Communications of the Korean Statistical Society Vol. 15, No. 4, 2008, pp. 633–642

벌점함수를 이용한 부분최소제곱 회귀모형에서의 변수선택

박종선[1], 문규종[2]

요 약

본 논문에서는 반응변수가 하나 이상이고 설명변수들의 수가 관측치에 비하여 상대적으로 많은 경우에 널리 사용되는 부분최소제곱회귀모형에 벌점함수를 적용하여 모형에 필요한 설명변수들을 선택하는 문제를 고려하였다. 모형에 필요한 설명변수들은 각각의 잠재변수들에 대한 최적해 문제에 벌점함수를 추가한 후 모의담금질을 이용하여 선택하였다. 실제 자료에 대한 적용 결과 모형의 설명력 및 예측력을 크게 떨어뜨리지 않으면서 필요없는 변수들을 효과적으로 제거하는 것으로 나타나 부분최소제곱회귀모형에서 최적인 설명변수들의 부분집합을 선택하는데 적용될 수 있을 것이다.

주요용어: 부분최소제곱 회귀모형; 설명변수선택; 벌점함수; 모의담금질.

1. 서론

화학분야와 식품분야에서 널리 사용되는 부분최소제곱법(partial least squares: PLS)은 두 변수집단간의 연관성을 잠재변수(또는 잠재벡터)를 통하여 모형화하는 일반적인 방법으로 Wold (1975)에 의하여 처음 소개되었다. 이 방법은 주성분 분석과 다중 회귀분석의 특징을 일반화한 방법으로 다중 설명변수와 다중 반응변수를 모형화할 수 있으며, 다중 반응변수를 가장 적절히 설명·예측하는 설명변수들의 선형 결합을 찾을 수 있다. 또한 설명변수들 간의 다중공선성을 처리할 수 있고, 잡음(Noise)이나 결측값이 있는 자료에도 로버스트(robustness)하며, 반응변수를 포함하는 교차타당성을 바탕으로 잠재설명변수들을 만들어 반응변수에 대한 정확한 예측을 할 수 있다.

최소제곱법은 주로 설명변수와 반응변수의 수가 많은 경우에 사용되며 이 경우에는 자료의 질도 중요한 문제이지만, 주어진 변수특성들이 부적절하거나 불필요할 때는 분석의 수행속도가 느려지며 많은 변수들이 모형에 포함되면 모형의 해석 측면에서도 어려움이 있다. 따라서 모형에 대한 해석을 보다 쉽게 하고 분석의 수행 속도를 좀 더 빠르게 하기 위해서 불필요한 변수들을 제거하는 작업이 필요할 것이다. 부분최소제곱법에 대한 설명변수들의 선택 문제는 기존의 최소제곱법을 사용하는 선형모형에서 널리 사용되는 전진, 후진, 및 단계선택법 등을 그대로 적용하거나 기타의 방법들을 부분최소제곱법에 적용할 수 있도록 변형한 방법들을 사용할 수 있다 (Gauchi와 Chagnon, 2001). 또한, 반응변수가 하

1) (110-745) 서울시 종로구 명륜동3가 53, 성균관대학교 통계학과, 교수.
교신저자: cspark@skku.edu
2) (110-745) 서울시 종로구 명륜동3가 53, 성균관대학교 통계학과, 대학원생.
E-mail: moongj97@nate.com

나인 회귀분석에 사용되는 PLS1에 대하여 다양한 측도들을 이용하여 반응변수와 상관관계가 높은 변수들을 선택하는 방법 (Höskuldsson, 2001)과 부분최소제곱 회귀에서 유전알고리즘을 이용하는 방법 (Leardi와 González, 1998) 등이 있으며 설명변수들과 함께 모형에 필요한 잠재변수들의 선택도 동시에 고려하는 방법 (Lazraq 등, 2003)도 있다. 그러나 지금까지 제시된 방법들은 반응변수가 하나인 경우에만 적용이 가능한 것들이 대부분이다.

