Journal of Korean Institute of Intelligent Systems
Vol. 25, No. 6, December 2015, pp. 592-599
http://dx.doi.org/10.5391/JKIIS.2015.25.6.592

ISSN(Print) 1976-9172
ISSN(Online) 2288-2324

JKIIS

주성분 분석을 이용한 빅데이터 분석

Big Data Analysis Using Principal Component Analysis

이승주*†
Seung-Joo Lee†

*청주대학교 통계학과
† Department of Statistics, Cheongju University

요 약

빅 데이터 환경에서 빅데이터를 분석하기 위한 새로운 방법의 필요성이 대두되고 있다. 데이터의 크기, 다양성, 그리고 적재 속도 등의 빅데이터 특성으로 인해 모집단의 추론에서 전체 데이터의 분석이 가능해졌기 때문이다. 그러나 전통적인 통계분석 방법은 모집단으로부터 추출된 확률표본에 초점이 맞추어져 있다. 따라서 기존의 통계적 접근방법은 빅데이터 분석에 적합하지 않은 경우가 발생한다. 이와 같은 문제점을 해결하기 위하여 본 논문에서는 빅데이터분석을 위한 새로운 접근방법에 대하여 제안하였다. 특히 대표적인 다변량 통계분석 기법인 주성분 분석을 이용하여 효율적인 빅데이터분석을 위한 방법론을 연구하였다. 제안방법의 성능평가를 위하여 통계적 모의실험을 실시하였다.

키워드 : 빅데이터, 주성분분석, 고유치, 빅데이터분석, 통계분석

Abstract

In big data environment, we need new approach for big data analysis, because the characteristics of big data, such as volume, variety, and velocity, can analyze entire data for inferring population. But traditional methods of statistics were focused on small data called random sample extracted from population. So, the classical analyses based on statistics are not suitable to big data analysis. To solve this problem, we propose an approach to efficient big data analysis. In this paper, we consider a big data analysis using principal component analysis, which is popular method in multivariate statistics. To verify the performance of our research, we carry out diverse simulation studies.

Key Words : Big Data, Principal Component Analysis, Eigenvalue, Big Data Analysis, Statistical Analysis

Received: May. 27 2015
Revised : Jun. 16, 2015
Accepted: Jun. 19, 2015
†Corresponding author
access@cju.ac.kr

이 논문은 2014-2015학년도에 청주대학교 산업과학연구소가 지원한 학술연구조성비(특별연구과제)에 의해 연구되었음.

1. 서 론

Pearson[1]이 창안한 주성분 분석은 100년 이상 오래된 수학적 기법으로 주성분의 해석뿐만 아니라 그 성질들이 많은 연구자들에 의해 연구되었다[2-4]. 주성분 분석은 통계학뿐만 아니라 컴퓨터 과학, 경제학, 심리학, 행동과학, 생태학, 기상학, 유전학 등 매우 다양한 분야에서 응용되고 있다. 최근에는 cDNA 마이크로어레이 실험에서 얻어진 데이터의 분석에도 응용되고 있다[5].

주성분 분석은 서로 상관되어 있는 p개의 변수 집합을 선형변환하여 주성분이라고 부르는 서로 상관되어 있지 않은 m개의 새로운 인공변수들을 도출한다. 이때 m개의 주성분이 전체 변이중 가능한 많은 양의 변이를 설명하도록 변환시킴으로써 p차원 변이를 m차원으로 차원 축소할 수 있다[4][6]. 차원 축소의 목적은 분석이나 해석을 쉽게 하고 데이터에 있는 대부분의 변이나 정보를 보유해야 한다. 주성분 분석을 수행할 때 가장 중요한 문제는 추출할 주성분의 수를 결정하는 것이다[7]. 보유할 적절한 주성분의 수 m을 결정하기

위한 많은 방법들이 휴리스틱 방법과 통계적 방법으로 제안되었으며 현재도 많은 연구자들이 연구하고 있다. 몇 가지 방법은 계산하기 쉽지만 어떤 방법들은 과도한 연산 작업이 필요하다. 보유할 주성분의 수 m을 결정하기 위한 방법으로는 Bartlett의 카이제곱 검정[8-9], Kaiser 규칙[10], Cattle의 스크리(scree) 검정[11], 병렬 분석(parallel analysis)[12], MAP 검정[13], 총분산비 등 매우 많은 방법들이 존재하지만 이러한 기준들은 동일한 결과를 제공하지는 않는다[4][6-7].

