Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 1, pp. 51 ~ 61

Combining Ridge Regression and Latent Variable Regression

Jong-Duk Kim[1)]

Abstract

Ridge regression (RR), principal component regression (PCR) and partial least squares regression (PLS) are among popular regression methods for collinear data. While RR adds a small quantity called ridge constant to the diagonal of $\boldsymbol{X}'\boldsymbol{X}$ to stabilize the matrix inversion and regression coefficients, PCR and PLS use latent variables derived from original variables to circumvent the collinearity problem. One problem of PCR and PLS is that they are very sensitive to overfitting. A new regression method is presented by combining RR and PCR and PLS, respectively, in a unified manner. It is intended to provide better predictive ability and improved stability for regression models. A real-world data from NIR spectroscopy is used to investigate the performance of the newly developed regression method.

Keywords : Partial Least Squares Regression, Principal Component Regression, Ridge Partial Least Squares Regression, Ridge Principal Component Regression, Ridge Regression

1. 서론

예측변수들 사이에 높은 다중공선성이 존재하면 회귀계수의 보통 최소제곱회귀(ordinary least squares regression, OLS) 추정은 매우 불안정해지고 예측의 정확도가 떨어진다. 또한 예측변수수가 샘플수보다 커지면 일반화역행렬에 의존해야 한다. 이런 문제를 극복하기 위해 개발된 대표적인 방법은 능형회귀(ridge regression, RR)로 추정된 회귀계수의 안정화를 도모하는 것이 주초점이다. 또한 잠재인자를 이용하는 주성분회귀(principal component regression, PCR)와 부분최소제곱회귀(partial least squares regression, PLS) 등이 개발되었다. PCR에서는 예측변수들에서의 설명된 공분산을 최대화하며 PLS에서는 예측변수들과 반응변수 간의 설명된 공분산을 최대화

1) 부산시 남구 우암동 산 55-1, 부산외국어대학교 응용통계학과 교수
E-mail : jdkim@pufs.ac.kr

하는 잠재인자를 구하고 이들을 예측변수로 사용하는 방식이다. 본 논문에서는 RR과 PCR의 각 장점을 결합한 회귀를 먼저 설명하고 이 방식을 연장하여 RR과 PLS를 결합한 새로운 회귀 방법을 보인다. 높은 다중공선성이 존재하는 대표적인 자료인 NIR (Near-infrared) 분광 데이터를 이용하여 이 회귀방법의 예측력을 다른 회귀방법들과 비교한다.
다음의 일반 선형회귀모형을 고려한다.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

여기서 $\boldsymbol{y}$는 크기가 $n \times 1$인 반응변수 벡터, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$은 모형의 미지 모수, $\boldsymbol{X}$는 크기가 $n \times p$인 예측변수행렬, $\boldsymbol{\epsilon}$은 크기가 $n \times 1$인 오차벡터이다. $\boldsymbol{X}$와 $\boldsymbol{y}$의 각 열은 중심화 또는 표준화된 것으로 간주한다. 논의의 편의상 $\boldsymbol{X}$는 완전열계수로 가정한다. 이때 행렬 $\boldsymbol{X}$의 비정칙값분해(SVD, singular value decomposition)는 $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$으로 표현될 수 있으며 여기서 $\boldsymbol{U} = (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_p)$는 열들이 $\boldsymbol{X}\boldsymbol{X}'$의 고유벡터로 구성된 행렬 $(n \times p)$, $\boldsymbol{V} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_p)$는 열들이 $\boldsymbol{X}'\boldsymbol{X}$의 고유벡터로 구성된 행렬$(p \times p)$, $\boldsymbol{D} = \mathrm{diag}(\delta_1, \cdots, \delta_p)$는 p개의 양의 비정칙값으로 구성된 대각행렬$(p \times p)$이다. 따라서 $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}$, $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}' = \boldsymbol{I}$이 성립한다.

2. 능형회귀와 주성분회귀의 결합

먼저 RR과 PCR의 장점을 결합하는 방법을 보도록 한다. 최소제곱법을 이용한 회귀추정량 $\hat{\boldsymbol{\beta}}_{ols} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$을 $\boldsymbol{X}$의 SVD를 이용하면

$$\hat{\boldsymbol{\beta}}_{ols} = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}'\boldsymbol{y}$$

가 되며, 이것은 p개 양의 비정칙값을 전부 사용한다는 의미이다. 예측변수들이 직교성에서 멀어질수록 행렬 $\boldsymbol{D}$의 대각에 있는 비정칙값들의 마지막 부분은 0에 가까워지고, 따라서 역은 매우 커져 여러 가지 문제가 발생한다. 예를 들면, 회귀계수 추정이 불안정하게 되어 실제 계수와는 매우 다른 계수 추정값을 얻게 될 수 있고, 예측력이 많이 나빠질 수 있다.

