碩士學位論文

다중회귀분석에서 희소주성분회귀법의 효율성 연구

韓國外國語大學校 大學院

統 計 學 科

李 昊 星

碩士學位論文

다중회귀분석에서 희소주성분회귀법의 효율성 연구

On the Estimation for Sparse Principal Component Regression Approach under Multiple Regression Problem

指導 李 碩 浩 教授

이 論文**을** 碩士學位請求論文**으로** 提出**합니다.**

2014 年 **12** 月

韓國外國語大學校 大學院
統 計 學 科
李 昊 星

이 論文**을** 李昊星**의** 碩士學位論文**으로** 認定**함**

2014 年 **12** 月 日

審 査 委 員 _____ **(인)**

審 査 委 員 _____ **(인)**

審 査 委 員 _____ **(인)**

韓國外國語大學校 大學院

요약

본 연구에서는 주성분 회귀분석에서 이용되는 주성분을 Lasso등 여러 희소벌점함수를 이용한 변수선택을 통해 예측율을 높이고 추정량의 정밀성을 높이는 방법을 제안한다. 다양한 상황에서의 회귀분석 문제에서 기존의 방법론과의 비교를 컴퓨터 모의실험을 통해 수행하였으며, 그 결과 모든 상황에서 우수한 성능을 보임을 확인하였다.

주요용어 :주성분 분석 , 능형회귀, Lasso회귀, SCAD회귀

목 차

표 목 차

그 림 목 차

1 서 론

회귀분석은 여러 응용분야에서 매우 흔히 사용되는 통계적 기법 중 하나이다. 여러 개의 설명변수가 하나의 종속변수에 영향을 미치는 다중회귀모형 $Y = X\beta + \epsilon$에서 회귀계수 β의 추정은 오차제곱합을 최소화하는 최소제곱추정법(least square estimation)에 의해 이루어지는 것이 일반적이다. 최소제곱추정법은 불편성(unbiasedness), 일치성(consistency), 최소분산(minimum variance)의 성질을 갖는 추정량을 제공하는 전통적이며 효율적인 추정법이다. 하지만 최소제곱추정량이 이러한 좋은 성질을 가지기 위해서는 자료의 개수가 설명변수의 개수보다 매우 커야 한다는 조건이 담보되는 경우에만 한정된다는 단점이 존재한다.

하지만 여러 응용분야에서 나타나는 최근의 자료들은 자료 취득에 대한 기술적 발전 및 저장 용량의 비용 감소 등의 이유로 인하여 관측치의 규모가 방대할 뿐만 아니라 고려하는 설명변수의 수도 자료의 수보다 많은 경우가 매우 빈번하다. 빅데이터 시대가 도래함에 따라 이러한 경향은 가속화 되고 있으며 IT, 생명정보학 분야에서는 매우 일반적인 현상이다. 이와 같은 경우, 고려하는 설명변수에 대한 면밀한 선택과정이 없다면 상호간의 연관성이 높은 설명변수들이 자료에 포함되는 것은 매우 빈번하며, 반드시 고려해야 하는 변수들이 본질적으로 연관성이 높지만 사전작업을 통해 변수를 취사선택하기 어려운 경우가 존재한다. 텍스트마이닝의 경우 연관성이 높은 두 단어가 나타나는지 여부를 변수로 포함하는 경우는 전자에 해당하며, 유전자 분석에서 동일한 생물학적 반응에 관련되는 두 개 혹은 그 이상의 유전자들의 발현정도를 설명변수로 포함하는 경우가 후자의 예에 해당한다. 또한 자료의 개수가 설명변수 개수보다 작은 경우에는 설명변수 간에 물리적 연관성이 없음이 명백하더라도 설명변수간의 강한 상관성이 반드시 나타나게 되어 최소제곱추정량의 안정성이 보장되지 않게 된다. 이처럼 설명변수간에 강한 연관성을 가지는 경우를 다중공선성(multicolinearity)이라 부른다.

다중공선성이 존재하는 경우 최소제곱법에 의한 회귀계수의 추정량의 분산이 팽창하

게 되어 결과적으로 추정량의 안정성이 담보되지 않는다. 회귀분석에서 다중공선성의 문제를 해결하는 방법으로 차원축소법(dimension reduction) 및 축소추정법(shrinkage method)이 이용된다.

차원축소방법은 가급적 자료 본래의 특성을 유지하면서 설명변수 공간의 차원을 훨씬 작도록 변환하여 축소된 공간 상의 정보만을 회귀분석에 이용하는 방법이다. 이러한 차원축소법은 설명변수의 공간을 직교화하여 각 차원에 대응하는 변수들의 상관계수를 0으로 만들고 설명변수간의 상관성이 높게 나타나는 부분공간을 분석에서 배제함으로써 다중공선성 문제를 해결한다. 회귀분석에서의 대표적인 차원축소방법으로는 주성분회귀법(Principal Component Regression; PCR)이 있다. 주성분회귀법은 설명변수의 공간을 주성분으로 분해하고, 설명변수의 변동을 잘 나타내는 소수의 상위주성분을 설명변수로 고려하여 회귀분석을 수행하는 방법이다. 하지만, 주성분회귀법은 선택된 상위주성분이 반응변수와 연관성이 적거나, 선택되지 않은 하위주성분이 반응변수와 연관성이 높은 경우 예측력 향상을 보장할 수 없다는 단점을 지닌다. 따라서 반응변수와의 연관성을 기준으로 설명변수의 차원을 줄이는 방법으로 부분최소제곱회귀법(Partial Least Squares Regression; PLSR)은 주성분회귀법의 단점을 보완할 수 있다. 자세한 내용은 2.1절에서 다루도록 한다.

다중공선성을 회피하는 가장 직관적인 해법은 서로 상관성이 높은 설명변수 중에서 반응변수와 연관성이 높은 하나, 혹은 몇 개의 설명변수만 모형에 포함하고 이들과 상관성이 높은 나머지 설명변수들을 모형에서 배제하는 방법이다. 이는 변수선택(variable selection)의 문제이며 이를 자료상황에 맞게 자동으로 수행하는 방법들을 통칭하여 모형선택(model selection)이라 부른다. 이러한 방법은 설명변수가 많은 경우에 비교해야할 가능한 설명변수의 조합수가 기하급수적으로 늘어나 실제문제에 적용하기 어렵다는 현실적 어려움이 있다. 이를 기술적으로 해결하기 위해 전진 선택법(forward selection), 후진 제거법(backward elimination) 등이 고안되었지만 이는 준최

적(suboptimal) 해법으로 최적의 결과를 기대할 수 없다. 더욱이 변수선택 과정 자체는 이산적인 과정(discrete selection procedure)에 기반하므로 결과의 안정성이 떨어진다는 이론적 제약을 안고 있다. 이러한 단점을 보완하는 연속적 선택과정(continuous selection procedure)으로써 벌점함수(penalty function)을 이용한 축소추정법이 활발히 연구되고 있다. 벌점함수는 회귀계수의 추정량의 크기를 축소하는 역할을 하며 이는 추정량의 공간을 유계공간(bounded space)으로 한정하여 추정하는 원리로 이루어진다. 대표적인 벌점회귀법으로는 사용되는 벌점함수에 따라 능형회귀법(ridge regression), lasso 회귀법, SCAD 회귀법 등이 있다. 자세한 내용은 2.2절에 기술한다.

본 논문에서는 회귀분석에서 다중공선성의 문제와 변수선택의 문제를 보완하기 위해 차원축소방법 중 주성분분석을 이용하되, 사용되는 설명변수로써 상위주성분만을 이용하는 대신 희소벌점함수를 이용하여 주성분을 선택하는 방법을 제안한다. 이는 차원축소의 효과를 기대할 수 있을 뿐만 아니라, 동시에 희소벌점함수를 이용하여 주성분을 선택함으로 상위주성분이 아닌 주성분이라고 할지라도 반응변수와 연관성이 높은 주성분을 선택할 수 있게 한다. 따라서 반응변수와의 연관성이 높은 모든 주성분을 모형에 포함하기에 예측력을 높일 수 있을 것으로 기대한다.

2 다중 회귀분석의 다양한 방법

다중회귀분석에 있어서 예측력을 향상시키기 위해서 차원축소법과 변수선택법이 사용된다. 차원축소법은 설명변수 공간의 차원을 줄이는 것을 의미한다. 대표적으로 주성분분석(Principal Component Analysis; PCA)을 이용하여 설명변수 공간을 주요한 몇개의 주성분으로 대표하여 이를 이용한 회귀분석에 이용하는 방법으로 주성분회귀법(Principal Component Regression; PCR)이 있으며, 주성분분석과 비슷하나 설명변수 뿐 아니라 반응변수를 고려한 부분최소제곱회귀법(Partial Least Squares Regression;

PLSR)이 있다. 변수선택방법은 설명변수 중 일부분만 모형에 넣는 방법을 의미하며, 반응변수와 연관성이 높은 변수만 취하고 연관성이 낮은 설명변수는 모형에서 제외하는 방법이다. 이는 모든 가능한 조합의 모형 중에서 가장 높은 예측력을 주는 모형을 취함으로써 최적의 설명변수 집합을 찾는 방법이다. 하지만, 설명변수가 많은 경우에는 너무 많은 조합수가 존재하므로 실용적이지 못한다는 문제점이 있다. 따라서 이런 문제점을 완화하기 위해 전진 선택법(forward selection), 후진 소거법(backward elimination) 등의 방법이 이용된다. 하지만 이러한 모형선택은 이산적인 방법으로 안정적이지 못하다는 단점이 있다. 이러한 변수선택법의 단점을 보완하는 방법으로 희소벌점함수(sparsity-inducing penalty function)를 이용한 축소추정법을 사용할 수 있다. 이 절에서는 차원축소법을 이용한 회귀분석과 축소추정법을 활용한 회귀분석에 대하여 간단히 살펴본다.

독립변수의 개수가 p인 아래의 다중회귀모형을 고려하자.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_p + \epsilon,$$

여기서 오차는 $\epsilon \sim N(0, \sigma^2)$를 가정한다. 일반적인 회귀분석은 잔차제곱합(residual sum of squares)을 최소로 하는 추정량을 찾는 최소제곱법(least squares method)을 사용하여 회귀계수(regression coefficient) 추정량을 구한다. 그러나 설명변수의 개수가 자료수와 비슷하거나 더 커지게 되는 경우, 최소제곱추정량은 여러가지 측면에서 만족스럽지 못한 결과를 가진다. 첫째로 추정량의 추정정확도(estimation accuracy)가 낮아진다. 설명변수의 수가 증가하면 설명변수들 간의 강한 상관관계에 의해 다중공선성(multicollinearity)이 존재할 수 있다. 이로 인하여 회귀계수 추정량의 분산(variance)이 커져 추정량의 정확도가 떨어지는 문제점이 발생할 수 있다. 둘째로, 변수수가 자료수보다 많기 때문에 회귀식으로부터 얻게 되는 예측값을 통한 예측정확도(prediction accuracy)가 현저히 떨어지게 된다. 이러한 경우 회귀식을 적합하는 자료의 반응변수값과 정확히 일치하는 예측값을 얻게 된다. 하지만 동일한 모형으로부터 나타나는 새로

운 자료에 모형을 이용하게 되면 예측력은 매우 낮아지게 된다. 이러한 현상을 훈련자료(training data)에 대한 과대적합(overfitting)이라 부르며, 과대적합이 일어나는 경우 검증자료(test data)에 대한 일반화(generalization)가 어렵게 된다.

