The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.5.569
33(5), 569–578

Divide and conquer kernel quantile regression for massive dataset

Sungwan Bang[a] · Jaeoh Kim[b,1]

[a]Department of Mathematics, Korea Military Academy;
[b]Center for Army Analysis and Simulation, HQs ROKA

(Received June 15, 2020; Revised July 23, 2020; Accepted July 27, 2020)

Abstract

By estimating conditional quantile functions of the response, quantile regression (QR) can provide comprehensive information of the relationship between the response and the predictors. In addition, kernel quantile regression (KQR) estimates a nonlinear conditional quantile function in reproducing kernel Hilbert spaces generated by a positive definite kernel function. However, it is infeasible to use the KQR in analysing a massive data due to the limitations of computer primary memory. We propose a divide and conquer based KQR (DC-KQR) method to overcome such a limitation. The proposed DC-KQR divides the entire data into a few subsets, then applies the KQR onto each subsets and derives a final estimator by aggregating all results from subsets. Simulation studies are presented to demonstrate the satisfactory performance of the proposed method.

Keywords: divide and conquer, kernel, quadratic programming, quantile regression

1. 서론

전통적인 회귀분석은 p차원의 설명변수 $\boldsymbol{x} \in R^p$가 주어졌을 때 반응변수 $Y \in R$의 조건부 평균 함수(conditional mean function)를 추정한다. 반면에 Koenker와 Bassett (1978)에 의해 제안된 분위수 회귀모형(quantile regression)은 최소절대추정법(least absolute estimation)을 일반화한 것으로서 반응변수의 조건부 분위수 함수(conditional quantile function)을 추정함으로써 반응변수의 조건부 분포에 대한 포괄적인 정보를 제공하는 이점을 지니고 있다. 분위수 회귀모형은 회귀계수 추정의 강건성과 유용성을 바탕으로 의학 (Cole과 Green, 1992; Heagerty와 Pepe, 1999), 경제 (Koenker과 Hallock 2001; Powell과 Wagner, 2014), 생존분석 (Koenker과 Geling, 2001; Bang 등, 2016), 마이크로어레이 연구 (Wang과 He, 2007; Yang과 Liu, 2016) 등 여러 다양한 분야에 적용되고 있다.

반응변수 Y에 대한 $100\tau\%$ 조건부 분위수 함수 $q_\tau(\boldsymbol{x})$는

$$P(Y \le q_\tau(\boldsymbol{x})|X = \boldsymbol{x}) = \tau, \quad \text{단 } 0 < \tau < 1 \tag{1.1}$$

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NO. 2020R1F1A1A01065107).
[1]Corresponding author: Center for Army Analysis and Simulation, HQs ROKA, 663, Gyeryongdae-ro, Sindoan-myeon, Gyeryong-si, Chungcheongnam-do 32800, Korea. E-mail: c14180@gmail.com

와 같이 정의 되며, 개체수가 N인 훈련자료 $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$가 주어졌을 때 Koenker과 Bassett (1978)은 체크(check) 손실함수 $\rho_\tau(t) = t(\tau - I(t < 0))$를 이용한 분위수 회귀모형의 적합식

$$\min_{q_\tau} \sum_{i=1}^{N} \rho_\tau(y_i - q_\tau(\boldsymbol{x}_i)) \tag{1.2}$$

을 통해 조건부 분위수 함수 $q_\tau(\boldsymbol{x})$를 추정하였다. 일반적으로 모형의 간결성, 계산의 편의성 등의 이점으로 선형 분위수 회귀모형에 대한 많은 연구가 진행되었다 (Li와 Zhu, 2008; Wu와 Liu, 2009; Bang과 Shin, 2016). 그러나 분위수 회귀모형의 추정에서 반응변수와 설명변수가 비선형 관계를 갖는 자료에 대해서는 선형 관계식을 가정하는 전통적 방법론의 사용은 제한된다. 또한 분석자료의 크기에 비해 설명변수의 수가 상대적으로 큰 경우 손실함수만을 고려한 추정법은 다중공선성(multi-collinearity) 및 과대적합(over-fitting)의 문제를 초래한다. 이러한 문제점을 해결하기 위하여 Li 등 (2007)은 반응변수와 설명변수의 비선형 관계식을 고려하기 위하여 양정치(positive definite) 커널함수(kernel function) $K : R^p \times R^p \to R$에 의해 만들어지는 재생 커널 힐버트 공간(reproducing kernel Hilbert space; RKHS)에서 분위수 함수 $q_\tau(\boldsymbol{x})$를 추정하는 커널 분위수 회귀모형(kernel quantile regression; KQR)을 제안하였다. 커널함수를 이용한 비모수 추정법은 비선형 함수 추정의 대표적인 기법으로 서포터 벡터 머신 (Vapnik, 1998), 커널 K-평균 군집화 (Dhillon 등, 2004) 등 여러 다양한 통계 분석에서 활용되고 있다.

