비대칭 라플라스 분포를 이용한 분위수 회귀

Quantile regression using asymmetric Laplace distribution

저자
(Authors)

박혜정
Hye Jung Park

출처
(Source)

한국데이터정보과학회지 20(6), 2009.12, 1093-1101(9 pages)

Journal of the Korean Data And Information Science Society 20(6), 2009.12, 1093-1101(9 pages)

발행처
(Publisher)

한국데이터정보과학회
The Korean Data and Information Science Society

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07244400

APA Style

박혜정 (2009). 비대칭 라플라스 분포를 이용한 분위수 회귀. 한국데이터정보과학회지, 20(6), 1093-1101

Journal of the Korean Data & Information Science Society
2009, **20**(6), 1093–1101

한국데이터정보과학회지

비대칭 라플라스 분포를 이용한 분위수 회귀

박혜정[1]

대구대학교 교양대학

접수 2009년 9월 1일, 수정 2009년 11월 7일, 게재확정 2009년 11월 21일

요 약

분위수 회귀모형은 확률변수들 사이에 확률적인 관계구조를 포함한 함수 모형을 좀 더 완벽하게 추정하도록 제공한다. 본 논문에서는 함수 추정에 로버스트하다고 알려져 있는 서포트벡터기계 기법과 이중벌칙커널기계를 이용하여 분위수 회귀모형을 추정하고자 한다. 이중벌칙커널기계는 고차원의 입력변수에 대한 분위수 회귀가 요구될 때 분위수 회귀모형을 잘 추정한다고 알려져 있다. 또한 본 논문에서는 광범위한 형태의 분위수 회귀모형 추정을 위해서 정규분포보다 비대칭 라플라스 분포를 이용한다. 본 논문에서 제안한 모형은 분위수 회귀모형 추정을 위해서 서포트벡터기계 기법에 이중벌칙커널기계를 이용하여 각각의 평균과 분산을 동시에 추정한다. 평균과 분산함수 추정을 위해 사용된 커널함수의 모수들은 최적의 값을 찾기 위해 일반화근사 교차타당성을 이용한다.

주요용어: 분위수 회귀, 비대칭 라플라스 분포, 서포트벡터기계, 이중벌칙커널기계, 일반화근사 교차타당성함수.

1. 서론

분위수 회귀모형은 확률변수들 사이에 확률적인 관계구조를 포함한 함수 모형을 좀 더 완벽하게 추정한다고 알려져 있다. 분위수 회귀모형은 반응변수의 조건부 분포에 대해 자세한 정보를 제공하며, 이상점이나 오차항 분포에 민감하게 반응하지 않고 로버스트하다고 알려져 있다. 또한 단조변환 (monotone transformation)에 대해 등분산성을 갖는 장점으로 경제관련 회귀모형의 경험적 연구에 유용하게 응용될 수 있음에도 불구하고 미분 불가능한 목적함수를 최소화시켜 추정량을 구해야 하는 계산상의 어려움으로 분위수 회귀추정량을 이용한 실증분석이 미흡했던 것이 사실이다. 그러나 최근 분위수 회귀추정량을 계산하는 알고리즘의 발전과 컴퓨터의 연산 처리장치의 급속한 향상으로 분위수 회귀추정법을 이용한 실증적 연구에 관심이 집중되고 있다. 조건부 기대치만을 계산하던 과거의 연구에서는 도출할 수 없었던 다양하고 유용한 실증분석 결과가 도출되고 있다. 분위수 회귀모형에 관련된 연구는 통계와 경제관련 분야에서 많이 이루어지고 있다.

Koenker와 Bassett (1978)이 소개한 분위수 회귀추정법은 이상점이나 오차항 분포에 민감하게 반응하지 않는 준모수적인 방법을 소개하고 있다. Koenker와 Bassett (1978)에서는 조건부 평균함수만을 추정하는 기존의 최소제곱추정법과 달리 조건부 분위수 함수를 추정하며, 선형 분위수 회귀추정량의 일치성과 점근적 정규성에 대해 증명하고 있다. Koenker와 Bassett (1978)의 분위수 회귀추정법은 조건부 분위수 함수를 추정함으로 반응변수의 조건부 분포의 특성을 잘 파악할 수 있는 장점이 있다. Bassett와 Koenker (1982)는 오차항이 i.i.d.인 선형모형의 경험적 분위수 회귀함수에 대해 설명하였다.

[1] (712-714) 경상북도 경산시 진량읍 내리리 15번지, 대구대학교 교양대학, 초빙교수.
E-mail: hyjpark@daegu.ac.kr

Powell (1986)이 처음으로 선형 분위수 회귀모형을 비선형모형의 특수한 형태인 중절회귀 (censored regression) 모형에 적용하여 추정량의 이론적 특성을 살펴보았으며, White (1992)는 일반적 비선형 회귀모형에 대한 분위수 추정량의 일치성을 증명하였다. Weiss (1991)가 비선형 동적모형 (dynamic models)을 위한 최소절대오차 (least absolute error) 추정량의 일치성과 점근적 정규성을 보여 주면서 비선형 분위수 회귀모형에 대한 연구가 본격화되었다고 볼 수 있다.