이러한 문제를 해결하기 위해서는 최소제곱추정법의 바람직한 성질 중 일정 부분을 완화하여 예측력을 높히는 방법이 활용된다. 차원축소법 및 축소추정법은 추정량의 불편성(unbiasedness)을 포기하는 대신에 추정량의 분산을 획기적으로 줄여 추정량의 안정성을 높이고 이로부터 반응값의 예측정확도를 높인다.

2.1 차원축소법

전술하였듯이, 차원축소법에는 주성분회귀법 및 부분최소제곱회귀법이 있다. 이 방법들을 살펴보기에 앞서 주성분회귀법의 기초가 되는 주성분분석에 대하여 간단히 서술하는 것으로 시작한다.

2.1.1 주성분분석

주성분분석(Principal Component Analysis; PCA)은 다차원의 데이터의 변동(variation)을 최대한으로 유지하면서 차원축소를 행하는 방법이다. 즉, 다차원 변수의 변동을 주성분(principal component)이라는 적은 수의 변수로 축소하되, 다차원 상의 주요 특징들이 주성분을 통하여 대변될 수 있도록 차원을 축소하는 방법이라 할 수 있다. 이때 주성분은 p개의 원변수(original variables) $X = (X_1, X_2, \cdots, X_p)^T$ 벡터의 선형결합이며 주성분벡터 Z는 $Z = \mathbf{A}X$의 형태로 나타낼 수 있다. 즉, 원변수의 변량을 최대한 유지하는 선형결합의 형태를 찾는 작업을 행하는 것이다. 따라서 위 식을 만족하는 행렬 A를 찾는 작업이 주성분분석의 목적이 된다.

원변수 X의 주성분은 원변수의 공분산행렬(Σ)에 의해서 도출된다. 원변수인 벡터

X는 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$인 고유치를 갖는 공분산행렬을 갖는다고 하면 공분산행렬로부터 계산된 고유치와 각각 고유치에 상응하는 고유벡터를 도출하여 이들을 각각 주성분의 분산과 주성분 변수의 계수로 사용한다. 원변수의 선형결합인 주성분변수의 생성에 있어 변수의 계수로 각각의 고유치에 대응하는 고유벡터를 사용한다는 것은 주성분변수간 독립성을 만족한다는 점에서 변수간에 강한 상관관계로 인해 다중공선성의 문제를 해결하기 위한 방안으로 많이 이용된다.

주성분 변수가 가지는 특성을 살펴보면 제1주성분의 경우 원변수 X를 가장 잘 설명하는 성분을 의미한다. 즉, 원변수 X의 변동(분산)을 최대화 하는 성분으로 정의한다. 제2주성분은 제1주성분과 직교한다는 제약조건을 가지면서 제1주성분 다음으로 원변수 X의 변동을 잘 설명할 수 있는 주성분을 의미한다. 같은 방식으로 k번째 ($k = 1, 2, \ldots, p$) 주성분은 앞의 주성분들과 직교성을 가지며 앞의 주성분들이 설명하지 못한 변동 중 가장 큰 변동을 설명하는 방향의 주성분이 된다. 따라서 모든 주성분을 사용할 경우, 원변수의 모든 변동을 설명할 수 있음을 알 수 있다.

크기가 p인 대칭행렬(symmetric matrix) $\mathbf{A} = (a_{ij})_{i,j=1,2,\ldots,p}$를 고려하자. 변수의 개수가 p인 확률벡터 X의 공분산행렬을 $\mathbf{A}$라고 간주해도 무방하다. 음이 아닌 스칼라 λ와 $\mathbf{0}$-벡터가 아닌 벡터 $\mathbf{v}$가

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

를 만족하는 경우 λ를 $\mathbf{A}$의 고유치(eigenvalue), $\mathbf{v}$를 고유벡터(eigenvector)라 한다. 위의 관계를 만족하는 고유벡터는 p개가 존재하며 각기 대응하는 고유치를 가지고 있다. p개의 고유치를 크기 순으로 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ 라 하고 대응되는 고유벡터를 $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p$라 하면, 행렬 $\mathbf{A}$는 다음과 같이 분해된다.

$$\mathbf{A} = \sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^T = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T. \tag{1}$$

식 (1)에서 $\mathbf{\Lambda} = \mathrm{diag}(\lambda_k)_{k=1,2,\ldots,p}$이고 $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p)$이다. $\mathbf{V}$는 직교행렬(orthogonal matrix)로써 $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_p$를 만족한다. 식 (1)에서 차원 p보다 작은 K

에 대해 행렬 $\mathbf{A}$를 다음과 같이 근사시킬수 있다.

$$\mathbf{A} \approx \sum_{k=1}^{K} \lambda_k \mathbf{v}_k \mathbf{v}_k^T = \mathbf{V}_K \mathbf{\Lambda}_K \mathbf{V}_K^T := \mathbf{A}_K,$$

여기서 $\mathbf{\Lambda}_K = \mathrm{diag}(\lambda_k)_{k=1,2,\ldots,K}$이고 $\mathbf{V}_K = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k)$이다. $\mathbf{A}_K$는 행렬 $\mathbf{A}$의 K차원으로 축소된 행렬로써 $\mathbf{A}$와 거리가 가장 가까운 K차원의 대칭행렬로 해석이 가능하다.

2.1.2 주성분회귀법

회귀분석에서 설명변수들 사이에 높은 상관관계가 존재하면 다중공선성(multicollinearity) 현상이 발생한다. 다중공선성이 존재하는 경우, 회귀계수 추정량의 분산은 커지므로 회귀계수 안정성을 보장할 수 없다. 이러한 문제점을 해결하기 위하여 주성분회귀법(Principal Component Regression; PCR)을 사용한다. 주성분회귀모형은 설명변수들에 대한 주성분분석을 통해 수행된다. 주성분분석은 다차원상의 점으로 표현되는 설명변수의 분산을 최대로 하는 선형결합(linear combination) 을 순차적으로 찾아준다. 설명변수의 총변동을 주성분분석을 통해 각각의 주성분 방향으로 분해할 수 있다.

주요 주성분은 본래의 설명변수의 특성을 최대한 유지하면서 낮은 차원에서의 묘사가 가능하므로 차원축소의 주요 방법으로 많이 이용된다. 하위 주성분은 설명변수 간에 강한 상관성을 가지는 차원을 나타낸다. 따라서 주성분회귀모형에서는 하위 주성분을 제거하고 주요 주성분만을 설명변수로 고려하여 회귀분석을 수행하므로 다중공선성의 문제를 해결할 수 있다. 주성분회귀법은 능형회귀법(ridge regression)과 더불어 다중공선성을 해결하는 방법으로 많이 이용되는 전통적인 기법이다. 주성분회귀법은 주성분의 일부만을 이용한다. 따라서 주성분 회귀모형을 통해 얻은 주성분 회귀계수는 편의추정량(biased estimator)이지만, 추정량의 분산을 줄이는 효과가 있어 예측력을 높인다는 점에서 장점이 있으며 이는 능형회귀와 유사한 성질을 갖는다.

위에서 설명한 주성분분석을 이용하여 회귀모형을 설명한다. n개의 자료와 p개의 설명변수로 이루어진 자료행렬을 $\mathbf{X} = (x_{ij})_{i=1,2,\ldots,n;j=1,2,\ldots,p}$라 하자. $\mathbf{X}$는 $n \times p$ 크기의 행렬이다. $\mathbf{X}$의 열은 자신의 표본평균을 빼어 중심화(centering)가 되어 있다고 가정한다. 또한 반응변수 $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ 도 중심화 되었다고 가정한다. 설명변수의 공분산행렬은 $\mathbf{A} = \mathbf{X}^T\mathbf{X}/(n-1)$로 표현된다. 공분산행렬에 대한 고유벡터를 $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p$라 하면 각 고유벡터 $\mathbf{v}_k$에 대하여 대응되는 주성분(principal component)은 다음과 같이 얻는다.

$$\mathbf{z}_k = \mathbf{X}\mathbf{v}_k. \qquad (k = 1, 2, \ldots, p)$$

크기 n인 벡터 $\mathbf{z}_k = (z_{1k}, z_{2k}, \cdots, z_{nk})^T$는 p차원 공간 상의 점으로 표현되는 n개의 자료의 설명변수를 고유벡터 $\mathbf{v}_k$방향으로 직교사영(orthogonal projection)된 점의 좌표이다. 주성분분석에서 $\mathbf{v}_k$를 k번째 주성분방향(PC loading)이라 부르며, 해당 $\mathbf{z}_k$를 k번째 주성분이라 부른다. 즉, i번째 자료의 k번째 주성분은 $z_{ik} = \mathbf{x}_i^T\mathbf{v}_k$로 주어지게 되며, 여기서 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$는 $\mathbf{X}$의 i번째 행이다. 전체 주성분은 다음의 행렬계산으로 쉽게 얻을 수 있다.

$$\mathbf{Z} = (z_{ik}) = \mathbf{X}\mathbf{V}. \qquad (2)$$

식 (2)에서 $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p)$이다. 또한, $\mathbf{Z}^T\mathbf{Z} = \mathbf{V}^T\mathbf{X}^T\mathbf{X}\mathbf{V} = (n-1)\mathbf{V}^T\mathbf{A}\mathbf{V} = (n-1)\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\mathbf{V} = (n-1)\mathbf{\Lambda}$로 주어짐을 확인할 수 있다. 중심화 되어 있는 설명변수와 반응변수로 이루어진 자료에 대한 선형회귀모형은 절편(intercept)이 없는 아래의 모형으로 표현가능하며, 이를 주성분행렬($\mathbf{Z}$)로 표현할 수 있다.

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{Z}\mathbf{V}^T\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}. \end{aligned} \qquad (3)$$

식 (3)에서 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$이고 $\boldsymbol{\gamma} = \mathbf{V}^T\boldsymbol{\beta}$이다. 즉 설명변수 대신 주성분을 설명변수로 하는 회귀분석의 결과는 본래의 설명변수로 표현된 회귀모형과 동등함을 알 수 있다. 변경된 회귀식의 모수 $\boldsymbol{\gamma}$의 추정치는 다음과 같이 주어진다.

