The Korean Journal of Applied Statistics (2017) DOI: http://dx.doi.org/10.5351/KJAS.2017.30.5.733
30(5), 733–745

Selection of bandwidth for local linear composite quantile regression smoothing

Myoungshic Jhun[a] · Jongkyeong Kang[a] · Sungwan Bang[b,1]

[a]Department of Statistics, Korea University;
[b]Department of Mathematics, Korea Military Academy

(Received July 27, 2017; Revised September 6, 2017; Accepted September 11, 2017)

Abstract

Local composite quantile regression is a useful non-parametric regression method widely used for its high efficiency. Data smoothing methods using kernel are typically used in the estimation process with performances that rely largely on the smoothing parameter rather than the kernel. However, L_2-norm is generally used as criterion to estimate the performance of the regression function. In addition, many studies have been conducted on the selection of smoothing parameters that minimize mean square error (MSE) or mean integrated square error (MISE). In this paper, we explored the optimality of selecting smoothing parameters that determine the performance of non-parametric regression models using local linear composite quantile regression. As evaluation criteria for the choice of smoothing parameter, we used mean absolute error (MAE) and mean integrated absolute error (MIAE), which have not been researched extensively due to mathematical difficulties. We proved the uniqueness of the optimal smoothing parameter based on MAE and MIAE. Furthermore, we compared the optimal smoothing parameter based on the proposed criteria (MAE and MIAE) with existing criteria (MSE and MISE). In this process, the properties of the proposed method were investigated through simulation studies in various situations.

Keywords: composite quantile regression, bandwidth selection, local linear regression, kernel function, mean integrated absolute error

1. 서론

일반적인 선형회귀분석에서는 회귀모형의 형태 및 오차항의 분포에 대한 강한 가정 하에서 회귀모형을 추정하게 된다. 그러나 자료의 형태가 가정을 만족하지 않거나 자료의 형태에 대하여 충분한 정보가 없는 상황에서의 일반적인 회귀분석은 그 효용성이 크게 퇴색된다. 이러한 경우 모형에 대해 약한 가정만이 요구되는 비모수적 회귀방법이 보다 유용할 수 있다.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRT) funded by the Ministry of Education (2016R1D1A1B0393114) for M. Jhun and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2015R1C1A1A02036473) for S. Bang.
[1]Corresponding author: Department of Mathematics, Korea Military Academy, 574 Hwarang Rd, Nowon Gu, Seoul 01805, Korea. E-mail: wan1365@gmail.com

반응변수 Y와 공변량 X에 대한 다음과 같은 비모수 회귀모형을 고려하자.

$$Y = m(X) + \sigma(X)\varepsilon, \tag{1.1}$$

여기서 $\sigma(X)$는 표준편차를 나타내는 양함수이고, $m(X) = E(Y|X)$는 매끄러운(smooth) 비모수 함수라고 가정하자. 편의상 오차 ε은 평균이 0이고 분산이 1이며 대칭인 분포를 따른다고 가정하자. 이러한 비모수 회귀모형의 추정에 있어서 가중 최소제곱법(weighted least squares)을 통해 국소적으로 다항 회귀모형에 적합시키는 국소 다항 회귀(local polynomial regression)가 널리 사용되고 있으며, 이에 대한 자세한 내용은 Fan과 Gijbels (1996)에 정리되어 있다. 그러나 이상치(outlier)가 존재하거나 라플라스 분포 등과 같이 오차항의 꼬리분포가 두터운 경우에는 최소제곱법에 기반한 국소 다항회귀는 그 실용성이 떨어지며, 국소 최소절대편차(least absolute deviation) 다항 회귀 (Fan 등, 1994; Welsh, 1996)가 이에 대한 좋은 대안이 될 수 있다. 하지만 오차항의 분포가 정규분포와 같은 특정한 경우에 있어서 국소 최소절대편차 다항회귀의 성능은 국소 최소제곱 다항회귀에 비해 크게 떨어지는 한계가 있다. 이에 Kai 등 (2010)은 Zou와 Yuan (2008)이 제안한 복합 분위수 회귀(composite quantile regression; CQR)방법을 이용하여 다양한 오차 분포에서 효과적으로 모형 적합을 가능하게 하는 국소 복합 분위수 회귀(local CQR)를 제안하였다. 국소 복합 분위수 회귀는 오차항의 분포가 정규분포를 따르는 경우에는 일반적인 국소 다항 회귀와 거의 유사한 효율성을 보이며, 오차항의 분포가 정규분포를 따르지 않는 경우에는 일반적인 국소 다항 회귀에 비해 훨씬 더 높은 효율성을 보임으로써 최근 그 활용성에 대한 연구가 활발히 진행되고 있다.

이런 비모수 회귀모형의 추정과정에 있어 커널함수(kernel function)를 사용한 자료 평활방법(smoothing method)이 대표적으로 사용되고 있으며, 그 성능은 커널함수보다는 띠너비(bandwidth) 혹은 평활계수(smoothing parameter)의 선택 크게 의존한다 (Yu와 Jones, 1998). 일반적으로 큰 평활계수는 분산을 줄이는 반면에 편향(bias)은 크게 하며, 작은 평활계수는 그 반대의 상황을 불러오므로, 편향과 분산의 균형(trade-off)을 반영한 적절한 평활계수의 선택이 중요한 문제로 부각된다. 한편, 회귀함수 추정방법의 성능을 평가하는 기준으로는 통상적으로 L_2-노름(norm)이 사용되어 점근평균제곱오차(asymptotic mean squared error; AMSE) 및 점근평균적분제곱오차(asymptotic mean integrated squared error; AMISE)를 최소화하는 평활계수의 선택에 대한 많은 연구가 있으며 이는 확률밀도함수 추정이나 평균 회귀모형 추정에서도 유사하게 진행되어왔다 (Fan과 Gijbels, 1992; Sheather, 2004). 이에 대한 타당한 대안으로, L_1-노름인 점근평균절대오차(asymptotic mean absolute error; AMAE) 및 점근평균적분절대오차(asymptotic mean integrated absolute error; AMIAE)의 사용 또한 그 해석의 유용성과 합리성 측면을 고려할 수 있지만 이론 및 계산상 어려움 등의 이유로 연구가 미뤄져 왔다.

본 논문에서는 국소 선형 복합 분위수 회귀를 활용한 비모수적 회귀함수의 추정에서 AMAE 및 AMIAE를 평가기준으로 삼아 최적의 평활계수를 구하고 그 유일성에 관하여 연구하였다. 나아가, 기존의 평가기준인 AMSE와 AMISE를 사용한 선택과의 관계를 파악하고, 그 성능을 비교하였다. 이러한 과정에서 다양한 상황에서의 모의실험을 통해 제안한 방법의 특성을 규명하였다. 본 논문의 구성은 다음과 같다. 먼저 2절에서는 국소 선형 복합 분위수 회귀에 대하여 간략하게 설명하였다. 3절에서는 제곱오차와 절대오차의 기준에서 국소 영역 및 전체 영역에서의 평활계수의 선택 방법을 다루었다. 4절에서는 모의실험의 분석 결과를 나타내었다.

