The Korean Journal of Applied Statistics
2022, Vol. 35, No. 2, 217–227

DOI: https://doi.org/10.5351/KJAS.2022.35.2.217
Print ISSN1225-066X / Online ISSN 2383-5818

Model selection via Bayesian information criterion for divide-and-conquer penalized quantile regression

Jongkyeong Kang[a], Seokwon Han[b], Sungwan Bang[1,b]

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

Abstract

Quantile regression is widely used in many fields based on the advantage of providing an efficient tool for examining complex information latent in variables. However, modern large-scale and high-dimensional data makes it very difficult to estimate the quantile regression model due to limitations in terms of computation time and storage space. Divide-and-conquer is a technique that divide the entire data into several sub-datasets that are easy to calculate and then reconstruct the estimates of the entire data using only the summary statistics in each sub-datasets. In this paper, we studied on a variable selection method using Bayes information criteria by applying the divide-and-conquer technique to the penalized quantile regression. When the number of sub-datasets is properly selected, the proposed method is efficient in terms of computational speed, providing consistent results in terms of variable selection as long as classical quantile regression estimates calculated with the entire data. The advantages of the proposed method were confirmed through simulation data and real data analysis.

Keywords: Bayesian information criterion, divide-and-conquer, large-scale data, quantile regression, variable selection

1. 서론

전통적인 평균 회귀 모형에 대한 대안으로서 Koenker와 Basset (1978)에 의해 제안된 분위수 회귀 모형은 변수에 숨겨진 복잡한 정보를 살펴보기 위한 효율적인 도구를 제공한다. 분위수 회귀 모형은 주어진 변수들의 전반적인 조건부 분포를 조사함으로서 한 변수들이 여러 분위수에 미치는 다양한 영향을 포착할 수 있으며, 오차의 분포에 대한 가정에도 자유로운 이점을 가진다. 이러한 유용성과 강건성을 바탕으로 분위수 회귀는 의학 (Cole과 Green, 1992; Heagerty와 Pepe, 1999), 경제학 (Okada과 Samreth, 2012; Powell과 Wagner, 2014), 사회과학 (Cook, 2014; Jiang 등, 2016), 생존분석 (Ning과 Tang, 2014; Bang 등, 2016), 그리고 마이크로 어레이 연구 (Alhamzawi, 2015; Yang과 Liu, 2016) 등 다양한 응용 분야에서 널리 활용되고 있다.

한편, 데이터 수집 및 탐색 기술의 발달은 빅데이터로 대표되는 현대의 대용량 데이터 분석에 대한 수요를 증폭시켰다. 이러한 대용량 데이터의 잠재적 가치를 발굴하기 위해서는 별도로 고안된 데이터 처리 및 분석 능력이 필수적이며, 기존의 알고리즘 또는 통계 패키지에 의한 분석은 시간 및 저장 공간적 측면에서 상당한 제약이 따른다. 랜덤화 샘플링 알고리즘 (Dasgupta 등, 2009)은 대용량 데이터 분석하는 대표적인 방법 중 하나로 저 왜곡 임베딩(low distortion embedding) 행렬을 이용하여 빠른 속도로 근사해를 구할 수 있다. 그러나 랜덤화 샘플링 알고리즘의 특성상 서브샘플링을 수행하기 위해서는 전체 데이터를 저장할 공간을 필요로

This work was supported by 2021 research fund of Korea Military Academy (Hwarangdae Research Institute)
[1] Corresponding Author: Department of Mathematics Korea Military Academy, 574 Hwarang-ro Nowon-gu, Seoul 01805, Korea. E-mail: wan1365@gmail.com

Published 30 April 2022 / journal homepage: http://kjas.or.kr

하며, 전체 데이터 가운데 일부만을 사용되는 알고리즘의 특성상 추정 과정에 있어 전체 데이터를 무시하는 한계가 있다. 이에 저장 공간의 문제를 해결하면서도 전체 데이터를 충분히 이용하는 블록 평균 기법 (Fan 등, 2007; Li 등, 2013), 분할-정복 기법 (Zhang 등, 2015; Chen과 Xie, 2014; Kang과 Jhun, 2020)등을 활용한 평균 회귀 모형 기법들이 개발되었다. 베이지안 통계에서도 자료를 분할하여 구한 하위사후분포등을 활용한 컨센서스 몬테 카를로 방법 (Scott 등, 2016)을 비롯한 다양한 분할 기반 기법들이 개발되었다 (Srivastava 등, 2015; Scott, 2017; Xue과 Liang, 2019). 이러한 연구를 바탕으로 대용량 데이터의 분석을 위한 분위수 회귀에 있어서도 랜덤화 샘플링 기반 (Yang 등, 2014), 블록 평균 기법 (Xu 등, 2020), 분할 정복-기법 (Chen과 Zhou, 2020)등이 최근 개발되었다.

