Journal of the Korean Data & Information Science Society
2013, **24**(3), 533–541

http://dx.doi.org/10.7465/jkdi.2013.24.3.533
한국데이터정보과학회지

라소를 이용한 간편한 주성분분석

박철용[1]

[1]계명대학교 통계학과

접수 2013년 4월 19일, 수정 2013년 5월 7일, 게재확정 2013년 5월 12일

요 약

이 연구에서는 라소를 이용한 간편한 주성분분석을 제안한다. 이 방법은 다음의 두 단계로 구성되어 있다. 먼저 주성분분석에 의해 주성분을 구한다. 다음으로 각 주성분을 반응변수로 하고 원자료를 설명변수로 하는 라소 회귀모형에 의한 회귀계수 추정량을 구한다. 이 회귀계수 추정량에 기반한 새로운 주성분을 사용한다. 이 방법은 라소 회귀분석의 성질에 의해 회귀계수 추정량이 보다 쉽게 0이 될 수 있기 때문에 해석이 쉬운 장점이 있다. 왜냐하면 주성분을 반응변수로 하고 원자료를 설명변수로 하는 회귀모형의 회귀계수가 고유벡터가 되기 때문이다. 라소 회귀모형을 위한 R 패키지를 이용하여 모의생성된 자료와 실제 자료에 이 방법을 적용하여 유용성을 보였다.

주요용어: 라소, 주성분분석, 회귀모형.

1. 머리말

Tibshirani (1996)에 의해 회귀계수 (regression coefficient)에 L^1 벌점 (penalty)을 부과하는 Lasso (least absolute shrinkage and selection operator) 기법이 소개되었다. Tibshirani (1996)에 의하면 Lasso 회귀 (Lasso regression)는 최소제곱법 (least squares method)에 의한 회귀나 능형 회귀 (ridge regression)와는 달리 회귀계수 추정값이 0인 모형이 보다 쉽게 선택되기 때문에 변수선택 (variable selection)의 효과가 있다고 하였다. 따라서 Lasso 회귀는 능형 회귀의 장점인 예측정확도 (prediction accuracy)와 변수선택의 장점인 해석력 (interpretation)을 어느 정도 겸비할 수 있는 장점이 있다고 알려져 있다.

Tibrashini (1996)에 의해 회귀모형에 대한 Lasso 방법이 소개된 이후 여러 분야에서 Lasso 방법이 사용되고 있다. 예를 들어 일반화선형모형 (generalized linear model; Friedman 등, 2008; Park과 Kye, 2013)과 다변량분석 (multivariate analysis)의 일종인 선형판별분석 (linear discriminant analysis)에 대한 연구 (Whitten과 Tibshirani, 2011) 등이 있었다.

이 연구는 주성분분석 (principal component analysis)에 의한 주성분 (principal component)의 계수에 L^1 벌점을 가하면 해석력과 예측정확도가 동시에 만족되는 새로운 주성분 결과가 나오지 않을까 하는 호기심에서 출발하였다. 실제로 반응변수 (response variable)를 주성분으로 하고 설명변수 (explanatory variable)를 원자료로 했을 때 회귀계수 추정량은 해당되는 고유벡터 (eigenvector)가 되는 것을 쉽게 보일 수 있다. 따라서 반응변수를 주성분으로 하고 설명변수를 원자료를 하는 Lasso 회귀를 통해 회귀계수 추정량을 구하면 Lasso 회귀 추정량의 장점을 누릴 수 있음을 알 수 있다.

이 연구의 장점으로는 다음과 같은 것을 생각해 볼 수 있을 것이다. 먼저 Lasso 회귀를 실행하는 기존의 프로그램만 있으면 간단히 실행할 수 있다는 것이다. 다음으로 반응변수를 주성분으로 하고 설명

[1] (704-701) 대구광역시 달서구 달구벌대로 1095, 계명대학교 통계학과, 교수. E-mail: cypark1@kmu.ac.kr

변수를 원자료로 했을 때 회귀계수 추정량이 해당되는 고유벡터가 되는 것을 쉽게 보일 수 있기 때문에, 이 논문으로서 이론적인 내용이 충분히 갖춰진 간편한 방법이라는 점이다. 이 간편한 방법을 실제 자료 (real data)와 모의생성된 자료 (simulated data)에 적용하고 유용성을 살펴보도록 하고자 한다.