2. 국소 선형 복합 분위수 회귀

$\{(x_i, y_i)\}_{i=1}^n$을 서로 독립이고 동일한 확률 밀도함수 $f(x, y)$를 갖는 분포로부터 추출한 확률표본이라고 하자. $X = x$에서 Y의 τ번째 조건부 분위수 함수 $q_\tau(x)$는 체크 손실 함수(check loss function)

$\rho_\tau(t) = t[\tau - I(t < 0)]$, $\tau \in (0,1)$을 통해 다음과 같이 정의 된다.

$$q_\tau(x) = \arg\min_a E\{\rho_\tau(Y-a)|X = x\}. \tag{2.1}$$

대칭인 오차분포를 가정했을 때 $m(X) = E(Y|X) = q_{0.5}(X)$이므로, 국소 최소절대 선형 회귀를 통해 식 (1.1)의 $m(X)$를 추정하는 문제는 $\tau = 0.5$에서 식 (2.1)의 $q_\tau(x)$를 추정하는 것과 같다. 이러한 추정을 위해 우선 다음과 같이 x_0에서의 국소 선형 함수로의 근사

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) \equiv a + b(x - x_0)$$

를 고려하면, x_0주변에서의 국소 최소절대 선형 회귀모형을 적합 시킬 수 있다. $K(\cdot)$를 매끄러운 커널함수라고 하면 $m(x_0)$의 국소 최소절대 선형 회귀 추정량은

$$\left(\hat{a}, \hat{b}\right) = \arg\min_{a,b} \left[\sum_{i=1}^{n} \rho_{0.5}\left(y_i - a_k - b(x_i - x_0)\right) K\left(\frac{x_i - x_0}{h}\right)\right]$$

을 만족하는 $\hat{a}$이며, 여기서 h는 평활계수이다. $\rho_{\tau_k}(t) = t[\tau_k - I(t < 0)]$, $k = 1, 2, \ldots, q$를 q개의 분위수 $\tau_k = k/(q+1)$에서의 체크 손실함수라고 하자. 다수의 통계적 모형의 추정에 있어서 이들이 공유하는 공통적인 정보를 동시에 고려하는 것은 매우 유용하다 (Breiman과 Frideman, 1997). 이러한 특성을 이용하여 Zou와 Yuan (2008)은 서로 다른 τ_k 값에 대한 여러 개(q)의 분위수 회귀모형을 동시적으로 적합하는 복합 분위수 회귀 방법을 제시하였으며, 이 때의 손실함수는 다음과 같이 정의 된다.

$$\sum_{k=1}^{q}\sum_{i=1}^{n} \rho_{\tau_k}(y_i - a_k - bx_i). \tag{2.2}$$

Kai 등 (2010)은 식 (2.2)의 CQR 손실함수와 매끄러운 커널함수 $K(\cdot)$를 결합하여 다음과 같은 국소 가중 CQR 손실함수

$$\sum_{k=1}^{q}\left[\sum_{i=1}^{n} \rho_{\tau_k}\left\{y_i - a_k - b(x_i - x_0)\right\} K\left(\frac{x_i - x_0}{h}\right)\right] \tag{2.3}$$

의 최소화 문제를 고려하였다. 식 (2.3)을 최소화 하는 값들을 $(\hat{a}_1, \ldots, \hat{a}_q, \widehat{b})$로 나타내자. 여기서 $(1/q)\sum_{k=1}^{q} \tau_k = 0.5$이므로, $m(x_0)$와 $m'(x_0)$의 국소 복합 분위수 회귀 추정량은 다음과 같다.

$$\hat{m}(x_0; h) = \frac{1}{q}\sum_{k=1}^{q} \hat{a}_k,$$
$$\tilde{m}'(x_0; h) = \hat{b}.$$

국소 복합 분위수 회귀의 점근적 성질을 규명하기 위하여 몇 가지 표기의 도입이 필요하다. $f(\cdot)$와 $F(\cdot)$를 각각 오차 분포의 확률밀도함수와 누적분포함수라고 정의하고, 공변량 X의 주변확률밀도함수를 $f_X(\cdot)$로 나타내자. 커널함수 $K(\cdot)$로 대칭형 확률밀도함수를 선택하고, 임의의 음이 아닌 정수 j에 대하여 $\mu_j = \int u^j K(u)du$, $\eta_j = \int u^j K^2(u)du$라고 하자. 또한,

$$R_1(q) = \frac{1}{q^2}\sum_{k=1}^{q}\sum_{k'=1}^{q} \frac{\tau_{kk'}}{f(d_k)f(d_{k'})}$$

로 정의하자. 여기서 $d_k = F^{-1}(\tau_k)$이고, $\tau_{kk'} = \tau_k \wedge \tau_{k'} - \tau_k\tau_{k'}$이다.

x_0를 $f_X(\cdot)$의 받침(support)의 내부점이라고 하면, 정칙조건(regularity condition) 하에서 $h \to 0$이면서 $nh \to \infty$일 때, 국소 복합 분위수 회귀 추정량 $\hat{m}(x_0; h)$의 기댓값과 분산은

$$E[\hat{m}(x_0; h)] = m(x_0) + b(x_0)h^2 + o_p\left(h^2\right),$$
$$\mathrm{Var}[\hat{m}(x_0; h)] = \frac{1}{nh}s^2(x_0) + o_p\left(\frac{1}{nh}\right) \tag{2.4}$$

와 같이 주어지며, 여기서

$$b(x_0) = \frac{1}{2}m''(x_0)\mu_2, \qquad s^2(x_0) = \frac{\eta_0\sigma^2(x_0)}{f_X(x_0)}R_1(q)$$

이다. 나아가

$$\sqrt{nh}\left\{\hat{m}(x_0; h) - m(x_0) - b(x_0)h^2\right\} \xrightarrow{d} N\left\{0, s^2(x_0)\right\} \tag{2.5}$$

이 성립하며, 여기서 $\xrightarrow{d}$는 분포수렴을 나타낸다.

식 (2.4)와 (2.5)에 대한 증명 및 이에 요구되는 정칙조건에 관한 자세한 내용은 Kai 등 (2010)에 나타나 있다.

3. 평활계수 선택

커널함수를 이용한 국소회귀모형의 추정에 있어서 최적인 평활계수를 선택하는 것은 추정 모형의 정확성을 결정짓는 매우 중요한 문제이다. x_0에서의 최적인 국소 평활계수는 $m(x_0)$와 그 추정량 $\hat{m}(x_0; h)$의 거리를 최소화하는 평활계수이다. 본 논문에서는 최적인 국소 평활계수의 선택을 위한 거리의 측도로 점근평균제곱오차(AMSE)와 점근평균절대오차(AMAE)를 고려하였다. 한편, 실제 추정모형의 정확성을 평가하는 데 있어서 특정 x_0에서의 성능을 살펴보는 것 보다, 가능한 전체 변수 영역에서의 추정 회귀선의 성능을 측정하는 것이 더 타당하다. 이 경우 점근평균적분제곱오차 $\mathrm{AMISE}(\hat{m}) = \int \mathrm{AMSE}\{\hat{m}(x)\}w(x)dx$ 또는 점근평균적분절대오차 $\mathrm{AMIAE}(\hat{m}) = \int \mathrm{AMAE}\{\hat{m}(x)\}w(x)dx$를 최소화하는 상수 평활계수를 선택할 수 있을 것이며, 여기서 $w(x) > 0$는 가중함수이다.

