Journal of Statistical
Theory & Methods
1998, Vol. 9, No. 1, pp. 29 ~ 36

선형 평활스플라인 함수 추정과 적용 [1]

윤용화 · 김경무 · 김종태 [2]

요약

본 논문은 Eubank (1994, 1997)에 의해 이론적으로 제안된 선형 평활스플라인 추정량에 대한 알고리즘을 개발함으로 선형 스플라인의 추정을 보다 쉽고 효율적으로 사용할 수 있도록 하는데 목적이 있다. 이 알고리즘을 이용하여 여러가지 모형의 예들에 대하여 추정량의 적합성을 조사하였고, 제시된 선형 평활스플라인 추정량이 비모수 함수 추정의 도구로서 잘 적합됨을 알 수 있었다.

주제어: 선형 평활스플라인, 커널 함수, 평활모수, 퓨리에 급수, 비모수 회귀분석, 일반교차타당성기준.

1. 서론

지난 30여년간 퓨리에 급수, 커널 함수 기법, 평활스플라인들과 같은 비모수 함수추정 기법들은 회귀진단이나 통계 추론에 있어서 중요한 도구로 사용되어 왔다. 최근에는 자료분석을 함에 있어서 비모수적 함수 추정 기법들이 고전적 모수 추론 방법들 보다 때로는 좋은 결과를 얻는다는 것이 많은 연구 문헌들을 통해서 증거되고 있다.

평활스플라인(smoothing spline) 기법은 비모수 회귀 함수 추정에 있어서 커널(kernel) 함수 기법과 함께 매우 중요한 도구로서 사용되고 있다. 그러나 유감스럽게도 평활스플라인 기법은 커널과 같은 다른 비모수적 함수추정 기법들과는 다르게 일반적으로 평활스플라인을 쉽게 설명할 수 있는 구체화된 형태들을 갖고 있지 않다. 이러한 점을 보충하기 위해 Eubank (1994, 1997)는 비모수적 추정량들에 대한 관점을 보다 넓히기 위한 목적으로 선형 평활스플라인들에 대한 구체화된 형태를 이론적으로 소개하였다. 본 논문은 Eubank (1997)가 제안한 이론적인 선형 평활스플라인(linear smoothing splines)에 대하여 제안된 추정량에 대한 프로그램 알고리즘을 개발함으로서 보다 쉽게 선형 평활스플라인 추정량을 적합시키는데 목적이 있다.

[1]이 논문은 1997년도 대구대학교 학술연구비 지원에 의한 연구임
[2](712-714) 경북 경산시 진량면 대구대학교 통계학과

다음의 비모수 회귀 모형에 대한 문제를 생각해 보자. 반응변수 $y_1, \ldots, y_n$이 아래 모형 (1)로 부터 동시에 발생하지 않는 설계점들 (non-coinsident design points), $0 \le t_1 < \cdots < t_n \le 1$, 에서 구해진다고 하자.

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

여기서 μ는 미지의 회귀함수이고, $\epsilon_1, \ldots, \epsilon_n$은 평균 0과 공통 분산 σ^2를 가지는 비상관 랜덤오차들 (uncorrelated random errors) 이다. 이 모형에서 비모수 함수 추정은 μ의 구체적인 형태에 대하여 가능한 최소한의 가정을 가지고 μ를 추정하는데 있다.

평활스플라인 추정량 $\hat{\mu}_\lambda$는 평활 모수 λ를 가지는 μ에 대하여 연속 미분 가능한 모든 함수들 f에 대해 아래의 식 (2)를 최소로 하는 것이다.

$$n^{-1}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda \int_0^1 f^m(t)^2 dt, \quad \lambda > 0. \tag{2}$$

평활스플라인은 자료의 적합에 있어서 보다 유연성을 제공하는 선형다항회귀(linear polynomial regression) 추정량의 일반적인 표현이다. 예를 들어 $m = 2$이고, 식 (2)에서 $\int_0^1 (f^2(t))^2 dt = 0$이면 모든 함수들에 대해 최소가 되는 추정량이 존재하고 이 추정량은 일반 단순 선형 회귀모형이다. 그러나 $\int_0^1 f^2(t)^2 dt > 0$일 때, 추정량은 기울기가 상수가 아닌 값을 갖게 되고, 그때 λ는 선형(linearity)으로부터 얼마나 떨어지는 정도를 조절한다.