PCR에서는 다중공선성이나 큰 변수수에서의 OLS의 문제를 해결하기 위해 원래의 p개 예측변수 대신에 소수의 독립된 m개 잠재변수를 사용한다. 즉 PCR에서는 원 p차원 공간의 문제를 m 차원의 공간으로 축소하여 악조건으로 인한 가역 문제를 해결하고 보다 안정된 추정 회귀계수를 얻게 되기를 기대한다. 기본 방법은 다음과 같다: $\boldsymbol{T} \equiv \boldsymbol{X}\boldsymbol{V}$와 $\boldsymbol{\alpha} \equiv \boldsymbol{V}'\boldsymbol{\beta}$로 두어 회귀모형 $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$을 정준형

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

으로 바꾼다. 계수 α의 OLS 추정량은

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{D}^{-2}\boldsymbol{T}'\boldsymbol{y}$$

이며, 이때

$$\boldsymbol{D}^{-2} = \mathrm{diag}\left(\frac{1}{\delta_1^2}, \cdots, \frac{1}{\delta_p^2}\right)$$

이다. 벡터 $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_p)'$의 원소에서 첫 m개만을 사용한다면

$$\hat{\boldsymbol{\alpha}}_{(m)} = \boldsymbol{D}_{(m)}^{-2}\boldsymbol{T}'\boldsymbol{y}$$

이며, 여기서

$$\boldsymbol{D}_{(m)}^{-2} = \mathrm{diag}\left(\frac{1}{\delta_1^2}, \cdots, \frac{1}{\delta_m^2}, 0, \cdots, 0\right) \tag{2.1}$$

이다.

이제 (2.1)의 대각선 원소의 분모에 작은 양의 값을 추가한 다음의 방법을 고려한다.

$$\boldsymbol{D}_{(m,c)}^{-2} = \mathrm{diag}\left(\frac{1}{\delta_1^2 + c}, \cdots, \frac{1}{\delta_m^2 + c}, 0, \cdots, 0\right) \tag{2.2}$$

식 (2.1) 대신에 (2.2)를 사용함으로써, 확실하게 잡음인 주성분을 먼저 제거하고, 여전히 남아있는 작은 고유값으로 인해 발생할 수 있는 문제점을 줄이기 위해 작은 양의 값을 첨가하여 PCR의 이점과 RR의 이점을 같이 가지기를 기대한다. 이것을 이용한 회귀를 능형 주성분회귀(ridge principal component regression, RPCR)이라 부르기로 한다(Vigneau 외, 1997).

위의 RPCR을 얻는 다른 접근방법은 다음과 같다. 첫 m개 인자의 PCR의 해 $\hat{\boldsymbol{\beta}}_{pcr}^{(m)} = \boldsymbol{V}\hat{\boldsymbol{\alpha}}_{(m)}$은

$$\hat{\boldsymbol{\beta}}_{pcr}^{(m)} = \boldsymbol{V}_m\boldsymbol{D}_m^{-1}\boldsymbol{U}_m'\boldsymbol{y}$$

로 쓸 수 있으며, 여기서 $\boldsymbol{U}_m = (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m)$, $\boldsymbol{D}_m = \mathrm{diag}(\delta_1, \cdots, \delta_m)$, $\boldsymbol{V}_m = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m)$이다. 이것의 대수식은

$$\hat{\boldsymbol{\beta}}_{pcr}^{(m)} = \sum_{j=1}^{m}\left(\frac{\boldsymbol{u}_j'\boldsymbol{y}}{\delta_j}\right)\boldsymbol{v}_j$$

이다. 이제 m개 인자의 PCR 해를 약간 다르게 표현하면

$$\hat{\boldsymbol{\beta}}_{pcr}^{(m)} = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{E}\boldsymbol{U}'\boldsymbol{y} \tag{2.3}$$