이 논문은 다음과 같이 구성되어 있다. 2절에서는 Lasso 로지스틱 회귀와 주성분분석에 대해 간략히 설명하고 제안된 방법인 Lasso를 이용한 간편한 주성분분석에 대해 설명한다. 3절에서는 제안된 방법을 모의생성된 자료와 실제 자료에 적용하고 유용성을 살펴본다. 4절의 결론 및 논의에서는 이 연구의 결과들을 요약하고 추후 연구과제를 정리한다.

2. Lasso를 이용한 주성분분석

2.1. Lasso 회귀

통상적으로 회귀분석 (regression analysis)에서 회귀계수 (regression coefficient)의 추정량을 구하기 위해서 잔차 (residual)의 제곱합을 최소로 하는 최소제곱법 (least squares method)을 사용한다. 그러나 설명변수 (explanatory variables)의 개수가 증가하면 설명변수들 사이의 강한 상관관계로 인한 다중공선성 (multicollinearity)이 존재할 수 있기 때문에 최소제곱 회귀계수 추정량의 분산이 커져 추정회귀식의 예측정확도가 떨어지는 문제점이 발생할 수 있다. 또한 설명변수의 개수가 증가하면 변수에 대한 해석력이 떨어진다. 다시 말해, 많은 설명변수 중 어떤 변수가 중요한 역할을 하는지에 대한 판단이 어려워진다.

Lasso 회귀 (Lasso regression)는 능형 회귀 (ridge regression)의 장점인 회귀계수 축소를 통해 예측정확도 (prediction accuracy)를 높이고, 동시에 영향력이 적은 회귀계수 값을 쉽게 0으로 만드는 변수선택 (variable selection)의 기능이 있어 해석력 (interpretability)을 높여준다. 따라서 Lasso 회귀는 능형회귀의 예측정확도와 변수선택의 해석력을 모두 갖출 수 있는 분석 방법으로 알려져 있다.

Lasso 회귀의 추정량은 다음의 식 (2.1)과 같이 구할 수 있다.

$$\text{argmin}_{\beta_1,\cdots,\beta_p}\left\{\sum_{i=1}^{n}(Y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\}. \tag{2.1}$$

여기서 회귀계수 $\beta_1,\cdots,\beta_p$의 값은 설명변수 x_{ij}의 척도 (scale)에 의존하기 때문에 회귀계수 값의 크기가 그 변수의 영향력을 반영하지 못하기 때문에 x_{ij}에 표준화된 값을 사용한다. 또한 β_0의 추정량은 항상 $\bar{Y}$가 되기 때문에 β_0에 대한 추정은 관심 대상에서 제외되고 $\beta_1,\cdots,\beta_p$에 대한 최소화 문제로 귀착된다.

식 (2.1)을 다음과 같은 제약조건이 주어진 최소화 문제로 표시할 수 있다 (Tibshirani, 1996).

$$\text{argmin}_{\beta_1,\cdots,\beta_p}\left\{\sum_{i=1}^{n}(Y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2\right\}\ \text{subject to}\ \sum_{j=1}^{p}|\beta_j|\le t. \tag{2.2}$$

위의 식 (2.2)의 제약조건인 $t\ (\ge 0)$는 회귀계수 값에 대하여 축소 정도를 조절하는 조절모수 (tuning parameter)이다. 이 조절모수 t값이 줄어들면 중요하지 않은 변수의 회귀계수 값은 축소되면서 순서대로 0으로 만들어져 변수선택이 되는 효과가 생긴다. 조절 모수 t값이 충분히 커지면 회귀계수 값에 대한 제약이 없어지므로 최소제곱 부분만 남아 Lasso 회귀추정량이 최소제곱 추정량이 된다.