3.1. AMSE / AMISE를 최소화하는 평활계수 선택

Kai 등 (2010)은 국소 복합 분위수 회귀에서의 최적의 평활계수의 선택을 위한 기준으로 AMSE와 AMISE를 고려하였다. 식 (2.4)의 결과로부터, x_0에서의 $m(x_0)$의 추정량 $\hat{m}(x_0; h)$의 AMSE는

$$\mathrm{AMSE}\{\hat{m}(x_0; h)\} = \mathrm{Bias}\{\hat{m}(x_0; h)\}^2 + \mathrm{Var}\{\hat{m}(x_0; h)\} = b^2(x_0)h^4 + \frac{1}{nh}s^2(x_0) \tag{3.1}$$

으로 주어진다. 식 (3.1)로부터 알 수 있듯이, 큰 평활계수의 선택은 분산을 줄이는 반면에 편향을 크게 하며, 작은 평활계수는 그 반대의 상황을 불러오는 균형(trade-off)의 문제가 발생한다. 따라서 편향과 분산을 동시에 고려하여 AMSE를 최소화하는 평활계수를 선택하는 것이 한 기준이 될 수 있다. 식 (3.1)의 $\mathrm{AMSE}\{\hat{m}(x_0; h)\}$는 $h > 0$에서 h에 대한 볼록(convex) 함수이므로, h에 대한 일차 미분을 통해 이를 최소화하는 평활계수를 구할 수 있다. $h_{\mathrm{opt}}^{\mathrm{MSE}}(x_0) = \arg\min_h \mathrm{AMSE}\{\hat{m}(x_0; h)\}$라고 하면,

$$h_{\mathrm{opt}}^{\mathrm{MSE}}(x_0) = \left(\frac{s(x_0)}{2b(x_0)}\right)^{\frac{2}{5}} n^{-\frac{1}{5}} = c_2(x_0)n^{-\frac{1}{5}} \tag{3.2}$$

가 되며, 여기서 $c_2(x_0)^{5/2} = s(x_0)/|2b(x_0)|$이다. 3.2절에서 다룰 AMAE를 최소화하는 평활계수와의 비교를 위해 $\xi_2 = \{c_2(x_0)\}^{5/2} b(x_0)/s(x_0)$라고 하자. 그러면, $|\xi_2| = 1/2$임을 알 수 있다.

식 (3.1)로부터 $m(X)$의 추정량 $\hat{m}(X;h)$의 AMISE는

$$\text{AMISE}\{\hat{m}(X;h)\} = \int \left[\left\{ \frac{1}{2} m''(x)\mu_2 \right\}^2 h^4 + \frac{1}{nh} \frac{\eta_0 \sigma^2(x)}{f_X(x)} R_1(q) \right] w(x)dx \tag{3.3}$$

로 주어진다. 식 (3.3)으로부터 적분을 통해 x를 제거시켰으므로 AMISE$\{\hat{m}(X;h)\}$는 오직 h에 대한 함수임을 알 수 있으며, AMSE$\{\hat{m}(x_0;h)\}$와 마찬가지로 AMISE$\{\hat{m}(X;h)\}$ 또한 $h > 0$에서 h에 대한 볼록(convex) 함수이다. 따라서 AMISE$\{\hat{m}(X;h)\}$의 h에 대한 일차 미분을 통해 이를 최소화하는 평활계수를 구할 수 있다. $h_{\text{opt}}^{\text{MISE}} = \arg\min_h \text{AMISE}\{\hat{m}(X;h)\}$라고 하면,

$$h_{\text{opt}}^{\text{MISE}} = \left(\frac{\eta_0 R_1(q) \int \sigma^2(x)\{f_X(x)\}^{-1} w(x)dx}{\mu_2^2 \int \{m''(x)\}^2 w(x)dx} \right)^{\frac{1}{5}} n^{-\frac{1}{5}} \tag{3.4}$$

가 된다.

3.2. AMAE / AMIAE를 최소화하는 평활계수 선택

평균절대오차는 적절한 변환에 대한 불변성 및 기하적 해석의 용이성 등 여러 장점을 갖고 있어 평균제곱오차의 좋은 대안으로 고려될 수 있다. 표준정규분포를 따르는 확률변수 Z에 대하여 $\varphi(y) = E(|Z - y|)$라고 하자. 한편, 확률변수 X가 평균이 μ이고 분산이 σ^2인 정규분포를 따른다면,

$$E|X| = E|\sigma Z + \mu| = \sigma E\left|Z + \frac{\mu}{\sigma}\right| = \sigma\varphi\left(-\frac{\mu}{\sigma}\right)$$

이 성립하므로, x_0에서의 $m(x_0)$의 추정량 $\hat{m}(x_0;h)$의 AMAE는 다음과 같이 주어진다.

$$\text{AMAE}\{\hat{m}(x_0;h)\} = (nh)^{-\frac{1}{2}} s(x_0)\varphi\left(-\frac{b(x_0)}{s(x_0)} (nh^5)^{\frac{1}{2}} \right). \tag{3.5}$$

3.1절에서와 마찬가지로, $h = cn^{-1/5}$의 형태를 고려하자. 그러면 h에 관해 나타낸 식 (3.5)를 $cn^{-1/5}$에 관해

$$\text{AMAE}\left\{\hat{m}\left(x_0; cn^{-\frac{1}{5}}\right)\right\} = c^{-\frac{1}{2}} n^{-\frac{2}{5}} s(x_0)\varphi\left(-\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}} \right) \tag{3.6}$$

와 같이 나타낼 수 있으며, 식 (3.5)를 h에 관해 최소화 하는 문제는 식 (3.6)을 c에 관해 최소화하는 문제와 동치임을 알 수 있다. AMSE와 마찬가지로 AMAE$\{\hat{m}(x_0;cn^{-1/5})\}$도 $c > 0$에서 c에 대한 볼록함수이므로, AMAE$\{\hat{m}(x_0;cn^{-1/5})\}$의 c에 대한 일차 미분을 통해 이를 최소화하는 평활계수를 구할 수 있다. $c_1(x_0) = \arg\min_c \text{AMAE}\{\hat{m}(x_0;cn^{-1/5})\}$라고 하면 AMAE를 최소화하는 평활계수는 $h_{\text{opt}}^{\text{MAE}}(x_0) = c_1(x_0)n^{-1/5}$이다. $c_1(x_0)$의 값과 관련하여 다음 정리가 성립한다.

정리 3.1 $\xi_1 = c_1^{5/2}(x_0)b(x_0)/s(x_0)$라고 하자. 그러면 ξ_1는 다음 방정식의 해이다.

$$\phi(-\xi_1) + 2\xi_1[2\Phi(-\xi_1) - 1] = 0, \tag{3.7}$$

여기서 $\phi(\cdot)$, $\Phi(\cdot)$는 각각 표준정규분포를 따르는 확률변수의 확률밀도함수와 누적분포함수이다.

식 (3.7)의 해는 폐쇄형으로 나타나지 않으나 수치해석 방법으로 소수점 셋째 자리까지 구하면 $|\xi_1| \approx 0.481$이다.