의 형태로 쓸 수 있는데 여기서 $p\times p$ 행렬 $\boldsymbol{E}=\begin{pmatrix}\boldsymbol{I}_m & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}$이며 $\boldsymbol{I}_m$은 $m\times m$ 단위행렬이다. 이 결과는 행렬 $\boldsymbol{V}, \boldsymbol{D}, \boldsymbol{U}$를 $\boldsymbol{V}=(\boldsymbol{V}_m, \boldsymbol{V}_{p-m})$, $\boldsymbol{D}=\mathrm{diag}(\boldsymbol{D}_m, \boldsymbol{D}_{p-m})$, $\boldsymbol{U}=(\boldsymbol{U}_m, \boldsymbol{U}_{p-m})$으로 분할하여 곱함으로써 쉽게 확인된다. 위의 회귀벡터의 식에서 $\boldsymbol{D}^{-1}\boldsymbol{E}$ 부분은

$$\boldsymbol{D}^{-1}\boldsymbol{E}=\mathrm{diag}\left(\frac{1}{\delta_1},\cdots,\frac{1}{\delta_m},0,\cdots,0\right)$$

으로 대각원소가 m개의 $1/\delta_j$와 $p-m$개의 0으로 구성된 대각행렬이다.

RR은 $\boldsymbol{X}'\boldsymbol{X}$의 대각에 작은 양의 값 c를 추가하여 회귀해 $\hat{\boldsymbol{\beta}}_{rr}=(\boldsymbol{X}'\boldsymbol{X}+c\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$를 계산함으로써 추정 회귀해의 안정을 도모한다. 여기서 $\boldsymbol{X}$의 SVD를 이용하면

$$\hat{\boldsymbol{\beta}}_{rr}=\boldsymbol{V}\boldsymbol{D}^{-1}(\boldsymbol{I}+c\boldsymbol{D}^{-2})^{-1}\boldsymbol{U}'\boldsymbol{y} \tag{2.4}$$

를 얻는다. 이 RR의 해 식에서 $\boldsymbol{D}^{-1}(\boldsymbol{I}+c\boldsymbol{D}^{-2})^{-1}$ 부분의 대수식은

$$\boldsymbol{D}^{-1}(\boldsymbol{I}+c\boldsymbol{D}^{-2})^{-1}=\mathrm{diag}\left(\frac{1}{\delta_1+c/\delta_1},\cdots,\frac{1}{\delta_p+c/\delta_p}\right)$$

이다. 이것은 각 비정칙값 δ_j에 일반적으로 작은 값인 c/δ_j를 추가로 조정해주는 셈인데, 큰 δ_j에 추가된 부분에서는 별 영향이 없으며 작은 δ_j에 추가된 부분에서 중요한 역할을 하게 된다.

이제 식 (2.3)과 (2.4)를 결합한 다음의 RPCR 해를 고려한다.

$$\hat{\boldsymbol{\beta}}_{rpcr}=\boldsymbol{V}\boldsymbol{D}^{-1}(\boldsymbol{I}+c\boldsymbol{D}^{-2})^{-1}\boldsymbol{E}\boldsymbol{U}'\boldsymbol{y} \tag{2.5}$$

여기서 $\boldsymbol{E}$는 앞에서와 마찬가지로 $p\times p$ 행렬 $\boldsymbol{E}=\begin{pmatrix}\boldsymbol{I}_m & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}$이다. (2.5)는 또한

$$\hat{\boldsymbol{\beta}}_{rpcr}^{(m)}=\boldsymbol{V}_m\boldsymbol{D}_m^{-1}(\boldsymbol{I}_m+c\boldsymbol{D}_m^{-2})^{-1}\boldsymbol{U}_m{}'\boldsymbol{y}$$

로 쓸 수 있다. 이것의 대수식은

$$\hat{\boldsymbol{\beta}}_{rpcr}^{(m)}=\sum_{j=1}^{m}\left(\frac{\boldsymbol{u}_j{}'\boldsymbol{y}}{\delta_j+c/\delta_j}\right)\boldsymbol{v}_j$$

이다. 이 회귀는 RR과 PCR의 장점을 각각 살린 방법이 될 수 있으리라 기대한다. 즉 PCR에서처럼 영에 가까운 비정칙값 즉 확실히 잡음인 부분을 제거하고 더불어 RR에서처럼 작은 비정칙값을 위해 적절한 값 c/δ_j로 추가 조정해주는 것이다.