2.2. 주성분분석

주성분분석 (principal component analysis)은 (평균이 0인) 연속형 확률변수 $X_1,\cdots,X_p$의 분산을 최대한 설명하는 선형결합 (linear combination)을 사용하여 차원축소 (dimension reduction)하는 통

계적 방법이다. 이 문제는 $X_1, \cdots, X_p$의 표본공분산행렬을 S라 했을 때 이 행렬의 고유값-고유벡터 짝 (eigenvalue-eigenvector pairs) $(\lambda_i, \underline{e}_i)$, $i = 1, \ldots, p$에 의해 수학적 해법이 존재한다 (여기서 $\lambda_1 \geq \cdots \geq \lambda_p$로 정렬되어 있다고 가정한다). 구체적으로 다음의 결과가 성립한다 (Johnson과 Wichern, 1992).

$$\max_{\underline{l} \neq 0} \frac{\underline{l}^T S \underline{l}}{\underline{l}^T \underline{l}} = \lambda_1 \text{ 이며 } \text{argmax}_{\underline{l} \neq 0} \frac{\underline{l}^T S \underline{l}}{\underline{l}^T \underline{l}} = \underline{e}_1.$$

$\underline{e}_1, \ldots, \underline{e}_k$ $(k = 1, \ldots, p-1)$에 직교인 $\underline{l}$에 대해

$$\max_{\underline{l} \neq 0} \frac{\underline{l}^T S \underline{l}}{\underline{l}^T \underline{l}} = \lambda_{k+1} \text{이며 } \text{argmax}_{\underline{l} \neq 0} \frac{\underline{l}^T S \underline{l}}{\underline{l}^T \underline{l}} = \underline{e}_{k+1}.$$

그런데 $\underline{l} \neq 0$에 대해 $\underline{u} = \underline{l}/\sqrt{\underline{l}^T \underline{l}}$이라 놓으면 길이 $|\underline{u}|_2 \equiv \sqrt{\underline{u}^T \underline{u}}$가 1이 되며 또한 $\underline{l}^T S \underline{l}/\underline{l}^T \underline{l} = \underline{u} S \underline{u}$를 만족한다. 따라서 첫 번째 주성분 (first principal component) $Y_1 = \underline{e}_1^T \underline{X}$는 $\underline{X}$의 (길이가 1인) 선형결합 중 최대의 표본분산인 λ_1을 설명한다. 또한 $k+1$번째 주성분 ($(k+1)$-th principal component) $Y_{k+1} = \underline{e}_{K+1}^T \underline{X}$는 앞에서 구한 $\underline{e}_1, \cdots, \underline{e}_k$에 직교인 $\underline{X}$의 (길이가 1인) 선형결합 중 최대의 표본분산인 λ_{k+1}을 설명한다.

2.3. Lasso를 이용한 간편한 주성분분석

이 연구에서 제안하는 Lasso를 이용한 간편한 주성분분석을 요약하면 다음과 같다. 우선 주성분분석에 의해 주성분 벡터 $\underline{y}_1, \ldots, \underline{y}_p$를 구한다. 다음으로 이 주성분벡터를 반응변수로 하고 (표준화된) $X_1, \cdots, X_p$를 설명변수로 하는 Lasso 회귀에 의해 얻어지는 회귀계수 추정량을 얻는다. 이 회귀계수 추정량을 길이가 1이 되도록 치환한 후 고유벡터 대신에 사용하여 새로운 주성분을 만든다.

상기 절차를 사용할 수 있는 근거가 되는 핵심적인 정리는 다음과 같다.

