응용통계연구 제 8권 2호
1995년 9월 pp. 201-216.

붓스트랩 방법을 이용한 이항분포자료에 대한 요인수 결정에 관한 연구

김 성 호[1], 정 미 숙[2]

요 약

본 연구에서는 관측변수들이 이항확률변수일 때, 요인의 갯수를 찾는 방법을 모색하였다. 이를 위해 붓스트랩 방법을 사용하여 요인수 결정 기준을 제시하였고, 모의실험을 통하여 이 제시된 기준의 유용성을 보였다.

1. 서 론

연속확률변수들에 대해서 요인분석을 할 때, 몇개의 요인을 사용하느냐는 중요한 문제이다. 요인의 갯수를 정하는 방법으로는 최대우도 측정법에 의한 축차적 요인수 검정법이 있고, Kaiser의 방법으로 알려져 있는 고유값 1을 기준으로 하는 판정법, 표본공분산행렬로 부터 얻어진 고유값의 누적값을 사용해서 요인의 갯수를 정하는 방법, 고유값의 변화추이를 관찰해서 정하는 Cartell의 방법, 잔차상관행렬에 의한 방법, 편상관행렬에 의한 방법, AIC(Akaike's Information Criterion)방법, SBC(Schwarz's Bayesian Criterion)방법, 그리고 표본 공분산행렬의 순위(rank)로써 요인의 갯수를 정하는 방법 등 여러가지 방법들이 연구되어 왔다. 그러나, 이 방법들은 기본적으로 관측자료가 연속확률모형을 가정하고 있다.

관측자료가 명목적이거나 이산적인 경우, 더 나아가서 이러한 관측자료에 대한 요인들 역시 명목적이거나 이산적인 확률변수로써 간주할 수 있는 경우가 많이 있다. 예를 들어, 학생들에 대한 평가시험에서 각 평가문항의 점수를 매길 때 틀렸으면 0, 맞았으면 1점으로 하는 경우이다. 이때, 어느 한 문제를 풀기 위해서는 두가지 능력이 있어야 한다고 하면, 이 두가지 능력이 학생에게 '있다', '없다'라는 이항확률변수로써 나타낼 수 있다.

위에서처럼, 관측가능변수와 관측불능변수(즉, 요인)가 모두 이산변수인 경우를 쉽게 접할 수 있다. 이 경우에 관측가능변수와 요인변수들 사이의 확률모형을 개발하기 전에 반드시 점검해야 될 사항으로는 과연 몇개의 요인변수들이 현재의 관측가능변수들에 영향을 미치느냐 하는 문제이다. 즉, 요인수가 몇개냐는 것이다.

본 연구의 주 목적은 관측변수들이 이항확률변수일 때에 요인수를 찾는 방법을 모색하는데 있다. 이항자료의 경우, 실제 요인수보다 적은 요인수들이 추천될 가능성이 많을 것이라는 것이다. 그 이유중의 하나로는, 앞에서 언급한 여러가지 방법들은 전제조건 중에 관측변수들에 대한 확률 오차들 간의 독립성이 포함되어 있는데 반해서, 관측변수들이 이항확률변수일 때는 이

1) (305-701) 대전광역시 유성구, 한국과학기술원 교양과정부 기초과학과정.
2) (609-735) 부산광역시 금정구 장전동, 부산대학교 통계학과.

독립성이 보장될 수가 없기 때문이다. 이에 대한 보완책으로, 붓스트랩 방법을 사용해서 표본상관행렬의 고유값의 평균값을 좀 더 정확히 구한 뒤, 경험적으로 얻은 요인수 결정 기준을 제시하고, 모의실험을 통하여 앞의 방법들과 비교하여 보려한다.