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}. \tag{4}$$

식 (4)를 이용하여 본래의 회귀계수의 추정치는 다음 식을 통해 얻을 수 있다.

$$\hat{\boldsymbol{\beta}} = \mathbf{V}\hat{\boldsymbol{\gamma}}.$$

주성분회귀모형은 설명변수의 변동의 대부분을 설명하는 주요 주성분을 사용하여 회귀분석을 수행하는 방법이다. 즉, 전체 p개의 주성분을 이용하는 대신, 상위 K개의 주성분 만을 설명하는 변수를 고려한다. 이를 이해하기 위해서 $\hat{\boldsymbol{\beta}}$의 분산을 계산하면 다음과 같다.

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^T = \sigma^2\sum_{k=1}^{p}\frac{1}{\lambda_k}\mathbf{v}_k\mathbf{v}_k^T.$$

하위 고유치값이 0에 가깝다면 추정량의 분산을 매우 커지게 됨을 알 수 있다. 따라서 고유치의 크기가 0에 가까운 하위 주성분을 제거한다면 추정치의 분산을 줄일 수 있다. 따라서 주성분회귀법은 하위 주성분을 제거함으로써 추정량의 분산을 줄여주며, 고차원의 자료를 저차원 상에서 분석할 수 있다는 장점을 가지고 있다. 하지만 차원을 줄인다고 해도 실질적으로 본래의 설명변수를 모두 가지고 있으므로 해석이 용이하지 않다는 단점을 가지고 있으며, 상위 주성분만을 가지고 분석하기 때문에 실질적으로 하위 주성분에 반응변수와 연관성이 높은 중요한 정보가 있다고 해도 그 정보를 사용하지 못하는 단점 또한 지니고 있다.

2.1.3 부분최소제곱회귀법

부분최소제곱회귀법(Partial Least Squares Regression; PLSR)은 반응변수와 설명변수의 수가 많은 경우에 주로 사용되는 방법이며 반응변수가 한 개인 일반적인 회귀분석 모형 하에서도 적용이 가능하다. 주성분분석이 설명변수의 분산만을 고려하여 변수들의 선형결합을 통해 새로운 변수를 유도하는 반면, 부분최소제곱회귀법은 설명변수와 반응변수를 동시에 고려하기 때문에 분석에 있어 복잡한 구조적인 문제가 발생하는 부분을 줄여줄 수 있다는 장점을 가지고 있다. 부분최소제곱법에서 사용되는 새로운 주성분 변수를 잠재변수(latent variable)라 하며 n개의 관측치와 q개의 반응변수로로 이루어진 $n \times q$행렬 $\mathbf{Y}$와 n개 관측치에 대하여 수집된 p개의 설명변수로 이루어진 $n \times p$ 행렬 $\mathbf{X}$를 두 잠재변수인 $\mathbf{T}$와 $\mathbf{U}$의 선형관계로 가정하는 모형을 사용하면 다음과 같이 표현된다.

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^T + \mathbf{E}, \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^T + \mathbf{F}. \end{aligned} \tag{5}$$

식 (5)에서 $n \times k$ 행렬 $\mathbf{T}$와 $\mathbf{U}$는 추측된 k개의 잠재변수(latent variables)들을 열로 가지며, $p \times k$ 행렬 $\mathbf{P}$와 $q \times k$ 행렬 $\mathbf{Q}$는 적재(loadings)행렬이라하고 나머지 $n \times p$행렬 $\mathbf{E}$와 $n \times q$행렬 $\mathbf{F}$는 잔차행렬이다.

부분최소제곱법은 비선형 반복 알고리즘을 이용하여 식 (5)의 모수들을 추정하게 되는데 이는 다음 식을 만족한다.

$$\max_{\mathbf{w},\mathbf{c}} [\mathrm{cor}(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c})]^2 = \max_{\mathbf{w},\mathbf{c}} [\mathbf{w}^T\mathbf{X}^T\mathbf{Y}\mathbf{c}] = \max_{\mathbf{t},\mathbf{u}} [\mathbf{t}^T\mathbf{u}],$$
$$\text{subject to } \mathbf{w}^T\mathbf{w} = 1 \text{ and } \mathbf{t}^T\mathbf{t} = 1$$

위 식에서 $\mathbf{w}$와 $\mathbf{c}$를 순차적으로 구하게 되면 $\mathbf{t}$는 $\mathbf{X}\mathbf{w}$이고 $\mathbf{u}$는 $\mathbf{Y}\mathbf{c}$를 만족한다. $\mathbf{P}$와 $\mathbf{Q}$를 구성하는 적재벡터 $\mathbf{p}$와 $\mathbf{q}$들은 각각 $\mathbf{X}$와 $\mathbf{t}$ 및 $\mathbf{Y}$와 $\mathbf{u}$의 선형회귀계수들로 구성되며

$\mathbf{p} = \mathbf{X}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$ 및 $\mathbf{q} = \mathbf{Y}^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})$로 주어진다. 이러한 과정하에 얻어진 $\mathbf{t}$와 $\mathbf{p}$는 다음 반복에서의 반응변수와 설명변수를 다음과 같이 감쇠(deflation) 과정에 이용되며, 감쇠된 $\mathbf{X}$와 $\mathbf{Y}$는 다음 성분을 얻기 위해 순차적으로 수행된다.

$$\mathbf{X} \leftarrow \mathbf{X} - \mathbf{t}\mathbf{p}^T,$$
$$\mathbf{Y} \leftarrow \mathbf{Y} - \mathbf{t}\mathbf{t}^T\mathbf{Y}/(\mathbf{t}^T\mathbf{t}).$$

부분최소제곱회귀법은 반응변수와 설명변수들에 대한 감쇠방법에 따라 다양한 알고리즘이 존재하는데 회귀분석을 목적으로 하는 경우, 위 식에서와 같은 감쇠방법을 사용하는 것이 일반적이다. 이 알고리즘을 회귀분석 목적으로 하는 경우 점수벡터 $\mathbf{t}$들은 $\mathbf{Y}$에 대한 설명변수들이며 점수벡터들 $\mathbf{t}$와 $\mathbf{u}$사이에는 $\mathbf{U} = \mathbf{T}\mathbf{D} + \mathbf{H}$의 선형관계가 성립한다. 위에서 $\mathbf{D}$는 $q \times q$인 대각행렬이고 $\mathbf{H}$는 잔차행렬이 된다. 추가된 가정에 의해서 부분최소제곱회귀모형은 다음과 같이 표현할 수 있다.

$$\mathbf{Y} = \mathbf{T}\mathbf{D}\mathbf{Q}^T + (\mathbf{H}\mathbf{Q}^T + \mathbf{F}) = \mathbf{T}\mathbf{C}^T + \mathbf{F}^*,$$

여기서 $\mathbf{C}^T = \mathbf{D}\mathbf{Q}^T$이며 이는 $p \times q$ 인 회귀계수 행렬이고 $\mathbf{F}^* = \mathbf{H}\mathbf{Q}^T + \mathbf{F}$는 잔차행렬이 된다. 이 식은 반응변수 $\mathbf{Y}$를 최소제곱법을 이용하여 서로 직교인 설명변수 $\mathbf{T}$로 분해한 것과 같다. 위 추정방법에 의하여 추정된 회귀계수 또한 주성분회귀분석과 마찬가지로 차원을 줄인다고 해도 실질적으로 본래의 설명변수를 모두 가지고 있으므로 해석이 용의 하지 않다는 단점을 가지고 있다. 하지만 차원축소 과정에서 반응변수와의 상관계수가 최대가 되도록 하는 방향을 찾으므로 주성분회귀에서와 같이 하위 주성분의 제거에 기인한 정보의 손실이 적다는 장점을 가지고 있다.

2.2 축소추정법

독립변수들 사이에 다중공선성이 존재하는 경우 최소제곱추정량의 분산이 매우 커지게 되어 추정량으로서의 활용도가 떨어지게 된다. 이러한 경우 작은 편이(bias)를 허용하여 회귀계수의 크기를 축소시킴으로써 모형의 안정화시키며 추정량의 분산을 작게 하는 작업이 도움이 된다. 또한 선형회귀분석에서 자료의 개수 n보다 설명변수의 개수 p가 더 큰 경우, 회귀계수의 추정량은 유일하게 결정되지 않는다. 또한 $n > p$일지라도, n이 p보다 월등히 크지 않은 경우 회귀계수 추정량의 분산을 매우 크게 나타난다. 이는 회귀계수 추정량의 신뢰성이 낮아지게 되고 결과적으로 모형으로 부터 높은 예측력을 기대하기가 어렵다. 따라서 이러한 문제를 해결하기 위하여 계수축소법(coefficient shrinkage)이 사용된다. 이 방법은 기존의 최소제곱추정법(Ordinary Least Squares; OLS) 방법에 벌점을 부여함으로써 회귀계수를 축소하는 방법이다. 본 논문에서는 계수축소의 다양한 방법 중 능형회귀법, lasso 회귀법, SCAD 회귀법을 살펴본다.

2.2.1 능형회귀

능형회귀(ridge regression)모형은 회귀계수 추정량에 약간의 편이(bias)를 주는 대신 분산을 크게 줄임으로써 예측력을 향상시켜 준다. 아래와 같은 선형회귀모형을 가정하자.

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \boldsymbol{\epsilon}_i. \qquad (i = 1, 2, \cdots, n)$$

최소제곱추정량(Ordinary Least Square Estimator)은 다음의 잔차제곱합(Residual Sum of Squares; RSS)을 최소로 하는 모수값으로 다음의 최적화 과정을 통해 얻을수 있다.

$$\underset{\beta_1, \cdots, \beta_p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2.$$

능형회귀추정량은 다음과 같이 특정 제약조건하에서 잔차제곱합을 최소로 하는 모수 값이다.

$$\underset{\beta_1,\cdots,\beta_p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2,$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \le t. \tag{6}$$

식 (6)에서 t는 양수로써 모수의 크기를 조절하는 조절모수(regularization parameter)이다. t가 충분히 크면 모수 크기에 대한 제약이 없기 때문에 최소제곱추정량과 동일한 추정량을 준다. 반면 조절모수 t가 작아지면 추정량의 크기가 작아지는 효과를 가져올 수 있다. 이러한 효과는 중요하지 않은 변수의 회귀계수는 0이 되는 것이 아니라 0에 가까워 진다. 따라서 능형회귀분석은 변수선택의 방법으로는 사용하기 힘들며, 모든 설명변수를 포함하기 때문에 해석에는 용이하지 않음을 알 수 있다. 또한 식 (6)에서의 능형회귀분석을 위한 최적화 문제는 아래와 같이 벌점함수를 이용한 목적함수의 최소화 문제로 표현 가능하다.

$$\underset{\beta_1,\cdots,\beta_p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

여기서 양수 λ는 추정량의 크기에 제약을 주는 조절모수로써, 식 (6)에서의 조절모수 t와 밀접한 연관성을 가진다. t가 0에 가까우면 λ는 $+\infty$에 가까워지며, t가 일정값 이상 크게 되면 λ는 0으로 주어지게 된다.

