(t) - \mathrm{E}\,\tilde{\mu}_\lambda(t) = O(\lambda)$가 되어진다. 그러므로 $\tilde{\mu}_\lambda$는 경계들에서 이차 추정량의 편의에 대하여 μ_λ의 각 점들에 있어서의 편의를 효과적으로 감소시킨다.

그림 3.1은 그림 1.1에서의 경계수정 추정량에 관계된 $\tilde{q}_0$와 $\tilde{q}_1$의 함수들의 그림이다. 이 경우에 있어서 λ의 GCV 선택은 (3.248)10^{-4}이고 따라서 우리들은 이러한 함수들이 주로 경계 지역들 $[\,0, \sqrt{\lambda}\,]$와 $[\,1-\sqrt{\lambda}, 1\,]$에 있어서 지원되었음을 볼 수 있다. (2.9)에 있는 계수 벡터의 요인들은 $b_0 = 1.1867$와 $b_1 = 7.627$ 이고 이것은 식 (3.7) - (3.8)의 결과로서 $\mu'(0) = 0$ 와 $\mu'(1) = 12.96$의 추정량으로서 인정되어 진다.

$\tilde{\mu}_\lambda$의 전체적인 특성들은 그것의 위험 (risk) 함수의 대표본에서의 특성을 공부함으로서 조사할 수 있다. 이러한 관점에서 우리는 다음의 결과들을 가지는데 이 정리의 증명은 다음절에서 소개한다.

정리 3.1 $n \to \infty$에 따라서 $\log n/n\lambda^2 \to 0$ 를 가지고 $\lambda \to 0$가 되고 그리고 $\mu \in C^3[0,1]$를 가정하자. 식 (2.5) - (2.6)에 있는 q_0와 q_1에 대하여 $\mu_0 = \mu - \mu'(0)q_0 - \mu'(1)q_1$로 정의 하자. 그러면

$$\begin{aligned}\tilde{\mathrm{R}}_n(\lambda) &= n^{-1}\sum_{i=1}^{n}\mathrm{E}\,(\tilde{\mu}_\lambda(t_i) - \mu(t_i))^2 \\ &= \lambda^2\int_0^1 \mu_0''(t)^2 dt + \frac{\sigma^2}{4n\sqrt{\lambda}}(1+o(1)) + O(\lambda^{5/2} + n^{-1}).\end{aligned}$$

이 정리로 부터의 $\tilde{\mu}_\lambda$에 대한 평활 모수의 점근적인 최적의 선택은 다음의 식에 의해 제공된다.

$$\lambda_n^* = \left(\sigma^2/4n \int_0^1 \mu_0''(t)^2 dt\right)^{2/5}.$$

그러므로 다음의 위험 함수를 만들 수 있다.

$$\tilde{\mathrm{R}}_n(\lambda_n^*) = 1.25\left(\frac{\sigma^2}{4n}\right)^{4/5}\left(\int_0^1 \mu_0''(t)^2 dt\right)^{1/5} + o(n^{-4/5}). \tag{3.10}$$

그러므로 $\tilde{\mu}_\lambda$은 그것의 위험함수의 점근적 성질에 있어서 경계수정된 이차 커널 추정량과 비슷한 특성을 지닌다. 그러나 여기에는 식 (3.10)이 커널 추정량으로 부터 기대 되는 $\int_0^1 \mu''(t)^2 dt$ 보다 오히려 $\int_0^1 \mu_0''(t)^2 dt$을 포함한다는 관점에서 중요한 차이점이 있다. 다음을 주목하자.

$$\int_0^1 \mu_0''(t)^2 dt = \int_0^1 \mu''(t)^2 dt - \left(\int_0^1 \mu''(t)^2 dt\right)^2 \leq \int_0^1 \mu''(t)^2 dt.$$

이때 $\mu'(0) \neq \mu'(1)$. 그러므로 μ_λ에 대하여 만들어진 경계수정은 실제로 위험함수 추정에 있어서 전체적인 강한 영향력을 가진다.