식 (2)로부터 얻은 추정량은 가중함수 $w_n(\cdot,\cdot)$에 대하여 $\hat{\mu}_\lambda(t) = \sum_{i=1}^n w_n(t, t_i) y_i$라는 관점에서 선형이다. 그러나 w_n에 대하여 일반적이며 구체화된 형태가 없다는 것이 다른 비모수 함수기법과의 비교에서 평활스플라인은 그자체를 이해시키는데 단점을 가진다. 예를 들어 비모수 회귀 문제에 대한 μ의 전형적인 커널 추정량은 다음과 같은 구체화된 형태를 가진다.

$$(nb)^{-1}\sum_{i=1}^{n} K(\frac{t - t_i}{b}) y_i \tag{3}$$

여기서 K는 커널함수이고 b는 띠폭(bandwidth)이다. 또 다른 비모수 함수 추정기법인 퓨리에(코사인) 급수 (Fourier (cosine) series)는 자료의 회귀에 의하여 구해지는 $p+1$항과 $1, \cos(\pi t), \ldots, \cos(p\pi t)$를 가지고 다음의 형태로 표현된다.

$$\bar{y} + \sum_{r=1}^{p} a_r \cos(r\pi t). \tag{4}$$

이때 $\bar{y} = n^{-1}\sum_{j=1}^n y_j$이고,

$$a_r = 2/n \sum_{j=1}^{n} y_j \cos(r\pi t_j), \qquad r = 1, \ldots, n-1. \tag{5}$$

이다. (비모수적 추정 기법에 대한 다양한 정보에 대해서는 Messer (1991), Silverman (1984), Eubank, R. L. (1988)를 참조.)

2. 선형 평활스플라인

본 절에서는 Eubank (1994)에서 제시한 선형 평활스플라인 추정량, 즉, $m = 1$일 때 모형 (1)로부터 자료를 적합시키는 평활스플라인을 소개한다. 제시된 구체적 형태의 선형 평활스플라인은 다소 한정된 실제값을 가지지만 일반적인 평활스플라인과 다른 추정량들의 특성을 보다 직관적으로 이해할 수 있게 된다.

식 (2)에서 $m = 1$인 경우를 가정하자. 그러면

$$n^{-1}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda\int_0^1 f'(t)^2dt, \quad \lambda > 0, \tag{6}$$

을 최소화하는 $\hat{\mu}_\lambda$에 대하여 구체화된 표현을 얻을 수 있다. 추정량 $\hat{\mu}_\lambda$를 선형 평활스플라인 (linear smoothing splines)이라 한다. 이 선형 평활스플라인은 λ가 클 때에는 자료에 대해 거의 상수에 가까운 적합이 이루어 지고 λ의 값이 작아 질수록 반응값들에 더 적합되는 성질을 지닌다. 이 추정량은 스플라인으로 불리어지는 활꼴다항형태(type of segmented polynomial)가 된다. Eubank (1997)는 선형 평활스플라인 추정량에 대한 다음과 같은 구체적인 형태를 제시한다.

정리 2.1. (Eubank (1994)) 제곱적분가능 일차 도함수를 가지는 모든 절대 연속 함수들에서 식 (6)를 최소로하는 유일한 추정량은 다음과 같다.

$$\hat{\mu}_\lambda(t) = \bar{y} + \sum_{j=1}^{n-1} a_j x_j(t)/(1 + \lambda\gamma_j). \tag{7}$$

여기서 $\bar{y}$는 반응값의 평균이고,

$$x_j(t) = \cos(j\pi t), \tag{8}$$

$$a_j = \frac{2}{n}\sum_{r=1}^{n} y_r \cos(j\pi t_r), \quad j = 1, \ldots, n-1, \tag{9}$$

$$\gamma_j = (2n\sin(j\pi/2n))^2, \quad 1 \le j \le n-1, \tag{10}$$

이다.