분위수 회귀모형 추정과 관련된 연구에서는 대부분 평균함수만을 추정한다. 추정하고자 하는 자료의 분포가 등분산인 경우에는 별문제가 없겠지만, 이분산인 경우에 관한 연구에서는 평균과 분산함수를 고려해야 한다. Hwang과 Shim (2005)은 분위수 회귀모형 추정에 함수추정에 로버스트하다고 알려져 있는 서포트벡터기계 (Support Vector Machine, SVM) 기법을 적용하였다. SVM은 Vapnik (1998)과 Smola와 Scholkopf (1998)에 잘 소개되어 있다. Shim 등 (2009a)에서는 분위수 회귀모형 추정을 위해 이중벌칙 커널기계 (doubly penalized kernel machine) 이용하여 분위수 회귀모형을 추정하고 있다. DPKM으로 분위수 회귀모형을 추정할 때 SVM기법인 SVQR-IRWLS (support vector quantile regression iteratively reweighted least squares)을 이용하여 추정하였다. DPKM에 관한 상세한 내용은 Yuan과 Wahba (2004)에 잘 설명되어 있다. 선행 연구된 논문들에서 분위수 회귀모형을 추정할 때에 대부분 오차항이 정규분포를 따른다는 가정하에 모형을 추정한다. 그러나 경제관련 실증자료들을 보면 정규분포보다는 꼬리가 더 두꺼운 형태의 분포 즉 라플라스 분포를 따르는 경우들이 많다. 본 논문에서는 비대칭인 라플라스 분포를 따르는 자료의 분위수 회귀모형을 추정하고자 한다. 분위수 회귀모형 추정을 위해 SVQR-IRWLS 분위수 회귀함수추정 기법을 이용한 DPKM을 적용하고자 한다. 본 논문은 다음과 같이 구성되어 있다. 2절에서는 분위수 회귀모형을 3절에서는 DPKM과 본 논문에서 제안한 분위수 회귀모형이 정리되어 있으며, 4절에서는 본 논문에서 제안한 분위수 회귀모형을 적용한 모의실험 결과가 설명된다. 마지막 절에서는 본 논문에 대한 결론으로 마무리하고 있다.

2. 선형 분위수 회귀

분위수의 개념을 회귀식에 적용하기 위해 Koenker와 Bassett (1978)에 의해 제안된 위치모형 (location model)은 다음과 같다.

$$y_i = \beta + \epsilon_i$$

여기서 y_i는 연속확률분포함수 F를 갖는 확률변수이며 β 는 중앙값이다. θ-번째 표본분위수 β_θ는 다음 식을 최소화하여 해를 구한다.

$$\min \sum_{i=1}^{n} \rho_\theta(y_i - \beta_\theta), \beta_\theta \in R \tag{2.1}$$

여기서 $\rho(\cdot)$은 점검함수 (check function)이며, $\rho_\theta(r) = \theta r I_{(r \geq 0)} + (1-\theta) r I_{(r<0)}$이다. 이 표본추정량의 개념을 선형회귀모형 (linear regression model)으로 표현하면 다음과 같다.

$$y_i = \boldsymbol{x}_i^{'} \boldsymbol{\beta} + \epsilon_i$$

여기서 ϵ_i은 0에 대해 대칭인 연속확률분포함수 F를 갖는 확률변수이며 $\boldsymbol{\beta}$는 $p \times 1$모수벡터이다. 일반적으로 선형회귀모형에 대한 최소제곱 추정량은 x가 주어질 때 y의 조건부 평균을 $p(y|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$로 정의하며 일반적인 최소제곱 추정치 $\widehat{\boldsymbol{\beta}}$는 손실함수를 최소화시켜 해를 구한다. 유사한 방법으로 식 (2.1)에서 구한 θ-번째 표본분위수의 개념을 선형회귀모형에 확장하여 $\boldsymbol{x}$가 주어질 때 y의 θ-번째 조

건부 분위수함수 (conditional quantile function)를 다음과 같이 정의할 수 있다.

$$Q_y(\theta|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}_\theta$$

회귀 분위수 추정치 (regression quantile estimate) $\widehat{\boldsymbol{\beta}}_\theta$는 식 (2.1)의 β대신 $x_i^{'}\beta$로 대체하여 다음의 최소화 문제를 만족하는 해로부터 구해진다.

$$\min \sum_{i=1}^{n} \rho_\theta(y_i - \boldsymbol{x}_i^{'}\boldsymbol{\beta}_\theta), \quad \boldsymbol{\beta}_\theta \in R^p$$

