이 성 건 교수지도

석사학위 청구논문

수정된 RV계수를 이용한 주성분 개수 결정에 관한 연구

2012

성신여자대학교 대학원

통 계 학 과

김 소 영

수정된 RV계수를 이용한 주성분 개수 결정에 관한 연구

이 성 건 교수지도

이 논문을 석사학위논문으로 제출함

2011년 11월

성신여자대학교 대학원

통 계 학 과

김 소 영

인 준 서

김소영의 석사학위 논문으로 인준함.

심사위원 _____인

심사위원 _____인

심사위원 _____인

성신여자대학교 대학원

논 문 개 요

다변량분석(Multivariate analysis)은 다변량 데이터의 변수들 간 상관관계를 이용하여 차원의 수를 축약하는 통계적 분석방법이다. 그중에서도 주성분분석(Principal Component Analysis: PCA)은 연속형 독립변수들의 선형결합을 통해 원데이터가 내포한 정보의 대부분을 설명하고 의미 있는 주성분이란 새로운 결합변수를 탐색하는 분석방법이다.

주성분분석의 목적은 다차원데이터에 대한 정보의 손실을 최소화하는 소수의 주성분을 선택하여 차원을 축소하는 것으로서 적절한 주성분의 개수를 추정하는데 있어서 여러 가지 방법들이 연구되어왔다. 방법들은 크게 두 분류로 나눠질 수 있는데 분포의 가정이 필요한 병렬분석, 바틀렛의 구형성검정이나 분포의 가정이 필요하지 않은 탐색적 방법으로 카이저규칙과 산비탈그림과 같은 방법 등이 널리 쓰이고 있다.

본 논문에서는 고차원 행렬간의 상관성 척도인 수정된 RV계수를 기반으로 적절한 주성분의 개수를 추정하고 그 개수가 적절한지에 대해 모의실험과 실제자료에 적용시켜봄으로서 효율성을 평가하였다. 그 결과 관측치의 수보다 변수의 수가 더 큰 경우에 수정된 RV계수를 이용한 주성분개수 결정방법이 기존의 방법보다 우수함을 확인하였다.

목 차

제 1 장 서 론

다변량분석(Multivariate analysis)은 변수들 간의 인과관계(casual relationship)를 규명, 분석하거나 변수들 간의 상관관계를 이용하여 변수를 축약(reduction), 개체들을 분류(classification)하는데 관련된 분석방법이다. 일반적으로 협의의 다변량분석이란 후자를 일컬으며 변수와 개체의 수가 많은 대용량이고 복잡한 데이터에 대한 분석방법이다(권세혁, 2008).

그중에서도 주성분분석(Principal Component Analysis: PCA)은 연속형 독립변수들의 선형결합을 통해 의미 있는 새로운 결합변수인 주성분(principal component)을 탐색하고 원데이터에 대한 정보의 손실을 최소화 하여 차원을 축소하는 분석방법이다. 주성분분석의 목적은 원데이터의 정보를 가장 많이 설명하는 즉, 정보의 손실을 최소화하는 적절한 개수의 주성분들을 선택함으로써 차원을 줄이는 것이다. 제 1 주성분은 원데이터의 정보를 가장 많이 반영하는 독립변수들의 결합이 되고 나머지 주성분들도 순차적으로 남은 정보를 최대한 설명하는 독립변수들의 선형결합이 된다. 결국 독립변수의 수만큼 생성된 주성분들은 원데이터의 정보를 모두 다 설명하고 서로 독립인 새로운 독립변수들이 된다.

주성분의 적절한 개수를 추정하는 방법들에 대해 많은 연구 진행되어왔다(Jackson, 1991; Peres-Neto, 2005; Jolliffe, 2002). 기존의 방법들은 자료행렬의 상관행렬로부터 고유값, 고유벡터를 구하고 이를 이용하여 주성분의 개수를 구하는 것이 대부분이었다. 하지만 이러한 방법들은 변수들 간의 상관계수, 개체, 차원 수에 매우 민감한 결과를 보이는 것으로 알려졌다(Peres-Neto, 2005). 이와 같은 단점을 보완하기 위해 상관행렬이 아닌 행렬간의 상관성 척도인 RV계수를 이용한 새로운 개수결정방법이 제안되었다(Dray, 2008). 특이값분해(Singular Value Decomposition: SVD)를 이용해 원자료행렬의 근사

(approximation)자료행렬과 잔차를 구하고 이들간의 RV계수인 RVDIM을 산출하여 순열검정(permutation test)을 기반으로 주성분의 수를 정하는 방법이다. 그러나 일반적으로 RV계수는 차원이 클수록 행렬간의 상관계수를 과대추정되는 경향이 있는 것으로 알려져 있다. 이와 같은 단점을 보완하기 위한 방법으로 고차원의 자료행렬간의 상관성의 척도로 수정된 RV계수가 제안되었다(Smilde, 2009).

본 논문에서는 고차원의 자료행렬간의 상관성 척도인 수정된 RV계수를 주성분의 개수 탐색에 활용하는 방법을 제안하고자 한다. 제안한 방법의 효율성을 살펴보기 위해 RVDIM을 이용한 프로세스와 수정된 RV계수를 이용한 새로 제안할 방법인 수정된 RVDIM을 모의실험과 실제자료 분석을 통해 서로 비교해보고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 주성분의 개수를 추정하는 기존 방법론들을 소개하고, 3장에서는 RVDIM과 수정된 RV계수를 소개하고 수정된 RVDIM을 제안한다, 4장에서는 RVDIM과 수정된 RVDIM을 이용하여 주성분의 개수가 알려져 있을때 얼마나 정확히 주성분의 개수를 추정하는 모의실험을 실시하여 결과를 비교해본다. 5장에서는 실제자료에 대해 적용하여 결과를 살펴보고, 6장에서는 본 연구의 한계점 및 향후 연구방향을 제시한다.

제 2 장 주성분분석 방법론

주성분분석의 목적은 서로 밀접한 관계가 있는 다수의 독립변수들로 구성된 다차원 데이터를 정보의 손실을 최소화하면서 차원의 수를 줄이는 것이다. 주성분분석은 독립변수들의 선형결합을 통해 서로 독립인 주성분이라 불리는 새로운 독립변수들을 생성하고, 그중에서 소수의 주성분들을 선택함으로써 원데이터의 정보를 최대한 반영하며 차원을 축소하는 방법이다.

원데이터의 공분산행렬(또는 상관행렬)로 부터 고유값, 고유벡터를 구하여 원데이터의 정보의 양인 변동성의 크기를 구하고 그 값을 기반으로 주성분을 선택하는 방법이 널리 쓰이고 있다. 고유값이 클수록 주성분의 기여도가 크고 기여도가 큰 주성분은 변동성의 많은 부분을 설명한다. 원데이터의 변수의 개수만큼 생성되는 주성분들 중에서 몇 개의 주성분이 선택되어야 주성분분석의 목적이 달성될 수 있을지에 대한 여러 방법들이 제안되어왔다.

Jackson(2002)은 주성분의 개수를 선택하는 방법들을 크게 두 개로 나누어 분포가정을 필요로 하는 방법과 탐색적 방법으로 분류하여 소개하였다. 분포의 가정을 필요로 하는 방법들로는 병렬분석(parallel analysis), 바틀렛 구형성 검정(Barlett's test of sphericity), Lawly검정(Lawley's test)등이 있는데 이들은 계산절차가 복잡하고 데이터가 다변량 정규성(multivariate normality)을 만족하지 못할 때 결과가 안정적이지 못한 단점이 있다. 탐색적 방법들로는 카이저규칙(Kaiser-Guttman rule), 총 분산의 누적점유율, 산비탈그림(scree plot)등이 있고 이들은 계산절차가 간단하고 개수를 정하는 기준이 간단하여 널리 쓰이고 있다.

본 장에서는 위에서 언급한 주성분의 개수를 결정하는 방법들에 대해 간략히 소개하고자 한다.

