The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.1.135
30(1), 135–145

Hierarchically penalized sparse principal component analysis

Jongkyeong Kang[a] · Jaeshin Park[a] · Sungwan Bang[a,1]

[a]Department of Mathematics, Korea Military Academy

(Received November 16, 2016; Revised January 12, 2017; Accepted January 25, 2017)

Abstract

Principal component analysis (PCA) describes the variation of multivariate data in terms of a set of uncorrelated variables. Since each principal component is a linear combination of all variables and the loadings are typically non-zero, it is difficult to interpret the derived principal components. Sparse principal component analysis (SPCA) is a specialized technique using the elastic net penalty function to produce sparse loadings in principal component analysis. When data are structured by groups of variables, it is desirable to select variables in a grouped manner. In this paper, we propose a new PCA method to improve variable selection performance when variables are grouped, which not only selects important groups but also removes unimportant variables within identified groups. To incorporate group information into model fitting, we consider a hierarchical lasso penalty instead of the elastic net penalty in SPCA. Real data analyses demonstrate the performance and usefulness of the proposed method.

Keywords: principal component analysis, sparse PCA, hierarchical penalty, grouped variable

1. 서론

대표적인 다변량 분석기법 중 하나인 주성분 분석(principal component analysis; PCA)은 고차원의 변수들을 선형 연관성이 없는 저차원 공간으로 축소, 요약하는 기법으로 서로 상관되어 있는 변수들 간의 복잡한 구조를 분석하는데 많이 사용되고 있다. 즉, 주성분 분석은 변수들의 선형변환을 통해 자료에 존재하는 원래의 변동(variation)을 가능한 한 많이 설명하는 새로운 인공 변수, 즉 주성분을 생성하는 것을 그 목적으로 한다. 주성분 분석은 자료의 변동성을 가장 많이 표현하는 처음 몇 개의 주성분만을 선택함으로써 결과적으로 정보의 손실을 최소화하면서 차원의 축소를 이루어 낸다. 그러나 일반적인 주성분 분석은 모든 변수들의 선형결합(linear combination)을 이용하기 때문에 이에 대응하는 적재(loading)는 일반적으로 0이 아니며, 이는 생성된 주성분의 특성에 관한 분석을 하고자 할 때 그 해석을 어렵게 한다. 특히, 근래 화두에 오른 유전배열(gene expression)자료 또는 빅데이터와 같이 변수가 매우 많은 경우에서의 전통적인 주성분 분석은 적용과 활용에 한계가 있다.

This research was supported by 2015 research fund of Korea Military Academy (20150501) for J. Kang and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2015R1C1A1A02036473) for S. Bang.
[1]Corresponding author: Department of Mathematics, Korea Military Academy, 574, Hwarang-ro, Nowon-gu, Seoul 01805, Korea. E-mail: wan1365@gmail.com

이러한 주성분 분석의 제한사항을 해결하기 위해 Zou 등 (2006)은 주성분 분석을 회귀모형 형태의 최적화 문제로 전환하여 성긴(sparse) 적재를 생성하는 sparse PCA(SPCA)를 제안하였다. SPCA는 회귀분석에서 보편적으로 사용되는 벌점화 방법인 lasso (Tibshirani, 1996)와 ridge (Hoerl과 Kennard, 1970) 형태의 벌점함수를 결합한 elastic net (Zou와 Hastie, 2003) 형태의 벌점함수를 주성분 분석에 적용하여 그 결과로 성긴 적재를 얻으며, 기존의 주성분 분석에 비해 설명력을 많이 잃지 않으면서도 보다 간결한 모형을 얻는 특징이 있다.

한편, 변수들이 그룹화되어 있는 경우 각각의 변수를 식별하는 것 보다 유의한 변수들의 그룹을 식별하는 것이 더 의미가 있을 수 있다. Lasso와 이를 일반화한 elastic net은 개별적인 변수들은 효과적으로 선택하지만, 변수들의 그룹구조를 이용하지 못하므로 변수들간의 상관관계가 높은 경우에는 변수선택 능력이 다소 떨어지며, SPCA 또한 elastic net 형태의 벌점함수를 적용하고 있기 때문에 이러한 제한점을 그대로 가지고 있다. 변수들이 그룹화되어 있는 자료의 분석에 있어서 Yuan과 Lin (2006)은 제곱손실함수를 이용한 회귀모형에서 group lasso의 사용을 제안했으며, Kang 등 (2016)과 Wang 등 (2009)은 각각 Cox 회귀모형과 분위수 회귀모형에서 계층적 벌점함수의 사용을 제안한 바 있다. 그리고 주성분 분석에서 변수들 간의 다블록(multiblock) 효과에 대한 선택 문제를 개선하기 위한 방법으로 Bernard 등 (2012)은 group lasso 벌점함수를 이용한 group SPCA(G-SPCA)를 제안한 바 있다. 그러나 G-SPCA는 group lasso 벌점함수의 특성으로 인하여 그룹 간 변수선택에는 효과적이나 그룹 내에서의 변수선택은 이루어지지 않는다.

본 논문에서는 변수들이 그룹화되어 있는 다변량자료의 분석에 적용 및 활용할 수 있는 주성분 분석 기법에 관하여 연구하였다. 그룹과 그룹 내 변수 구조를 모델 적합에 동시에 이용하기 위하여, lasso 또는 group lasso 형태의 벌점함수 대신에 계층적 벌점함수 (Kang 등, 2016; Wang 등, 2009)를 적용하는 새로운 주성분 분석 기법을 제안하였으며, 모형적합을 위한 계산 알고리즘으로 이차계획법(quadratic programming)을 이용한 반복 추정방법을 제시하였다. 본 논문의 구성은 다음과 같다. 2절에서는 벌점함수를 이용한 주성분 분석 방법인 SPCA와 G-SPCA에 대해 간략히 설명하고, 3절에서는 본 논문의 제안방법인 계층적 벌점함수(hierarchical penalty function)를 이용한 SPCA(H-SPCA)의 소개와 계산 알고리즘에 대하여 논의하였다. 4절에서는 실제 자료의 분석 결과를 나타내었고, 마지막으로 5절에서는 결론과 더불어 차후 연구방향을 제시하였다.

