로버스트주성분회귀에서 최적의 주성분선정을 위한 기준

김부용^{1,a}

^a숙명여자대학교 통계학과

요약

회귀모형에 연관성이 높은 설명변수들이 포함되면 다중공선성의 문제가 야기되며, 동시에 자료에 회귀 이상점들이 포함되면 최소자승추정량에 바탕을 둔 제반 통계적 추론은 심각한 결함을 갖게 된다. 이러한 현상들은 데이터마이닝 분야에서 많이 볼 수 있는데, 본 논문에서는 두 가지 문제를 동시에 해결하기 위한 방안으로서 로버스트주성분회귀를 제안하였다. 특히 최적의 주성분을 선정하기 위한 새로운 기준을 개발하였는데, 설명변수들의 표본공분산 대신에 MVE-추정량을 기반으로 하였으며, 고유치가 아니라 상태지수의 크기에 바탕을 둔 선정기준을 제안하였다. 그리고 주성분모형에서의 추정을 위하여 회귀이상점에 대해 로버스트한 LTS-추정을 도입하였다. 제안된 선정기준이 기존의 기준들보다 다중공선성과 이상점이 유발하는 문제들을 잘 해결할 수 있음을 모의실험을 통하여 확인하였다.

주요용어: 다중공선성, 회귀이상점, 로버스트주성분회귀, MVE-추정량, 상태지수, LTS-추정.

1. 서론

데이터마이닝 분야에서 변수간의 관계를 규명하거나 변수 값을 예측하기 위하여 회귀분석이 많이 활용되고 있다. 그런데 이 분야에서의 자료들은 그 규모가 방대하여 설명변수로 채택될 수 있는 후보 변수들을 많이 가지고 있을 뿐만 아니라, 기본적으로 상호연관성이 높은 변수들을 다수 포함하고 있다는 특징을 가지고 있다. 그러므로 회귀분석의 출발점에서 적절한 변수선정 기준 및 방법에 의해 설명변수가 선정되었더라도 모형에 도입된 설명변수들 사이에 높은 연관성이 존재하는 경우가 있다. 이러한 현상을 다중공선성이라 하는데, 회귀분석에서 주로 사용되는 최소자승추정량의 분산이 지나치게 팽창되는 심각한 문제를 야기한다. 그러므로 이 추정량에 바탕을 둔 제반 통계적 추론들은 신뢰성을 보장받을 수 없는 것이다. 이러한 다중공선성의 문제를 해결하기 위하여 사용되는 기법 중의 한 가지가 주성분회귀(PCR; principal component regression)이다.

한편, 데이터마이닝 분야에서는 사전에 자료수집 계획이 수립되지 않는 경우가 있고, 수집계획이 있더라도 그에 따른 철저한 통제 하에 자료가 수집되지 않는 상황도 있다. 더욱이 다양한 특성을 갖는 인간을 대상으로 한 자료가 주로 수집되기 때문에 나쁜 지렛점(bad leverage point), 좋은 지렛점(good leverage point), 그리고 수직이상점(vertical outlier)과 같은 회귀이상점이 자료에 다수 포함될 수 있다. 이와 같이 자료에 다중공선성과 회귀이상점이 동시에 존재하는 경우에 두 가지 문제점을 분리하여 대응하는 해결방법은 완전할 수가 없다. 왜냐하면 지렛점은 다중공선성의 심각성을 더 악화시킬 수도 있고 반대로 완화시킬 수도 있기 때문이다. 따라서 다중공선성과 회귀이상점의 문제를 동시에 해결할 수 있는 방안으로 로버스트주성분회귀분석(RPCR; robust principal component regression)을 고려할 수 있다.

본 연구는 숙명여자대학교 2010년도 교비연구비 지원에 의해 수행되었음.

¹ (140-742) 서울특별시 용산구 청파동, 숙명여자대학교 통계학과, 교수. E-mail: buykim@sookmyung.ac.kr

RPCR의 실행과정에서 가장 핵심적인 내용은 최적의 주성분을 선정하는 것인데, 이를 위해 적절한 선정기준을 설정하는 것이 매우 중요한 과제다. 그런데 RPCR을 위한 선정기준에 관한 연구는 활발하게 이루어지지 않았기 때문에 PCR이나 주성분분석에서 적용하는 기존의 선정기준을 원용할 수 있다. 그러나 기존의 선정기준들은 지렛점에 대해 매우 취약할 수밖에 없는 표본공분산의 고유치에 바탕을 둔 것이기 때문에, 기존의 방법들에 의해서는 주성분의 선정이 적절하게 이루어지기 어렵다. 따라서 표본공분산 대신에 지렛점에 대해 매우 로버스트한 추정량인 MCD(minimum covariance determinant)-추정량이나 MVE(minimum volume ellipsoid)-추정량을 기반으로 하는 선정기준을 제안하되, 자료의 존성이 큰 고유치 대신에 상태지수(condition index)에 바탕을 둔 주성분 선정기준을 개발하고자 한다.

제안된 기준에 의해 최적의 주성분이 선정되더라도 주성분모형에서의 최소자승추정량은 회귀이상점의 영향에 의해 심각하게 왜곡될 수밖에 없다. 따라서 로버스트 회귀추정인 LMS(least median of squares)-추정이나 LTS(least trimmed squares)-추정을 주성분모형에서 회귀계수추정에 적용하려 한다.