ξ_1과 3.1절에서 구한 ξ_2는 모두 $c^{5/2}b(x_0)/s(x_0)$의 형태이며, AMSE와 AMAE를 최소화시키는 평활계수 역시 $h = cn^{-1/5}$의 형태이다. 또한 $c_1(x_0)/c_2(x_0)$는 $b(x_0)$와 $s(x_0)$에 의존하지 않으므로, 임의의 커널함수와 오차 및 공변량의 확률분포 그리고 복합 분위수 회귀에 사용하는 분위수의 개수에 관계없이 이 비율은 동일한 값을 갖는다. 따라서

$$\frac{h_{\text{opt}}^{\text{MAE}}}{h_{\text{opt}}^{\text{MSE}}} = \frac{c_1(x_0)}{c_2(x_0)} = \left(\frac{\xi_1}{\xi_2}\right)^{\frac{2}{5}} = \left(\frac{0.481}{0.5}\right)^{\frac{2}{5}} = 0.985 \tag{3.8}$$

의 관계가 성립하며, AMAE를 최소화하는 평활계수는 AMSE를 최소화하는 평활계수보다 약 1.5% 작음을 알 수 있다. 이러한 결과는 Schucany (1989)가 확률밀도함수의 커널 추정법에서 밝힌 것과 유사하다.

다음으로 $m(X)$의 추정량 $\hat{m}(X; cn^{-1/5})$의 AMIAE를 나타내기 위하여 AMAE에서 논의한 것과 같은 방법으로 $h = cn^{-1/5}$를 고려하자. 식 (3.5)로부터 $\hat{m}(X; cn^{-1/5})$의 AMIAE는

$$\text{AMIAE}\left\{\hat{m}\left(X; cn^{-\frac{1}{5}}\right)\right\} = \int c^{-\frac{1}{2}} n^{-\frac{2}{5}} s(x)\varphi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right) w(x)dx \tag{3.9}$$

로 주어진다. 그러면 식 (3.9)를 최소화하는 상수 평활계수 h를 $h_{\text{opt}}^{\text{MIAE}} = c_{\text{opt}}^{\text{MIAE}} n^{-1/5}$로 나타낼 수 있으며, 여기서 $c_{\text{opt}}^{\text{MIAE}} = \arg\min_c \text{AMIAE}\{\hat{m}(X; cn^{-1/5})\}$이다. 즉, AMIAE$\{\hat{m}(X; h)\}$를 최소화하는 h를 찾는 문제는 AMIAE$\{\hat{m}(X; cn^{-1/5})\}$를 최소화하는 c를 찾는 문제와 동일하다.

정리 3.2 $f(c, x) = c^{-1/2}n^{-2/5}s(x)\varphi(-(b(x)/s(x))c^{5/2})w(x)$가 $c > 0$와 $f_X(\cdot)$의 받침(SUPPORT)에서 연속이면, AMIAE$\{\hat{m}(X; cn^{-1/5})\}$를 최소화하는 $c = c_0$는 유일하게 존재한다.

AMIAE$\{\hat{m}(X; cn^{-1/5})\}$의 이계도함수가 존재하므로, 뉴턴 방법(Newton's method)을 이용한 방정식

$$\frac{\partial}{\partial c}\text{AMIAE}\left\{\hat{m}\left(X; cn^{-\frac{1}{5}}\right)\right\} = 0$$

의 해 $c = c_0$를 구할 수 있으며, 이를 통해 식 (3.9)를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MIAE}} = c_{\text{opt}}^{\text{MIAE}} n^{-1/5}$를 구할 수 있다. AMSE와 AMAE를 최소화시키는 평활계수의 비 $h_{\text{opt}}^{\text{MAE}}/h_{\text{opt}}^{\text{MSE}}$가 약 0.985로 임의의 커널함수와 오차 및 공변량의 확률분포에 관계없이 동일한 값을 갖는 것과 달리, $h_{\text{opt}}^{\text{MIAE}}$와 $h_{\text{opt}}^{\text{MISE}}$의 비 $h_{\text{opt}}^{\text{MIAE}}/h_{\text{opt}}^{\text{MISE}}$는 식 (3.4)와 식 (3.9)의 형태로부터 커널함수, 공변량 X의 분포 $f_X(\cdot)$ 및 표준편차 함수 $\sigma(X)$와 오차항의 분포 $f(\cdot)$, 그리고 복합 분위수 회귀에 사용한 분위수의 개수 q에 따라 달라질 수 있음을 알 수 있다. 정리 3.1과 정리 3.2에 대한 증명은 부록에서 확인할 수 있다.

4. 모의실험

이 장에서는 3장에서 살펴본 다양한 평가 기준 즉, AMSE / AMISE와 AMAE / AMIAE를 최소화하는 국소 복합 분위수 회귀 방법에서의 평활계수의 선택문제를 몇 가지 실험모형을 통해 다뤄보고자 한다. 이 장의 각 예제에서 커널함수는 Epanechnikov 커널함수 $K(u) = (3/4)(1-u^2)_+$를 이용하였으며, 따라서 $\mu_2 = \int u^2 K(u)du = 1/5$, $\eta_0 = \int K^2(u)du = 3/5$로 주어진다. 또한 AMISE와 AMIAE에서의 가중함수 $w(x)$는 공변량 X의 확률밀도함수 $f_X(\cdot)$로 가정하였다.

Table 4.1. Optimal bandwidth selection based on AMAE and AMSE for Example 4.1

ε	q	x_0	Optimal bandwidth			AMAE($\times 10^2$)		AMSE($\times 10^2$)	
			$h_{\text{opt}}^{\text{MAE}}$	$h_{\text{opt}}^{\text{MSE}}$	$h_{\text{opt}}^{\text{MAE}}/h_{\text{opt}}^{\text{MSE}}$	$h_{\text{opt}}^{\text{MAE}}$	$h_{\text{opt}}^{\text{MSE}}$	$h_{\text{opt}}^{\text{MAE}}$	$h_{\text{opt}}^{\text{MSE}}$
$N(0,1)$	1	-1	0.3334	0.3386	0.9845813	7.4687	7.4707	0.8702	0.8698
		0	0.1059	0.1075	0.9845813	13.2530	13.2570	2.7401	2.7388
Laplace	5	0.2	0.2442	0.2481	0.9845813	8.6960	8.6982	1.1797	1.1791
		0.5	0.2970	0.3017	0.9845813	7.8858	7.8880	0.9701	0.9697
$t(3)$	9	0.8	0.3303	0.3355	0.9845813	7.9687	7.9708	0.9906	0.9901
		2	0.3675	0.3733	0.9845813	7.5543	7.5563	0.8903	0.8898

AMAE = asymptotic mean absolute error; AMSE = asymptotic mean squared error.

예제 4.1: $Y = \sin(2X) + 2\exp(-16X^2) + 0.5\varepsilon$.

본 예제는 Kai 등 (2010)에서 다룬 것과 같으며, 여기서 X는 표준정규분포를 따른다. 이 예제에서는 오차항 ε의 분포로 표준정규분포($N(0,1)$), 라플라스분포(Laplace), 그리고 자유도가 3인 t분포(t_3)를 고려하였다. 복합 분위수 회귀에 사용하는 분위수의 개수 q는 $q = 1, 5, 9$를 고려하였으며, 표본의 크기는 $n = 200$이다.

Table 4.1에는 특정 x_0지점에서의 AMAE와 AMSE를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MAE}}$, $h_{\text{opt}}^{\text{MSE}}$와 그 평활계수를 통해 계산한 AMAE값과 AMSE값이 정리되어 있다. Table 4.1에서 보는 바와 같이, AMAE를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MAE}}$와 AMSE를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MSE}}$의 비 $h_{\text{opt}}^{\text{MAE}}/h_{\text{opt}}^{\text{MSE}}$는 오차항의 분포 및 복합 분위수 회귀에 사용되는 분위수의 수, 그리고 최소화 시키고자 하는 지점 x_0와 관계없이 항상 0.9846의 값을 갖는 것을 확인할 수 있으며, 이러한 결과는 3.2절에서 살펴본 식 (3.8)과 일치한다. 또한 두 평활계수 값의 근소한 차이로 인해 각 평활계수를 사용하였을 때의 AMAE와 AMSE값은 거의 유사하게 나타난다.