데이터에서 보유할 주성분의 수를 결정하는 것은 연구자가 해야 할 가장 중요한 의사결정중 하나이다. 그러나 분석자가 분석을 위해 선택한 데이터가 빅 데이터로 데이터가 너무 방대해서 분석이나 데이터 마이닝 절차를 진행하는데 어려움이 있다면 분석이나 마이닝 결과를 나쁘게 하지 않으면서 데이터 집합의 크기를 줄이는 방법으로 데이터 축소(data reduction) 방법이 사용된다[14].

본 논문에서는 빅 데이터를 사용하여 주성분의 수를 결정할 때 전체 데이터를 사용하지 않고 표본추출 방법을 사용하여 데이터의 양을 축소하는 수량축소를 먼저 수행하고 축소된 데이터를 이용하여 보유할 주성분의 수를 결정하는 방법을 제안하고자 한다. 따라서 본 논문에서는 먼저, 주성분의 수를 결정하는 몇 가지 방법을 간단히 고찰하고, 몬테칼로 모의실험을 통하여 표본추출방법을 사용하여 수량축소를 한 후 주성분의 수를 결정하는 방법의 성능을 평가하고자 한다.

2. 표본추출 방법을 이용한 주성분 분석

빅데이터는 크기(volume), 다양성(variety), 그리고 속도(velocity)로 특징되며 특허, 항만교통, 내트워크 패킷 등 다양한 분야에서 생성, 분석되고 있다 [15-17]. 또한 기존의 다양한 방법과 결합되어 새로운 방법론을 제공하고 있다 [18]. 따라서 빅데이터의 분석은 모든 분야에서 중요한 이슈로 떠오르고 있다 [19-20]. 전통적으로 데이터분석은 통계학에 기반하고 있다. 통계적 분석은 모집단으로부터 추출된 표본에 기반한 분석을 수행하기 때문에 소규모 표본데이터의 분석에 적합하다. 그러므로 빅데이터 분석을 위해 기존의 통계분석 기법의 이와 같은 문제점을 해결해야 한다. 본 논문에서는 표본추출방법을 이용하여 빅데이터분석에서 발생하는 전통적인 통계분석의 문제점을 해결하였다. 특히 여러 분야에서 다양하게 사용되는 다변량 통계분석 기법인 주성분분석을 빅데이터에 적용하였다. 즉, 빅 데이터로부터 보유할 주성분의 수를 직접 추정하지 않고 데이터의 양을 축소하는 수량축소 방법으로 빅 데이터의 양을 축소한 후 주성분의 수를 추정하는 방법을 제안하고자 있다. 수량축소 방법은 표본추출 방법을 고려할 수 있다. 표본추출 방법은 단순확률추출법, 층화확률추출법, 계통추출법, 집락추출법 등[21] 여러 가지 추출 방법이 있으나 본 연구에서는 비복원 단순확률추출법을 사용하여 빅 데이터로부터 표본을 추출하고 표본을 통해서 얻을 수 있는 정보의 양을 조절하여 주성분의 수를 추정하고자 한다. 빅 데이터로부터 비복원 단순확률추출법을 사용하여 작은 표본을 추출하고 추출된 표본을 사용하여 주성분의 수를 결정하고 전체 빅 데이터를 분석할 때 사용할 수 있는 순서도는 그림 1과 같다.

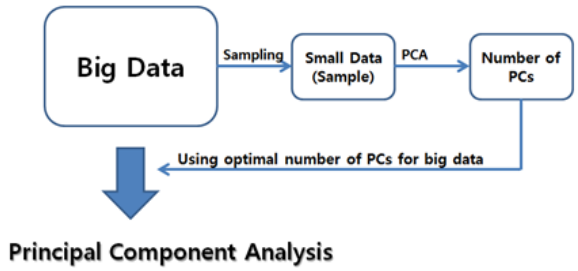

그림 1. 주성분의 수를 결정하기 위한 순서도
Fig. 1. Flow chart of determining the number of principal components

3. 주성분의 수 결정 기준

보유할 주성분의 수를 결정하기 위해 많은 기준들이 제안되었다[8-13]. 이러한 기준들은 동일한 결과를 제공하지는 않는다[12][22]. 본 절에서는 주성분의 수를 결정하는 몇 가지 방법을 간단히 고찰하고자 한다. 이 방법들은 Bartlett의 카이제곱 검정, Kaiser 규칙(K1), 스크리(scree) 검정, 병렬분석(parallel analysis; PA) 방법, MAP(minimum average partial) 검정 등 널리 사용되는 방법이다.