최근에는 네트워크 등 과학기술의 발전으로 다양한 분야에서 새로운 유형의 대용량 자료(massive data)가 생성, 수집 및 저장되고 있으며, 이러한 대용량 자료의 출현은 많은 통계적 분석 방법론의 직접적인 적용을 불가능하게 하고 있다. 특히 KQR은 커널함수를 이용하여 비선형 분위수 함수를 효과적으로 추정하나, 그 계산 알고리즘이 이차계획법(quadratic programming; QP)으로 공식화 되므로 많은 계산 비용으로 인하여 대용량 자료의 분석에는 그 사용이 제한된다. 이러한 대용량 자료의 분석을 위하여 전체 자료를 분할(divide)한 후 분할된 자료의 추정 결과를 통합(conquer)하는 분할정복 알고리즘(divide and conquer algorithm; DC)이 최근 전통적인 선형 분위수 회귀모형에서 활용되어 추정의 계산 효율을 증대시키는 방법론이 연구되고 있다 (Chen 등, 2018; Jiang 등, 2018; Xu 등, 2020; Chen과 Zhou, 2020). 따라서 본 논문에서는 대용량 자료의 분석을 위하여 분할정복 알고리즘을 활용한 커널 분위수 회귀모형(DC-KQR)과 그 계산 알고리즘을 제안한다.

논문의 구성은 다음과 같다. 2절에서는 커널 분위수 회귀모형(KQR)에 대하여 간략히 소개하고 계산의 효율을 증대시키기 위하여 분할정복 알고리즘을 활용한 커널 분위수 회귀모형(DC-KQR)을 제안하였다. 3절과 4절에서는 모의실험과 실제자료의 분석을 통해 기존의 KQR과 제안한 DC-KQR의 성능 및 특성을 비교하였으며, 제안한 방법론의 활용가능성을 보였다. 마지막으로 5절에서는 결론과 더불어 차후 연구방향을 제시하였다.

2. 대용량 자료의 분석을 위한 분할정복 커널 분위수 회귀모형

본 절에서는 먼저 재생 커널 힐버트 공간(RKHS)에서 비선형 분위수 함수를 추정하는 KQR 추정법과 그 계산 알고리즘을 간략히 소개하고, 분할정복 알고리즘을 활용하여 대용량 자료의 분석에서 계산 효율을 향상시키기 위한 분할정복 커널 분위수 회귀모형(DC-KQR)을 제안하기로 한다.

2.1. 커널 분위수 회귀모형의 소개

조건부 분위수 함수의 추정에서 반응변수와 설명변수의 비선형 관계식을 고려하기 위하여 Li 등

(2007)은 양정치 커널함수 $K : R^p \times R^p \to R$에 의해 만들어지는 재생 커널 힐버트 공간(RKHS)에서 분위수 함수 $q_\tau(\boldsymbol{x})$를 추정하는 커널 분위수 회귀모형(KQR)을 제안하였다. KQR의 적합식은 분위수 회귀모형의 적합식 (1.2)에 재생 커널 힐버트 공간 H_K에서 정의된 분위수 함수의 벌칙항 $||q_\tau||^2_{H_K}$를 추가한 다음의 식 (2.1)과 같다.

$$\min_{q_\tau \in H_K} \sum_{i=1}^{N} \rho_\tau \left(y_i - q_\tau(\boldsymbol{x}_i)\right) + \lambda ||q_\tau||^2_{H_K} \tag{2.1}$$

이때 representer 정리 (Kimeldorf와 Wahba, 1971)를 이용하면 적합식 (2.1)의 해는

$$q_\tau(\boldsymbol{x}) = \sum_{i=1}^{N} \beta_{i,\tau} K\left(\boldsymbol{x}_i, \boldsymbol{x}\right) + b_\tau \tag{2.2}$$

와 같은 형태로 표현되고, 이를 이용하여 KQR의 적합식 (2.1)을

$$\min_{\boldsymbol{\beta}_\tau, b_\tau} \sum_{i=1}^{N} \rho_\tau \left(y_i - \sum_{j=1}^{N} \beta_{j,\tau} K(\boldsymbol{x}_i, \boldsymbol{x}_j) - b_\tau \right) + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} \beta_{i,\tau} \beta_{j,\tau} K(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{2.3}$$

와 같이 다시 표현할 수 있다. 여기서 $\boldsymbol{\beta}_\tau = (\beta_{1,\tau}, \beta_{2,\tau}, \ldots, \beta_{N,\tau})^T \in R^N$이며 $\lambda > 0$는 조율모수(tuning parameter)이다. 커널함수는 사용의 편의성, 계산의 효율성 및 유연성 등과 같은 장점을 바탕으로 비선형 함수의 추정에 많이 활용되고 있으며, 본 논문에서는 일반적으로 많이 사용되는 가우시안 커널(Gaussian kernel) 함수 $K(\boldsymbol{x}, \boldsymbol{x}') = \exp(||\boldsymbol{x} - \boldsymbol{x}'||^2/2\sigma^2)$를 주로 사용하였다.