2. 벌점함수를 이용한 주성분 분석 기법 비교

2.1. Sparse PCA(SPCA)

주성분 분석에서는 서로 연관 가능성이 있는 고차원의 변수들을 저차원 공간의 새로운 변수들로 변환하기 위해 직교 변환을 이용한다. $\boldsymbol{X}$를 $n \times p$ 자료 행렬이라 하자. 여기서 n과 p는 각각 관측치와 변수의 개수를 나타내고, 자료 행렬 $\boldsymbol{X}$의 각 열은 평균이 0으로 중심화되어 있다고 가정하자. 자료행렬 $\boldsymbol{X}$는 특이값 분해(singular value decomposition; SVD)를 통해

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \tag{2.1}$$

와 같이 표현할 수 있다. 여기서 $\boldsymbol{U}$는 $n \times r$ 행렬, $\boldsymbol{V}$는 $p \times r$ 행렬, 그리고 $\boldsymbol{D}$는 $d_1, \ldots, d_r$을 대각원소로 갖는 $r \times r$ 대각행렬이며, $\boldsymbol{U}$와 $\boldsymbol{V}$의 각 열은 정규직교이다. 그러면 $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}$는 주성분이 되고, $\boldsymbol{V}$의 각 열은 해당 주성분에 대응하는 적재(loading)가 되며, $\boldsymbol{D}$의 대각원소 $d_1 \geq d_2 \geq \cdots \geq d_r > 0$은 해당 주성분의 표본분산을 나타낸다. 임의의 i에 대하여, $\boldsymbol{z}_i = \boldsymbol{u}_i d_i$라고 하면, $\boldsymbol{z}_i =$ 는 i번째 주성분이 된다. 처음 몇 개의 주성분을 선택함으로써 우리는 최소한의 정보 손실만을 가져오면서 전체 자료의 변동을 효

과적으로 설명할 수 있다. 그러나 식 (2.1)을 통해 구한 주성분의 적재값들은 일반적으로 0이 아니며, 이는 추정된 주성분의 해석을 어렵게 한다. 이러한 점을 보완하기 위해 Zou 등 (2006)은 성긴 적재를 갖는 수정된 주성분을 얻기 위한 SPCA를 제안하였다. 이 방법론에서의 주요점은 주성분 분석이 회귀모형의 형태로 공식화될 수 있다는 것이다.

양의 λ값에 대한 다음과 같은 ridge 회귀모형을 고려하자.

$$\hat{\boldsymbol{\beta}}_i^{\text{Ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ ||\boldsymbol{z}_i - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_2^2 \right\}, \tag{2.2}$$

여기서 $\|\cdot\|_2$는 L^2 노름(norm)이다. 한편 $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$이고, $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$이므로, 추정된 회귀계수 $\hat{\boldsymbol{\beta}}_i^{\text{Ridge}}$는

$$\hat{\boldsymbol{\beta}}_i^{\text{Ridge}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{v}_i) = \boldsymbol{v}_i \frac{d_i^2}{d_i^2 + \lambda} \tag{2.3}$$

가 된다. 즉, $\hat{\boldsymbol{\beta}}_i^{\text{Ridge}}$를 표준화한 $\hat{\boldsymbol{v}}_i = \hat{\boldsymbol{\beta}}_i / ||\hat{\boldsymbol{\beta}}_i||_2$는 $\boldsymbol{v}_i$, 즉 i번째 주성분에 대응하는 적재가 되며, $\boldsymbol{X}\hat{\boldsymbol{v}}_i$는 i번째 주성분의 추정치가 된다. 이러한 사실은 ridge 형태의 회귀모형으로의 변환을 통해 각각의 주성분에 대한 적재를 구할 수 있음을 의미한다. 또한 양의 값을 갖는 ridge 형태의 벌점함수 $\lambda||\boldsymbol{\beta}||_2^2$은 어떤 경우에서든 회귀방법론을 통해 주성분을 계산할 수 있음을 보장한다. 예를 들어, $n < p$인 경우 $\boldsymbol{X}$는 계수(rank)가 n이 되지 않으므로, 일반적인 회귀방법론으로는 계수를 추정할 수 없지만, ridge 형태의 벌점함수의 도움으로 회귀계수를 추정할 수 있기 때문이다. 여기서 $\hat{\boldsymbol{v}}_i$는 표준화로 인해 λ의 선택과는 독립적인 값을 갖는다. 다시 말해서 ridge 형태의 벌점함수는 계수의 추정에 벌점을 가한다기 보다는 $n < p$인 경우에서도 회귀계수를 추정할 수 있게 한다는 특징을 갖는다.

이러한 사실로부터 Zou 등 (2006)은 식 (2.2)의 ridge 형태의 회귀모형에 추가적으로 lasso 형태의 벌점함수를 적용함으로써 성긴 적재를 얻을 수 있는 SPCA를 다음과 같이 제안하였다.

$$\hat{\boldsymbol{\beta}}_i^{\text{SPCA}} = \arg\min_{\boldsymbol{\beta}} \left\{ ||\boldsymbol{z}_i - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_2^2 + \lambda_1 ||\boldsymbol{\beta}||_1 \right\}, \tag{2.4}$$

여기서 $\|\cdot\|_1$는 L^1 노름(norm)이다. 이 경우, $\hat{\boldsymbol{v}}_i = \hat{\boldsymbol{\beta}}_i / ||\hat{\boldsymbol{\beta}}_i||_2$는 $\boldsymbol{v}_i$의 추정치가 되며, $\boldsymbol{X}\hat{\boldsymbol{v}}_i$는 i번째 주성분의 추정치가 된다.

2.2. Group SPCA(G-SPCA)

Zou 등 (2006)이 제안한 SPCA는 주성분 분석을 회귀모형 형태의 문제로 전환하였다는 데에 그 첫 번째 의의가 있다. 다시 말해서, 자료의 형태 및 구조에 따라 ridge 또는 lasso 형태의 벌점함수가 아닌 다른 형태의 벌점함수의 적용이 가능하다는 것이다. 변수들이 그룹화되어 있는 경우, 자료의 그룹구조를 모형 적합에 적용하는 것이 보다 타당하다.