2. 회귀이상점과 다중공선성

회귀이상점은 최소자승추정에 막대한 영향을 미치기 때문에 회귀분석을 실행하기에 앞서 지렛점이나 수직이상점의 존재 여부를 확인하는 과정을 거쳐야 한다. 회귀이상점들을 식별하기 위해서는 Rousseeuw와 Zomeren (1990), Hadi와 Simonoff (1993), McKean 등 (1993), Rocke와 Woodruff (1997), Kim과 Kim (2002), Kim과 Oh (2004) 등의 식별방법을 활용할 수 있다. 특히, 회귀이상점 중에서 지렛점은 회귀추정량뿐만 아니라 다중공선성의 심각성에도 직접적인 영향을 미친다. 즉 좋은 지렛점들은 다중공선성의 정도를 더 강화시키고, 나쁜 지렛점들은 다중공선성의 정도를 상당히 완화시키는 작용을 한다. Mason과 Gunst (1985)는 지렛점들에 의해 강화된 다중공선성이 회귀계수 추정치에 어떤 영향을 미치는지 연구하였다.

본 연구에서는 지렛점들이 다중공선성의 정도에 얼마나 영향을 미치는지 실증적으로 분석하기 위하여 모의실험을 실행하였다. 회귀모형에 포함된 설명변수들 사이에 다중공선성이 어느 정도 심각한지를 판단하기 위해서는 공분산행렬의 고유치 혹은 상태지수의 크기를 측정해야 하는데, 본 연구에서는 고유치보다 상대적인 비교에 적합한 상태지수를 측정하여 다중공선성의 정도를 판단하였다. 김부용과 신명희 (2010)의 방법에 의해 다양한 설명변수의 개수(2~9개)와 다중공선성에 관련된 변수의 수(2~5개)별로 관찰치(100~450개)를 생성하였으며, 각 자료에서 일정한 비율(5%, 10%)의 관찰치를 좋은 지렛점 또는 나쁜 지렛점으로 교체하였다. 좋은 지렛점은 다중공선성에 관련된 설명변수 모두에 대해 생성된 각 자료의 범위만큼을 증가시켜 얻었으며, 나쁜 지렛점은 다중공선성에 관련된 설명변수 중의 일부에 대해 생성된 각 자료의 범위만큼을 증가시켜 얻었다. 각 프로파일별로 1,000개의 자료를 임의의 생성하고 각 자료별로 표본공분산행렬의 상태지수를 계산한 후 평균을 구하였다. 모의실험 결과로부터 다음과 같은 사실을 확인할 수 있었다. 즉, 나쁜 지렛점들이 포함된 자료에서는 정상점만 존재하는 자료에서보다 상태지수의 크기가 상당히 작아지는 현상을 볼 수 있었으며, 반면에 좋은 지렛점들이 포함된 자료에서는 상태지수의 크기가 상당히 커졌는데, 특히 강력한 수준의 좋은 지렛점들이 포함된 자료에서는 상태지수가 매우 커지는 현상을 볼 수 있었다. 이와 같이 지렛점의 종류와 강도에 따라 표본공분산 행렬의 상태지수는 민감하게 영향을 받고, 그에 바탕을 둔 주성분 선정과정도 큰 영향을 받게 된다. 따라서 다중공선성과 지렛점의 문제를 동시에 극복할 수 있는 방법이 필요한데, 지렛점에 대해서 매우 로버스트한 추정량인 MCD-추정량이나 MVE-추정량에 바탕을 둔 RPCR을 채택할 수 있다.

3. 로버스트주성분회귀

일반적인 선형회귀분석은 설명변수들 사이에 상호연관성이 존재하지 않는다는 가정에 기초를 둔

다. 이 가정이 충족되지 않는 상황, 즉 설명변수들 간에 상당히 높은 수준의 다중공선성이 존재하는 경우에는 최소자승법에 의한 회귀계수 추정량의 분산은 매우 크게 팽창되므로 추정치가 불안정하고 심지어 추정치가 반대의 부호로 얻어지는 문제를 야기하기도 한다. 그러므로 분산분해비율과 같은 다중공선성 진단척도를 적용하여 자료에 다중공선성이 존재한다는 사실이 확인되면, 이 문제를 극복할 수 있는 적절한 방안을 강구해야 한다. PCR이 그 중의 한 가지 방법인데 간단히 요약하면 다음과 같다.

3.1. 주성분회귀

PCR은 주성분분석과 회귀분석을 결합한 기법인데, 회귀추정량에 어느 정도의 편의를 허용하는 대신에 추정량의 분산이 지나치게 팽창되는 것을 예방하려는 의도에서 사용된다. 선형회귀모형

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (3.1)$$

(단, $\mathbf{y}$ 는 반응변수 n -벡터, $\mathbf{X}$ 는 설명변수 $n \times p$ -행렬, $\boldsymbol{\beta}$ 는 회귀계수 p -벡터, $\boldsymbol{\epsilon}$ 는 $N(0, \sigma^2)$ 을 따르는 오차 n -벡터)를 바탕으로 PCR을 실행하려면, 우선 $\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$ (단, $\tilde{\mathbf{X}}$ 는 $\mathbf{X}$ 를 중심화 한 행렬)의 각 고유치($\lambda_j : j = 1, \dots, k$)에 대응하는 고유벡터($\mathbf{v}_1, \dots, \mathbf{v}_k$)로 구성된 직교행렬 $\mathbf{V}_{k \times k}$ 을 구하는 것에서 시작된다. 다음은 행렬 $\tilde{\mathbf{X}}$ 를 $\mathbf{V}$ 에 의해 회전시킨 행렬 $\mathbf{Z} = \tilde{\mathbf{X}}\mathbf{V}$ 을 얻을 수 있는데, $\mathbf{Z}$ 를 주성분행렬이라 하고 $\mathbf{Z}$ 의 각 열($\mathbf{z}_1, \dots, \mathbf{z}_k$)을 주성분이라 한다. 그리고 주성분들에 의해 표현된 회귀모형

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad \boldsymbol{\gamma} = \mathbf{V}^T \boldsymbol{\beta} \quad (3.2)$$