정리 2.1 표본공분산행렬 S의 고유값-고유벡터 짝을 $(\lambda_i, \underline{e}_i)$, $i = 1, \ldots, p$라고 하자. 이 때 $n \times p$인 자료행렬 X에 대해 $\underline{y}_i = X\underline{e}_i$를 i-번째 주성분 벡터라고 하면 다음이 성립한다.

$$\text{argmin}_{\underline{\beta}} \left\{ |\underline{y}_i - X\underline{\beta}|_2^2 \right\} = \underline{e}_i.$$

증명: 다중회귀모형 (multiple regression model) $\underline{y}_i = X\underline{\beta} + \underline{\epsilon}$의 회귀계수 $\underline{\beta}$의 최소제곱 추정량 (least squares estimator)은 다음과 같다.

$$\hat{\underline{\beta}}_i = (X^T X)^{-1} X^T \underline{y}_i = (X^T X)^{-1} X^T X \underline{e}_i = \underline{e}_i.$$

이것으로 간단히 정리가 증명된다. □

따라서 이 연구에서 사용하고자 하는 Lasso를 이용한 주성분분석은 다음과 같이 구할 수 있다.

$$\tilde{\underline{\beta}}_i^* = \text{argmin}_{\underline{\beta}} \left\{ |\underline{y}_i - X\underline{\beta}|_2^2 + \lambda_1^* |\underline{\beta}|_1 \right\}. \tag{2.3}$$

여기서 $|\underline{\beta}|_1 \equiv \sum_{i=1}^p |\beta_i|$와 $|\underline{\beta}|_2 \equiv \sqrt{\sum_{i=1}^p \beta_i^2}$ 각각 β의 L^1과 L^2 노름 (norm)이다. 그러나 상기 추정량은 길이가 1인 조건을 만족하지 못하기 때문에 길이가 1이 되도록 조정한 $\tilde{\underline{\beta}}_i = \tilde{\underline{\beta}}_i^* / |\tilde{\underline{\beta}}^*|_2$를 사용하게 된다. 따라서 Lasso 회귀의 성질에 의해 추정량 $\tilde{\underline{\beta}}$는 원래의 고유벡터 $\hat{\underline{\beta}}_i = \underline{e}_i$보다 0의 값을 많이 가질 것이기 때문에 주성분의 해석이 용이하게 된다.

3. 적용 예제

이 절에서는 하나의 실제 자료 (real data)와 하나의 모의생성된 자료 (simulated data)에 앞에서 제안한 Lasso를 이용한 간편한 주성분분석을 적용한다. 자료분석에는 R을 사용하였으며 Lasso 회귀를 위해 R 패키지인 lars나 elasticnet을 사용할 수 있으나 보다 일반적인 elasticnet을 이용하였다.

3.1. 10종 경기 자료

실제 적용 예로 사용된 자료는 R 패키지인 FactoMinerR에 있는 decathlon 자료이다. 이 자료는 2004년 올림픽과 2004 Decastar의 41명 선수의 기록이다. 원래 13개 변수가 있으나 경기 기록에 해당되는 다음의 자료가 사용되었다.

Table 3.1 Variables used in decathlon data

Variable name	Explanation	Unit
100m	100 meters	seconds
Long.jump	Long jump	meters
Shot.put	Shot put	meters
High.jump	High jump	meters
400m	400 meters	seconds
110m.hurdle	110 meters hurdles	seconds
Discus	Discus throw	meters
Pole.vault	Pole vault	meters
Javeline	Javelin throw	meters
1500m	1500 meters	seconds

10개 경기의 기록이기 때문에 원변수 대신에 표준화된 변수를 사용하는 것이 적절하다고 판단되어 표준화된 변수를 사용하였다. 그리고 달리기 경기인 100m, 400m, 110m.hurdle, 1500m는 다른 경기와 달리 값이 작은 것이 기록이 좋은 것이라 표준화 과정에 -1을 곱해 모든 변수들의 값이 크면 기록이 좋은 것으로 변환하였다.

먼저 고유값 (eigenvalue)을 구하였더니 0.327, 0.174, 0.140, 0.106, 0.068, 0.060, 0.045, 0.040, 0.021, 0.018이 나왔다. 평균 분산인 1보다 큰 분산을 설명하는 주성분이 4개이기 때문에 이 기준에 의해 4개의 주성분만 관심을 가지고 살펴보도록 한다.