우리는 선형 평활스플라인과 다른 이차 추정량들의 대표본 위험함수 사이의 효과적인 비교를 위하여 식 (3.10)을 사용할 수 있다. 예를 들어 Epanechnikov 커널을 사용한 이차 경계수정 커널 추정량을 위한 점근적 최적 위험함수는 다음과 같다.

$$R^* = 1.25(.349)\left(\frac{\sigma^2}{n}\right)^{4/5}\left(\int_0^1 \mu''(t)^2 dt\right)^{1/5} + o(n^{-4/5}).$$

그러므로 커널과 경계수정 평활스플라인 추정량의 타당성 있는 점근적 위험 효율은 다음과 같다.

$$\mathrm{R}_n(\lambda_n^*)/R^* \sim .945\left(\int_0^1 \mu_0''(t)^2 dt / \int_0^1 \mu''(t)^2 dt\right)^{1/5}.$$

결론적으로 경계수정 선형 평활스플라인은 심지어 경계수정을 전혀 필요로 하지 않는 최적 커널 추정량 보다 대략 5.5% 이상의 점근적인 효율이 있다. 이것은 Epanechnikov 커널의 최적성을 부인하는 것이 아니다 그 이유는 선형 평활스플라인의 점근적 특성에서 Epanechnikov 커널 추정량으로서 표현 되어 질수 있기 때문이다. 그리고 그것의 관계되어진 동일한 추정량은 무한한 지지를 가진 Laplace 확률밀도 함수이기 때문이다.

4. 증명들

이 절에서는 정리 3.1을 증명할 것이다. 우리들은 증명에 필요한 여러 lemma들을 가지고 시작한다. 모든 결과들은 균등 설계 $t_i = (2i-1)/2n$, $i = 1, \ldots, n$의 경우에 국한 시켰다.

함수 g에 대하여 g_λ는 다음을 최소화 함으로 구해진 선형 평활스플라인 근사라 하자.

$$n^{-1}\sum_{i=1}^{n}(g(t_i)-f(t_i))^2+\lambda\int_0^1 f'(t)^2dt.$$

여기서 f는 제곱 적분가능 도함수를 가지는 모든 절대적으로 연속인 함수이다. 편리함을 위해 다음의 표현을 사용한다.

$$\tilde{g}=g-g_\lambda.$$

이것은 g_λ에 의해 근사적인 g에서의 오차를 의미한다. 함수 $\tilde{g}$는 또한 회귀함수 g를 가지고 데이타에 선형 평활스플라인 적합으로 부터의 편의들이다.

우리들의 세개의 lemma들은 선형 평활스플라인 편의와 편의 내적들 (inner products)에 대한 점근적인 근사들을 제시한다.

LEMMA 4.1 만약 $\gamma\in(0,1)$에 대하여 $g\in C^3[0,1]$와 $\lambda\asymp n^{-\gamma}$ 이라면,

$$\tilde{g}(t)=\sqrt{\lambda}g'(t)h_{1\lambda}(t)-\lambda g''(t)-\lambda g''(t)h_{2\lambda}(t)+O(\frac{\log n}{n\lambda}+\lambda^{5/2}).$$

여기서 $h_{1\lambda}(t)=e^{(t-1)/\sqrt{\lambda}}-e^{-t/\sqrt{\lambda}}$ 그리고 $h_{2\lambda}(t)=((t-1)/\sqrt{\lambda})e^{(t-1)\sqrt{\lambda}}-(t/\sqrt{\lambda})e^{-t/\sqrt{\lambda}}$에대하여균등하게 $t\in[0,1]$이다.

증명: Eubank (1997)을 보라. □

LEMMA 4.2 $\gamma\in(0,1)$ 와 $g\in C^2[0,1]$에 대하여, 만약 $\lambda\asymp n^{-\gamma}$이라면,

$$n^{-1}\sum_{i=1}^{n}g(t_i)h_{1\lambda}^k(t_i)=\begin{cases}\frac{\sqrt{\lambda}}{k}\left(g(1)+(-1)^jg(0)\right)+O(\lambda+\frac{1}{n\sqrt{\lambda}})\\ O(\lambda^{3/2}+\frac{1}{n\sqrt{\lambda}}),\ \text{if } g'(0)=g'(1)=0.\end{cases}$$

또한,

$$n^{-1}\sum_{i=1}^{n}g(t_i)h_{2\lambda}(t_i)h_{r\lambda}(t_i)=O(\lambda^{j/2}+\frac{1}{n\sqrt{\lambda}}).$$

여기서 $r=1,2$에 대하여 $g(0)=g(1)=0$ 일때 $j=2$가 되고 그외에는 $j=1$가 된다.