3. 비선형 분위수 회귀

3.1. 등분산 비선형 분위수 회귀

주어진 자료가 $(\boldsymbol{x}_i, y_i)$, $i = 1, \cdots, n$로 구성된 훈련 집합을 D라고 할 때, 입력변수 $\boldsymbol{x}_i$ ($\boldsymbol{x}_i \in R^d$)은 상수 1을 포함하며, 반응변수 y_i ($y_i \in R$)는 입력변수 $\boldsymbol{x}_i$와 연관이 있다. 여기서 비선형 특징사상함수 $\phi(\cdot)$: $R^d \to R^{d_f}$는 입력 공간을 고차원 특징 공간으로 사상시킨다. 특징 공간에서의 사상함수의 내적은 입력 공간에서 커널 $\phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_i) = k(\boldsymbol{x}_i, \boldsymbol{x}_i)$을 사용한 것과 동일하다 (Mercer, 1909). 점검함수 $\rho_\theta(\cdot)$을 가진 θ-번째 분위수 회귀함수의 추정값 $q_\theta(\boldsymbol{x}_i)$은 다음 식과 같이 최적화함으로 정의할 수 있다.

$$\min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n} \rho_\theta(y_i - q_\theta(\boldsymbol{x}_i)).$$

여기서 $\rho_\theta(r) = \theta r I_{(r\geq 0)} + (1-\theta) r I_{(r<0)}$이다. 서포트벡터 분위수 회귀 (support vector quantile regression)형식으로 회귀문제를 표현하면 다음과 같다.

$$\min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n} (\theta\xi_i + (1-\theta)\xi_i^*)$$

제약 조건은 다음과 같다.

$$\begin{aligned} y_i - \boldsymbol{w}'\phi(\boldsymbol{x}_i) &\leq \xi_i \\ \boldsymbol{w}'\phi(\boldsymbol{x}_i) - y_i &\leq \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned}$$

여기서 C는 훈련 오차를 벌칙화하는 정규화 모수이다. 라그랑즈 함수 (Lagrange function)로 표현하면 다음과 같다.

$$\begin{aligned} L =& \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n}(\theta\xi_i + (1-\theta)\xi_i^*) - \sum_{i=1}^{n}\alpha_i(\xi_i - y_i + \boldsymbol{w}'\phi(\boldsymbol{x}_i)) \\ & - \sum_{i=1}^{n}\alpha_i^*(\xi_i^* + y_i - \boldsymbol{w}'\phi(\boldsymbol{x}_i)) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*). \end{aligned} \tag{3.1}$$

여기서 $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$을 만족해야 한다. 식 (3.1)을 원변수 (primal variables) $(\boldsymbol{w}, \xi_i, \xi_i^*)$로 편미분한 후에 식 (3.1)에 다시 적용하여 최적화시킨 식은 다음과 같다.

$$\max -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i - e\sum_{i}^{n}(\alpha_i + \alpha_i^*)$$

제약 조건은 다음과 같다.

$$\alpha_i \in [0, \theta C], \alpha_i^* \in [0, (1-\theta)C].$$

제약조건을 가진 위의 식을 풀면 최적의 라그랑즈 배수 α_i, α_i^*을 구할 수 있다. 입력변수 $\boldsymbol{x}$가 주어졌을 때 θ-번째 분위수 회귀함수의 추정량은 다음과 같다.

$$\widehat{q}_\theta(\boldsymbol{x}) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\boldsymbol{x}_i, \boldsymbol{x}).$$

3.2. 이분산 비선형 분위수 회귀

3.2절에서는 분산을 고려한 비모수 분위수 회귀모형을 추정하고자 한다. 비모수 위치-척도 모형(location-scale model)은 일반적으로 다음 식과 같다.

$$y_i = \mu(\boldsymbol{x}_i) + \sigma(\boldsymbol{x}_i)\epsilon_i, \tag{3.2}$$

여기서 $\boldsymbol{x}_i$, $i = 1, \cdots, n$은 d차원의 공변량 벡터이고, ϵ_i은 관측되지 않은 확률변수이며, 평균이 0이고 분산이 1인 어떤 분포 F_o로부터 i.i.d.라고 가정하며, $\boldsymbol{x}_i$와 독립적으로 분포된다. $\boldsymbol{x}_i$안에 첫 번째 원소로 모형의 절편을 포함하기 위해 1을 포함한다. $\mu(\boldsymbol{x})$와 $\sigma(\boldsymbol{x})$는 알려져 있지 않다. 위치-척도 모형에 관한 연구는 통계학과 경제학 관련 논문에서 많이 다루고 있으며, He (1997), Heagerty와 Pepe (1999), Koenker (2005), Shim 등 (2009b)와 Shim 등 (2009c) 등에 잘 안내되어 있다. 대부분의 비모수적인 방법들은 관측치 벡터 $(\boldsymbol{x}_i, y_i)$의 임의 표본으로부터 $\mu(\boldsymbol{x})$를 추정하는 것에 초점을 맞추고 있다. 본 논문에서는 $\mu(\boldsymbol{x})$와 $\sigma(\boldsymbol{x})$를 SVM기법을 활용한 DPKM을 활용하여 동시에 추정하고자 한다. 위치-척도 모형에 대해 반응변수 y_i의 θ-번째 조건부 분위수 회귀함수는 다음과 같다.