$n \times p$ 자료행렬 $\boldsymbol{X}$에서 p개의 변수들이 g개의 인자(factor)로 그룹화되어 있다면, j번째 인자에 해당하는 변수들의 그룹은 $\boldsymbol{x}_{(j)}^T = (x_{j1}, x_{j2}, \ldots, x_{jp_j})$와 같이 표현되며, 이때 $\sum_{j=1}^g p_j = p$이다. 또한 p개의 변수들을 $(x_1, x_2, \ldots, x_p) = (\boldsymbol{x}_{(1)}^T, \boldsymbol{x}_{(2)}^T, \ldots, \boldsymbol{x}_{(g)}^T)$와 같이 나타내면 자료행렬 $\boldsymbol{X} = [\boldsymbol{X}_{(1)}\boldsymbol{X}_{(2)}\cdots\boldsymbol{X}_{(g)}]$로 표현할 수 있으며, 여기서 $\boldsymbol{X}_{(j)}$는 $\boldsymbol{x}_{(j)}^T$에 해당하는 $n \times p_j$ 행렬이다. Bernard 등 (2012)은 주성분 분석에서 그룹구조를 이용하기 위하여 식 (2.4)의 lasso 형태의 벌점함수 대신 group lasso (Yuan과 Lin, 2006) 형태의 벌점함수를 적용한 G-SPCA를 사용할 것을 제안하였다.

$$\hat{\boldsymbol{\beta}}_i^{\text{Group Lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \left\| \left\| \boldsymbol{z}_i - \sum_{j=1}^{g} \boldsymbol{X}_{(j)} \boldsymbol{\beta}_{(j)} \right\| \right\|_2^2 + \lambda_g \sum_{j=1}^{g} \left\| \left\| \boldsymbol{\beta}_{(j)} \right\| \right\|_2 \right\},$$

여기서 $\boldsymbol{\beta}_{(j)} = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp_j})^T$는 j번째 그룹 또는 인자에 대한 회귀계수 벡터이며, $\boldsymbol{\beta} = (\boldsymbol{\beta}_{(1)}^T, \boldsymbol{\beta}_{(2)}^T, \ldots, \boldsymbol{\beta}_{(g)}^T)^T$이다. SPCA와 마찬가지로 $\hat{\boldsymbol{v}}_i = \hat{\boldsymbol{\beta}}_i / ||\hat{\boldsymbol{\beta}}_i||_2$는 $\boldsymbol{v}_i$의 추정치가 되며, $\boldsymbol{X}\hat{\boldsymbol{v}}_i$는 i번째 주성분의 추정치가 된다. 적합식 (2.2)의 벌점함수 $\lambda_g \sum_{j=1}^{g} ||\boldsymbol{\beta}_{(j)}||_2$는 그룹 내에서는 ridge 형태의 축소추정이, 그룹 간에는 lasso 형태의 축소추정이 이루어지게 한다.

3. 계층적 벌점함수(hierarchical penalty)를 이용한 SPCA(H-SPCA)

G-SPCA의 적합식 (2.2)는 group lasso 형태의 벌점함수를 이용하기 때문에, 변수들의 그룹구조를 모형적합에 활용한다는 장점은 있지만, 그룹 내 개별 변수들에 대해서는 축소추정을 하지 못하는 한계가 있다. 따라서 G-SPCA는 그룹별 변수선택에는 효율적이나 선택된 그룹 내에서는 변수선택이 이루어지지 않는다. 이러한 점을 보완하기 위하여 본 논문에서는 계층적 벌점함수를 이용하여 그룹 간과 그룹 내에서의 변수선택이 동시에 이루어지는 H-SPCA를 제안하고자 한다.

3.1. H-SPCA의 적합식

그룹화된 변수들의 구조를 이용하기 위해 회귀계수 β_{jk}의 재매개화된 형태

$$\beta_{jk} = \gamma_j \theta_{jk}, \quad (j = 1, \ldots, g,\ k = 1, \ldots, p_j) \tag{3.1}$$

를 고려하자. 여기서 $\gamma_j \geq 0$은 그룹식별 모수이며, 이러한 인수분해 형태는 모수 γ_j를 통해 계층의 첫 번째 단계에서 그룹 효과를 제어하고, 모수 θ_{jk}를 통해 계층의 두 번째 단계에서 j번째 그룹에 속한 변수들의 영향력의 차이를 반영하게 한다. $\boldsymbol{\theta}_{(j)}$를 $\boldsymbol{\theta}_{(j)} = (\theta_{j1}, \ldots, \theta_{jp_j})^T$로 나타내면 $\boldsymbol{\beta}_{(j)} = \gamma_j \boldsymbol{\theta}_{(j)}$가 된다.

그룹구조를 가진 모수 각각을 축소추정하기 위하여 주성분 분석의 회귀모형 형태를 고려하자. 회귀모형 형태의 주성분 분석에서의 손실함수 $||\boldsymbol{z}_i - \boldsymbol{X}\boldsymbol{\beta}||_2^2$는 식 (3.1)과 같은 그룹구조의 변수와 모수형태를 이용하면

$$\left\| \left\| \boldsymbol{z}_i - \sum_{j=1}^{g} \boldsymbol{X}_{(j)} \gamma_j \boldsymbol{\theta}_{(j)} \right\| \right\|_2^2 \tag{3.2}$$

와 같이 나타낼 수 있다. 이제 식 (3.2)를 이용하여 양의 λ_γ와 λ_θ값에 대하여 다음 벌점화된 최적화 문제를 고려하자.

$$\min_{\boldsymbol{\gamma}, \boldsymbol{\theta}} \left\{ \left\| \left\| \boldsymbol{z}_i - \sum_{j=1}^{g} \boldsymbol{X}_{(j)} \gamma_j \boldsymbol{\theta}_{(j)} \right\| \right\|_2^2 + \lambda_\gamma ||\boldsymbol{\gamma}||_1 + \lambda_\theta ||\boldsymbol{\theta}||_1 \right\}, \tag{3.3}$$