최근에 Jolliffe 등 (2003)은 주성분분석에서 주성분들의 계수 중 0에 가까운 계수들을 제거하기 위해 벌점 함수(penalty function)를 추가하는 방법을 제안하였다. 이 방법은 벌점 함수가 추가되기 때문에 주성분으로 설명할수 있는 자료의 변동량이 줄어드는 문제가 있긴 하지만 불필요한 변수를 0으로 만들어 해석을 좀 더 쉽게 해주는 좋은 성질을 가지고 있다. 여러 종류의 제안된 벌점 함수 중에서 HARD 벌점 함수는 불편성을 유지하면서 필요 없는 계수를 0으로 만드는 좋은 성질을 가지고 있다.

본 논문에서는 주성분분석과 부분최소제곱회귀의 밀접한 연관성 (Frank와 Friedman, 1993)에 착안하여 부분최소제곱법에서 필요없는 설명변수들을 벌점함수를 추가하여 제거하는 방법을 고려하고자 한다. 더불어 HARD 벌점 함수의 기본 성질처럼 부분최소제곱회귀에서도 HARD 벌점함수를 추가 하였을 때 모형에 포함되는 변수들에 대한 점수들의 불편성을 유지하며 설명력과 예측력을 적절히 유지하는지 알아보고자 한다.

본 논문의 구성은 다음과 같다. 제 2절에서는 부분최소제곱법과 부분최소제곱회귀모형에 대한 내용을 포함하였으며 제 3절에는 벌점함수에 대한 소개와 더불어 벌점함수가 적용된 부분최소제곱 회귀문제를 해결하기 위한 방법으로 Metropolis-Hastings 알고리즘 (Metropolis 등, 1953)을 이용한 모의담금질(Simulated Annealing: Kirkpatrick 등, 1983) 방법에 대한 개념과 모의담금질 방법을 적용해 변수를 선택하는 알고리즘에 대하여 설명하였다. 벌점함수의 적용이 모형에 미치는 결과를 쉽게 알아 볼 수 있는 크기가 작은 자료와 설명변수의 수가 많은 실제적인 문제에 대한 적용 결과를 제 4절에 포함하였고 제 5절에서 실제 자료들에 대한 분석결과을 바탕으로 본 연구를 정리하였다.

2. 부분최소제곱법과 부분최소제곱 회귀모형

부분최소제곱법에서는 n개의 관측치와 q개의 반응변수로 이루어진 $n \times q$행렬 $\boldsymbol{Y}$와 n개 관측치에 대하여 수집된 p개의 설명(예측)변수로 이루어진 $n \times p$행렬 $\boldsymbol{X}$를

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}} + \boldsymbol{E}, \\ \boldsymbol{Y} &= \boldsymbol{U}\boldsymbol{Q}^{\boldsymbol{T}} + \boldsymbol{F} \end{aligned} \tag{2.1}$$

로 분해하게 된다. 여기서, $n \times k$행렬 $\boldsymbol{T}$와 $\boldsymbol{U}$는 추출된 k개의 잠재변수(latent variables)이며 $p \times k$행렬 $\boldsymbol{P}$와 $q \times k$행렬 $\boldsymbol{Q}$는 적재(loadings)행렬들을 그리고 나머지 $n \times p$행렬 $\boldsymbol{E}$와 $n \times q$행렬 $\boldsymbol{F}$는 잔차행렬들이 된다.

부분최소제곱법은 비선형 반복 알고리즘으로

$$\begin{aligned} &\max_{w,c}[\,\mathrm{corr}(\boldsymbol{X}\boldsymbol{w}, \boldsymbol{Y}\boldsymbol{c})\,]^2 = \max_{w,c}[\,\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}\boldsymbol{c}\,] = \max_{t,u}[\,\boldsymbol{t}^{\boldsymbol{T}}\boldsymbol{u}\,], \\ &\text{여기서 } \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{w} = \boldsymbol{1} \text{이고 } \boldsymbol{t}^{\boldsymbol{T}}\boldsymbol{t} = \boldsymbol{1} \end{aligned} \tag{2.2}$$

