Journal of the Korean Data & Information Science Society
2017, **28**(4), 721–731

http://dx.doi.org/10.7465/jkdi.2017.28.4.721
한국데이터정보과학회지

페널티 방법을 이용한 주성분분석 연구

박철용[1]

[1]계명대학교 통계학전공

접수 2017년 6월 5일, 수정 2017년 7월 8일, 게재확정 2017년 7월 11일

요 약

이 연구에서는 Lasso 페널티 방법을 이용한 주성분분석 방법을 소개한다. 주성분분석에 Lasso 페널티를 적용하는 방법으로 흔히 사용되는 방법은 크게 두 가지가 있다. 첫 번째 방법은 주성분을 반응변수로 놓고 원 자료행렬을 설명변수로 하는 회귀분석의 회귀계수를 이용하여 최적의 선형결합 벡터를 구할 때 Lasso 페널티 (일반적으로 elastic net 페널티)를 부과하는 방법이다. 두 번째 방법은 원 자료행렬을 비정칙값 분해로 근사하고 남은 잔차행렬에 Lasso 페널티를 부과하여 최적의 선형결합 벡터를 구하는 방법이다. 이 연구에서는 주성분 분석에 Lasso 페널티를 부과하는 이 두 가지 방법들을 자세하게 개관하는데, 이 방법들은 변수 숫자가 표본크기보다 큰 경우에도 적용가능한 장점이 있다. 또한 실제 자료분석에서 R 프로그램을 통해 두 방법을 적용하고 그 결과를 비교한다. 구체적으로 변수 숫자가 표본크기보다 큰 Ahamad (1967)의 crime 자료에 적용한다.

주요용어: 라소, 비정칙값 분해, 주성분분석, 페널티, elastic net.

1. 머리말

Lasso (least absolute shrinkage and selection operator) 기법은 Tibshirani (1996)에 의해 회귀모형에서 L^1 벌점 (penalty)을 부과하여 회귀계수를 구하는 방법으로 제안되었다. Tibshirani (1996)에 의하면 Lasso 회귀 (Lasso regression)는 최소제곱 회귀 (least squares regression)와는 달리 회귀계수 추정값이 보다 쉽게 0이 되는 희박성 (sparsity)이 생기기 때문에 변수선택 (variable selection)의 효과가 있으며, 동시에 능형 회귀 (ridge regression)의 장점인 예측정확도 (prediction accuracy)도 겸비하고 있는 것으로 알려져 있다. 회귀모형에 Lasso 방법이 소개된 이후 일반화선형모형 (generalized linear model; Friedman 등, 2008; Park과 Kye, 2013)과 선형판별분석 (linear discriminant analysis; Witten과 Tibshirani, 2011) 등의 다양한 분야에 적용되었다. Lasso 방법에 대한 자세한 개관은 Kwon 등 (2013)을 참조하기 바란다.

Lasso 방법은 주성분분석 (principal component analysis)에도 적용가능하다. 다시 말해 원 변수의 선형결합으로 주성분 (principal component)을 구할 때, 선형결합 계수에 L^1 페널티를 부과하면 변수선택의 해석력과 예측정확도가 동시에 만족되는 새로운 주성분을 구할 수 있다. 이것을 직접 구현한 알고리즘이 ScoTLass 방법 (Jolliffe 등, 2003)이다 (자세한 것은 2.1에 설명되어 있다). 이 ScoTLass 방법은 R을 이용한 프로그램이 존재하지 않으며, 또한 아래에서 설명되는 두 번째 방법으로 효율적으로 ScoTLass 알고리즘이 구현될 수 있는 것으로 알려져 있어 (Witten 등, 2009) 이 연구에서는 더 이상 추구하지 않는다.

[1] (42601) 대구광역시 달서구 달구벌대로 1095, 계명대학교 통계학전공, 교수. E-mail: cypark1@kmu.ac.kr

주성분분석에서 Lasso 페널티를 부과하는 첫 번째 방법으로 고려되고 있는 것은 주성분에 elastic net (Zou와 Hastie, 2005) 회귀를 적용하는 방법이다. 실제로 반응변수를 주성분으로 하고 설명변수를 원 자료행렬로 지정하면 회귀계수 추정량이 바로 해당 고유벡터 (eigenvector)가 된다 (Park, 2013). 따라서 Lasso 회귀 관점에서 보다 0이 많아지는 희박성을 가진 고유벡터를 구할 수 있다. 그러나 이 방법은 변수의 수가 개체수보다 큰 경우에는 적용되지 못하는 단점이 있다. 그런데 L^2 페널티를 부과한 상태에서 반응변수를 주성분으로 하고 설명변수를 원 자료행렬로 한 회귀계수를 구하면 해당 고유벡터에 비례하게 된다 (Zou 등, 2006). 따라서 Lasso에 해당되는 L^1 페널티에 능형에 해당되는 L^2 페널티를 추가하는 elastic net을 통해 변수의 수가 개체수보다 큰 경우에도 적용가능한 희박성을 가진 고유벡터를 구할 수 있다.

