Journal of the Korean Data & Information Science Society
2017, **28**(3), 533–545

http://dx.doi.org/10.7465/jkdi.2017.28.3.533
한국데이터정보과학회지

다변량 분위수 회귀나무 모형에 대한 연구[†]

김재오[1] · 조형준[2] · 방성완[3]

[12]고려대학교 통계학과 · [3]육군사관학교 수학과

접수 2017년 4월 18일, 수정 2017년 5월 24일, 게재확정 2017년 5월 27일

요 약

분위수 회귀모형은 반응변수의 조건부 분포에 대하여 포괄적이고 유용한 통계적 정보를 제공한다. 그러나 많은 실제 자료는 설명변수와 반응변수가 비선형의 관계를 갖고 있어 전통적인 선형 분위수 회귀모형은 왜곡되고 잘못된 결과를 초래할 수 있다. 또한 자료의 복잡성이 증가하여 반응변수가 여러 개인 다변량 자료의 분석에 대한 보다 정확한 예측과 더불어 풍부한 해석에 대한 요구가 증가하고 있다. 이러한 이유로 본 연구에서는 다변량 분위수 회귀나무 모형을 제안하였다. 본 연구에서는 기존의 다변량 회귀나무 모형의 분할변수 선택 알고리즘의 문제점을 지적하고 향상된 분할변수 선택 알고리즘을 제안하였다. 제안한 알고리즘은 합리적인 계산시간으로 적용 가능하며 분할변수 선택에서 편향 발생의 문제를 갖지 않는 동시에 기존 방법보다 더 정확하게 분할변수를 선택할 수 있었다. 본 연구에서는 모의실험과 실증 예제를 통해 제안한 방법의 우수한 성능과 유용성을 확인하였다.

주요용어: 다변량 자료분석, 데이터마이닝, 분위수 회귀모형, 회귀나무 모형.

1. 서론

분위수 회귀 (quantile regression)는 반응변수의 조건부 평균 (conditional mean)을 추정하는 최소제곱법 (ordinary least squares)에 비하여 설명변수에 대한 반응변수의 다양한 분포 형태를 추정할 수 있다는 점에서 매우 유용한 통계적 방법으로 각광받고 있다. 특히 자료에 이상치 (outlier)가 존재하거나 오차항 분포의 꼬리 부분이 두꺼운 경우 최소제곱법에 의한 추정은 바람직하지 않은 것으로 알려져 있으나, 분위수 회귀는 상대적으로 강건한 (robust) 추정 결과를 제공한다. 통계적 유용성과 강건한 추정의 장점을 바탕으로 분위수 회귀모형은 경제 (economics), 생물통계 (biostatistics), 교육 (education) 및 생존분석 (survival analysis) 등 매우 다양한 분야에서 활용되고 있다 (Koenker, 2005). 특히 Shim과 Hwang (2012)의 소지역 추정을 위한 분위수 커널 회귀와 Jeremiah와 Jung (2014)의 한국의 세대 간 경제적 이동성에 대한 연구 등과 같이 많은 사회과학 분야에 적용되고 있다.

그러나 때때로 분위수 회귀모형이 가정하는 선형성 (linearity)은 모형의 유연성 (flexibility) 측면에서 완화 (relaxation)할 필요가 강하게 대두된다. 비선형 분위수 회귀모형을 위한 비모수적 (non-parametric) 방법은 매우 다양하게 연구되고 있다. 대표적인 연구로 Yu와 Jones (1998)과 Hallin 등 (2009)이 제안한 국소 선형분위수 회귀 (local linear quantile regression)방법, Koenker와 Mizera

[†] 본 연구는 2015년 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF2015R1C1A1A02036473, NRF-2015R1D1A1A09058602).
[1] (02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과, 박사수료.
[2] (02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과, 교수.
[3] 교신저자: 서울특별시 노원구 화랑로 574, 육군사관학교, 부교수. E-mail: wan1365@gmail.com

(2004)의 분위수 평활 스플라인 (quantile smoothing splines)을 활용한 방법, Li 등 (2007)과 Liu와 Wu (2011)의 커널함수 (kernel function) 방법 등이 있다. 특히 Chaudhuri와 Loh (2002)는 나무구조 모형 (tree-structured model)에 분위수 회귀모형을 적용한 분위수 회귀나무모형 (quantile regression tree) 방법을 제안한 바 있다.

한편 현실의 많은 문제는 여러 개의 반응변수를 갖는 다변량 (multivariate) 자료의 분석을 필요로 한다. 분위수 회귀도 최소제곱법과 마찬가지로 자연스럽게 다변량 분위수 회귀로 확장이 가능하다. 다변량 자료에 대한 다양한 분석방법이 개발되어 있으나 본 연구에서는 다변량 회귀나무 모형으로 한정한다. 회귀나무는 분석과정을 나무구조로 도식화한 대표적인 데이터마이닝 (data mining) 기법이다. 이 방법은 일반적으로 뛰어난 해석 (interpretation)과 높은 예측 (prediction)이 가능하여 다양한 현실 문제에 활용이 되고 있다. 다변량 자료에 대한 나무모형의 핵심 사항은 나무모형의 해석력을 극대화하기 위하여 하나의 나무모형으로 모형을 구축하는 것이다. 다변량 회귀나무 모형에 대한 기존 연구 고찰은 2.2절에서 상세하게 제시한다.

대표적인 나무모형 방법인 CART (classification and regression tree)와 같이 가능한 모든 분할점 또는 분할집합을 점검하여 나무모형을 구축하는 완전 탐색 (exhaustive search) 방법은 알려진 중대한 문제가 있다. 먼저 설명변수에 따라 과도하게 많은 계산시간이 소요되는 것이다. 예를 들어 범주 20개를 갖는 설명변수에 대해 최적의 분할집합을 선택하기 위해서는 524,287 ($2^{19}-1$)번 적합식을 계산해야 한다. 둘째 분할변수 선택간 편향(bias)이 발생하는 것으로 알려져 있다 (Loh 2002; 2009). 이러한 편향은 가능한 분할점 또는 분할집합이 많은 경우 주로 발생하며 연속형 변수의 경우 양 끝에 편향되는 것으로 알려져 있다. 이러한 문제를 극복하기 위하여 Loh와 Vanichsetakul (1988)의 FACT (fast algorithm for classification tree), Kim과 Loh (2011)의 CRUISE (classification rule with unbiased interaction selection and estimation), Loh (2002; 2009)의 GUIDE (generalized, unbiased, interaction detection and estimation) 등이 개발되어 왔다.