본 논문은 총 5절로 구성되어 있다. 2절은 요인분석의 기본적인 모형과 연속확률분포에 대한 요인수 결정의 전통적인 판정기준을 압축정리하였고 3절은 이항확률분포에 대한 요인수 결정방법을 제시하였다. 4절은 모의실험결과를 통해 2절의 기준방법들과 3절에서 제시한 방법을 비교 분석하였고, 5절은 결론으로 본 논문에서 제시한 판정기준에 대하여 논의하였다.

2. 연속확률변수들의 요인모형과 요인수 결정방법들

이 절에서 우리는 연속확률변수들에 대한 요인모형에 있어서 지금까지 알려진 요인결정 방법들을 간략히 정리하고자 한다. 이 방법들에 익숙한 독자들은 3절로 바로 넘어가도 무방하리라 본다. 그러나, 익숙치 못한 독자들은 본 절의 압축정리된 내용이 본 논문 전체를 이해하는데 큰 도움이 되리라 생각한다.

2.1 모형

확률벡터 $X'=(X_1,X_2,\cdots,X_P)$의 평균은 μ이고 공분산행렬은 Σ라고 하자. 요인모형을 행렬로 나타내면 다음과 같다.

$$\underset{(px1)}{\underline{X}} = \underset{(px1)}{\underline{\mu}} + \underset{(pxm)}{L} \cdot \underset{(mx1)}{\underline{F}} + \underset{(px1)}{\underline{\varepsilon}} \qquad (2.1)$$

단,

$$\underline{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_P \end{bmatrix}, \quad \underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_P \end{bmatrix}, \quad \underline{F} = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_P \end{bmatrix}$$

$$L = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ a_{P1} & a_{P2} & \cdots & a_{Pm} \end{bmatrix}, \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_P \end{bmatrix}$$

여기에서 F_j는 j번째 공통요인 확률변수(관찰불가능한 확률변수)를, a_{ij}는 i번째 반응값 구성에서 j번째 요인의 영향을 나타내는 모수를, ε_i는 X_i에 대한 오차(특수요인)를 나타내며, L은 부하량 행렬이라고 부른다.이 (2.1) 모형에서 F와 ε는 서로 독립이고, $E(F)=0$, $Cov(F)=I_m$는 mxm 항등행렬이며, $E(\varepsilon)=0$이고, $Cov(\varepsilon)=\Psi$는 대각 행렬이라 가정한다. 이런 가정을 가질 때의 모형을 직교요인모형이라고 부른다.

식 (2.1)에서 공분산 $\Sigma = Cov(X) = LL' + \Psi$ 이고, 공분산행렬의 대각선상의 i번째 값인

$Var(X_i)$은 $\sigma_{ii} = \sum_{j=1}^{m} a_{ij}^2 + \psi_i$ 으로 공통성(communality : 공통요인 부하량의 제곱합)과 특수분산(ψ_i)의 합으로 나타나며, i번째와 k번째 반응값들의 공분산 $Cov(X_i,X_k)$는 $\sigma_{ik} = \sum_{j=1}^{m} a_{ij}a_{kj}$ 이며, 또한 $a_{ij} = Cov(X_i,F_j)$이다.

부하량 행렬(L)과 특수분산(Ψ)의 추정은 주성분법, 주축인자법, 최대우도법등을 사용하여 $\Sigma = LL' + \Psi$에서 L과 Ψ을 추정하여 요인분석에 필요한 여러 문제들을 처리한다. 이와 관련해서는 Richard & Dean(1982)을 참고하기 바란다.

2.2 연속확률분포에 대한 요인수 결정의 전통적인 판정기준

Kim & Charles(1978)와 김기영, 전명식(1989)에서도 언급한 것처럼 요인의 수를 결정할 때 여러 전통적인 판정기준들이 있는데, 이들 기준들은 서로 상호 보완적이므로 모든 기준들을 독립적으로 검토한 뒤 여러 기준에 의해 지지되는 결과를 선택하는 방법이 타당하며, 또한 요인분석의 여러 가지 불확실성과 복합성을 검토하여 이론적인 바탕 위에 가장 합리적으로 판정이 이루어져야 할 것이다.