빅데이터가 갖는 또 하나의 특징은 많은 설명변수들을 포함하고 있다는 것이다. 일반적으로 모형화 과정의 초기 단계에서는 예측 모형의 잠재적 편향을 줄이기 위해 다수의 후보 설명변수들을 포함하게 된다 (Fan과 Li, 2001). 그러나 반응변수와 관련 없는 변수가 최종 모형에 포함될 경우, 모형의 해석이 어려울 뿐만 아니라 예측 능력도 감소할 수 있다. 따라서 의미있는 정보를 제공하는 변수만을 식별하는 것은 고차원 데이터 분석에 있어 핵심적인 단계이다. 전진선택법, 후진제거법 등의 순차적 방법은 변수선택을 위한 대표적인 고전적 방법으로서 여전히 실제 문제에의 적용에 널리 활용되고 있으며, 이론적으로는 모든 경우의 수를 고려한 최고의 부분집합을 선택하는 것이 바람직할 것이다. 그러나 이와 같은 방법들은 설명변수의 수가 증가할 수록 계산 속도의 측면에서 활용성이 떨어진다. 이러한 문제를 해결하기 위해 변수 선택과 모형 적합을 동시에 하는 벌점화 방법론인 LASSO (Tibshirani, 1996)가 개발되었으며, 축소 추정에 의한 벌점화 방법론은 현재까지도 고차원 데이터를 분석하는 각 분야의 연구자들로부터 많은 관심을 받고 있다. Smoothly clipped absolute deviation (SCAD) (Fan과 Li, 2001)은 가장 인기있는 벌점화 방법론 중 하나로, 정칙 조건 하에서 SCAD 벌점화 추정량은 실제 유의한 변수만으로 얻은 추정량과 점근적으로 동일한 오라클 속성을 가진다. 분위수 회귀에서의 벌점화 방법론 역시 Li (2008), Wu와 Liu (2009)를 필두로 많은 연구가 이루어져 왔으며, 데이터 표본의 크기가 크지 않은 경우에서의 효율적인 벌점화 분위수 회귀 추정 기법에 대한 연구도 활발히 진행되어 왔다 (Belloni와 Chernozhukov, 2011; Wang 등, 2012; Peng과 Wang, 2015). 그러나 이러한 방법들 역시 표본의 크기가 큰 경우 높은 계산비용으로 인해 활용이 제한적일 수 있다.

벌점화 방법론을 통해 추정한 모형의 성능은 조율모수 선택에 의해 결정되며, 따라서 적절한 조율모수의 선택이 매우 중요하다. 교차검증법(cross-validation)은 조율모수의 선택을 위한 일반적인 방법으로서 직관적이고 구현하기 쉬운 장점이 있다. 그러나 교차검증법은 종종 과대적합을 초래할 뿐만 아니라 (Wang 등, 2007), 표본의 크기가 크거나 변수의 수가 많은 경우 계산 속도측면에서 비효율 적이다. 이에 베이즈 정보기준(Bayesian information criterion, BIC)이 고차원-대용량 데이터에서의 조율모수의 선택을 통해 축소량을 결정하는 유용한 방법으로 주목을 받고 있다. BIC는 교차검증법에 비해 계산 속도 측면에서 효율적이며, BIC를 최소로 하는 조율모수의 선택은 점근적으로 모형 선택에 일치성을 가진다 (Wang과 Leng, 2007; Wang 등, 2007; Wang 등, 2009; Lee 등, 2014).

본 논문에서는 고차원-대용량 데이터에서의 효율적인 변수 선택을 위해 BIC를 활용한 분할-정복 벌점화 분위수 회귀 방법(divide-and-conquer penalized quantile regression)에 관하여 연구하였다. 제안 방법은 분할 수를 적절히 선택했을 때 일반적인 벌점화 분위수 회귀에 비해 작은 계산 비용을 가지면서도 예측 정확성 측면에서 동등한 성능을 보여준다. 본 논문의 구성은 다음과 같다. 먼저 2장에서는 벌점화 분위수 회귀에 관하여 간략히 설명하였다. 3장에서는 제안 방법인 BIC를 활용한 분할-정복 벌점화 분위수 회귀 방법을 소개하였다. 4장에서는 모의실험과 실제 데이터 분석을 통해 제안 방법의 유용성을 확인하였다.

2. 벌점화 분위수 회귀

반응변수 Y와 p차원 설명변수 $\mathbf{X} = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$로 이루어진 n개의 개체에 대한 데이터 $\{(\mathbf{x}_1, y_1), \ldots,$

$(\mathbf{x}_n, y_n)\}$를 고려하자. 이 때 $\mathbf{x}$가 주어졌을 때의 반응변수 Y의 τ 조건부 분위수는,

$$P\left(Y \leq f^{\tau}(\mathbf{x}) \mid \mathbf{X} = \mathbf{x}\right) = \tau$$

로 정의되며 선형 분위수 회귀 모형에서는 다음과 같이 τ조건부 분위수가 설명변수의 선형 함수로 주어진다고 가정한다.

$$f^{\tau}(\mathbf{x}) = \beta_0^{\tau} + \beta_1^{\tau} x_1 + \cdots + \beta_p^{\tau} x_p.$$

이러한 분위수 회귀 모형의 적합은 체크 손실 함수 $\rho_t\tau(t) = t(\tau - I(t < 0))$를 이용하여,

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)$$

을 최소화하는 회귀계수 벡터 $\boldsymbol{\beta}^{\tau} = (\beta_0^{\tau}, \beta_1^{\tau}, \ldots, \beta_p^{\tau})^{\top}$를 추정함으로써 이뤄진다.