을 주성분모형이라 한다. $\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$ 의 고유치들을 원소로 갖는 대각행렬을 Λ 라 하면 모형 (3.2)에서의 최소 제곱추정량 $\hat{\boldsymbol{\gamma}}$ 의 분산은 $\sigma^2 \Lambda^{-1}$ 이므로, 매우 작은 고유치들은 모형 (3.2)에서는 물론 모형 (3.1)에서도 최소자승추정량의 분산을 크게 팽창시키게 된다. 그러므로 매우 작은 고유치에 대응하는 주성분들은 모형에서 제외시키고 나머지 주성분만으로 구성된 모형을 채택하는 것이다. 적절한 기준에 의해 선정된 일부 주성분만 포함된 모형에서 회귀계수 추정치를 구한 후, 직교행렬 $\mathbf{V}$ 에 의해 모형 (3.1)에서의 추정치로 변환시키고 절편을 조정함으로써 다중공선성이 야기하는 문제를 해결할 수 있게 된다.

3.2. 로버스트주성분회귀

PCR의 출발점에서 설명변수의 표본공분산 행렬로부터 고유벡터와 고유치를 계산하는데, 표본공분산 행렬은 지렛점에 대해 매우 민감하기 때문에 이들에 바탕을 둔 주성분 선정결과 역시 지렛점에 의해 많은 영향을 받는다. 이러한 문제를 극복하기 위하여 설명변수 공분산의 로버스트 추정량에 해당하는 MCD-추정량이나 MVE-추정량에 바탕을 둔 RPCR을 고려할 수 있다. Marden (1999)은 주성분분석에 MVE-추정량을 적용할 것을 제안하였는데, 지렛점이 심각하지 않은 상황에서는 MVE-추정량에 바탕을 둔 결과가 순위나 부호에 바탕을 둔 것보다 우수함을 모의실험을 통하여 확인하였다. 그리고 Hubert와 Verboven (2003)은 MCD-추정량에 바탕을 둔 RPCR을 제안하였다.

MCD-추정량과 MVE-추정량에 관한 연구로는 Rocke와 Woodruff (1997)가 있는데, MCD-추정량 ($\hat{\boldsymbol{\mu}}_J, \hat{\boldsymbol{\Sigma}}_J$)은 $h(= \lfloor (n+k+1)/2 \rfloor)$ 는 최대정수함수)개의 관찰치로 구성된 모든 공분산행렬 중에서 그 행렬식이 최소인 것으로부터 구한 평균과 공분산행렬, $\hat{\boldsymbol{\Sigma}}_J = (1/h) \sum_{i \in J} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_J)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_J)^T$, $\hat{\boldsymbol{\mu}}_J = (1/h) \sum_{i \in J} \mathbf{x}_i$ (단, J 는 관찰치의 수가 h 인 모든 집합 K 에 대해 $|\hat{\boldsymbol{\Sigma}}_J| \leq |\hat{\boldsymbol{\Sigma}}_K|$ 을 만족시키는 집합)로 정의된다. 그리고 MCD-추정량의 일치성을 위하여 수정인자를 반영할 수 있는데 Fauconnier와 Haesbroeck (2009)은 수정인자로서 $h/(nP(\chi_{k+2}^2 < \chi_{k,h/n}^2))$ 을 적용할 것을 제안하였다. 한편, MVE-추정량 ($\hat{\boldsymbol{\mu}}_E, \hat{\boldsymbol{\Sigma}}_E$)은 $e = \lfloor n/2 \rfloor + 1$ 개의 관찰치로 구성된 모든 타원체들 중에서 그 부피가 최소인 것(E)의 중심과 공분산행렬, $\hat{\boldsymbol{\Sigma}}_E = (1/e) \sum_{i \in E} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_E)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_E)^T$, $\hat{\boldsymbol{\mu}}_E = (1/e) \sum_{i \in E} \mathbf{x}_i$ 에 일치추정량의 성질을 위한 수정

인자 $(\chi_{k,0.5}^2)^{-1}$ 을 적용한 결과로 정의된다. MCD-추정량과 MVE-추정량의 로버스트성의 척도인 붕괴점(breakdown point)은 $[(n-k+1)/2]/n$ 이고 점근붕괴점은 0.5이기 때문에 두 추정량은 지렛점에 대해 가장 로버스트한 추정량으로 알려져 있다. MCD-추정치와 MVE-추정치를 구하기 위해서는 막대한 계산이 요구되는데 Woodruff와 Rocke (1994)와 Rousseeuw와 Driessen (1999)은 계산효율성이 향상된 알고리즘을 개발하였다.

MCD-추정량 및 MVE-추정량에 바탕을 둔 상태지수가 어떤 특성을 갖는지 확인하기 위하여 제2장에서와 동일한 방법으로 모의실험을 실행하였다. 모의실험 결과로부터, 표본공분산에 바탕을 둔 상태지수의 크기는 지렛점의 종류나 강도에 따라 크게 변동하는 반면에 MCD-추정량 및 MVE-추정량에 바탕을 둔 상태지수는 나쁜 지렛점은 물론 좋은 지렛점의 영향을 거의 받지 않고 일정한 값으로 측정됨을 확인할 수 있었다. 그러므로 지렛점이 포함된 자료에서 표본공분산을 바탕으로 주성분을 선정하는 것보다 MCD-추정량이나 MVE-추정량과 같은 로버스트 추정량에 바탕을 둔 기준에 따라 주성분을 선정하는 것이 타당하다는 사실을 확인하였다. 그런데 Rousseeuw와 Leroy (2003)는 MCD-추정량보다 MVE-추정량이 다소 우수한 추정량이라 하였다. 따라서 본 연구에서는 MVE-추정량에 바탕을 둔 주성분선정 기준을 개발하고자 한다.