정리 2.1를 이용한 선형 평활스플라인 추정량을 정리하면 다음과 같다.

$$\hat{\mu}_\lambda(t) = \bar{y} + \sum_{j=1}^{n-1} a_j \cos(j\pi t)/(1 + \lambda(2n\sin(j\pi/2n))^2). \tag{11}$$

여기서

$$a_j = \frac{2}{n}\sum_{r=1}^{n} y_r \cos(j\pi t_r), \quad j = 1, \ldots, n-1. \tag{12}$$

여기서 $t_r = (2r-1)/(2n)$, $r = 1, \ldots, n$. 다음의 4절에서 소개되는 각 예들에 대한 추정량의 적합은 위의 식 (11)을 가지고 조사되었다. 식 (6)의 평활스플라인을 보다 잘 이해하기 위해서 평활 모수 λ의 값을 가지는 식 (11)의 선형평활스플라인 추정량을 사용할 수 있음을 그림 2를 통하여 알 수 있다.

μ에 대한 선형 평활스플라인 추정량 $\hat{\mu}_\lambda$가 위의 식 (11)로 부터 계산되고 난 후에 자료에 대한 올바른 선형 평활스플라인의 적합을 이루기 위해서 추정량에 대한 평활 모수 λ의 최적값을 구하여야 한다. 다음은 제안된 추정량을 계산하는 방법을 소개할 것이다.

최적화된 평활 모수를 구하기 위한 판정기준으로 일반교차타당성기준을 사용한다 (Wahba, 1985). 일반교차타당성기준을 계산하기 위하여 먼저 다음의 잔차의 제곱합 (residual sum of square, RSS)을 계산한다.

$$RSS(\lambda) = \sum_{i=1}^{n} (y_i - \hat{\mu}_\lambda(t_i))^2. \tag{13}$$

위의 잔차의 제곱합을 이용하여 λ의 최적값을 구하기 위해 다음의 일반교차타당성 (generalized cross validation, GCV) 기준을 최소화 시키는 λ의 값을 사용하는 것이다.

$$GCV = \frac{nRSS(\lambda)}{tr(I - H_\lambda)^2} \tag{14}$$

여기서 I는 단위행렬 (unit matrix)이고 H_λ는 $\hat{\mu}_\lambda$에 대한 해트행렬 (hat matrix) 혹은 평활행렬이다. (이 평활행렬에 대해서는 Eubank (1988, 30쪽)을 참고.) 그리고 위의 식 (14)에서 $tr(I - H_\lambda)^2 = n - \lambda(2n\sin(j\pi/2n))^2$로서 계산되어 진다. 그러므로 우리는 다음의 일반교차타당성기준을 최소로 하는 최적 λ 값을 찾을 수 있다.

$$GCV = \frac{nRSS(\lambda)}{n - \lambda(2n\sin(j\pi/2n))^2} \tag{15}$$

우리는 위에서 제시한 선형 평활스플라인의 최적화 추정량을 구하는 알고리즘을 계산하기 위해서 FORTRAN 프로그램 언어를 사용하였다. 이 선형 평활스플라인의 FORTRAN 알고리즘은 기존의 자연 삼차스플라인의 알고리즘 보다 훨씬 쉽고 간편하게 사용할 수 있다. 또한 자료에 대한 적합에 있어도 삼차 스플라인에 뒤떨어지지 않는 매우 훌륭한 적합이 이루어짐을 다음절에서의 여러가지 모형들에 대한 추정량의 적합에서 볼 수 있다.