벌점화 방법은 손실함수를 최소화하는 과정에서 계수의 추정값에 대한 벌점항을 부과함을 통해 중요한 변수를 선택함과 동시에 모형의 계수를 추정하는 기계학습 분석 방법의 한 형태이다. 분위수 회귀 모형에서의 벌점화 방법은 양의 조율모수 $\lambda > 0$와 벌점함수 $p_{\lambda}(|\beta_j|)$ 를 도입하여

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right) + \sum_{j=1}^{p} p_{\lambda}\left(|\beta_j|\right) \tag{2.1}$$

를 최소화하는 회귀계수 벡터 $\boldsymbol{\beta}^{\tau} = (\beta_0^{\tau}, \beta_1^{\tau}, \ldots, \beta_p^{\tau})^{\top}$를 추정한다. 가장 직관적인 벌점함수의 형태로서 회귀계수의 절댓값의 q제곱을 한 $p_{\lambda}(|\beta|) = |\beta|^q$ 형태를 생각할 수 있을 것이다 (Frank와 Friedman, 1993). 대표적인 축소추정법인 LASSO (Tibshirani, 1996)는 $q = 1$일 때의 벌점함수 $p_{\lambda}(|\beta|) = |\beta|$를 이용하며, $q = 2$일 때의 벌점함수 $p_{\lambda}(|\beta|) = |\beta|^2$을 이용한 결과는 릿지 회귀를 제공한다. 그러나 이러한 형태의 벌점함수는 $q > 1$에서는 축소추정을 하지 못하며, $q < 1$에서는 추정량이 0에서 불연속이다. 또한, $q = 1$일 때의 LASSO 벌점함수는 0이 아닌 회귀계수의 추정에 있어 항상 편향이 존재하는 문제가 있다. 이에 Fan과 Li (2001)은 좋은 벌점함수의 성질로서 유의한 회귀계수의 추정에 있어서의 불편성(unbiasedness), 유의하지 않은 회귀계수를 0으로 축소추정하는 희소성(sparsity), 그리고 모형 예측에서의 불안정성을 방지하기 위한 연속성(continuity)을 가져야 함을 주장하며 SCAD 벌점함수의 사용을 제안하였다. SCAD 벌점함수는 $a > 2$에 대하여 벌점함수로

$$p_{\lambda}(|\beta|) = \lambda|\beta| I(0 \leq |\beta| < \lambda) + \frac{\frac{a\lambda|\beta| - (\beta^2 + \lambda^2)}{2}}{a - 1} I(\lambda \leq |\beta| \leq a\lambda) + \frac{(a+1)\lambda^2}{2} I(|\beta| > a\lambda),$$

를 이용한다. 실제 모형 적합에서는 $p_{\lambda}(|\beta|)$의 도함수

$$p'_{\lambda}(|\beta|) = \lambda I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)} I(|\beta| > \lambda),$$

를 이용하여 축소추정을 수행하며, 여기서 $(t)_+ = tI(t > 0)$이다. SCAD 벌점함수를 이용한 추정결과는 불편성, 희소성, 연속성을 만족할 뿐만 아니라, 실제 유의한 변수만으로 얻은 추정량과 점근적으로 동일한 오라클 속성을 가진다.

회귀분석에 있어 과소 적합된 모형은 심각하게 편향된 결과를 가져오는 반면, 과도하게 적합된 모형은 추정 효율에서 상당한 손실을 초래한다. 따라서 좋은 예측 능력을 가진 간결한 모형을 찾는 것은 중요한 작업이다. 벌점화 방법에서는 조율모수 λ의 선택을 통해 모형을 선택하게 되며, 최적의 조율모수를 찾는 것은 유의한 변수만을 선택함과 동시에 모형의 편향과 분산을 조정하여 정확한 예측 모형 생성하는 데 매우 중요한 역할을 한다. 베이즈 정보 기준(BIC)은 전통적인 회귀 분석에서의 모형 선택을 위해 사용되는 방법론 중

하나로서 계산의 간편성과 최적 모형 선택의 일치성 등을 바탕으로 널리 이용되고 있다. 조율모수가 λ일 때, 식(2.1)을 최소화하는 식을 $\hat{f}^{\tau}_{\lambda}(\cdot)$라고 하고, 이 때 모형에 포함된 설명변수의 수를 $|S|_{\lambda}$라고 하면, BIC는 다음과 같이 정의된다.

$$\mathrm{BIC}(\lambda) = \log\left(\sum_{i=1}^{n}\rho_{\tau}\left(y_i - \hat{f}^{\tau}_{\lambda}(\mathbf{x}_i)\right)\right) + \frac{\log(n)}{2n}|S|_{\lambda}. \tag{2.2}$$

이와 같이 BIC는 모형의 복잡성에 따라 벌점을 부과하여 간결한 모형을 선택하며, 다수의 조율모수 λ를 이용하여 적합한 후보 모형들 가운데 최소의 BIC값을 갖게하는 λ에서의 모형을 최종 모형으로 선택한다.
한편, p가 n이 커짐에 따라 함께 발산하는 경우에 기존의 BIC (2.2)는 유의한 변수 선택의 측면에서 일치성을 보장할 수 없다. 이에 Lee 등 (2014)은 고차원 데이터에서의 BIC로서,

$$\mathrm{hBIC}(\lambda) = \log\left(\sum_{i=1}^{n}\rho_{\tau}\left(y_i - \hat{f}^{\tau}_{\lambda}(\mathbf{x}_i)\right)\right) + C_n\frac{\log(n)}{2n}|S|_{\lambda} \tag{2.3}$$

를 사용할 것을 제안하였으며, 여기서 C_n은 n이 커짐에 따라 발산하는 상수이다. Lee 등 (2014)은 p가 n에 비해 크지 않은 때에도, 기존의 BIC (2.2)보다 hBIC가 변수 선택의 측면에서 우수함을 보였다.