KQR의 적합식 (2.3)은 이차계획법으로 공식화 될 수 있다. 이를 위하여 제약식 $y_i - \sum_{j=1}^{N} \beta_{j,\tau} K(\boldsymbol{x}_i, \boldsymbol{x}_j) - b_\tau = u_i - v_i$를 만족하는 $2N$개의 잉여변수(slack variable) $\{(u_i, v_i),\ i = 1, \ldots, N\}$를 사용하면 KQR의 계산 알고리즘은 다음의 식 (2.4)–(2.6)과 같이 공식화 된다. 이때 잉여변수는 $u_i \geq 0$, $v_i \geq 0$을 만족하며, $\mathbf{K}$는 (i, j)번째 원소로 $\mathbf{K_{ij}} = \mathbf{K}(\boldsymbol{x_i}, \boldsymbol{x_j})$를 갖는 $N \times N$ 커널행렬(kernel matrix)이다.

$$\min_{\boldsymbol{\beta}_\tau, b_\tau} \sum_{i=1}^{N} \left(\tau u_i + (1-\tau) v_i\right) + \lambda \boldsymbol{\beta}_\tau^T \mathbf{K} \boldsymbol{\beta}_\tau \tag{2.4}$$

subject to

$$y_i - \sum_{j=1}^{N} \beta_{j,\tau} K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) - b_\tau = u_i - v_i, \quad \text{for } i = 1, 2, \ldots, N \tag{2.5}$$

$$u_i \geq 0,\ v_i \geq 0 \quad \text{for } i = 1, 2, \ldots, N. \tag{2.6}$$

2.2. 분할정복 커널 분위수 회귀모형의 제안

최적화 식 (2.4)–(2.6)에서 보는바와 같이 KQR의 적합식은 이차계획법으로 공식화 되며 이는 $O(N^3)$의 계산비용을 필요로 한다. 따라서 대용량 자료의 분석에서 메모리 능력이 제한된 하나의 컴퓨터로 KQR 추정량을 구하기는 불가능하다. 이러한 경우 컴퓨터 과학 분야에서 많이 사용되는 분할정복 알고리즘을 이용하여 컴퓨터의 계산 능력을 초과하는 대용량 자료를 분석할 수 있다. 분할정복 알고리즘은 먼저 전체 자료를 분할한 후 분할된 자료의 결과를 통합(conquer or merge)하는 기법으로 빅데이터 분석에서 널리 사용되고 있으며, 최근에는 대용량 자료의 분석을 위해 전통적인 통계분석 방법론에 활용되고 있다. Fan 등 (2007)은 선형회귀 모형의 최소제곱추정에 분할정복 알고리즘을 적용하였으며, Lin과

Xi (2011)은 대용량 자료에서 추정 방정식(estimating equations)을 이용한 분할정복 추정법을 개발하였다. 또한 분할정복 알고리즘은 일반화 선형모형(generalized linear model)의 추정에 활용되었으며 (Chen과 Xie, 2014), 대용량 자료의 분석을 위한 선형 분위수 회귀모형의 추정에도 적용되어 연구되었다 (Chen 등, 2018; Jiang 등, 2018; Xu 등, 2020; Chen과 Zhou, 2020). 특히 Zhang 등 (2015), Kang과 Jhun (2020)은 분할정복 알고리즘을 활용하여 커널 릿지 회귀모형의 효율적인 계산법을 제안하였다.

분할정복 알고리즘은 컴퓨터의 성능 제한으로 대용량 자료의 분석이 전통적인 통계적 방법론으로 불가능 할 때 계산의 효율을 향상시킬 수 있으며, 특히 KQR과 같이 많은 계산 비용을 요구하는 추정법에 매우 효과적으로 적용될 수 있다. 따라서 본 논문에서는 대용량 자료의 분석을 위하여 분할정복 알고리즘을 활용한 커널 분위수 회귀모형(DC-KQR)을 제안한다. 제안한 DC-KQR은 먼저 N개의 전체 훈련자료를 서로 배반인 동일한 크기($n = N/K$)의 훈련자료로 구성된 K개의 부분집합으로 무작위로 분할한 후, 각각의 부분집합에 대하여 커널 분위수 회귀함수를 추정하고 이들의 산술 평균을 이용하여 최종적인 추정량으로 통합하는 기법으로 구체적인 알고리즘은 다음과 같다.