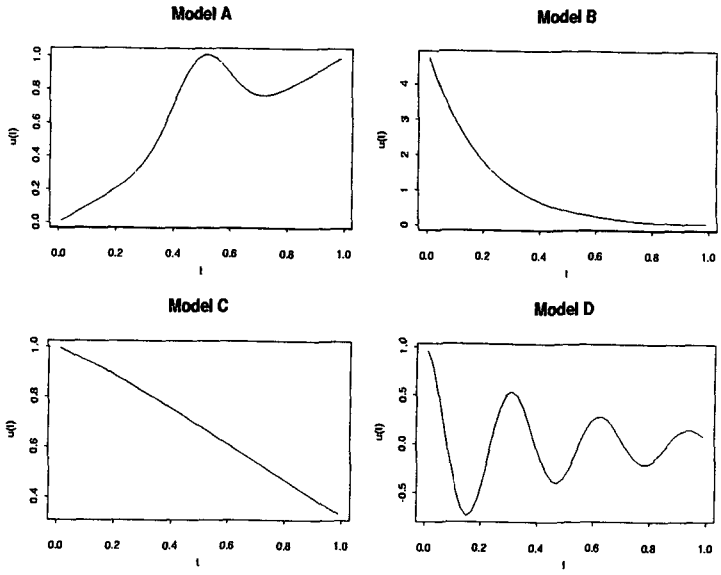
다음 절에서는 우리들의 선형 평활스플라인의 FORTRAN 알고리즘을 가지고 실험을 이용하여 일반교차타당성기준에 의하여 구한 λ를 가지고 μ의 여러가지 함수에 대하여 제시한 선형 평활 추정량 $\hat{\mu}_\lambda$가 얼마나 잘 적합되는지를 조사할 것이다.

추정량의 적합성 조사

이 절에서는 위의 식 (11)에서 $\hat{\mu}_\lambda$가 자료에 대해 어떻게 작용하는가 혹은 평활 방법을 이해하는데 어떻게 사용되는가를 조사한다. 이것은 진동수(frequency)나 시간 영역(time

domain)에 추정량을 적용함으로서 얻어진다. 보다 정확하게 알기 위하여, 자료의 진동수에 추정량이 어떻게 작용하는지를 조사하거나 혹은 추정량이 어떻게 반응값들을 스스로 평가하는지를 조사한다. 이러한 조사방법은 추정량의 특성에 대하여 둘 모두 유용한 통찰력을 제공할 것이다. $\hat{\mu}_\lambda$의 특성을 조사하기 위해 $t_r = (2r-1)/2n,\ r = 1, \ldots, n,$ 에 대하여 다음의 네가지 모형을 사용하였다.

그림1 : 각 예들에 대한 실험함수 μ의 분포



모형 A:

$$\mu(t) = t + a_1 \exp(-a_2(t-a_3)^2), \quad (a_1 = 0.5, a_2 = 50, a_3 = 0.5). \tag{16}$$

모형 B:

$$\mu(t) = a_1 \exp(-a_2 t), \quad (a_1 = 5, a_2 = 5). \tag{17}$$

모형 C:

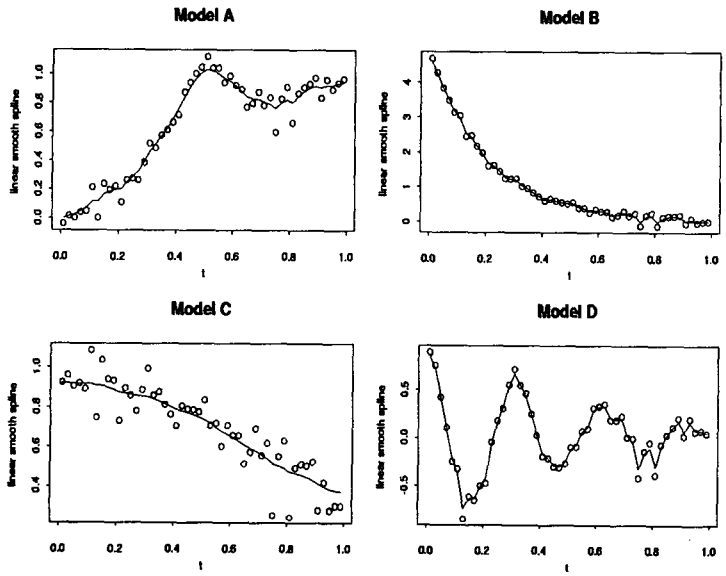
$$\mu(t) = \exp(-t/a_1)\cos(t/a_2), \quad (a_1 = 2.0, a_2 = 1.0). \tag{18}$$