본 연구는 다변량 분위수 회귀나무 모형의 알고리즘을 제안하기 위한 것으로 다음과 같은 의의를 가진다. 첫째 분위수 회귀의 선형 가정을 완화하여 다양한 자료에 대해 적용 가능성을 확대함과 동시에 최종 노드에서는 선형 적합을 하여 기존 비모수적 방법에 비하여 직관적인 해석이 가능하다. 둘째 여러 개의 반응변수와 설명변수로 이루어진 복잡한 다변량 자료에 대하여 통합된 하나의 나무모형을 제공하여 데이터 마이닝 측면에서 많은 효용성을 확인할 수 있다. 특히 분할변수 선택에 있어 Loh와 Zheng (2013)이 제안한 다변량 회귀나무 모형의 분할변수 선택 알고리즘의 단점을 보완한 새로운 알고리즘을 제안하고 모의실험을 통하여 그 우수성을 입증하였다. 제안하는 알고리즘은 일변량 자료에 적용하는 GUIDE 알고리즘의 일부를 적용하여 합리적인 계산시간과 선택편향이 거의 발생하지 않는 특징이 있다. 본 연구에서 제안한 방법의 구현 및 모의실험, 실증예제는 모두 R 프로그래밍 3.2.2를 이용하였다.

본 연구의 구성은 다음과 같다. 2절에서 분위수 회귀나무 모형과 다변량 회귀나무 모형에 대한 기존 연구를 고찰한 뒤, 3절에서 나무모형의 적합식, 불순도 함수 (impurity function), 분할규칙 (split rule) 및 나무모형의 크기를 결정하는 방법을 포함하는 다변량 분위수 회귀나무 모형을 제안한다. 4절에서 기존 방법과 분할변수 선택 및 예측력을 비교한 모의실험 결과를 제시하고, 5절에서 시멘트의 점성 및 강성에 대한 세 가지 반응변수를 가지는 자료에 대해 적용함으로써 제안된 방법의 활용 가능성을 보인다. 마지막 6절에서 결론 및 향후 연구방향을 제시한다.

2. 기존 연구에 대한 고찰

본 절에서는 분위수 회귀나무 모형 및 다변량 회귀나무 모형에 대한 기존 연구를 고찰한다. 특히 Chaudhuri와 Loh (2002)가 제안한 분위수 회귀나무 모형과 Loh와 Zheng (2013)이 제안한 다변량 회

귀나무 모형의 분할변수 선택 알고리즘에 대하여 상세히 알아본다.

2.1. 분위수 회귀나무 모형

p차원 설명변수 $\boldsymbol{x} \in R^p$와 1차원 반응변수 $y \in R$로 이루어진 크기가 n인 훈련자료 $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$가 주어졌다고 하자. 반응변수 y에 대한 $100\tau\%$ 조건부 분위수 함수 (conditional quantile function of $y|\boldsymbol{x}$) $q_\tau(y|\boldsymbol{x})$는

$$P(Y \le q_\tau(\boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}) = \tau \quad \text{for} \quad 0 < \tau < 1 \tag{2.1}$$

와 같이 정의되며, Koenker와 Bassett (1978)은 조건부 선형 분위수 함수 $q_\tau(\boldsymbol{x}) = \mathbf{x}^T\boldsymbol{\beta}_\tau$를 가정하여

$$\min_{\boldsymbol{\beta}_\tau} \sum_{i=1}^{n} \rho_\tau \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau \right) \tag{2.2}$$

와 같은 최적화 문제의 최적해로서 회귀계수 벡터 $\boldsymbol{\beta}_\tau = (\beta_{0,\tau}, \beta_{1,\tau}, \cdots, \beta_{p,\tau})^T$를 추정하였다. 여기서 $\rho_\tau(u) = u(\tau - I(u < 0))$는 체크 손실함수 (check loss function)이며 $\mathbf{x} = (1, \boldsymbol{x}^{\boldsymbol{T}})^T$는 절편 (intercept)를 포함한 $p+1$차원 설명변수를 나타낸다. 서론에서 언급한 바와 같이 선형 분위수 회귀모형은 최소제곱 (least square) 손실함수를 사용하는 일반적인 평균에 대한 회귀모형에 비해 이상치의 존재와 오차항의 확률분포에 크게 영향을 받지 않는 강건한 특성을 가진다.

그러나 많은 현실 문제에 있어서 선형 가정 (linearity assumption)의 완화가 필요한 경우가 빈번히 발생한다. 비선형 분위수 함수의 추정을 위해 Chaudhuri와 Loh (2002)는 회귀나무 구조에 조각별 다항식 (piecewise polynomial) 분위수 회귀식을 노드 (node)마다 적합하는 분위수 회귀나무 모형을 제안하였다. 이 방법에서는 가능한 모든 분할점 또는 분할집합을 대상으로 점검하는 완전 탐색 계산 방법과는 상반되는 GUIDE 알고리즘을 적용하였으며, 세부적인 GUIDE 알고리즘은 다음과 같이 요약할 수 있다.

1단계 분위수 회귀모형을 적합한 뒤 잔차를 계산한다.

2단계 잔차 값이 0보다 크면 1을, 그렇지 않으면 -1을 부여한 부호벡터 (sign vector) S를 만든다. 이 벡터의 값을 행 (row)으로 하고 분할후보변수 Z를 적절한 간격으로 구분하여 열 (column)로 설정한다. 이러한 두 변수에 대해 분할후보변수별 분할표 (contingency table)를 생성한다.

3단계 2단계에서 생성한 분할표에 대해 카이제곱검정 (chi-square test) 결과 유의확률을 계산한다. 이때 Wilson-Hilferty 근사방법을 이용하여 모든 분할후보변수의 자유도를 1로 조정한다.

4단계 3단계의 유의확률 중 최소값을 갖는 분할후보변수를 분할변수로 결정한다.