3. 능형회귀와 부분최소제곱회귀의 결합

m개 잠재인자의 PLS의 해는 다음과 같이 설명할 수 있다. 여기서는 $\boldsymbol{T} \equiv \boldsymbol{X}\boldsymbol{W}(\boldsymbol{P}'\boldsymbol{W})^{-1}$와 $\boldsymbol{\alpha} \equiv (\boldsymbol{P}'\boldsymbol{W})\boldsymbol{W}'\boldsymbol{\beta}$로 두는데, $\boldsymbol{W}$와 $\boldsymbol{P}$는 Wold의 PLS 알고리듬(Wold, 1966; Martens과 Naes, 1989)에서 얻어지는 가중값행렬과 적재행렬이다(편의상 PCR에서와 같은 부호 $\boldsymbol{T}$와 $\boldsymbol{\alpha}$를 사용한다). 회귀모형 $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$을 정준형

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

으로 바꾸고, 여기서 PLS 추정량

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{y}$$

를 얻는다. 이 $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_p)'$의 원소에서 m개만을 선택한 것을

$$\hat{\boldsymbol{\alpha}}_{(m)} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_m, 0, \cdots, 0)'$$

이라 하면 이 때의 PLS의 해는

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{pls}^{(m)} &= \boldsymbol{W}(\boldsymbol{P}'\boldsymbol{W})^{-1}\hat{\boldsymbol{\alpha}}_{(m)} \\ &= \boldsymbol{W}_m(\boldsymbol{W}_m'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{W}_m)^{-1}\boldsymbol{W}_m'\boldsymbol{X}'\boldsymbol{y} \end{aligned}$$

임을 보일 수 있다.

이 m개 잠재인자의 PLS 해는 다음과 같이 쓸 수도 있다(Kim, 2003).

$$\hat{\boldsymbol{\beta}}_{(pls)}^{(m)} = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{G}_{(m)}\boldsymbol{U}'\boldsymbol{y} \tag{3.1}$$

여기서 $\boldsymbol{G}_{(m)} = \boldsymbol{Q}_m(\boldsymbol{Q}_m'\boldsymbol{Q}_m)^{-1}\boldsymbol{Q}_m'$, $\boldsymbol{Q}_m = (\boldsymbol{D}^2\boldsymbol{U}'\boldsymbol{y}, \boldsymbol{D}^4\boldsymbol{U}'\boldsymbol{y}, \cdots, \boldsymbol{D}^{2m}\boldsymbol{U}'\boldsymbol{y})$이고, $\boldsymbol{V}$, $\boldsymbol{D}$, $\boldsymbol{U}$는 1절에서 정의된 대로이다. 이것의 대수식은

$$\hat{\boldsymbol{\beta}}_{pls}^{(m)} = \sum_{j=1}^{p}\left(\frac{1}{\delta_j}\sum_{l=1}^{p} g_{jl}\boldsymbol{u}_l'\boldsymbol{y}\right)\boldsymbol{v}_j$$

인데 여기서 g_{jl}은 행렬 $\boldsymbol{G}_{(m)}$의 (j,l)번째 원소이고 나머지는 앞서 정의된 것과 동일하다.

이제 앞 절에서 시도한 RR과 PCR의 결합과 비슷한 접근법을 사용하고자 한다. 식 (2.3)과 (3.1)을 비교하면 PCR 식의 $\boldsymbol{E}$와 PLS 식의 $\boldsymbol{G}_{(m)}$은 대응되는 위치에 있으며 유사한 역할을 한다. 따라서 RR과 PLS의 결합인 다음의 '능형 부분최소제곱회귀'(ridge partial least squares regression, RPLS)를 고려한다.

$$\hat{\boldsymbol{\beta}}_{rpls}^{(m)} = \boldsymbol{V}\boldsymbol{D}^{-1}(\boldsymbol{I}+c\boldsymbol{D}^{-2})^{-1}\boldsymbol{G}_{(m)}\boldsymbol{U}'\boldsymbol{y}$$

이것의 대수식은

$$\hat{\boldsymbol{\beta}}_{rpls}^{(m)} = \sum_{j=1}^{p}\left(\frac{\sum_{l=1}^{p} g_{jl}\boldsymbol{u}_l'\boldsymbol{y}}{\delta_j + c/\delta_j}\right)\boldsymbol{v}_j$$

임을 보일 수 있다. 이 추정량은 PLS의 장점과 RR의 장점을 같이 가지리라 기대한다.