2.1. 분포의 가정을 필요로 하는 방법론

분포의 가정이 필요한 방법들은 개수를 추정하는데 있어서는 비교적 정확하나 가정이 비현실적(예를 들어, 데이터가 다변량 정규성을 만족해야함)이고 차원을 과대 추정하는 경향이 있다. 또한 계산이 복잡할 뿐만 아니라 다변량 정규성을 만족하지 못하면 결과의 신뢰성이 떨어지고 표본크기에 민감하다는 것이 알려져 있다.

2.1.1. 병렬분석(parallel analysis)

병렬분석은 관찰된 데이터로부터 구한 고유값이 랜덤데이터로부터 구한 고유값보다 큰 주성분을 선택하는 방법이다(Horn, 1965). 이때 랜덤데이터들은 관찰된 데이터의 차원과 개체수가 동일하고 몬테카를로 방법(Monte carlo method)에 의해 생성된 독립적인 정규분포확률변수(independent normally distributed variables)로 구성된다. 다수의 랜덤데이터들로 부터 고유값들을 구하고, 고유값들로부터 경험적(empirical)분포를 생성하고 분포의 백분위수 구간(percentile interval)을 이용하여 주성분을 선택한다. 병렬분석의 순서는 다음과 같다.

1) 평균은 0 분산은 1인 표준정규분포로부터 서로 독립인 정규분포확률변수(independent normally distributed variable)들로 구성된 랜덤데이터 생성. 여기서 랜덤데이터들은 관찰된 데이터와 동일한 개체수와 차원수로 구성되어야 함

2) 1)에서 생성된 랜덤데이터의 상관행렬의 고유값을 산출

3) 1), 2)의 과정을 총 1000번 반복. 단, 반복수는 분석가의 주관에 따라 달라질 수 있음

4) 각 주성분에 대한 고유값을 평가하는데 임계치(critical value)로 사용될 경험적 백분위수(일반적으로 95%) 구간을 계산

5) 만약 관찰된 데이터의 고유값이 경험적 백분위수 구간보다 크다면 주성분으로 선택(즉, 95번째 분위수의 값보다 관찰된 데이터의 값이 크다면 주성분으로 선택)

병렬분석은 주성분의 개수를 결정하는 여러 가지 방법들 중에서 비교적 정확하게 주성분의 개수를 추정한다고 알려져 있지만 다변량 정규성에 강건(robust)하다는 단점이 있다.

2.1.2. 바틀렛 구형성검정(Bartlett's test of sphericity)

바틀렛 구형성검정(Bartlett's test of sphericity)은 '상관행렬이 단위행렬이다'라는 귀무가설을 검정하여 요인분석시 사용이 적합하고 공통된 요인이 있는지를 검정하는 방법이다.

주성분 개수의 추정에 있어서 바틀렛 구형성검정은 순차적인 각각의 고유값과 나머지 고유값들 간에 유의한 차이가 있는지를 검정하는 방법이다(Pimentel 1979). 만약 귀무가설이 'k번째 주성분의 고유값이 나머지 $p-k$개의 주성분들의 고유값들이 같다'라면 다음의 검정통계량값에 따라 귀무가설의 기각 또는 채택여부가 결정된다.

$$\chi_k^2 = \left[n-k- \frac{2(p-k)+7+\frac{2}{(p-k)}}{6} + \sum_{j=1}^{k}\left(\frac{\overline{\lambda}}{(\lambda_j - \overline{\lambda})} \right)^2 \right] \times \left[-\ln \prod_{j=k+1}^{p} \lambda_j + (p-k)\ln \overline{\lambda} \right], \quad (2.1)$$

여기서 $\overline{\lambda} = \sum_{j=k+1}^{p} \frac{\lambda_i}{(p-k)}$이고 p는 자료행렬의 변수의 개수를, λ_k는 k번째 주성분의 고유값, n은 개체의 수를 나타내고 검정통계량 χ_k^2는 근사적으로 자유도가 $0.5(p-k-1)(p-k+2)$ 인 카이제곱 분포를 따른다. k는 1부터 p까지의 값을 갖고 순차적으로 χ_k^2가 유의하지 않을 때까지 검정한다. 예를 들어, χ_k^2은 유의하나 χ_{k+1}^2은 유의하지 않다면 k개의 주성분을 선택할 수 있다.

상관행렬로부터 구한 제 1 주성분의 고유값이 나머지 $p-1$개 주성분들의 고유값들과 유의한 차이가 있는지 검정하는 방법도 연구되었다(Bartlett, 1954). 귀무가설은 '모든 변수들은 서로 독립이다'이고 검정통계량은 다음과 같이 구할 수 있다.

$$\chi^2 = -\left[n - \frac{1}{6}(2p+11) \right] \ln |R|, \quad (2.2)$$

여기서 n은 개체의 수를, p는 변수의 수, 그리고 $|R|$은 관찰된 데이터로부터 구한 상관행렬의 행렬식(determinant)이다. 검정통계량은 근사적으로 자유도가 $0.5p(p-1)$ 인 카이제곱분포를 따른다. 이 검정법은 제 1 주성분에 대해서만 평가하므로 주성분 개수 결정에 위의 바틀렛 구형성 검정에 비해 많은 계산시간을 줄일 수 있다. 그러나 단지 한 개의 주성분에 대해 검정한다는 한계점이 존재한다.

2.1.3. 제 2 주성분에 대한 Lawly검정

Lawly(1956)는 제 2 고유값이 나머지 $p-2$개의 고유값들과 유의한 차이가 있는지 평가하는 방법을 제안하였다. 즉, 귀무가설은 '적어도 두 개 이상의 변수들은 서로 상관되어있고 제 2 고유값은 나머지 $p-2$개의 고유값들과 유의한 차이가 없다.'이고 검정통계량은 다음과 같이 구할 수 있다.

$$\chi^2 = \frac{n-1}{\lambda}\sum_{\substack{i=1 \\ i\neq j}}^{p}\sum_{\substack{j=1 \\ i\neq j}}^{p}(r_{ij}-\bar{r})^2 - \mu\sum_{k=1}^{p}(\overline{r_k}-\bar{r})^2, \tag{2.3}$$

여기서 r_{ij}는 변수 i와 j간의 상관계수이고,

$$\begin{aligned}
\bar{r} &= \frac{2}{p(p-1)}\sum_{i=k+1}^{p}\sum_{j=1}^{p} r_{ij}, \\
\lambda &= 1-\bar{r}, \\
\mu &= \frac{(p-1)^2(1-\lambda^2)}{p-(p-2)\lambda^2}, \\
\overline{r_k} &= \frac{1}{p-1}\sum_{i=1, i\neq k}^{p} r_{ik}, \qquad k = 1\cdots p
\end{aligned} \tag{2.4}$$

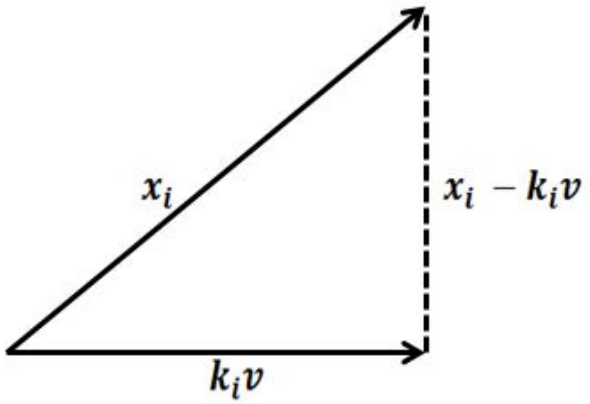
와 같이 계산될 수 있으며 검정통계량은 근사적으로 자유도가 $0.5(p+1)(p-2)$인 카이제곱분포를 따른다.

비록 Lawly검정은 처음 두 개의 주성분에 대한 평가만을 제공하는 한계점이 존재하나 2개의 주성분에 의해 데이터의 변동성이 대부분 설명되는 경우에는 유용한 검정방법이다.

2.2. 탐색적 방법론

본 절에서는 분포의 가정이 필요하지 않은 탐색적 방법들을 소개하고자 한다. 탐색적 방법들은 앞 절에서 소개한 방법들에 비해 직관적이고 이해가 쉽기 때문에 널리 쓰이고 있다. 또한 주성분의 개수를 결정하는 기준이 간단하고 명확하므로 빠른 의사결정이 가능하다.