5단계 선정된 분할후보변수에 대해 가능한 모든 탐색점을 점검하여 분할점 또는 분할집합을 선정한다.

GUIDE 알고리즘은 분할변수를 선택한 뒤 분할점 또는 분할집합을 선택하는 과정으로 구분되어 분할변수 선택간 발생될 수 있는 선택 편향을 제거하고 계산량을 획기적으로 줄이는 방법으로 알려져 있다.

2.2. 다변량 회귀나무 모형

반응변수가 여러 개로 구성되는 회귀나무 모형, 즉 다변량 회귀나무 모형으로의 확장 연구는 지속적으로 있어왔다. Segal (1992)은 경시적 (longitudinal) 자료에 대해 CART를 적용하였으며 Zhang (1998)은 다변량 이항 (binary) 반응변수로 CART를 확장하였다. Zhang과 Ye (2008)는 Zhang (1998)의 연구를 순서변수 (ordinal variable)로 추가 확장하였다. 또한 다변량 회귀나무 모형에 대

한 R 프로그래밍 패키지로는 MVPART (De'ath, 2012)가 있다. 그러나 이러한 연구는 완전 탐색 계산을 수행하는 CART에 기반하여 앞서 언급한 선택편향 및 과도한 계산시간의 문제를 내재한다.

이러한 문제를 해결하기 위해 Loh와 Zheng (2013)은 GUIDE 알고리즘을 확장하여 경시적 자료와 다변량 자료에 적합한 회귀나무 모형을 제안하였다. 이 방법은 조건부 평균에 대한 것으로 분위수 모형과는 관계가 없으며 다음과 같이 요약할 수 있다.

1단계 반응변수 벡터별 평균을 구하여 특정 반응변수의 값이 평균보다 크면 1을 그렇지 않으면 -1을 부여한 부호벡터 를 만든다.

2단계 부호벡터 의 모든 조합된 값을 열로, 분할후보변수 를 적절한 간격으로 구분한 것을 행으로 설정한 분할표를 생성한다.

3단계 $\sim$ 5단계 2.1절의 GUIDE 알고리즘과 동일하게 처리한다.

Loh와 Zheng (2013)의 확장된 GUIDE 알고리즘은 반응변수의 차원이 증가함에 따라 때때로 매우 큰 분할표 (big contingency table; BC)를 생성한다. 예를 들어, 다섯 개의 반응변수를 갖는 경우 2^5개 열을 갖게 된다. 이러한 이원 (two-way) 분할표는 열과 행의 변수가 서로 독립 (independence)이라는 귀무가설 (null hypothesis)을 검정하기 위한 검정 통계량 (test statistic)의 계산을 부정확하게 유도할 가능성이 매우 높다. 또한 회귀나무 모형이 반복적으로 표본을 분할하여 표본의 크기가 급속히 줄어드는 것을 고려할 때 BC방법과 같이 표본 (sample)의 크기에 크게 의존하는 방법은 바람직하지 않다. 따라서 본 연구에서는 BC방법의 단점을 보완한 새로운 다변량 회귀나무 모형을 제안한다.

3. 다변량 분위수 회귀나무 모형

본 절에서는 나무모형의 주요 구성요소인 불순도 함수, 분할규칙, 나무모형의 크기 결정 방법을 중심으로 다변량 분위수 회귀나무 모형에 대해 설명한다. 특히 2절에서 설명한 BC방법을 대신하는 새로운 분할변수 선택 알고리즘을 제안한다.

3.1. 적합식 및 불순도 함수

임의의 노드 (arbitrary node) t에서의 조건부 분위수 함수 $q_\tau(\boldsymbol{x}, t)$를

$$q_\tau(\boldsymbol{x}, t) = \mathbf{x}^T \boldsymbol{\beta}_\tau(t) = \beta_{0,\tau}(t) + x_1\beta_{1,\tau}(t) + \cdots + x_p\beta_{p,\tau}(t) \tag{3.1}$$

와 같이 정의하자. 여기서 $\boldsymbol{\beta}_\tau(t)$는 노드 t에서 분위수 τ에 대한 회귀계수 벡터이다. 노드 t에서 추정된 선형 분위수 회귀모형에 대한 잔차는

$$e_{i,\tau}(t) = y_i - \mathbf{x}_i^T \hat{\beta}_\tau(t) \tag{3.2}$$

와 같으며, 본 연구에서 q차원 반응변수 벡터 $\boldsymbol{y} = (y_1, y_2, \cdots, y_q)^T \in R^q$를 고려하는 다변량 분위수 회귀모형에 대한 잔차를 자연스럽게

$$e_{ik,\tau_k}(t) = y_{ik} - \mathbf{x}_i^T \hat{\beta}_{k,\tau_k}(t) \text{ for } k = 1, 2, \cdots, q \tag{3.3}$$

와 같이 확장할 수 있다. 불순도 함수 $R(t)$는 식(3.3)과 체크 손실함수를 이용하여

$$R_\tau(t) = \sum_{i \in t} \sum_{k=1}^{q} \rho_{\tau_k}(e_{ik,\tau_k}(t)) \tag{3.4}$$

으로 정의할 수 있다. 불순도 감소량 $\Delta(t) = R_\tau(t) - \{R_\tau(t_L) + R_\tau(t_R)\}$로 계산되며 일반적으로 불순도 감소량이 가장 크도록 모형을 유도한다. 이때 반응변수는 모두 표준화함을 가정하고 t_L과 t_R은 각각 노드 t의 좌측, 우측 자식노드를 나타낸다.

3.2. 분할규칙 및 나무모형 크기 결정 방법

본 연구의 분할규칙은 GUIDE 알고리즘의 프레임을 빌려와 2단계로 분할규칙을 구분한다. 첫 번째 단계에서 완전 탐색 계산 방법 대신 통계적 유의성에 기반하여 분할변수를 선택하고, 두 번째 단계에서 선택된 분할변수에 대해서만 분할점 (또는 분할집합)을 탐색한다. 분할변수의 선택 알고리즘은 다음과 같다.

1단계 분위수 회귀모형을 적합한 뒤 잔차를 계산한다.