우선 MVE-추정량의 고유치($\lambda_j^* : j = 1, \dots, k$)와 상태지수($c_j = \lambda_{\max}^*/\lambda_j^*$)를 구한 후, 각 고유치에 대응하는 고유벡터들($\mathbf{v}_1^*, \dots, \mathbf{v}_k^*$)로 구성된 직교행렬 $V_{k \times k}^*$ 을 얻는다. 그리고 행렬 V^* 에 의해 회전된 주성분행렬 $Z^* = XV^*$ 을 구성하고 Z^* 의 각 열($\mathbf{z}_1^*, \dots, \mathbf{z}_k^*$)인 주성분 중에서 적절한 기준치(3.3절에서 제시할 예정임)를 초과하는 c_j 에 대응하는 s 개의 주성분은 모형에서 제외시키고 $g(=k-s)$ 개의 주성분만으로 구성된 모형

$$\mathbf{y} = Z_p^* \boldsymbol{\gamma}_{PC}^* + \boldsymbol{\epsilon}, \quad Z_p^* = [\mathbf{1}_n \mid Z_{n \times g}^* \mid \mathbf{0}_{n \times s}] \quad (3.3)$$

을 설정한다. 다음은 모형 (3.3)의 회귀계수 $\boldsymbol{\gamma}_{PC}^*$ 을 추정하는 과정이 필요한데, 앞 단계에서 로버스트추정량에 바탕을 둔 주성분 선정이 적절하게 이루어졌더라도 최소자승추정량은 지렛점과 수직이상점의 악영향을 받게 된다. 따라서 최소자승추정 대신 로버스트 회귀추정을 적용해서 $\boldsymbol{\gamma}_{PC}^*$ 을 추정해야 한다. 로버스트 회귀추정으로 Rousseeuw (1984)가 제안한 LMS-추정이나 LTS-추정법이 적용될 수 있다. 모형 (3.3)에서의 LMS-추정량은 $\text{median}_i \epsilon_i^2$ (단, $\epsilon_i = y_i - \mathbf{z}_{pi}^{*T} \boldsymbol{\gamma}_{PC}^*$)을 최소화하는 $\boldsymbol{\gamma}_{PC}^*$ 로 정의되며, 이 추정량의 붕괴점은 $([n/2]-p+2)/n$ 이다. LTS-추정량은 $\sum_{i=1}^u \epsilon_{i:n}^2$ (단, $u = [(n+p+1)/2]$, $\epsilon_{1:n} \leq \dots \leq \epsilon_{n:n}$)을 최소화하는 $\boldsymbol{\gamma}_{PC}^*$ 로 정의되는데, 이 추정량의 붕괴점은 $([(n-p)/2] + 1)/n$ 로 알려져 있다. 그러므로 LMS-추정량과 LTS-추정량의 점근붕괴점은 0.5로서 회귀이상점에 대하여 가장 로버스트한 추정량이라 할 수 있다. 그런데 LTS-추정량은 점근적 정규성을 가지므로 LMS-추정량보다 통계적 효율성이 더 좋다고 알려졌다 (Rousseeuw와 Driessen, 2006). 더욱이 LMS-추정에 비하여 LTS-추정의 목적함수가 더 평활하여 국소효과에 덜 민감하므로 LTS-추정 알고리즘이 LMS-추정 알고리즘보다 수렴성이 좋고 (Rousseeuw와 Leroy, 2003), 계산효율성이 개선된 LTS-추정 알고리즘이 개발되었기 때문에 본 연구에서는 LTS-추정을 채택하였다. 한편 모형 (3.3)에서 구한 LTS-추정치 $\hat{\boldsymbol{\gamma}}_{PC}^*$ 은 다음의 관계식,

$$\hat{\boldsymbol{\beta}}_{PC} = \bar{V} \hat{\boldsymbol{\gamma}}_{PC}^*, \quad \bar{V} = \begin{bmatrix} \mathbf{1} & \mathbf{0}_k^T \\ \mathbf{0}_k & V^* \end{bmatrix}$$

에 의해 모형 (3.1)에서의 RPCR-추정치로 변환된다.

3.3. 로버스트 주성분 선정기준

PCR은 물론 RPCR에서 가장 중요한 과정은 모형에 포함될 최적의 주성분을 선정하는 것이다. 이는 최적의 주성분선정이 이루어져야 설명변수가 갖는 정보의 손실이 최소화될 수 있기 때문이다. 즉

지나치게 작은 개수의 주성분을 선정하면 정보의 손실이 크게 되고, 반면에 너무 많은 개수의 주성분을 선정하면 다중공선성 문제를 해결해야 하는 주성분회귀의 목적을 달성하기 어렵다. 그러므로 최적의 주성분이 선정될 수 있는 기준을 적용하는 것이 매우 중요하다. RPCR에서의 주성분선정기준에 관한 연구는 거의 이루어지지 않고 있는 실정인데, 일단 주성분분석이나 PCR에서 사용되고 있는 기존의 기준들을 원용할 수 있다.