3. BIC를 활용한 분할-정복 벌점화 분위수 회귀 방법

분할-정복은 그 이름에서와 같이 주어진 문제를 동일하거나 유사한 하위 문제들로 작게 분할한 뒤 각각의 개별 문제의 결과들을 결합하여 원래 문제에 대한 해답을 찾는 알고리즘 설계 기법이다. 분할 정복 기법은 다중 프로세서 또는 서로 다른 프로세서에서 하위 문제를 개별적으로 실행할 수 있기 때문에 자연적으로 병렬화 연산이 가능하다. 본 연구에서는 BIC를 활용한 분할-정복 벌점화 분위수 회귀 방법을 제안하고자 한다. 구체적인 알고리즘은 다음과 같다.

• 알고리즘 1: BIC를 활용한 분할-정복 벌점화 분위수 회귀

단계 1) 크기가 크기가 n인 데이터 $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$를 서로 배반인 t개의 부분 집합 $G_1, \ldots, G_t$로 나눈다.

단계 2) 조율모수 λ를 고정한 후, 각각의 $k = 1, 2, \ldots, t$에 대하여 다음식을 최소로 하는 국소적 분위수 회귀 추정량 $\hat{f}^{\tau}_{k,\lambda}$를 계산한다.

$$\min_{\boldsymbol{\beta}}\frac{1}{|G_k|}\sum_{(\mathbf{x},y)\in G_k}\rho_{\tau}\left(y_i - f^{\tau}(\mathbf{x}_i)\right) + \sum_{j=1}^{p}p_{\lambda}\left(|\beta_j|\right). \tag{3.1}$$

단계 3) t개의 분할-정복 추정 결과들을 이용하여 다음과 같이 BIC를 계산한다.

$$\mathrm{hBIC}^{\mathrm{DC}}(\lambda) = \log\left(\sum_{k=1}^{t}\sum_{(\mathbf{x},y)\in G_k}\rho_{\tau}\left(y_i - \hat{f}^{\tau}_{k,\lambda}(\mathbf{x}_i)\right)\right) + C_n\frac{\log(n)}{2n^2}\sum_{k=1}^{t}|G_k||S_k|_{\lambda}. \tag{3.2}$$

여기서 $|S_k|_{\lambda} = \{j : \beta_{j,k} \neq 0, 1 \leq j \leq p\}$는 $\hat{f}^{\tau}_{k,\lambda}$에 포함된 설명변수의 수이다.

단계 4) 여러 후보 λ에 대해 [단계 2]와 [단계 3]을 통해 식 (3.2)를 구한 뒤, 가장 작은 $\mathrm{hBIC}^{\mathrm{DC}}$값을 갖는 $\lambda = \hat{\lambda}$를 선택한다. $k = 1, \ldots, t$, $j = 1, \ldots, p$에 대해 $\hat{\lambda}$에서의 추정량을 $\hat{\beta}_{j,k}$로 정의한다.

단계 5) $j = 1, \ldots, p$에 대해

$$\hat{\beta}_j = \begin{cases} 0, & \text{if, } \sum_{k=1}^{t}|G_k|I(\hat{\beta}_{j,k} \neq 0) < n\alpha, \\ \frac{\sum_{k=1}^{t}|G_k|\hat{\beta}_{j,k}I(\hat{\beta}_{j,k}\neq 0)}{\sum_{k=1}^{t}|G_k|I(\hat{\beta}_{j,k}\neq 0)}, & \text{elsewhere,} \end{cases} \tag{3.3}$$

를 최종 추정량으로 결정한다.

단계 1에서 개별 부분집합의 크기 $|G_k|$는 서로 거의 동일하도록 n/t와 가까운 수가 되도록 하며, $|G_1|+\cdots+|G_t| = n$ 를 만족해야 한다. 만약 $|G_1| = \cdots = |G_t| = n/t$라면, [단계 3]에서 식 (3.2)의 두번째 항은 $\log(n)/2nt \sum_{k=1}^{t} |S_k|_\lambda$ 가 된다. 즉, 식 (2.2)에서의 $|S|_\lambda$에 대한 추정량으로, t개의 $|S_k|_\lambda$값의 평균이 이용된다. 또한 t개의 분할의 크기가 모두 같다면, 단계 5에서 식 (3.3)에서의 $\hat{\beta}_j$의 최종 추정량도

$$\hat{\beta}_j = \begin{cases} 0, & \text{if, } \sum_{k=1}^{t} I(\hat{\beta}_{j,k} \neq 0) < t\alpha \\ \frac{\sum_{k=1}^{t} \hat{\beta}_{j,k} I(\hat{\beta}_{j,k} \neq 0)}{\sum_{k=1}^{t} I(\hat{\beta}_{j,k} \neq 0)}, & \text{elsewhere,} \end{cases}$$