DC-KQR 알고리즘	
단계 1	크기가 N인 훈련자료 $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$를 서로 배반인 K개의 동일한 크기($n = N/K$)의 부분집합 $S_1, S_2, \ldots, S_K$로 무작위로 분할한다.
단계 2	각각의 $k = 1, 2, \ldots, K$에 대하여 KQR 국소 추정량(local estimator)을 $$\hat{q}_{\tau,k}(\boldsymbol{x}) := \arg\min_{q_\tau \in H_K} \sum_{(\boldsymbol{x},y)\in S_k} \rho_\tau(y_i - q_\tau(\boldsymbol{x}_i)) + \lambda ||q_\tau||^2_{H_K} \quad \text{for} \quad k = 1, 2, \ldots, K$$ 와 같이 계산한다.
단계 3	K개의 KQR 국소 추정량들을 평균하여 DC-KQR 추정량을 $$\bar{q}_\tau(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^K \hat{q}_{\tau,k}(\boldsymbol{x})$$ 와 같이 최종적으로 계산한다.

DC-KQR 알고리즘의 단계 2에서 KQR 국소 추정량의 실질적인 계산은 최적화 식 (2.4)–(2.6)을 이용하여 이루어진다. 따라서 국소 KQR 추정량 $\hat{q}_{\tau,k}(\boldsymbol{x})\,(k = 1, 2, \ldots, K)$은 $n\,(= N/K)$차원의 이차계획법(QP)에 의해 구해지고, 이러한 계산을 K번 반복하므로 DC-KQR 추정량 $\bar{q}_\tau(\boldsymbol{x})$의 계산 비용은 $O(K(N/K)^3) = O(N^3/K^2)$으로 기존의 KQR에 비해 계산이 매우 효율적이다. 또한 단계 2의 계산은 분할정복 알고리즘의 특성상 병렬 분산처리가 가능하므로 다중 프로세서 또는 GPU 기반의 병렬 연산 환경에서 계산할 경우 상당한 계산 속도의 향상을 기대할 수 있다.

3. 모의실험

본 절에서는 제안한 DC-KQR의 분위수 함수의 추정 정확도와 계산 효율성을 기존의 KQR과 비교하기 위하여 모의실험을 시행하였다. 회귀모형으로는 비선형 함수

$$y = 2 + 4\sin(\pi x_1) + 4(x_2 - 0.5)^2 + \epsilon \tag{3.1}$$

을 고려하였으며, 설명변수 x_1과 x_2는 서로 독립적으로 균등분포 $U(-1, 1)$를, 오차항 ϵ은 표준정규분포 $N(0, 1)$를 따르는 것으로 가정하였다. 모형적합(model fitting)은 중위수(median) 함수($\tau = 0.5$)와 90% 분위수 함수($\tau = 0.9$)를 추정하였으며, 이를 위해 훈련자료(training data)의 크기 N은 $N \in \{2^8, 2^9, \ldots, 2^{15}\}$을 고려하였고 훈련자료의 균등분할 수 K는 $K \in \{2^2, 2^3, \ldots, 2^6\}$을 고려하였다.

Table 3.1. Mean absolute errors (MAE) and computation times as a function of number of partitions K and data size N for the simulated example with $\tau = 0.5$

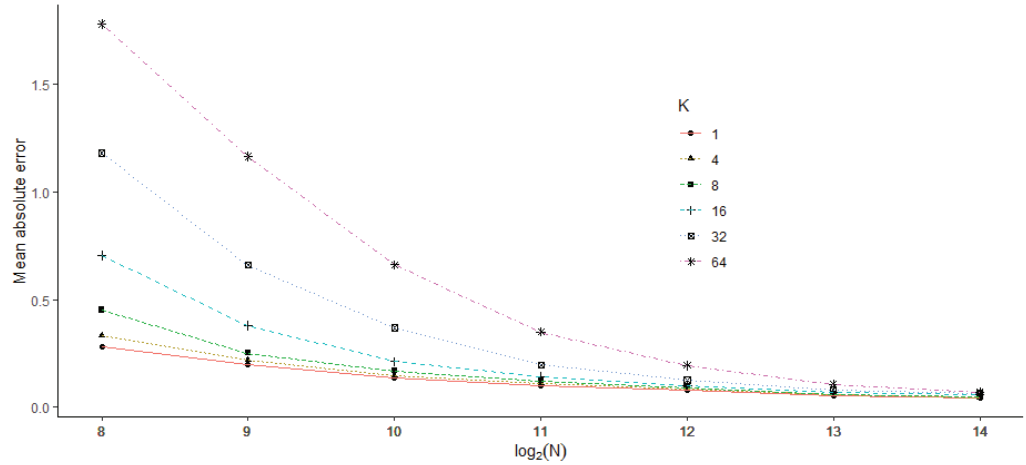
N		K					
		1	2^2	2^3	2^4	2^5	2^6
2^8	MAE	0.281	0.331	0.451	0.705	1.179	1.778
	Time(sec)	0.215	0.046	0.073	0.106	0.193	0.359
2^9	MAE	0.199	0.219	0.252	0.379	0.659	1.163
	Time(sec)	0.509	0.212	0.094	0.143	0.216	0.385
2^{10}	MAE	0.138	0.147	0.169	0.213	0.365	0.662
	Time(sec)	1.253	0.886	0.410	0.194	0.295	0.436
2^{11}	MAE	0.102	0.112	0.122	0.141	0.197	0.350
	Time(sec)	3.549	1.714	1.734	0.803	0.387	0.596
2^{12}	MAE	0.079	0.083	0.092	0.103	0.126	0.194
	Time(sec)	20.412	3.192	3.228	3.389	1.521	0.617
2^{13}	MAE	0.054	0.057	0.061	0.070	0.080	0.108
	Time(sec)	85.412	15.329	6.980	6.783	6.806	3.200
2^{14}	MAE	0.043	0.044	0.047	0.057	0.066	0.071
	Time(sec)	408.453	92.823	29.190	13.715	13.569	13.643