주성분선정을 위한 기존의 방법들은 주로 공분산행렬의 고유치들의 상대적 크기를 기준으로 삼는 것들이다. Pidot (1969)은 $\lambda_j < 1.0$ 에 해당하는 고유치들에 대응하는 주성분을, Jolliffe (1972)는 $\lambda_j < 0.71$ 에 해당하는 고유치들에 대응하는 주성분을 모형에서 제외시키는 방법을 제시하였다. 그리고 Marquardt (1970)는 고유치들의 누적 비율이 일정한 수준을 초과하지 않는 범위($\sum_{j=1}^g \lambda_j / \sum_{j=1}^k \lambda_j < \delta$, 예를 들면, $\delta = 0.9$)내의 고유치들에 해당하는 주성분을 모형에 포함시키는 방법을 제안하였다. Karlis 등 (2003)은 고유치들의 평균 $\sum_{j=1}^k \lambda_j / k$ 보다 큰 고유치에 대응하는 주성분을 선정하는 방법을 개선하여, 고유치의 신뢰구간을 바탕으로 하는 선정기준을 제시하였다. 즉, 고유치 중에서 $\lambda_j > 1 + 2\sqrt{(k-1)/(n-1)}$ 을 충족시키는 고유치에 대응하는 주성분을 모형에 포함시키는 방법이다. 그러나 이러한 기준들을 RPCR에 적용하면 주성분 전체를 모형에서 제외시키게 되는 과도한 차원축소의 문제가 발생할 수 있다. 한편, Legendre와 Legendre (1998)는 broken-stick 모형을 바탕으로 $\lambda_j > (1/k) \sum_{i=j}^k (1/i)$ 을 충족시키는 고유치에 대응하는 주성분을 모형에 포함시키는 방법을 제안하였으나, 이 기준을 적용하면 매우 작은 고유치에 대응하는 주성분은 모형에 포함시키고 오히려 중간 정도 크기의 고유치에 대응하는 주성분을 제외시키는 상황이 발생할 수 있기 때문에 적절한 기준이라 할 수 없다.

한편, Marden (1999)과 Hubert와 Verboven (2003)은 각각 로버스트 추정량을 도입한 주성분분석과 PCR를 제안하였지만, 주성분 선정기준을 구체적으로 제시하지는 않았다. 따라서 본 연구에서는 로버스트 추정량에 바탕을 두되, 고유치를 기준으로 하는 선정방법의 한계를 극복하기 위하여 상태지수를 기준으로 하는 주성분 선정기준을 개발하고자 한다. 상태지수의 최소치는 항상 1.0이므로 상태지수를 선정기준으로 채택하면 모든 주성분이 모형에서 제외되는 상황을 피할 수 있다. RPCR에서의 주된 관심은 설명변수의 분산을 대부분 차지하는 일부의 주성분을 선정하는 것에 있는데, 일부의 주성분만을 선정하면 RPCR-추정량의 분산은 축소되지만 동시에 편의가 증가하는 것을 막을 수 없다. 따라서 RPCR-추정량의 평균제곱오차 $MSE(\hat{\beta}_{PC})$ 을 최소가 되게 하는 선정방법에 초점을 맞추기로 한다. 그런데 $\hat{\beta}_{PC}$ 는 벡터이므로 각 회귀계수에 대한 RPCR-추정량인 $\hat{\beta}_{PCj}$ 의 평균제곱오차를 합제한 총평균제곱오차(TMSE; total mean square error)를 최적화 대상으로 하였다. 김부용과 신명희 (2010)에 의하면 $TMSE(\hat{\beta}_{PC})$ 는 총분산과 총제곱편의의 합으로 나타낼 수 있는데, 총분산은 g 에 대해 단조증가함수이고 총제곱편의는 g 에 대해 단조감소함수이다. 즉, 모형에 포함되는 주성분의 개수인 g 의 크기가 증가함에 따라 RPCR-추정량의 총분산은 증가하고 반면에 총제곱편의는 감소한다. 그러므로 RPCR-추정량의 분산이 축소되는 것과 편의가 증가되는 것을 적절한 수준에서 조정할 수 있는 선정기준이 요구된다.

RPCR에서 주성분 선정을 위한 상태지수의 기준은 Monte Carlo 모의실험 결과를 바탕으로 설정하였다. 모의실험을 위한 설명변수 자료는 김부용과 신명희 (2010)의 방법에 의해서 생성하였으며, 다양한 규모와 특성을 갖는 자료를 생성하기 위하여 설명변수의 수(3 (1) 9), 설명변수의 수에 비례한 관찰치의 수(30 (10) 90배), 다중공선성에 관련된 변수의 수(2 (1) 5)의 모든 조합에 해당하는 자료를 정규분포로부터 생성하였다. 오차 역시 정규분포로부터 생성하였으며, 회귀계수 값으로 $\beta^0 = 1_p$ 을 사전에 지정하였다. 다양한 자료를 생성하기 위하여 매 반복마다 상이한 seed를 지정하였다. 역시 2장과 같은 방법에 의해 각 자료의 일부분을 좋은 지렛점과 나쁜 지렛점으로 교체하였다. 생성된 자료마다 2에서 100까지의 상태지수 경계치를 순차적으로 적용하여 각 경계치에 대응하는 $TMSE(\hat{\beta}_{PC})$ 의 추정치를 $\sum_{m=1}^{1000} (\hat{\beta}_{PCj}^{(m)} - \beta_j^0)^2 / 1000$ (단, $\hat{\beta}_{PCj}^{(m)}$ 은 m 번째 반복에서 j 번째 RPCR-추정치)에 의해 구하였다.

각 자료별로 $TMSE(\hat{\beta}_{PC})$ 의 추정치를 최소가 되게 하는 최적 경계치인 c^* 를 파악하였다. c^* 의 크기에 영향을 미치는 속성들 중에서 사전에 파악이 가능한 설명변수의 수(k), 관찰치의 수(n), 다중공선성에 관련된 변수의 수(m)를 설명변수로 채택하고, 자료별로 파악된 c^* 을 반응변수로 삼아 다음과 같은 선형모형,

$$c_F = 1.12618k + 0.00933n + 3.61326m - 2.11854 \quad (3.4)$$

을 구축하였다. 그런데 모형 (3.4)에서 필요한 m 의 크기는 분산분해비율을 활용하여 구할 수 있다. 즉, 상태지수가 30보다 크면서 분산분해비율이 0.5를 초과하는 경우의 수를 파악하면 m 의 크기를 알 수 있다. 주성분 선정 단계에서 식 (3.4)에 의해 경계치 c_F 를 구하고 자료로부터 상태지수 c_j 를 계산하여, $c_F < c_j$ 에 해당되는 주성분은 강력한 다중공선성에 관련된 것으로 판단하여 모형에서 제외시키고, $c_j \leq c_F$ 에 해당되는 주성분은 다중공선성과의 관련이 약한 것으로 인정하여 모형에 포함시킴으로써 주성분 선정과정이 완료된다.

