



중간 발표

Team 2: 김종민, 김민주, 김준엽, 박정현, 조수연, 조준태

중간 발표: 논문 내용 요약 발표

최종 발표: simulation 결과 발표

Remarks on Parallel Analysis (1992, Buja and Eyuboglu)

Contents

1. PA란 무엇인가?
2. 3가지 방법에 PA의 적용 및 비교
3. Loadings에 적용

Introduction

- What is Parallel Analysis? : a selection rule for the number of factors in principal components & principal factor analysis, 위 저자들은 **Permutation** 방법을 도입하여 기존의 Parallel Analysis와 차이점을 설명
- 이것이 왜 좋은가? : significance level을 조정하여 minor components의 영향을 줄일 수 있음 & tabulation (테이블화)이 불가능한 복잡한 경우에 유용함 this implies non-parametric version of PA

Multivariate Data Permutation

permutation이 무엇인가?

이것이 왜 좋은가?

실험은 어떻게 setting하는가?

Permutation principle

임의의 통계량에 대해 conditional null distribution(조건부 귀무분포)를 생성하는 방법

* 조건부 귀무분포 : 주어진 조건 하에서의 통계량 분포로, 조건부 유의수준과 p-values 계산 하는데 사용됨

표기법 정리

- $x_{i,j} (i = 1, 2, \dots, N, j = 1, 2, \dots, P)$: 큰 모집단에서 추출된 N명의 피험자로부터 얻어진 것으로 가정
- 랜덤 변수 $X_{i,j}$ 의 실현값(realizations)이며, 랜덤 벡터 $X_{i,1}, X_{i,2}, \dots, X_{i,P}$ 는 $i = 1, 2, \dots, N$ 에 대해 독립적이고 동일하게 분포(iid 가정 만족), 정규성과 같은 추가 가정은 하지 않음
- 독립변수 $X_{1,j}, X_{2,j}, \dots, X_{N,j}$ 의 common marginal distribution F_j 은 임의적이며, 상관관계의 존재 보장을 위해서 분산이 0이면 안된다는 제약이 존재함

$$\mathbf{X} = \begin{bmatrix} X_{1,1} & X_{1,2} & X_{1,3} & \dots & X_{1,P} \\ X_{2,1} & X_{2,2} & X_{2,3} & \dots & X_{2,P} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ X_{N,1} & X_{N,2} & X_{N,3} & \dots & X_{N,P} \end{bmatrix}$$

→ j 번째 열 내의 변수들은 common marginal distribution F_j 를 공유하며, 이값은 칼럼별로 다름

- Null assumption

- 확률변수 간 독립(stochastic independent) → 한 변수 값이 다른 변수에 대한 정보를 제공하지 않음. 즉, $X_{i,1}, X_{i,2}, \dots, X_{i,P}$ 는 independently distributed
- PFA로 도출한 모집단 eigenvalues는 모두 0인 반면, PCA로 도출한 모집단 eigenvalues는 +1 을 가짐을 의미

변수들의 joint null distribution은 column내에서의 permutation(순열)에 대해 불변

In shorts, N개의 데이터 & P개의 변수. 각 변수들은 독립이고 같은 변수 내의 데이터들은 같은 분포에서 나왔다고 가정

$$\mathbf{X}_\pi = \begin{bmatrix} X_{1,1} & X_{\pi_2(1),2} & X_{\pi_3(1),3} & \dots & X_{\pi_P(1),P} \\ X_{2,1} & X_{\pi_2(2),2} & X_{\pi_3(2),3} & \dots & X_{\pi_P(2),P} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ X_{N,1} & X_{\pi_2(N),2} & X_{\pi_3(N),3} & \dots & X_{\pi_P(N),P} \end{bmatrix}$$

- $\pi_j = (\pi_j(1), \pi_j(2), \dots, \pi_j(N))$ N개의 인덱스 1, 2, ..., N에 대한 임의의 순열을 나타냄 (각각의 j에 대해 j=2, 3, ..., P) → 변수들의 순열된 배열은 원래의 배열 X와 동일한 결합 분포를 가지게 됨
 - 변수들 간의 순서를 변경해도 결합분포는 변하지 않는다는 것을 의미
 - ⇒ 열 내에서 변수들의 위치를 바꾸더라도 관련된 확률분포는 변하지 않는다는 특성을 나타냄
- 특정열의 불변성: 첫 번째 열과 같이 특정 열은 귀무가설 없이도 주어진 N명의 주체의 순서를 변경해도 분포가 변하지 않음 → 어떤 특정 변수나 열에 대해 순열을 적용할 필요가 없음을 의미
- X의 행 내에서의 stochastic dependence는 permutation symmetry(순열대칭성)를 파괴함
 - 순열 대칭성? 귀무가설이 성립할 때 순열된 데이터 배열들이 동일한 확률로 나타남
 - 각 순열된 데이터 배열이 실제 관측된 배열과 동일한 확률로 발생한다는 것을 의미