2.2.2 Lasso회귀

Tibshirani(1996)는 능형회귀모형에서 회귀계수가 0이 되는 확률이 작음을 보완하기 위하여 능형회귀모형의 L_2 제약조건 대신 L_1 제약조건을 제시하였다. 이러한 제약식 하에서 수행되는 회귀분석을 lasso 회귀법이라 한다. Lasso 회귀모형으로부터 얻게되는 추정량은 능형회귀추정량보다 쉽게 0을 가지게 된다. 이는 결과적으로 변수선택

(variable selection)의 효과를 주게 되며, 모형선택(model selection)기법으로 최근 많이 활용되고 있다. Lasso 회귀추정량은 다음과 같이 특정 제약조건하에서 잔차제곱합을 최소로하는 모수값이다.

$$\underset{\beta_1,\cdots,\beta_p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2,$$
$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \le t. \tag{7}$$

식 (7)에서 t는 양수로써 능형회귀법과 유사하게 모수의 크기를 조절하는 조절모수(regularization parameter)이다. t가 충분히 크면 모수에 대한 크기 제약이 없기 때문에 최소제곱추정량과 동일한 추정량을 준다. 반면에 t가 충분히 작다면 추정량의 크기는 작아지므로 중요하지 않은 변수의 회귀계수는 0이 된다. 즉, $t = 0$을 만족한다면 모든 추정치는 0이 된다는 것을 의미한다. 한편 적절한 크기의 t가 주어진다면 추정치의 일부분은 정확히 0으로 주어지게 된다. 이러한 점은 능형회귀법과의 중요한 차이점이다. 식 (7)을 lasso 회귀분석을 위한 최소화 함수는 벌점함수를 이용한 목적함수의 최소화 함수로 표현가능하다.

$$\underset{\beta_1,\cdots,\beta_p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

위 식에서 $\lambda(\ge 0)$는 조절모수로써 식 (7)에서의 조절모수 t와 같은 유사한 역할을 한다. λ가 0에 가까워지면 벌점함수의 역할이 축소되어 자유롭게 $\boldsymbol{\beta}$가 추정되므로 최소제곱추정량과 동일한 추정량을 얻게되며, $\lambda \to \infty$이면 모든 회귀계수가 0으로 수렴하게 되다. 결과적으로 lasso회귀모형은 몇 개의 덜 유의한 회귀계수가 0이 되어 축소되고, 축소된 회귀계수는 편이(bias)가 발생하게 되지만 분산(variance)은 매우 작아지게 되어 전반적으로 예측 정확도가 높아진다. 또한 변수선택과 모형선택이 용이하고 능형회귀추정량의 분산을 크게 줄이는 효과를 가지게 되어 능형회귀모형 보다는 더 좋은 추정치를 가진다 할 수 있다.

2.2.3 능형회귀와 lasso회귀의 비교

능형회귀분석과 lasso회귀분석을 비교하는 과정에서 이해를 쉽게 하기 위해 $p = 2$인 경우를 가정한다. 그림 1은 최소제곱추정치에 따른 능형회귀추정치와 lasso추정치의 변화를 나타낸 것이다. 그림 1에서 붉은색 실선은 능형회귀 및 lasso회귀의 추정치이며 검정색 점선은 최소제곱추정치와의 비교를 위한 참조선이다. 최소제곱추정치에 비해 두 추정치가 일괄적으로 작게 추정되며 최소제곱추정치가 일정수준 이하의 값을 가지게 되면 lasso회귀 추정치는 정확히 0으로 주어지는 것을 볼 수 있다. 이는 능형회귀분석은 회귀계수를 축소시키는 역할을 하지만 lasso 회귀분석은 추정치 값을 정확히 0으로 만들어 변수선택의 역할을 하는 것을 알 수 있다.

그림 1: 최소제곱회귀추정치와 lasso회귀 및 능형회귀 추정치의 비교그림.

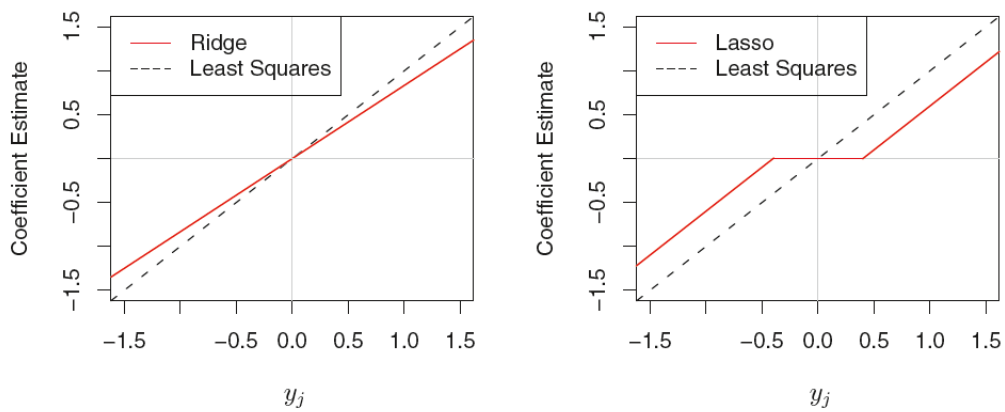
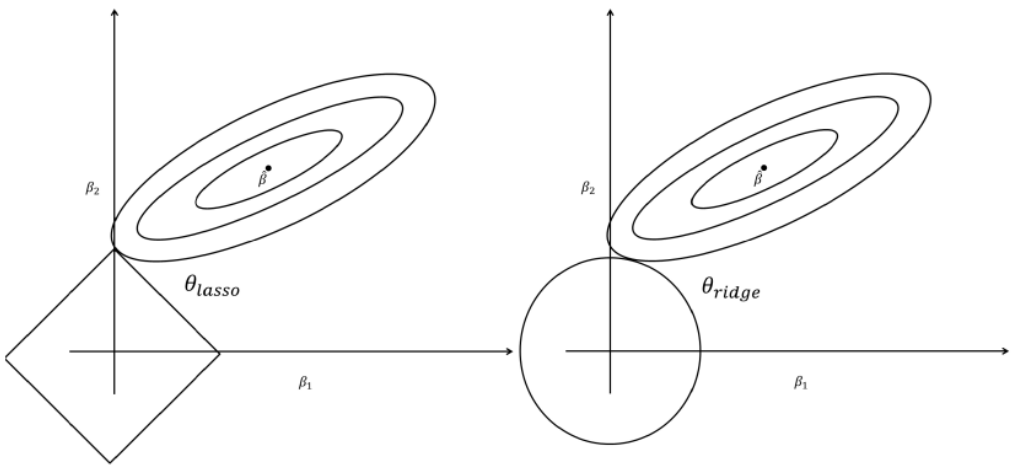


그림 2는 능형회귀모형의 L_2 제약식 $\beta_1^2 + \beta_2^2 \leq t$와 lasso 회귀모형의 L_1 제약식 $|\beta_1| + |\beta_2| \leq t$ 의 영역 하에서 회귀계수추정량을 얻는 과정에 대한 기하학적 묘사를 주고 있다. 두 방법은 모두 동일한 RSS를 가지고 있으며 이는 그림상에서 타원형의

그림 2: Lasso회귀법 및 능형회귀법의 기하학적 형태의 비교그림.



등고선으로 표현된다. 그림에 등고선의 가장 낮은 위치에 표현된 점은 최소제곱추정치($\hat{\beta}$)에 대응된다. 각 방법의 추정치는 제약식 영역하에서 RSS 값을 최소로 하는 점에서 표현된다. Lasso 회귀분석의 경우 제약식이 마름모 형태로 표현될 수 있으며, 능형회귀분석의 경우는 원형으로 주어지게 된다. 그림 2에서 제약영역 하에서 RSS의 최소값을 탐색하기 위해 등고선을 최소제곱추정량으로부터 점차 변화시켜 제약영역에 처음으로 만나는 점을 찾는다. 이때 제약영역의 점 중에서 등고선을 최초로 만나게 되는 점이 각 방법에 대응하는 회귀계수 추정치가 된다. 따라서 제약식이 마름모 형태를 가지는 lasso 회귀분석은 모서리가 축 위에 놓이며 뾰족하게 튀어나와 있어 등고선이 제약영역의 꼭지점에 닿을 확률이 높음을 알 수 있고 꼭지점은 축의 일부분 값이 0에 대응하는 점에 놓여 있으므로 lasso 회귀계수추정량은 일부분이 0이 될 가능성이 높게 된다. 반면 능형회귀분석의 경우 제약식이 원형의 형태를 가지므로 축에 등고선이 만날 확률이

낮다. 즉, 능형회귀계수 추정량의 일부분이 0이 될 가능성이 매우 낮음을 알 수 있다.

2.2.4 SCAD회귀

SCAD(Smoothly Clipped Absolute Deviation: SCAD) 벌점함수는 Fan and Li(2001)에 의해 제안되었다. Lasso 회귀법은 0에 가까운 추정치를 본질적으로 0과 차이가 없다고 판단하여 정확히 0으로 주어지는 장점이 있지만, 0이 아닌 추정치 또한 일괄적으로 0 방향으로 크기를 줄인다. 이는 잠재적으로 회귀계수추정량으로 하여금 과도하게 편이(bias)를 주게 되고, 결과적으로 전반적인 예측력에 영향을 주게 된다. SCAD 회귀분석은 lasso 회귀분석의 단점을 보완하도록 고안되었는데, 일정값 이상 큰 추정량에 대하여 편이를 전혀 주지 않도록 설계되어 있다.

SCAD 벌점함수는 다음과 같이 정의 된다.

$$P_\lambda(\theta) = \lambda \left\{ I(\theta \le \lambda) + \frac{a\lambda - \theta}{(a-1)\lambda} + I(\theta > \lambda) \right\} \tag{8}$$

위 식에서 a와 λ는 벌점함수의 형태 및 크기를 조절하는 조절모수이다. SCAD 벌점함수를 이용하여 회귀분석의 목적함수는 다음과 같이 표현된다.

$$\underset{\beta_1,\cdots,\beta_p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \sum_{j=1}^{p} P_\lambda(\beta_j)$$

SCAD 회귀분석은 lasso 회귀분석의 추정량의 성질을 개선한 회귀모형으로 많이 이용되고 있다.