네 개의 큰 고유값인 0.327, 0.174, 0.140, 0.106에 해당되는 고유벡터 (eigenvector)는 Table 3.2와 같다.

Table 3.2 Eigenvectors corresponding to four largest eigenvalues

Variable name	PC1	PC2	PC3	PC4
100m	-0.4283	-0.1420	0.1556	-0.0368
Long.jump	-0.4102	-0.2621	0.1537	-0.0990
Shot.put	-0.3441	0.4539	-0.0197	-0.1854
High.jump	-0.3162	0.2658	-0.2189	0.1319
400m	-0.3757	-0.4320	-0.1109	0.0285
110m.hurdle	-0.4126	-0.1736	0.0782	0.2829
Discus	-0.3054	0.4600	0.0362	0.2526
Pole.vault	-0.0278	-0.1368	0.5836	-0.5365
Javeline	-0.1532	0.2405	-0.3287	-0.6929
1500m	-0.0321	-0.3598	-0.6599	-0.1567

다른 통계패키지와는 달리 R에서는 첫 고유벡터의 값이 모두 음수로 나타나 있어 특이하다. elastic-net 패키지에서 Lasso 회귀를 실행하는 함수인 enet에는 LARS-EN 알고리즘 (Zhou와 Hastie, 2005)을 사용하기 때문에 10단계에 걸쳐 하나씩 0이 되는 변수가 추가되는 과정을 부가적으로 쉽게 구할 수 있다. 이 장점을 이용하면 각 주성분별로 계수가 0인 변수의 개수인 희박성 (sparsity)에 따른 누적 설명분산 비율 (CPEV; cumulative proportion of explained variance)을 시각적으로 살펴보면서 적절한 희박성을 선택할 수 있다. 이 논문에서는 Shen과 Huang (2008)에 의해 제안된 CPEV를 사용하였다. 네 개의 주성분을 각각 반응변수로 하고 10개의 표준화된 경기 성적을 설명변수로 하는 Lasso 회귀의 희박성과 누적 설명분산 비율을 그림으로 나타낸 것이 Figure 3.1이다.