여기서 $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_g)^T$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}^T, \boldsymbol{\theta}_{(2)}^T, \ldots, \boldsymbol{\theta}_{(g)}^T)^T$이다. $\lambda_\gamma \geq 0$과 $\lambda_\theta \geq 0$은 조율모수로 λ_γ는 그룹단계에서의 추정을 조율하며 유의하지 않은 그룹을 제거하는 역할을 한다. 즉, γ_j가 0으로 축소되면, j번째 그룹에 속한 모든 β_{jk}는 0으로 추정되며 따라서 j번째 그룹 변수들은 최종 모형에서 제외된다. λ_θ는 그룹 내 변수선택 단계에서의 추정을 조율하는 역할을 한다. 즉, γ_j가 0으로 추정되지 않았다고 하더라도, 어떤 θ_{jk}가 0으로 축소될 경우, 그에 따른 β_{jk}가 0으로 추정되는 것이다.

식 (3.3)에서 계산된 국소최저점을 $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}})$라고 하면, $j = 1, \ldots, g$, $k = 1, \ldots, p_j$에 관해 $\hat{\beta}_{jk} = \hat{\gamma}_j \hat{\theta}_{jk}$로 나타낼 수 있다. 따라서 추정된 회귀계수 $\hat{\boldsymbol{\beta}}_i$를 표준화한 $\hat{\boldsymbol{v}}_i = \hat{\boldsymbol{\beta}}_i / ||\hat{\boldsymbol{\beta}}_i||_2$는 $\boldsymbol{v}_i$, 즉 i번째 주성분에 대응하는 적재가 된다. 여기서 $\boldsymbol{X}\hat{\boldsymbol{v}}_i$는 i번째 주성분의 추정치가 된다.

3.2. H-SPCA의 계산 알고리즘

H-SPCA를 이용하여 성긴 적재를 구하기 위해서는 먼저 전통적인 주성분 분석을 실시한다. 그리고 식 (3.3)을 이용하여 적절한 성긴 근사치를 찾아낸다. 구체적인 알고리즘은 다음과 같다.

Step 1. A를 처음 k개의 주성분의 적재로 이루어진 행렬로 정의한다. 즉, $A = V[\,, 1:k]$이다.

Step 2. $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k]$를 고정하고, 각각의 $i = 1, \ldots, k$에 대하여 다음의 최적화문제

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\gamma}, \boldsymbol{\theta}} \left\{ \left\| \boldsymbol{z}_i - \sum_{j=1}^{g} \boldsymbol{X}_{(j)} \gamma_j \boldsymbol{\theta}_{(j)} \right\|_2^2 + \lambda_\gamma ||\boldsymbol{\gamma}||_1 + \lambda_\theta ||\boldsymbol{\theta}||_1 \right\}. \tag{3.4}$$

Step 3. $\hat{\boldsymbol{B}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_k)$를 고정하고, $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$의 SVD를 계산한다. 그리고 $\boldsymbol{A}$를 $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{V}^T$로 업데이트 한다.

Step 4. A가 수렴할 때까지 Step 2와 Step 3을 반복하고, 수렴하면 각각의 $i = 1, \ldots, k$에 대하여 $\hat{\boldsymbol{v}}_i = \hat{\boldsymbol{\beta}}_i / ||\hat{\boldsymbol{\beta}}_i||_2$로 정의한다.

Step 2에서의 γ_j와 θ_{jk}의 추정을 위하여, 본 논문에서는 다음과 같은 반복적 추정방법을 이용하였다.

Step 2-1. 각각의 γ_j에 대하여 초기값 $\gamma_j^{(0)}$을 설정한다 ($j = 1, \ldots, g$). 예를 들어, 본 연구에서는 $\gamma_j^{(0)} = 1$로 설정하였다.

Step 2-2. $\tilde{\boldsymbol{X}}_{(j)} = \gamma_j \boldsymbol{X}_{(j)}$ ($j = 1, \ldots, g$)로 정의하고 $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}^T, \ldots, \boldsymbol{\theta}_{(g)}^T)^T$를 다음과 같이 추정한다.

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ \left\| \boldsymbol{z}_i - \sum_{j=1}^{g} \tilde{\boldsymbol{X}}_{(j)} \boldsymbol{\theta}_{(j)} \right\|_2^2 + \lambda_\theta ||\boldsymbol{\theta}||_1 \right\}. \tag{3.5}$$

Step 2-3. $\widehat{\boldsymbol{X}}_{(j)} = \boldsymbol{X}_{(j)} \boldsymbol{\theta}_{(j)}$ ($j = 1, \ldots, g$)로 정의하고, $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_g)^T$를 다음과 같이 추정한다.

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \left\{ \left\| \boldsymbol{z}_i - \sum_{j=1}^{g} \widehat{\boldsymbol{X}}_{(j)} \gamma_j \right\|_2^2 + \lambda_\gamma ||\boldsymbol{\gamma}||_1 \right\}. \tag{3.6}$$

Step 2-4. $\boldsymbol{\gamma}$와 $\boldsymbol{\theta}$가 수렴할 때까지 Step 2-2와 Step 2-3을 반복한다. 수렴했으면 $\hat{\beta}_{jk} = \hat{\gamma}_j \hat{\theta}_{jk}$를 최종해로 설정한다.

Step 2의 반복추정의 각 단계를 살펴보면 우선 γ_j를 고정시킨 상태에서 θ_{jk}를 추정하고, 추정된 θ_{jk}를 고정시킨 상태에서 γ_j를 추정하였으며, γ_j와 θ_{jk}가 수렴할 때까지 이를 반복하였다. 이러한 반복추정의 과정에서 식 (3.4)에서의 목적함수의 값은 비증가(nonincreasing)하므로 위의 알고리즘은 언제나 수렴한다. Step 2-2에서의 식 (3.5)에 관한 최적화문제는 lasso 형태, Step 2-3에서의 식 (3.6)에 관한 최적화 문제는 nonnegative garrote 형태의 문제로, 여유 변수(slack variable)들의 도입을 통하여 이차계획법(quadratic programming)의 문제로 재구성할 수 있다. 이러한 이차계획법 문제의 해를 찾기 위하여 R에서 제공하는 quadprog 패키지를 사용하였다.