주성분분석에서 Lasso 페널티를 부과하는 두 번째 방법으로 고려되고 있는 것은 비정칙값 분해 (singular value decomposition)에 의해 원 자료 행렬을 근사시킬 때 Lasso 벌점을 부과하는 페널티행렬분해 (penalized matrix decomposition) 방법이다 (Witten 등, 2009). 구체적으로 차수가 $n \times p$인 원 자료 행렬 X를 차수가 1인 비정칙값 분해 ($\underline{u}\, d\, \underline{v}^t$; $\underline{u}, \underline{v}$는 각각 n-벡터, p-벡터이며 d는 실수값)로 근사시킬 때 고유벡터에 해당되는 $\underline{v}$에 Lasso 페널티를 부과하며, 이 후에는 이렇게 구한 잔차행렬 (residual matrix; $X - \sum_{i=1}^{k} \underline{u}_i\, d_i\, \underline{v}_i^t$) ($k = 1, 2, \ldots, p-1$)에 대해 동일한 방식으로 희박성이 있는 고유벡터 $\underline{v}_k$를 구하는 방법이다. 이 방법은 주성분분석에서 $p > n$인 경우에도 적용가능하며, 또한 정준상관분석 (canonical correlation analysis)에도 적용가능한 장점이 있다 (Witten 등, 2009).

이 연구에서는 앞에서 간략히 설명한 주성분분석에 L^1 페널티를 부과하는 두 가지 방법을 자세히 살펴보고, R 프로그램을 통해 $p > n$ 인 실제 자료에 적용하는 예제를 보여주고자 한다. 이를 위해 이 논문은 다음과 같이 구성하고자 한다. 2절에서는 먼저 주성분분석과 ScoTLass 방법을 간략히 설명하고, elastic net 회귀와 페널티행렬분해에 의해 L^1 페널티를 부과하는 주성분분석 방법을 자세히 설명한다. 3절에서는 R 프로그램을 통해 이 두 가지 방법을 $p > n$인 실제 자료인 crime 자료 (Ahamad, 1967)에 적용하고 비교한다. 4절의 결론에서는 이 연구의 결과들을 정리한다.

2. Lasso 페널티에 기반한 주성분분석 방법

2.1. 주성분분석과 ScoTLass 방법

평균이 0인 연속형 확률변수 $X_1, \cdots, X_p$의 차원축소 (dimension reduction) 방법 중 선형적인 방법인 주성분분석 (principal component analysis)이 많이 사용되고 있다. 구체적으로 주성분분석 방법은 $\underline{X} = (X_1, \cdots, X_p)^T$의 선형결합 $\underline{l}^T \underline{X}$ (상첨자 T는 전치)의 분산을 최대화시키는 선형결합 벡터를 구하여, 이 선형결합 벡터로 계산되는 합성변수인 주성분 (principal component)을 이용하여 차원축소하는 방법이다. 이 문제에 대한 수학적 해법은 $X_1, \cdots, X_p$의 표본공분산행렬 S의 고유값-고유벡터 쌍 (eigenvalue-eigenvector pairs) $(\delta_i, \underline{v}_i)$, $i = 1, \ldots, p$에 의해 결정된다 (여기서 $\delta_1 \geq \cdots \geq \delta_p$로 정렬되어 있다고 가정한다). 구체적인 수학적 해법은 다음과 같이 주어진다 (Johnson과 Wichern, 1992).

$$\max_{\underline{l} \neq 0} \frac{\underline{l}^T S \underline{l}}{\underline{l}^T \underline{l}} = \delta_1 \text{이며 } \operatorname{argmax}_{\underline{l} \neq 0} \frac{\underline{l}^T S \underline{l}}{\underline{l}^T \underline{l}} = \underline{v}_1.$$

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

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

그런데 $\underline{l} \neq 0$에 대해 $\underline{v} = \underline{l}/\sqrt{\underline{l}^T\underline{l}}$를 새로운 선형결합 벡터라고 놓으면 길이가 $|\underline{v}|_2 \equiv \sqrt{\underline{v}^T\underline{v}} = 1$이 되기 때문에 선형결합 벡터 $\underline{l}$은 길이가 1이라고 가정해도 무방하다. 이렇게 선형결합 벡터를 구할 때 L^1 페널티를 부과하게 되면 ScoTLass 방법 (Jolliffe 등, 2003)이 된다. 구체적으로 L^1 페널티를 부과한 주성분 $X\underline{v}_k$ $(k = 1, \ldots, p)$의 선형결합 벡터 $\tilde{\underline{v}}_k$는 다음과 같이 구할 수 있다.

$$\tilde{\underline{v}}_1 = \text{argmax}_{|\underline{v}|_2=1} \left(\underline{v}^T S \underline{v} + \lambda_1 |\underline{v}|_1 \right). \tag{2.1}$$

$$\tilde{\underline{v}}_1, \ldots, \tilde{\underline{v}}_k \ (k = 1, \ldots, p-1)\text{에 직교인 } \underline{v}\text{에 대해}$$
$$\tilde{\underline{v}}_{k+1} = \text{argmax}_{|\underline{v}|_2=1} \left(\underline{v}^T S \underline{v} + \lambda_1 |\underline{v}|_1 \right).$$

여기서 $|\underline{v}|_1 = \sum_{i=1}^p |v_i|$는 $\underline{v} = (v_1, \ldots, v_p)^T$의 L^1 노름 (norm)이다.