4. 수치 예

수치예로서 Naes(1989)의 NIR 분광 데이터 셋을 이용하였다(이 데이터 셋은 부록에 주어져 있다). 이것은 균질 고기의 자료이며 비교적 넓은 범위의 실제 샘플을 다룬 것이다. 각 샘플(관측값)은 미리 지정된 19개의 NIR 파장을 log(1/반사율)로 변환한 것과 단백질의 비율(%)을 측정한 것으로 구성되어 있다. 고기에서의 다른 주요 성분인 지방과 물은 각각 (2% − 25%)와 (60% − 80%)의 구간에 있다. log(1/반사율)은 NIR 분석에서 주로 사용되는 것인데, 가정된 회귀모형의 선형성이 잘 맞는 것이 확인되었다. 원래의 총 샘플 수는 28이나 Naes는 회귀진단을 수행하기 위해 3개의 샘플(2, 27, 28번째 관측값)은 고의로 변경시켰기 때문에 본 논문에서는 이들을 제외한 25개의 샘플을 이용하였다.

모든 변수는 상관변환(단위길이표준화)하여 처리하였다. 즉 각 변수에서 그것의 평균을 빼고 제곱합(평균에서의)의 제곱근을 나눈다. 따라서 변환후의 $\boldsymbol{X}'\boldsymbol{X}$는 x변수들간의 상관행렬이 되고 $\boldsymbol{X}'\boldsymbol{y}$는 x변수들과 y간의 상관벡터가 된다. 이 데이터의 조건수를 구해보면

$$\phi = \frac{\delta_{\max}}{\delta_{\min}} = \frac{4.3404995}{0.0004674} = 9286$$

으로 매우 높은 다중공선성이 존재함을 알 수 있다.

예측력의 비교는 leave-one-out 방법에 의한 PRESS를 이용한다. 최종 비교는 다음의 RMSEP을 사용한다.

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_{i(i)})^2}{n}}$$

여기서 y_i는 i번째 y 관측값, $\hat{y}_{i(i)}$는 i번째 관측값을 제외한 나머지 $n-1$개의 관측값에 의한 적합 회귀식으로 구한 y_i의 예측값이다. 그리고 능형상수 c의 최적값은 Hoerl 외(1975)가 제의한 $c = p\widehat{\sigma^2}/\hat{\beta}_{ols}'\hat{\beta}_{ols}$을 이용한다. 여기서 p는 모형에서 β_0을 제외한 모수의 수, $\widehat{\sigma^2}$은 OLS에서 얻어진 잔차평균제곱, $\hat{\beta}_{ols}$는 OLS에서 얻어진 회귀벡터이다. 이것은 추정량의 MSE를 최소화하는 기준으로 구해진 식이다.

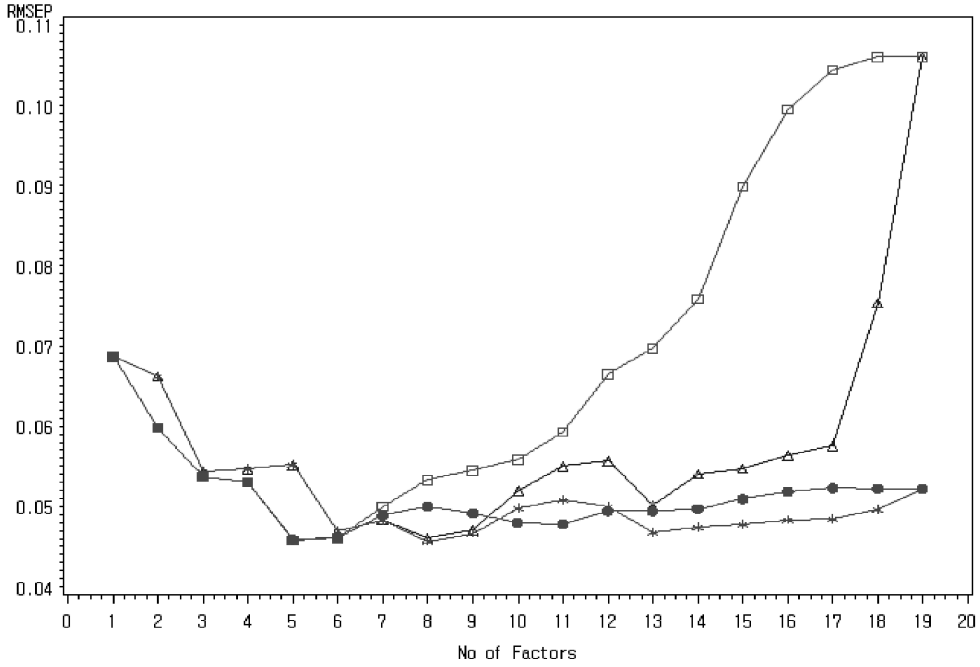
각 회귀방법에서 상관변환과 PRESS 계산의 자세한 알고리듬은 다음과 같다.