인 $\boldsymbol{w}$와 $\boldsymbol{c}$를 순차적으로 구하게 되며 $\boldsymbol{t}$는 $\boldsymbol{Xw}$이고 $\boldsymbol{u}$는 $\boldsymbol{Yc}$이다. 식 (2.1)을 이용하여 $\boldsymbol{P}$와 $\boldsymbol{Q}$를 구성하는 적재벡터 $\boldsymbol{p}$와 $\boldsymbol{q}$들은 각각 $\boldsymbol{X}$와 $\boldsymbol{t}$, $\boldsymbol{Y}$와 $\boldsymbol{u}$의 선형회귀계수들로 구성되며 $\boldsymbol{p} = \boldsymbol{X^T t}/(\boldsymbol{t^T t})$, $\boldsymbol{q} = \boldsymbol{Y^T u}/(\boldsymbol{u^T u})$가 된다.

위의 극대화 과정은 구해진 $\boldsymbol{w}$를 통하여 다음 반복에서의 반응변수와 설명변수를

$$\begin{aligned} \boldsymbol{X} &\leftarrow \boldsymbol{X} - \boldsymbol{tp^T}, \\ \boldsymbol{Y} &\leftarrow \boldsymbol{Y} - \boldsymbol{tt^T Y}/(\boldsymbol{t^T t}) \end{aligned} \tag{2.3}$$

와 같이 감쇠(deflation)시킨 후 순차적으로 이루어 진다. 부분최소제곱법에는 반응변수와 설명변수들에 대한 감쇠방법에 따라 다양한 알고리즘이 존재하는데 회귀분석을 목적으로 하는 경우 식 (2.3)에서와 같은 감쇠방법을 사용하는 것이 일반적이다. 이 알고리즘을 특별히 PLS1과 PLS2라 하며 PLS1은 반응변수가 하나인 경우이고 PLS2는 반응변수가 2개 이상인 경우를 의미한다.

회귀분석을 목적으로 하는 PLS1과 PLS2에서는 일반적인 PLS에 다음의 두가지 가정을 추가한 것이다.

1. 점수벡터 $\boldsymbol{t}$들은 $\boldsymbol{Y}$에 대한 설명변수들이며
2. 점수벡터들 $\boldsymbol{t}$와 $\boldsymbol{u}$사이에는 $\boldsymbol{U} = \boldsymbol{TD} + \boldsymbol{H}$의 선형관계가 성립한다.

위에서 $\boldsymbol{D}$는 $q \times q$인 대각행렬이고 $\boldsymbol{H}$는 잔차행렬이 된다. 추가된 가정들을 PLS 모형 식 (2.1)에 적용하면

$$\boldsymbol{Y} = \boldsymbol{TDQ^T} + (\boldsymbol{HQ^T} + \boldsymbol{F}) = \boldsymbol{TC^T} + \boldsymbol{F^*} \tag{2.4}$$

와 같이 표현할 수 있다. 여기서 $\boldsymbol{C^T} = \boldsymbol{DQ^T}$이며 이는 $p \times q$인 회귀계수 행렬이고 $\boldsymbol{F^*} = \boldsymbol{HQ^T} + \boldsymbol{F}$는 잔차행렬이 된다. 위의 식 (2.4)는 반응변수 $\boldsymbol{Y}$를 최소제곱법을 이용하여 서로 직교인 설명변수 $\boldsymbol{T}$로 분해한 것과 같다.

다음 절에서는 부분최소제곱법을 위한 최적화 문제 (2,2)에 설명변수들의 가중치 $\boldsymbol{w}$에 대한 벌점을 추가하여 각각의 잠재변수에서 필요하지 않은 설명변수들을 제거하는 방법에 대하여 살펴보기로 한다.