2단계 각 반응변수에 대한 잔차에 대해 -1과 1로 이루어진 부호벡터를 생성하여 이 벡터의 값을 행으로 하고, 분할후보변수 를 사전 설정된 기준에 의한 간격으로 구분하여 열로 설정한다. 이러한 두 변수에 대해 분할후보변수별 분할표를 생성한다.

3단계 2단계에서 생성한 분할표에 대해 카이제곱검정 (chi-square test) 결과 유의확률을 계산한다. 이때 Wilson-Hilferty 근사방법을 이용하여 자유도를 1로 조정한다.

4단계 3단계에서 얻은 유의확률 중 최소값을 갖는 분할후보변수를 분할변수로 결정한다.

이 알고리즘과 BC방법의 분할변수 선택 알고리즘과 구별점은 BC방법은 모든 잔차의 부호 조합에 대해 큰 분할표를 생성하는 반면, 이 알고리즘에서는 각 잔차의 부호에 대해 분할표를 반응변수별로 생성하여 독립성 검정을 수행하는 것이다. 가장 유의미한 변수를 분할변수로 선택하는 측면에서 본 연구에서는 이 방법을 MS (maximum significance)라 부르기로 한다.

Table 3.1은 세 개의 반응변수에 대한 부호벡터 $S_k(k = 1, 2, 3)$와 세 개의 분할변수 $Z_s(s = 1, 2, 3)$에 MS방법을 적용한 예제이다. Z_1은 어떤 반응변수와도 유의하지 않음을 알 수 있고, Z_2는 세 개의 반응변수와 적당하게 유의한 관계를 갖고 있다. 반면 Z_3는 첫 번째 반응변수와 매우 유의하며 다른 두 개의 반응변수와는 크게 유의한 관계가 없는 것으로 보인다. 또한 반응변수와 분할변수간 독립성 검정 결과 유의확률의 평균은 Z_2가 가장 유의하고 Z_3 이어서 Z_1 순서이다. 이 경우 MS방법은 가장 유의한 독립성 관계를 갖는 S_1과 Z_3에 주목하고 Z_3를 분할변수로 선택한다. 일반적인 회귀나무 모형이 재귀적 분할 (recursive partitioning) 과정을 통하여 하나의 노드가 두 개의 자식노드 (children nodes)만을 갖는 이진형 (binary) 나무모형임을 고려할 때 특정 분할시 가장 유의미한 분할변수를 선택하여 분할된 자료에 대해 회귀식으로 적합하는 것은 자연스러운 방법이다.

일반적으로 나무모형의 크기는 모형의 분산 (variance)을 결정하며 편향-분산 트레이드 오프 (bias-variance tradeoff)의 관계가 있다. 즉 나무의 크기가 너무 크면 과적합 (overfitting)되어 낮은 편향(bias)과 높은 분산을 가지며, 나무가 너무 성장하지 못하면 과소적합 (underfitting)된다. 나무모형의 주요 단점으로 알려진 과적합과 과소적합의 문제는 최적 나무모형의 크기를 결정함으로서 해결이 가능하다. 대표적인 나무모형의 크기 결정 알고리즘은 Breiman 등 (1984)이 제안한 가지치기 (pruning)로 나무모형을 최대한 성장시킨 뒤 교차검증 (cross-validation) 방법으로 나무모형의 크기를 결정하는 것이다. 이 방법의 우수성은 널리 알려져 있으나 계산량이 과도하게 많은 문제를 갖는다. 본 연구에서는 가지치기 방법과 정지규칙 (stopping rule)을 혼합하여 우수한 성능을 보인 Eo와 Cho (2014)의 M-단계 방법을 적용한다. 이것은 특정 노드에서 더 이상 유의한 향상이 없더라도 나무모형의 성장을 정지하지 않고 M-단계 후 상황까지 고려하여 결정하는 방법이다.

Table 3.1 Results for chi-square test between a sign vector S and split variable Z. The significant probabilities by approximating the degree of freedom to 1

	Z_1	Z_2	Z_3
S_1	0.88	0.09	0.02
S_2	0.45	0.12	0.29
S_3	0.65	0.11	0.18
Average	0.66	0.11	0.16

4. 모의실험

본 절에서는 나무모형의 분할변수 선택과 예측력에 대하여 기존의 BC방법과 제안하는 MS방법을 비교하고자 한다. 여기서 X_1과 X_2를 적합변수 (fitting variable)로 설정하고, 5개의 분할변수 Z_s ($s = 1, 2, \cdots, 5$)를 고려하였다. 적합변수 $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$이며 여기서 $N(0,1)$은 표준정규분포이다. 또한 $Z_1 \sim C_2$, $Z_2 \sim C_{12}$, $Z_3 \sim N(0,1)$, $Z_4 \sim \text{Exp}(1)$, $Z_5 \sim DU(1, 2, \cdots, 6)$이며 C_L은 순서가 없는 $1, 2, \cdots, L$에 대해 동일한 확률 $1/L$을 갖는 분포이며 $Exp(\lambda)$는 모수 (rate parameter)가 λ인 지수분포 (exponential distribution)이고 $DU(1, 2, \cdots, U)$는 집합 $\{1, 2, \cdots, U\}$에 대한 이산균등분포 (discrete uniform distribution)를 나타낸다. 특히 $DU(1, 2, \cdots, U)$와 C_L은 가능한 분할집합의 수에 있어서 매우 큰 차이가 있다. 가령 $U = L = 12$인 경우 $DU(1, 2, \cdots, 12)$에서는 $12 - 1$개가 가능한 분할집합의 수인 반면, C_{12}에서는 $2^{(12-1)} - 1$개가 가능한 분할집합의 수가 된다.