(1) 유의성 검정

이 검정을 위해서는 변수들의 근사적 정규분포성이나, χ^2-test를 위한 조건들 등 최대우도 추정법에서 요구되는 여러 가정들이 충족된다는 전제 하에서 만족스런 결과가 유도된다.

이 기준은 m개의 요인수를 갖는 요인 모형의 적합성을 검정하기 위해 귀무가설(H_m : $\Sigma = LL' + \Psi$))과 대립가설(K_m : Σ은 임의의 양정치행렬)을 축차적으로 고려하여 유의한 m을 결정하는 방법을 채택한다.

H_m을 검정하기 위한 우도비 검정 통계량으로

$$-2\ln \frac{\text{maximized likelihood under } H_m}{\text{maximized likelihood}} = n \ln \frac{|\hat{L}\hat{L}' + \hat{\Psi}|}{|S_n^2|} \qquad (2.2)$$

을 사용하는 데, 여기서 $\hat{L},\hat{\Psi}$ 는 L,Ψ 의 최대 우도추정값, n은 표본크기이고 S_n^2은 표본분산값이다.

Bartlett(1954)은 (2.2)식에서 n대신 $[n-1-(2p+4m+5)/6]$을 대치하면 자유도 $((p-m)^2-(p+m))/2$을 갖는 χ^2분포에 근사한다는 것을 보였다.

m=0로부터 시작하여 어떤 m에 대해 기각될 때까지 (혹은 최대우도추정법의 반복적인 처리과정이 수렴하지 못할 때까지) 계속적으로 m을 하나씩 증가시켜 나가는 축차적인 방법을 따르는데, 이 기법은 유의성 검정을 실시함으로써 매우 바람직하나, 여러 가지 모의실험을 통해 분석한 결과 유의하지 않고(insignificant) 사소하다고 판단되는 변수들 간의 관계를 설명하기 위해서 요인을 더 추가하게 되는 경향이 있는것으로 알려져 있다. 또한, 이 기준은 축차적인 검정절차를 취하기 때문에 비용면에서 비효율적일 수도 있다.

(2) 고유값기준

이 기준은 표본 상관행렬로부터 요인을 유도할 경우 고유값이 1보다 큰 것들의 수효에 해당하는 만큼의 요인을 보유하는 기준으로, 주성분분석을 요인분석의 한 변형으로 고려하는 데에서 출발하였다. 우선 변수들이 서로 독립이라면 R(표본상관행렬)= I_P가 되어 모든 주성분의 합은 변수의 갯수와 동일하게 된다. 이 때 주성분들의 분산은 모두 1이 되므로 각 주성분이 가지는 분산들의 평균도 1이 된다. 따라서 1보다 작은 분산을 가지는 주성분은 원래의 변수들 중의 어느 하나보다 더 작은 정보를 가진다는 기준인데, 이는 Kaiser(1974)의 규칙으로 알려져 있다. 이 기준은 아주 단순하나 일반적으로 기대하는 요인의 수와 일치하는 정도가 크다. 하지만, 보유되어야 할 요인보다도 작은 수의 요인이 얻어질 수 있음에 유의할 필요가 있다.

(3) 요인의 공헌도

이 기준은 한 요인이 중요한 의미를 가지기 위해 모든 요인들의 전체 변이에 대해 이 요인이 가지는 공헌도가 최소한 얼마 이상이 되어야 한다는 것이다.

$$X_i-\mu_i = \sum_{j=1}^{m} a_{ij}F_j + \varepsilon_i$$

$$Var(X_i-\mu_i) = Var(\sum_{j=1}^{m} a_{ij}F_j+\varepsilon_i) = Var(\sum_{j=1}^{m} a_{ij}F_j) + \psi_i = \sum_{j=1}^{m} a_{ij}^2 + \psi_i$$