ScoTLass 방법은 R을 이용한 프로그램이 존재하지 않으며, 또한 다음 절에서 설명될 페널티행렬분해 (penalized matrix decomposition) 방법에 의한 주성분분석이 ScoTLass 방법을 효과적으로 구현하는 하나의 알고리즘을 제공하는 것으로 알려져 있어 (Witten 등, 2009) 이 연구에서는 이 방법을 더 이상 다루지 않기로 한다.

2.2. elastic net 회귀를 이용한 주성분분석 방법

elastic net (Zou와 Hastie, 2005) 회귀를 이용한 주성분분석 방법을 요약하면 다음과 같다. 먼저 주성분분석에 의해 p개 주성분에 해당되는 벡터 $\underline{y}_1, \ldots, \underline{y}_p$를 구한다. 그 다음 이 주성분벡터를 반응변수로 하고 (평균이 0인) $X_1, \cdots, X_p$를 설명변수로 하는 elastic net 회귀에 의해 회귀계수 추정량을 구한다. 이 회귀계수 추정량을 길이가 1이 되도록 치환한 후 이것을 계수벡터로 이용한 주성분을 구한다.

상기 절차를 사용할 수 있는 핵심적인 근거는 다음과 같다.

정리 2.1 (Zou 등, 2006) 표본공분산행렬 S의 고유값-고유벡터 쌍을 $(\delta_i, \underline{v}_i)$, $(i = 1, \ldots, p)$라고 하자. 이 때 S의 계수 (rank)를 r $(r \leq p)$이라 가정하면 $\delta_1 \geq \cdots \geq \delta_r > 0 = \delta_{r+1} = \cdots = \delta_p$이라고 둘 수 있다. $n \times p$인 평균이 0인 자료행렬 X에 대해 $\underline{y}_i = X\underline{v}_i$를 i-번째 주성분 벡터라고 하면 다음이 성립한다.

$$\text{argmin}_{\underline{\beta}} \left\{ |\underline{y}_i - X\underline{\beta}|_2^2 + \lambda_2 |\underline{\beta}|_2^2 \right\} \propto \begin{cases} \underline{v}_i & \text{if } i \leq r, \\ 0 & \text{otherwise.} \end{cases}$$

증명: 이 정리는 공분산행렬 S 대신에 X^TX에 대해 ($i \leq r$인 경우에 한해) Zhou 등 (2006)에서 Theorem 1으로 주어져 있다. 그러나 이 연구에서 필요한 핵심 근거 중의 하나이기 때문에 자체적인 증명을 제시하도록 하겠다. 다중회귀모형 (multiple regression model) $\underline{y}_i = X\underline{\beta} + \underline{\epsilon}$의 회귀계수 $\underline{\beta}$의 능형 추정량 (ridge estimator)은 다음과 같다.

$$\tilde{\underline{\beta}}_i = (X^TX + \lambda_2 I)^{-1} X^T \underline{y}_i = (X^TX + \lambda_2 I)^{-1} X^T X \underline{v}_i = (X^TX + \lambda_2 I)^{-1} \delta_i \underline{v}_i.$$

따라서 $i > r$에 대해서는 $\delta_i = 0$이기 때문에 $\tilde{\underline{\beta}}_i = 0$이 된다.

$i \leq r$에 대해서는 다음과 같이 $\tilde{\underline{\beta}}_i$를 직접 계산한다. $S = X^TX/(n-1)$이기 때문에 스펙트럼분해 (spectral decomposition)에 의해 $X^TX = (n-1)V\Delta V^T$가 된다. 여기서 $\Delta = \text{diag}(\delta_1, \ldots, \delta_r)$이며 $V = (\underline{v}_1, \ldots, \underline{v}_r)$이다. 따라서 다음이 성립한다.

$$(X^TX + \lambda_2 I)^{-1} = \left(V((n-1)\Delta + \lambda_2 I)V^T \right)^{-1} = V((n-1)\Delta + \lambda_2 I)^{-1} V^T.$$

그러므로 다음이 성립한다.

$$(X^TX+\lambda_2 I)^{-1}\delta_i \underline{v}_i = V\left((n-1)\Delta+\lambda_2 I)\right)^{-1}V^T\delta_i\underline{v}_i$$
$$= \left(\sum_{j=1}^{r}\frac{\delta_i}{(n-1)\delta_j+\lambda_2}\underline{v}_j\underline{v}_j^T\right)\underline{v}_i = \frac{\delta_i}{(n-1)\delta_i+\lambda_2}\underline{v}_i.$$

이것으로 증명이 마무리된다. □

참고로 X^TX의 역행렬이 존재하면 (다시 말해 $r=p\le n$이면) $\text{argmin}_{\underline{\beta}}|\underline{y}_i - X\underline{\beta}|_2^2 = \underline{v}_i$가 된다 (Park, 2013). 그러나 $p>n$ (혹은 $r<p$)과 같은 경우에는 X^TX의 역행렬이 존재하지 않기 때문에 최소제곱 회귀를 적용할 수 없고, 정리 2.1과 같은 능형 회귀를 통해 주성분의 계수벡터를 구할 수 있다.