방법론들을 살펴보기에 앞서 자료행렬 $\boldsymbol{X}$가 평균 0, 분산 1로 표준화된 $(p\times1)$인 n개의 개체벡터 $\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}$들로 이루어졌다고 가정하자(즉, $\boldsymbol{X}$는 $(n\times p)$인 자료행렬). 그리고 개체벡터들을 $(p\times1)$인 단위벡터 $\boldsymbol{v}$에 사영(projection)시키기로 하자($\boldsymbol{v^tv}=1$). 그렇다면 $\boldsymbol{x_i}=\boldsymbol{k_iv}+(\boldsymbol{x_i}-\boldsymbol{k_iv})$(단, $\boldsymbol{k_i}$는 실수)로 나타낼 수 있고, $\boldsymbol{k_iv}$는 $\boldsymbol{x_i}$가 단위벡터에 사영된 크기를 나타내고 $\boldsymbol{x_i}-\boldsymbol{k_iv}$는 잔차가 된다. p차원공간에서 잔차들의 평균제곱을 최소화하는 1차원 공간을 찾을 수 있다. 이는 [그림 1]로 나타낼 수 있다.



[그림 1] 벡터 $\boldsymbol{x_i}$를 단위벡터 $\boldsymbol{v}$에 사영

주성분분석에서 고유값을 산출하기 위해 사용되는 자료행렬 $\boldsymbol{X}$의 공분산행렬은 $\boldsymbol{C}=(1/(n-1)\boldsymbol{X}^t\boldsymbol{X})$으로 정의할 수 있다($\boldsymbol{X}$는 표준화된 데이터들로 이루어졌기 때문에 공분산행렬은 상관행렬과 동일하다.). $\boldsymbol{C}$는 고유방정식에 의해 $\boldsymbol{Cv}=\boldsymbol{\lambda v}$로 분해될 수 있고 여기서 $\boldsymbol{v}$는 $(p\times 1)$인 고유벡터, $\boldsymbol{\lambda}$는 고유값으로서 이들의 짝은 자료행렬의 변수의 수(즉, p개)만큼 도출된다.

i번째 주성분의 분산은 i번째 고유값 λ_i와 같고 총 p개의 고유값들의 크기는 $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p$와 같이 나타낼 수 있다. p개의 고유값에 의해 데이터의 모든 변동성이 설명될 수 있고 이는 관찰된 데이터의 변수들이 서로 독립인 변수들로 변환된 것을 뜻한다. 제 1 주성분의 분산은 λ_1이므로 관찰된 데이터의 변동을 가장 많이 설명하고 데이터의 변동성 대부분은 처음 몇 개의 주성분들에 의해 설명된다. 그러므로 주성분분석은 여러 가지 기준을 적용하여 소수의 주성분을 선택함으로써 정보의 손실을 최대한 줄이며 차원을 축소하는 분석기법이다.

2.2.1. 카이저 규칙(Kaiser-Guttman rule)

카이저 규칙(Kaiser-Guttman rule)은 일반적으로 널리 사용되는 주성분의 개수 결정방법 중 하나로서 고유값들의 평균보다 큰 고유값을 갖는 주성분을 선택하는 방법이다.

$\boldsymbol{X}$는 표준화 되어있으므로 공분산행렬(즉, 상관행렬) $\boldsymbol{C}$는 $(p\times p)$이고 대각원소가 자기 자신의 상관계수이므로 모두 1이 된다. 그러므로 상관행렬의 대각원소의 모든 합 p는 $\boldsymbol{X}$의 분산이다. 고유값들의 총합은 주성분의 분산과 동일하므로

$$Var(\boldsymbol{X}) = \sum_{i=1}^{p} \lambda_i = p\,, \tag{2.5}$$

가 되고 고유값의 평균은 1이 된다. 고유값이 1보다 큰 주성분은 평균적으로 다른 주성분에 비해 원데이터에 대한 변동성을 많이 설명한다고 볼 수 있다. 만약 공분산행렬을 기반으로 고유값을 구한 경우에는 카이저규칙을 적용 할 때 1이 아닌 고유값들의 평균을 기준으로 두고 주성분을 선택할 수 있다.

2.2.2. 총 분산의 누적점유율

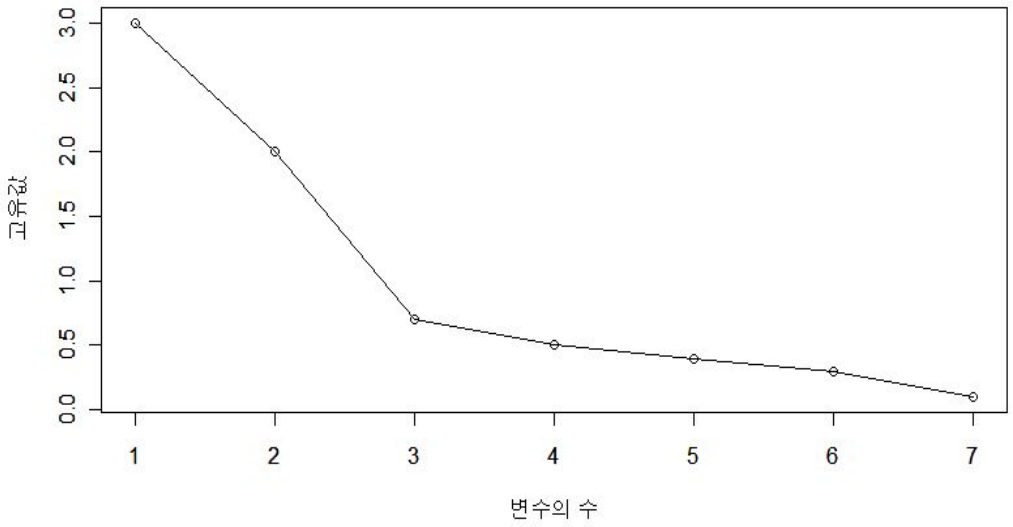
총 분산의 누적점유율방법은 관찰된 데이터의 총 분산(변동)중에서 주성분들의 분산의 누적점유율이 80~90% 이상을 차지하는 처음 몇 개의 주성분들을 선택하는 방법이다. k번째 주성분까지의 누적 점유율은 다음과 같이 구할 수 있다.

$$100 \times \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} = \; 100 \times \frac{\sum_{i=1}^{k} \lambda_i}{p}, \tag{2.6}$$

식(2.6)에서 오른쪽 항은 자료행렬 $\boldsymbol{X}$가 표준화된 경우에만 만족한다. 만약 λ_1이 관찰된 데이터의 총변동 중에서 80%이상을 설명한다면 제 1주성분은 원데이터의 80%이상을 반영한 축으로서 p차원인 $\boldsymbol{X}$는 20%이하의 정보의 손실을 감안하더라도 1차원으로 축소될 수 있다. 기준으로 제시한 80~90%의 누적점유율은 실제 적용에 있어서 60~70%등 완화된 기준을 적용할 수 있다.

2.2.3. 산비탈그림(scree plot)

산비탈그림(scree plot)은 카이저 규칙과 총 분산의 누적점유율만큼 주성분의 개수를 구하는데 있어서 널리 쓰이는 방법 중 하나이다(Cattel, 1966). 관찰된 데이터의 변수의 개수만큼 생성된 p개의 주성분의 고유값들을 그래프를 통해 시각화함으로써 주성분의 개수를 결정하는 방법이다.



[그림 2] 산비탈그림(변수의 수=7)

[그림 2]는 7개의 변수로 이루어진 데이터의 상관행렬로부터 구한 고유값들에 대한 산비탈그림이다. 이처럼 산비탈그림은 고유값들을 직선으로 이어서 나타내 직선의 기울기가 급격하게 꺾이는 지점에서 주성분의 개수를 선택한다. [그림 2]에서는 제 3 주성분의 위치에서 기울기가 급격히 완만해 지므로 산비탈 그림을 통해서 우리는 3개의 주성분을 선택할 수 있다.