3 벌점회귀에 기반한 주성분회귀

주성분회귀에서는 상위주성분을 설명변수로 하여 차원축소를 통해 추정량의 성능을 개선하지만, 하위주성분이 반응변수와 연관성이 높은 경우 이를 활용하지 않음으로 인하여 효율성이 떨어지는 단점을 지적하였다. 이를 개선하는 직관적인 방법은 주성분을 설명변수로 사용하되 반응변수와 연관성이 높은 주성분을 선택하여 이를 회귀모형에 활용하는 것이다. 본 연구에서는 주성분을 선택하는 방법으로써 lasso 혹은 SCAD 회귀법을 통하여 주성분을 자동으로 선별하는 방법을 제안한다. 이를 통해 반응변수에 유의한 영향을 주는 주성분은 하위주성분이라 하더라도 모형에 들어감으로써 설명력의 감소를 방지할 수 있고, 상위주성분이라 하더라도 반응변수에 대한 설명력이 작으면 벌점최적화에 의해 최종모형에서 배제함으로써 최적의 주성분 집합을 구성한다.

크기 $n \times p$인 자료행렬 $\mathbf{X} = [\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_p^T]$는 n개의 자료로부터 얻은 p개의 설명변수로 구성된 자료행렬이라 하자. 즉, $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$는 i번째 관측치의 설명변수의 모임으로 해석할 수 있다. 이를 이용하여 다중선형회귀모형

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

으로 표현할 수 있다. 설명변수를 주성분으로 변환하면

$$\mathbf{Z} = \mathbf{X}\mathbf{V}$$

로 표현할 수 있으며, 변환한 새로운 변수를 이용하여 다음과 같은 모형을 얻을 수 있다.

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{Z}\mathbf{V}^T\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}. \end{aligned} \tag{9}$$

여기서 $\boldsymbol{\gamma} = \mathbf{V}^T\boldsymbol{\beta}$이다. 행렬 $\mathbf{Z}$를 구성하는 열벡터 $\tilde{\mathbf{z}}_j$ $(j = 1, 2, \ldots, p)$는 j번째 주성분으로 주어진다. 회귀분석시 이 주성분의 전부를 설명변수로 삼고 반응변수에 회귀적합을

한다면 주성분 $\mathbf{Z}$에 대한 회귀계수 $\boldsymbol{\gamma}$를 추정할 수 있다. 그 추정치를 원래 모형의 회귀계수 $\boldsymbol{\beta}$로 역변환하여 추정량을 구하면 정확히 최소제곱회귀추정량이 된다. (9)에서 $\boldsymbol{\gamma} = \mathbf{V}^T\boldsymbol{\beta}$의 관계로부터 $\hat{\boldsymbol{\beta}}$의 최소제곱추정량은

$$\hat{\boldsymbol{\beta}} = \mathbf{V}\hat{\boldsymbol{\gamma}}$$

임을 알 수 있다. 여기서 모수 $\boldsymbol{\beta}$에 대하여 추정할 때 모든 주성분을 사용하여 $\boldsymbol{\gamma}$를 추정 한다면 최소제곱추정량과 동일한 결과를 얻는다.

일반적인 주성분회귀분석은 작은 고유치와 대응되는 몇 개의 주성분을 제거시켜 차원축소를 한다. 하지만 이러한 주성분의 선택은 작은 고유치에대한 기준이 존재하지 않아 주관적인 결정을 해야만 하는 단점이 있으며, 하위주성분의 일괄적인 배제로 인하여 효율성이 떨어질 우려가 있다. 따라서 이러한 주성분 선택의 방법을 본 논문에서는 벌점함수의 도입으로 선택하고자 한다. 이러한 분석법을 본 연구에서는 벌점주성분회귀분석법(Penalized Principal Component Regression; PPCR) 이라고 부른다. 본 연구에서 벌점함수로써 능형벌점(ridge penalty), lasso 벌점(lasso penalty), SCAD 벌점(SCAD penalty)등을 이용한다. 벌점함수의 형태는 다음과 같다.

- 능형벌점함수

$$\lambda \sum_{j=1}^{p} \beta_j^2$$

- Lasso벌점함수

$$\lambda \sum_{j=1}^{p} |\beta_j|$$

- SCAD벌점함수

$$\sum_{j=1}^{p} P_\lambda(\beta_j)$$

SCAD벌점함수에서 $P_\lambda(\beta_j)$는 식 (8)로 주어진다. 위의 벌점함수를 적용한 방법을 각각 PPCR-R(Penalized Principal Component Regression with Ridge), PPCR-L(Penalized Principal Component Regression with Lasso), PPCR-S(Penalized Principal Component Regression with SCAD)라고 부르기로 한다.

4 모의 실험

본 장에서는 3장에서 소개한 벌점회귀에 기반한 주성분의 선별방법을 모의실험에 적용하고 각각의 분류법들의 성능을 비교한다.

4.1 모의 실험 설계

본 논문에서는 통계프로그램 R을 사용하여 다음과 같이 모의실험 상황을 설정하였다. 자료의 수를 n, 변수의 개수가 p인 설명변수 행렬을 $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_n^T)^T$라 하자. 설명변수와 반응변수의 관계는 다음과 같은 선형식을 가정한다.

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \qquad (i = 1, 2, \ldots, n)$$

컴퓨터 모의실험을 위하여 설명변수 $\mathbf{X}$, 회귀계수 $\boldsymbol{\beta}$ 그리고 반응변수 $\mathbf{y}$를 생성하기 위해 다음 작업을 수행한다.

(step 1) $\mathbf{U} = (\mathbf{u}_1^T, \mathbf{u}_2^T, \cdots, \mathbf{u}_{5000}^T)^T = (u_{ij}) \in \mathbb{R}^{5000 \times p}$의 생성 :

$5000 \times p$개의 확률변수 $u_1, u_2, \cdots, u_{5000\times p}$를 $N(0,1)$로부터 랜덤하게 생성하여 $\mathbf{U}$를 만든다.

(step 2) 위의 5000개의 샘플에 대한 표본공분산행렬을 계산 :

$\boldsymbol{\Sigma}^* = \frac{1}{n}\sum_{i=1}^{5000}(\mathbf{u}_i - \bar{\mathbf{u}})(\mathbf{u}_i - \bar{\mathbf{u}})^T$를 계산한다. 여기서 $\bar{\mathbf{u}} = \frac{1}{5000}\sum_{i=1}^{5000}\mathbf{u}_i$이다.

(step 3) 고유치분해(eigen decomposition)을 통한 직교행렬 $\mathbf{V}$ 획득 :

$\mathbf{\Sigma}^* = \mathbf{V}\mathbf{D}\mathbf{V}^T$ 계산하여 $\mathbf{V}$를 이어지는 계산에 활용한다. 여기서 d_ℓ은 ℓ 번째 고유치, $\mathbf{v}_\ell$은 해당 고유벡터이며, $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p)$이고, $\mathbf{D} = \mathrm{diag}(d_1, d_2, \ldots, d_p)$이다.

(step 4) n개의 자료에 대한 주성분점수의 생성 :

n개의 주성분 $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$을 $N_p(\mathbf{0}, \mathbf{\Lambda})$로부터 랜덤하게 생성하여 주성분행렬 $\mathbf{Z} = (\mathbf{z}_1^T, \mathbf{z}_2^T, \cdots, \mathbf{z}_n^T)^T$를 구성한다. 여기서 $\mathbf{\Lambda}$는 $\lambda_j = 2^{1-j}$ $(j = 1, 2, \ldots, p)$를 원소로 하는 대각행렬로 정의한다. λ_j는 j번째 주성분의 분산이다.

(step 5) $\mathbf{X} = \mathbf{Z}\mathbf{V}^T$의 변환을 통한 설명변수의 생성 :

$\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_n^T)^T$이며 $\mathbf{x}_i = \mathbf{V}\mathbf{z}_i$ $(i = 1, 2, \ldots, n)$이다.

(step 6) $\mathbf{Z}$를 이용하여 반응변수 $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$을 생성 :

미리 설정한 $\boldsymbol{\gamma}$를 이용하여 반응변수를 $y_i = \mathbf{z}_i^T\boldsymbol{\gamma} + \epsilon_i$ $(i = 1, 2, ..., n)$를 통해 생성한다. 여기서 오차는 $\epsilon_i \sim N(0, 0.1^2)$로 생성한다.

위에서 반응변수 $\mathbf{y}$는 $\mathbf{Z}$와 $\boldsymbol{\gamma}$에 의해 생성된다. 이와 같은 정보를 갖는 설명변수는 $\mathbf{X}$이고 해당 회귀계수는 $\boldsymbol{\beta} = \mathbf{V}\boldsymbol{\gamma}$로 주어진다. 컴퓨터 모의실험에서는 주성분 및 주성분에 대응하는 회귀계수 $\boldsymbol{\gamma}$를 알고 있지만, 일반적인 회귀분석을 수행하게 되는 경우 설명변수 $\mathbf{X}$ 및 반응변수 $\mathbf{y}$만 주어지고 이를 기반으로 회귀분석을 수행하게 된다. 위 과정은 $\mathbf{x}_i$들을 공분산 $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$이고 평균이 $\mathbf{0}$인 다변수정규분포에서 생성한 것과 동일함을 알 수 있다.

모의실험 자료를 생성하기 위해서는 $\boldsymbol{\gamma}$를 설정하여야 한다. γ_j $(j = 1, 2, \ldots, p)$의 값은 j번째 주성분과 반응변수의 연관성의 크기로 나타내며, $\gamma_j = 0$은 j번째 주성분이 반응변수에 대한 설명력이 전혀 없음을 의미한다. 주성분과 반응변수 간의 관계를 아래와 같이 모형화 한다.

(T1) 상위 3개의 주성분만이 반응변수에 동일한 설명력을 가지는 경우 :
$\boldsymbol{\gamma} = (1, 1, 1, 0, \ldots, 0)^T$로 설정한다.

(T2) 4, 5, 6번째 주성분만이 반응변수에 동일한 설명력을 가지는 경우:
$\boldsymbol{\gamma} = (0, 0, 0, 1, 1, 1, 0, \ldots, 0)^T$로 설정한다.

(T3) 상위 6개의 주성분 중 랜덤하게 3개만을 선택하여 반응변수에 동일한 설명력을 가지는 경우:
$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_6, 0, \ldots, 0)^T$로 두되, $\gamma_1, \ldots, \gamma_6$ 중 랜덤하게 3개를 선택하여 1을 부여하고 나머지는 0을 부여한다.

위의 3가지 상황(**T1**∼**T3**)에 대하여 각 자료를 동등하게 n개씩 훈련자료(training data)와 검증자료(test data)를 생성한다. 표본의 개수(n)와 변수 개수(p)는 아래와 같이 설정하였다.

- 자료 개수가 변수 개수보다 많은 경우: $p = 10$, $n = 100, 500, 1000, 1500$
- 변수 개수가 자료 개수보다 많은 경우: $n = 100$, $p = 100, 500, 1000, 1500$

위와 같이 생성한 자료를 기반으로 본 연구에서 제안한 방법과 다양한 회귀분석 방법을 적용하여 성능을 비교한다. 비교하고자 하는 8가지 방법들은 다음과 같다.

(M1) PPCR-L : 주성분을 설명변수로 하고 lasso penalty를 이용하여 주성분을 선정하고 회귀계수를 추정함. glmnet 패키지 이용.