⇒ 종속성이 있으면 순열 배열들이 서로 다른 확률로 나타날 수 있기 때문에 순열대칭성이 파괴

Approximate permutation distribution (근사 순열 분포)

몬테카를로 시뮬레이션을 이용하며, 합리적인 횟수 R 번만큼 열 순열이 복원추출됨

- 대부분의 경우 $R=9$ 랜덤 순열만으로도 strong structure 찾기 충분함
- 이외에도 $R = 99, R = 499$ 등의 홀수로 설정
 - 실제 데이터와 함께 R 개의 시뮬레이션 데이터 세트는 귀무가설 하에서 동일한 확률로 발생하는 $R+1$ 개의 데이터 세트 풀을 형성하기 때문
 - 순위 매기기 용이 → 10개 중, 100중, 500 중 몇위

Permutation tests

- 순열 검정은 비모수적 독립성 귀무가설 하에서 정확한 유의성을 달성하는 특성이 있음 = 실제로 나타나는 분포를 고려하여 정확한 결과를 얻을 수 있음
- 근사 순열 검정에도 위와 같은 특성이 유지됨
- 적은 수의 몬테카를로 샘플을 사용하는 경우에는 일반적으로 검정력이 낮을 수 있지만, 몬테카를로 샘플링을 충분히 많이 사용하면 검정력을 높일 수 있음

앞으로 위처럼 permutation PA를 적용할 것인데, 28명의 subject에게 15개의 질문으로 구성된 데이터를 사용할 것

Eigenvalues from PCA

Interpretation of Table 1

Table 1: Sample eigenvalues of 15-item data (28 subjects) from principal component analysis (PCA), with permutation null quantiles and p-values.

Order	Observed Eigenvalue	Permutation Quantiles					Permutation P-values	Bartlett P-values
		Median	75%	90%	95%	99%		
1	5.54	2.53	2.67	2.83	2.94	3.10	0.002	0.000
2	2.46	2.10	2.21	2.32	2.37	2.53	0.026	0.000
3	1.81	1.81	1.89	1.96	2.01	2.12	0.524	0.000
4	1.35	1.56	1.62	1.69	1.73	1.79	0.984	0.000
5	1.00	1.34	1.41	1.47	1.51	1.55	1.000	0.002
6	0.67	1.16	1.21	1.27	1.29	1.36	1.000	0.022
7	0.54	0.99	1.04	1.08	1.11	1.15	1.000	0.049
8	0.43	0.84	0.88	0.93	0.95	1.00	1.000	0.082
9	0.34	0.70	0.75	0.79	0.82	0.86	1.000	0.108
10	0.31	0.58	0.62	0.66	0.68	0.70	1.000	0.117
11	0.19	0.46	0.49	0.53	0.56	0.60	1.000	0.322
12	0.15	0.36	0.39	0.42	0.44	0.48	1.000	0.291
13	0.11	0.27	0.30	0.33	0.35	0.39	1.000	0.324
14	0.06	0.18	0.21	0.24	0.25	0.29	1.000	0.493
15	0.04	0.11	0.13	0.15	0.17	0.20	0.994	*****

→ 각 Observed Eigenvalue와 Permutation Quantiles를 비교 & p-value 관찰

- 가장 큰 고유값(5.54): p값 0.002 / Quantiles 값과 상관없이 유의미함
- 두번째 고유값(2.46): p값 0.026 / 99%, 95% Quantiles 사이에 있지만 유의미함
- 세번째 고유값(1.81): p값 0.524 / Median Quantiles 근처에 위치
- 네번째 고유값부터는 Median Quantiles보다 작은 값

일반적인 상황에서는 **고유값의 합=총분산**이므로 값이 큰 값이 있으면 작은 값이 나와야함

- Order 낮은 고유값이 클수록, 나머지 고유값은 작으므로 유의적이지 않게 됨
- **Bartlett Test** (Large Order 효과 고려하는 방법)를 이용하면 p값이 전체적으로 작게 계산되어서 유의미한 고유값의 개수가 많아짐(10% 유의수준이면 4~8번째 고유값도 의미해짐)

Bartlett's test of sphericity¹, which is often done prior PCA or factor analysis, tests whether the data comes from multivariate normal distribution with zero covariances. (Note please, that the standard asymptotic version of the test is not at all robust to the departure from multivariate normality. One might use bootstrapping with nongaussian cloud.) To put it equivalently, the null hypothesis is that the population correlation matrix is identity matrix or that the covariance matrix is diagonal one.