와 같이 보다 간단하게 나타낼 수 있다. 단계 5의 식 (3.3)에서 사용된 α는 일종의 한계점(threshold)으로서 $100\alpha\%$ 이상의 분할에서 유의하다고 포함된 변수에 대해서만 회귀계수를 계산하는 역할을 한다. 예를 들어 $\alpha = 0.5$인 경우, 과반수 이상의 분할에서 0이 아니라고 추정된 선택한 한해서만 그들의 가중 평균으로 $\hat{\beta}_j$를 추정하고, 만일 0이 아니게 추정된 분할이 절반에 미치지 못한다면, $\hat{\beta}_j = 0$으로 설정하였다. α의 값이 작을수록 미세한 신호도 모형에 포함될 가능성이 높아지지만 모형의 변동성이 커질 수 있으며, α의 값이 클 수록 보다 강건한 모형을 제공한다. 본 연구에서는 Lee 등 (2014)이 제안한 hBIC (2.3)를 사용하였으며, 본 논문에는 싣지 않았으나, 기존의 BIC (2.2)를 사용했을 때보다 변수 선택 및 올바른 모형 식별 측면에서 더 우수한 성능을 보임을 확인하였다.

4. 모의실험 및 실제 데이터 분석

이 장에서는 제안된 분할-정복 벌점화 분위수 회귀의 효율성을 기존의 벌점화 회귀와 비교하기 위해 모의실험을 수행하고, 실제 데이터의 분석을 통해 활용 가능성을 살펴보았다. 모의실험 및 실제 데이터 분석에서는 $a = 3.7$ 인 SCAD 벌점함수를 사용하였으며 (Fan과 Li, 2001), 2.2GHz의 속도를 내는 20개의 프로세스가 장착된 리눅스 환경에서 R 프로그래밍을 통해 진행되었다.

4.1. 모의실험

먼저 설명변수 $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$를 다변량 정규분포 $N_p(\mathbf{0}, \mathbf{\Sigma})$ 로부터 추출하였으며, 여기서 $\mathbf{\Sigma} = (\sigma_{jk})_{p\times p}$, $\sigma_{jk} = 0.5^{|j-k|}$이다. 즉 인접한 변수들 간의 공분산은 0.5이며, 변수의 인덱스간의 차가 클수록 상관관계가 적어지도록 설계하였다. 본 논문의 모의실험에서는 다음 선형 회귀 모형을 고려하였다.

$$Y = X_6 + X_{12} + X_{15} + X_{20} + 0.7\Phi(X_1)\epsilon.$$

여기서 오차항 ϵ은 표준정규분포를 따르며, $\Phi(X)$는 표준정규분포의 누적분포함수이다. 이 실험모형은 Wang 등 (2012), Peng과 Wang (2015)이 사용한 모형과 동일하다. X_1은 반응변수의 조건부 분포에서 중요한 역할을 하지만 조건부 분포의 중위수에는 직접적인 영향을 미치지 않는다. 후보 변수의 수는 $p = 100$을, 표본의 크기는 $n = 10,000$으로 설정하였으며, 분위수 τ는 0.5와 0.75를 고려하였다. 제안한 분할-정복 분위수 회귀에서의 분할수로는 $t = 50, 100$을 이용하였다. 조율모수의 선택을 위해서 20개의 후보 λ 를 도입하였으며, 식(2.3)와 식 (3.2)에서의 hBIC를 최소로 하는 λ를 선택하였다. hBIC에서 사용한 상수 C_n은 Lee 등 (2014)의 제안에 따라 $\log(p)$를 이용하였다. 모형의 평가를 위해 선택한 변수의 수(Size), $(X_6, X_12, X_{15}, X_{20})$를 모두 선택한 비율(P1), X_1을 선택한 비율(P2), 실제 회귀계수와의 절대 예측 오차 $\sum_{j=1}^{p} |\hat{\beta}_j - \beta_j|$를 계산하였다. 추가적으로 선택한 조율모수 λ의 값(Selected lambda λ)과 20개의 λ에 대해 추정하는 데 걸리는 시간(Time (s))을 함께 측정하였다.

Table 1: Simulation results

Method	Size	P1	P2	AE	Selected λ	Time(s)
Oracle($\tau = 0.5$)	4	100%	0%	0.0046(0.0021)	-	0.10(0.02)
$t = 1(\tau = 0.5)$	4	100%	0%	0.0047(0.0021)	0.103(0.055)	32.97(1.69)
$t = 50(\tau = 0.5)$	4	100%	0%	0.0084(0.0030)	0.158(0.015)	2.69(0.16)
$t = 100(\tau = 0.5)$	4	100%	0%	0.0099(0.0038)	0.158(0.009)	2.71(0.14)
Oracle($\tau = 0.75$)	5	100%	100%	0.0054(0.0024)	-	0.10(0.02)
$t = 1(\tau = 0.75)$	5	100%	100%	0.0141(0.0050)	0.0897(0.006)	41.97(4.59)
$t = 50(\tau = 0.75)$	5	100%	100%	0.0149(0.0055)	0.0820(0.007)	4.245(1.21)
$t = 100(\tau = 0.75)$	4.99(0.1)	100%	99%	0.0173(0.0068)	0.1228(0.006)	3.97(0.63)