(M2) PPCR-R : 주성분을 설명변수로 하고 ridge penalty를 이용하여 회귀계수를 추정함. glmnet 패키지 이용.

(M3) PPCR-S : 주성분을 설명변수로 하고 SCAD penalty를 이용하여 주성분을 선정하고 회귀계수를 추정함. lqa 패키지 이용.

(M4) PCR : 원변수 $\mathbf{X}$를 설명변수로 하는 주성분회귀분석. pls 패키지 이용.

(M5) REG-L : 원변수 $\mathbf{X}$를 사용한 lasso 회귀분석. glmnet 패키지 이용.

(M6) REG-R : 원변수 $\mathbf{X}$를 사용한 능형회귀분석. glmnet 패키지 이용.

(M7) REG-S : 원변수 $\mathbf{X}$를 사용한 SCAD 회귀분석. lqa 패키지 이용.

(M8) PLSR : 원변수 $\mathbf{X}$를 부분최소제곱회귀분석. pls 패키지 이용.

방법 **(M1)**∼**(M3)**는 본 연구에서 제안한 방법으로서 주성분을 설명변수로 사용한다. 여기서 주성분은 **(step 4)**에서 자료를 생성하기 위해 준비한 $\mathbf{Z}$를 사용하지 않고 설명변수 행렬 $\mathbf{X}$의 주성분분석을 통해 얻은 주성분을 사용하였다. 이는 실제자료가 주어졌을 때 주성분을 추정해서 사용해야 하는 현실적인 상황을 반영하기 위함이다. 고려하는 방법들 중 PCR과 PCLR을 제외한 모든 방법은 각기 모형선택(model selection)을 위한 조절모수를 포함하고 있다. 이를 10-fold cross-validation(CV)을 통해 최적의 모수를 선택하였다. 이 경우 가급적 사용하는 패키지에서 제공되는 기본옵션을 이용하였다.

위의 각 방법을 통해 얻은 회귀계수추정량을 $\hat{\boldsymbol{\beta}}$라 하고, 반응변수의 예측값을 $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$라 하자. 각 방법의 성능을 평가하기 위해 추정오차(estimation error, EE)와 예측오차(prediction error, PE)를 다음과 같이 계산하였다.

$$\text{EE} = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = \sqrt{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})},$$
$$\text{PE} = \|\hat{\mathbf{y}} - \mathbf{y}\| = \sqrt{\textstyle\sum_{i=1}^{n}(\hat{y}_i - y_i)^2/n}.$$

위의 추정오차가 작다는 것은 추정된 회귀계수 $\hat{\boldsymbol{\beta}}$가 실제 $\boldsymbol{\beta}$와 큰 차이가 없으며 이는 회귀계수를 잘 추정하였다는 것을 의미한다. 또한 예측오차가 작다는 것은 모형으로부터 예측된 반응변수가 실제 관측된 반응변수와 큰 차이가 없으며 이는 모형의 예측력이

좋다는 것을 의미한다. 따라서 위에서 구한 EE와 PE값이 작을수록 해당 모형이 우수하다고 할 수 있다. 모의실험을 반복할 때마다 비교결과가 항상 동일하지 않기 때문에 위의 실험을 반복하여 수행하였다. 각 상황에 대하여 1,000회의 모의실험을 반복하고 각 결과로부터 얻어진 1,000개의 추정오차 및 예측오차의 평균값을 구하여 각 모형의 성능을 비교하였다.

4.2 모의 실험 결과

위 절에서 제안한 8가지 방법에 대하여 모의실험 설계에 의해서 얻었다. 모의실험 설계에서 제안한 3가지 상황(**T1**∼**T3**)에 대하여 1,000회의 자료를 반복생성하고 각 방법을 이용하여 EE와 PE값을 구하였다. 각 방법에 대한 EE와 PE의 평균은 6절에 제공한 표 1 ∼ 표 4에 수록하였다. 표에서 볼드체로 표기한 수치는 각 상황에서 가장 작은 값을 의미한다. 표 1은 변수수를 $p = 10$으로 고정한 경우 자료수($n = 100, 500, 1000, 1500$)가 변화하는 상황에서의 각 상황별 EE 값을 나타내며, 표 2는 자료수를 $n = 100$으로 고정한 경우 변수수($p = 100, 500, 1000, 1500$)가 변화하는 상황에서의 각 상황별 EE 값을 나타낸다. 표 3은 표 1과 동일한 조건에서의 PE 값을 나타내며, 표 4는 표 2와 동일한 조건에서의 각 상황별 PE 값을 나타낸다.

표 1은 동일한 변수수일 때 자료수가 커짐에 따라 EE의 평균이 작아짐을 보여주고 있다. 또한 각 방법에 대하여 주성분의 크기를 규제하거나 선별하는 경우(PPCR 방법들)가 본래의 설명변수를 이용한 경우(REG 방법들)에 비해 상대적으로 낮은 EE값을 가지는 것을 볼 수 있다. 상위 주성분만을 취하는 주성분회귀법(PCR)나 반응변수와의 설명력을 기반으로 차원을 축소하는 부분최소제곱회귀법(PLSR)에 비해 PPCR이 전체적으로 나은 EE값을 보여주고 있지는 않지만, lasso를 사용한 PPCR-L은 대부분의 경우에 있어서 낮은 EE값을 보여주어 회귀계수의 추정오차 측면에서 PPCR-L이 가장 좋은 성능을 보여주고 있다. 구체적으로 살펴보면, 상황 **T1**에서는 PCR이 좋은 성능을

보여줄 수 있는 상황이지만, PPCR-L이 가장 좋은 성능을 보여주고 있으며, 반면 PCR과 PLSR이 PPCL-R 및 PPCL-S 보다 더 좋은 추정오차를 보여주고 있다. 이는 능형회귀에 기반한 PPCL-R은 주성분의 크기에 대한 규제를 취할 뿐 주성분 자체에 대한 선별이 이루어 지지 않으므로 좋은 성능을 기대하기 어렵기 때문으로 판단되며, 선별능력이 있는 SCAD에 기반한 PPCL-S가 그리 좋은 성능을 보여지 못하는 것은 의외의 결과이다. 상황 **T2** 및 **T3**에서도 **T1**상황과 마찬가지로 자료수가 커짐에 따라 EE의 평균이 작아짐을 확인할 수 있다. $(n, p) = (100, 10)$ 등 몇몇의 예외적 상황을 제외하곤 원 설명변수를 이용한 경우보다 주성분을 이용한 경우가 약간의 낮은 EE값을 가짐을 알 수 있다. 그러나 이 경우를 제외하고 모두 상황 **T1**과 같은 결과를 가짐을 알 수 있다. 따라서 자료의 수가 많다면 **T2**에서도 본 논문에서 주장하는 주성분을 기반한 벌점회귀(PPCR-L)가 더 정확한 결과를 가진다는 것을 알 수 있다.

표 2는 상대적으로 변수(p)가 자료수(n)보다 큰 고차원 자료의 경우로, 자료수를 $n = 100$으로 고정하고 변수(p)를 달리 하였을 경우의 EE를 보여준다. 여기서 변수수가 1000과 1500인 경우 SCAD를 이용한 방법인 PPSR-S와 REG-S에 빈칸으로 주어져 있는데, 이는 SCAD 기법을 적용하는데 이용된 `lqa` 패키지가 변수수 1000 이상인 경우 적합하는데 매우 오랜 시간이 필요하거나 알고리즘의 수렴이 이루어지지 않음을 확인할 수 있었고 이로 인해 알고리즘이 개선된 패키지가 제공되지 않는 한 현실문제에 적용하기에는 어렵다고 판단되므로 결과에서 제외하였다. 각 방법에 대하여 주성분을 이용한 경우가 본래의 설명변수를 이용한 경우에 비해 능형회귀법을 이용한 경우를 제외하곤 대략적으로 작은 EE 값을 가짐을 확인할 수 있다. 이들과는 대조적으로 PCR과 PLSR는 추정오차가 매우 큼을 볼 수 있다. 이는 두 방법을 통해 얻어지는 필요한 차원의 수가 실제 차원인 3보다 매우 크게 얻어지게 되어 회귀계수에 대한 추정이 신뢰하기 어렵다는 결과를 보여준다. 하지만 표 3에서 제공되는 반응값에 대한 예측오차가 그리 크지 않는 것으로 나타나 예측에는 효과적이나 추정에서는 신뢰하기 어렵다고 판단된다. 전

체적으로 자료수에 비해 변수수가 큰 고차원 자료의 경우 원 설명변수를 이용한 경우와 상위 주성분만을 취하는 회귀분석에 비해 주성분을 기반한 벌점회귀(lasso 벌점회귀)가 가장 낮은 추정오차값을 가짐으로 가장 정확한 회귀계수추정을 한다는 것을 알 수 있다.

표 3은 고정된 변수수 $p = 10$에 대하여 상대적으로 큰 자료수(n)를 가지는 경우에서의 반응변수의 예측오차값(PE)을 보여준다. 표 1에서의 추정오차(EE)에 비해 모든 방법이 비교적 동등한 예측오차를 보여주고 있음을 알 수 있다. 특히, PCR 및 PLSR의 예측오차가 PPCL 방법들에 비해 나는 경우가 많은데 이는 추정치가 정확하지 않더라도 예측력이 좋게 나타나는 다중공선성의 문제를 여전히 보여주는 예라고 할 수 있다. 이러한 현상은 변수수가 자료수보다 큰 경우인 표 4에서 현저히 나타남을 알 수 있다. 즉, 실제 차원수에 비해 변수수가 매우 큰 경우 PCR 및 PLSR은 여전히 다중공선성의 문제를 해결할 수 없다는 것을 실증하는 예라고 볼 수 있다.

모의 실험 결과를 종합적으로 해석해 볼 때, 본 연구에서 제안한 주성분을 기반으로 벌점함수의 도입은 반응변수에 영향을 미치는 주성분이 상위 주성분이 아닌 경우에 추정과 예측에서 좋은 성능을 보임을 알 수 있다. 특히 변수의 수가 자료의 수보다 매우 많을 경우 비슷한 반응변수의 예측력을 가지면서 모수의 추정에서 주성분회귀법 및 부분최소제곱추정법보다 훨씬 향상된 결과를 줌을 알 수 있다. 그림 3 ∼ 그림 13에 모의실험 결과의 비교를 쉽게 하기 위한 추정오차 및 예측오차의 비교그림을 제시하였다.