예제 4.2: $Y = X\sin(2\pi X) + \sigma(X)\varepsilon$.

예제 2는 AMIAE와 AMISE를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MIAE}}$와 $h_{\text{opt}}^{\text{MISE}}$의 관계를 알아보기 위해 고안되었다. $h_{\text{opt}}^{\text{MIAE}}$와 $h_{\text{opt}}^{\text{MISE}}$의 비 $h_{\text{opt}}^{\text{MIAE}}/h_{\text{opt}}^{\text{MISE}}$는 공변량 X의 분포 $f_X(\cdot)$ 및 표준편차 함수 $\sigma(X)$와 오차항의 분포 $f(\cdot)$, 그리고 복합 분위수 회귀에서의 분위수의 개수 q에 따라 달라질 수 있으며, 이들 가운데 어떠한 특성이 $h_{\text{opt}}^{\text{MIAE}}/h_{\text{opt}}^{\text{MISE}}$에 영향을 주는지 파악할 필요가 있다. 먼저 오차항 ε의 분포를 표준정규분포로, 표준편차 함수는 $\sigma(X) = 1$로 고정한 상태에서 다양한 공변량 X의 분포 $f_X(\cdot)$에 따라 $h_{\text{opt}}^{\text{MIAE}}/h_{\text{opt}}^{\text{MISE}}$가 어떻게 달라지는 지 실험하였다. 비교를 위해 공변량 X의 분포로 균등분포(Unif$(0,1)$), 모수로 $\alpha = \beta = 2$를 갖는 베타 분포(Beta$(2,2)$), 그리고 모수로 $\alpha = 2$, $\beta = 4$를 갖는 베타 분포(Beta$(2,4)$)를 고려하였으며, 복합 분위수 회귀에서의 분위수의 개수 q는 $q = 1, 5, 9$를 사용하였다.

Table 4.2에는 표본의 크기가 $n = 400$일 때 다양한 공변량 X의 분포 $f_X(\cdot)$에서의 AMIAE와 AMISE를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MIAE}}$, $h_{\text{opt}}^{\text{MISE}}$와 그 평활계수를 통해 계산한 AMIAE의 값과 AMISE의 값이 정리되어 있다. 먼저 복합 분위수 회귀에서의 분위수의 개수 q를 다르게 사용하였을 때 최적 평활계수의 비 $h_{\text{opt}}^{\text{MIAE}}/h_{\text{opt}}^{\text{MISE}}$살펴보면, 공변량 X의 분포 $f_X(\cdot)$에 관계없이 소수점 넷째 자리까지 동일한 값을 가졌으며, 그 비는 1에 가까움을 알 수 있다. 또한 공변량 X의 분포 $f_X(\cdot)$가 달라져도, 최적 평활계수의 비 $h_{\text{opt}}^{\text{MIAE}}/h_{\text{opt}}^{\text{MISE}}$는 큰 차이를 보이지 않고 역시 1에 가까운 값을 가졌으며, 최적 평활계수에서

Table 4.2. Optimal bandwidth selection based on AMIAE and AMISE for Example 4.2 with various distributions of the covariate

$f_X(\cdot)$	q	Optimal bandwidth			AMIAE($\times 10^2$)		AMISE($\times 10^2$)	
		$h_{\text{opt}}^{\text{MIAE}}$	$h_{\text{opt}}^{\text{MISE}}$	$h_{\text{opt}}^{\text{MIAE}}/h_{\text{opt}}^{\text{MISE}}$	$h_{\text{opt}}^{\text{MIAE}}$	$h_{\text{opt}}^{\text{MISE}}$	$h_{\text{opt}}^{\text{MIAE}}$	$h_{\text{opt}}^{\text{MISE}}$
Unif$(0,1)$	1	0.1729	0.1741	0.9929493	10.3960	10.3960	1.6919	1.6918
	5	0.1607	0.1618	0.9929178	8.9796	8.9801	1.2624	1.2623
	9	0.1593	0.1604	0.9929175	8.8295	8.830	1.2205	1.2204
Beta$(2,2)$	1	0.1756	0.1768	0.9931248	10.3150	10.3150	1.6659	1.6657
	5	0.1632	0.1643	0.9931007	8,9099	8.9103	1.2430	1.2428
	9	0.1618	0.1630	0.9931006	8.7609	8.7614	1.2017	1.2016
Beta$(2,4)$	1	0.1987	0.2002	0.9924861	9.6960	9.6966	1.4716	1.4715
	5	0.1846	0.1860	0.9924714	8.3753	8.3758	1.0980	1.0979
	9	0.1831	0.1845	0.9924912	8.2352	8.2357	1.0616	1.0615

AMIAE = asymptotic mean integrated absolute error; AMISE = asymptotic mean integrated squared error.

Table 4.3. Optimal bandwidth selection based on AMIAE and AMISE for various heteroscedastic error distributions

$\sigma(X)$	ε	Optimal bandwidth			AMIAE($\times 10^2$)		AMISE($\times 10^2$)	
		$h_{\text{opt}}^{\text{MIAE}}$	$h_{\text{opt}}^{\text{MISE}}$	$h_{\text{opt}}^{\text{MIAE}}/h_{\text{opt}}^{\text{MISE}}$	$h_{\text{opt}}^{\text{MIAE}}$	$h_{\text{opt}}^{\text{MISE}}$	$h_{\text{opt}}^{\text{MIAE}}$	$h_{\text{opt}}^{\text{MISE}}$
1	$N(0,1)$	0.1607	0.1618	0.9929178	8.9797	8.9801	1.2624	1.2623
	Laplace	0.1726	0.1739	0.9929478	10.3670	10.3680	1.6826	1.6824
	t	0.1750	0.1763	0.9928824	10.6570	10.6580	1.7781	1.7779
$\lvert X\cos(2\pi X)\rvert$	$N(0,1)$	0.1210	0.1287	0.9402747	5.2592	5.2797	0.5088	0.5052
	Laplace	0.1300	0.1383	0.9402760	6.0717	6.0955	0.6781	0.6733
	t	0.1318	0.1402	0.9402763	6.2416	6.2661	0.7166	0.7115
$e^{\sin(2\pi x)}$	$N(0,1)$	0.1568	0.1696	0.9246137	9.7105	9.7701	1.5418	1.5245
	Laplace	0.1685	0.1823	0.9246102	11.2110	11.2810	2.0551	2.0319
	t	0.1709	0.1848	0.9246085	11.5250	11.5960	2.1717	2.1472

AMIAE = asymptotic mean integrated absolute error; AMISE = asymptotic mean integrated squared error.

의 AMIAE 및 AMISE는 큰 차이를 보이지 않았다. 따라서 AMIAE를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MIAE}}$와 AMISE를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MISE}}$의 비 $h_{\text{opt}}^{\text{MIAE}}/h_{\text{opt}}^{\text{MISE}}$는 공변량 X의 분포 $f_X(\cdot)$와 복합 분위수 회귀에서의 분위수의 개수 q에 큰 영향을 받지 않음을 확인하였다.