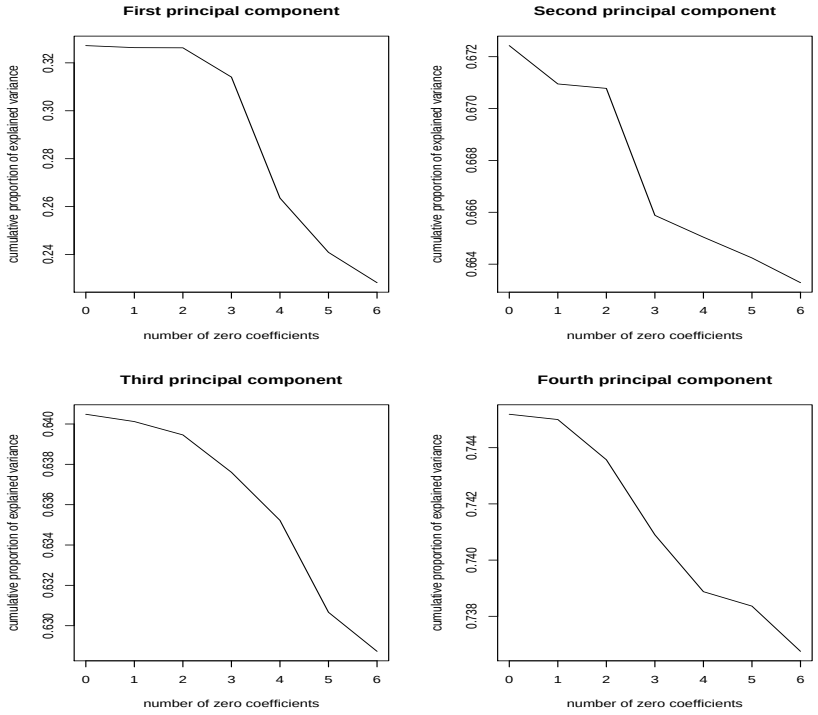

Figure 3.1 CPEV versus sparsity for the four principal components

위의 그림에서 보면 첫 번째 주성분에서는 2개 변수까지 0을 놓더라도 CPEV가 거의 변동이 없기 때문에 희박성을 2로 선택하였고, 두 번째 주성분에서는 희박성을 1로 하더라도 CPEV가 다소 떨어지는 관계로 희박성을 0으로 선택하였다. 세 번째와 네 번째 주성분에서는 희박성을 각각 2와 1로 선택하였다.

이렇게 희박성을 2, 0, 2, 1로 선택하였을 때 해당되는 계수벡터 $\underline{\tilde{\beta}}_i, i = 1, 2, 3, 4$는 다음 Table 3.3과 같다.

Table 3.3을 보면 각 계수벡터의 0인 변수의 숫자가 정확히 2, 0, 2, 1로 나와 있음을 알 수 있다. 세 번째와 네 번째의 계수벡터에서 절대값이 0.1도 되지 않는 것이 추가로 2개 변수씩 발견되고 있으나 누적 설명 분산비율이 떨어져 추가적인 희박성을 확보하기는 힘든 상태이다. 이렇게 선택된 계수벡터에 의한 CPEV는 0.3263, 0.5000, 0.6395, 0.7450로 나타나 원래 주성분의 CPEV인 0.3272, 0.5009, 0.6414, 0.7471에 비해 아주 미약하게 값이 작아져 있는 것을 알 수 있다.

Table 3.3 Coefficients vectors for the new principal components with sparsity 2, 0, 2, 1

Variable name	$\tilde{\underline{\beta}}_1$	$\tilde{\underline{\beta}}_2$	$\tilde{\underline{\beta}}_3$	$\tilde{\underline{\beta}}_4$
100m	-0.4190	-0.1420	0.1298	-0.0240
Long.jump	-0.4170	-0.2621	0.1320	-0.0803
Shot.put	-0.3593	0.4539	0.0000	-0.1699
High.jump	-0.3091	0.2658	-0.2051	0.1203
400m	-0.3926	-0.4320	-0.0510	0.0000
110m.hurdle	-0.4127	-0.1736	0.0653	0.2789
Discus	-0.2819	0.4600	0.0000	0.2498
Pole.vault	0.0000	-0.1368	0.5723	-0.5389
Javeline	-0.1485	0.2405	-0.3141	-0.7072
1500m	0.0000	-0.3598	-0.7005	-0.1374

3.2. 모의생성된 자료

이 예제는 Zou 등 (2006)에 사용된 모의생성된 자료 형태를 사용하였다. 구체적인 변수생성 형태는 다음과 같다.

$$\begin{aligned} X_i &= V_1 + \epsilon_i,\ i = 1, 2, 3, 4 \\ X_i &= V_2 + \epsilon_i,\ i = 5, 6, 7, 8 \\ X_i &= V_3 + \epsilon_i,\ i = 9, 10 \end{aligned}$$

여기서

$$V_1 \sim N(0, 290),\ V_2 \sim N(0, 300),\ V_3 = -0.3V_1 + 0.925V_2 + \epsilon,$$

이며 $\epsilon, \epsilon_i\ (i = 1, 2, \ldots, 10)$은 표준정규분포에서의 확률표본이다.

표본크기를 5000으로 하여 모의생성 자료를 만들고 원변수에 대한 주성분분석을 실시하였다. 그 결과 고유값은 1804.49, 1187.74, 2.34, 1.06, 1.05, 1.03, 0.99, 0.99, 0.97, 0.94가 나와 2개의 주성분이면 충분하다는 명확한 결과가 나왔다. 처음 두 개의 고유값에 대응되는 고유벡터는 Table 3.4와 같다.