따라서 elastic net 회귀를 이용한 주성분분석의 계수벡터는 다음과 같이 구할 수 있다.

$$\tilde{\underline{\beta}}_i^* = \text{argmin}_{\underline{\beta}}\left\{|\underline{y}_i - X\underline{\beta}|_2^2+\lambda_1|\underline{\beta}|_1+\lambda_2|\underline{\beta}|_2^2\right\},\ (i\le r). \tag{2.2}$$

그러나 상기 추정량은 길이가 1인 조건을 만족하지 못할 수 있기 때문에 길이가 1이 되도록 조정한 $\tilde{\underline{\beta}}_i^{**} = \tilde{\underline{\beta}}_i^*/|\tilde{\underline{\beta}}^*|_2,\ (i\le r)$를 사용하게 된다. 따라서 Lasso 회귀의 성질에 의해 추정량 $\tilde{\underline{\beta}}_i^{**},\ (i\le r)$는 능형 회귀에서의 추정량 $\tilde{\underline{\beta}}_i \propto \underline{v}_i$보다 0의 값을 많이 가질 것이기 때문에 주성분의 해석이 용이하게 된다.

Zhou 등 (2006)에서는 앞에서와 같은 2단계를 하나의 단계로 통합하는 주성분분석의 계수벡터 반복계산 알고리즘을 소개하고 있으나, 이 연구에서는 그 방법을 사용하지 않도록 한다. 왜냐하면 그 방법에 의할 경우 여러 주성분의 계수벡터들을 동시에 추정하기 때문에, 고유값이 큰 순서에 따라 순차적으로 각 주성분에 적절한 희소성 (sparsity)을 가지는 계수벡터를 탐색하는 것이 훨씬 어렵기 때문이다.

2.3. 페널티행렬분해를 이용한 주성분분석 방법

이 절에서는 특이값 분해 (singular value decomposition)에 의한 차수가 $n\times p$인 (평균이 0인) 원자료행렬 X의 근사와 그 근사과정에서 Lasso 페널티를 부과하는 페널티행렬분해 (peneralized matrix decomposition) 방법을 설명한다. 먼저 X의 특이값 분해를 나타내기 위한 표기법을 정의하자. 차수가 $n\times p$인 (평균이 0인) 원 자료행렬 X의 계수 (rank)가 r $(r\le p)$이면 특이값 분해에 의해 다음과 같이 나타낼 수 있다.

$$X = U\,D\,V^T,$$

여기서 $D=\text{diag}(d_1,\ldots,d_r)$은 대각원소가 $d_1\ge\ldots,\ge d_r>0$인 대각행렬, $U=(\underline{u}_1,\ldots,\underline{u}_r)$은 길이가 1인 직교 열벡터 (orthogonal column vector)로 구성된 $n\times r$ 행렬 (즉, $U^TU=I$ 만족)이며, 행렬 $V=(\underline{v}_1,\ldots,\underline{v}_r)$는 길이가 1인 직교 열벡터로 구성된 $p\times r$ 행렬 (즉, $V^TV=I$)이다. 앞의 정리 1의 증명에서 사용했던 스펙트럼분해 $S=V\,\Delta\,V^T$ 관점에서 표현하면 V는 동일하고 $\Delta=D^2/(n-1)$가 성립되는 것을 알 수 있다.

그러면 계수 k $(k\le r)$인 X의 근사에 대한 다음의 사실이 잘 알려져 있다 (Eckart와 Young, 1936).

$$\sum_{i=1}^{k} d_i\,\underline{u}_i\,\underline{v}_i^T = \text{argmin}_{\hat{X}\in M(k)}|X-\hat{X}|_2^2,$$

여기서 $M(k)$은 계수가 k인 차수가 $n\times p$ 행렬의 집합이며, $|M|_2^2\equiv\text{tr}(M^TM)$는 행렬 M의 제곱합이다. 따라서 계수가 1인 X의 근사는 다음과 같이 표시할 수 있다.

$$(d_1, \underline{u}_1, \underline{v}_1) = \text{argmin}_{(d, |\underline{u}|_2^2=1, |\underline{v}|_2^2=1)} |X - d\,\underline{u}\,\underline{v}^T|_2^2.$$

여기에 $\underline{v}$에 L^1 페널티를 부과하게 되면 주성분 $\underline{z} = X\underline{v}$을 구할 때 계수벡터에 희소성 (sparsity)이 생겨 주성분을 해석하기 쉽게 된다. 그런데 조건 $|\underline{v}|_2^2 = 1, |\underline{v}|_2^2 = 1$은 볼록이 아니기 때문에 일반적으로 다음과 같은 볼록 조건을 부여한 문제로 치환하여 해를 구할 수 있다.

$$\text{minimize}_{d,\underline{u},\underline{v}} |X - d\,\underline{u}\,\underline{v}^T| \text{ subject to } |\underline{u}|_2^2 \le 1, |\underline{v}|_2^2 \le 1, |\underline{v}|_1 \le c, \tag{2.3}$$

여기서 $|\underline{v}|_1 = \sum_{i=1}^p |v_i|$은 $\underline{v}$의 L^1 노름이다. 그런데 다음이 성립한다.

$$|X - d\,\underline{u}\,\underline{v}^T|_2^2 = \text{tr}(X^TX - 2dX^T\underline{u}\underline{v}^T + d^2\underline{v}\underline{u}^T\underline{u}\underline{v}^T) = \text{tr}(X^TX) - 2d\,\underline{u}^TX\underline{v} + d^2.$$

따라서 $\underline{u}$, $\underline{v}$이 주어지면 $d = \underline{u}^TX\underline{v}$로 계산되기 때문에, d를 제외한다면 앞의 (2.3)의 문제는 다음과 같이 표현할 수 있다.

$$\text{maximize}_{\underline{u},\underline{v}} \underline{u}^TX\underline{v} \text{ subject to } |\underline{u}|_2^2 \le 1, |\underline{v}|_2^2 \le 1, |\underline{v}|_1 \le c. \tag{2.4}$$

앞의 (2.3)를 만족하는 알고리즘을 표현하기 위해 다음과 같은 소프트분계점 연산자 (soft threshold operator)를 정의하자.

$$S(a, c) = \text{sign}(a)\max(|a| - c, 0).$$

그러면 앞의 (2.3)를 만족하는 계수가 1인 근사 $(d, \underline{u}, \underline{v})$는 다음의 알고리즘에 의해 계산될 수 있다 (Zou 등, 2006).