Table 4.1. Blue crab data: 25 trace elements

Ag	Silver	Co	Cobalt	Li	Lithium	Ni	Nickel	Sn	Tin
Al	Aluminum	Cr	Chrome	Mg	Magnesium	P	Phosphor	Ti	Titanium
As	Arsenicum	Cu	Copper	Mn	Manganese	Pb	Lead	U	Uranium
Ca	Calcium	Fe	Iron	Mo	Molybdene	Se	Selenium	V	Vanadium
Cd	Cadmium	K	Potassium	Na	Natrium/salt	Si	Silicium/Silica	Zn	Zinc

Step 2의 최적화 문제는 Step 2-1의 초기값 $\gamma_j^{(0)}$의 선정에 따라 최종해가 달라질 수 있다. 예를 들어 $\gamma_j^{(0)}$값이 큰 경우, 그에 따른 $\boldsymbol{\theta}_{(j)} = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{jp_j})$의 값들은 작게 추정되므로, 식 (3.5)의 lasso 벌점함수에서 작은 벌점을 갖게 되어 축소추정이 잘 이루어지지 않는다. 이와 반대로 $\gamma_j^{(0)}$의 값이 작은 경우에는 $\boldsymbol{\theta}_{(j)} = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{jp_j})$의 값들은 크게 추정되어 큰 벌점을 갖게 된다. 본 연구에서의 $\gamma_j^{(0)} = 1$ $(j = 1, \ldots, g)$의 설정은 첫 반복추정의 Step 2-2에서의 적합식을 일반적인 lasso 형태의 문제로 부터 시작하는 것과 동일하다.

한편, 본 연구에서 사용한 반복적 추정 과정에 있어서 재매개화한 모수 $\beta_{jk} = \gamma_j \theta_{jk}$는 크기에 대한 교환, 즉 $\|\beta_{jk}\| = \gamma_j \|\theta_{jk}\|$를 만족하게 되어 반복 추정시 무한 루프에 빠지는 교착상태에 이를 가능성이 있다. 이런 문제를 예방하기 위해 매 교차추정시마다 $\|\theta_{jk}\| = 1$로 제약하는 방법을 생각할 수 있다. 그러나 이는 이차 제약식을 갖는 이차계획법(quadratically constrained quadratic program)의 문제로 귀결되며 많은 계산 비용을 초래한다. 또한 γ_j는 변수그룹의 포함여부를 결정짓는 모수로서 0 또는 1의 값을 갖는 것이 적절하므로, γ_j가 0 또는 1만 갖도록 제약을 가하는 방법을 고려할 수 있다. 그러나 이 방법은 0–1 정수계획법(integer program) 문제에 해당하므로 많은 계산 비용을 요구한다. 이에 대한 대안으로 본 연구의 실제 자료 분석에서는 $\gamma_j \leq 1$ $(j = 1, \ldots, g)$의 제약을 추가로 설정하였으며, 대부분의 γ_j의 값이 0 또는 1의 값을 갖는 것을 확인하였다.

4. 실제자료 분석

4.1. 바다게 자료

바다게(blue crab)는 미국 노스 캐롤라이나(North Carolina) 근안에서 서식하는 식용 꽃게로 매우 상업적 가치가 높은 종이다. 그러나 1986년 팜리코 강(Pamlico River) 지역에서 잡힌 바다게들은 매우 심각한 질병을 갖고 있었으며 학계에 굉장한 관심을 불러일으켰다. Gemperline 등 (1992)는 이 문제에 관하여 환경적 스트레스가 바다게를 약화시켜 정상적인 면역 반응이 키틴인산염(chitnoclastic) 박테리아에 의한 기회감염을 막지 못했다고 주장하였다. 이러한 부류의 박테리아는 5–25 mm의 병변과 함께 바다 껍질을 침투한 것으로 간주된다. Gemperline과 그의 동료들은 어떤 미량 원소들이 이 질병의 발생과 관련이 있는지를 조사하기 위하여 바다게의 아가미, 간췌장, 그리고 근육으로부터 세포를 채집하였다. 조사 대상은 노스 캐롤라이나의 알베말만(Albermarle Sound) 지역의 바다게, 그리고 팜리코 강의 병든 바다게와 건강한 바다게 각각 16마리로, 총 48마리이다. 각 바다게로부터 채집된 미량원소(trace element)는 25가지이며, 따라서 48개의 세포조직 표본이 3개의 세포조직 형태에 따라 25가지의 미량원소 종류에 대한 분석이 이루어졌다. 분석한 25가지의 미량원소는 Table 4.1과 같으며, 원자료(raw data) 및 자료에 대한 설명은 http://www.leidenuniv.nl/fsw/three-mode/data/bluecrabsinfo.htm에서 확인할 수 있다.

제안방법인 H-SPCA와의 비교를 위해 전통적인 주성분 분석과, SPCA, 그리고 G-SPCA를 이용하여 각각 자료를 분석하였다. 각 변수들의 분산이 다르므로 분석에는 자료의 상관행렬이 이용되었다 (Zou