4.1. 분할변수 선택에 대한 모의실험

분할변수 선택에 대한 모의실험을 위해 최종노드에서 적합되는 선형모형을 각각 다르게 설정하여 한 번 분할되는 나무모형을 가정하였다. 여기서 분할변수의 종류를 연속형 및 범주형으로 구분하고, 영향받는 반응변수의 수를 한 개 또는 모든 반응변수로 제한하였다. 또한 반응변수의 수를 세 개와 다섯 개에 대해 실험을 실시하여 다음과 같은 총 8개의 모형에 대해 모든 모형에서 절편과 기울기가 일정하게 모두 다르도록 고려하였다. 모든 모형에서 오차항은 표준정규분포를 따른다. 여기서 $Z_1 \in \{A\}$는 Z_1이 순서가 없는 두 개의 범주를 갖는 변수이므로 이 범주변수임을 의미한다.

$$y_k = \begin{cases} X_1 + X_2 + \epsilon_k, & \text{if } Z_1 \in \{A\} \\ 0.5 + 0.5X_1 + 0.5X_2 + \epsilon_k, & \text{otherwise} \end{cases} \quad \text{for } k = 1, 2, 3. \tag{4.1}$$

$$y_k = \begin{cases} X_1 + X_2 + \epsilon_k, & \text{if } Z_3 \le 0 \\ 0.5 + 0.5X_1 + 0.5X_2 + \epsilon_k, & \text{otherwise} \end{cases} \quad \text{for } k = 1, 2, 3. \tag{4.2}$$

$$y_1 = \begin{cases} X_1 + X_2 + \epsilon_1, & \text{if } Z_1 \in \{A\} \\ 0.5 + 0.5X_1 + 0.5X_2 + \epsilon_1, & \text{otherwise} \end{cases}$$
$$y_k = X_1 + X_2 + \epsilon_k \quad \text{for } k = 2, 3. \tag{4.3}$$

$$y_1 = \begin{cases} X_1 + X_2 + \epsilon_1, & \text{if } Z_3 \le 0 \\ 0.5 + 0.5X_1 + 0.5X_2 + \epsilon_1, & \text{otherwise} \end{cases}$$
$$y_k = X_1 + X_2 + \epsilon_k \quad (4.4)$$
$$\text{for } k = 2, 3.$$

$$y_k = \begin{cases} X_1 + X_2 + \epsilon_k, & \text{if } Z_1 \in \{A\} \\ 0.5 + 0.5X_1 + 0.5X_2 + \epsilon_k, & \text{otherwise} \end{cases} \quad (4.5)$$
$$\text{for } k = 1, 2, \cdots, 5.$$

$$y_k = \begin{cases} X_1 + X_2 + \epsilon_k, & \text{if } Z_3 \le 0 \\ 0.5 + 0.5X_1 + 0.5X_2 + \epsilon_k, & \text{otherwise} \end{cases} \quad (4.6)$$
$$\text{for } k = 1, 2, \cdots, 5.$$

$$y_1 = \begin{cases} X_1 + X_2 + \epsilon_1, & \text{if } Z_1 \in \{A\} \\ 0.5 + 0.5X_1 + 0.5X_2 + \epsilon_1, & \text{otherwise} \end{cases}$$
$$y_k = X_1 + X_2 + \epsilon_k \quad (4.7)$$
$$\text{for } k = 2, 3, 4, 5.$$

$$y_1 = \begin{cases} X_1 + X_2 + \epsilon_1, & \text{if } Z_1 \le 0 \\ 0.5 + 0.5X_1 + 0.5X_2 + \epsilon_1, & \text{otherwise} \end{cases}$$
$$y_k = X_1 + X_2 + \epsilon_k \quad (4.8)$$
$$\text{for } k = 2, 3, 4, 5.$$

Table 4.1은 위에서 고려한 8개의 모형에 대한 분할변수에 대한 선택확률을 나타낸 것이다. 분할변수 선택의 난이도가 상대적으로 쉬운 Z_1의 경우 BC나 MS방법이 모두 우수한 성능을 나타내며 연속형 분할변수 Z_3에 대해서는 MS방법이 더 좋은 성능을 보였다. 반응변수 중 일부에만 영향을 미치는 경우에는 범주형이나 연속형임에 관계없이 MS방법이 더 정확한 분할변수를 선택함을 알 수 있다. 또한 반응변수의 개수가 많은 경우 본 연구에서 제안하는 MS방법의 우수성이 더욱 확연히 드러난다.

4.2. 예측력에 대한 모의실험

예측력을 비교하기 위한 모의실험을 위해 4.1절에서 설정한 동일한 분할변수에 대해 다음과 같이 나

Table 4.1 Estimated selection probabilities for split variable for Models (4.1) ∼ (4.8)

Model	BC					MS				
	Z_1	Z_2	Z_3	Z_4	Z_5	Z_1	Z_2	Z_3	Z_4	Z_5
(4.1)	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
(4.2)	0.03	0.04	0.82	0.06	0.05	0.03	0.02	0.90	0.03	0.02
(4.3)	0.85	0.04	0.03	0.03	0.05	0.90	0.02	0.03	0.04	0.01
(4.4)	0.08	0.09	0.59	0.10	0.14	0.05	0.04	0.79	0.05	0.07
(4.5)	0.99	0.00	0.00	0.01	0.00	0.99	0.00	0.00	0.01	0.00
(4.6)	0.03	0.09	0.72	0.06	0.10	0.02	0.01	0.95	0.02	0.00
(4.7)	0.53	0.10	0.05	0.11	0.21	0.89	0.02	0.02	0.04	0.03
(4.8)	0.13	0.15	0.39	0.07	0.26	0.06	0.02	0.75	0.06	0.11

무모형이 두 번 분할하여 최종노드가 세 개인 모형을 고려하였다.

$$y_j = \begin{cases} X_1 + X_2 + \epsilon_j, & \text{if } Z_1 \in \{A\} \text{ and } Z_3 \geq 0 \\ 0.5 + 0.5X_1 + 0.5X_2 + \epsilon_j, & \text{if } Z_1 \in \{B\} \\ -0.5 - 0.5X_1 - 0.5X_2 + \epsilon_j, & \text{otherwise} \end{cases} \tag{4.9}$$

for $k = 1, 2, 3$.

여기서 적합변수 $X_1 \sim U(-1, 1)$, $X_2 \sim U(-1, 1)$이며 U는 균등분포 (uniform distribution)를 나타내며 오차항은 $N(0, 0.5)$를 따르는 것으로 설정하였다. 훈련자료 (training data)의 크기는 100이며 동일한 조건에서 평가자료 (test data)를 1,000개 생성하여 식 (4.10)의 평균절대오차 (mean absolute error; MAE)를 이용하여 예측력을 비교하였다. 이 과정은 100회 독립적으로 반복 시행하였다.

$$\text{MAE} = \frac{1}{3}\sum_{k=1}^{3}\left(\frac{1}{1000}\sum_{r=1}^{|\tilde{T}|}\sum_{i\in\tilde{t_r}}\left|q_{\tau_k}(\boldsymbol{x}_i, \tilde{t_r}) - \hat{q}_{\tau_k}(\boldsymbol{x}_i, \tilde{t_r})\right|\right), \tag{4.10}$$

여기서 $|\tilde{T}|$는 최종노드 (terminal node)의 총 개수이며 $\tilde{t_r}$은 특정 최종노드이다.