1. 데이터에서 한 개의 관측값을 제외한 나머지 $n-1$개의 관측값에서 y와 x 변수들의 각각을 상관변환을 이용하여 척도화한다. 즉 각 변수에서 평균을 빼고 SS의 제곱근으로 나눈다. 여기서 SS는 평균에서의 제곱합이다.
2. 단계 1에서 얻어진 척도화된 $n-1$개의 관측값을 이용하여 회귀벡터를 계산한다.
3. 제외된 관측값에서의 각 x 변수 값을 단계 1에서 얻어진 평균과 SS의 제곱근을 사용하여 척도화하고 단계 2의 회귀벡터로써 y의 예측값을 구한다.
4. 제외된 점의 y 관측값을 단계 1에서 구해진 평균과 SS의 제곱근으로 척도화하고, 이 값과 단계 3에서 구한 y의 예측값과의 차이를 계산하고 그 결과를 제곱한다.
5. 각 관측값에 대해 단계 1-4를 반복하고 단계 4의 차이제곱값을 더해나가는 과정을 관측값을 모두 한 번씩 제외시키면서 반복한다.

이상 설명된 데이터 표준화와 PRESS 알고리듬에 의해 구해진 PCR, RPCR, PLS, RPLS의 RMSEP 값은 <표 1>과 같고 이것을 그래프로 그린 것이 <그림 1>이다.

<표 1> Naes 데이터의 PCR, RPCR, PLS, RPLS의 RMSEP 값

인자수	PCR	RPCR	PLS	RPLS
1	0.06881	0.06881	0.06872	0.06872
2	0.06625	0.06625	0.05983	0.05983
3	0.05430	0.05430	0.05374	0.05375
4	0.05469	0.05467	0.05303	0.05305
5	0.05518	0.05516	0.04580	0.04590
6	0.04698	0.04700	0.04601	0.04610
7	0.04835	0.04821	0.05005	0.04895
8	0.04608	0.04561	0.05329	0.05002
9	0.04709	0.04664	0.05452	0.04917
10	0.05196	0.04980	0.05588	0.04796
11	0.05504	0.05074	0.05928	0.04779
12	0.05570	0.04994	0.06652	0.04946
13	0.05008	0.04671	0.06972	0.04942
14	0.05405	0.04740	0.07589	0.04972
15	0.05475	0.04776	0.08994	0.05094
16	0.05635	0.04829	0.09952	0.05188
17	0.05753	0.04838	0.10449	0.05231
18	0.07536	0.04956	0.10608	0.05225
19	0.10604	0.05223	0.10604	0.05223



<그림 1> 인자수에 따른 RMSEP 값. 삼각 △은 PCR, 별표 *는 RPCR, 정사각 □은 PLS, 점 •은 RPLS의 값을 나타낸다.

PCR과 RPCR은 인자수 $k=8$, PLS와 RPLS는 $k=5$에서 가장 작은 RMSEP 값을 보이므로 각각 최적모형으로 선택할 수 있을 것이다. 그런데 이때의 RMSEP 값은 비슷함을 알 수 있다. 그러나 인자수가 커짐에 따라 PCR과 PLS의 RMSEP 값은 전체적으로 계속 증가하여 예측력이 많이 떨어지는 반면 RPCR과 RPLS의 경우는 그 증가가 미미하며 여전히 좋은 예측력을 보인다. RPLS는 식의 도입에서 기대한 바와 같이 PLS와 RR의 장점을 같이 지니는 것으로 보인다. 즉 PLS의 기법에 의해 적은 수의 인자에서 예측력이 좋은 모형이 되고 능형상수를 도입함으로서 인자수가 커지더라도 예측력이 그다지 나빠지지 않는 모형이 되는 것이다.

OLS의 RMSEP 값은 바로 PCR이나 PLS의 마지막($k=19$) 값이며 예측력이 상대적으로 떨어짐을 알 수 있다. 또한 RR의 RMSEP 값은 RPCR이나 RPLS의 마지막($k=19$) 값이며 PCR, RPCR, PLS, RPLS의 최적 경우보다는 약간 떨어지나 OLS 보다는 훨씬 나음을 알 수 있다.