3. 벌점함수를 적용한 부분최소제곱 회귀모형

앞 절에서 살펴본 부분최소제곱회귀의 목적함수에 벌점함수를 추가하는 과정을 살펴보자. 가중치 $\boldsymbol{w}$에 대한 벌점함수를 $p_\lambda(\boldsymbol{w})$라고 하고 이를 앞 절에서의 목적함수 (2.2)에 추가하면

$$\max_{w,c} \left[\boldsymbol{w^T X^T Y c} + n\Sigma_{j=1}^{p} p_\lambda(|w_j|) \right], \quad \text{여기서 } \boldsymbol{w^T w} = \boldsymbol{1} \text{이고 } \boldsymbol{t^T t} = \boldsymbol{1} \tag{3.1}$$

이 된다. 여기서 w_j는 잠재변수 $\boldsymbol{w}$의 j번째 원소이다. 첫 번째의 잠재변수를 구한 후 순차적인 잠재변수들은 앞의 식 (2.3)에서와 동일한 방법으로 감쇠된 후 같은 최적화 과정을 적

용하여 구할 수 있다. 본 논문에서 고려할 HARD 벌점함수를 포함한 대표적인 벌점함수들은 다음과 같다.

L_p 벌점함수: $p_\lambda(|w_j|) = \lambda|w_j|^p$, $p = 1$인 경우 LASSO(Least Absolute Shrinkage and Selection Operator: Tibshirani, 1996)가 됨

HARD 벌점함수: $p_\lambda(w_j) = \lambda^2 - (|w_j| - \lambda)^2 I(|w_j| < \lambda)$

SCAD 벌점함수:

$$p_\lambda(w_j) = \begin{cases} \lambda w_j, & \text{if } w_j < \lambda, \\ -\dfrac{{w_j}^2 - 2a\lambda w_j + \lambda^2}{2(a-1)}, & \text{if } \lambda \le w_j < a\lambda, \\ \dfrac{(a+1)\lambda^2}{2}, & \text{if } w_j \ge a\lambda. \end{cases}$$

Fan과 Li (2001)는 바람직한 벌점함수가 가져야 할 성질들로 불편성, 희박성(Sparsity) 그리고 연속성을 꼽았으며 일반화선형모형 등을 포함하는 회귀모형에 가장 적절한 벌점함수로 SCAD(Smoothly Clipped Absolute Deviation)를 제시하였다. 하지만 앞의 벌점함수들 중 이 세 가지의 성질들을 모두 만족하는 벌점함수는 없으며 L_p 및 SCAD 벌점함수는 식 (2.2)와 같이 파라미터들에 대한 제약조건 $\boldsymbol{w^T w = 1}$이 있는 최적화 문제에 적용할 경우 편의가 발생하게 된다. 반면, HARD 벌점함수는 연속이지는 않으나 불편성과 희박성을 만족하며 넓은 범위의 λ값에 대하여 파라미터 추정치의 편의가 거의 없었으며 SCAD 벌점함수와 같이 2개인 조율 파라미터 a와 λ의 최적해를 찾아야 하는 문제점도 없어 향후 HARD 벌점함수를 적용한 경우만을 다루기로 한다. 식 (3.1)의 최적화 문제의 해결에는 Kirkpatric 등 (1983)에 의하여 소개된 다변량 함수에 대한 몬테칼로 최적화 방법인 모의담금질을 사용할 수 있다. 모의담금질 기법은 판매원여행문제 등의 조합최적화 문제에 효과적인 것으로 알려져 있으며 연속형인 다변량 인수들을 가지는 최적화 문제에도 적용이 가능한 것으로 알려져 있다. 이 기법은 다변량 최적화를 위한 목적함수를 온도 파라미터를 포함하는 Boltzman 분포로 변환한 후 이 온도를 서서히 낮추면서 주변의 해들을 탐색하는 알고리즘으로 존재할 수 있는 국소 최적해에 빠지지 않고 전역 최적해에 가까운 해를 얻을 수 있는 알고리즘이다.

모의담금질 기법을 적용하기 위하여 벌점함수가 포함된 목적함수 (3.1)을 포함하는 다음과 같은 분포를 가정하자.

$$u(\boldsymbol{w}) = C \exp\left(\frac{1}{\gamma} D(\boldsymbol{w})\right), \tag{3.2}$$

여기서 $D(\boldsymbol{w}) = \boldsymbol{w^T X^T Y c} + n\Sigma_{j=1}^{p} p_\lambda(|w_j|)$이며 C는 정규화상수이고 γ는 보통 온도라고 불리는 파라미터이다. 첫 번째 잠재변수를 구하는 알고리즘은 다음과 같다.