4. 주성분 선정기준의 평가

제안된 선정기준을 비교평가하기 위하여 Monte Carlo 모의실험을 실행하였다. 모의실험을 위한 자료는 3.3절과 동일한 방법으로 생성하였는데, 설명변수 개수에 비례한 관찰치의 수는 30, 60, 90배, 다중공선성에 관련된 변수의 수는 2, 3개의 조합에 따라 회귀자료를 생성하였다. 모형을 개발하기 위해 생성한 추정용 자료와 상이한 평가용 자료를 생성하기 위하여 난수생성 seed를 다르게 지정하였다. 그리고 각 자료에 회귀이상점을 심기 위하여 생성된 자료의 일부를 임의로 오염시켰다. 즉, 나쁜 지렛점이 포함된 자료(BL)와 좋은 지렛점이 포함된 자료(GL)를 생성하기 위하여 설명변수 관찰치의 10%를 나쁜 지렛점과 좋은 지렛점으로 대체하였으며, 나쁜 지렛점과 수직이상점이 동시에 포함된 자료(BL+VO)와 좋은 지렛점과 수직이상점이 동시에 포함된 자료(GL+VO)를 생성하기 위하여 BL과 GL의 반응변수 관찰치의 5%를 수직이상점으로 대체하였다.

모의실험 반복수는 1,000회이며 프로그램은 SAS/IML을 사용하였다. 기존의 선정기준들과 새로운 선정기준을 적용한 RPCR-추정량의 성능을 MSE의 관점에서 비교 평가하였다. 즉, 다양한 규모의 자료에 Pidot (1969)의 기준(PIDOT), Marquardt (1970)의 기준(MARQ1: $\delta = 0.9$ 을 적용한 경우, MARQ2: $\delta = 0.99$ 을 적용한 경우), Jolliffe (1972)의 기준(JOLLI), Karlis 등 (2003)의 기준(KARLI)과 본 연구에서 제안한 선정기준(RPCR)을 적용하여 RPCR-추정치들을 구하였으며, 비교 목적을 위한 MSE는 $TMSE/p$ 로 추정하였다. 그런데 PIDOT를 적용하면 $k = 3 \sim 7$ 인 자료에서 주성분 전체가 모형에서 제외되는 경우가 발생하는데, 특히 $k = 3, m = 2$ 인 자료의 경우에 22% 이상의 자료에서 주성분 모두가 모형에서 제외되는 현상을 확인하였다. 역시 JOLLI를 적용하면 $k = 3, m = 2$ 인 경우 모든 주성분이 모형에서 제외되는 상황을 피할 수 없었다. 그리고 MARQ1의 경우 $k = 3$ 일 때 가장 큰 고유치의 비율이 90%를 초과하면 모든 주성분이 제외되는 상황이 발생하였다. KARLI 기준에서도 $k = 3 \sim 9$ 의 자료에서 주성분 전체가 모형에서 제외되는 경우가 발생하는데, 특히 $k = 3, m = 2$ 인 자료의 경우에 51% 이상의 자료에서 주성분 모두가 모형에서 제외되는 문제점을 확인하였다.

각 반복 중에서 주성분 전체가 모형에서 제외되어 LTS-추정치를 구할 수 없는 자료의 경우는 MSE 추정과정에서 제외되었으며, 선정기준별 RPCR-추정량의 MSE 추정 결과는 표 1과 표 2에 수록되었다. 선정기준별로 MSE 추정치에 유의적인 차이가 있는지 검정하기 위하여 분산분석을 실행하였다. 분석결과를 살펴보면 이상점의 형태(BL, BL+VO, GL, GL+VO)와 선정기준(PIDOT, MARQ1, MARQ2, JOLLI, KARLI, RPCR) 사이에 교호작용은 없는 것으로 확인되었다. 그리고 BL과 BL+VO, GL과 GL+VO 사이에는 유의적인 차이가 없었기 때문에 BL과 BL+VO를 통합하여

표 1: 선정기준별 RPCR-추정량의 평균제곱오차 추정치(나쁜 지렛점의 경우)

자료규모		PIDOT		MARQ1		MARQ2		JOLLI		KARLI		RPCRK	
$k(m)$	n	BL	BL+VO	BL	BL+VO	BL	BL+VO	BL	BL+VO	BL	BL+VO	BL	BL+VO
3(2)	90	.390	.390	.445	.447	.277	.277	.246	.247	.394	.396	.139	.134
	180	.337	.337	.392	.394	.233	.235	.192	.193	.329	.329	.109	.105
	270	.326	.326	.387	.388	.227	.228	.188	.188	.313	.315	.094	.092
4(2)	120	.477	.477	.353	.354	.232	.232	.222	.223	.531	.530	.115	.113
	240	.443	.442	.321	.321	.203	.204	.170	.188	.483	.482	.092	.089
	360	.446	.447	.320	.321	.197	.197	.149	.171	.488	.490	.079	.078
5(3)	150	.308	.308	.231	.233	.098	.097	.101	.102	.419	.419	.102	.100
	300	.297	.297	.214	.214	.073	.073	.069	.068	.400	.401	.076	.077
	450	.303	.303	.201	.201	.067	.067	.060	.060	.394	.394	.067	.066
6(3)	180	.360	.360	.219	.219	.083	.082	.101	.101	.513	.515	.086	.082
	360	.360	.361	.198	.198	.065	.065	.063	.063	.496	.497	.066	.067
	540	.357	.357	.185	.185	.059	.058	.052	.052	.482	.482	.063	.062
7(3)	210	.392	.392	.199	.199	.070	.071	.106	.106	.581	.580	.077	.074
	420	.385	.386	.178	.177	.058	.059	.065	.066	.564	.565	.064	.060
	630	.390	.390	.174	.174	.053	.053	.051	.051	.552	.553	.059	.059
8(3)	240	.417	.417	.193	.193	.069	.069	.119	.119	.630	.631	.058	.059
	480	.411	.412	.163	.163	.054	.054	.066	.066	.606	.607	.039	.039
	720	.416	.416	.163	.163	.047	.047	.051	.051	.599	.600	.036	.038
9(3)	270	.447	.447	.190	.191	.065	.065	.119	.119	.674	.674	.051	.052
	540	.442	.442	.166	.166	.050	.050	.067	.067	.653	.654	.036	.036
	810	.442	.442	.153	.153	.045	.045	.047	.047	.648	.649	.034	.034
평균		.388 ^b	.388 ^b	.240 ^c	.241 ^c	.111 ^d	.111 ^d	.110 ^d	.112 ^d	.512 ^a	.513 ^a	.073 ^d	.072 ^d
BL 평균		.388 ^b		.240 ^c		.111 ^d		.111 ^d		.512 ^a		.073 ^d	