Imagine now that multivariate cloud is perfectly spherical (i.e. its covariance matrix is proportional to the identity matrix). Then 1) any arbitrary dimensions can serve principal components, so PCA solution is not unique; 2) all the components have the same variances (eigenvalues), so PCA cannot help to reduce the data.

Imagine the second case where multivariate cloud is ellipsoid with oblongness strictly along the variables' axes (i.e. its covariance matrix is diagonal: all values are zero except the diagonal). Then the rotation implied by PCA transformation will be zero; principal components are the variables themselves, only reordered and potentially sign-reverted. This is a trivial result: no PCA was needed to discard some weak dimensions to reduce the data.

Figure 2: The PA scree plot. Scree plot plus the parallel analysis...



Download scientific diagram | The PA scree plot. Scree plot plus the parallel analysis results. from publication: The Scree Test and the Number of Factors: a Dynamic Graphics Approach | Exploratory Factor Analysis

 https://www.researchgate.net/figure/The-PA-scrree-plot-Scree-plot-plus-the-parallel-analysis-results_fig2_276934102

Permutation quantiles

Permutation 분포는 Median과 5% 차이에서 알 수 있듯이 narrow (order가 커질수록 더 narrow)

- 의미: Median, 5% 등 '유의수준을 어떻게 결정하는 것'이 cutoff에 큰 영향을 미치지 않음 (Median이면 세번째, 5%이면 두번째 고유값까지)
- 해당 예제뿐만 아니라 $N \times P$ 에도 적용 가능

Horn's PA? 가장 처음 제안된 PA 방식으로 **Median**, 즉 유의수준 50%를 사용

→ Liberal한 방법론: 유의수준이 크므로 (논문의) 1%, 5%일 때 대비해서 'Minor Component' 포함

⇒ 이에 대한 보완책으로 본 논문에서 *Permutation quantiles*를 도입했다고 이해 가능

Eigenvalues from Principal Factor Analysis and a Comparison with Principal Components

PCA와의 결과 비교를 중심으로

PCA vs PFA

*diagonal entries of correlation matrix → estimates of shared variance, 주로 R_j^2 (결정 계수 → 설명력)

Table 1: Sample eigenvalues of 15-item data (28 subjects) from principal component analysis (PCA), with permutation null quantiles and p-values.

PCA

Order	Observed Eigenvalue	Permutation Quantiles					Permutation P-values	Bartlett P-values
		Median	75%	90%	95%	99%		
1	5.54	2.53	2.67	2.83	2.94	3.10	0.002	0.000
2	2.46	2.10	2.21	2.32	2.37	2.53	0.026	0.000
3	1.81	1.81	1.89	1.96	2.01	2.12	0.524	0.000
4	1.35	1.56	1.62	1.69	1.73	1.79	0.984	0.000
5	1.00	1.34	1.41	1.47	1.51	1.55	1.000	0.002
6	0.67	1.16	1.21	1.27	1.29	1.36	1.000	0.022
7	0.54	0.99	1.04	1.08	1.11	1.15	1.000	0.049
8	0.43	0.84	0.88	0.93	0.95	1.00	1.000	0.082
9	0.34	0.70	0.75	0.79	0.82	0.86	1.000	0.108
10	0.31	0.58	0.62	0.66	0.68	0.70	1.000	0.117
11	0.19	0.46	0.49	0.53	0.56	0.60	1.000	0.322
12	0.15	0.36	0.39	0.42	0.44	0.48	1.000	0.291
13	0.11	0.27	0.30	0.33	0.35	0.39	1.000	0.324
14	0.06	0.18	0.21	0.24	0.25	0.29	1.000	0.493
15	0.04	0.11	0.13	0.15	0.17	0.20	0.994	*****

Table 2: Sample eigenvalues of 15-item data from principal factor analysis (PFA), with permutation null quantiles and p-values.