STEP1: 해의 초기화

1. $\boldsymbol{w}$에 대한 적절한 초기값을 설정
2. $k = 0$(단계값)
3. 초기 온도 γ_0 설정($k = 0$ 단계에서의 온도)
4. 초기 반복회수 L_0 설정($k = 0$에서의 표본수)
5. $\gamma = \gamma_0$ 설정

STEP2: 수렴할 때까지 다음을 반복

1. $l = 1$부터 L_k까지
   (a) $\boldsymbol{w}^{\text{old}}$의 이웃에서 $\boldsymbol{w}^{\text{new}}$를 설정
   (b) 만일 $D(\boldsymbol{w}^{\text{new}}) \leq D(\boldsymbol{w}^{\text{old}})$이면 $\boldsymbol{w}^{\text{old}} = \boldsymbol{w}^{\text{new}}$
   (c) 만일 $D(\boldsymbol{w}^{\text{new}}) > D(\boldsymbol{w}^{\text{old}})$이고 $\exp([D(\boldsymbol{w}^{\text{old}}) - D(\boldsymbol{w}^{\text{new}})]/\gamma_k) > U[0,1]$이면 $\boldsymbol{w}^{\text{old}} = \boldsymbol{w}^{\text{new}}$
2. $k = k + 1$
3. $\gamma_k = \gamma_0 \times (0.9)^k$

출력: $\boldsymbol{w}^{\text{old}}$

최적해를 구하는 위의 전 과정을 모의담금질 방법이라고 하며 STEP2에서 기존 해의 이웃에서 새로운 해를 탐색하여 (b) 또는 (c)의 조건에 따라 이를 채택 또는 기각하는 과정을 Metropolis-Hastings 알고리즘이라 한다. 이 알고리즘은 최적화를 위한 목적함수를 포함하는 분포 (3.2)에서 난수를 구하는 과정이다. 모든 알고리즘은 R을 이용하여 구현하였다.

모의담금질이 적절히 수행되기 위해서는 적절한 파라미터들 및 초기값의 설정이 매우 중요하다. 목적함수에서는 $\boldsymbol{w}$에 대한 초기값과 다음의 반복을 위한 이웃들에 대한 선택이 주의깊게 이루어져야 하며 $\boldsymbol{w}$에 대한 초기값으로는 벌점함수가 포함되지 않은 부분최소제곱문제의 해를 초기값으로 사용할 수 있다. 각각의 잠재변수에 대한 최적의 λ값은 문제에 따른 목적함수의 크기에 의존하는데 목적함수의 값은 두 변수집단간의 공분산에 의존하는 동시에 잠재변수의 수가 추가될 때마다 $\boldsymbol{Y}$와 $\boldsymbol{X}$에 대한 감쇠가 수행되므로 점점 더 작아지게 된다. 따라서 모든 문제에 일반적으로 적용할 수 있는 최적값의 범위 등을 제시할 수 없으며 주어진 문제와 잠재변수의 수에 따라 시행착오를 거치면서 최적해를 탐색하여야 한다. 최적의 λ값은 검정자료에서 반응변수에 대한 예측값들과 실제값들의 차이를 최소화하는 교차타당성 방법 등을 사용하여 구할 수 있으나 본 연구에서는 최적해를 탐색하지 않고 몇 개의 값에 대한 결과만을 포함하였다. 기타 모의담금질 알고리즘에서 파라미터들의 결정과 관련한 상세한 내용은 Aarts와 Korst (1989) 등에서 찾을 수 있다.