KQR과 제안한 DC-KQR을 이용하여 비선형 함수를 추정하기 위하여 일반적으로 많이 사용되는 가우시안 커널 함수 $K(\boldsymbol{x}, \boldsymbol{x}') = \exp(||\boldsymbol{x} - \boldsymbol{x}'||^2/2\sigma^2)$를 사용하였으며, 커널 모수 σ^2은 훈련자료의 개체들 간의 유클리드 제곱거리 $||\boldsymbol{x} - \boldsymbol{x}'||^2$의 중위수를 이용하여 선택하였다 (Caputo 등, 2002). 또한 모형의 적합에서 조율모수 λ는 훈련자료를 이용한 5-겹 교차타당법(5-fold cross validation)으로 선택하였다. 모형평가(model assessment)를 위해 크기가 10,000인 평가자료(test data)를 독립적으로 생성하였으며, 평가자료를 이용한 평균절대오차(mean absolute error; MAE)

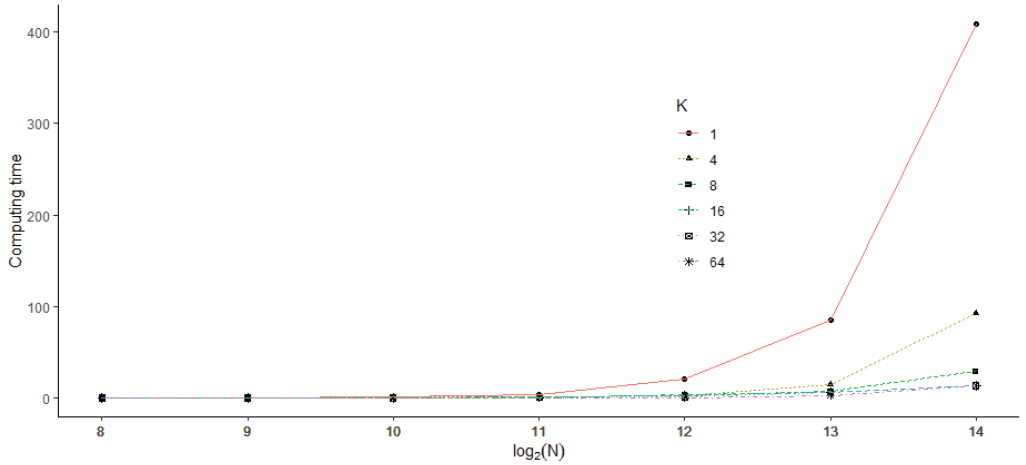
$$\text{MAE} = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} |q_\tau(\boldsymbol{x}_i) - \bar{q}_\tau(\boldsymbol{x}_i)| \tag{3.2}$$

를 계산하였다. 이러한 절차는 20회 독립 반복시행 하였으며, 그 결과는 평균값으로 Table 3.1에 제공하였다. 본 실험은 10개의 코어 및 20개의 스레드(thread)를 가진 2.2GHz의 프로세서와 64GB 메모리를 가진 리눅스 환경에서 R 프로그램을 통해 수행되었으며, 식 (2.4)–(2.6)의 최적화 문제는 R 프로그램의 "kernlab" 패키지 (Karatzoglou 등, 2004)에서 제공하는 kqr( ) 함수를 사용하여 계산하였다. 본 논문에서 사용한 DC-KQR 추정법에 대한 R 코드는 차후 연구에 도움이 되도록 요청 시 제공할 것이다.