다음으로 오차항 ε의 분포 및 표준편차 함수 $\sigma(X)$의 형태에 따라 $h_{\text{opt}}^{\text{MIAE}}/h_{\text{opt}}^{\text{MISE}}$가 어떻게 달라지는 지 실험하였다. 먼저 공변량 X의 분포는 구간 (0, 1)에서 정의된 균등분포로 고정하였으며, 복합 분위수 회귀에서의 분위수의 개수는 $q = 5$를 이용하였다. 비교를 위한 오차항 ε의 분포로 실험모형 1에서와 같이 표준정규분포($N(0,1)$), 라플라스분포(Laplace), 그리고 자유도가 3인 t분포(t_3)를, 표준편차 함수 $\sigma(X)$로는 1, $|X\cos(2\pi X)|$, $e^{\sin(2\pi x)}$의 형태를 고려하였다.

Table 4.3에는 표본의 크기가 $n = 400$일 때 다양한 오차항 ε의 분포 및 표준편차 함수 $\sigma(X)$에 따른 AMIAE와 AMISE를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MIAE}}$, $h_{\text{opt}}^{\text{MISE}}$와 그 평활계수를 통해 계산한 AMIAE값과 AMISE값이 정리되어 있다. 표준편차 함수 $\sigma(X)$가 동일할 때, 오차항 ε의 분포가 달라져도 AMIAE를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MIAE}}$와 AMISE를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MISE}}$의 비 $h_{\text{opt}}^{\text{MIAE}}/h_{\text{opt}}^{\text{MISE}}$는 크게 달라

지지 않았으며, 이는 Table 4.2에서의 결과와 유사하다. 하지만, 표준편차 함수 $\sigma(X)$가 상수값을 갖지 않는 경우에는 두 평활계수의 비 $h_{\text{opt}}^{\text{MIAE}}/h_{\text{opt}}^{\text{MISE}}$가 1로부터 다소 멀어졌으며, 이에 따라 AMIAE값과 AMISE값 역시 차이가 나타났다. 즉, 오차항에 이분산성이 존재할 경우, 모형의 최적성을 평가하는 기준에 따라 이를 최소화하는 평활계수가 달라질 수 있음을 알 수 있다.

5. 결론

본 논문에서는 국소 선형 복합 분위수 회귀에서의 평활계수를 선택하는 기준으로 L_1-노름을 활용한 AMAE와 AMIAE를 제시하고, 이를 최소화하는 평활계수가 유일하게 존재함을 이론적으로 밝혔다. AMAE를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MAE}}$와 AMSE를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MSE}}$의 비 $h_{\text{opt}}^{\text{MAE}}/h_{\text{opt}}^{\text{MSE}}$는 약 0.985로 커널함수와 오차 및 공변량의 확률분포 그리고 복합 분위수 회귀에 사용하는 분위수의 개수에 관계없이 동일함을 수치해석적 방법을 이용하여 확인하였다. 또한 AMIAE를 최소화하는 평활 계수 $h_{\text{opt}}^{\text{MIAE}}$와 AMISE를 최소화하는 평활계수 $h_{\text{opt}}^{\text{MISE}}$의 비 $h_{\text{opt}}^{\text{MIAE}}/h_{\text{opt}}^{\text{MISE}}$는 공변량 및 오차항의 분포에 크게 영향을 받지 않으며, 오차항이 등분산성을 갖는 경우 거의 1에 가까운 값을 가지지만, 오차항이 이분산성을 갖는 경우에는 달라질 수 있음을 모의실험을 통하여 확인하였다.

부록

보조정리 1.

$$\text{AMAE}\left\{\hat{m}\left(x_0; cn^{-\frac{1}{5}}\right)\right\} = c^{-\frac{1}{2}} n^{-\frac{2}{5}} s(x_0)\varphi\left(-\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}}\right)$$

는 볼록함수이다.

증명: 편의상 $\psi_n^{L_1}(x_0, c) = \text{AMAE}\{\hat{m}(x_0; cn^{-1/5})\} = c^{-1/2} n^{-2/5} s(x_0)\varphi(-(b(x_0)/s(x_0))c^{5/2})$로 나타내자.

$$\begin{aligned}
\frac{\partial}{\partial c}\psi_n^{L_1}(x_0, c) &= -\frac{1}{2} c^{-\frac{3}{2}} n^{-\frac{2}{5}} s(x_0)\varphi\left(-\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}}\right) \\
&\quad + c^{-\frac{1}{2}} n^{-\frac{2}{5}} s(x_0)\left(-\frac{5}{2}\frac{b(x_0)}{s(x_0)} c^{\frac{3}{2}}\right)\varphi'\left(-\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}}\right) \\
&= -\frac{1}{2} c^{-\frac{3}{2}} n^{-\frac{2}{5}} s(x_0)\varphi\left(-\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}}\right) - \frac{5}{2} cn^{-\frac{2}{5}} b(x_0)\varphi'\left(-\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}}\right), \\
\frac{\partial^2}{\partial^2 c}\psi_n^{L1}(x_0, c) &= \frac{3}{4} c^{-\frac{5}{2}} s(x_0)\varphi\left(-\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}}\right) + \frac{5}{4} b(x_0)\varphi'\left(-\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}}\right) \\
&\quad - \frac{5}{2} b(x_0)\varphi'\left(-\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}}\right) + \frac{25}{4} c^{\frac{5}{2}} \frac{b^2(x_0)}{s(x_0)}\varphi''\left(-\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}}\right) \\
&= \frac{3}{4} b(x_0)\frac{s(x_0)}{b(x_0)} c^{-\frac{5}{2}}\varphi\left(-\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}}\right) - \frac{5}{4} b(x_0)\varphi'\left(-\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}}\right) \\
&\quad + \frac{25}{4} b(x_0)\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}}\varphi''\left(-\frac{b(x_0)}{s(x_0)} c^{\frac{5}{2}}\right).
\end{aligned}$$

$\xi_1 = c^{5/2} b(x_0)/s(x_0)$라고 하고, $(\partial^2/\partial^2 c)\psi_n^{L1}(x_0, c)$를 ξ_1에 관한 함수 $(\partial^2/\partial^2 c)\psi_n^{L1}(x_0, \xi_1)$로 나타내

면,

$$\begin{aligned}\frac{\partial^2}{\partial^2 c}\psi_n^{L1}(x_0,\xi_1) &= \frac{3}{4}b(x_0)\xi_1^{-1}\varphi(-\xi_1) - \frac{5}{4}b(x_0)\varphi'(-\xi_1) + \frac{25}{4}b(x_0)\xi_1\varphi''(-\xi_1) \\ &= \frac{b(x_0)}{4}\left[3\xi_1^{-1}\left\{2\phi(-\xi_1) - \xi_1(2\Phi(-\xi_1)-1\right\} - 5\left\{2\Phi(-\xi_1)-1\right\} + 25\xi_1\left\{2\phi(-\xi_1)\right\}\right] \\ &= \frac{b(x_0)}{2}\left[3\xi_1^{-1}\phi(-\xi_1) - 4\left\{2\Phi(-\xi_1)-1\right\} + 25\xi_1\phi(-\xi_1)\right] \qquad \text{(A.1)}\end{aligned}$$