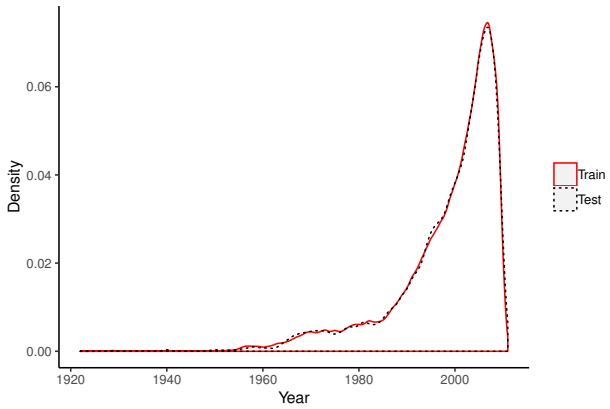

Figure 1: *Empirical density of million songs data.*

Table 1은 위의 모의실험을 100회 반복한 결과를 정리해서 보여주고 있다. 여기서 AE, Selected λ, Time (s)는 100번 반복했을 때의 평균값을 나타내며, 괄호 안은 표준편차를 나타낸다. 비교를 위해 실제 중요한 변수들만을 이용한 일반적인 분위수 회귀 결과(Oracle)도 함께 나타내었으며, 이 때의 계산시간은 1회의 계산에 소요된 시간이다. $t = 1$일 때의 결과는 SCAD 벌점화를 이용한 일반적인 분위수 회귀를 나타낸다. $\tau = 0.5$인 중위수 회귀의 경우 모든 방법들이 올바르게 변수들을 선택하였으며, 특히 $t = 1$일 때의 일반적인 SCAD 벌점화 분위수 회귀는 예측 오차 측면에서 중요한 변수만을 이용하여 적합한 결과와 거의 동등하였다. $t = 50$일 때와 $t = 100$일 때에는 $t = 1$일 때에 비해 10배 가량 계산 속도를 단축할 수 있었다, 분할 수가 증가함에 따라 더 큰 조율모수를 선택하는 경향이 있었으며, 예측 오차도 증가하는 경향을 보였다. $\tau = 0.75$일 때의 실험에서도 $\tau = 0.5$에서와 마찬가지로 모든 방법이($X_6, X_{12}, X_{15}, X_{20}$) 변수들을 모두 선택하였으며, $t = 100$일 때에 단 1회 X_1을 선택하지 못하였다. $t = 50$일 때의 실험결과는 $t = 1$일 때에 비해 약 10배 가량 효율적인 계산을 수행하면서도 예측오차 측면에서 거의 동등한 결과를 보여줬다.

4.2. 실제 데이터 분석

제안 방법의 유용성을 확인하기 위해 Bertin-Mahieux 등 (2011)의 Million Songs 데이터를 분석하였다. 이 데이터는 1922년부터 2011년 사이에 발매된 곡에 대한 $n = 463,715$개의 훈련 데이터와 51,630개의 시험 데이터로 구성되어 있으며, 노래를 구성하는 음성적 특성을 바탕으로 노래가 발매된 연도를 예측하는 모형을 구축하는 것을 분석 목적으로 한다. Figure 1은 노래의 발매연도에 대한 경험적 밀도를 훈련 데이터와 시험

Table 2: Comparison results for million songs data

Method	Number of variables	MCE	Selected λ	Time (s)
$t = 1(\tau = 0.5)$	59	3.26	0.0041	1836
$t = 1000(\tau = 0.5)$	22	3.30	0.0291	1502
$t = 2000(\tau = 0.5)$	13	3.34	0.0271	99
$t = 1(\tau = 0.75)$	60	2.20	0.0031	3570
$t = 1000(\tau = 0.75)$	22	2.21	0.0291	2060
$t = 2000(\tau = 0.75)$	13	2.23	0.0271	715

데이터에 따라 나타내고 있다. 훈련 데이터와 시험 데이터는 거의 동일한 분포로 나눠져 있으며, 2007년을 봉우리로 하여 왼쪽으로 꼬리가 긴 분포를 하고 있다. 여기서는 노래의 발매년도를 반응변수 $Y \in [1922, 2011]$로 하고, 노래의 음색에 대한 12가지 정보의 평균 및 공분산으로부터 생성된 90개의 특성을 설명변수로 이용하여 $\tau = 0.5$와 $\tau = 0.75$일 때의 분위수 회귀 분석을 수행하였다. 분석에 앞서 설명변수들은 모두 표준편차가 1이 되도록 표준화하였다.

제안한 분할-정복 분위수 벌점화 회귀에서의 분할수로는 $t = 1000, 2000$을 이용하였다. 이보다 적은 분할수에서는 일반적인 SCAD 벌점화 분위수 회귀에 비해 계산 시간 측면에서 효율적이지 않았다. 조율모수의 선택을 위해서 30개의 후보 λ 를 도입하였으며, 모의실험에서와 마찬가지로 (2.3)와 식 (3.2)에서의 hBIC를 최소로 하는 λ를 선택하였고, hBIC에서 사용한 상수 $C_n = \log(p)$를 이용하였다. 모형의 정확성을 평가하기 위해 시험 데이터를 이용하여 평균 체크 오차(Mean check error, MCE) $\sum_{i=1}^{n} \rho_\tau(y_i - \hat{f}^\tau(\mathbf{x}))$를 측정하였으며, 비교를 위해 선택한 변수의 수(Number of variables), 선택한 조율모수 λ의 값(Selected lambda λ)과 30개의 λ에 대해 추정하는 데 걸리는 시간(Time (s))를 함께 기록하였다. Million Songs 데이터의 분석결과는 Table 2에 나타나있다.