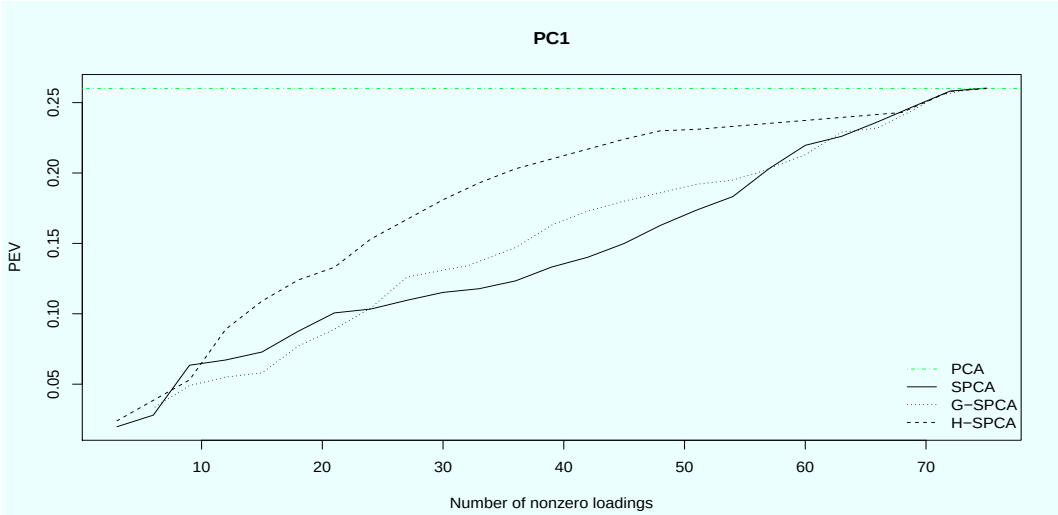

Figure 4.1. Blue crab data: plot of percentage of explained variance (PEV) as a function of number of non-zero loadings for the first PC (PCA = principal component analysis; SPCA = Sparse PCA; G-SPCA = Group SPCA; H-SPCA = hierarchically penalized SPCA).

등, 2006). Figure 4.1은 각 방법에서의 첫 번째 주성분의 0이 아닌 적재의 수에 따른 주성분 분석의 설명력(percentage of explained variance; PEV)의 변화를 보여준다. Figure 4.1에서 볼 수 있는 바와 같이 설명력과 성김의 정도에는 트레이드-오프(trade-off) 관계가 필연적으로 발생한다. 본 자료에서의 조율모수 선택은 Figure 4.1에 따라 전통적인 주성분의 설명력에 비해 너무 많이 떨어지지 않는 설명력을 가지면서 최대로 성긴 적재를 갖게 하도록 이루어졌으며, 이는 Shen과 Huang (2008)과 Zou 등 (2006)이 제안한 방법과 유사하다. 0이 아닌 적재의 수를 특정한 값으로 하기 위한 조율모수의 선택에는 격자검색 방법이 활용되었다. Figure 4.1을 통해 제안 방법인 H-SPCA가 SPCA 및 G-SPCA에 비해 동일한 조건, 즉 0이 아닌 적재의 수에 대해서 높은 설명력을 가짐을 알 수 있으며, 같은 설명력을 갖게 하기 위해서 더 적은 수의 0이 아닌 적재만을 필요로 함을 확인 할 수 있다. Figure 4.1로부터 각 방법에서의 첫 번째 주성분이 약 22% 정도의 분산을 설명할 수 있도록 조율모수를 선택하였으며, 첫 번째 주성분이 선택된 이후 두 번째 주성분의 0이 아닌 적재의 수에 따른 주성분 분석의 설명력의 변화는 Figure 4.2에 나타나 있다. Figure 4.2에서 볼 수 있듯이, 두 번째 주성분도 첫 번째 주성분과 마찬가지로 제안 방법인 H-SPCA가 기존 방법인 SPCA와 G-SPCA보다 같은 수의 적재에서 더 많은 설명력을 가지며, 같은 설명력을 갖는 경우 보다 성긴 적재만을 필요로 함을 확인할 수 있다.

4.2. 세계 경제 위기 자료

2008년 말 미국의 초대형 모기지론 대출회사들의 연이은 파산으로부터 세계 경제 위기가 발생한 이후 전 세계 경제는 현재까지도 어려움을 겪고 있다. 이에 Rose와 Spiegel (2011)은 107개 국가들의 여러 경제 지표를 나타내는 119개 변수들의 횡단면 분석(cross-section analysis)을 통해 세계 경제 위기의 원인과 그 징후를 모형화 하고자 하였다. 주성분 분석을 위하여 본 연구에서는 결측치가 있는 국가와 이산형 변수들을 제외한 72개 국가의 37개의 변수들을 이용하였다. 분석에 포함된 37개의 변수들은 이론적 배경으로부터 10개의 그룹으로 분류될 수 있으며 이는 Table 4.2와 같다. 원자료 및 자료에 대한 설명은 http://faculty.haas.berkeley.edu/arose에서 확인할 수 있다.

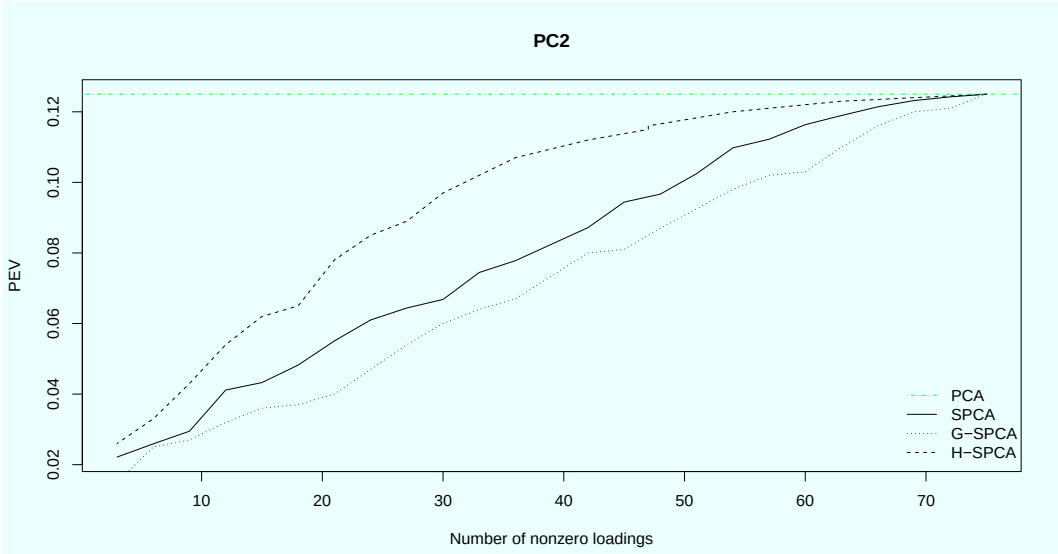

Figure 4.2. Blue crab data: plot of percentage of explained variance (PEV) as a function of number of non-zero loadings for the second PC (PCA = principal component analysis; SPCA = Sparse PCA; G-SPCA = Group SPCA; H-SPCA = hierarchically penalized SPCA).