3.1 Bartlett의 카이제곱 검정

Bartlett(1950, 1951)는 나머지 $p-m$ 개의 고유치가 모두 동일하다는 귀무가설의 통계적 검정을 개발하였다[8-9]. 각각의 고유치는 근사 카이제곱 검정 결과 귀무가설이 채택될 때까지 순차적으로 제거되며 제거한 m개 고유치에 대응하는 주성분을 보유하게 된다. Bartlett의 카이제곱 검정은 표본 크기에 민감한 것으로 알려져 있다. 표본 크기가 증가함에 따라서 모집단 고유치들의 추정치는 더 정확하게 된다. Zwick과 Velicer(1982)는 Bartlett의 카이제곱 검정이 소표본에서 보다 대표본에서 더 정확함을 보였다[22].

3.2 Kaiser 규칙

주성분분석에서 보유할 주성분의 수를 결정하기 위해 가장 많이 사용하는 일반적인 규칙은 1보다 큰 고유치를 사용하는 Kaiser(1960)의 규칙(K1)이다[4][10][23]. 변수들의 측정단위가 다른 경우 대부분의 주성분분석에서 상관행렬을 사용한다. 만약 주성분분석이 상관행렬에 기초한다면, 고유치들의 합은 변수의 수와 같게 되며, 상관행렬의 대각원소는 1이므로 주성분의 분산인 고유치들의 평균은 1이 된다. 따라서 1보다 작은 고유치를 갖는 주성분은 원래의 반응변수 중 어느 하나의 변수 보다 더 작은 정보를 가지므로 보유할 가치가 없게

된다. 따라서 Kaiser의 규칙은 고유치가 1보다 큰($\lambda > 1.0$) 주성분만 분석에 사용한다[4]. 많은 연구자들은 Kaiser의 규칙이 보유된 주성분의 수를 종종 과대추정(overestimate)함을 보였다[12][22]. Kaiser(1960)와 Gorsuch(1983)는 Kaiser의 규칙에 의해 보유된 주성분의 수는 $p/3$와 $p/5$ 또는 $p/6$ 사이에 있어야 한다고 보고하였으며[10][24] Zwick과 Velicer (1982)는 모의실험을 통하여 이 결과를 지지하였다[22].

3.3 Scree 검정

Cattle(1966)은 인자분석에서 공통인자의 수를 결정하기 위해 고유치들의 그래프를 이용하는 스크리 검정(scree test)을 제안하였다[11]. 그러나 스크리 검정은 주성분분석에서 널리 사용되고 있다. 고유치 (k,λ_k), $k=1,2,\cdots,p$들을 플롯하고 점을 연결한 기울기가 k값 왼쪽은 가파르고 오른쪽은 완만하면 이 k값을 보유할 주성분의 수 m으로 결정하는 규칙이다[4].

3.4 MAP 검정

Velicer(1976)는 편상관행렬을 이용한 방법을 제안하였다. Velicer(1976)가 제안한 MAP(Minimum Average Partial) 검정은 첫 m개 주성분이 주어졌다는 조건하에서 p개 변수들 간의 편상관계수들의 제곱의 평균을 계산하고 이 값의 최소값에 해당하는 m을 보유할 주성분의 수로 하는 것이다. 편상관계수들의 제곱의 평균은 잔차행렬이 항등행렬과 거의 유사할 때 최소가 된다[13]. Zwick과 Velicer(1982)는 MAP 규칙은 K1이나 Bartlett 검정보다 알려진 성분의 수를 식별하는데 더 정확하다고 보고하였다[22].

3.5 병렬 분석

Horn(1965)은 Kaiser 규칙의 대안으로 보유할 주성분의 수를 결정하기 위한 병렬 분석(parallel analysis; PA) 방법을 제안하였다[12]. 이 방법은 분석할 데이터 행렬에서 구한 고유치들과 랜덤하게 생성된 데이터 집합에서 구한 고유치들을 비교하여 주성분의 수를 결정하는 방법이다. 병렬분석은 분석할 $n\times p$ 데이터 집합에서 $p\times p$ 표본상관행렬의 고유치를 계산한다. p개의 독립인 정규변량에서 임의로 추출된 $n\times p$ 모의실험 데이터 집합에 대해 $p\times p$ 상관행렬을 계산하고 고유치를 구하여 크기순으로 나열하며 이 단계를 r번 반복하고 r개 고유치들의 중앙값이나 평균을 구한다. 이때 관측 데이터의 고유치들과 모의실험 데이터의 고유치들의 평균을 비교하여 관측 데이터의 고유치가 모의실험 데이터의 고유치보다 크면 주성분으로 보유한다[12][25]. 모의실험한 랜덤 데이터에서 구한 고유치들의 임계값은 랜덤 데이터에서 도출한 고유치들의 분포의 95 백분위수를 사용할 것을 추천하고 있다[26].