증명: 표준 구적법 개요는 $O(1/n\sqrt{\lambda})$ 나머지를 산출하고 적분에 의해 근사되어 질수 있는 합들을 이용한다. 증명의 나머지 부분은 적분과 평균값 정리에 의해 구해진다. □

LEMMA 4.3 $n \to \infty$가 됨에 따라, 만약 $\log n/n\lambda^2 \to 0$와 $f, g \in C^3[0,1]$를 가지고 $\lambda \to 0$이라면,

$$n^{-1}\sum_{i=1}^{n}\tilde{g}(t_i)\tilde{f}(t_i) = \frac{\lambda^{3/2}}{2}\left(g'(1)f'(1) + g'(0)f'(0)\right) + O(\lambda^2).$$

만약 $g'(0)f'(0) + g'(1)f'(1) = 0$이라면,

$$n^{-1}\sum_{i=1}^{n}\tilde{g}(t_i)\tilde{f}(t_i) = \lambda^2\int_0^1 g''(t)f''(t)dt + O(\lambda^{5/2}).$$

증명: 이 lemma의 증명은 매우 지루하다. 우리는 lemma 4.2를 사용하여 각각 계산된 16개 항들을 포함하는 $n^{-1}\sum_{i=1}^{n}\tilde{g}(t_i)\tilde{f}(t_i)$를 구하기 위해 lemma 4.1을 사용한다. 예를들면, 전형적인 항은 $\lambda n^{-1}\sum_{i=1}^{n} g'(t_i)f'(t_i)h_{1\lambda}^2(t_i)$이다. 만약 $g'(0)f'(0) + g'(1)f'(1) \neq 0$인 경우, 전형적인 항의 값은 $2^{-1}\lambda^{3/2}\left(g'(0)f'(0) + g'(1)f'(1)\right) + O(\lambda^2)$으로 표현된다. 그 외의 경우에 있어서, 그것의 값은 $O(\lambda^{5/2})$가 된다. 이와 유사한 방법이 다른 15항들에 대하여도 적용된다. □

마지막 lemma는 선형 평활스플라인으로 부터 해트행렬 혹은 평활행렬 $\mathbf{S}_\lambda$의 고유화(trace)의 근사에 관계된다.

LEMMA 4.4 $n \to \infty$이고, 만약 $\log n/n\lambda^2 \to 0$을 가지고 $\lambda \to 0$이 된다면, $n^{-1}\mathrm{tr}\mathbf{S}_\lambda = (4n\sqrt{\lambda})^{-1}(1 + o(1))$이다.

증명: Eubank (1997)로 부터 우리는 $\mathbf{S}_\lambda^2$의 i번째 대각 요소들은 $h_{3\lambda}(t) = e^{2(t-1)/\sqrt{\lambda}}(1 - \frac{2(t-1)}{\sqrt{\lambda}})$와 $h_{4\lambda}(t) = e^{-2t/\sqrt{\lambda}}(1 - \frac{2t}{\sqrt{\lambda}})$대하여 다음과 같다.

$$(4n\sqrt{\lambda})^{-1}\{1 + h_{3\lambda}(t_i) + h_{4\lambda}(t_i) + o(1)\}.$$

증명은 lemma 4.2를 이용하여 얻어진다.