4. 자료분석

이 절에서는 두 개의 실제 자료에 대하여 벌점함수를 이용한 변수선택 방법을 적용하여 보고 그 결과를 살펴보기로 한다. 예제 4.1에서는 벌점함수의 적용에 따라 선택되는 변수

표 4.1: 반응변수 Y 행렬

Wine	Hedonic	Goes with meat	Goes with dessert
1	14	7	8
2	10	7	6
3	8	5	5
4	2	4	7
5	6	2	4

표 4.2: 설명변수 X 행렬

Wine	Price	Sugar	Alcohol	Acidity
1	7	7	13	7
2	4	4	14	7
3	10	5	12	5
4	16	7	11	3
5	13	3	10	3

표 4.3: 벌점함수 적용전과 적용후(괄호안)의 W 행렬

	w_1	w_2
Price	−0.5137(−0.5030)	−0.3379(−0.2327)
Sugar	0.2010(0.0000)	−0.9400(−0.9724)
Alcohol	0.5705(0.5786)	−0.0188(−0.0128)
Acidity	0.6085(−0.6420)	0.0429(0.0000)

들과 이들이 모형에 미치는 결과를 쉽게 일아볼 수 있도록 크기가 작은 자료를 사용하였다. 예제 4.2에서는 설명변수의 수가 상대적으로 많은 실제적인 문제에 대하여 제시한 방법의 효과를 살펴보았다.

예제 4.1 예제에서 사용된 자료 (Abdi, 2003)는 반응변수가 3개, 설명변수가 4개이며 관측치의 수는 5개인 작은 크기의 자료이다. 예측을 위한 반응변수는 호감도(Hedonic), 육류와의 어울림(Goes with meat) 그리고 디저트와의 어울림(Goes with dessert)에 대한 주관적 평가점수로 표 4.1에 포함하였으며 설명변수는 가격(Price), 당도(Sugar), 알코올도수(Alcohol) 그리고 산도(Acidity)이며 표 4.2에 포함하였다.

첫 번째 잠재변수에서는 $\lambda = 0.9$인 경우 당도(Sugar) 변수가 제거 되었으며 표 4.3의 왼쪽 열에 벌점함수 적용 전과 후의 점수벡터값들을 포함하였다. 제거된 당도 변수 이외에 다른 변수들에 대한적재값은 큰 차이가 없어 편의가 크지 않음을 알 수 있다. 첫 번째 잠재변수에서 당도를 제거한 후 두 번째 잠재변수에서는 산도(Acidity) 변수가 제외되었음을 볼 수 있다. 이 경우 λ 값은 0.025를 사용하였으며 두 잠재변수에서 하나씩의 설명변수를 제거한 후 반응변수와 설명변수에 대한 분산의 설명비율이 제거하기 전의 비율과 큰 차이가 없음을 표 4.4를 통하여 알 수 있다.

표 4.4: 벌점함수 적용전과 적용후(괄호안) X 및 Y의 분산에 대한 설명비율(%)

잠재변수	설명된 X의 분산	설명된 X의 누적분산	설명된 Y의 분산	설명된 Y의 누적분산
1	70.45(71.83)	70.45(71.83)	61.00(59.34)	61.00(59.34)
2	27.90(26.51)	98.35(98.35)	24.27(25.91)	85.26(85.25)

표 4.5: 벌점함수 적용전과 적용후(괄호안) X 및 Y의 분산에 대한 설명비율(%)과 검정자료에 대한 RMSE

잠재변수	λ	포함된 변수의 개수	RX^2(%)	RY^2_{adj}(%)	RMSE(검정자료)
1	0.0	117	13.78	27.39	-
	10.0	81	13.43	28.37	-
	15.0	60	13.07	29.37	-
	20.0	50	12.47	31.05	-
2	0.0	117	8.40	20.48	0.7961
	5.0	74	7.41	28.43	0.7382
	10.0	39	6.93	28.97	0.7147

예제 4.2 두 번째 예제에서 사용한 자료 (Gauchi와 Chagnon, 2001)는 종이 코팅용으로 사용되는 라텍스(latex)의 제조 공정에서 얻어진 것으로 설명변수들은 온도, 시간, 단위체 입력율, 촉매수준, 반응농도 등으로 모두 117개로 구성되어 있다. 반응변수는 전문가들이 일반적으로 "불용성 부산물"이라 부르는 이차 부산물의 양이다. 전문가들은 위의 다양한 설명변수들을 이용하여 부산물의 양을 적절히 설명하는 모형을 찾아 궁극적으로 최종 반응 후에 남는 부산물의 양을 최소화하는 것이 목적이다.