5 결론

본 연구를 통해, 주성분을 사용한 다중회귀모형과 벌점함수를 사용한 변수선택을 응용하여 새로운 회귀법을 제시하였다. 제시한 방법론의 성능을 확인하기 위해 다양한 상황에서의 컴퓨터 모의실험을 수행하였다. 현실에서 관측되는 많은 자료에서는 변수간의 강한 상관관계를 가진 다중공선성의 문제를 가지고 있거나, 자료수보다 설명변수의 수가 많은 고차원의 자료들이 빈번하게 나타난다. 본 연구에서는 이러한 문제들로부터 수반되는 추정과 예측이 나빠짐을 해결하기 위해 주성분회귀분석의 개념 및 변수선택의 개념을 회귀분석에 도입하였다. 본 연구에서는 이를 자연스럽게 결합하였으며, 모의실험을 이용하여 추정과 예측이 향상될 수 있음을 보였다.

6 부록

6.1 표

표 1: 변수수가 $p = 10$이고 자료수 n이 변하는 상황에서의 추정오차(EE).

상황	방법	100	500	1000	1500
(T1)	PCR	0.5229	0.1477	0.1612	0.1127
	PLSR	0.4910	0.1887	0.1405	0.1265
	PPCR-L	**0.2300**	**0.0534**	**0.0395**	**0.0352**
	PPCR-R	0.6592	0.6429	0.6434	0.6424
	PPCR-S	1.0620	0.2698	0.1359	0.0915
	REG-L	0.9719	0.8604	0.8512	0.8517
	REG-R	0.5732	0.5537	0.5539	0.5529
	REG-S	1.1699	1.1329	1.1356	1.0289
(T2)	PCR	1.2170	0.4992	0.3591	0.2960
	PLSR	**1.1008**	0.5839	0.3362	0.3255
	PPCR-L	1.2721	**0.5039**	**0.3110**	**0.2423**
	PPCR-R	1.2478	1.1996	1.1950	1.1967
	PPCR-S	1.7321	1.7320	1.5319	1.4742
	REG-L	1.1158	0.6420	0.5467	0.5065
	REG-R	1.1756	1.0718	1.0442	1.0342
	REG-S	1.7320	1.7321	1.7321	1.7320
(T3)	PCR	0.8623	0.4141	0.2749	0.2672
	PLSR	1.0258	0.3822	0.2942	0.2749
	PPCR-L	**0.6912**	**0.2963**	**0.2515**	**0.2412**
	PPCR-R	1.1252	1.1204	1.1200	1.1196
	PPCR-S	1.4295	1.2203	1.0599	1.0037
	REG-L	1.0139	0.7440	0.7109	0.7098
	REG-R	1.0652	1.0459	1.0437	1.0426
	REG-S	1.5411	1.4830	1.3904	1.3531

표 2: 자료수가 $n = 100$이고 변수수 p가 변하는 상황에서의 추정오차(EE).

상황	방법	100	500	1000	1500
(T1)	PCR	1891738	45937.93	45.8521	17186.95
	PLSR	562.7411	5618.413	671.7326	10006.72
	PPCR-L	**0.2356**	**0.2996**	**0.2949**	**0.2937**
	PPCR-R	0.6578	1.6543	1.6542	1.6542
	PPCR-S	1.0619	1.0612	-	-
	REG-L	4.0108	8.4819	10.2755	12.3423
	REG-R	0.3312	1.3974	1.3122	1.2719
	REG-S	4.9203	9.8435	-	-
(T2)	PCR	1915015	33638515	66705588	17206.76
	PLSR	1416.564	4208.36	2123.915	12851.12
	PPCR-L	3.1024	**0.4586**	**0.4593**	**0.4872**
	PPCR-R	1.2503	1.7218	1.7221	1.7219
	PPCR-S	1.7321	1.7321	-	-
	REG-L	4.0811	6.9743	8.4259	10.1285
	REG-R	**0.9431**	1.6747	1.6566	1.6416
	REG-S	1.7321	1.7321	-	-
(T3)	PCR	357706.4	31291.58	34435.18	1299.605
	PLSR	6615.269	9875.761	181.6539	13335.9
	PPCR-L	**0.8748**	**0.8406**	**0.8437**	**0.8377**
	PPCR-R	1.1238	1.6874	1.6873	1.6872
	PPCR-S	1.4304	1.4332	-	-
	REG-L	3.5009	6.3393	8.2272	9.8986
	REG-R	0.9008	1.5416	1.4879	1.4649
	REG-S	3.1289	5.3966	-	-

표 3: 변수수가 $p = 10$이고 자료수 n이 변하는 상황에서의 예측오차(PE).

상황	방법	100	500	1000	1500
(T1)	PCR	**0.1028**	**0.1004**	**0.1002**	**0.1001**
	PLSR	**0.1028**	**0.1004**	**0.1002**	**0.1001**
	PPCR-L	0.1257	0.1014	0.1005	0.1003
	PPCR-R	0.2392	0.2334	0.2335	0.2331
	PPCR-S	1.5120	1.5967	1.6091	1.6142
	REG-L	0.1264	0.1024	0.1014	0.1013
	REG-R	0.2082	0.2017	0.2017	0.2013
	REG-S	0.3409	0.1485	0.1487	0.1265
(T2)	PCR	0.1041	**0.1006**	**0.1003**	0.1002
	PLSR	0.1041	**0.1006**	**0.1003**	0.1002
	PPCR-L	**0.1039**	**0.1006**	**0.1003**	**0.1001**
	PPCR-R	0.1271	0.1220	0.1216	0.1216
	PPCR-S	0.1752	0.1745	0.1836	0.1908
	REG-L	0.1038	0.1007	0.1003	0.1002
	REG-R	0.1228	0.1153	0.1139	0.1133
	REG-S	0.1746	0.1746	0.1746	0.1746
(T3)	PCR	0.7445	0.7440	0.7440	0.7441
	PLSR	0.7445	0.7440	0.7440	0.7441
	PPCR-L	0.7373	0.7427	0.7428	0.7429
	PPCR-R	**0.7069**	**0.7082**	**0.7081**	**0.7083**
	PPCR-S	0.9366	1.0010	1.0150	1.0201
	REG-L	0.7354	0.7432	0.7433	0.7434
	REG-R	0.7130	0.7149	0.7149	0.7150
	REG-S	0.7093	0.7278	0.7266	0.7297

표 4: 자료수가 $n = 100$이고 변수수 p가 변하는 상황에서의 예측오차(PE).

상황	방법	100	500	1000	1500
(T1)	PCR	0.1035	**0.1033**	**0.1030**	0.1036
	PLSR	**0.1030**	**0.1033**	0.1031	**0.1034**
	PPCR-L	0.1258	0.1353	0.1341	0.1336
	PPCR-R	0.2389	1.0554	1.0551	1.0551
	PPCR-S	1.5121	1.5084	-	-
	REG-L	0.1253	0.1313	0.1299	0.1310
	REG-R	0.1427	0.7566	0.6668	0.6273
	REG-S	0.2480	0.3734	-	-
(T2)	PCR	0.1051	0.1050	0.1048	0.1052
	PLSR	0.1048	0.1050	0.1050	0.1051
	PPCR-L	0.1041	**0.1036**	**0.1037**	0.1039
	PPCR-R	0.1273	0.1736	0.1737	0.1736
	PPCR-S	0.1754	0.1752	-	-
	REG-L	**0.1039**	**0.1036**	0.1038	**0.1037**
	REG-R	0.1121	0.1684	0.1664	0.1646
	REG-S	0.1747	0.1747	-	-
(T3)	PCR	0.7447	0.7369	0.7450	0.7447
	PLSR	0.7446	0.7369	0.7449	0.7447
	PPCR-L	0.7376	0.7251	0.7327	0.7328
	PPCR-R	**0.7066**	0.7226	0.7285	0.7284
	PPCR-S	0.9367	0.9384	-	-
	REG-L	0.7374	0.7283	0.7321	0.7342
	REG-R	0.7252	**0.6839**	**0.6861**	**0.6847**
	REG-S	0.7314	0.7408	-	-

6.2 그림

그림 3: **(T1)** 에서의 p가 고정되고 n이 변하는 상황에서의 모든 EE.

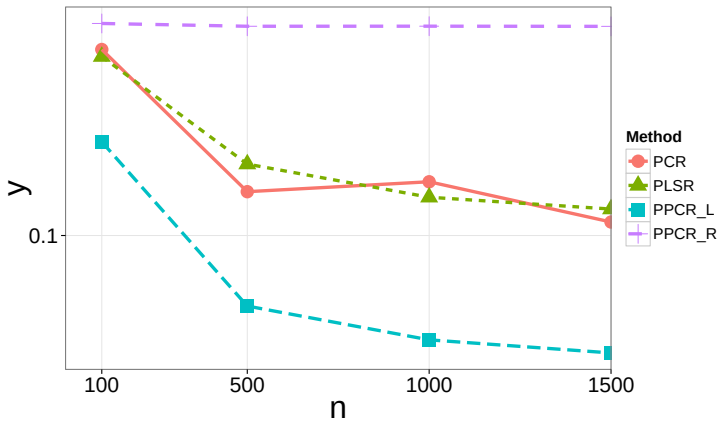


그림 4: **(T2)** 에서의 p가 고정되고 n이 변하는 상황에서의 모든 EE.

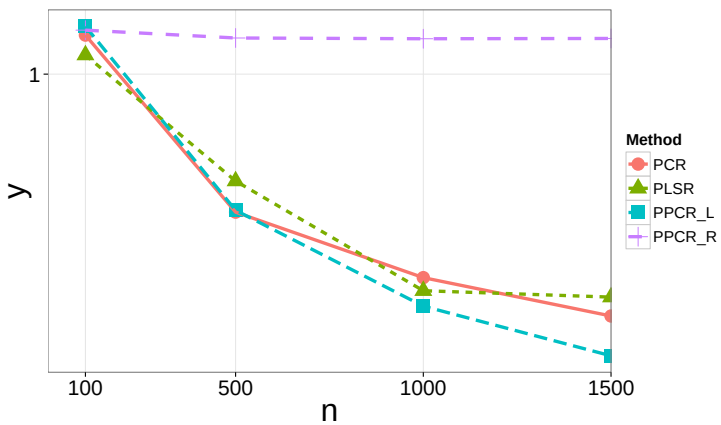


그림 5: **(T3)** 에서의 p가 고정되고 n이 변하는 상황에서의 모든 EE.

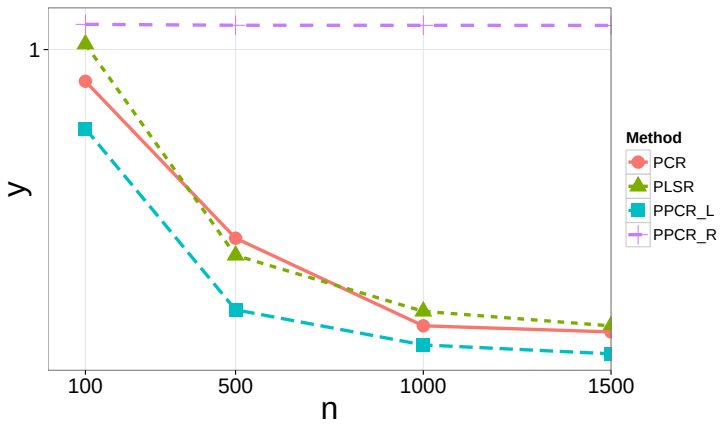


그림 6: **(T1)** 에서의 n이 고정되고 p가 변하는 상황에서의 모든 EE.