$\tau = 0.5$ 에서의 실험결과 변수 선택의 측면에서 $t = 1$일 때의 방법이 59개의 변수를 선택한 반면, $t = 1000$일 때의 방법은 22개의, $t = 2000$개의 방법은 13개의 변수만을 선택하여, 제안한 분할-정복 벌점화 분위수 회귀 방법이 더 성긴 모형을 생성하였다. 제안 방법들은 더 적은 변수만을 선택하였기 때문에 $t = 1$일 때에 비해 평균 체크 오차가 더 크게 측정되었다. 반면 $t = 1000$일 때의 계산 시간은 $t = 1$일 때의 약 0.82배, $t = 2000$일 때의 계산 시간은 $t = 1$일 때의 약 0.05배로 제안한 방법이 계산 속도의 측면에서 매우 효율적임을 확인할 수 있다. $\tau = 0.75$에서의 실험에서 제안 방법은 $\tau = 0.5$에서의 실험에서와 동일한 변수들을 선택하였으며, $t = 1$일 때의 방법은 1개의 변수를 추가적으로 선택하였다. 이 실험에서도 $t = 1$일 때의 평균 체크 오차가 가장 작게 나타났다. 하지만 $t = 1000$일 때의 결과는 $t = 1$일 때에 비해 38개의 변수를 덜 선택하고, 0.58배의 계산 시간만을 소모하면서 단지 0.01의 평균 체크 오차만을 증가시켰다. 또한 $t = 2000$일 때의 결과는 $t = 1$일 때에 비해 47개의 변수를 덜 선택하고, 0.20배의 계산 시간만을 소모하면서 0.03의 평균 체크 오차만을 증가시켰다. 반응변수의 측도가 연도인 것을 감안할 때, 소숫점 둘째 자리에서의 평균 체크 오차의 차이는 실질적으로 그리 크지 않다고 할 수 있으며, 보다 간결한 모형을 효율적인 계산 시간을 통해 제공한다는 측면에서 제안 방법의 활용성을 확인할 수 있었다.

5. 결론

본 연구에서는 고차원-대용량 데이터에서의 효율적인 변수 선택을 위해 BIC를 활용한 분할-정복 벌점화 분위수 회귀방법을 제안하였다. 제안 방법은 분할 수를 적절하게 선택하였을 때, 전체 데이터를 한 번에 분석하는 기존의 방법에 비해 적은 계산 비용으로 중요한 변수를 올바르게 선택함을 확인하였다. 표본의 크기가 크고 변수의 수가 작은 경우에 대해서는 Frisch–Newton 알고리즘 (Portnoy와 Koenker, 1997; Koenker와 Ng, 2005)

등을 이용하여 빠른 모형 적합이 가능하므로, 보다 정확하고 효율적인 모형 적합을 위한 사전 변수 선별 방법으로서 제안 방법을 사용할 수 있을 것이다. 또한, 분할 정복 기법의 특성상 병렬 계산이 용이하기 때문에 다중 프로세서 또는 GPU 기반 병렬 컴퓨팅 환경에서 상당한 속도 향상으로 이어질 수 있다 (Jung과 Lim, 2017). 본 연구에서 제안하는 분할 정복 기법은 서포트 벡터 머신 (Kim과 Shin, 2017)과 같이 대규모 데이터를 다루는 다른 형태의 통계적 문제를 분석함에 있어 계산 시간과 저장 공간을 절약할 수 있는 가능성이 있다고 판단되며, 본 논문의 결과가 관련 연구의 탐색에 유용할 것으로 기대한다.

References

Alhamzawi R (2015). Model selection in quantile regression models, *Journal of Applied Statistics*, **42**, 445–458.

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.

Belloni A and Chernozhukov V (2011). L1-penalized quantile regression in high-dimensional sparse models, *The Annals of Statistics*, **39**, 82–130.

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

Chen L and Zhou Y (2020). Quantile regression in big data: A divide and conquer based strategy, *Computational Statistics & Data Analysis*, **144**, 106892.

Chen X and Xie MG (2014). A split-and-conquer approach for analysis of extraordinarily large data, *Statistica Sinica*, 165–1684.

Cole T and Green P (1992). Smoothing reference centile curves: The LMS method and penalized likelihood, *Statistics in Medicine*, **11**, 1305–1319.

Cook L (2014). Gendered parenthood penalties and premiums across the earnings distribution in Australia, the United Kingdom, and the United States. *European Sociological Review*, **30**, 360–372.

Dasgupta A, Drineas P, Harb B, Kumar R, and Mahoney MW (2009). Sampling algorithms and corsets for l_p regression, *SIAM Journal on Computing*, **38**, 2060–2078.

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American statistical Association*, **96**, 1348–1360.

Fan TH, Lin DK, and Cheng KF (2007). Regression analysis for massive datasets, *Data & Knowledge Engineering*, **61**, 554–562.