Table 4.2는 BC와 MS방법의 예측력을 세 가지 분위수 수준에서 비교한 결과의 평균 및 표준오차 (standard error)이다. 실험을 실시한 모든 분위수에서 MS방법이 더 작은 MAE를 보였다. 특히 제 1, 3사분위수 (first, third quartile)에서 BC방법의 성능은 MS방법의 그것과 더욱 비교됨을 알 수 있다. 이러한 결과는 BC방법의 큰 분할표가 갖는 불안정성에 대한 것으로 설명할 수 있다.

Table 4.2 Mean absolute errors and standard errors for Model (4.9)

(τ_1, τ_2, τ_3)	Method	MAE (s.e.)
(0.25, 0.25, 0.25)	BC	1.055(0.002)
	MS	0.681(0.001)
(0.50, 0.50, 0.50)	BC	0.520(0.001)
	MS	0.485(0.001)
(0.75, 0.75, 0.75)	BC	0.917(0.003)
	MS	0.616(0.001)

5. 실증예제 : 시멘트 점성 및 강성 자료

본 절에서는 시멘트 점성 (viscosity)과 강성 (strength)에 대한 다변량 자료를 제안한 방법에 적용하였다. 이 자료는 UCI machine learning repository (Asuncion과 Newman, 2007)에서 다운로드가 가

능하다. 시멘트 점성과 강성에 대한 자료는 총 103개의 관측치가 있으며, 'cement, slag, fly ash, water, superplasticizer (SP), coarse aggregate (CA), fine aggregate (FA)'의 일곱 개 설명변수와 'slump, flow, strength'의 세 개 반응변수로 구성되어 있다. 반응변수 중 'slump'는 슬럼프 콘에 새로운 시멘트를 충전하고, 탈형했을 때 자중에 의해 변형되면서 상면이 밑으로 내려 앉는 양, 즉 새로운 시멘트의 유동성 정도를 표시하는 것이고 'flow'는 동일한 실험에서 새로운 시멘트가 좌우로 흩어지는 정도를 나타낸다. 본 절에서는 반응변수 세 개가 결합되어 설명변수와 어떤 관계가 있는지 확인하기 위하여 먼저 0.25, 0.5, 0.75분위수에 대해 각각 일변량 선형 분위수 모형과 일변량 분위수 회귀나무 모형을 적합한 결과를 살펴본다. 이어서 언급한 세 가지 분위수에 대해 본 연구에서 제안하는 다변량 분위수 회귀나무 모형의 결과와 비교한다.

Table 5.1 Summary of separate linear quantile regression models

	Slump		Flow		Strength	
$\tau = 0.25$	Estimate	p-value	Estimate	p-value	Estimate	p-value
(intercept)	-325.588	0.350	-559.166	0.357	198.445	0.099
cement	0.055	0.632	0.136	0.491	0.048	0.196
slag	0.079	0.613	0.104	0.704	-0.046	0.374
fly ash	0.055	0.617	0.15	0.441	0.031	0.441
water	0.586	0.097	1.156	0.048	-0.309	0.017
SP	0.029	0.965	0.619	0.595	-0.059	0.762
CA	0.124	0.364	0.188	0.429	-0.075	0.098
FA	0.115	0.404	0.187	0.454	-0.065	0.18
$\tau = 0.5$	Estimate	p-value	Estimate	p-value	Estimate	p-value
(intercept)	149.864	0.551	352.413	0.467	105.613	0.341
cement	-0.054	0.534	-0.161	0.33	0.073	0.044
slag	-0.096	0.450	-0.281	0.195	-0.011	0.83
fly ash	-0.061	0.501	-0.149	0.361	0.064	0.084
water	-0.032	0.892	0.214	0.642	-0.208	0.072
SP	-0.554	0.272	-1.241	0.214	0.21	0.258
CA	-0.065	0.511	-0.168	0.38	-0.042	0.316
FA	-0.043	0.667	-0.141	0.474	-0.026	0.57
$\tau = 0.75$	Estimate	p-value	Estimate	p-value	Estimate	p-value
(intercept)	-91.795	0.527	33.14	0.937	41.755	0.791
cement	0.046	0.373	0.015	0.917	0.09	0.088
slag	0.064	0.417	-0.026	0.903	0.013	0.862
fly ash	0.05	0.362	0.027	0.857	0.09	0.077
water	0.14	0.286	0.346	0.367	-0.131	0.407
SP	-0.084	0.783	-0.316	0.739	0.308	0.194
CA	0.037	0.516	-0.031	0.849	-0.02	0.741
FA	0.045	0.438	-0.023	0.893	0.001	0.987

Table 5.1은 시멘트 점성 및 강성 자료에 대한 일변량 선형 분위수 모형의 결과를 요약한 것이다. flow에 대해 water를 제외하고 점성을 나타내는 반응변수에 대해 유의한 설명변수를 찾을 수 없음을 알 수 있다. 또한 strength와는 cement, water가 유의한 변수로 보인다. 이러한 결과는 Loh와 Zheng (2013)이 일변량 회귀모형을 적용한 결과와 크게 다르지 않다. Figure 5.1은 일변량 분위수 회귀나무 모형을 각 설명변수와 세 가지 분위수에 대해 적합한 결과이다. 나무구조의 선형 분위수 모형을 해석하면 Table 5.1에서 제시한 선형 분위수 모형보다 풍부한 해석이 가능하다. 그러나 반응변수별로 상이한 나무모형에서 해석하는 것은 반응변수 세 개가 결합되어 설명변수와 어떤 관계가 있는지를 명확하게 설명하지 못하는 제한사항이 있다.