Algorithm 2.1: Rank 1 approximation of X

1. Initialize $\underline{v}$ to have L^2 norm 1.
2. Iterate until converngece:
   - $\underline{u} \leftarrow X\underline{v}/|X\underline{v}|_2$.
   - $\underline{v} \leftarrow S(X^T\underline{u}, \delta)/|S(X^T\underline{u}, \delta)|_2$, where $\delta = 0$ if this results in $|\underline{v}|_1 \le c$; otherwise δ is chosen to be a positive number such that $|\underline{v}|_1 = c$.
3. $d \leftarrow \underline{u}^TX\underline{v}$.

앞의 Algorithm 2.1의 2단계는 (2.4)을 만족시키는 $(\underline{u}, \underline{v})$를 구하는 과정이다. 여기서 2단계의 첫 번째 계산 $\underline{u} \leftarrow X\underline{v}/|X\underline{v}|_2$를 (2.4)에 대입하면 다음과 표현된다.

$$\text{maximize}_{\underline{v}} \underline{v}^TX^TX\underline{v} \text{ subject to } |\underline{v}|_2^2 \le 1, |\underline{v}|_1 \le c. \tag{2.5}$$

이것을 라그랑즈승수 (Lagrange multiplier) 형태로 표현하면 (2.1)에 주어진 ScoTLass 방법 (Jollife 등, 2006)과 같아져, 결국 Algorithm 2.1은 ScoTLass 방법을 효과적으로 구현하는 알고리즘이 되는 것을 알 수 있다.

계수가 $r = \text{rank}(X)$인 X의 근사는 다음의 순차적 잔차 알고리즘에 의해 계산될 수 있다 (Zou 등, 2006).

Algorithm 2.2: Rank $r = \text{rank}(X)$ approximation of X

1. Let $X^1 = X$.
2. For $k = 1, \ldots, r$:
   - Find $(d_k, \underline{u}_k, \underline{v}_k)$ by applying Algorithm 2.1 to data X^k.
   - $X^{k+1} \leftarrow X^k - d_k \, \underline{u}_k \, \underline{v}_k^T$.

이 연구에서 페널티행렬분해를 이용한 주성분분석 방법으로 사용하는 것이 바로 Algorithm 2.2이다.

3. 실제 자료 적용 예제

이 절에서는 하나의 실제 자료 (real data)에 elastic net 회귀를 이용한 주성분분석 방법과 페널티 행렬분해를 이용한 주성분분석 방법을 적용한다. 자료분석에는 R을 사용하였으며 구체적으로 elastic-net과 PMA 패키지를 이용하였다.

실제 적용 예제로 사용된 자료는 Ahamad (1967)의 crime 자료이다. 이 자료를 선택한 이유는 일반인이 비교적 쉽게 이해할 수 있는 예제로서 $p > n$을 만족하기 때문이다. 구체적으로 이 자료는 1950년에서 1963년 사이에 잉글랜드와 웨일즈에 발생한 18개 종류의 범죄 (indictable offenses 12개: homicide, woundings, homosexual offenses, heterosexual offenses, breaking and entering, robbery, larceny, frauds and false pretences, receiving stolen goods, malicious injuries to property, forgery, blackmail; non-indictable offenses 6개: assault, malicious damage, revenue laws, intoxication laws, indecent exposure, taking motor vehicle without consent)에 대한 자료이다.

우선 페널티가 주어지지 않은 주성분분석을 실행하였다. 참고로 R에서 $p > n$이면 princomp 함수를 사용할 수 없고 prcomp 함수를 사용하여야 한다. crime 자료의 상관행렬에 대해 값이 제일 큰 고유값 5개를 살펴보았더니 12.896, 2.712, 0.957, 0.684, 0.322가 나와 처음 2개 혹은 3개의 주성분을 사용하는 것이 적절하다고 판단된다. 여기서는 페널티에 따른 효과를 좀 더 자세히 살펴보기 위해 처음 3개의 주성분을 사용하도록 한다. 처음 3개의 주성분에 해당되는 고유벡터는 다음과 같다.