이와 같은 반응에 있어서의 어려움은 일반적으로 상관관계가 높은 많은 수의 설명변수들이 포함되어 있어 예를 들면 특정한 온도의 수준과 이 수준에서의 지속시간을 독립적으로 조정할 수 없다는 점이다. 따라서 부산물의 양을 최소화하는 최적의 조합을 찾기 위하여 가장 간단한 모형을 찾는 것이 필요하게 된다. 자료의 표본수는 262개이고 이 중 무작위로 추출된 210개를 훈련자료로 하고 나머지 52개는 검정자료로 사용하였다.

표 4.5의 결과를 살펴보면 우선 첫 번째 잠재변수의 경우 λ값이 증가할수록 제거되는 변수의 수가 늘어나는 것을 볼 수 있으며 λ가 20인 경우 50개의 변수만이 모형에 포함되었다. 이 경우 설명변수에 대한 설명부분은 감소하지만 반응변수에 대한 설명부분은 수정된 R^2의 경우 증가하는 것을 볼 수 있다. λ가 20보다 큰 경우에는 수정된 R^2 또한 감소하므로 20을 최종 λ값으로 사용하였다.

두 번째 잠재변수의 경우에는 λ가 10인 경우 78개의 설명변수가 제거 되었으며 이 때 검정자료에 대한 RMSE는 약 72%로 모든 변수를 포함하였을 경우에 비교하여 약 8%정도 줄어드는 것으로 나타났다.

Gauchi와 Chagnon (2001)은 제외된 변수의 수에 의존하는 Stone-Geisser 지수 (Geisser, 1974; Stone, 1974)와 비슷한 기준을 제시하였으며 이를 라텍스 자료에 적용한 결과 24개의 변수를 포함하는 모형을 최적모형으로 제시하였다. 본 연구에서 적용한 방법은

기존의 방법들과 달리 모든 잠재변수 즉, 전체 모형에 대한 변수선택 방법이 아닌 각각의 잠재변수에 대한 변수선택의 의미를 가지므로 직접 비교하는 것은 무리이나 Gauchi와 Chagnon (2001)의 결과와 비교하여 상대적으로 많은 수의 변수들이 모형에 포함된 것을 알 수 있다.

라텍스 자료에 대한 결과에서 자료에 포함된 설명변수들에 대한 상세한 내용을 알 수 없어 잠재변수들에 대하여 변수선택에 따른 의미를 부여할 수는 없었지만 제시된 방법의 결과는 각각의 잠재변수에 포함되는 변수들의 특성을 분석하여 의미를 부여하는데 유용하게 사용될 수 있을 것이다. 또한 모든 잠재변수에서 동시에 제외된 변수들은 최종 모형에서 제외하여 전체 모형에 대한 변수선택 방법으로 활용하는 것도 가능할 것이다.

5. 결론

본 연구에서는 부분최소제곱 회귀모형에 벌점함수를 적용하여 각각의 잠재변수별로 모형에 필요 없는 설명변수들을 제거하는 방법에 대하여 살펴보았다. 두 개의 실제 자료에 대하여 제시한 방법을 적용하면 반응변수 및 설명변수에 대한 분산의 설명력과 검정자료에 대한 예측력이 크게 떨어지지 않는 범위에서 각각의 잠재변수에 대한 최적의 설명변수 집합을 선택할 수 있음을 보았다. 제시한 방법은 앞에서 언급한 설명변수선택 방법들에서와 같이 모든 잠재변수에 동일하게 적용되는 설명변수들의 집합을 선택하는 방법이 아니므로 전체 모형에 필요한 변수들의 집단을 탐색하는 경우에는 적합하지 않다. 하지만 각각의 잠재변수들에서 공통적으로 포함되지 않는 변수들을 제거하는 방법을 통하여 어느 정도 전체 모형에 대한 변수선택의 의미를 가질 수 있도록 변형하여 사용할 수 있을 것이다. 따라서 본 논문에서 제시하는 방법은 전체 모형에 대한 변수선택 보다는 주성분 분석에서의 적용예 (Jolliffe 등, 2003)의 경우와 비슷하게 각각의 잠재변수에서 필요없는 변수를 제거함으로써 각 잠재변수들에 대한 의미의 해석 등을 명확하게 하는데 도움이 될 것으로 사료된다.