4. 모의실험 방법

본 연구의 목적은 빅 데이터에서 작은 표본을 추출하여 보유할 주성분의 수를 결정하는 방법을 몬테칼로 방법을 이용하여 보이고자 한다. 이러한 목적을 위해 주성분의 수를 결정하는 네가지 방법을 몬테칼로 모의실험을 수행하여 비교하고자 한다. 주성분의 수를 결정하는 방법의 정확도는 정확한 주성분의 수를 추정하는 방법을 효과적으로 평가함으로써 결정된다. 주성분의 수 m을 결정하는 방법은 표본 크기(n), 변수의 수(p), 주성분 적재의 크기인 포화도(ℓ_{ij}) 등에 영향을 받는 것으로 알려져 있다[4][7]. 모집단 주성분 적재 행렬에서 각 주성분은 변수의 수를 같게 하였으며 영이 아닌 주성분 적재의 크기는 동일하고 나머지 적재의 크기는 영으로 하였다.

변수의 수(p)는 36과 72로 고정하였다. 이러한 변수의 수는 작은 변수 집합과 큰 변수 집합을 나타내기 위해 사용하였다. 큰 변수 집합은 MAP와 Bartlett 검정에 긍정적인 영향을 미치며 Kaiser 규칙에는 부정적인 영향을 미친다[22].

데이터 또는 표본 크기(n)는 100,000으로 고정하였다. 100,000의 값은 빅 데이터를 나타내기 위해 선택하였다. 이 표본에서 단순확률추출법을 사용하여 비복원추출하며 각 표본에서 1%, 2%, 3%, 4%, 5%를 표본 추출하여 주성분의 수를 추정하는데 사용하였다.

모집단 주성분의 수 m은 3, 6, 9와 12를 사용하였다. 즉, 변수의 수 p가 36일 때에는 m은 3, 6, 9를 사용하며, p가 72일 때에는 m은 3, 6, 9, 12를 사용하였다. 따라서 p가 36인 경우 각 주성분당 변수의 수($p:m$)는 4, 6, 12 이며 p가 72인 경우 각 주성분당 변수의 수는 6, 8, 12, 24가 된다. 본 연구에서는 하나의 주성분을 정의하는데 필요한 변수의 수를 최소 4개로 정하였으며 $p:m$이 4인 경우는 하나의 주성분을 정의하는데 필요한 변수의 수가 4개를 의미한다.

관측 변수와 각 주성분 사이의 상관계수를 나타내는 주성분 적재의 크기인 포화도는 0.4, 0.6과 0.8을 사용하였다. 여기서 0.4는 작은 적재, 0.6은 중간 정도의 적재를 나타내고 0.8은 높은 적재를 나타내기 위해 사용하였다.

데이터 생성을 위해 먼저 모집단 상관행렬을 변수의 수, 주성분의 수와 주성분 적재 크기의 각 조합에 대하여 생성하였다. 각각의 모집단 상관행렬은 다음과 같이 생성하였다. 모집단 주성분 적재행렬 L은 변수의 수, 모집단 주성분의 수와 적재의 크기에 따라 생성하고 L의 전치행렬을 곱하여 LL'를 구하고 LL'의 대각원소를 1로 대체하여 모집단 상관행렬 R을 생성하였다. 이와 같이 생성된 모집단 상관행렬을 이용하여 다변량 정규분포로부터 $n=100{,}000$개 데이터를 생성하여 $n\times p$ 데이터 행렬을 만들었다. 이 데이터 행렬로부터 단순확률추출법을 사용하여 비복원추출하여 주성분의 수를 추정하는데 사용하였다. 따라서 추출된 1%, 2%, 3%, 4%, 5% 표본(s)에 대하여 4가지 방법(Bartlett 검정, K1, MAP, PA)에 따라 주성분의 수를 계산하고 이와 같은 추정 단계를 100번 반복하여 주성분의 수를 100번 추정하였다. Bartlett 검정은 두 가지 유의수준 0.05와 0.01에서 검정(Bart5, Bart1)하였으며, 병렬분석은 p와 s에서 100번 반복한 표본상관행렬로부터 고유치를 구하여 5, 50, 95 백분위수(PA05, PA50, PA95)에서

추정하였다. 몬테칼로 모의실험 도구는 R 3.1.2를 사용하였다[27].

5. 모의실험 결과

본 논문에서 제안된 모의실험 방법을 이용하여 4가지 기준에 따라 보유할 주성분의 수를 100번 추정하였다. 이 결과로부터 각 기준의 추정값의 평균을 구하여 모집단 주성분의 수 m으로부터의 평균 차이 d를 계산하였다. 따라서 평균 차이 d가 양수이면 주성분의 수를 과대추정, d가 음수이면 과소추정을 의미한다. 모의실험 결과는 표 1에서 표 3과 같다. 각각의 표는 변수의 수 p, 모집단 주성분의 수 m, 주성분당 변수의 수 $p:m$, 표본 크기 s, 추출률 $s:n$과 4가지 기준에 따라 보유할 주성분의 수를 추정한 평균 차이 d를 나타낸다. 표 1에서 표 3의 형식은 동일한 형식으로 표를 작성하였으며 단지 포화도만 차이가 있다. 따라서 표 1의 모의실험 결과만을 자세히 고찰하고자 한다.