Table 4.2. Crisis data: classification of groups

Group	Abbreviation	Number of variables
Principal factors	PF	3
Financial policies	FP	3
Financial conditions	FC	4
Asset price appreciation	APA	2
Macroeconomic policies	MP	1
Institutions	INS	10
Geography	GEO	1
Financial linkage	FL	1
Exports linkage	EL	6
Trade linkage	TL	6

Figure 4.3에는 각 방법에 대한 첫번째 주성분의 0이 아닌 적재의 수에 따른 주성분의 설명력이 나타나 있다. Figure 4.3에서 볼 수 있는 바와 같이 제안 방법인 H-SPCA가 SPCA 및 G-SPCA에 비해 더 적은 수의 0이 아닌 적재를 갖고도 높은 설명력을 가짐을 확인할 수 있다. 본 자료의 조율모수 역시 바다게 자료에서와 마찬가지로 Figure 4.3에 따라 전통적인 주성분의 설명력에 비해 너무 많이 떨어지지 않는 설명력을 가지면서 최대로 성긴 적재를 갖도록 선택하였다. Table 4.3은 바다게 자료에서와 같은 방법으로 두 번째 조율모수를 선택한 뒤 각 방법에 대한 처음 두 개의 주성분 적재와 그에 대한 분산을 정리한 결과를 보여준다. SPCA의 경우 그룹 정보를 모형적합에 반영한 G-SPCA와 제안방법인 H-SPCA보다 많은 수의 0이 아닌 적재를 갖고도 낮은 설명력을 보였다. 또한 G-SPCA와 H-SPCA는 Table 4.3과 같이 0이 아닌 적재를 선택하였을 때 두 주성분이 각각 서로 다른 그룹의 변수들만 사용하였으며, H-SPCA가 G-SPCA에 비해 더 적은 수의 그룹을 선택하고, 더 성긴 적재를 가지면서도 더 높은 설명력을 가짐을 확인할 수 있었다.

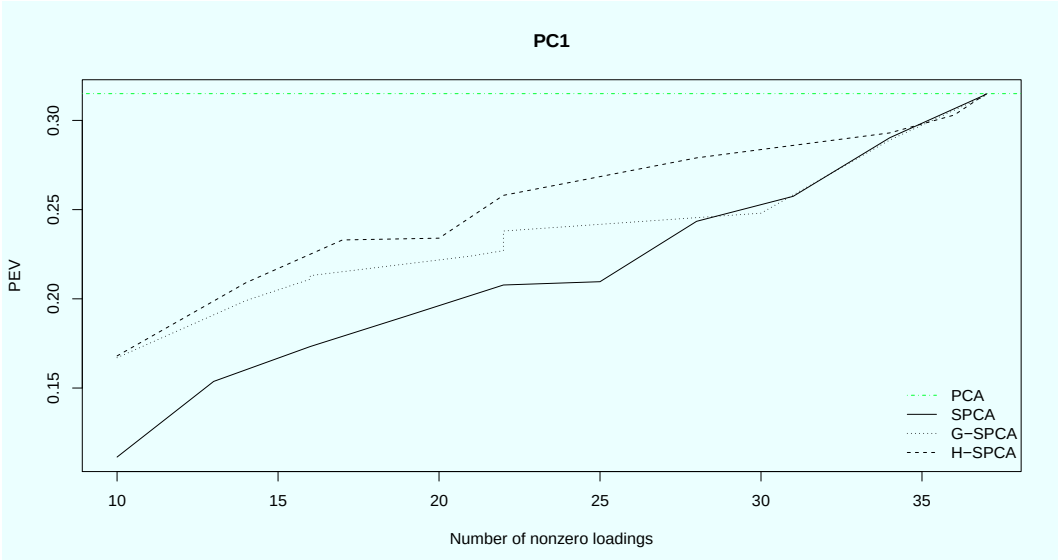

Figure 4.3. Crisis data: Plot of percentage of explained variance (PEV) as a function of number of non-zero loadings for the first PC (PCA = principal component analysis; SPCA = Sparse PCA; G-SPCA = Group SPCA; H-SPCA = hierarchically penalized SPCA).

Table 4.3. Crisis data: loadings of the first two PCs by PCA, SPCA, G-SPCA and H-SPCA

Group	PCA		SPCA		G-SPCA		H-SPCA	
	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
	−0.111	0.068			0.071		0.116	
PF	0.245	0.011	0.183		−0.144		−0.255	
	−0.046	−0.177	−0.062	−0.044	0.048		0.089	
FP	0.178	0.037	0.246		−0.131		−0.210	
⋮				⋮				
	−0.049	0.274		0.357		0.405		0.351
	0.157	−0.018	0.152			−0.088		−0.079
TL	0.152	−0.314	0.046	−0.253		−0.502		−0.385
	−0.029	0.356		0.589		0.499		0.370
	0.029	−0.124		−0.095		−0.195		−0.197
	0.039	−0.237		−0.028		−0.294		−0.273
Selected groups	10	10	8	6	7	3	5	2
Non-zero loadings	37	37	28	15	24	13	22	12
Variance(%)	31.5	11.4	23.5	9.1	26.7	8.4	27.0	9.5
Cummulative variance(%)	31.5	42.9	23.5	32.6	26.7	35.1	27.0	36.5

PCA = principal component analysis; SPCA = sparse PCA; G-SPCA = group SPCA; H-SPCA = hierarchically penalized SPCA; PF = principal factors; FP = financial policies; TL = trade linkage. Empty cells have zero loadings.