제 3 장 RVDIM과 수정된 RV계수

일반적으로 상관계수는 두 변수간의 상관성을 나타내는 값으로서 값이 1에 가까울수록 두 변수는 서로 상관이 높고, 0에 가까우면 상관이 없다고 볼 수 있다(즉, 독립). 두 변수뿐만 아니라 두 자료행렬간의 상관성을 측정하는 값들도 존재하는데 그중에서도 같은 수의 개체를 갖지만 차원수가 다른 두 자료행렬간의 상관성은 RV계수를 통해 평가될 수 있다(Escoufier, 1973).

다변량 피어슨 상관계수라 볼 수 있는 RV계수는 상관계수와 같이 1에 가까울수록 두 자료행렬간 상관이 높다는 것을 뜻한다. 반면 0에 가까우면 두 자료행렬은 서로 독립이라고 볼 수 있다(즉, 두 자료행렬의 변수들끼리는 서로 독립이다). RV계수를 이용하여 주성분의 개수를 정하는 RVDIM[1]이 제안되었다(Dray, 2008). RV계수 계산시 특이값분해(Singular Value Decomposition: SVD)를 이용해 원자료행렬의 근사(approximation)자료행렬과 잔차를 구하고 이들간의 RV계수인 RVDIM을 통해 순열검정(permutation test)을 기반으로 하여 주성분의 수를 정하는 방법이다.

RV계수는 개체수와 차원의 수에 영향을 많이 받는 값으로서 특히 RV계수는 자료행렬들의 변수의 수가 많을수록 행렬간의 상관을 과대추정하는 경향이 있는데 이러한 RV계수의 약점을 보안하기 위한 수정된 RV계수가 제안되었다(Smilde, 2009). 수정된 RV계수는 개체수에 비해 차원의 수가 훨씬 큰 자료행렬들간의 상관성을 구할 때 RV계수에 비해 덜 큰 값을 갖는다.

RVDIM역시 개체수와 변수의 수에 영향을 많이 받으므로 고차원의 자료행렬에서는 주성분의 개수선택에 있어 불안정한 결과를 도출할 것으로 예상된다다. 본 절에서는 RVDIM과 수정된 RV계수를 함께 이용한 수정된 RVDIM[2]을

1) 본 논문에서는 RVDIM을 이용하여 주성분의 개수를 선택하는 프로세스를 통틀어 일컬음
2) 수정된 RVDIM을 이용하여 주성분의 개수를 선택하는 프로세스를 통틀어 일컬음

제안하고자 한다. RVDIM과 수정된 RV계수에 대해 살펴보기 전 두 개의 자료행렬간의 연관성 측도인 RV, COI, RLS계수부터 살펴보고자 한다.

3.1. 두 행렬간의 연관성척도

만약 자료행렬 $\boldsymbol{X}$, $\boldsymbol{Y}$의 상관계수가 1이라면 $\boldsymbol{X}$, $\boldsymbol{Y}$는 단지 직교회전(orthogonal rotation)이 다르다는 것을 뜻한다(즉, $\boldsymbol{X}=\boldsymbol{YQ}$, $\boldsymbol{Q}^t\boldsymbol{Q}=\boldsymbol{I}$)[3]. 이러한 자료행렬간의 연관성의 척도로소 여러 값이 존재하는데 차원이 서로 다른 두 자료행렬간의 상관성에 관한 척도로 RV계수는 그 값이 1에 가까울수록 두 자료행렬은 상관성이 높다고 볼 수 있다.

자료행렬 $\boldsymbol{X}$와$\boldsymbol{Y}$가 각각 중심화 된 $(n\times 1)$인 p, q개의 열벡터들로 구성되어 있다고 가정하자(즉, $\boldsymbol{X}(n\times p)$, $\boldsymbol{Y}(n\times q)$). 그렇다면 RV계수는 다음과 같이 정의할 수 있다.

$$RV(\boldsymbol{X},\boldsymbol{Y}) = \frac{tr(\boldsymbol{X}\boldsymbol{X}^t\boldsymbol{Y}\boldsymbol{Y}^t)}{\sqrt{tr(\boldsymbol{X}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{X}^t)tr(\boldsymbol{Y}\boldsymbol{Y}^t\boldsymbol{Y}\boldsymbol{Y}^t)}}, \qquad (3.1)$$

RV계수는 0에서 1사이의 값을 갖는데 $\boldsymbol{X}$의 변수들과 $\boldsymbol{Y}$의 변수들 간에 상관이 없다면(즉, 두 자료행렬의 각 변수들은 서로 독립) 0의 값을 갖고, 반대의 경우에는 1의 값을 갖는다. 만약 자료행렬 $\boldsymbol{X}$와$\boldsymbol{Y}$가 모두 1차원($p=q=1$)이라면 RV계수는 단순 상관계수(ρ)의 제곱 값을 갖는다($RV=\rho^2$).

RV계수의 분자부분은 $\boldsymbol{X}$와 $\boldsymbol{Y}$사이의 연관성의 척도인 co-interia criterion(Dray, 2003)과 일치한다.

---

3) Smilde(2009)의 논문을 참조.

$$COI(\boldsymbol{X},\boldsymbol{Y}) = tr(\boldsymbol{X}\boldsymbol{X}^t\boldsymbol{Y}\boldsymbol{Y}^t) = tr(\boldsymbol{X}^t\boldsymbol{Y}\boldsymbol{Y}^t\boldsymbol{X}), \tag{3.2}$$

COI값이 클수록 두 데이터의 상관은 높다고 볼 수 있다.

또 다른 상관성의 척도로 Gower(1971) 와 Lingoes & Schonemann (1974)에 의해 제안된 RLS계수는 다음과 같이 정의할 수 있다.

$$RLS(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sqrt{tr(\boldsymbol{X}\boldsymbol{X}^t\boldsymbol{Y}\boldsymbol{Y}^t)}}{\sqrt{tr(\boldsymbol{X}^t\boldsymbol{X})tr(\boldsymbol{Y}^t\boldsymbol{Y})}}, \tag{3.3}$$

RLS계수도 RV계수처럼 0에서 1사이의 값을 갖고, 값이 1에 가까울수록 두 자료행렬은 두 자료행렬은 상관성이 높다고 볼 수 있다.

3.2. RVDIM

선형대수학에서 특이값분해는 행렬의 스펙트럼 이론을 임의의 직사각행렬에 대해 일반화한 방법으로 신호처리와 통계학 분야에서 많이 쓰인다.

특이값분해로 얻어진 $\boldsymbol{X}^* = (\frac{1}{\sqrt{n}})\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}_t$ 을 살펴보자. 여기서 $\boldsymbol{D}$는 r개의 고유값이 내림차순으로 정렬된 $(r\times r)$ 대각행렬이다. 행렬 $\boldsymbol{U}$는 r개의 $(n\times 1)$ 열벡터로 이루어진 $(n\times r)$행렬이고, 행렬 $\boldsymbol{V}$는 r개의 $(p\times 1)$ 열벡터로 이루어진 $(p\times r)$행렬이다. 특이값분해는 계수(rank)가 m인 $\hat{\boldsymbol{X}}_m$에 의한 자료행렬 $\boldsymbol{X}$의 근사(approximation)와도 매우 친밀히 연관되어있다. $\boldsymbol{X}$의 최량근사(best approximation)는 최소제곱법의 관점에서(예를 들어, $\| \boldsymbol{X} - \widehat{\boldsymbol{X}_m} \|^2$ 을 최소화) 다음과 같이 나타낼 수 있다(Good, 1969).

$$\hat{\boldsymbol{X}}_{\boldsymbol{m}} = \sum_{i=1}^{m} d_i \boldsymbol{u_i} \boldsymbol{v_i^t} = \sum_{i=1}^{m} X_i \,, \tag{3.4}$$

여기서 $X_i = d_i \boldsymbol{u_i} \boldsymbol{v_i^t}$ 이고 잔차는 다음과 같이 정의할 수 있다.

$$\begin{aligned} R_i = \ \boldsymbol{X} - \sum_{j=1}^{i-1} d_j \boldsymbol{u_j} \boldsymbol{v_j^t} \ &= \sum_{j=1}^{r} d_j \boldsymbol{u_j} \boldsymbol{v_j^t} - \sum_{j=1}^{i-1} d_j \boldsymbol{u_j} \boldsymbol{v_j^t} \\ = \sum_{j \neq i}^{r} d_j \boldsymbol{u_j} \boldsymbol{v_j^t} &= \sum_{j=i}^{r} X_j \,, \end{aligned} \tag{3.5}$$