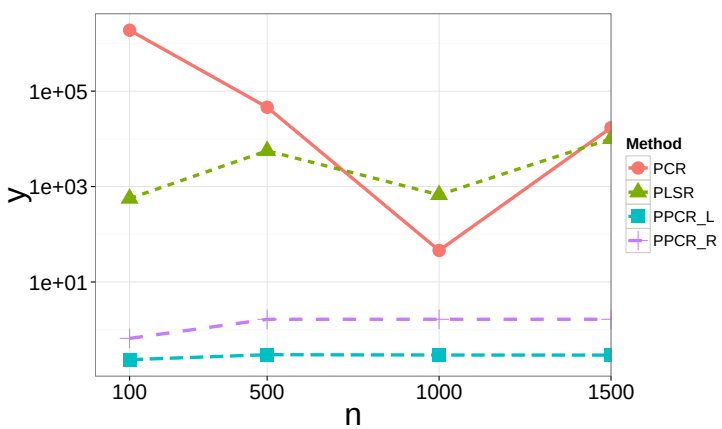


그림 7: **(T1)** 에서의 n이 고정되고 p가 변하는 상황에서의 모든 EE.

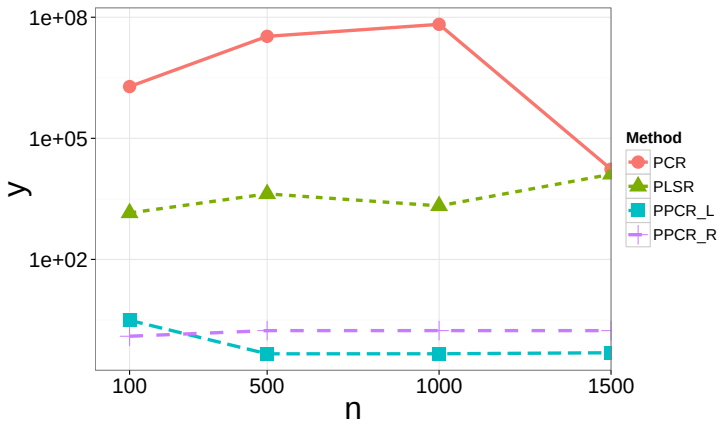


그림 8: **(T1)** 에서의 n이 고정되고 p가 변하는 상황에서의 모든 EE.

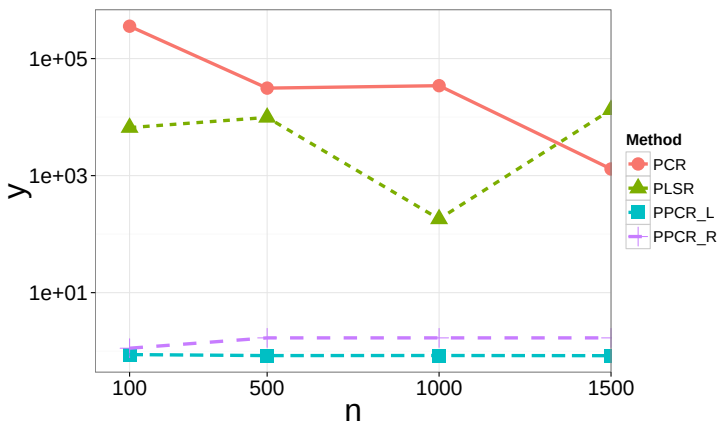


그림 9: **(T1)** 에서의 p가 고정되고 n이 변하는 상황에서의 모든 PE.

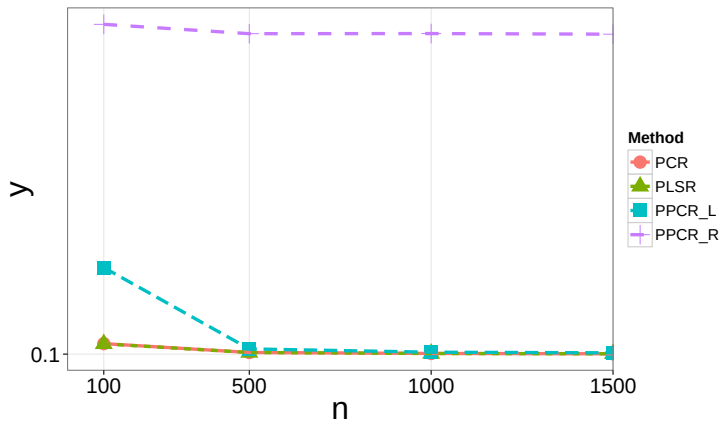


그림 10: **(T2)** 에서의 p가 고정되고 n이 변하는 상황에서의 모든 PE.

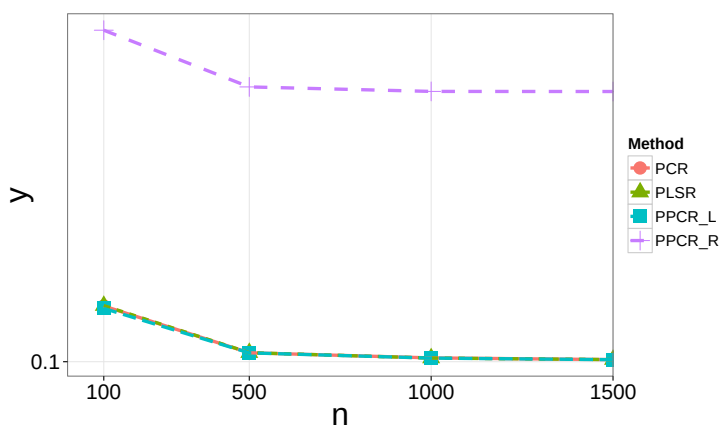


그림 11: **(T3)** 에서의 p가 고정되고 n이 변하는 상황에서의 모든 PE.

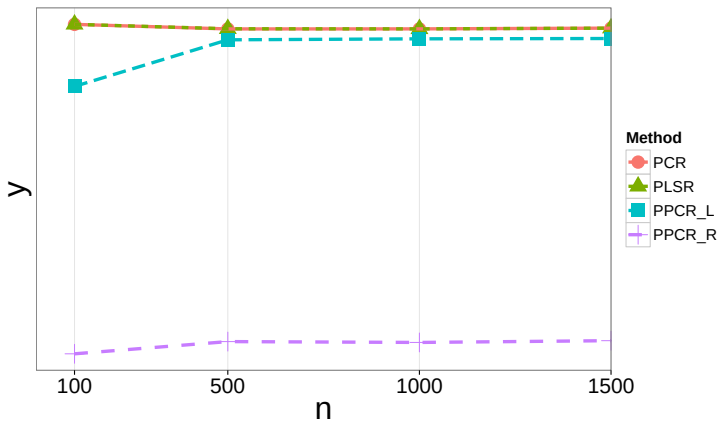


그림 12: **(T1)** 에서의 n이 고정되고 p가 변하는 상황에서의 모든 PE.

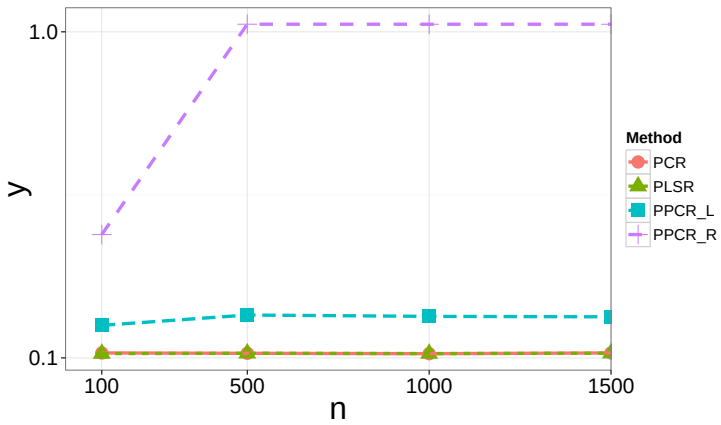
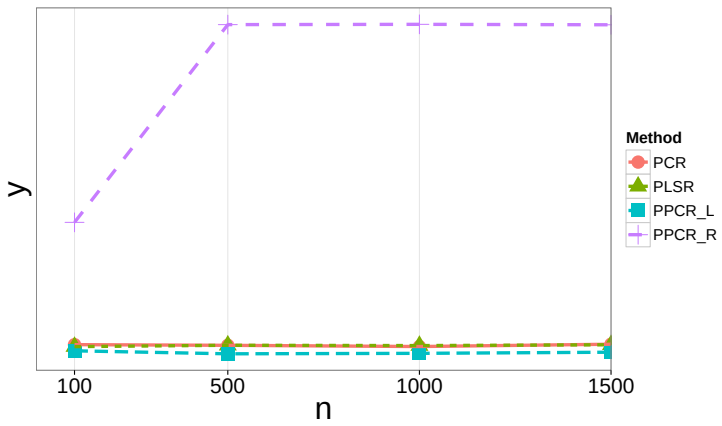


그림 13: **(T1)** 에서의 n이 고정되고 p가 변하는 상황에서의 모든 PE.



참고문헌

DAHINDEN, C., KALISCH, M. and B¨UHLMANN, P. (2010). Decomposition and model selection for large contingency tables. Biometrical Journal 52 233–252.

ANDERSON, T. (1984). An Introduction to Multivariate Statistical Analysis. 2nd ed. Wiley.

Bache, K., and Lichman, M. (2013) UCI machine learning repository [http://archive.ics.uci.edu/ml] Irvine, CA: University of California, School of Information and Computer Science.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2008) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

Hastie, T., Buja, A., and Tibshirani, R. (1995) Penalized discriminant analysis. *The annals of Statisitcs*, **23**, 73–102.

Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Hoerl, A. E. and Kennard, R. W. (1970a) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hoerl, A. E. and Kennard, R. W. (1970b) Ridge regression: Iterative estimation of the baising parameter. *Technometrics*, **12**, 77–88.

Kim, K., and Lee, S. (2013) Logistic regression classification by principal component selection *Submitted*.

Kondylis, A., and Whittaker, J. (2008) Spectral preconditioning of Krylow spaces: combining pls and pc regression. *Computational Statistics & Data Analysis*, **52**, 2588–2603.

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

On the Estimation for Sparse Principal Component Regression Approach under Multiple Regression Problem

Abstract

In this study, we suggest a penalized principal component regression by selecting principal components under sparsity-inducing penalties, including lasso penalty. Under various situations, the proposed method is tested and compared with the existing methods through simulation studies. Simulation results demonstrate that the proposed method outperforms the existing methods.