Table 3.1은 훈련자료의 크기 N과 훈련자료의 균등분할 수 K에 따른 $\tau = 0.5$에서의 KQR과 DC-KQR의 평균절대오차와 계산시간을 20회 반복의 평균값으로 나타내고 있다. 또한 Figure 3.1의 (a)와 (b)는 각각 훈련자료의 크기 N에 따른 $\tau = 0.5$에서의 KQR과 DC-KQR의 평균절대오차와 계산 속도를 그림으로 비교하고 있다. 여기서 균등분할 수 $K = 1$일 때는 기존의 KQR 추정법을 나타낸다. Table 3.1과 Figure 3.1로부터 균등분할 수 K가 증가함에 따라 DC-KQR의 추정 정확도는 감소하는 반면, DC-KQR의 계산 속도는 현저히 향상되는 것을 알 수 있다. 특히 균등분할 수 $K \leq 2^4$일 때 DC-KQR의 추정 정확도는 다소 안정적인(stable) 경향을 나타내었으며, 훈련자료의 크기 $N \geq 2^{12}$일 때는 균등분할 수 $K = 2^6$에서도 수용 가능한 추정 정확도의 성능을 나타내었다. 또한 예상한 바와 같이 균등 분할된 부분집합의 훈련자료의 수 $n\,(= N/K)$이 동일할 때 균등분할 수 K가 클수록 추정의 정확도가 높게 나타나고 있으며, 이러한 결과는 데이터 스트림(stream)의 형태로 수집되는 대용량 자료의 분석에서 제안한 DC-KQR의 활용 가능성을 보여주고 있다. 나아가 균등 분할된 K개의 부분집합에 대하



(a) Mean absolute against data size



(b) Computing time against data size

Figure 3.1. Mean absolute errors and computing times against data size for the simulated example with $\tau = 0.5$.

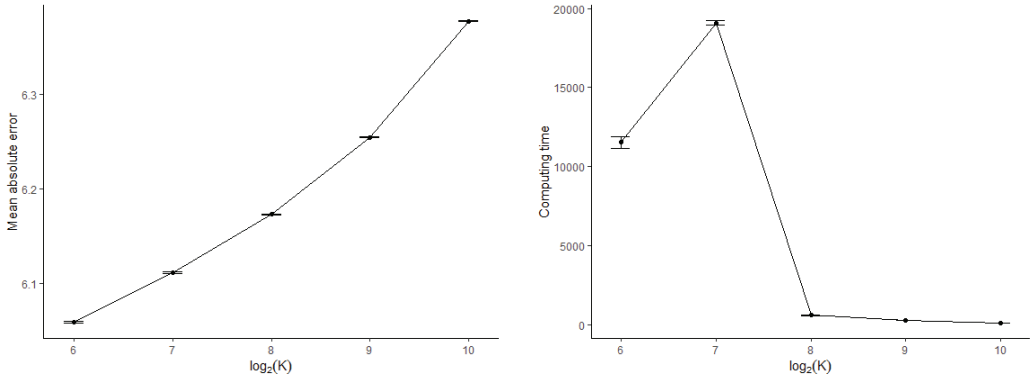
여 분위수 함수의 추정을 병렬처리하게 되면 계산 속도를 더욱 향상시킬 수 있을 것으로 기대한다. 90% 분위수 함수($\tau = 0.9$)의 추정에서 KQR과 DC-KQR의 성능은 중위수 함수($\tau = 0.5$)의 추정에서와 유사한 경향을 나타내었으며 그 결과는 Table 3.2에 제시하였다.

4. 실제 자료분석

본 논문에서 제안하는 DC-KQR 추정법의 유용성을 확인하기 위하여 Bertin-Mahieux 등 (2011)의 Million Song 자료를 분석하였다. 이 자료는 1922년부터 2011년 사이에 발매된 515,345개의 곡에 대한 발매년도와 음색에 관한 90개의 설명변수로 구성되어 있으며, 음색의 정보를 이용하여 발매년도를 추정하는 모형을 구축하는 것이 분석의 목적이다.

Table 3.2. Mean absolute errors and computation times as a function of number of partitions K and data size N for the simulated example with $\tau = 0.9$

N		K					
		1	2^2	2^3	2^4	2^5	2^6
2^8	MAE	0.381	0.519	0.814	1.556	2.525	3.916
	Time(sec)	0.286	0.039	0.058	0.095	0.166	0.344
2^9	MAE	0.259	0.317	0.413	0.824	1.353	2.649
	Time(sec)	0.553	0.234	0.081	0.121	0.188	0.340
2^{10}	MAE	0.192	0.208	0.253	0.382	0.741	1.325
	Time(sec)	1.112	1.166	0.471	0.162	0.244	0.380
2^{11}	MAE	0.131	0.132	0.165	0.214	0.355	0.766
	Time(sec)	3.995	2.019	2.297	0.950	0.321	0.493
2^{12}	MAE	0.099	0.105	0.119	0.138	0.210	0.356
	Time(sec)	22.141	3.005	3.084	3.350	1.385	0.446
2^{13}	MAE	0.070	0.074	0.081	0.087	0.117	0.190
	Time(sec)	98.681	14.968	6.329	6.405	7.209	3.002
2^{14}	MAE	0.055	0.058	0.061	0.068	0.078	0.113
	Time(sec)	408.786	108.568	24.070	11.580	11.844	13.133



(a) Mean absolute error against number of partition (b) Computing time against number of partition

Figure 4.1. Mean absolute errors and computing times against number of partition for the Million Song data.