5. 결론

높은 다중공선성의 데이터에서 RR은 추정회귀계수의 안정화에 효과적이며 PCR과 PLS는 많은 원 예측변수 대신에 적은 수의 잠재인자를 사용함으로써 모형화에서 강점을 지닌다. 그러나 이들 잠재인자회귀는 과대적합에 종종 매우 민감하며 적절한 잠재인자수를 결정하는 것이 쉽지 않은 문제점이 있다. 본 논문에서는 RR과 PCR의 장점만을 결합한 RPCR을 유도하고 유사한 방식으로 RR과 PLS를 결합한 RPLS를 개발하였다. 높은 다중공선성을 지닌 실제 NIR 분광 데이터를 이용하여 새로 개발된 두 회귀와 기존의 OLS, RR, PCR, PLS의 예측력을 RMSEP 값으로 비교하였다. RPCR과 RPLS는 소수의 잠재인자 식에서 좋은 예측력을 가질 뿐 아니라 과대적합 시에도 여전히 괜찮은 예측력을 지니는 것으로 나타났다. 본 논문에서는 언급하지 않은 몇 가지 다른 NIR 데이터 셋에서의 조사에서도 유사한 패턴이 나타났다. 따라서 RPCR과 RPLS는 기존의 PCR과 PLS에 비해 잠재인자의 수에 덜 민감하고 보다 안정된 예측력을 주는 회귀방법으로 기대된다.

부록

<표> Naes 데이터 셋 (첫 19개 예측변수는 NIR 측정이고 마지막 반응변수는 단백질 %)

샘플 \ 변수	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	1.6709	1.6778	1.6648	1.6471	1.6286	1.6269	1.6286	1.6337	1.6292	1.6578
2	1.6161	1.6255	1.6112	1.5827	1.5561	1.5566	1.5604	1.5724	1.5629	1.6250
3	1.6430	1.6524	1.6378	1.6123	1.5878	1.5878	1.5909	1.6016	1.5927	1.6468
4	1.5868	1.5992	1.5850	1.5455	1.5124	1.5130	1.5189	1.5363	1.5235	1.6131
5	1.6024	1.6129	1.5989	1.5643	1.5343	1.5338	1.5388	1.5530	1.5414	1.6134
6	1.5672	1.5791	1.5670	1.5293	1.4955	1.4953	1.5013	1.5178	1.5047	1.5868
7	1.5590	1.5712	1.5576	1.5171	1.4828	1.4832	1.4895	1.5072	1.4930	1.5865
8	1.5286	1.5427	1.5309	1.4842	1.4452	1.4465	1.4534	1.4732	1.4585	1.5650
9	1.5699	1.5834	1.5682	1.5262	1.4903	1.4911	1.4973	1.5165	1.5019	1.6019
10	1.4868	1.5051	1.4934	1.4328	1.3832	1.3846	1.3937	1.4199	1.4001	1.5465
11	1.5096	1.5262	1.5128	1.4558	1.4106	1.4114	1.4194	1.4443	1.4257	1.5602
12	1.4681	1.4875	1.4774	1.4102	1.3563	1.3573	1.3670	1.3943	1.3741	1.5291
13	1.4818	1.5007	1.4879	1.4246	1.3740	1.3752	1.3841	1.4111	1.3906	1.5430
14	1.4426	1.4639	1.4553	1.3804	1.3199	1.3206	1.3296	1.3587	1.3374	1.5049
15	1.4493	1.4718	1.4598	1.3836	1.3259	1.3266	1.3364	1.3673	1.3443	1.5235
16	1.4459	1.4666	1.4589	1.3854	1.3258	1.3266	1.3364	1.3651	1.3440	1.5092
17	1.4493	1.4726	1.4650	1.3820	1.3162	1.3166	1.3262	1.3583	1.3341	1.5174
18	1.4197	1.4467	1.4404	1.3443	1.2677	1.2672	1.2775	1.3130	1.2868	1.4963
19	1.4557	1.4806	1.4758	1.3865	1.3120	1.3106	1.3191	1.3513	1.3276	1.5115
20	1.4152	1.4432	1.4386	1.3397	1.2608	1.2605	1.2705	1.3053	1.2796	1.4887
21	1.4432	1.4691	1.4593	1.3677	1.2962	1.2970	1.3071	1.3429	1.3165	1.5253
22	1.3909	1.4199	1.4183	1.3133	1.2238	1.2218	1.2312	1.2676	1.2402	1.4579
23	1.3925	1.4188	1.4157	1.3195	1.2401	1.2389	1.2479	1.2824	1.2569	1.4557
24	1.3971	1.4274	1.4283	1.3159	1.2205	1.2170	1.2265	1.2628	1.2355	1.4557
25	1.4004	1.4283	1.4280	1.3241	1.2378	1.2355	1.2447	1.2809	1.2534	1.4591