여기서 $\boldsymbol{X}$는 다시 다음과 같이 나타낼 수 있다.

$$\boldsymbol{X} = R_1 = X_1 + R_2 = X_1 + X_2 + R_3 = \sum_{j=1}^{i} X_j + R_{i+1}. \tag{3.6}$$

특이값분해를 통해 주성분분석에서의 주성분 개수 결정에 대한 문제는 $\boldsymbol{X}$의 근사의 문제로 볼 수 있다. 즉, $\boldsymbol{X}$를 추정하는데 있어서 $\hat{X}_i$가 $\hat{X}_{i-1}$보다 더 유의한 근사인지에 판단하는 것이다. X_i가 랭크가 $i-1$인 분해행렬 $\hat{X}_{i-1}$에 연관성 있는 정보를 더하는지 여부를 알기위해 Dray(2008)는 RVDIM을 제안했다. i번째 주성분에 의해 형성된 1차원 공간에서 X_i는 앞의 $i-1$ 개의 주성분들에 의해 설명되고 남은 변동성을 가장 많이 설명하는 축을 뜻한다. 반면 R_i는 나머지 $(r-i+1)$개의 주성분들에 형성된 $(r-i+1)$차원 공간으로 잔차를 뜻한다.

RVDIM은 X_i와 R_i의 상관성에 기반을 두고 있는데 그들의 상관성은 RV계수로 측정될 수 있고 RVDIM는 다음과 같이 정의할 수 있다.

$$RVDIM(i) = RV(X_i, R_i) = \frac{tr(X_i X_i^t R_i R_i^t)}{\sqrt{tr(X_i X_i^t X_i X_i^t) tr(R_i R_i^t R_i R_i^t)}}, \quad (3.7)$$

RV계수를 이용하여 X_i와 R_i의 상관을 통해 RVDIM이 클수록 R_i의 변동성의 많은 부분을 X_i가 설명하는 것을 뜻하고 결론적으로 $\hat{X}_{i-1}$에 비해 $\hat{X}_i$가 $\boldsymbol{X}$의 더 좋은 근사임을 수 있다.

근사적으로 RVDIM은 또한 다음과 같이 나타낼 수 있다.

$$RVDIM(i) = \frac{d_i^4}{\sqrt{d_i^4 \sum_{j=i}^{r} d_j^4}} = \frac{\lambda_i^2}{\lambda_i \sqrt{\sum_{j=i}^{r} \lambda_j^2}} = \frac{\lambda_i}{\sqrt{\sum_{j=i}^{r} \lambda_j^2}}, \quad (3.8)$$

위의 식 (3.8)은 총 분산의 누적점유율 식과 비슷한데, i번째 주성분의 고유값이 클수록 분자의 값은 커지고 분모의 값은 작아지므로 결국 $RVDIM(i)$은 커진다.

RVDIM을 통한 차원결정의 순서는 다음과 같다:

1) 자료행렬 $\boldsymbol{X}$에 대해 특이값분해 실시

2) $i(i=1,\cdots,p)$번째 축에 대해 관찰된 $RVDIM(i)$ 계산

3) i번째 축에 대해 다음을 반복함(예를 들어 999번):

① X의 각 열안에서 개체들을 랜덤화(randomization)시킴

② 순열행렬(permuted matrix)에 대해 특이값분해 실시

③ 순열행렬의 i번째 축에 대한 $RVDIM(i)$ 계산

4) i번째 축에 대한 p값 p_i를 추정

$$p_i = \frac{\text{관찰된 } RVDIM\text{보다 크거나 같은 랜덤값}(random\ value)\text{들의 개수} + 1}{999 + 1}$$

5) i번째 축에 대한 유의수준 α_i를 선택한 후에 $p_i < \alpha_i$ 이고 $p_{i+1} > \alpha_{i+1}$를 만족하면 1부터 i번째 축을 선택.

5)번째 단계에서 만약 $p_1 > \alpha_1$이면 어느 축도 선택되지 않는다. 즉, 데이터를 설명하는 주성분이 없다고 볼 수 있다.

Dray(2008)는 모의실험을 통해 본페로니 수정(Bonferroni adjustment)에 의한 RVDIM이 가장 정확한 주성분의 개수를 추정하는 확인할 수 있었다. 그러나 실제자료의 적용에 있어서는 다소 불안정안 결과를 보였다.

3.3. 수정된 RV계수

자료행렬 $\boldsymbol{X}(n \times p)$와 $\boldsymbol{Y}(n \times q)$의 개체들은 표준정규분포로부터 생성된 랜덤자료행렬(fully random matrix)이라 가정하자. 그렇다면 RV계수는 근사적으로 다음과 같이 나타낼 수 있다.

$$\begin{aligned} RV(X,Y) &\approx \frac{pq}{\sqrt{(p^2+2p+(n-1)p)(q^2+2q+(n-1)q)}} \\ &= \frac{pq}{\sqrt{(p^2+(n+1)p)(q^2+(n+1)q)}}, \end{aligned} \tag{3.9}$$

식 (3.9)을 살펴보면 개체수 n이 작을수록, 변수의 개수 p, q가 클수록 RV계수의 값은 커지는 것을 알 수 있다. 즉, 개체의 수는 적고 차원이 높은 자료행렬들을 이용하여 RV계수를 구하면 본래의 상관성보다 과대추정된다. 이러한 점을 보완하기 위해 고차원의 자료행렬들간의 RV계수인 수정된 RV계수가 제안되었다(Smilde, 2009).

수정된 RV계수는 RV계수와 달리 자료행렬과 전치한 자료행렬의 곱(즉, 벡터들 간의 외적)인 양의 준정부호 행렬(positive semi-definite) $\boldsymbol{XX^t}$의 대각원소를 모두 0으로 만듦으로써 $\boldsymbol{XX^t}$가 아닌 $\widetilde{\boldsymbol{XX^t}}=[\boldsymbol{XX^t}-diag(\boldsymbol{XX^t})]$를 이용하여 식(3.9)과 같이 구할 수 있다.

$$RV(\boldsymbol{X},\boldsymbol{Y})_{\mathrm{mod}} = \frac{tr(\widetilde{\boldsymbol{XX}^t}\widetilde{\boldsymbol{YY}^t})}{\sqrt{tr(\widetilde{\boldsymbol{XX}^t}\widetilde{\boldsymbol{XX}^t})tr(\widetilde{\boldsymbol{YY}^t}\widetilde{\boldsymbol{YY}^t})}}, \qquad (3.10)$$

RV계수는 0에서 1사의 값을 가졌지만 수정된 RV계수는 일반적인 상관계수와 같이 -1에서 1사이의 값을 갖고 -1의 값을 가질 경우에는 $\boldsymbol{X}$의 개체들 간의 연관성은 $\boldsymbol{Y}$의 개체들 간의 연관성에 반비례 한다. 즉, 상관계수가 -1일 때와 동일한 해석이 가능하다.

수정된 RV계수는 대각원소의 값을 모두 0으로 만듦으로써 고차원 자료행렬에서 RV계수에 비해 더 안정적인 값을 보이는 것으로 연구결과 나타났다.

3.4. 수정된 RVDIM

Smilde(2009)는 수정된 RV계수를 자료행렬이 마이크로어레이(microarray)인 경우에 대해서만 적용하였으나, 본 연구에서는 주성분 개수의 선택을 위해 수

정된 RV계수를 사용하는 것을 제안하고자 한다.