$$q_\theta(\boldsymbol{x}_i) = \widehat{\mu}(\boldsymbol{x}_i) + \widehat{F}_0^{-1}(\theta)\widehat{\sigma}(\boldsymbol{x}_i), \tag{3.3}$$

여기서 $\widehat{\mu}(\boldsymbol{x}_i)$와 $\widehat{\sigma}(\boldsymbol{x}_i)$는 각각 $\mu(\boldsymbol{x}_i)$와 $\sigma(\boldsymbol{x}_i)$의 추정치이며, $\widehat{F}_0$은 잔차분포함수의 추정치이다. 즉 잔차 $r(\boldsymbol{x}_i) = (y_i - \widehat{\mu}(\boldsymbol{x}_i))/\widehat{\sigma}(\boldsymbol{x}_i)$의 분포함수이다. 본 논문에서는 $\mu(\boldsymbol{x})$와 $\sigma(\boldsymbol{x})$를 동시에 추정하기 위해 DPKM을 이용한다. 일반적인 분위수 회귀모형을 추정할 때는 자료가 정규분포를 따른다는 가정 하에서 추정하게 된다. 그러나, 실증 분석 자료들을 보면 정규분포에 비해 꼬리부분이 좀 더 두꺼운 형태, 즉 라플라스 분포로 분포되어 있는 경우들을 접하게 된다. 본 논문에서는 y_i가 비대칭 라플라스 분포 (θ, $\mu(\boldsymbol{x}_i)$, $\sigma_i/\sqrt{2}$)를 따른다고 가정하여 이분산 분위수 회귀함수를 추정하고자 한다. 여기서 $\mu(\boldsymbol{x}_i)$는 대칭 라플라스 분포 ($\mu_0(\boldsymbol{x}_i)$, $\sigma_i/\sqrt{2}$)를 가정한 경우의 θ-번째 분위수 회귀함수와 동일하다.

비대칭 라플라스 분포함수는 다음과 같다.

$$p(y_i|\boldsymbol{x}_i) = \frac{\sqrt{2}}{\sigma}\theta(1-\theta)e^{-\frac{\sqrt{2}}{\sigma}\rho_\theta(y_i - \mu(\boldsymbol{x}_i))}$$

식 (3.2)의 우도함수는 다음과 같다.

$$L(\mu, \sigma) = \sum_{i=1}^{n} \left(\frac{\sqrt{2}\rho_\theta(y_i - \mu(\boldsymbol{x}_i))}{\sigma(\boldsymbol{x}_i)} + \log(\sigma(\boldsymbol{x}_i)) \right)$$

여기서 $\mu(\boldsymbol{x}_i)$와 $\sigma(\boldsymbol{x}_i)$는 커널기계를 이용하여 추정한다. θ-번째 분위수 회귀함수는 선형모형, $\mu(\boldsymbol{x}) = \boldsymbol{w}_u'\boldsymbol{\phi}_u(\boldsymbol{x})$에 의해 추정되며 고차원 특징 공간 F_μ ($\boldsymbol{\phi}_\mu : R^d \to F_\mu$)에서 재구성된다. F_μ는 Mercer (1909)에 의해 커널 $K_\mu(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\phi}_\mu(\boldsymbol{x}_i)'\boldsymbol{\phi}_\mu(\boldsymbol{x}_j)$ ($K_\mu : R^d \times R^d \to R$)로 정의한다. 표준

편차는 $\log\sigma(\boldsymbol{x}) = \boldsymbol{w}_g^{'}\boldsymbol{\phi}_g(\boldsymbol{x})$에 의해 추정되며 Mercer (1909)에 정의된 커널 K_σ에 의해 특징 공간 F_σ에 구성된다. 커널로 다시 표현하면 각각 $\mu(\boldsymbol{x}_i) = \sum_{i=1}^n \boldsymbol{K}_\mu(\boldsymbol{x}_i,\boldsymbol{x})\boldsymbol{\alpha}_{\mu i}$ 와 $g(\boldsymbol{x}_i) = \log\sigma = \sum_{i=1}^n \boldsymbol{K}_g(\boldsymbol{x}_i,\boldsymbol{x})\boldsymbol{\alpha}_{gi}$ $(\sigma = e^g)$이다. 본 논문에서는 가우시안 커널 함수를 사용하였다.