Table 3.4 Eigenvectors corresponding to two largest eigenvalues

Variable name	PC1	PC2
X_1	0.1049	-0.4808
X_2	0.1045	-0.4810
X_3	0.1044	-0.4808
X_4	0.1050	-0.4816
X_5	-0.3987	-0.1357
X_6	-0.3989	-0.1356
X_7	-0.3986	-0.1358
X_8	-0.3984	-0.1358
X_9	-0.4002	0.0188
X_{10}	-0.4003	0.0187

실제 자료분석에서와 마찬가지로 두 주성분에 대해서 적절한 희박성을 탐색하는 그림을 그렸더니 Figure 3.2와 같았다.

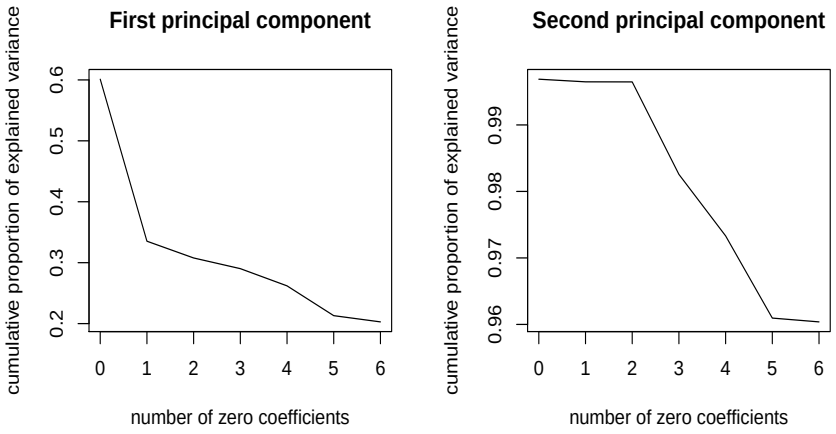

Figure 3.2 CPEV versus sparsity for the two principal components

따라서 두 주성분에 대한 희박성은 0과 2가 선택되었다. 첫 번째 주성분에 대해서 하나의 변수의 계수만 0이 되어도 CPEV (누적 설명분산 비율)가 너무 떨어져 희박성이 0으로 선택되었다. 이런 현상은 하나의 변수에 대한 계수만 0으로 만들어도 처음 네 개의 변수 X_i $(i = 1, 2, 3, 4)$에 대한 계수가 동시에 0에 가깝게 되고 또한 그 아래 네 개의 변수 X_i $(i = 5, 6, 7, 8)$에 대한 계수도 -0.1 정도가 되어 나머지 두 개의 변수 X_9, X_{10}으로만 설명되기 때문에 발생하였다. 두 번째 주성분에 대해서는 희박성 2까지 CPEV가 거의 변동이 없어 희박성 2가 선택되었다. 이렇게 선택된 희박성에 대응되는 계수벡터 $\underline{\tilde{\beta}}_i$, $i = 1, 2$는 Table 3.5와 같다.

Table 3.5 Coefficients vectors for the new principal components with sparsity 0, 2

Variable name	$\underline{\tilde{\beta}}_1$	$\underline{\tilde{\beta}}_2$
X_1	0.1049	-0.4824
X_2	0.1045	-0.4842
X_3	0.1044	-0.4832
X_4	0.1050	-0.4846
X_5	-0.3987	-0.1271
X_6	-0.3989	-0.1268
X_7	-0.3986	-0.1271
X_8	-0.3984	-0.1270
X_9	-0.4002	0.0000
X_{10}	-0.4003	0.0000

Table 3.5의 계수벡터에 대한 CPEV는 0.6012, 0.9965가 나와 원래의 두 개의 주성분에 대한 CPEV인 0.6012, 0.9969와 거의 차이가 없었다.