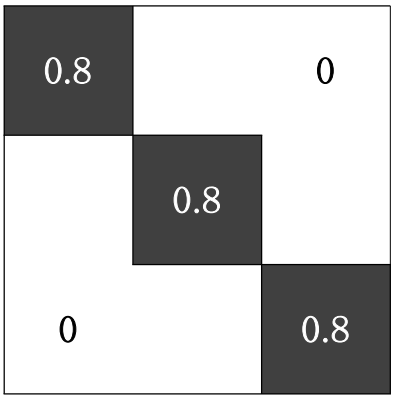
RVDIM과 수정된 RV계수를 결합한 새로운 방법인 수정된 RVDIM은 특이값분해를 통해 얻어진 분해된 행렬X_i와 X_i^t의 곱인 $X_iX_i^t$의 대각원소를 제외한 $\widetilde{X_iX_i^t}=[X_iX_i^t-diag(X_iX_i^t)]$를 이용하여 다음과 같이 구할 수 있다.

$$RVDIM(X_i,R_i)_{\mathrm{mod}}= \frac{tr(\widetilde{X_iX_i^t}\widetilde{R_iR_i^t})}{\sqrt{tr(\widetilde{X_iX_i^t}\widetilde{X_iX_i^t})tr(\widetilde{R_iR_i^t}\widetilde{R_iR_i^t})}}, \qquad (3.11)$$

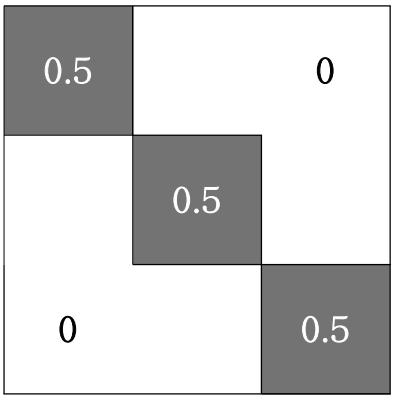
수정된 RVDIM역시 대각원소를 제거하여 사용함으로서 고차원 자료행렬에서 좀 더 안정된 결과를 보일 것이다. 다음 4장에서는 개체의 수는 고정하고 변수의 수가 많아질수록 RVDIM과 수정된 RVDIM 중에서 어느 계수가 주성분 개수를 추정하는데 있어서 효율적인지 비교해보고자 한다.

제 4 장 모의실험

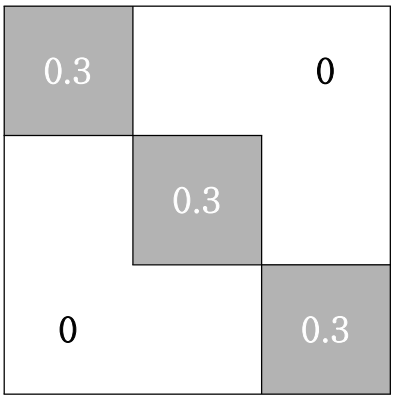
본 장에서는 개체수의는 고정되어 있고 주성분의 개수가 알려져 있을 때, 변수의 수에 따라 3장에서 소개한 RVDIM과 제안한 수정된 RVDIM을 이용한 주성분개수 결정방법의 수행능력을 모의실험을 통해 비교해보고자 한다.



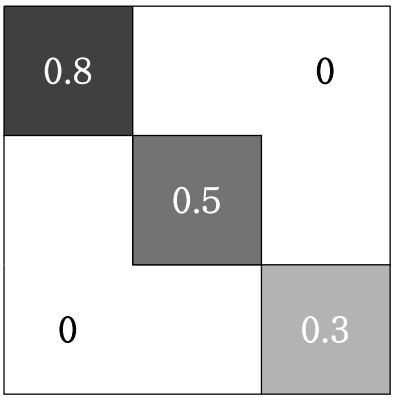
행렬A



행렬B



행렬C



행렬D

[그림 3] 변수 그룹 내 상관계수에 따른 상관행렬

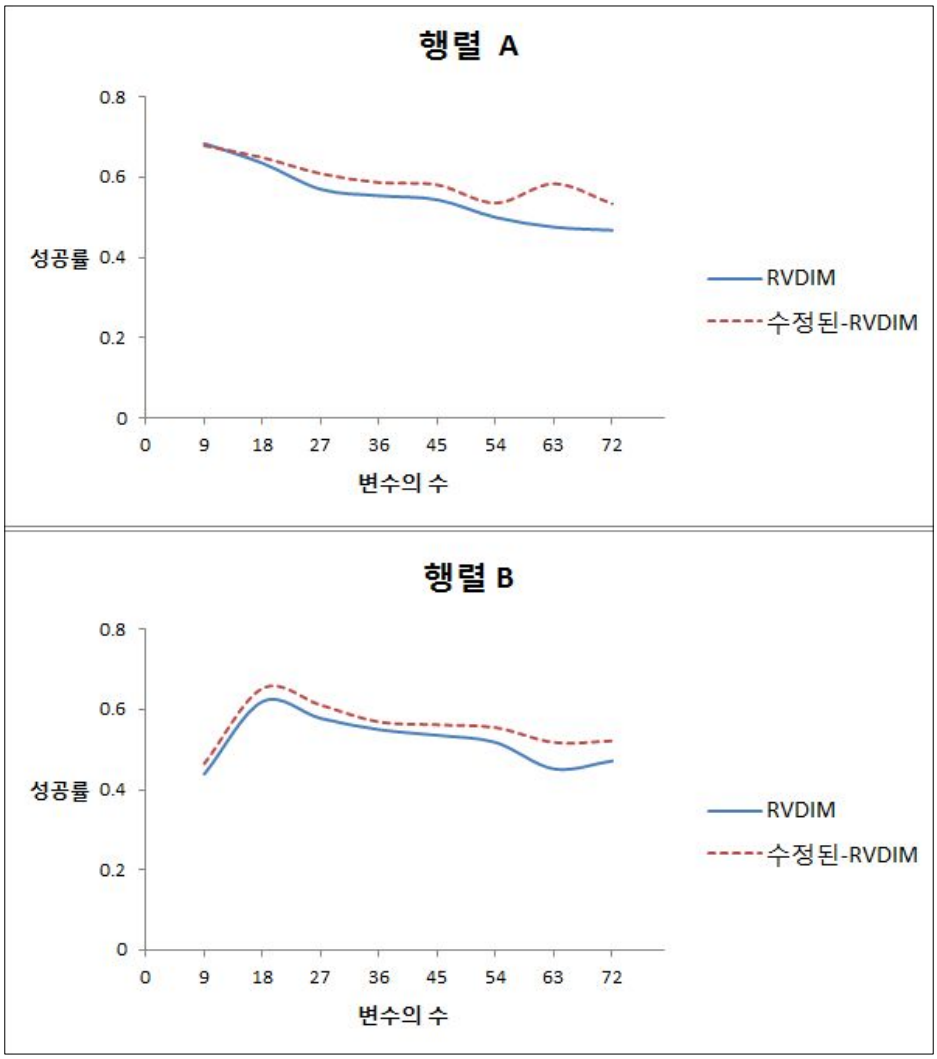
평균은 0, 분산은 1인 표준정규분포로부터 생성된 30×1의 열벡터와 각각 9, 18, 27, 36, 45, 54, 63, 75개의 변수로 이루어진 8개의 자료행렬을 생성하였다. 8개의 자료행렬과 4개의 상관계수 행렬의 조합을 통해 총 32(8×4)개의 시나리오를 설정하였다. 모의실험에서 고려한 상관행렬은 [그림 3]과 같다(즉, 4개의 상관행렬을 갖는 다변량 표준정규분포를 생성하였다.). 모든 상관행렬의 변수그룹을 3개로 두고 각 그룹 내 변수의 수는 모두 동일하게 설정하였다(예를 들어, 변수의 수가 36개인 시나리오에서는 변수그룹들은 각각 12개의 변수로 이루어짐). 32개의 시나리오에서 주성분 개수는 3개로 고정하였다. 변수 그룹 내 상관계수는 각각 0.8, 0.5, 0.3 이거나 모두 0.8, 0.5, 0.3으로 두고 변수 그룹 간 상관계수는 모두 0을 갖도록 정하였다(즉, 변수 그룹들은 서로 독립이다.).

각 시나리오에서 순열(permutation)의 수는 999를, 유의수준 α는 0.05로 정하고 1000개의 표본을 생성하였다. RVDIM과 수정된 RVDIM의 효율성을 비교하기위해 추정된 차원의 수가 3인 경우를 성공으로 정의하고 성공의 횟수를 표본의 수로 나눈 성공률로 평가하였다. 즉, 성공률=(성공의 횟수/1000)으로 계산하였고 각 시나리오별 성공률의 결과는 [표 1]와 같다.