표 1의 첫 번째 행은 $n=100{,}000$, $p=36$, $m=3$, $p:m=12$이고 $\ell_{ij}=0.4$일 때, n개 데이터에서 $s=1{,}000$개의 표본을 추출하면 추출률은 $s:n=0.01$이므로 100,000개

표 1. 모집단 주성분의 수로부터의 평균 차이 ($n=100{,}000$, 포화도=0.4)

Table 1. Mean difference from the number of population principal components ($n=100{,}000$, saturation=0.4)

Pattern	n	p	m	$p:m$	s	$s:n$	Bart5	Bart1	K1	MAP	PA05	PA50	PA95
1	100,000	36	3	12	1,000	0.01	0.02	0.00	5.90	0.00	0.00	0.00	0
2					2,000	0.02	0.00	0.00	2.72	0.00	0.00	0.00	0
3					3,000	0.03	0.04	0.01	0.80	0.00	0.00	0.00	0
4					4,000	0.04	0.02	0.00	0.02	0.00	0.00	0.00	0
5					5,000	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0
6			6	6	1,000	0.01	0.00	0.00	4.60	-5.99	0.00	0.00	0
7					2,000	0.02	0.00	0.00	1.82	-6.00	0.00	0.00	0
8					3,000	0.03	0.03	0.00	0.31	-6.00	0.00	0.00	0
9					4,000	0.04	0.03	0.00	0.00	-6.00	0.00	0.00	0
10					5,000	0.05	0.03	0.00	0.00	-6.00	0.00	0.00	0
11			9	4	1,000	0.01	-0.13	-0.35	3.45	-9.00	0.05	0.01	0
12					2,000	0.02	0.01	0.00	1.03	-9.00	0.00	0.00	0
13					3,000	0.03	0.01	0.00	0.09	-9.00	0.00	0.00	0
14					4,000	0.04	0.03	0.00	0.00	-9.00	0.00	0.00	0
15					5,000	0.05	0.01	0.00	0.00	-9.00	0.00	0.00	0
16		72	3	24	1,000	0.01	0.02	0.01	17.79	0.00	0.00	0.00	0
17					2,000	0.02	0.00	0.00	12.89	0.00	0.00	0.00	0
18					3,000	0.03	0.01	0.00	9.02	0.00	0.00	0.00	0
19					4,000	0.04	0.00	0.00	6.29	0.00	0.00	0.00	0
20					5,000	0.05	0.01	0.00	3.93	0.00	0.00	0.00	0
21			6	12	1,000	0.01	0.01	0.00	16.43	0.00	0.00	0.00	0
22					2,000	0.02	0.00	0.00	11.70	0.00	0.00	0.00	0
23					3,000	0.03	0.02	0.01	5.07	0.00	0.00	0.00	0
24					4,000	0.04	0.02	0.00	5.43	0.00	0.00	0.00	0
25					5,000	0.05	0.00	0.00	3.18	0.00	0.00	0.00	0
26			9	8	1,000	0.01	0.00	0.00	14.90	-0.11	0.00	0.00	0
27					2,000	0.02	0.00	0.00	10.45	0.00	0.00	0.00	0
28					3,000	0.03	0.01	0.00	7.13	0.00	0.00	0.00	0
29					4,000	0.04	0.02	0.00	4.57	0.00	0.00	0.00	0
30					5,000	0.05	0.02	0.00	2.56	0.00	0.00	0.00	0
31			12	6	1,000	0.01	-0.23	-0.52	13.34	-11.95	0.05	0.01	0
32					2,000	0.02	0.01	0.00	9.18	-12.00	0.00	0.00	0
33					3,000	0.03	0.00	0.00	6.00	-12.00	0.00	0.00	0
34					4,000	0.04	0.01	0.00	3.63	-12.00	0.00	0.00	0
35					5,000	0.05	0.01	0.00	1.75	-12.00	0.00	0.00	0