<표> (계속)

샘플 \ 변수	x11	x12	x13	x14	x15	x16	x17	x18	x19	y
1	1.4868	1.4806	1.6424	1.4671	1.6635	1.4450	1.4346	1.5829	1.3973	17.04
2	1.3852	1.3804	1.5893	1.3660	1.6343	1.3402	1.3314	1.5352	1.2747	16.22
3	1.4248	1.4236	1.6167	1.4094	1.6551	1.3846	1.3763	1.5617	1.3219	15.71
4	1.3330	1.3313	1.5612	1.3196	1.6253	1.2964	1.2926	1.5119	1.2174	15.90
5	1.3593	1.3566	1.5727	1.3440	1.6225	1.3198	1.3144	1.5202	1.2453	15.46
6	1.3206	1.3203	1.5393	1.3097	1.5984	1.2871	1.2841	1.4924	1.2066	15.89
7	1.3008	1.2995	1.5325	1.2882	1.5984	1.2646	1.2612	1.4842	1.1841	15.31
8	1.2630	1.2642	1.5015	1.2550	1.5784	1.2339	1.2331	1.4583	1.1466	15.34
9	1.3032	1.3020	1.5437	1.2904	1.6145	1.2664	1.2630	1.4963	1.1854	15.13
10	1.1840	1.1875	1.4577	1.1813	1.5646	1.1609	1.1654	1.4180	1.0578	14.77
11	1.2060	1.2077	1.4800	1.1984	1.5766	1.1753	1.1766	1.4369	1.0776	13.94
12	1.1522	1.1581	1.4344	1.1533	1.5479	1.1349	1.1410	1.3952	1.0276	14.24
13	1.1717	1.1750	1.4502	1.1673	1.5612	1.1464	1.1496	1.4101	1.0458	13.91
14	1.1194	1.1274	1.4001	1.1248	1.5255	1.1095	1.1180	1.3604	0.99641	13.71
15	1.1177	1.1233	1.4119	1.1178	1.5458	1.0990	1.1049	1.3751	0.99247	13.39
16	1.1258	1.1338	1.4062	1.1320	1.5300	1.1161	1.1258	1.3668	0.99938	13.51
17	1.1136	1.1228	1.4028	1.1217	1.5402	1.1059	1.1174	1.3668	0.98396	12.91
18	1.0626	1.0753	1.3607	1.0776	1.5231	1.0648	1.0798	1.3259	0.93333	12.83
19	1.1175	1.1302	1.3934	1.1332	1.5354	1.1213	1.1368	1.3522	0.99006	12.32
20	1.0617	1.0748	1.3524	1.0782	1.5165	1.0671	1.0829	1.3170	0.93452	12.63
21	1.0850	1.0946	1.3934	1.0932	1.5504	1.0768	1.0855	1.3584	0.95268	12.56
22	1.0310	1.0482	1.3132	1.0562	1.4885	1.0492	1.0690	1.2777	0.90219	11.83
23	1.0486	1.0634	1.3272	1.0684	1.4820	1.0585	1.0762	1.2928	0.92260	11.85
24	1.0252	1.0445	1.3068	1.0549	1.4866	1.0495	1.0724	1.2672	0.89733	11.44
25	1.0494	1.0666	1.3251	1.0744	1.4871	1.0669	1.0876	1.2927	0.92155	11.16

참고문헌

1. Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). Ridge Regression: Some Simulations, *Communication in Statistics*, 4, 105-123.
2. Kim, J. D. (2003). Unified Non-iterative algorithm for Principal Component Regression, Partial Least Squares and Ordinary Least Squares, *Journal of Korean Data & Information Science Society*, 14, 355-366.
3. Martens, H. and Naes, T. (1989). *Multivariate Calibration*, John Wiley & Sons.
4. Naes, T. (1989). Leverage and Influence Measures for Principal Component Regression. *Chemometrics and Intelligent Laboratory Systems*, 5, 155-168.
5. Vigneau, E., Devaux, M. F., Qannari, E. M., and Robert, P. (1997). Principal Component Regression, Ridge Regression and Ridge Principal Component Regression in Spectroscopy Calibration, *Journal of Chemometrics*, 11, 239-249.
6. Wold, H. (1966). *Estimation of Principal Components and Related Models by Iterative Least Squares, in Multivariate Analysis* (ed. Krishnaiah, P. R.), 391-420, Academic Press, New York.

[2007년 1월 접수, 2007년 2월 채택]