이며, $(\partial^2/\partial^2 c)\psi_n^{L1}(x_0,\xi)$은 모든 $\mathrm{sgn}(b(x_0))\xi_1 > 0$에서 양의 값을 갖는다. 한편 $\mathrm{sgn}(b(x_0)) = \mathrm{sgn}(\xi_1)$이고, $b(x_0) = 0$일 때, $(\partial^2/\partial^2 c)\psi_n^{L1}(x_0,c) = (3/4)c^{-5/2}s(x_0)\varphi(0) > 0$이므로, $(\partial^2/\partial^2 c)\psi_n^{L1}(x_0,c)$는 모든 $c>0$에서 양의 값을 가진다. 따라서 $\psi_n^{L_1}(x_0,c)$는 볼록함수이다. □

정리 3.1. $\xi_1 = c_1^{5/2}(x_0)b(x_0)/s(x_0)$라고 하자. 그러면 ξ_1는 다음 방정식의 해이다.

$$\phi(-\xi_1) + 2\xi_1[2\Phi(-\xi_1)-1] = 0.$$

증명: 보조정리 1에 의해 $\psi_n^{L_1}(x_0,c)$는 볼록함수이므로 $\psi_n^{L_1}(x_0,c)$를 최소화하는 c의 값을 찾는 것은 $(\partial/\partial c)\psi_n^{L_1}(x_0,c)=0$을 만족하는 $c=c_1(x_0)$를 찾는 것과 동치이다. 보조정리 1로부터,

$$\begin{aligned}\frac{\partial}{\partial c}\psi_n^{L_1}(x_0,c) &= -\frac{1}{2}c^{-\frac{3}{2}}n^{-\frac{2}{5}}s(x_0)\varphi\left(-\frac{b(x_0)}{s(x_0)}c^{\frac{5}{2}}\right) + c^{-\frac{1}{2}}n^{-\frac{2}{5}}s(x_0)\left(-\frac{5}{2}\frac{b(x_0)}{s(x_0)}c^{\frac{3}{2}}\right)\varphi'\left(-\frac{b(x_0)}{s(x_0)}c^{\frac{5}{2}}\right) \\ &= \left(-\frac{1}{2}c^{-\frac{3}{2}}n^{-\frac{2}{5}}s(x_0)\right)\left[\varphi\left(-\frac{b(x_0)}{s(x_0)}c^{\frac{5}{2}}\right) - 5\frac{b(x_0)}{s(x_0)}c^{\frac{5}{2}}\varphi'\left(-\frac{b(x_0)}{s(x_0)}c^{5/2}\right)\right]\end{aligned}$$

이다. $n>0$, $s(x_0)>0$, $c>0$이므로, $(\partial/\partial c)\psi_n^{L_1}(x_0,c)=0$을 만족하는 $c=c_1(x_0)$를 찾는 것은

$$\varphi\left(-\frac{b(x_0)}{s(x_0)}c^{\frac{5}{2}}\right) - 5\frac{b(x_0)}{s(x_0)}c^{\frac{5}{2}}\varphi'\left(-\frac{b(x_0)}{s(x_0)}c^{\frac{5}{2}}\right) = 0 \qquad \text{(A.2)}$$

를 만족하는 $c = c_1(x_0)$를 찾는 것과 동치이다. 여기서, $\xi_1 = c_1^{5/2}(x_0)b(x_0)/s(x_0)$라고 하면, 식 (A.2)는

$$\varphi(-\xi_1) + 5\xi_1\varphi'(-\xi_1) = 0 \qquad \text{(A.3)}$$

으로 표현할 수 있다. 한편,

$$\varphi(y) = E(|Z-y|) = 2\phi(y) + y\{2\Phi(y)-1\}, \qquad \varphi'(y) = 2\Phi(y)-1 \qquad \text{(A.4)}$$

이므로, 식 (A.3)을 식 (A.4)를 통해 나타내면

$$\phi(-\xi_1) + 2\xi_1[2\Phi(-\xi_1)-1] = 0$$

과 같다. □

보조정리 2.

$$\mathrm{AMIAE}\left\{\hat{m}\left(X; cn^{-\frac{1}{5}}\right)\right\} = \int c^{-\frac{1}{2}}n^{-\frac{2}{5}}s(x)\varphi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)w(x)dx$$

는 볼록함수이다.

증명: $f(c,x) = c^{-1/2}n^{-2/5}s(x)\varphi(-(b(x)/s(x))c^{5/2})w(x)$가 $c>0$와 $f_X(\cdot)$의 받침에서 연속이므로,

$$\begin{aligned}\frac{\partial^2}{\partial c^2}\mathrm{AMIAE}\left\{\hat{m}\left(X;cn^{-\frac{1}{5}}\right)\right\} &= \frac{\partial^2}{\partial c^2}\int c^{-\frac{1}{2}}n^{-\frac{2}{5}}s(x)\varphi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)w(x)dx \\ &= \int \frac{\partial^2}{\partial c^2}\left\{c^{-\frac{1}{2}}n^{-\frac{2}{5}}\varphi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)\right\}s(x)w(x)dx \qquad \text{(A.5)}\end{aligned}$$

이고, 보조정리 1의 결과를 이용하면,

$$\begin{aligned}&\frac{\partial^2}{\partial c^2}\left\{c^{-\frac{1}{2}}n^{-\frac{2}{5}}\varphi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)\right\} \\ &= \frac{3}{4}b(x)\frac{s(x)}{b(x)}c^{-\frac{5}{2}}\varphi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right) - \frac{5}{4}b(x)\varphi'\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right) + \frac{25}{4}b(x)\frac{b(x)}{s(x)}c^{\frac{5}{2}}\varphi''\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right) \\ &= \frac{b(x)}{2}\left[3\frac{s(x)}{b(x)}c^{-\frac{5}{2}}\phi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right) - 4\left\{2\Phi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)-1\right\} + 25\frac{b(x)}{s(x)}c^{\frac{5}{2}}\phi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)\right] \qquad \text{(A.6)}\end{aligned}$$

이다. 여기서 $\xi(x) = (b(x)/s(x))c^{5/2}$라고 하면 식 (A.6)는,

$$\frac{b(x)}{2}\left[3\xi(x)^{-1}\phi\left(-\xi(x)\right) - 4\left\{2\Phi\left(-\xi(x)\right)-1\right\} + 25\xi(x)\phi(-\xi(x))\right] \qquad \text{(A.7)}$$

이다. 식 (A.7)은 식 (A.1)과 마찬가지로 모든 $\mathrm{sgn}(b(x))\xi(x) > 0$ 인 x에서 양의 값을 가지며, 모든 x에서 $\mathrm{sgn}(b(x)) = \mathrm{sgn}(\xi(x))$이다. 또한 $s(x)>0$, $w(x)>0$이므로, 식 (A.5)의 적분구간에 속하는 모든 x에서 피적분함수는 양의 값을 갖는다. 따라서 $c>0$에서 $(\partial^2/\partial c^2)\mathrm{AMIAE}\{\hat{m}(X;cn^{-1/5})\} > 0$이므로 $\mathrm{AMIAE}\{\hat{m}(X;cn^{-1/5})\}$는 볼록함수이다. □

정리 3.2. $f(c,x) = c^{-1/2}n^{-2/5}s(x)\varphi(-(b(x)/s(x))c^{5/2})w(x)$가 $c>0$와 $f_X(\cdot)$의 받침에서 연속이면, $\mathrm{AMIAE}\{\hat{m}(X;cn^{-1/5})\}$를 최소화하는 $c=c_0$는 유일하게 존재한다.