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp\left(-\frac{1}{\gamma^2}||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right) \tag{3.4}$$

여기서 γ는 커널 모수이다. 벌칙 음-로그 함수 (penalized negative log likelihood)는 다음과 같다.

$$L(\boldsymbol{\alpha}_\mu,\boldsymbol{\alpha}_g) = \sum_{i=1}^n \left(\sqrt{2}\rho_\theta(y_i - \mu(\boldsymbol{x}_i))e^{-g(\boldsymbol{x}_i)} + g(\boldsymbol{x}_i)\right) + \frac{\lambda_\mu}{2}||\boldsymbol{w}_\mu||^2 + \frac{\lambda_g}{2}||\boldsymbol{w}_g||^2 \tag{3.5}$$

여기서 λ_μ와 λ_g는 벌칙 모수이다. θ-번째 분위수 회귀함수와 표준편차의 모형을 위해 편향-분산 (bias-variance)을 독립적으로 조절한다. $\mu(\boldsymbol{x}_i) = \sum_{i=1}^n \boldsymbol{K}_\mu(\boldsymbol{x}_i,\boldsymbol{x})\boldsymbol{\alpha}_{\mu i}$와 $g(\boldsymbol{x}_i) = \log\sigma = \sum_{i=1}^n \boldsymbol{K}_g(\boldsymbol{x}_i,\boldsymbol{x})\boldsymbol{\alpha}_{gi}$로 표현할 수 있으며 $\boldsymbol{w}_\mu$와 $\boldsymbol{w}_g$는 식 (3.5)을 최소화함으로 얻어진다. Kimeldorf와 Wahba (1971)에 따라 $\boldsymbol{w}_\mu$와 $\boldsymbol{w}_g$를 확장하면 다음 식과 같다.

$$L(\alpha_\mu,\alpha_g) = \sum_i \boldsymbol{K}_{gi}\boldsymbol{\alpha}_g + \sum_i \sqrt{2}\rho_\theta(y_i - \boldsymbol{K}_{\mu i}\boldsymbol{\alpha}_\mu)e^{-\boldsymbol{K}_{gi}\boldsymbol{\alpha}_g} + \frac{\lambda_\mu}{2}\boldsymbol{\alpha}_\mu^{'}\boldsymbol{K}_\mu\boldsymbol{\alpha}_\mu + \frac{\lambda_g}{2}\boldsymbol{\alpha}_g^{'}\boldsymbol{K}_g\boldsymbol{\alpha}_g \tag{3.6}$$

여기서 $\boldsymbol{K}_\mu$와 $\boldsymbol{K}_g$는 각각 (i,j)번째 원소 $\boldsymbol{K}_\mu(\boldsymbol{x}_i,\boldsymbol{x}_j)$와 $\boldsymbol{K}_g(\boldsymbol{x}_i,\boldsymbol{x}_j)$를 가지는 $n\times n$행렬로 구성되어 있다. $\boldsymbol{\alpha}_\mu$와 $\boldsymbol{\alpha}_g$는 각각 i번째 원소가 $\alpha_{\mu i}$와 α_{gi}인 $n\times 1$벡터이다. $\boldsymbol{\alpha}_\mu$와 $\boldsymbol{\alpha}_g$는 식 (3.6)을 최소화하여 구한다.

θ -번째 분위수 회귀함수와 표준편차 모형의 모수 $\boldsymbol{\alpha}_\mu$와 $\boldsymbol{\alpha}_g$는 SVM과 단순 뉴톤-랩손 반복 처리를 통해서 구할 수 있다. 먼저 $\boldsymbol{\alpha}_g$ $(g=0,\,\sigma=1)$를 고정시킨 후 θ-번째 분위수 회귀함수 모형의 최적의 모수 $\boldsymbol{\alpha}_\mu$는 식 (3.6)을 모수 $\boldsymbol{\alpha}_\mu$로 미분시켜 최소화하여 구한다. 식은 다음과 같다.

$$L(\boldsymbol{\alpha}_\mu) = \sqrt{2}\sum_i \rho_\theta(y_i - \boldsymbol{K}_{\mu i}\boldsymbol{\alpha}_\mu)e^{-g_i} + \frac{\lambda_\mu}{2}\boldsymbol{\alpha}_\mu^{'}\boldsymbol{K}_\mu\boldsymbol{\alpha}_\mu \tag{3.7}$$

$\boldsymbol{\alpha}_\mu$를 추정한 후 $\mu(\boldsymbol{x}_i) = \boldsymbol{K}_{\mu i}\boldsymbol{\alpha}_\mu$를 추정할 수 있다. 다음으로 $r_i = \rho_\theta(y_i - \boldsymbol{K}_{\mu i}\boldsymbol{\alpha}_\mu)$를 계산하여 고정시킨 후 식 (3.6)을 모수 $\boldsymbol{\alpha}_g$로 미분시켜 최소화하여 모수 $\boldsymbol{\alpha}_g$를 구한다. 식은 다음과 같다.