제시하는 방법은 부분최소제곱 회귀모형이 아닌 일반적인 최소제곱법의 경우 두 집단의 변수의 개수가 모두 2개 이상인 경우 $\boldsymbol{w}$와 $\boldsymbol{c}$에 벌점함수를 동시에 적용하여 두 변수집단에서의 변수선택문제로 확장하여 적용할 수 있을 것이다. 또한 결론의 앞부분에서 언급한 것처럼 각각의 모형에 포함되는 잠재변수들에서 공통적으로 포함되지 않는 설명변수들만을 제거하고 분석을 수행하는 경우에 설명력 및 예측력 등에 어떤 영향을 미치는지에 대한 추가적인 연구도 필요한 것으로 판단된다.

참고문헌

Aarts, E. and Korst, J. (1989). *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing,* John Wiley & Sons, New York.

Abdi, H. (2003). Partial least squares(PLS) regression, In Lewis-Beck M., Bryman, A. and Futing, T. (eds.), *Encyclopedia of Social Sciences Research Methods,* Thousand Oaks (CA): Sage.

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.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics*, **35**, 109–135.

Gauchi, J. P. and Chagnon, P. (2001). Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data, *Chemometrics and Intelligent Laboratory Systems*, **58**, 171–193.

Geisser, S. (1974). A predictive approach to the random effect model, *Biometrika*, **61**, 101–107.

Höskuldsson, A. (2001). Variable and subset selection in PLS regression, *Chemometrics and Intelligent Laboratory Systems*, **55**, 23–38.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003). A modified principal component technique based on the lasso, *Journal of Computational and Graphical Statistics*, **12**, 531–547.

Kirkpatrick, S., Gelatt, C. D. Jr. and Vecchi, M. P. (1983). Optimization by simulated annealing, *Science*, **220**, 671–680.

Lazraq, A., Cléroux, R. and Gauchi, J. P. (2003). Selecting both latent and explanatory variables in the PLS1 regression model, *Chemometrics and Intelligent Laboratory Systems*, **66**, 117–126.

Leardi, R. and González, A. L. (1998). Genetic algorithms applied to feature selection in PLS regression: How and when to use them, *Chemometrics and Intelligent Laboratory Systems*, **41**, 195–207.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**, 1087–1092.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society, Series B*, **36**, 111–147.

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

Wold, H. (1975). Path models with latent variables: The NIPALS approach, In H.M. Blalock *et al.*, *Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building*, pages 307–357, Academic Press, New York.

[2008년 6월 접수, 2008년 7월 채택]

Variable Selection in PLS Regression with Penalty Function

Chongsun Park[1], Guy Jong Moon[2]

Abstract

Variable selection algorithm for partial least square regression using penalty function is proposed. We use the fact that usual partial least square regression problem can be expressed as a maximization problem with appropriate constraints and we will add penalty function to this maximization problem. Then simulated annealing algorithm can be used in searching for optimal solutions of above maximization problem with penalty functions added. The HARD penalty function would be suggested as the best in several aspects. Illustrations with real and simulated examples are provided.

Keywords: PLS regression; penalty function; variable selection; simulated annealing.

1) Professor, Department of Statistics, Sungkyunkwan University, 3-53 Myungnyun-Dong, Jongno-Gu, Seoul 110-745, Korea. Correspondence: cspark@skku.edu

2) Graduate Student, Department of Statistics, Sungkyunkwan University, 3-53 Myungnyun-Dong, Jongno-Gu, Seoul 110-745, Korea. E-mail: moongj97@nate.com