PFA

Order	Observed Eigenvalue	Permutation Quantiles					Permutation P-Values
		Median	75%	90%	95%	99%	
1	5.34	2.06	2.22	2.39	2.52	2.77	0.002
2	2.23	1.64	1.76	1.87	1.95	2.06	0.002
3	1.59	1.33	1.44	1.53	1.57	1.70	0.048
4	1.02	1.08	1.18	1.25	1.29	1.38	0.664
5	0.73	0.85	0.95	1.01	1.04	1.12	0.838
6	0.38	0.66	0.74	0.80	0.85	0.91	0.998
7	0.30	0.48	0.55	0.62	0.66	0.75	0.966
8	0.20	0.32	0.39	0.45	0.48	0.53	0.932
9	0.14	0.19	0.24	0.30	0.33	0.39	0.724
10	0.06	0.05	0.10	0.15	0.18	0.24	0.410
11	-0.04	-0.06	-0.02	0.03	0.05	0.09	0.382
12	-0.07	-0.14	-0.11	-0.08	-0.06	-0.02	0.076
13	-0.11	-0.21	-0.19	-0.17	-0.16	-0.13	0.006
14	-0.13	-0.27	-0.25	-0.23	-0.22	-0.20	0.002
15	-0.18	-0.31	-0.29	-0.27	-0.26	-0.24	0.002

Obvious difference from PCA

1) **occurrence of negative eigenvalues** → prevents interpretation of eigenvalues as variances

- 음수인 고유값이 나오는 경우 = 과도한 요인 추출을 나타내는 것으로 해석될 수 있음
 - 주어진 데이터에서 설명할 수 있는 것보다 더 많은 요인을 추출했다는 것을 의미
- 분석에서 제외하거나 데이터의 잡음이나 오차를 나타내는 것으로 간주

2) *second eigenvalue-significant beyond convention level / third-significant at 5% level*

⇒ **3 factor 선택** ↔ PCA-2개 선택 ⇒ **PFA= more liberal, tend to include more factors** 이러한 경향을 아래 수식으로 설명 가능

$$\frac{1}{P} \sum_j R_j^2 - \frac{P-1}{N-1} = \text{average } R^2 - \text{null mean of } R^2$$

- expected average null R^2 = expected null eigenvalue for PFA = pure chance capitalization
- observed average R^2 = observed average eigenvalue for PFA

→ 위의 식 = discrepancy in overall levels of observed and null profiles

⇒ average R^2 adjusted for chance capitalization

⇒ PCA에 위 결과를 더하면, PFA와 비슷해진다!

(+) PCA의 목표=데이터의 분산을 최대한 보존하는 축을 찾는 것 ↔ PFA= 공통 요인을 찾아내는 것

- PCA는 각 변수들을 고려하지만, PFA는 각 변수들이 공통 요인과 어떻게 관련되어 있는지를 파악
- PFA는 PCA에 비해 추가적인 것이 과정에 들어간다. 그 추가적인 것, 즉 공통 요인을 찾아내는 것?을 나타낸 것이 바로 위의 식이라고 해석할 수 있을 것이다.

Principal Factor Analysis (PFA) and PA

PFA에 PA를 활용하는 것에 대한 의문

=PCA 때도 minor component를 포함해서 문제였는데, 이것보다 더한 PFA의 경우에는 clearly negligible한 factor를 포함할 수도 있다!

EX) third-추후에 채택되지 말아야 하는 것으로 밝혀지지만, PFA는 5% 수준에서 유의미하다고 판단

second-minor component 정도로 여겨져야 하는 정도지만, PFA는 유의미하다고 판단

PCA와 PFA 간의 차이는 P(변수의 수)가 증가하며 감소(단, N은 고정)

1) null mean $\frac{(P-1)}{(N-1)}$ 이 1로 접근하며 결정계수 R_j^2 가 증가

- P 변수의 수, 예측 변수가 증가하면서 결정계수의 null mean도 최대값인 1로 접근하고, 결정계수도 점점 증가

→ 결정계수가 증가, 즉 변수들 간의 상관관계가 증가한 것에 따른 결과, PFA와 PCA의 차이도 감소

2) Correlation matrix의 eigendecomposition에 대한 대각 요소의 영향 감소

- P가 증가함에 따라, $P^2 - P$ (비대각 요소)가 P (대각 요소)보다 더 빠르게 증가하므로, 고유값 분해에 대한 대각 요소의 영향이 감소!
 - 비대각선 요소는 두 변수 간의 공분산, 즉 변수 간의 상관 관계를 나타냄
 - 비대각선 요소의 영향이 증가한다는 것은 변수들 간의 상관관계가 증가했다는 것을 의미

→ PFA와 PCA의 차이도 감소

Eigenvalues of Resistant Correlation Matrices

Resistant (correlation)이란 무엇인가?