Table 3.1 Eigenvectors corresponding to the first three largest eigenvalues

Variable name	PC1	PC2	PC3
Homicide	0.0240	-0.3277	0.8157
Woundings	0.2703	0.1207	-0.0557
Homosexual	-0.0868	0.4907	0.1189
Heterosexual	0.2554	0.1787	-0.1671
Breaking	0.2761	-0.0309	0.0213
Robbery	0.2718	0.0249	-0.0939
Larceny	0.2762	-0.0076	-0.0302
Frauds	0.2736	-0.0694	-0.0318
Receiving	0.2690	-0.0763	0.0248
Property	0.1788	-0.3433	0.1135
Forgery	0.2711	0.0481	0.0860
Blackmail	0.2677	-0.0036	0.0839
Assault	-0.1176	-0.4464	-0.3558
Damage	0.2625	0.1070	-0.1419
Revenue	0.2654	-0.1409	0.0758
Intoxication	0.2631	0.1663	0.0480
Exposure	0.0940	-0.4566	-0.3088
TakingMotor	0.2676	0.0781	0.0153

3.1. elastic net 회귀를 이용한 주성분분석 방법

crime 자료에 elastic net을 적용하기 위해서는 L^2에 해당되는 λ_2 페널티를 양수값으로 지정해야 한다. 여기서는 상대적으로 작은 0.5로 고정하여 사용하였다. elasticnet 패키지에서 elastic net 회귀를 실행하는 함수인 enet에는 고정된 L^2 페널티에 따라 LARS-EN 알고리즘 (Zhou Hastie, 2005)을 사용한다. 따라서 계수가 0이 되는 변수의 숫자인 희박성 (sparsity) 값들 사이에는 누적 설명분산 비율 (CPEV; cumulative proportion of explained variance)이 선형함수가 되어 적절한 희박성을 선택하는 데 용이하다. 이 연구에서는 Shen과 Huang (2008)에 의해 제안된 CPEV를 사용하였다. 세 개의 주성분을 각각 반응변수로 하고 18개의 표준화된 범죄수를 설명변수로 하는 elastic net 회귀의 희박성과 누적 설명분산 비율을 그림으로 나타낸 것이 Figure 3.1이다.

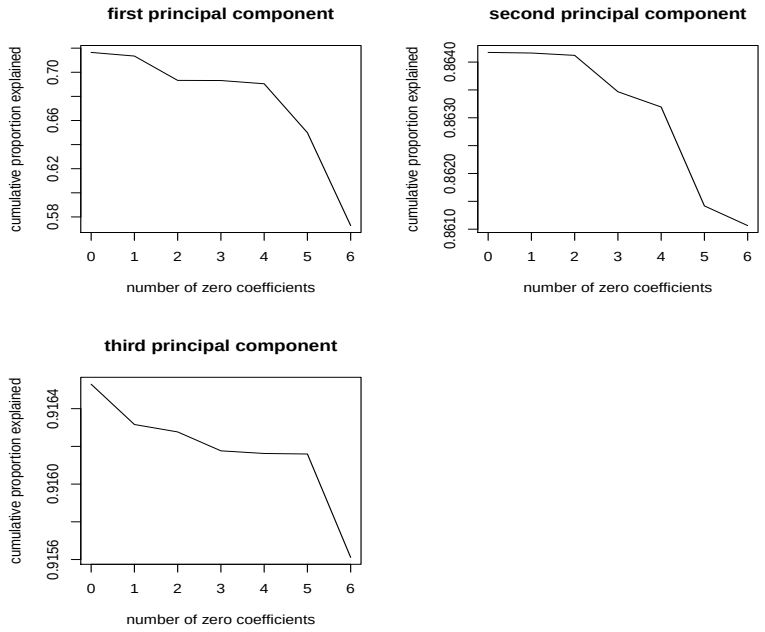

Figure 3.1 CPEV versus sparsity for the first three principal components using elestic net regression

위의 그림에서 보면 첫 번째 주성분에서는 CPEV가 최대 대비 0.01이내로 감소하는 희박성 1을 선택하였고, 두 번째 주성분에서는 CPEV가 최대 대비 0.001이내로 감소하는 희박성 4, 세 번째 주성분에서는 CPEV가 최대 대비 0.0005이내로 감소하는 희박성 5를 선택하였다. 이렇게 희박성을 1, 4, 5로 선택하였을 때 해당되는 길이가 1인 계수벡터 $\tilde{\beta}_i^{**}, i = 1, 2, 3$은 다음 Table 3.2와 같다.

Table 3.2을 보면 희박성, 즉 계수벡터의 값이 0인 변수의 숫자가 정확히 1, 4, 5가 된 것을 알 수 있다. 이렇게 선택된 계수벡터에 의한 CPEV는 0.7135, 0.8631, 0.9162로 나타나 원래 주성분의 CPEV인 0.7165, 0.8672, 0.9204에 비해 아주 미약하게 작아진 것을 알 수 있다.

elasticnet 패키지에 있는 cv.enet 함수를 이용하여 spar (sparsity parameter; mode="frac") 값을 구했더니 0.6364, 0.7172, 1.0이 나왔다. 그런데 이 spar 값에 해당되는 주성분의 희박성은 각각 4, 9, 0, CPEV가 0.6188, 0.7626, 0.8161이 나와 제1, 2주성분에 과도한 희박성이 생기고 CPEV가 상당히 많이 감소한 것을 알 수 있다.