표 2: 선정기준별 RPCR-추정량의 평균제곱오차 추정치(좋은 지렛점의 경우)

자료규모		PIDOT		MARQ1		MARQ2		JOLLI		KARLI		RPCRK	
$k(m)$	n	GL	GL+VO	GL	GL+VO	GL	GL+VO	GL	GL+VO	GL	GL+VO	GL	GL+VO
3(2)	90	.379	.381	.437	.440	.276	.278	.234	.235	.384	.386	.127	.127
	180	.338	.338	.388	.390	.234	.234	.187	.188	.331	.331	.094	.094
	270	.325	.325	.385	.386	.227	.227	.182	.183	.310	.310	.078	.078
4(2)	120	.457	.459	.344	.347	.228	.229	.209	.211	.529	.531	.099	.101
	240	.440	.439	.319	.318	.203	.203	.165	.165	.486	.487	.075	.075
	360	.443	.444	.320	.320	.195	.196	.146	.146	.490	.491	.063	.063
5(3)	150	.302	.302	.232	.232	.085	.086	.099	.099	.414	.414	.075	.075
	300	.294	.294	.212	.212	.067	.067	.067	.067	.398	.398	.056	.056
	450	.300	.300	.199	.200	.060	.060	.059	.059	.390	.391	.048	.048
6(3)	180	.354	.354	.224	.225	.077	.078	.100	.100	.509	.508	.065	.065
	360	.356	.357	.198	.198	.061	.061	.061	.061	.490	.491	.047	.047
	540	.354	.354	.186	.186	.052	.052	.051	.051	.479	.479	.040	.040
7(3)	210	.385	.385	.216	.215	.068	.068	.106	.106	.573	.574	.056	.056
	420	.381	.381	.182	.182	.055	.055	.064	.064	.559	.559	.040	.040
	630	.389	.389	.179	.179	.051	.051	.050	.050	.547	.548	.035	.035
8(3)	240	.417	.418	.213	.213	.069	.069	.114	.115	.625	.626	.052	.052
	480	.413	.414	.186	.186	.054	.054	.065	.065	.605	.605	.037	.037
	720	.416	.416	.179	.179	.048	.048	.051	.051	.598	.598	.032	.032
9(3)	270	.443	.443	.218	.218	.065	.065	.115	.115	.666	.667	.049	.049
	540	.437	.437	.191	.191	.054	.054	.066	.066	.648	.648	.032	.032
	810	.438	.438	.178	.178	.047	.047	.048	.048	.642	.643	.029	.029
평균		.384 ^b	.384 ^b	.247 ^c	.247 ^c	.108 ^d	.109 ^d	.107 ^d	.107 ^d	.508 ^a	.509 ^a	.059 ^d	.059 ^d
GL 평균		.384 ^b		.247 ^c		.109 ^d		.107 ^d		.509 ^a		.059 ^e	
전체 평균		.386 ^b		.244 ^c		.110 ^d		.109 ^d		.510 ^a		.066 ^e	

표 1에, GL과 GL+VO를 통합하여 표 2에 함께 수록하였다. 실험자료 전체에서 선정기준들 사이에는 유의적인 차이가 있는 것으로 나타났다. 즉 나쁜 지렛점만 포함된 자료와 나쁜 지렛점과 수직이상점이 함께 포함된 자료에서는 물론 좋은 지렛점만 포함된 자료와 좋은 지렛점과 수직이상점이 함

께 포함된 자료에서 선정기준별로 유의적인 차이가 있는 것으로 나타났다. 한편, 다중비교를 실행하였는데, RPCRK를 적용하면 전체 자료에서 가장 작은 MSE를 얻을 수 있는 것으로 나타났다. 그런데 BL과 BL+VO에서 MARQ2와 JOLLI 그리고 RPCRK 사이에 유의적인 차이가 없는 것으로 나타났으나, MARQ2와 JOLLI의 경우는 주성분 전체가 모형에서 제외되는 자료를 제외하고 MSE를 추정한 결과이기 때문에 RPCRK가 가장 우수한 기준이라 할 수 있다. 특히, 다른 기준들에 비하여 RPCRK는 좋은 지렛점의 경우에 MSE가 많이 축소되었는데, 이는 RPCRK가 상대지수를 바탕으로 하기 때문에 좋은 지렛점에 의해 강화된 다중공선성에 민감하게 반응한 것으로 해석된다.