$$L(\boldsymbol{\alpha}_g) = \sqrt{2}\sum_i \boldsymbol{K}_{gi}\boldsymbol{\alpha}_g + r_i e^{-\boldsymbol{K}_{gi}\boldsymbol{\alpha}_g} + \frac{\lambda_g}{2}\boldsymbol{\alpha}_g^{'}\boldsymbol{K}_g\boldsymbol{\alpha}_g \tag{3.8}$$

$\boldsymbol{\alpha}_g$를 추정한 후 $g(\boldsymbol{x}_i) = \boldsymbol{K}_{gi}\boldsymbol{\alpha}_g$를 추정할 수 있다. 커널함수의 최적 커널모수와 벌칙모수 (γ_μ, λ_μ, γ_g, λ_g)는 Xiang와 Wahba (1996)가 개발한 GACV (generalized approximate cross validation)를 이용하여 최적의 모수를 선택한다. 제안한 기법의 진행과정은 표 3.1과 같다.

4. 모의실험

이번 절에서는 2종류의 모의실험을 통해 본 논문에서 제안한 분위수 회귀모형의 성능을 평가하고자 한다. Shim 등 (2009a)에서 사용한 모의실험과 Sohn 등 (2008)에서 사용된 실험을 이용하여 분석하였다. 첫 번째 실험을 위해 준비한 자료는 다음과 같다. 입력변수 x_i는 $U(0,2)$로부터 150개의 자료를 난수 발생시켰으며, 반응변수 y_i는 $y_i = \mu(x_i) + \sigma(x_i)\epsilon_i$과 같다. 여기서 $\mu(x_i) = \sin(2\pi x_i)$, $\sigma(x_i) =$

표 3.1 제안한 분위수 회귀함수 추정절차

[1단계] :	$\boldsymbol{\alpha}_g^{(0)} = 0$을 초기값으로 지정
[2단계] :	(i) $\boldsymbol{\alpha}_g^{(t)}$값이 주어진 상태에서 GACV를 이용하여 γ_μ와 λ_μ를 선택 (ii) $\boldsymbol{\alpha}_g^{(t)}$, γ_μ와 λ_μ이 주어진 상태에서 $\boldsymbol{\alpha}_\mu$에 대해 식 (3.7)을 최소화하여 $\boldsymbol{\alpha}_\mu^{(t+1)}$을 업데이트
[3단계] :	(i) $\boldsymbol{\alpha}_\mu^{(t+1)}$값이 주어진 상태에서 GACV를 이용하여 γ_g와 λ_g를 선택 (ii) $\boldsymbol{\alpha}_\mu^{(t+1)}$, γ_g와 λ_g이 주어진 상태에서 $\boldsymbol{\alpha}_g$에 대해 식 (3.8)을 최소화 하여 $\boldsymbol{\alpha}_g^{(t+1)}$을 업데이트
[4단계] :	$\frac{1}{n}\|\|\boldsymbol{\alpha}_\mu^{(t)} - \boldsymbol{\alpha}_\mu^{(t+1)}\|\|^2 < \epsilon$와 $\frac{1}{n}\|\|\boldsymbol{\alpha}_g^{(t)} - \boldsymbol{\alpha}_g^{(t+1)}\|\|^2 < \epsilon$가 동시에 만족될 때까지 [2단계]와 [3단계] 반복 ($\epsilon = 10^{-6}$)
[5단계] :	잔차의 분포 함수 F_o을 추정
[6단계] :	각 분위수 θ에 따른 조건부 분위수 회귀함수 식 (3.3) 추정

$\sqrt{(2.1 - x_i)/4}$, ϵ_i는 $\chi^2_{(2)} - 2$ 이다. $\chi^2_{(2)}$는 자유도가 2인 카이제곱분포이다. 실험의 절차는 표 3.1과 같으며, 함수 추정을 위해 식 (3.4)의 가우시안 커널함수를 사용하였다. 분위수 θ는 (0.1, 0.5, 0.9)로 지정한 후 분위수 회귀모형을 추정하였다. 실험 결과는 그림 4.1과 같다. 제안한 분위수 회귀모형의 커널모수와 벌칙모수 (λ_μ, γ_μ, λ_g, γ_g)는 GACV를 통해 구하였으며 결과는 표 4.1와 같다. 또한 SVQR-IRWLS와 제안한 분위수 회귀모형의 MSE은 표 4.1와 같다. 그림 4.1에서 실선은 추정하고자 하는 목표값이며, 점선은 SVQR-IRWLS로 추정한 결과이며, 절취선은 본 논문에서 제안한 분위수 회귀모형에 의해 추정된 결과이다. 그림 4.1의 결과를 통해 본 논문에서 제안한 분위수 회귀모형이 SVQR-IRWLS 모형보다 목표값에 더 근사한 것을 알 수 있으며, 또한 표 4.1의 MSE에 대해서도 제안한 분위수 회귀모형이 더 작은 수치를 가지고 있음을 확인할 수 있었다. 본 논문에서 제안한 분위수 회귀모형이 SVQR-IRWLS보다 목표값에 더 근사하며 함수추정을 잘 한다고 할 수 있다.