4. 결론 및 추후 연구과제

이 연구에서는 Lasso를 이용한 간편한 주성분분석 (principal component analysis)을 제안하였다. 이 방법은 두 단계로 구성되어 있다. 먼저 주성분분석에 의해 주성분 (principal component)을 구한

다. 다음으로 각 주성분을 반응변수 (response variable)로 하고 원자료를 설명변수 (explanatory variable)로 하는 Lasso 회귀에 의한 회귀계수 추정량을 구한다. 이 회귀계수 추정량에 기반한 새로운 주성분을 새로운 주성분으로 사용하였다. 이렇게 할 수 있는 근거가 되는 사실은 주성분을 반응변수로 하고 원자료를 설명변수로 했을 때 회귀계수 추정량이 해당되는 고유벡터 (eigenvector)가 되기 때문이다.

R 패키지인 elastic net에 있는 enet 함수를 이용하여 10종 경기 자료와 모의생성 자료에 대해 적용하여 보았다. enet 함수에서 제공하는 0인 회귀계수의 숫자인 희박성 (sparsity)에 대응되는 회귀계수를 이용하여 Shen과 Huang (2008)의 CPEV (누적 설명분산 비율)을 계산하여 각 주성분별로 적절한 희박성을 선택하였다. 그 결과 원래 주성분분석에 대응되는 CPEV와 거의 차이가 없으면서도 약간의 희박성이 확보되는 것을 확인할 수 있었다.

이 연구의 추후 연구로서 다음을 생각해 볼 수 있을 것이다. 식 (2.3)에서 주어진 추정량 대신에

$$\underline{\tilde{\beta}}_i^* = \text{argmin}_{\underline{\beta}} \left\{ |\underline{y}_i - X\underline{\beta}|_2^2 + \lambda_1^* |\underline{\beta}|_1 + \lambda_2^* |\underline{\beta}|_2^2 \right\}$$

방식을 사용하는 것이다. 식 (2.3)에 $\lambda_2^* |\underline{\beta}|_2^2$가 추가된 것으로 elastic net 패키지의 enet 함수를 이용하면 추정량을 구할 수 있다. $\lambda_2^* > 0$를 사용할 때의 장점으로는 $p > n$인 경우에도 적용가능하며 또한 상관이 높은 변수가 동시에 선택되는 효과가 있는 것으로 알려져 있다 (Zou 등, 2006). 따라서 이러한 추정 방법을 통해 Lasso를 이용한 간편한 주성분분석이 적용될 수 있는 범위를 확장할 수 있으리라 생각된다.

References

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

Johnson, R. A. and Wichern, D. W. (1992). *Applied multivariate statistical analysis*, 3rd Ed., Prentice Hall, New Jersey.

Park, C. and Kye, M. J. (2013). Penalized logistic regression models for determining the discharge of dyspnea patients. *Journal of the Korean Data & Information Science Society*, **24**, 125-133.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, **99**, 1015-1034.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **21**, 279-289.

Whitten, D. A. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society B*, **73**, 753-772.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, **67**, 301-320.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**, 265-286.

Journal of the Korean Data & Information Science Society
2013, **24**(3), 533–541

http://dx.doi.org/10.7465/jkdi.2013.24.3.533
한국데이터정보과학회지

Simple principal component analysis using Lasso

Cheolyong Park[1]

[1]Department of Statistics, Keimyung University

Received 19 April 2013, revised 7 May 2013, accepted 12 May 2013

Abstract

In this study, a simple principal component analysis using Lasso is proposed. This method consists of two steps. The first step is to compute principal components by the principal component analysis. The second step is to regress each principal component on the original data matrix by Lasso regression method. Each of new principal components is computed as the linear combination of original data matrix using the scaled estimated Lasso regression coefficient as the coefficients of the combination. This method leads to easily interpretable principal components with more 0 coefficients by the properties of Lasso regression models. This is because the estimator of the regression of each principal component on the original data matrix is the corresponding eigenvector. This method is applied to real and simulated data sets with the help of an R package for Lasso regression and its usefulness is demonstrated.

Keywords: Lasso, principal component analysis, regression model.

[1] Professor, Department of Statistics, Keimyung University, Daegu 704-701, Korea.
E-mail: cypark1@kmu.ac.kr