한편, 제안된 기준 RPCRK의 일반화 문제를 점검하였다. 즉 설명변수의 수를 10개로 하고 관찰치의 수를 2,000개 및 10,000개로 확대하여 RPCRK를 실행하고 MSE를 추정하였다. 그 결과 나쁜 지렛점이 포함된 자료에서는 $n = 2,000$, $n = 10,000$ 인 경우에 MSE는 각각 0.023, 0.068이고, 좋은 지렛점이 포함된 자료에서는 $n = 2,000$, $n = 10,000$ 인 경우 MSE는 각각 0.018, 0.025이었다. 이 값들은 표 1과 표 2에 제시된 평균치보다 작은 값들이기 때문에, 모의실험의 범위를 벗어난 크기의 자료에도 RPCRK를 적용할 수 있는 것으로 판단된다.

5. 결론

데이터마이닝분야에서 회귀분석이 많이 활용되는데, 모형에 포함된 설명변수들의 특성상 다중공선성의 문제가 야기될 가능성이 높고, 자료수집 과정이 철저하게 계획되고 관리되지 않는 경우가 있기 때문에 회귀이상점이 다수 포함될 가능성이 높다. 본 논문에서는 다중공선성과 회귀이상점의 문제를 동시에 해결할 수 있는 RPCR에 관하여 연구하였는데, 특히 주성분 선정을 위한 새로운 기준을 제시하였다. 기존의 선정방법들이 갖는 한계점을 극복하기 위하여 표본공분산의 고유치 대신에 MVE-추정량의 상대지수를 바탕으로 삼은 기준을 제안하였는데, RPCR-추정량의 평균제곱오차를 최소화하는 상대지수의 경계치를 설정하고, 자료로부터 계산된 상대지수가 경계치보다 크면 해당 주성분을 모형에서 제외시키고, 경계치보다 작으면 해당 주성분을 모형에 포함시키는 방법이다. 한편, 지렛점은 물론 회귀이상점의 영향도 줄이기 위하여 최소자승추정 대신에 LTS-추정을 적용하여 회귀계수를 추정하는 과정을 도입하였다. 모의실험을 통해 기존의 선정기준들보다 제안된 기준이 상대적으로 우수한 것으로 평가되었다.

참고 문헌

- 김부용, 신명희 (2010). 주성분회귀분석에서 주성분선정을 위한 새로운 방법, <응용통계연구>, **23**, 967-975.
- Fauconnier, C. and Haesbroeck, G. (2009). Outliers detections with the minimum covariance determinant estimator in practice, *Statistical Methodology*, **6**, 363-379.
- Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models, *Journal of the American Statistical Association*, **88**, 1264-1272.
- Hubert, M. and Verboven, S. (2003). A robust PCR method for high-dimensional regressors, *Journal of Chemometrics*, **17**, 438-452.
- Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I: artificial data, *Applied Statistics*, **21**, 160-173.
- Karlis, D., Saporta, G. and Spinakis, A. (2003). A simple rule for the selection of principal components, *Communications in Statistics-Theory and Methods*, **32**, 643-666.
- Kim, B. Y. and Kim, H. Y. (2002). Hybrid algorithm for identification of regression outliers, *The Korean Communications in Statistics*, **9**, 291-304.

- Kim, B. Y. and Oh, M. H. (2004). Identification of regression outliers based on clustering of LMS-residual plots, *The Korean Communications in Statistics*, **11**, 485–494.
- Legendre, P. and Legendre, L. (1998). *Numerical Ecology*, Elsevier Science, Amsterdam.
- Marden, J. I. (1999). Some robust estimates of principal components, *Statistics & Probability Letters*, **43**, 349–359.
- Marquardt, D. W. (1970). Generalized inverse, ridge regression, biased linear estimation, and nonlinear estimation, *Technometrics*, **12**, 591–612.
- Mason, R. L. and Gunst, R. F. (1985). Outlier-induced collinearities, *Technometrics*, **27**, 401–407.
- McKean, J. W., Sheather, S. J. and Hettmansperger, T. P. (1993). The use and interpretation of residuals based on robust estimation, *Journal of the American Statistical Association*, **88**, 1254–1263.
- Pidot, Jr., G. B. (1969). A principal components of the determinants of local government fiscal patterns, *The Review of Economics and Statistics*, **51**, 176–188.
- Rocke, D. M. and Woodruff, D. L. (1997). Robust estimation of multivariate location and shape, *Journal of Statistical Planning and Inference*, **57**, 245–255.
- Rousseeuw, P. J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871–880.
- Rousseeuw, P. J. and Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41**, 212–223.
- Rousseeuw, P. J. and Driessen, K. (2006). Computing LTS regression for large data sets, *Data Mining and Knowledge Discovery*, **12**, 29–45.
- Rousseeuw, P. J. and Leroy, A. M. (2003). *Robust Regression and Outlier Detection*, Wiley-Interscience.
- Rousseeuw, P. J. and Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, **85**, 633–639.
- Woodruff, D. L. and Rocke, D. M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators, *Journal of the American Statistical Association*, **89**, 888–896.

A Criterion for the Selection of Principal Components in the Robust Principal Component Regression

Bu-Yong Kim^{1,a}

^aDepartment of Statistics, Sookmyung Women's University

Abstract

Robust principal components regression is suggested to deal with both the multicollinearity and outlier problem. A main aspect of the robust principal components regression is the selection of an optimal set of principal components. Instead of the eigenvalue of the sample covariance matrix, a selection criterion is developed based on the condition index of the minimum volume ellipsoid estimator which is highly robust against leverage points. In addition, the least trimmed squares estimation is employed to cope with regression outliers. Monte Carlo simulation results indicate that the proposed criterion is superior to existing ones.

Keywords: Multicollinearity, outlier, robust principal components regression, minimum volume ellipsoid estimator, condition index, least trimmed squares estimation.

This research was supported by Sookmyung Women's University Research Grants (2010).

¹ Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.
E-mail: buykim@sookmyung.ac.kr