변수의 수가 많아질수록 RVDIM과 수정된 RVDIM 모두 성공률은 감소하는 것을 확인할 수 있었다. 즉, 두 방법모두 주성분의 개수를 추정하는데 있어서 변수의 수가 많을수록 수행능력은 감소했다. 또한 대부분의 시나리오에서 수정된 RVDIM의 성공률은 RVDIM에 비해 높은 것을 확인할 수 있다. 상관행렬 C보다 상관행렬 B가, 상관행렬 B 보다는 상관행렬 A를 갖는 시나리오의 성공률이 높았다. 다시 말해서, 변수 그룹 내 상관계수가 높을수록 성공률이 높은 것을 알 수 있었다. 그러나 상관행렬 B의 경우보다 C의 경우의 성공률이 높은 경우도 발생하였다. 반면, 상관행렬 D의 경우는 차원수가 증가하여도 나머지 3가지 행렬에 비해 확연히 낮은 성공률을 보였다. 그러므로 3개의 변수 그룹 내 상관계수가 서로 상이할수록 RVDIM과 수정된 RVDIM의 수행능력은 현저히 떨어지는 것을 알 수 있다.

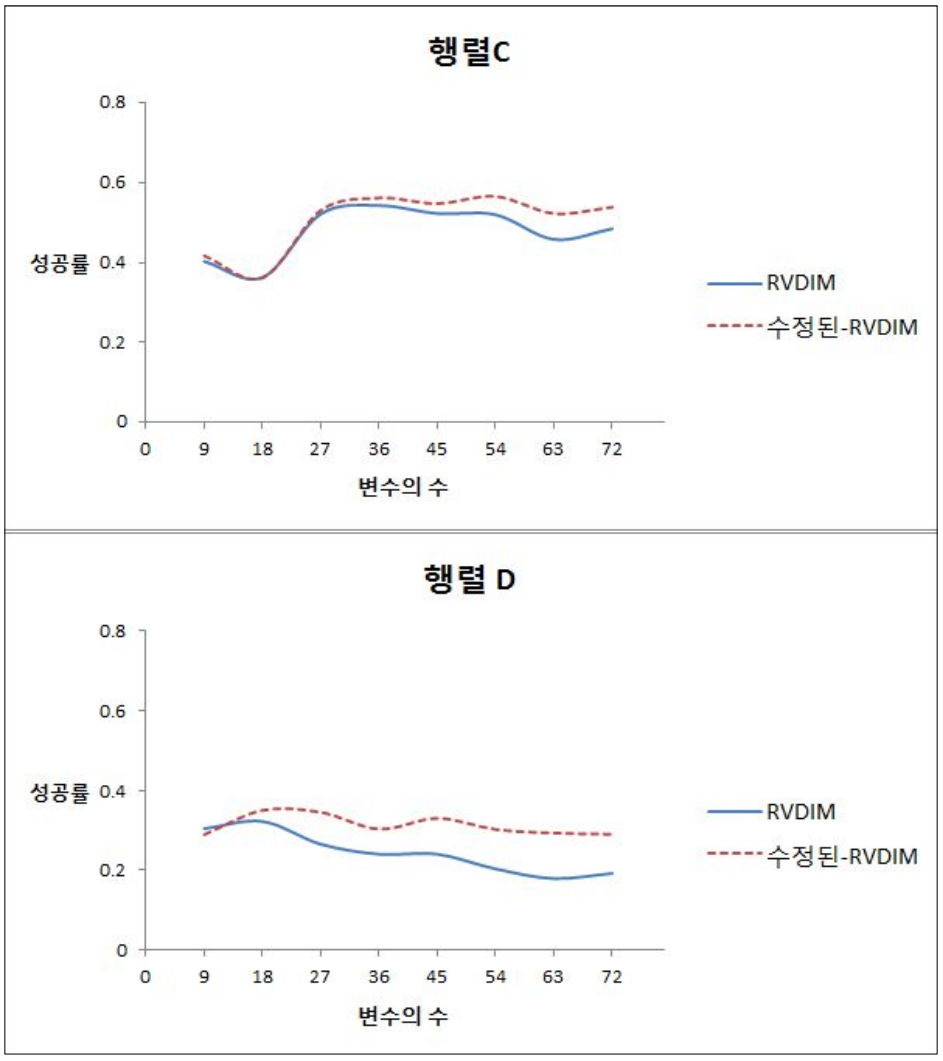
[표 1] 변수의 수에 따른 성공률

	변수의 수=9			
	A행렬	B행렬	C행렬	D행렬
RVDIM	0.684	0.439	0.403	0.304
수정된-RVDIM	0.679	0.465	0.417	0.289
	변수의 수=18			
RVDIM	0.635	0.620	0.363	0.322
수정된-RVDIM	0.649	0.653	0.362	0.350
	변수의 수=27			
RVDIM	0.570	0.578	0.522	0.265
수정된-RVDIM	0.609	0.611	0.531	0.345
	변수의 수=36			
RVDIM	0.554	0.55	0.543	0.240
수정된-RVDIM	0.587	0.569	0.562	0.303
	변수의 수=45			
RVDIM	0.544	0.536	0.523	0.240
수정된-RVDIM	0.581	0.562	0.548	0.330
	변수의 수=54			
RVDIM	0.500	0.518	0.52	0.203
수정된-RVDIM	0.536	0.555	0.566	0.302
	변수의 수=63			
RVDIM	0.476	0.452	0.458	0.179
수정된-RVDIM	0.584	0,518	0.523	0.293
	변수의 수=72			
RVDIM	0.468	0.472	0.485	0.192
수정된-RVDIM	0.534	0.522	0.539	0.290



[그림 4] 변수의 수에 따른 성공률(상관행렬A, B)

[그림 4], [그림 5]는 상관행렬의 형태와 변수의 수에 따른 시나리오들의 성공률을 나타낸 그래프이다. 실선은 RVDIM을 점선은 수정된 RVDIM을 나타내는데, 전반적으로 RVDIM에 비해 수정된 RVDIM의 성공률이 높고 변수의 수가 커질수록 두 계수간의 성공률 차이는 커진다. 특히 상관행렬이 D일 경우



[그림 5] 변수의 수에 따른 성공률(상관행렬C, D)

에는 성공률의 차이가 크지만 낮은 성공률을 보인다. 결론적으로 차원의 수가 높을수록 변수그룹내 상관계수가 높을수록 수정된 RVDIM이 RVDIM에 비해 주성분의 개수를 추정하는데 있어서 좀 더 안정된 결과를 보인다는 것을 확인할 수 있다.

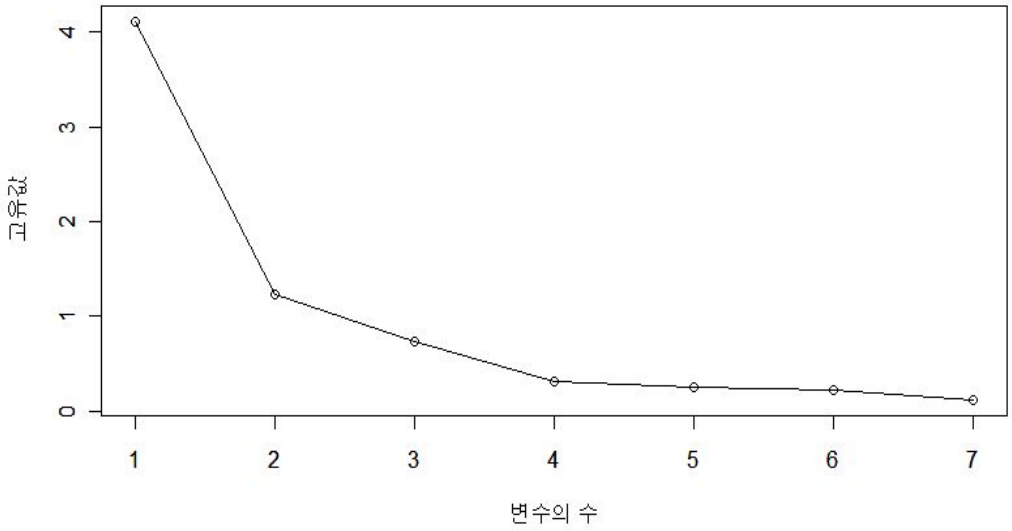
제 5 장 실제 데이터의 적용

본 장에서는 RVDIM과 수정된 RVDIM을 이용한 주성분의 개수 선택방법을 실제 데이터에 적용해봄으로서 두 지표의 수행능력을 평가하고자 한다. 분석에 사용된 데이터는 1977년 미국의 50개 주에서 발생한 7개 범행인 살인(murder), 강간(rape), 강도(robbery), 습격(assault), 불법목적주거침입(burglary), 절도(larceny), 자동차절도(auto theft)의 10만 명당 범죄율로 이루어져 있다(50×7). 원데이터에 대한 주성분분석의 결과는 다음과 같다(단, 분석시 원데이터는 표준화 하여 분석을 실시하였다.).