장점 및 단점 & 적용 예시

28*15 데이터에 적용 결과

What is resistant?

데이터의 outlier에 대해서 어떤 통계량이 면역력이 있으면 이는 resistant 하다

- Raw sample correlation 사용하는 경우, not resistant. Why? 한 개의 극단적인 포인트가 correlation의 변동을 야기할 수 있음
- 따라서 correlation 대신 resistant correlation을 구하여 사용할 것
 - (Gnanadesikan and Kettenring) 다듬어진 분산 (trimmed variances)와 분산의 관점에서 상관관계의 re-expression에 기반한 방법
- 우리는 이 저항 상관관계를 구하는 과정에서 원하는 정도의 저항성을 설정할 수 있는데 (called trimmed factor), 과거의 경험이나 현재 가진 데이터의 초기 탐색 과정을 기반으로 이를 결정

Pros and Cons of resistant correlation

왜 사용하는가?

1. 사람 분석가가 outlier를 판단하는 것보다 보통 뛰어남
2. Mis-recording, 실험실패, 피험자의 비협조 등으로 발생한 small subset 데이터의 분석에 유용함
 - PCA, PFA 등에서 이러한 데이터는 minor factor로 판단될 수 있지만, RC를 쓰면 이것들 방지 가능 & raw correlation과 비교해서 잠재적인 문제를 발견 가능

적용하면 어떻게 되는가? (사례)

- **setting:** 협상 과정에서 두 사람의 서로에 대한 perceived power를 측정
 - 사람들이 특정 개인이나 그룹이 자신의 행동을 통제하거나 행동에 영향을 미칠 수 있는 권한과 능력을 가지고 있다고 믿는 정도로, negative correlation을 기대함

(-0.29, 2.62), (-0.94, 2.84), (-2.35, 3.43), (1.97, 0.84), (-0.94,-0.88), (0.84, 0.84), (0.84, 2.02), (-4.94, 4.90), (-0.29, 0.84), (-0.88, 1.12), (-2.35, 2.56), (-1.76, 0.84), (-5.82,-3.82), (-4.69, 3.43).

the raw correlation is -0.001. With a trimming factor of 1/14 (allowing for about 1 outlier in 14), the resis-

- **result:** raw correlation = -0.001 vs. resistant correlation = -0.528 (trimming factor = 1/14, this implies 14개중 1개의 outlier는 허용. 1개 outlier가 있을 것으로 예상한다는 의미인듯)

- (-5.82, -3.82)가 outlier인데, 이것을 빼면 예상대로 raw cor = -0.663, 다른 것을 하나씩 빼면 -0.025 and 0.236 사이 → 13번째 저것이 outlier가 맞다고 예측 가능
- resistant 적용해도 무엇이 outlier인지 알려주지는 않으나, outlier 존재의 단서 제공 가능

단점은 없는가?

- Mildly negative eigenvalue가 나타날 수 있음. 그러나 PFA의 R2-adjust보다는 심하지 않다

Application to used data

- *Trimming factor = 1/14 (28개 데이터 중 2개의 outlier가 trim 됨), PCA 시행*

Order	Observed Eigenvalue	Median	Permutation Quantiles				Permutation P-Values
			75%	90%	95%	99%	
1	6.75	2.75	2.95	3.12	3.20	3.43	0.002
2	2.44	2.27	2.40	2.52	2.59	2.66	0.190
3	1.85	1.94	2.03	2.12	2.17	2.31	0.760
4	1.17	1.66	1.73	1.80	1.83	1.92	1.000
5	0.84	1.42	1.48	1.54	1.57	1.66	1.000
6	0.68	1.20	1.27	1.33	1.36	1.41	1.000
7	0.59	1.00	1.06	1.12	1.15	1.20	1.000
8	0.29	0.83	0.87	0.93	0.97	1.03	1.000
9	0.25	0.66	0.72	0.77	0.80	0.84	1.000
10	0.14	0.51	0.56	0.60	0.63	0.69	1.000
11	0.11	0.38	0.43	0.47	0.50	0.54	1.000
12	0.04	0.25	0.29	0.33	0.36	0.40	1.000
13	0.00	0.14	0.18	0.22	0.23	0.27	0.996
14	-0.06	0.03	0.07	0.10	0.12	0.15	0.950
15	-0.11	-0.09	-0.05	-0.01	0.00	0.05	0.658