Million Song 자료는 $N = 463{,}715$개의 훈련자료와 51,630개의 평가자료로 구분되어 있으며, 훈련자료의 균등분할 수 K는 $N \in \{2^6, 2^7, \ldots, 2^{10}\}$을 고려하였다. 3절의 모의실험과 동일하게 모형적합은 중위수 함수($\tau = 0.5$)를 추정하였으며, 가우시안 커널 함수의 커널 모수 σ^2은 훈련자료의 개체들 간의 유클리드 제곱거리의 중위수를 이용하여 선택하였다. 또한 모형의 적합에서 조율모수 λ는 훈련자료를 이용한 5-겹 교차타당법으로 선택하였다. 모형평가를 위해 크기가 51,630인 평가자료를 이용하여 평균절대오차를 계산하였으며, 이러한 절차는 10회 독립 반복시행 하였다. Figure 4.1은 평균절대오차와 모형적합에 소요되는 계산 시간의 평균값을 나타내고 있다. 예상한 바와 같이 균등분할 수 K가 증가함에 따라 DC-KQR의 평균절대오차는 다소 증가하지만, DC-KQR의 계산 속도는 현저히 감소되는 것을 확인할 수 있다. 균등분할 수 $K = 2^7$에서의 계산 시간이 $K = 2^6$에서 보다 크게 소요된 것은 하나의 컴퓨터로 균등분할 된 부분집합의 반복 추정으로 인한 것으로 병렬처리 환경에서는 균등분할 수 K가 증가할수록 계산 시간이 더욱 향상될 것이다.

5. 결론

분위수 회귀는 반응변수의 조건부 분포에 대한 포괄적인 정보를 제공하는 이점을 바탕으로 의학, 경제, 생존분석 등 여러 다양한 분야에서 널리 이용되고 있다. 특히 반응변수와 설명변수의 비선형 관계식을 고려하는 KQR은 커널함수를 이용하여 비선형 분위수 함수를 보다 정확하게 추정하나, 많은 계산 비용으로 인하여 대용량 자료의 분석에는 그 사용이 제한된다. 따라서 본 논문에서는 대용량 자료의 분석을 위하여 전체 자료를 분할한 후 분할된 자료에서 추정된 분위수 함수들을 통합하는 DC-KQR 추정법을 제안하였다.

모의실험과 실제자료 분석을 통해 균등분할 수 K가 증가함에 따라 제안한 DC-KQR의 추정 정확도는 다소 감소하는 반면, DC-KQR의 계산 속도는 현저히 향상되어 대용량 자료의 분석에 적용 가능한 것을 알 수 있다. 특히 균등 분할된 부분집합의 훈련자료의 수가 동일할 때에는 균등분할 수 K가 클수록 추정의 정확도가 높게 나타났으며, 이로부터 제안한 DC-KQR은 데이터 스트림의 형태로 수집되는 대용량 자료의 분석에서 활용될 수 있을 것으로 판단된다. 이처럼 본 논문의 모의실험과 실제자료 분석에서는 DC-KQR 추정법의 성능(추정량의 평균절대오차와 계산시간)이 균등분할 수 K에 따라 달라지는 것을 확인하였다. 따라서 차후에는 자료의 특성에 따라 적합한 균등분할 수 K를 결정하는 방법론이 개발되기를 기대해 본다.

References

Bang, S., Eo, S-H., Cho, Y., Jhun, M., and Cho, H. (2016). Non-crossing weighted kernel quantile regression with right censored data, *Lifetime Data Analysis*, **22**, 100–121.

Bang, S. and Shin, S. (2016). A comparison study of multiple linear quantile regression using non-crossing constraints, *The Korean Journal of Applied Statistics*, **29**, 773–786.

Bertin-Mahieux, T., Ellis, D., Whitman, B., and Lamere, P. (2011). The million song dataset, In Proceedings of the 12^{th} International Conference on Music Information Retrieval(IS-MIR).

Caputo, B., Sim, K., Furesjo, F., and Smola, A. (2002). Appearance-based object recognition using SVMs: Which kernel should I use?. In *Proceedings of INPS workshop on Statistical methods for computational Experiments in Visual Processing and Computer Vision*, 149–158.

Chen, X., Liu, W., and Zhang, Y. (2018). Quantile regression under memory constraint, arXiv preprint arXiv:1810.08264.

Chen, L. and Zhou, Y. (2020). Quantile regression in big data: A divide and conquer based strategy, *Computational Statistics and Data Analysis*, **144**, 1–17.

Chen, X., and Xie, M. G. (2014). A split-and-conquer approach for analysis of extraordinarily large data, *Statistica Sinica*, **24**, 1655–1684.

Cole, T. and Green, P. (1992). Smoothing Reference Centile Curves: The LMS Method and Penalized Likelihood, *Statistics in Medicine*, **11**, 1305–1319.