Frank I and Friedman J (1993). A statistical view of some chemometrics regression tools, *Technometrics*, **35**, 109–148.

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, Qian W, and Zhou Z (2016). Single-index composite quantile regression with heteroscedasticity and general error distributions, *Statistical Papers*, **57**, 185–203.

Jung BH and Lim DH (2017). Comparison analysis of big data integration models, *Journal of the Korean Data & Information Science Society*, **28**, 755–768.

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

Kim KH and Shin SJ (2017). Adaptive ridge procedure for L_0-penalized weighted support vector machines,

Journal of the Korean Data & Information Science Society, **28**, 1271–1278.

Koenker R and Ng P (2005). A Frisch-Newton algorithm for sparse quantile regression, *Acta Mathematicae Applicatae Sinica*, **21**, 225–236.

Lee E, Noh H and Park B (2014). Model selection via Bayesian information criterion for quantile regression models, *Journal of the American Statistical Association*, **109**, 216–229.

Li R, Lin DK, and Li B (2013). Statistical inference in massive data sets, *Applied Stochastic Models in Business and Industry*, **29**, 399–409.

Li Y (2008). L1-norm quantile regression, *Journal of Computational and Graphical Statistics*, **17**, 163–185.

Ning Z and Tang L (2014). Estimation and test procedures for composite quantile regression with covariates missing at random, *Statistics & Probability Letters*, **95**, 15–25.

Okada K and Samreth S (2012). The effect of foreign aid on corruption: a quantile regression approach, *Economics Letters*, **115**, 240–243.

Peng B and Wang L (2015). An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression, *Journal of Computational and Graphical Statistics*, **24**, 676–694.

Portnoy S and Koenker R (1997). The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators, *Statistical Science*, **12**, 279–300.

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.

Scott SL, Blocker AW, Bonassi FV, Chipman HA, George EI, and McCulloch RE (2016). Bayes and big data: The consensus Monte Carlo algorithm, *International Journal of Management Science and Engineering Management*, **11**, 78–88.

Scott SL (2017). Comparing consensus Monte Carlo strategies for distributed Bayesian computation, *Brazilian Journal of Probability and Statistics*, **31**, 668–685.

Srivastava S, Cevher V, Dinh Q, and Dunson D (2015). WASP: Scalable Bayes via barycenters of subset posteriors, *In Artificial Intelligence and Statistics*, 912–920.

Tibshirani R (1996). Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society: Series B*, **58**, 267–288.

Wang H and Leng C (2007). Unified Lasso estimation by least squares approximation, *Journal of the American Statistical Association*, **102**, 1418–1429.

Wang H, Li B, and Leng C (2009). Shrinkage tuning parameter selection with a diverging number of parameters, *Journal of the Royal Statistical Society, Series B*, **71**, 671–683.

Wang H, Li R, and Tsai CL (2007). Tuning parameter Selectors for the Smoothly Clipped Absolute Deviation Method, *Biometrika*, **94**, 553–568.

Wang L, Wu Y, and Li R (2012). Quantile regression for analyzing Heterogeneity in Ultra-High Dimension, *Journal of American Statistical Association*, **107**, 214–222.

Wu Y and Liu Y (2009). Variable selection in quantile regression. *Statistica Sinica*, 801–817.

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

Xue J and Liang F (2019). Double-parallel Monte Carlo for Bayesian analysis of big data, *Statistics and Computing*, **29**, 23–32

Yang J, Meng X, and Mahoney MW (2014). Quantile regression for large-scale applications, *SIAM Journal on Scientific Computing*, **36**, 78–110.

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

Received October 19, 2021; Revised December 02, 2021; Accepted January 03, 2022

베이즈 정보 기준을 활용한 분할-정복 벌점화 분위수 회귀

강종경[a], 한석원[b], 방성완[1,b]

[a]강원대학교 정보통계학전공; [b]육군사관학교 수학과

요 약

분위수 회귀 모형은 변수에 숨겨진 복잡한 정보를 살펴보기 위한 효율적인 도구를 제공하는 장점을 바탕으로 많은 분야에서 널리 사용되고 있다. 그러나 현대의 대용량-고차원 데이터는 계산 시간 및 저장공간의 제한으로 인해 분위수 회귀 모형의 추정을 매우 어렵게 만든다. 분할-정복은 전체 데이터를 계산이 용이한 여러개의 부분집합으로 나눈 다음 각 분할에서의 요약 통계량만을 이용하여 전체 데이터의 추정량을 재구성하는 기법이다. 본 연구에서는 분할-정복 기법을 벌점화 분위수 회귀에 적용하고 베이즈 정보기준을 활용하여 변수를 선택하는 방법에 관하여 연구하였다. 제안 방법은 분할 수를 적절하게 선택하였을 때, 전체 데이터로 계산한 일반적인 분위수 회귀 추정량만큼 변수 선택의 측면에서 일관된 결과를 제공하면서 계산 속도의 측면에서 효율적이다. 이러한 제안된 방법의 장점은 시뮬레이션 데이터 및 실제 데이터 분석을 통해 확인하였다.

주요용어: 대용량 데이터, 베이즈 정보 기준, 변수 선택,분위수 회귀, 분할-정복

본 논문은 육군사관학교 화랑대연구소의 2021년도(21-군학-19) 연구활동비 지원을 받아 연구되었음.

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