의 데이터에서 1%의 표본을 추출하여 각 방법의 평균차이 d를 추정한 결과이다. 따라서 5% 유의수준에서 검정한 Bartlett 검정 결과 평균차이는 0.02이므로 주성분의 수를 아주 약간 과대추정하며 1% 유의수준에서 검정한 Bartlett 검정 결과는 0.00이므로 정확히 추정하였다. Kaiser 방법의 평균차이는 5.90이므로 Kaiser의 방법은 주성분의 수를 매우 크게 과대추정하며, MAP 검정과 병렬분석의 5, 50, 95 백분위수(PA05, PA50, PA95) 추정치의 평균차이는 모두 0이므로 모집단 주성분의 수를 정확히 추정한다. 표 1의 11 번째 행은 $n=100{,}000$, $p=36$, $m=9$, $p:m=4$이고 $\ell_{ij}=0.4$일 때, $s=1{,}000$개의 표본을 추출하면 Bartlett 검정(-0.13, -0.35)과 MAP 검정(-9)은 주성분의 수를 매우 과소추정하며 Kaiser 방법(3.45)은 과대추정하고 병렬분석은 정확히 추정하는 경향이 있다. 모의실험 결과 $p=72$인 경우에도 비슷한 결과가 도출되었다. 따라서 포화도가 $\ell_{ij}=0.4$일 때 Bartlett 검정은 주성분의 수를 약간 과대추정하거나 과소추정하며 Kaiser 방법은 매우 과대추정하며, MAP 검정은 $p:m$이 6이하일 때 매우 과소추정하고 $p:m$이 6보다 크면 주성분이 수를 정확히 추정하며, 병렬분석은 주성분의 수를 거의 정확히 추정함을 알 수 있다.

표 2는 $\ell_{ij}=0.6$일 때 모의실험 결과에서 Bartlett 검정은 아주 사소하게 과대추정하며 Kaiser 방법은 $p=72$이고 $s:n$

표 2. 모집단 주성분의 수로부터의 평균 차이 ($n=100{,}000$, 포화도=0.6)

Table 2. Mean difference from the number of population principal components ($n=100{,}000$, saturation=0.6)

Pattern	n	p	m	$p:m$	s	$s:n$	Bart5	Bart1	K1	MAP	PA05	PA50	PA95
1	100,000	36	3	12	1,000	0.01	0.05	0.00	0	0	0	0	0
2					2,000	0.02	0.03	0.00	0	0	0	0	0
3					3,000	0.03	0.09	0.01	0	0	0	0	0
4					4,000	0.04	0.02	0.00	0	0	0	0	0
5					5,000	0.05	0.01	0.00	0	0	0	0	0
6			6	6	1,000	0.01	0.01	0.00	0	0	0	0	0
7					2,000	0.02	0.03	0.00	0	0	0	0	0
8					3,000	0.03	0.05	0.01	0	0	0	0	0
9					4,000	0.04	0.04	0.01	0	0	0	0	0
10					5,000	0.05	0.06	0.02	0	0	0	0	0
11			9	4	1,000	0.01	0.04	0.00	0	0	0	0	0
12					2,000	0.02	0.05	0.01	0	0	0	0	0
13					3,000	0.03	0.01	0.00	0	0	0	0	0
14					4,000	0.04	0.05	0.01	0	0	0	0	0
15					5,000	0.05	0.03	0.00	0	0	0	0	0
16		72	3	24	1,000	0.01	0.02	0.01	0.59	0	0	0	0
17					2,000	0.02	0.01	0.00	0.00	0	0	0	0
18					3,000	0.03	0.04	0.00	0.00	0	0	0	0
19					4,000	0.04	0.04	0.00	0.00	0	0	0	0
20					5,000	0.05	0.03	0.00	0.00	0	0	0	0
21			6	12	1,000	0.01	0.01	0.00	0.22	0	0	0	0
22					2,000	0.02	0.02	0.00	0.00	0	0	0	0
23					3,000	0.03	0.03	0.01	0.00	0	0	0	0
24					4,000	0.04	0.02	0.00	0.00	0	0	0	0
25					5,000	0.05	0.02	0.00	0.00	0	0	0	0
26			9	8	1,000	0.01	0.03	0.00	0.04	0	0	0	0
27					2,000	0.02	0.03	0.00	0.00	0	0	0	0
28					3,000	0.03	0.01	0.00	0.00	0	0	0	0
29					4,000	0.04	0.06	0.01	0.00	0	0	0	0
30					5,000	0.05	0.06	0.00	0.00	0	0	0	0
31			12	6	1,000	0.01	0.01	0.00	0.00	0	0	0	0
32					2,000	0.02	0.04	0.01	0.00	0	0	0	0
33					3,000	0.03	0.05	0.00	0.00	0	0	0	0
34					4,000	0.04	0.07	0.01	0.00	0	0	0	0
35					5,000	0.05	0.07	0.00	0.00	0	0	0	0

표 3. 모집단 주성분의 수로부터의 평균 차이 ($n=100,000$, 포화도=0.8)

Table 3. Mean difference from the number of population principal components ($n=100,000$, saturation=0.8)