Dhillon, I., Guan, Y., and Kulis, B. (2004). Kernel k-means, spectral clustering and normalized cuts, *KDD 2004*, 551–556.

Fan, T., Lin, D., and Cheng, K. (2007). Regression analysis for massive datasets, *Data and Knowledge Engineering*, **61**, 554–562.

Heagerty, P. and Pepe, M. (1999). Semiparametric Estimation of Regression Quantiles with Application to Standardizing Weight for Height and Age in U.S. Children, *The Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **48**, 533–551.

Jiang, R., Hu, X., Yu, K., and Qian, W. (2018). Composite quantile regression for massive datasets, *Statistics*, **52**, 980–1004.

Kang, J. and Jhun, M. (2020). Divide-and-conquer random sketched kernel ridge regression for large-scale data, *Journal of the Korean Data & Information Science Society*, **31**, 15–23.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab-An S4 package for kernel methods

in R, *Journal of Statistical Software*, **11**, 1–20.

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

Koenker, R. and Bassett, G. (1978). Regression quantiles, *Econometrica*, **4**, 33–50.

Koenker, R. and Geling, R. (2001). Reappraising Medfly Longevity: A Quantile Regression Survival Analysis, *Journal of the American Statistical Association*, **96**, 458–468.

Koenker, R. and Hallock, K. (2001). Quantile Regression, *Journal of Economic Perspectives*, **15**, 143–156.

Li, Y., Liu, Y., and Zhu, J. (2007). Quantile regression in reproducing kernel Hilbert spaces, *Journal of the American Statistical Association*, **102**, 255–268.

Li, Y. and Zhu, J. (2008). L1-norm quantile regression, *Journal of Computational and Graphical Statistics*, **17**, 1–23.

Lin, N. and Xi, R. (2011). Aggregated estimating equation estimation, *Statistics and Its Interface*, **4**, 73–83.

Powell, D. and Wagner, J. (2014). The exporter productivity premium along the productivity distribution: evidence from quantile regression with nonadditive firm fixed effects, *Review of World Economics*, **150**, 763–785.

Vapnik, V. N. (1998). *Statistical Learning Theory*, Wiley, New York.

Wang, H. and He, X. (2007). Detecting differential expressions in genechip microarray studies: A quantile approach, *Journal of the American Statistical Association*, **102**, 104–112.

Wu, Y. and Liu, Y. (2009). Stepwise multiple quantile regression estimation using non-crossing constraints, *Statistics and Its Interface*, **2**, 299–310.

Xu, Q., Cai, C., Jiang, C., Sun, F., and Huang, X. (2020). Block average quantile regression for massive dataset, *Statistical Papers*, **61**, 141–165.

Yang, H. and Liu, H. (2016). Penalized weighted composite quantile estimators with missing covariates, *Statistical Papers*, **57**, 69–88.

Zhang, Y., Duchi, J., and Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates, *Journal of Machine Learning Research*, **16**, 3299–3340.

대용량 자료의 분석을 위한 분할정복 커널 분위수 회귀모형

방성완[a] · 김재오[b,1]

[a]육군사관학교 수학과, [b]군본부 빅데이터분석센터

(2020년 6월 15일 접수, 2020년 7월 23일 수정, 2020년 7월 27일 채택)

요 약

분위수 회귀모형은 반응변수의 조건부 분위수 함수를 추정함으로써 반응변수와 예측변수의 관계에 대한 포괄적인 정보를 제공한다. 특히 커널 분위수 회귀모형은 비선형 관계식을 고려하기 위하여 양정치 커널함수(kernel function)에 의해 만들어지는 재생 커널 힐버트 공간(reproducing kernel Hilbert space)에서 비선형 조건부 분위수 함수를 추정한다. 그러나 KQR은 이차계획법으로 공식화되어 많은 계산비용을 필요로 하므로 컴퓨터 메모리 능력의 제한으로 대용량 자료의 분석은 불가능하다. 이러한 문제점을 해결하기 위하여 본 논문에서는 분할정복(divide and conquer) 알고리즘을 활용한 KQR 추정법(DC-KQR)을 제안한다. DC-KQR은 먼저 전체 훈련자료를 몇 개의 부분집합으로 무작위로 분할(divide)한 후, 각각의 부분집합에 대하여 KQR 분위수 함수를 추정하고 이들의 산술 평균을 이용하여 최종적인 추정량으로 통합(conquer)하는 기법이다. 본 논문에서는 모의실험과 실제자료 분석을 통해 제안한 DC-KQR의 효율적인 성능과 활용 가능성을 확인하였다.

주요용어: 분할정복 알고리즘, 커널, 이차계획법, 분위수 회귀모형

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NO. 2020R1F1A1A01065107).

[1]교신저자: (32800) 충남 계룡시 신도안면 계룡대로663 사서함 501-8, 육군본부 빅데이터분석센터 분석장교.
E-mail: c14180@gmail.com