모형 D:

$$\mu(t) = \exp(-t/a_1)\cos(t/a_2), \quad (a_1 = 0.5, a_2 = 0.05). \tag{19}$$

그림 1은 모형 A - 모형 D까지의 네개의 함수 μ에 대한 진동수나 시간영역에서 분포를 보인 것이다. 식 (1)의 오차 ϵ의 분포는 각 모형들에 대하여 평균이 0인 정규 분포를 주었고 분산의 값은 모형 A에 대해서 0.005 모형 B, C, D에 대해서 0.01로 가정 하였다.

그림 2 : 예에 의한 추정량의 적합성



추정량의 적합성에 대한 결과는 그림 2와 같다. 모의 실험의 과정에서 모형 A에서 식 (15)의 일반교차타당성기준을 최소로하는 평활모수 λ의 값은 0.05이고, 모형 B에서는 평활모수의 값이 0.007087, 모형 C에서는 0.35274, 모형 D에서는 0.0040556 값을 각각 얻었다. 그림 2에서 보듯이 정리 2.1의 선형 평활 스플라인 추정량은 진동수나 시간의 영역에 의존하는 자료들에 대하여 대체로 잘 적합되어짐을 알 수 있다. 일반적으로 평활 스플라인 기법을 사용할 때 2차 평활스플라인 (cubic smoothing spline)을 많이 사용하나 이 결과에서 보듯이 선형 평활 스플라인도 좋은 결과를 얻을 수 있음을 알 수 있다. 그러나 그림 2의 모형 D에서 선형 평활스플라인의 추정량이 지엽적인 부분들에 대해 다소 거칠음을 알 수 있다. 이것은 일반교차타당성을 최소로 하는 평활모수의 값이 다소 적은 값으로 구해졌음을 의미한다.

식 (6)의 평활스플라인을 보다 잘 이해하기 위해서 평활 모수 λ의 값을 가지는 식 (11)의 선형평활스플라인 추정량을 사용할 수 있음을 그림 2를 통하여 알 수 있다.

참고문헌

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

2. Eubank, R. L. (1994), *A Simple Smoothing Spline I*, American Statistician, Vol. 48, 103-106

3. Eubank, R. L. (1997), *A Simple Smoothing Spline II*, Manuscript.

4. Messer, K. (1991), " A Comparison of a Spline Estimate to Its 'Equivalent' Kernel Estimate," *The Annals of Statistics*, 19, 898-916

5. Silverman, B. (1984) "Spline Smoothing: The Equivalent Variable Kernel Method," *The Annals of Statistics*, 12, 898-916

6. Wahba, G. (1985), "A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem," *The Annals of Statistics*, 13, 1378-1402

A Linear Smoothing Spline Estimation and Applications [3]

Yonghwa Yoon · Kyungmu Kim · Jongtae Kim [4]

Abstract

A Simple, closed form expression is derived for a linear smoothing spline by Eubank (1994, 1997). We introduced his estimator and studied how well examples are fitted to this estimator with the values of minimum generalized cross validation.

Key Words and Phrases : Fourier series, Kernel estimators, Smoothing splines, Nonparametric regression, generalized cross validation

[3]This paper was supported by research fund, Taegu University, 1997

[4]Dept. of Statistics, Taegu University, Kyungbuk 712-714