Table 3.2 Coefficients vectors for the new principal components with sparsity 1, 4, 5 using elastic net regression

Variable name	PC1	PC2	PC3
Homicide	0.0000	-0.3327	0.8367
Woundings	0.2671	0.1056	-0.0276
Homosexual	-0.0611	0.5052	0.1063
Heterosexual	0.2414	0.1723	-0.1498
Breaking	0.2839	0.0000	0.0000
Robbery	0.2715	0.0000	-0.0720
Larceny	0.2814	0.0000	0.0000
Frauds	0.2820	-0.0348	0.0000
Receiving	0.2778	-0.0471	0.0000
Property	0.1794	-0.3477	0.0995
Forgery	0.2779	0.0215	0.0673
Blackmail	0.2748	0.0000	0.0578
Assault	-0.0791	-0.4499	-0.3676
Damage	0.2535	0.0931	-0.1234
Revenue	0.2779	-0.1147	0.0530
Intoxication	0.2598	0.1525	0.0218
Exposure	0.0668	-0.4672	-0.2980
TakingMotor	0.2680	0.0520	0.0000

3.2. 페널티행렬분해를 이용한 주성분분석 방법

페널티행렬분해를 이용한 주성분분석 방법은 PMA 패키지를 이용하여 분석한다. 구체적으로 SPC 함수를 이용하는데 앞의 enet과는 달리 희박성 값들 사이에 선형성이 만족되지 않기 때문에 spar 값들을 변동시키면서 CPEV를 구하여 적절한 spar 값을 선택해야 한다. 따라서 이 연구에서는 spar 값들이 1.2에서 부터 4.0까지 0.2씩 증가시키는 격자 (grid) 값들을 고려하도록 한다. 세 개의 주성분을 각각 반응변수로 하고 18개의 표준화된 범죄수를 설명변수로 하는 SPC에서 spar 값과 누적 설명분산 비율을 그림으로 나타낸 것이 Figure 3.2이다.

Figure 3.2에서 보면 첫 번째 주성분에서는 CPEV가 최대 대비 0.1이내로 감소하는 spar 3.8을 선택하였고, 두 번째 주성분에서는 CPEV가 최대 대비 0.002이내로 감소하는 spar 2.8, 세 번째 주성분에서는 CPEV가 최대 대비 0.001 근처인 spar 2.0으로 선택하였다. 그런데 세 번째 주성분에서는 spar 값 1.8에서 2.0 사이에 CPEV가 급격히 변하는 관계로 이 부분의 격자 값을 세분하여 집중 검토한 결과 CPEV가 최대 대비 0.002 이내인 spar 1.88을 선택하였다 (여기서는 보이지 않음). 이렇게 spar를 3.8, 2.8, 1.88으로 선택하였을 때 해당되는 길이가 1인 계수벡터 $\tilde{v}_i, i = 1, 2, 3$는 Table 3.3과 같다.

Table 3.3을 살펴보면 희박성이 앞의 elastic net 회귀와 마찬가지로 1, 4, 5이며, 이 중 elastic net 회귀와 동시에 계수값이 0으로 나타난 변수는 1, 3, 4개가 되는 것을 알 수 있다. 그리고 처음 두 개의 주성분의 계수값들도 elastic net 회귀와 아주 비슷하게 나타나고 있으며, 세 번째 주성분의 계수값은 다소 차이가 차이가 있으나 계수값이 0에서 0.2이상 벗어난 경우에는 부호가 반대로 나타나는 경향이 있음을 알 수 있다. 이렇게 선택된 계수벡터에 의한 CPEV는 0.7060, 0.8585, 0.9102가 되어 elastic net 회귀를 이용한 결과보다 CPEV가 약간 작게 나타나고 있다.

PMA 패키지에 있는 SPC.cv 함수를 이용하여 bestsumabsv, 즉 spar 값을 구했더니 각각 4.0, 3.1, 1.6 (해당 희박성은 0, 0, 12)이 나왔다. 이 spar 값들에 따른 CPEV가 각각 0.7165, 0.8671, 0.9172로 나와 Table 3.3보다 약간 개선되기는 했지만 희소성이 세 번째 주성분에 집중적으로 몰리는 단점이 발견되고 있다.

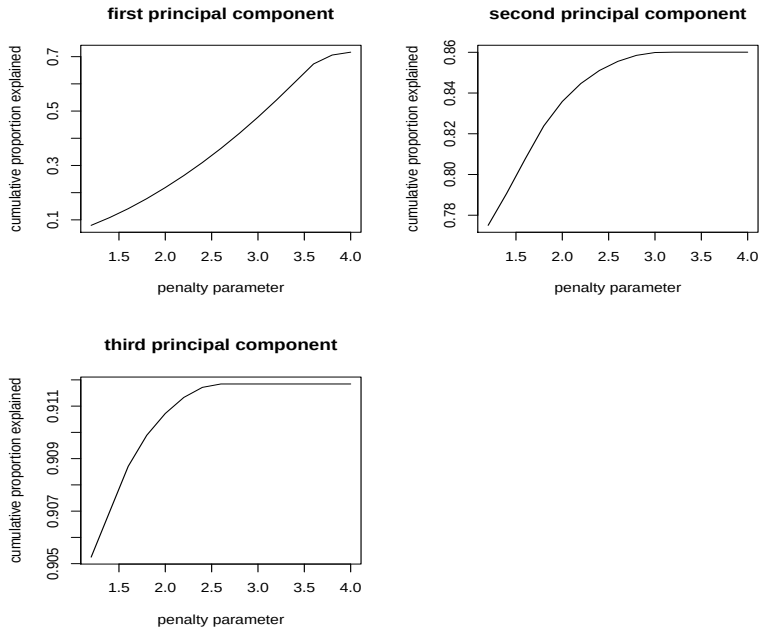

Figure 3.2 CPEV versus sparsity parameter for the first three principal components using penalized matrix decomposition