Figure 5.2는 MS방법을 세 가지 분위수에 대해 각각 적용한 결과이다. 시멘트의 점성 및 강성을 나타내는 세 가지 반응변수에 대해 water가 매우 중요함을 알 수 있다. 이것은 Loh와 Zheng (2013)의

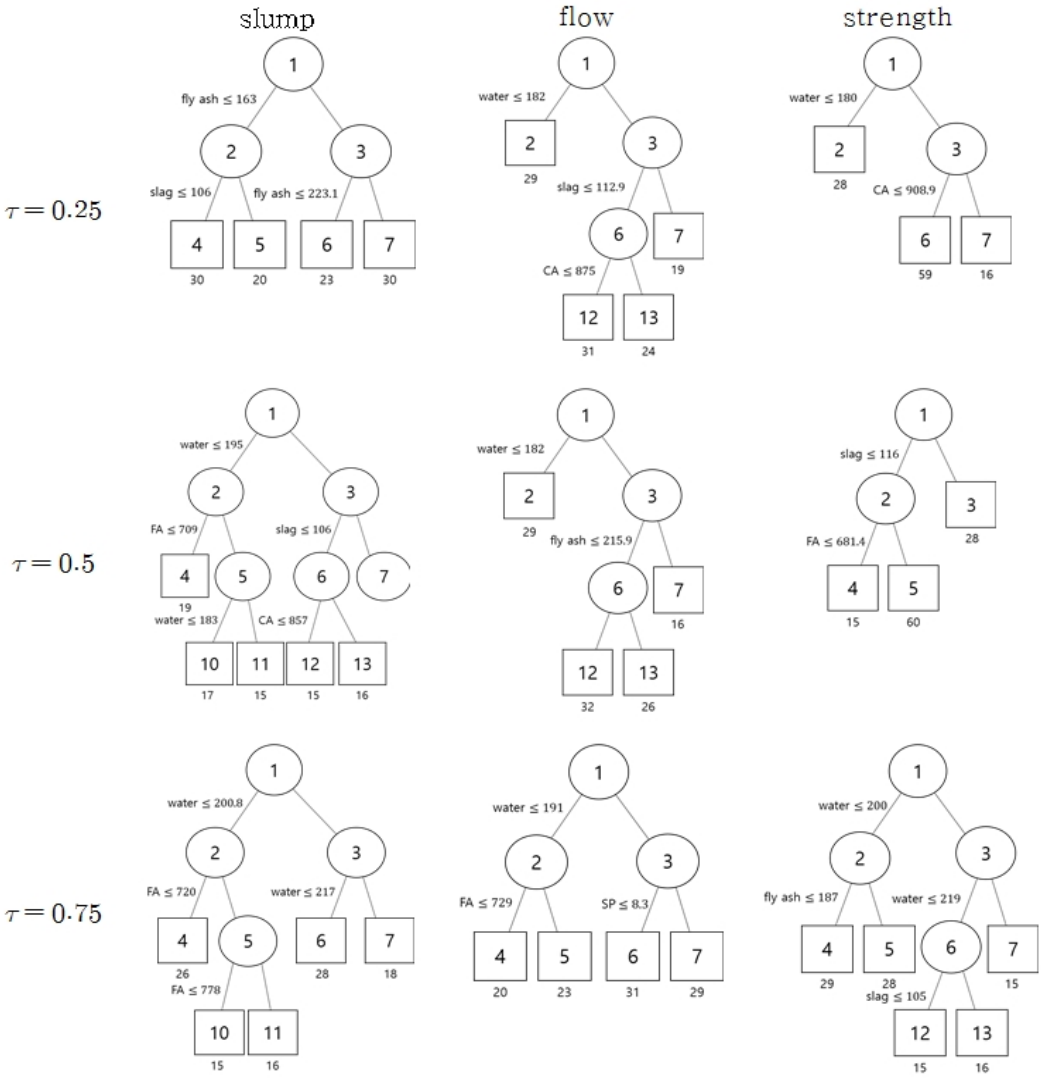

Figure 5.1 Univariate quantile regression trees for the concrete data at three quantile levels. An observation goes to the left node if the condition is satisfied at the intermediate nodes; otherwise, it goes to the right. Sample sizes are beneath. The numbers inside the circle are the node number.

평균에 대한 다변량 회귀나무 모형에서도 동일하게 나타난 결과이다. 그러나 다변량 분위수 회귀를 통해 다음과 같은 추가적인 해석이 가능하다. 상대적으로 낮은 점성과 강성보다 높은 점성과 강성에 대해 water는 더 중요한 의미를 가지는 것으로 보인다. 또한 시멘트의 점성과 강성이 하위 25% 미만인 경우 SP와 slag에 영향을 받는 것으로 보인다. 본 연구의 나무모형에서는 최종노드에서 조각별 선형(piecewise linear) 적합을 한다. 그러므로 Figure 5.2의 최종노드에서는 시멘트의 점성 및 강성을 나타내는 세 가지 반응변수와 일곱 개의 설명변수에 대한 각각 다른 적합식을 제공한다. 예를 들어 하위 25% 분위수에 대해 4번 최종노드에서는 water, SP, FA 등의 변수가 유의한 반면, 12번 최종노드에서는

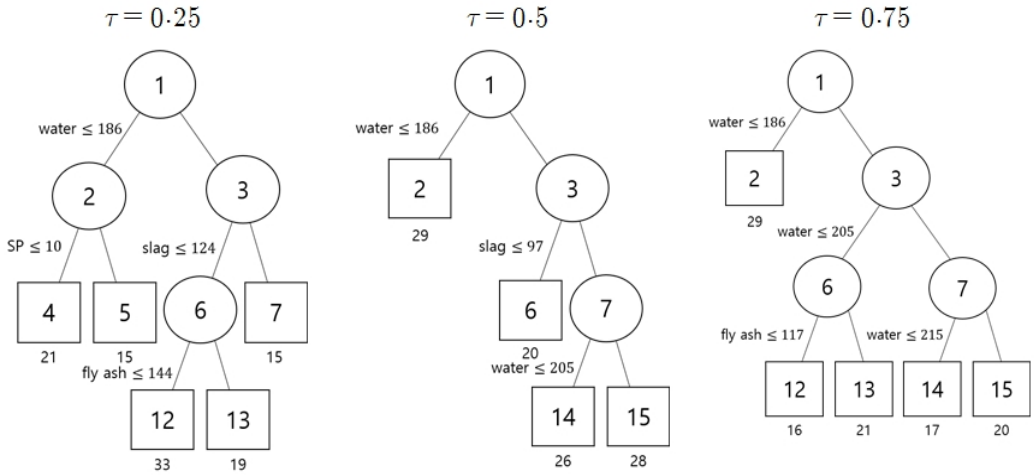

Figure 5.2 Multivariate quantile regression trees for the concrete data at three quantile levels using MS method

cement, fly ash 변수가 더욱 유의한 결과를 나타냈다.