5. 결론 및 향후 연구방향

본 연구는 변수들이 그룹화되어 있는 다변량 자료의 주성분 분석에서의 변수 선택 및 추정 방법으로 계층적 축소추정법을 제안하였다. 제안 방법은 그룹 변수가 있는 주성분 분석에서의 기존의 방법인 G-

SPCA와 마찬가지로 그룹 효과를 모형 적합에 반영하였다. 기존의 방법은 그룹 내 변수선택을 하지 못하는 한계가 있었으나, 본 제안 방법은 선택 된 그룹 내에서도 개별적인 변수들을 선택하는 장점을 갖고 있다. 또한 개개의 변수에 관하여 성긴 적재를 생성하는 기존의 방법인 SPCA에 비해서도 그룹 효과를 모형 적합에 반영하여 더 높은 설명력을 가지면서도 비슷한 수준의 성긴 적재를 생성하는 강점이 있다. 이러한 제안 방법의 강점은 실제 자료인 바다게 자료 및 세계 경제 위기 자료의 분석을 통해 확인할 수 있었으며, 실제 자료에의 적용을 통해 변수들이 그룹화되어 있는 다양한 실제 자료의 분석에 이용할 수 있음을 보였다.

기존의 연구 (Kang 등, 2016; Wang 등, 2009)에 따르면 계층적 벌점함수를 적용한 회귀 문제에서는 두 개의 조율모수 λ_γ, λ_θ를 모두 사용하는 대신 $\lambda = \lambda_\gamma \lambda_\theta$를 만족하는 하나의 조율모수만을 사용하여도 동일한 결과를 얻을 수 있다. 그러나 계층적 벌점함수를 적용한 주성분 분석의 경우, 하나의 조율모수만을 사용하게 되면 각 반복에서의 수렴속도가 현저하게 낮아지는 단점이 실험을 통해 발견되었다. 본 연구에서는 두 개의 조율모수 λ_γ와 λ_θ를 사용하였다. 그러나 두 개의 조율모수의 사용은 높은 계산 비용을 요구하는 단점이 있다. 따라서 하나의 조율모수를 사용하면서도 빠른 수렴속도를 가지는 알고리즘의 개발이 필요할 것이다.이에 따라 0이 아닌 적재 수의 특정값을 갖게 하기 위한 조율모수를 찾는 효과적인 알고리즘의 개발도 가능할 것이다. 최근 벌점화 회귀분석 이론에서 널리 사용되고 있는 적응적 조율모수의 적용도 고려해볼 필요가 있다. 예비 실험에 따르면 적응적 조율모수를 사용한 주성분 분석 방법은 더 성긴 적재를 생성하는 장점은 있으나, 변수의 선택과 설명력의 트레이드-오프 관계로 인하여 설명력이 현저하게 떨어지는 한계를 동시에 보였다. 따라서 적응적 조율모수를 사용하여 더 성긴 적재를 생성하면서도 설명력의 희생이 크지 않은 방법론의 개발 역시 본 제안 방법 및 기존의 연구들을 획기적으로 발전시킬 수 있는 연구가 될 것이다.

References

Bernard, A., Guinot, C., and Saporta, G. (2012). Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis. In *Proceedings of 20th International Conference on Computational Statistics* (pp. 99–106).

Gemperline, P. J., Miller, K. H., West, T. L., Weinstein, J. E., Hamilton, J. C., and Bray, J. T. (1992). Principal component analysis, trace elements, and blue crab shell disease, *Analytical Chemistry*, **64**, 523–531.

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

Kang, J., Bang, S., and Jhun, M. (2016). Hierarchically penalized quantile regression, *Journal of Statistical Computation and Simulation*, **86**, 340–356.

Rose, A. K. and Spiegel, M. M. (2011). Cross-country causes and consequences of the crisis: an update, *European Economic Review*, **55**, 309–324.

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 Series B (Methodological)*, **58**, 267–288.

Wang, S., Nan, B., Zhou, N., and Zhu, J. (2009). Hierarchically penalized Cox regression with grouped variables, *Biometrika*, **96**, 307–322.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society Series B (Methodological)*, **68**, 49–67.

Zou, H. and Hastie, T. (2003). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B (Methodological)*, **67**, 301–320.

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

계층적 벌점함수를 이용한 주성분분석

강종경[a] · 박재신[a] · 방성완[a,1]

[a]육군사관학교 수학과

(2016년 11월 16일 접수, 2017년 1월 12일 수정, 2017년 1월 25일 채택)

요 약

주성분 분석(principal component analysis; PCA)은 서로 상관되어 있는 다변량 자료의 차원을 축소하는 대표적인 기법으로 많은 다변량 분석에서 활용되고 있다. 하지만 주성분은 모든 변수들의 선형결합으로 이루어지므로, 그 결과의 해석이 어렵다는 한계가 있다. sparse PCA(SPCA) 방법은 elastic net 형태의 벌점함수를 이용하여 보다 성긴(sparse) 적재를 가진 수정된 주성분을 만들어주지만, 변수들의 그룹구조를 이용하지 못한다는 한계가 있다. 이에 본 연구에서는 기존 SPCA를 개선하여, 자료가 그룹화되어 있는 경우에 유의한 그룹을 선택함과 동시에 그룹 내 불필요한 변수를 제거할 수 있는 새로운 주성분 분석 방법을 제시하고자 한다. 그룹과 그룹 내 변수 구조를 모형 적합에 이용하기 위하여, sparse 주성분 분석에서의 elastic net 벌점함수 대신에 계층적 벌점함수 형태를 고려하였다. 또한 실제 자료의 분석을 통해 제안 방법의 성능 및 유용성을 입증하였다.

주요용어: 주성분 분석, sparse PCA, 계층적 벌점함수, 그룹변수

본 논문은 육군사관학교 화랑대연구소의 2015년도(20150501) 연구활동지원에 의해 출간되었으며(강종경), 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2015R1C1A1A02036473)(방성완).

[1]교신저자: (01805) 서울시 노원구 화랑로 574, 육군사관학교 수학과. E-mail: wan1365@gmail.com