RPCA 결과

Order	Observed Eigenvalue	Median	Permutation Quantiles				Permutation P-values
			75%	90%	95%	99%	
1	5.54	2.53	2.67	2.83	2.94	3.10	0.002
2	2.46	2.10	2.21	2.32	2.37	2.53	0.026
3	1.81	1.81	1.89	1.96	2.01	2.12	0.524
4	1.35	1.56	1.62	1.69	1.73	1.79	0.984
5	1.00	1.34	1.41	1.47	1.51	1.55	1.000
6	0.67	1.16	1.21	1.27	1.29	1.36	1.000
7	0.54	0.99	1.04	1.08	1.11	1.15	1.000
8	0.43	0.84	0.88	0.93	0.95	1.00	1.000
9	0.34	0.70	0.75	0.79	0.82	0.86	1.000
10	0.31	0.58	0.62	0.66	0.68	0.70	1.000
11	0.19	0.46	0.49	0.53	0.56	0.60	1.000
12	0.15	0.36	0.39	0.42	0.44	0.48	1.000
13	0.11	0.27	0.30	0.33	0.35	0.39	1.000
14	0.06	0.18	0.21	0.24	0.25	0.29	1.000
15	0.04	0.11	0.13	0.15	0.17	0.20	0.994

PCA 결과

- 더 중요한 낮은 order에서의 quantile 값들이 조금씩 올라간 것을 볼 수 있음
 - 2개의 outlier를 reject했으므로, effective sample size가 줄어들음
- 첫번째 eigenvalue만 significant하다고 판단
 - 값이 5.54 → 6.75로 증가: 소수의 데이터가 첫 요소의 강도를 조금 모호하게 할 수 있음 의미
 - 두번째 eigenvalue는 insignificant해짐
- 요약하자면, *trimming factor*에 의해 null distribution의 raise이 발생하여 low order의 quantile 값이 상승 & 첫번째 component의 강도가 증가

⇒ PFA (R2-adjust) 보다 PCA (Raw correlation)가, 그리고 이보다 PCA (Resist correlation)가 덜 liberal하여 적은 수의 component를 선택

앞선 3가지 방법의 비교

	PCA	PFA	PCA w/ RC
significant	1, 2	1, 2, 3	1
minor component	med	high, but OK as P increases	low

Assessment of Loadings

About Loadings and Table 5

*Loadings : 어떤 변수가 특정 요인과 가장 밀접하게 연관되어 있는지를 보여주는 지표

Factor Loading (요인적재값) → 각 변수와 해당 요인간의 관계계수로, 추출된 요인을 회전하여 요인구조를 명확히 알 수 있음

- 각 변수가 특정 요인에 ‘로드되는 양’을 나타내며 이를 통해 어떤 변수가 특정 요인과 가장 밀접하게 연관되어 있는지를 이해할 수 있음
- 흔히 +1과 -1 사이의 값을 가지며 흔히 +.5 이상일 때 실제적 유의성을 갖는 것으로 받아들임
- Factor Loading Matrix를 추정할 때 Eigenvalue를 이용

⇒ 따라서 **unrotated loading**에 대한 평가 역시 PA를 통해 할 수 있을 것

가장 큰 3개의 Principle component에 대해 15개 items 각각의 loadings를 계산

Table 5: Permutation p-values and average null quantiles for the absolute loadings of the largest three principal components (PCA) of the 15-item data.