Table 3.3 Coefficients vectors for the first three principal components with sparsity parameter 3.8, 2.8, 2.0 using penalized matrix decomposition

Variable name	PC1	PC2	PC3
Homicide	0.0000	-0.3180	-0.8844
Woundings	0.2786	0.0705	0.0074
Homosexual	-0.0102	0.5276	0.0000
Heterosexual	0.2414	0.1723	-0.1498
Breaking	0.2830	-0.0147	0.0000
Robbery	0.2787	0.0000	0.0308
Larceny	0.2838	0.0000	0.0000
Frauds	0.2782	-0.0579	0.0053
Receiving	0.2718	-0.0642	0.0000
Property	0.1394	-0.3614	-0.0061
Forgery	0.2772	0.0000	-0.0192
Blackmail	0.2721	0.0000	-0.0745
Assault	-0.0713	-0.4236	0.2666
Damage	0.2681	0.0546	0.0660
Revenue	0.2654	-0.1354	-0.0400
Intoxication	0.2694	0.1213	-0.0167
Exposure	0.0200	-0.4936	0.3463
TakingMotor	0.2734	0.0248	0.0000

4. 결론

이 연구에서는 Lasso 페널티 방법을 이용한 주성분분석 방법을 소개하였다. 구체적으로 주성분분석

에 Lasso 페널티를 부과하는 방법으로 두 가지를 고려하였다. 첫 번째 방법은 주성분을 반응변수로 놓고 원 자료행렬을 설명변수로 하는 회귀분석의 회귀계수를 구할 때 elastic net 페널티를 부과하는 방법이다. 두 번째 방법은 원 자료행렬을 비정칙값 분해로 근사하고 남은 잔차행렬에 Lasso 페널티를 부과하여 최적의 선형결합 벡터를 구하는 방법이다. Lasso 페널티를 가하는 이 두 가지 방법들은 변수 숫자가 표본크기보다 큰 경우에도 적용가능한 장점이 있다. R 프로그램을 통해 두 방법을 적용하고 그 결과를 비교하기 위해서 Ahamad (1967)의 crime 자료에 적용하는 실제 적용 예제를 제공하였다. 그 결과 두 방법은 처음 두 개의 주성분에서 희박성과 계수값에서 아주 유사한 결과를 얻었으며, 세 번째 주성분에서도 계수값이 큰 경우 부호가 반대로 나오기는 하지만 희박성이 아주 비슷한 결과를 얻었다.

elastic net 회귀를 이용할 때의 장점은 희박성을 선택할 때 희박성 값 사이에 CPEV 값이 선형관계가 유지되기 때문에 시각적으로 확인하면서 적절한 희박성을 쉽게 선택할 수 있다는 점이다. 그에 반해 페널티행렬분해를 이용할 때의 장점은 (1차 계수 근사에서) ScoTLass 방법의 효율적인 알고리즘을 제공하며, 또한 조율파라미터 (tuning parameter)가 없다는 점이다. 또한 페널티행렬분해를 이용할 때는 주성분이 서로 직교가 되게 만들 수 있다는 장점도 있다. 이러한 장점을 참조하여 분석하는 자료의 형태와 분석목적에 따라 적절한 주성분분석 방법을 선택하면 바람직한 자료분석을 행할 수 있으리라 기대해 본다.

References

Ahamad, B. (1967). An analysis of crimes by the method of principal components. *Applied Statistics*, **16**, 17-35.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of low rank. *Psychometrika*, **1**, 211-218.

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.

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

Kwon, S., Han S. and Lee, S. (2013). A small review and further studies on the LASSO. *Journal of the Korean Data & Information Science Society*, **24**, 1077-1088.

Park, C. (2013). Simple principal component analysis using Lasso. *Journal of the Korean Data & Information Science Society*, **24**, 533-541.

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.

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

Witten, D. A., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with application to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515-534.

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
2017, **28**(4), 721–731

http://dx.doi.org/10.7465/jkdi.2017.28.4.721
한국데이터정보과학회지

A study on principal component analysis using penalty method

Cheolyong Park[1]

[1]Major in Statistics, Keimyung University

Received 5 June 2017, revised 8 July 2017, accepted 11 July 2017

Abstract

In this study, principal component analysis methods using Lasso penalty are introduced. There are two popular methods that apply Lasso penalty to principal component analysis. The first method is to find an optimal vector of linear combination as the regression coefficient vector of regressing for each principal component on the original data matrix with Lasso penalty (elastic net penalty in general). The second method is to find an optimal vector of linear combination by minimizing the residual matrix obtained from approximating the original matrix by the singular value decomposition with Lasso penalty. In this study, we have reviewed two methods of principal components using Lasso penalty in detail, and shown that these methods have an advantage especially in applying to data sets that have more variables than cases. Also, these methods are compared in an application to a real data set using R program. More specifically, these methods are applied to the crime data in Ahamad (1967), which has more variables than cases.

Keywords: Elastic net, lasso, penalty, principal component analysis, regression model.

[1] Professor, Major in Statistics, Keimyung University, Daegu 42601, Korea. E-mail: cypark1@kmu.ac.kr