□

위의 lemma들의 결과들을 가지고 우리는 정리 3.1을 증명할 수 있다. 2절에 정의된대로 $\mathbf{Q}$를 규정하고 $\boldsymbol{\mu}_0 = (\mu_0(t_1), \ldots, \mu_0(t_n))^T$이라 두자. 그러면

$$\boldsymbol{\mu} - \mathrm{E}\,\tilde{\boldsymbol{\mu}}_\lambda = \tilde{\boldsymbol{\mu}}_0 - \tilde{\mathbf{Q}}(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T\tilde{\boldsymbol{\mu}}_0.$$

여기서 $\tilde{\boldsymbol{\mu}}_0 = (\mathbf{I} - \mathbf{S}_\lambda)\boldsymbol{\mu}_0$를 가진다. Lemma 4.3을 사용하여 다음을 구한다.

$$n^{-1}\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}} = \frac{\lambda^{3/2}}{2}\begin{bmatrix} 1 + O(\sqrt{\lambda}) & O(\sqrt{\lambda}) \\ O(\sqrt{\lambda}) & 1 + O(\sqrt{\lambda}) \end{bmatrix}$$

그리고

$$n^{-1}\tilde{\boldsymbol{\mu}}_0^T\tilde{\mathbf{Q}} = (O(\lambda^2), O(\lambda^2)).$$

이 이유는 $\mu_0'(0) = \mu_0'(1) = q_0'(1) = q_1'(0) = 0$ 와 $q_0'(0) = q_1'(1) = 1$이기 때문이다. 그러므로,

$$\begin{aligned} n^{-1}(\mu - \mathrm{E}\tilde{\mu}_\lambda)^T(\mu - \mathrm{E}\tilde{\mu}_\lambda) &= n^{-1}\tilde{\mu}_0^T\tilde{\mu}_0 - n^{-1}\tilde{\mu}_0^T\tilde{\mathbf{Q}}(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T\tilde{\mu}_0 \\ &= \lambda^2\int_0^1 \mu_0''(t)^2dt + O(\lambda^{5/2} + n^{-1}). \end{aligned}$$

$\tilde{\mathrm{R}}_n(\lambda)$의 분산 에 대하여 우리는 다음과 같이 쓸 수 있다.

$$\mathrm{tr}\mathbf{H}_\lambda = \mathrm{tr}\mathbf{S}_\lambda^2 + 2 - \mathrm{tr}\mathbf{S}_\lambda{}^2\tilde{\mathbf{Q}}(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T.$$

$\mathrm{tr}\mathbf{S}_\lambda{}^2\tilde{\mathbf{Q}}(\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T$은 2보다 크지 않다. 그 이유는 $\mathbf{S}_\lambda$의 고유값이 모두 1에 의해 제한되기 때문이다. lemma 4.4를 사용함으로서 증명은 마친다.

참고문헌

[1] Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker: New York.

[2] Eubank, R. L. (1997). A simple smoothing spline, II. Manuscript.

[3] Eubank, R. L. and Speckman, P. L. (1991). A bias reduction theorem with applications in nonparameteric regression. *Scandinavian Journal of Statistics. 18,* 211 - 222.

[4] Hutchinson, M. F. and de Hoog, F. R. (1985). Smoothing noisy data with spline functions. *Numerical Mathematics. 47,* 99-106.

[5] Lyche, T. and Schumaker, L. L. (1973). Computation of smoothing and interpolating natural splines via local bases. *SIAM J. Numerical Analysis. 10,* 1027-1038.

[6] Nychka, D. (1995) Splines as local smoothers. *Annals of Statistics. 23,* 1175–1197.

[7] Oehlert, G. W. (1992) Relaxed boundary smoothing splines. *Annals of Statistics. 20,* 146–160.

[1998년 1월 접수, 1998년 5월 최종수정]

A Second Order Smoother *

Jongtae Kim [1)]

ABSTRACT

The linear smoothing spline estimator is modified to remove boundary bias effects. The resulting estimator can be calculated efficiently using an $O(n)$ algorithm that is developed for the computation of fitted values and associated smoothing parameter selection criteria. The asymptotic properties of the estimator are studied for the case of a uniform design. In this case the mean squared error properties of boundary corrected linear smoothing splines are seen to be asymptotically competitive with those for standard second order kernel smoothers.

* This paper was supported by research fund, Taegu University, 1998

1) Dept. of Statistics, Taegu University, Kyunsan, Kyungpook 712-714, Korea.