증명: 보조정리 2에 의해 $\mathrm{AMIAE}\{\hat{m}(X;cn^{-1/5})\} = \int c^{-1/2}n^{-2/5}s(x)\varphi(-(b(x)/s(x))c^{5/2})w(x)dx$는 볼록함수이다. 따라서, $(\partial/\partial c)\mathrm{AMIAE}\{\hat{m}(X;cn^{-1/5})\} = 0$을 만족하는 $c=c_0$가 존재하면, 그 점에서 $\mathrm{AMIAE}\{\hat{m}(X;cn^{-1/5})\}$는 유일한 최솟값을 가진다.

$$\begin{aligned}&\frac{\partial}{\partial c}\mathrm{AMIAE}\left\{\hat{m}\left(X;cn^{-\frac{1}{5}}\right)\right\} \\ &= \int\left(-\frac{1}{2}c^{-\frac{3}{2}}\right)n^{-\frac{2}{5}}s(x)\varphi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)w(x)dx + \int c^{-\frac{1}{2}}n^{-\frac{2}{5}}s(x)\left(-\frac{5}{2}\frac{b(x)}{s(x)}c^{\frac{3}{2}}\right)\varphi'\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)w(x)dx \\ &= -\frac{1}{2}c^{-\frac{3}{2}}n^{-\frac{2}{5}}\int s(x)\varphi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)w(x)dx - \frac{5}{2}cn^{-\frac{2}{5}}\int b(x)\varphi'\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)w(x)dx \\ &= c^{-\frac{3}{2}}n^{-\frac{2}{5}}\left[-\frac{1}{2}\int\varphi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)s(x)w(x)dx - \frac{5}{2}\int c^{\frac{5}{2}}b(x)\varphi'\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)w(x)dx\right] \\ &= c^{-\frac{3}{2}}n^{-\frac{2}{5}}\int\left[-\frac{1}{2}\varphi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right) - \frac{5}{2}\frac{b(x)}{s(x)}c^{\frac{5}{2}}\varphi'\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)\right]s(x)w(x)dx\end{aligned}$$

이다. $\Delta(c) = \varphi(-(b(x)/s(x))c^{5/2}) + 5\frac{b(x)}{s(x)}c^{5/2}\varphi'(-(b(x)/s(x))c^{5/2})$라고 하자. 그러면 $\Delta(c)$는 식 (A.3)의 형태와 같으므로, $\Delta(c) = 2\phi(-(b(x)/s(x))c^{5/2}) + 4(b(x)/s(x))c^{5/2}[2\Phi(-(b(x)/s(x))c^{5/2}) - 1]$로 바꿔 나타낼 수 있다. 즉,

$$\frac{\partial}{\partial c}\mathrm{AMIAE}\left\{\hat{m}\left(X;cn^{-\frac{1}{5}}\right)\right\} = -\int\left[\phi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right) + 2\frac{b(x)}{s(x)}c^{\frac{5}{2}}\left\{2\Phi\left(-\frac{b(x)}{s(x)}c^{\frac{5}{2}}\right)-1\right\}\right]s(x)w(x)dx$$

이다.

$\vartheta(c) = -\int [\phi(-(b(x)/s(x))c^{5/2}) + 2(b(x)/s(x))c^{5/2}\{2\Phi(-(b(x)/s(x))c^{5/2}) - 1\}]s(x)w(x)dx$라고 하자. $c \to 0$에 따라 $\vartheta(c)$는 $\vartheta(0) = -\int \phi(0)s(x)w(x)dx = -(1/\sqrt{2\pi})\int s(x)w(x)dx < 0$으로 수렴하고, $c \to \infty$에 따라 $\vartheta(c)$도 양의 무한대로 발산하므로, $\vartheta(c) = 0$의 해 $c = c_0$가 존재한다. 따라서, $(\partial/\partial c)\text{AMIAE}\{\hat{m}(X; cn^{-1/5})\} = c^{-3/2}n^{-2/5}\vartheta(c) = 0$의 해 $c = c_0$가 유일하게 존재한다. □

References

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression, *Journal of the Royal Statistical Society, Series B*, **59**, 3–54.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers, *The Annals of Statistics*, **20**, 2008–2036.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.

Fan, J., Hu, T. C., and Truong, Y. K. (1994). Robust non-parametric function estimation, *Scandinavian Journal of Statistics*, **21**, 433–446.

Kai, B., Li, R., and Zou, H (2010). Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression, *Journal of the Royal Statistical Society, Series B*, **72**, 49–69.

Schucany, W. R. (1989). Locally optimal window widths for kernel density estimation with large sample, *Statistics and Probability Letters*, **7**, 401–405.

Sheather, S. J. (2004). Density estimation, *Statistical Science*, **19** 588–597.

Welsh, A. H. (1996). Robust estimation of smooth regression and spread functions and their derivatives, *Statistica Sinica*, **6**, 347–366.

Yu, K. and Jones, M. C. (1998). Local linear quantile regression, *Journal of the American Statistical Association*, **93**, 228–237.

Zou, H. and Yuan, M. (2008). Composite quantile regression and the Oracle model selection theory, *The Annals of Statistics*, **36**, 1108–1126.

국소 선형 복합 분위수 회귀에서의 평활계수 선택

전명식[a] · 강종경[a] · 방성완[b,1]

[a]고려대학교 통계학과, [b]육군사관학교 수학과

(2017년 7월 27일 접수, 2017년 9월 6일 수정, 2017년 9월 11일 채택)

요 약

국소복합분위수 회귀모형을 활용한 비모수적 함수 추정방법이 높은 효율성과 더불어 활발히 연구되고 있다. 이러한 추정과정에 커널을 사용한 자료 평활방법이 대표적으로 사용되고 있으며, 그 성능은 커널보다는 평활계수의 선택 크게 의존한다. 한편, 회귀함수 추정방법의 성능을 평가하는 기준으로는 통상적으로 L_2-노름이 사용되어 평균제곱오차 또는 평균적분제곱오차를 최소화하는 평활계수의 선택에 대한 많은 연구가 진행되어 왔다. 본 논문에서는 국소 선형 복합 분위수 회귀방법을 활용한 비모수 회귀모형 추정량의 성능을 결정하는 평활계수 선택의 최적성에 관해 연구하였다. 특히, 여러 장점을 가졌으나 수리적 어려움으로 연구가 미흡한 평균절대오차 및 평균적분절대오차를 최적의 기준으로 삼아 최적의 평활계수를 구하고 그 유일성에 관해 연구하였다. 나아가 기존의 평가기준인 평균제곱오차 및 평균적분제곱오차를 사용한 선택과의 관계를 파악하고 그 성능을 비교하였다. 이러한 과정에서 다양한 상황에서의 모의실험을 통해 제안한 방법의 특성을 규명하였다.

주요용어: 복합 분위수 회귀, 평활계수 선택, 국소 선형 회귀, 커널 함수, 평균적분절대오차

이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업이며 (NRF-2016R1D1A1B03931114) (전명식), 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (2015R1C1A1A02036473) (방성완).

[1]교신저자: (01805) 서울시 노원구 화랑로 574, 육군사관학교 수학과. E-mail: wan1365@gmail.com