Item	First Component		Second Component		Third Component	
	Loading	P-Value	Loading	P-Value	Loading	P-Value
1	0.14	0.802	0.81	0.006	0.14	0.730
2	-0.12	0.826	0.51	0.180	-0.49	0.182
3	0.79	0.014	0.29	0.484	0.00	0.984
4	0.56	0.180	-0.21	0.672	0.16	0.686
5	0.61	0.112	0.22	0.574	0.37	0.316
6	0.67	0.086	-0.25	0.570	0.56	0.114
7	0.48	0.342	-0.45	0.286	-0.53	0.092
8	0.71	0.054	-0.43	0.288	0.17	0.658
9	0.83	0.004	-0.14	0.776	0.22	0.592
10	0.11	0.806	0.85	0.002	0.37	0.298
11	0.62	0.138	-0.02	0.964	-0.29	0.464
12	0.77	0.008	0.06	0.898	-0.47	0.228
13	0.64	0.106	0.35	0.410	-0.49	0.174
14	0.81	0.004	0.20	0.618	-0.09	0.824
15	0.52	0.292	-0.03	0.932	0.13	0.748
Average Quantiles						
50%	0.35		0.30		0.26	
90%	0.64		0.60		0.57	
95%	0.70		0.67		0.65	
99%	0.79		0.78		0.77	

Components

- First component는 15개 중 12개의 item에서 weighted
- Second component는 나머지 3개 (1, 2, 10) item에 가장 큰 weight 보이고 있음
→ 비록 First component에 대한 보충 정도만 할 수 있지만, 어쨌든 해석 가능함
- Third component는 절댓값 0.5 이상의 loading을 지니는 item이 2개 뿐이고 이들 역시 이미 First component로 해석이 가능

⇒ 따라서 resistant parallel analysis 결과와 종합해보았을 때 *First = major*, *Second = minor*, *Third = negligible*이라는 결론 내릴 수 있음

Additional Interpretation

Inferences

- First component는 전반적인 weight가 높고 Eigenvalue의 p-value가 0.002였음에도 불구하고 검정 결과 유의하다고 판단된 item이 12개 중 4개밖에 없었음 (0.05 기준)
- 반면 Second component는 Eigenvalue의 p-value는 0.026이지만 loading에 대해 검정했을 때는 3개 중 2개의 item이 굉장히 유의 (0.01 기준)


⇒ *Summary statistic 기반의 Eigenvalue와 constituents 기반 loading 각각에 대한 inference가 어느정도 상이할 수 있음을 암시하는 결과*

*Strong eigenvalue는 상대적으로 약하지만 많은 수의 loadings에 의한 결과물일수도, 몇 안되는 강한 loadings에 의한 결과물일수도 있음

Sampling Stability

- *큰 magnitude의 loading ≠ 높은 sampling stability (샘플링 변동성이 클 수 있음)*
- Loading의 샘플링 안정성은 그 size가 아닌 모집단 고유값이 얼마나 well separated인지에 따라 결정됨

APA PsycNet

 <https://psycnet.apa.org/record/1968-03169-001>

⇒ Loading의 유의성을 샘플링 안정성으로 오인해서는 안됨

- 다만 특정 magnitude의 loading이 우연에 의한 것일 가능성은 낮음

*물론 사분위수는 표본 크기, 변수의 개수, 구성 요소의 순서 등의 parameter에 의해 달라질 수 있으나, loading의 size를 판단하기 위한 conventional thresholds는 이들과 무관

Discussion

- 일반적으로 loading의 larger magnitude가 more significant함을 의미한다고 생각할 수 있지만 (smaller p-values) 때로는 loadings와 p-value가 counterintuitive (0.79 & 0.77 of first component)
- 각 loading에 대해 별도의 사분위수 또는 p-value를 계산하는 대신, 특정 component의 모든 loading에 대한 단일 null-distribution을 얻는 것으로 이를 개선할 수 있음
- Table의 Average quantile이 이러한 접근 방식에 해당

*이러한 방식은 marginal distribution이 null distribution에 미치는 영향을 무시할 수 있다는 사실을 통해 정당화 (추후 섹션에서 다뤄짐)

More discussions

1. Horn씨의 전통적인 방법에서는 normal로 sampling해서 PA를 했는데, normal 분포가 아닌 non-normal 분포(e.g. 카이제곱, normal³ 등등)로 PA를 실험 (N, P) : (20, 5), (20, 15), (100, 10), (100, 50)

eigenvalue의 차이가 크지 않음 & 가장 큰 차이가 나는 때는 N:P 비율이 작은 경우((20,15), (100,50))

따라서 많은 non-normal분포의 경우에도, normal distribution을 가정하고 사용하는 것이 좋은 approximation이 된다는 insight
2. N과 P에 따른 loadings 값 & quantiles에 따른 eigenvalues값 table화 해서 첨부 → 이 것이 놀라운 결과를 나타냈다 (P, 변수, 의 개수가 증가함에따라 loading의 각 quantiles의 차이가 줄어듦)