Pattern	n	p	m	$p:m$	s	$s:n$	Bart5	Bart1	K1	MAP	PA05	PA50	PA95
1	100,000	36	3	12	1,000	0.01	0.10	0.02	0	0	0	0	0
2					2,000	0.02	0.06	0.00	0	0	0	0	0
3					3,000	0.03	0.14	0.05	0	0	0	0	0
4					4,000	0.04	0.11	0.01	0	0	0	0	0
5					5,000	0.05	0.08	0.00	0	0	0	0	0
6			6	6	1,000	0.01	0.07	0.02	0	0	0	0	0
7					2,000	0.02	0.05	0.00	0	0	0	0	0
8					3,000	0.03	0.15	0.04	0	0	0	0	0
9					4,000	0.04	0.11	0.02	0	0	0	0	0
10					5,000	0.05	0.13	0.04	0	0	0	0	0
11			9	4	1,000	0.01	0.12	0.04	0	0	0	0	0
12					2,000	0.02	0.14	0.14	0	0	0	0	0
13					3,000	0.03	0.02	0.02	0	0	0	0	0
14					4,000	0.04	0.10	0.10	0	0	0	0	0
15					5,000	0.05	0.11	0.11	0	0	0	0	0
16		72	3	24	1,000	0.01	0.13	0.02	0	0	0	0	0
17					2,000	0.02	0.11	0.01	0	0	0	0	0
18					3,000	0.03	0.10	0.01	0	0	0	0	0
19					4,000	0.04	0.07	0.04	0	0	0	0	0
20					5,000	0.05	0.07	0.02	0	0	0	0	0
21			6	12	1,000	0.01	0.11	0.01	0	0	0	0	0
22					2,000	0.02	0.09	0.00	0	0	0	0	0
23					3,000	0.03	0.10	0.04	0	0	0	0	0
24					4,000	0.04	0.11	0.02	0	0	0	0	0
25					5,000	0.05	0.09	0.02	0	0	0	0	0
26			9	8	1,000	0.01	0.12	0.01	0	0	0	0	0
27					2,000	0.02	0.12	0.01	0	0	0	0	0
28					3,000	0.03	0.06	0.01	0	0	0	0	0
29					4,000	0.04	0.11	0.03	0	0	0	0	0
30					5,000	0.05	0.14	0.03	0	0	0	0	0
31			12	6	1,000	0.01	0.06	0.01	0	0	0	0	0
32					2,000	0.02	0.07	0.02	0	0	0	0	0
33					3,000	0.03	0.14	0.02	0	0	0	0	0
34					4,000	0.04	0.13	0.05	0	0	0	0	0
35					5,000	0.05	0.10	0.03	0	0	0	0	0

이 0.01일 때 과대추정하고 나머지의 경우는 정확히 추정하였으며 MAP 검정과 병렬분석은 모두 모집단의 주성분의 수 m을 정확히 추정하였다. 표 3은 $\ell_{ij}=0.8$일 때 모의실험 결과이며 $\ell_{ij}=0.6$일 때 모의실험 결과와 비슷한 결과이지만 Kaiser 방법, MAP 검정과 병렬분석 모두 모집단의 주성분의 수를 매우 정확히 추정하였다.

6. 결론

모집단으로부터 소규모 표본을 추출하여 분석하는 전통적인 통계적 분석방법은 빅데이터 환경에서 분석의 어려움이 있다. 이와 같은 문제점을 해결하기 위하여 본 논문에서는 효율적인 표본추출을 통하여 이와 같은 문제를 해결하려고

노력하였다. 특히 많은 분야에서 다양하게 사용되는 다변량 통계분석 기법인 주성분분석에 대하여 빅데이터분석을 위한 표본추출방법을 제안하였다. 제안방법의 성능평가를 위하여 본 연구에서는 다양한 모의실험을 실시하였다. 모의실험 결과를 통하여 빅데이터로부터 추출된 일부 표본을 분석한 후 이를 전체와 비교하는 반복된 모의실험을 통하여 제안하는 방법이 빅데이터분석에 효과적으로 적용될 수 있음을 확인하였다. 향후 연구과제에서는 주성분분석 뿐만 아니라 다양한 통계분석 기법들 전체에 적용할 수 있는 일반화된 방법론에 대하여 연구할 것이다.

References

[1] K. Pearson, "On lines and planes of closest fit to systems of points in space", *Phil Mag*, vol. 2, pp. 559-572, 1901.

[2] J. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis", *Biometrika*, vol. 53, pp. 325-338, 1966.

[3] G. Arnold and A. Collins, "Interpretation of transformed axes in multivariate analysis", *Applied Statistics*, vol. 42, pp. 381-400, 1993.