표 4.1 커널 및 벌칙모수와 MSE

	λ_μ	γ_μ	λ_g	γ_g	제안한 모형MSE	SVQR-IRWLS MSE
θ =0.1	0.1	0.5	1	2	0.0046	0.0073
θ =0.5	0.01	0.5	0.01	1	0.0183	0.0195
θ =0.9	0.01	0.5	1	2	0.1345	0.1881

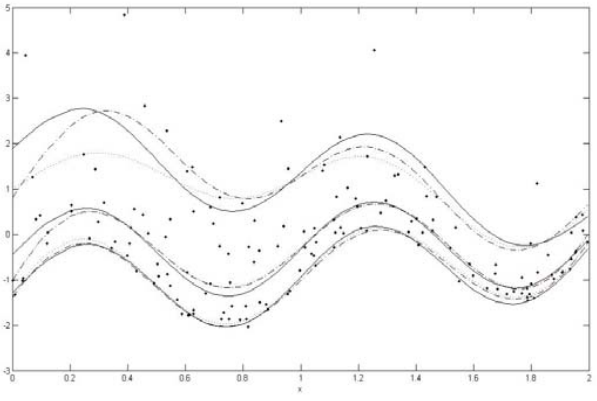

그림 4.1 첫 번째 모의실험 자료를 이용한 함수 추정

두 번째 실험을 위해 준비한 자료는 다음과 같다. 입력변수 x_i는 $U(0,1)$로부터 200개의 자료를 난수 발생시켰으며, 반응변수 y_i는 $y_i = \mu(x_i) + \sigma(x_i)\epsilon_i$과 같다. 여기서 $\mu(x_i) = 2\sin(2\pi x_i)$, $\sigma(x_i) = e^{(x_i+0.5)}$, $\epsilon_i = \sigma(x_i)/\sqrt{2}$이다. 실험의 절차는 표 3.1과 같으며 함수 추정을 위해 식 (3.4)의 가우시안 커널함수를 사용하였다. 분위수 θ는 (0.1, 0.5, 0.9)로 지정한 후 분위수 회귀모형을 추정하였다. 실험 결과는 그림 4.2와 같다. 그림 4.2에서 실선은 추정하고자 하는 목표값이며, 점선은 SVQR-IRWLS로 추정한 결과이며, 절취선은 본 논문에서 제안한 분위수 회귀모형에 의해 추정된 결과이다. 그림 4.2의 왼쪽그림은 등분산이라 가정할 경우 SVQR-IRWLS와 제안한 분위수 회귀모형을 추정한 결과로 거의 유사함을 확인할 수 있다. 또한 그림 4.2의 오른쪽 그림을 통해 알 수 있듯이 본 논문에서 제안한 분위수 회귀모형이 SVQR-IRWLS보다 목표값에 더 근사함을 확인할 수 있다.

표 4.2 커널 및 벌칙모수와 MSE

	λ_μ	γ_μ	λ_g	γ_g	제안한 모형MSE	SVQR-IRWLSMSE
θ =0.1	0.01	0.5	0.01	1	0.4683	0.8933
θ =0.5	0.01	0.5	0.01	1	0.1080	0.1490
θ =0.9	0.01	0.5	1	4	0.6401	1.1262

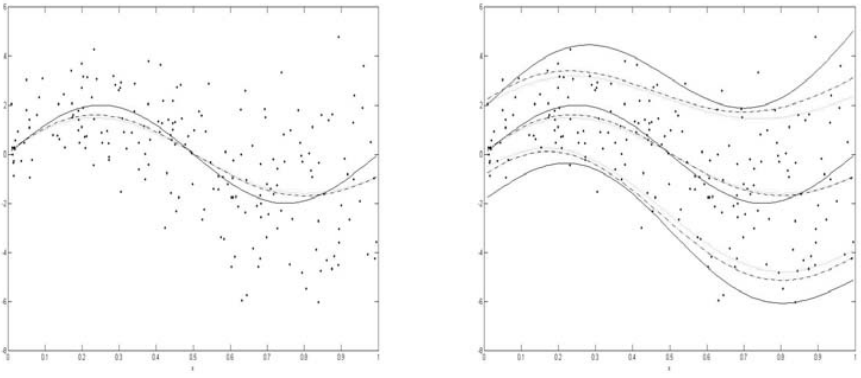

그림 4.2 두 번째 모의실험 자료를 이용한 함수 추정 (왼쪽: 등분산, 오른쪽: 이분산)