[표 2] 상관행렬을 통해 구한 고유값과 총 누적분산비율

변수의 수	고유값	비율	누적비율
1	4.11	59%	59%
2	1.24	18%	76%
3	0.73	10%	87%
4	0.32	5%	91%
5	0.26	4%	95%
6	0.22	3%	98%
7	0.12	2%	100%

[표 2]는 자료행렬의 상관행렬을 통해 구한 고유값과 총 누적분산비율을 나타내고 있다. 제 2 주성분의 고유값은 1보다 크나 제 3 주성분의 고유값은 0.73으로서 카이저규칙에 의해서는 2개의 주성분을 선택할 수 있다. 누적비율을 살펴보면 제 2 주성분은 원데이터의 76%를, 제 3 주성분은 87%를 설명하



[그림 6] 실제 데이터에 대한 산비탈 그림

므로 총 분산의 누적점유율방법에 의하면 2개 또는 3개의 주성분을 선택할 수 있다.

[그림 6]은 [표 2]의 고유값들을 나타낸 산비탈 그림을 나타내는데, 변수의 수가 2개 일 때 즉, 제 2 고유값의 위치에서 급격히 꺾이는 것을 알 수 있다. 그러므로 산비탈 그림을 통해서는 2개의 주성분을 선택할 수 있다.

[표 3] RVDIM과 수정된 RVDIM의 p값

	1	2	3	4	5	6	7
RVDIM	0.001	0.001	0.001	0.228	0.260	0.030	0.873
수정된 RVDIM	0.001	0.001	0.001	0.315	0.356	0.033	0.849

실제데이터를 이용하여 순열검정을 통해 RVDIM과 수정된RVDIM의 p값을 구한 결과는 [표 3]과 같다. RVDIM과 수정된RVDIM 모두 3번째 축(즉, 주성분)의 p값은 유의수준 0.05보다 작으나 4번째 축의 p값은 모두 0.05보다 크므로 우리는 3개의 주성분을 선택할 수 있다.

변수의 수가 개체수보다 많은 경우($n < p$)에 대해 적절한 실제 자료 획득에 어려움이 있어 적용하지 못하였으나, 기존의 알려진 자료에서 관측지를 임의추출하여 자료를 모의실험한 결과, 변수의 수가 관측치보다 작은 경우에 비해 효율성이 RVDIM, 수정된 RVDIM 모두 떨어지는 것으로 나타났다.

제 6 장 결 론

본 논문에서는 개체수가 변수의 수보다 많은 고차원의 두 자료행렬간 상관정도를 나타내는 수정된 RV계수를 이용하는 수정된 RVDIM을 통해 축소될 차원수를 추정하고 이를 검정하는 방법을 제안하였다. 모의실험을 통해 기존에 제안된 RVDIM과 본 논문에서 제안한 수정된 RVDIM을 모의실험과 실제자료에 적용시켜봄으로서 향상된 정도를 파악한 결과, 수정된 RVDIM을 이용한 검정법이 변수의 수가 관측치의 수보다 큰 경우 기존의 방법보다 우수함을 확인하였다.

모의실험시 변수 그룹 내 상관계수가 변수 그룹간에 서로 다른 값을 갖는 경우에 RVDIM과 수정된 RVDIM의 성공률이 유난히 낮았던 것을 알 수 있다. 그러므로 변수 그룹내 상관이 변수 그룹간에 다를 경우 두 방법을 이용한 주성분의 개수추정은 불안정한 결과를 도출하는 것을 알 수 있다. 변수그룹간의 상관계수는 0인 경우에 대해서만 고려하였기 때문에 변수그룹간의 상관이 존재하는 경우에도 성공률이 낮을 것으로 예상된다. 추후 분석에서는 더 다양한 상관행렬을 이용한 분석 및 단점을 보완할 수 있는 방법이 요구된다. 또한, RV계수뿐만 아니라 다른 행렬간의 상관성 척도를 통해 주성분의 개수를 추정하는 방법에 대해서도 생각해 볼 수 있을 것이다.

RV계수의 불안정함이 수정된 RVDIM에서는 안정되는 모습을 모의실험 결과 보였으나 실제데이터 적용에서는 다소 불안정한 결과를 도출하였으므로 이에 대한 보완이 필요할 것으로 보인다.

참 고 문 헌

[1] 권세혁 (2008), *다변량데이터 분석과 활용*, 자유아카데미, 서울

[2] Bartlett, M. S. (1954). A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological),* 16, 296 - .298.

[3] Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research,* 1, 245-276.

[4] Dray, S., Chessel, D., Thioulouse, J. (2003). Co-inertia analysis and the linking of ecological data tables. *Ecology,* 84, 3078 - 3089.

[5] Dray, S. (2008). On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *computational Statistics & Data Analysis,* 52, 2228 - 2237.

[6] Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics* 29, 750 - .760.

[7] Good, I. (1969). Some applications of the singular decomposition of a matrix. *Technometrics*, 11, 823 - 831.

[8] Gower, J. (1971). *Statistical methods of comparing different multivariate analyses of the same data.* In: Hodson, F., Kendall, D., Tautu, P. (Eds.), Mathematics in the Archaeological and Historical Sciences. Edinburgh University Press, Edinburgh, pp. 138 - 149.

[9] Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179 - 185.

[10] Jackson, J. E. (1991). *A User's Guide to Principal Components.* Wiley, NewYork.

[11] Jolliffe, I. T. (2002). *Principal Component Analysis*. 2nd Edition. Springer, NewYork.

[12] Lawley, D. N. (1956). Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika*, 43, 128-136.

[13] Lingoes, J., Schonemann, P. (1974). Alternative measures of fit for the Schonemann - .Carrol matrix fitting algorithm. *Psychometrika*, 39, 423 - 427.

[14] Peres-Neto, P., Jackson, D., Somers, K. (2005). How many principal components? stopping rules for determining the number of non-trivial axes revisited. Computational Statistics & Data Analysis, 49, 974 - 997.

[15] Pimentel, R. A. (1979). *Morphometrics: the multivariate analysis of biological data*. Kendall-Hunt, Dubuque, Iowa, USA.

[16] Smilde, A. K., Kiers, H. A. L., Bijlsma, S., Rubingh, C. M., van Erk, M. J. (2009). Matrix correlations for high-dimensional data: the modified RV-coefficient. *Bioinformatics,* 25, 401 - 405.

ABSTRACT

A study on the selection of the number of principal components by using modified RV-coefficient

SOYOUNG KIM
Department of Statistics
The Graduate School
Sungshin Women's University

Principal Component Analysis(PCA) is one of the multivariate techniques that analyze a data in which observations are described by several inter-correlated quantitative variables. As the purpose of PCA is determining the number of non-trivial principal components, many methods to estimate the number of non-trivial axes has been proposed.

In this study, we suggested a new technique to evaluate the dimensionality in PCA. It is based on a similarity measurement, singular value decomposition(SVD), and permutation procedures. The RV-coefficient was introduced as a measure of similarity between matrices and as a theoretical tool to analyze multivariate techniques. Modified RV-coefficient can be used in high-dimensional data analysis studies as a measurements of common information of two datasets. The technique

proposed used the RV-coefficient as the similarity measure between data matrix and residual matrix of original data.

The simulation study and real application showed that the suggested method based on modified RV-coefficient is more accurate than based on th RV-coefficient when the number of variables is lager than observations ($n < p$).