[4] I. Jolliffe, *Principal component analysis*, Springer, 2002.

[5] M. Oleksiak, J. Roach, and D. Crawford, "Natural variation in cardiac metabolism and gene expression in fundulus heteroclitus", *Nature Genetics*, vol. 37, pp. 62-72, 2005.

[6] Johnson, R. A. and Wichern, D. W., *Applied multivariate statistical analysis*, Prentice-Hall, NJ, 1982.

[7] W. R. Zwick and W. F. Velicer, "Comparison of five rules for determining the number of components to retain", *Psychological Bulletin*, vol. 99, pp. 432-442, 1986.

[8] M. S. Bartlett, "Tests of significance in factor analysis", *British Journal of Psychology*, vol. 3, pp. 77-85, 1950.

[9] M. S. Bartlett, "A further note on tests of significance in factor analysis", *British Journal of Psychology*, vol. 4, pp. 1-2, 1951.

[10] H. F. Kaiser, "The application of electronic computers to factor analysis", *Educational and Psychological Measurement*, vol. 20, pp. 141-151, 1960.

[11] R. B. Cattle, "The scree test for the number of factors", *Multivariate Behavioral Research*, vol. 1, pp. 245-276, 1966.

[12] J. L. Horn, "A rationale and test for the number of factors in factor analysis", *Psychometrika*, vol. 30, pp. 179-185, 1965.

[13] W. F. Velicer, "Determining the number of components from the matrix of partial correlations", *Psychometrika*, vol. 41, pp. 321-327, 1976.

[14] J. Han and M. Kamber, *Data mining: concepts & techniques*, 2nd ed., Elsevier Inc., New York, 2006.

[15] S. Jun, "A Big Data Learning for Patent Analysis", *Journal of Korean Institute of Intelligent Systems*, Vol. 23, No. 5, pp. 406-411, 2013.

[16] B. Choi, J. Kong, and M. Han, "The Model of Network Packet Analysis based on Big Data", *Journal of Korean Institute of Intelligent Systems*, Vol. 23, No. 5, pp. 392-399, 2013.

[17] K. Kim, J. Jeong, and G. Park, "Assessment of External Force Acting on Ship Using Big Data in Maritime Traffic", *Journal of Korean Institute of Intelligent Systems*, Vol. 23, No. 5, pp. 379-384, 2013.

[18] S. Hong, and M. Han, "The Efficient Method of Parallel Genetic Algorithm using MapReduce of Big Data", *Journal of Korean Institute of Intelligent Systems*, Vol. 23, No. 5, pp. 385-391, 2013.

[19] H. C. Cho, and Y. J. Jung, "Probabilistic Modeling of Photovoltaic Power Systems with Big Learning Data Sets", *Journal of Korean Institute of Intelligent Systems*, Vol. 23, No. 5, pp. 412-417, 2013.

[20] J. H. Cho, D. J. Lee, J. I. Park and M. G. Chun, "Feature Extraction and Classification of High Dimensional Biomedical Spectral Data", *Journal of Korean Institute of Intelligent Systems*, Vol. 19, No. 3, pp. 297-303, 2009.

[21] W. G. Cochran, *Sampling techniques*, 3rd ed., New York, Wiley, 1977.

[22] W. R. Zwick and W. F. Velicer, "Factors influencing four rules for determining the number of components to retain", *Multivariate Behavioral Research*, vol. 17, pp. 253-269, 1982.

[23] N. Cliff, "The eigen value greater than one rule and the reliability of components", *Psychological Bulletin*, vol. 103, pp. 276-279, 1988.

[24] R. L. Gorsuch, *Factor analysis*, 2nd ed., Lawrence Erlbaum Associates, Inc., 1983.

[25] B. P. O'Connor, "SPSS and SAS programs for determining the number og components using parallel analysis and Velicer's MAP test", *Behavioral Research Methods Instruments & Computers*, vol. 32, pp. 396-402, 2000.

[26] L. W. Glorfeld, "An improvement on Horn's parallel

analysis methodology for selecting the correct number of factors to rertain", *Educational and Psychological Measurement*, vol. 55, pp. 377-393, 1995.

[27] R Development Core Team, *R: A language and environment for statistical computing, R Foundation for statistical computing*, http://www.R-project.org, 2011.

저 자 소 개

이승주(Seung-Joo Lee)

1985년 : 청주대학교 응용통계학과 경제학사

1987년 : 동국대학교 통계학과 이학석사

1994년 : 동국대학교 통계학과 이학박사

1995년~현재 : 청주대학교 통계학과 교수

관심분야 : Multivariate Analysis, Data Mining

Phone : +82-43-229-8204

E-mail : access@cju.ac.kr