5. 결론

본 논문에서는 분위수 회귀모형 추정을 위해 비대칭 라플라스 분포를 이용한 SVQR-IRWLS 기법의 DPKM모형을 제안하였다. 평균함수와 분산함수를 동시에 추정할 수 있는 장점을 가진 모형으로 모의실험을 통해 성능을 평가하였다. 분위수 회귀모형 추정에 로버스트하다고 알려져 있는 SVQR-IRWLS 기법과 성능을 비교하였으며 그림 4.1과 그림 4.2를 통해 그 결과를 확인할 수 있었다. 그림 4.1과 그림 4.2의 결과를 통해 본 논문에서 제안한 분위수 회귀모형이 SVQR-IRWLS보다 목표값에 더 근사하게 잘 추정하고 있음을 확인할 수 있었다. 표 4.1의 MSE값 비교를 통해 본 논문에서 제안한 SVQR-IRWLS 기법의 DPKM모형이 기존에 알려져 있는 SVQR-IRWLS보다 더 작은 값을 가짐을 확인 할 수 있었다. 실험 결과를 통해 본 논문에서 제안한 모형이 분위수 회귀모형 추정에 좋은 결과를 보여줌을 알 수 있었다.

참고문헌

Basset, G. and Koenker, R. (1982). An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association,* **77**, 407-415.

He, X. (1997). Quantile curves without crossing. *The American Statistician,* **51**, 186-192.

Heagerty, P. J. and Pepe, M. S. (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Applied Statistics*, **48**, 533-551.

Hwang, C. and Shim, J. (2005). A simple quantile regression via support vector machine. *Lecture Notes in Computer Science*, **3610**, 512-520.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and its Applications*, **33**, 82-95.

Koenker, R. (2005). *Quantile regression,* Cambridge University Press, London.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, 33-50.

Mercer, J. (1909). Functions of positive and negative and their connection with the theory of integral equations. *Philosphical Transactions of the Royal Society*, **A**, 415-446.

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics*, **32**, 143-155.

Shim, J., Hwang, C. and Seok, K. (2009). Non-crossing quantile regression via doubly penalized kernel machine. *Computational Statistics*, **24**, 83-94.

Shim, J., Park, H. and Hwang, C. (2009). A kernel machine for estimation of mean and volatility functions. *Journal of the Korean Data & Information Science Society*, **20**, 905-912.

Shim, J., Park, H. and Seok, K. (2009). Variance function estimation with LS-SVM for replicated data. *Journal of the Korean Data & Information Science Society*, **20**, 925-931.

Smola, A. J. and Scholkopf, B. (1998). *A tutorial on support vector regression*. NeuroCOLT2 Technical Report, NeuroCOLT.

Sohn, I., Kim, S., Hwang, C., Lee, J. W. and Shim, J. (2008). Support vector machine quantile regression for detecting differentially expressed genes in microarray analysis. *Methods of Information in Medicine*, **47**, 459-467.

Vapnik, V. N. (1998). *Statistical learning theory*, Springer.

Weiss, A. (1991). Estimating nonlinear dynamic models using least absolute error estimation. *Econometric Theory*, **7**, 46-68.

White, H. (1992). *Nonparametric estimation of conditional quantile using neural networks*, in H. White, eds., Artificial Neural Networks: Approximation and Learning Theory, Blackwell, Oxford, 191-205.

Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, **6**, 675-692.

Yuan, M. and Wahba, G. (2004). Doubly penalized likelihood estimator in heteroscedastic regression. *Statistics and Probability Letter*, **69**, 11-20.

Journal of the Korean Data & Information Science Society
2009, **20**(6), 1093–1101

한국데이터정보과학회지

Quantile regression using asymmetric Laplace distribution

Hye Jung Park[1]

College of General Education, Daegu University

Received 1 September 2009, revised 7 November 2009, accepted 21 November 2009

Abstract

Quantile regression has become a more widely used technique to describe the distribution of a response variable given a set of explanatory variables. This paper proposes a novel modelfor quantile regression using doubly penalized kernel machine with support vector machine iteratively reweighted least squares (SVM-IRWLS). To make inference about the shape of a population distribution, the widely popularregression, would be inadequate, if the distribution is not approximately Gaussian. We present a likelihood-based approach to the estimation of the regression quantiles that uses the asymmetric Laplace density.

Keywords: Asymmetric Laplace distribution, doubly penalized kernel machine, generalized approximate cross validation, quantile regression, support vector machine.

[1] Invitation Professor, College of General Education, Daegu University, Gyeongbuk 712-714, Korea. E-mail: hyjpark@daegu.ac.kr