6. 결론

분위수 회귀모형의 통계적 유용성은 이미 많은 실제 사례 적용과 확장 연구를 통해서 입증되었으나 여전히 선형 가정을 완화한 비모수적 분위수 회귀모형에 대한 요구는 지속되고 있다. 한편 실제 자료의 복잡성이 증가할수록 반응변수가 여러 개인 다변량 자료의 분석과 더불어 다양한 분위수에 대한 반응변수의 조건부 분포를 알고자 하는 필요성이 대두된다.

본 연구에서는 이러한 실정을 고려하여 다변량 분위수 회귀나무 모형을 제안하였다. 본 연구에서 제안한 방법은 일반적인 평균에 대한 회귀나무 모형보다 풍부한 해석이 가능한 분위수 회귀에 대한 것이다. 또한 기존의 다변량 회귀나무 모형의 단점을 보완한 새로운 분할변수 선택 알고리즘을 제안하였다. 본 연구에서 제안한 분할변수 선택 알고리즘의 우수성은 여러 가지 모형을 포함하는 모의실험과 실제자료의 분석을 통해 입증하였다. 결론적으로 본 연구에서 제안한 방법을 통하여 다변량 자료에 대한 보다 정확한 예측 결과와 풍부한 해석을 제공하는 데이터 마이닝이 가능하다.

References

Asuncion, A. and Newman, D. (2007). *UCI machine learning repository*, Available at `ttp://www.ics.uci.edu/~mlearn/MLRepository.`tml.

Breiman, L., Friedman, J., Stone, C. and Olshen, R. (1984). *Classification and regression trees*, Chapman & Hall/CRC, Belmont, CA.

Chaudhuri, P. and Loh, W. Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, **8**, 561-576.

De'ath, G. (2012). *Mvpart: multivariate partitioning*. R package version 1.6-0.

Eo, SH. and Cho, H. (2014). Tree-structured mixed-effects regression modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, **23**, 740-760.

Hallin, M., Lu, Z. and Yu, K. (2009). Local linear spatial quantile regression. *Bernoulli*, **15**, 659-686.

Richey, J. and Jung, K. H. (2014). Intergenerational economic mobility in Korea using a quantile regression analysis. *Journal of the Korean Data & Information Science Society*, **25**, 715-725.

Kim, H. and Loh, W. Y. (2011). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, **96**, 589-604.
Koenker, R. (2005). *Quantile regression*, Cambridge University Press, Cambridge, UK.
Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, **46**, 33-50.
Koenker R. and Mizera I. (2004). Penalized triograms: total variation regularization for bivariate smoothing. *Journal of the Royal Statistical Society B*, **66**, 145-163.
Li, Y., Liu, Y. and Zhu, J. (2007). Quantile regression in reproducing kernel hilbert spaces. *Journal of the American Statistical Association*, **102**, 255-268.
Liu Y. and Wu Y. (2011). Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *Journal of Nonparametric statistics*, **23**, 415-437.
Loh, W. Y. (2002). Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica*, **12**, 361-386.
Loh, W. Y. (2009). Improving the precision of classification trees. *The Annals of Applied Statistics*, **3**, 1710-1737.
Loh, W. Y. and Wei, Z. (2013). Regression trees for longitudinal and multiresponse data. *The Annals of Applied Statistics*, **7**, 495-522.
Loh, W. Y. and Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association*, **83**, 715-725.
Segal, M. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, **87**, 407-418.
Shim, J. Y. and Hwang, C. H. (2012). M-quantile kernel regression for small area estimation. *Journal of the Korean Data & Information Science Society*, **23**, 749-756.
Yu, K. and Jones, M. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, **93**, 228-237.
Zhang, H. (1998). Classification trees for multiple binary responses. *Journal of the American Statistical Association*, **93**, 180-193.
Zhang, H. and Ye, Y. (2008). A tree-based method for modeling a multivariate ordinal response. *Statistics and its Interface*, **1**, 169.

Journal of the Korean Data & Information Science Society
2017, **28**(3), 533–545

http://dx.doi.org/10.7465/jkdi.2017.28.3.533
한국데이터정보과학회지

Multivariate quantile regression tree[†]

Jaeoh Kim[1] · HyungJun Cho[2] · Sungwan Bang[3]

[12]Department of Statistics, Korea University
[3]Department of Mathematics, Korea Military Academy

Received 18 April 2017, revised 24 May 2017, accepted 27 May 2017

Abstract

Quantile regression models provide a variety of useful statistical information by estimating the conditional quantile function of the response variable. However, the traditional linear quantile regression model can lead to the distorted and incorrect results when analysing real data having a nonlinear relationship between the explanatory variables and the response variables. Furthermore, as the complexity of the data increases, it is required to analyse multiple response variables simultaneously with more sophisticated interpretations. For such reasons, we propose a multivariate quantile regression tree model. In this paper, a new split variable selection algorithm is suggested for a multivariate regression tree model. This algorithm can select the split variable more accurately than the previous method without significant selection bias. We investigate the performance of our proposed method with both simulation and real data studies.

Keywords: Data mining, multivariate data analysis, quantile regression, regression tree.

[†] This research was supported by 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 and by the Ministry of Education (NRF-2015R1D1A1A09058602) for H. Cho.

[1] Ph.D. candidate, Department of Statistics, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, 02841, Korea.

[2] Professor, Department of Statistics, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, 02841, Korea.

[3] Corresponding author: Associate professor, Department of Mathematics, Korea Military Academy, 574, Hwarang-ro, Nowon-gu, Seoul, Korea. E-mail: wan1365@gmail.com