이므로 X_i의 분산중 F_j에 의해 설명되는 부분이 a_{ij}^2이 되고 모든 변수들의 공통분산에 대해 F_j에 의해 설명되는 부분은 $\sum_{i=1}^{p} a_{ij}^2$으로 이 요인의 종합적인 공헌도가 된다. 따라서, 표본으로부터 추정된 각 요인의 공헌도에 관한 추정값은

$$\hat{V}_j = \sum_{i=1}^{p} \hat{a}_{ij}^2 \quad j=1, 2, \cdots, m$$

인데, 여기서 $\hat{a}_{ij}$ 는 a_{ij}의 추정치이다.

실제로 Kaiser의 판정기준은 100/p%를 설명할 수 있는 요인을 유도하는 것과 동일하다. $\hat{\lambda}_j$를 표본상관행렬의 j번째 고유값이라 할때, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_j \geq \hat{\lambda}_{j+1} \geq \cdots \geq \hat{\lambda}_p$에서, 선택된 요인은($\lambda_j / \sum_{i=1} \lambda_i$) ≥ 100/p% 해당하는 고유값 만큼의 수이다.

(4) Scree - test

Cartell(1966)에 의해 제안된 이 기준은 R(혹은 S_n^2)의 고유값 λ_i를 크기순으로 배열하여 $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$라 할때, {(i, λ_i), i=1,2,···,p}의 도형(scree 도형)을 근거로 적절한 요인수를 결정하는 것이다. Scree 도형에서, (i, λ_i) 왼쪽에서 급격한 하강을 보이며, 그 다음 값부터 완만한 기울기를 가지고 최소 고유값까지 평평한 직선 모양일 때 i-1이 요인수가 된다. 이 기준은 주관적일 수 있으나 그림이 가지는 시각적인 효과를 얻을 수 있다. 만약 사소한 요인들이 상당히 존재하는 경우 주가 되는 요인들을 찾는 것이 주목적이라면 다른 방법들 보다 우수하다는 것이 여러 실험 결과 밝혀져 있다.

(5) 기타 판정기준

① 잔차상관행렬

이 행렬은 표본상관행렬 R에서 유도된 요인및 특수분산에 의해 재 생성된 상관행렬 $\hat{L}\hat{L}'+\hat{\Psi}$을 고려한 $R-(\hat{L}\hat{L}'+\hat{\Psi})$을 나타낸 것으로 만약 자료가 요인모형에 잘 부합된다면, 이 행렬의 비대각원소가 아주 작을 것이다.

② 편상관행렬

요인들을 유도한 후, 이들이 주어졌다는 조건하에서 변수들간의 모든 편상관계수를 구한후, 요인모형이 적절하다면, 그 값들이 아주 작을 것이라 기대한다.

③ AIC와 SBC

AIC(Akaike's Information Criterion)은 최대우도추정법이 사용될 때 모형에 포함된 요인의 수를 잘 추정한다. 유의성 검정처럼 통계적으로는 유의하나 실제와는 다른 요인수를 포함하는 경향이 있다. 기준은 통계량이 최소가 될 때 대응하는 요인수가 가장 적절하다는 판단이다.

SBC(Schwarz's Bayesian Criterion)은 AIC와 비슷하고 이 또한 통계량이 최소일 때의 요인수가 가장 적절하다는 판단이다. SBC는 AIC또는 유의성 검정보다는 사소한 요인을 덜 포함하는 것 같다.

3. 이항자료에 대한 요인수 결정

관측자료가 명목적이거나 이산적인 경우, 더 나아가서 이러한 관측자료에 대한 요인들 역시 명목적이거나 이산적인 경우를 생각하자. 예컨대, 학생들에 대한 평가시험에서 각 평가문항의 점수를 매길 때 틀렸으면 0점, 맞췄으면 1점으로 하는 경우, 어느 한 문항을 풀기 위해서는 그 문항에 대한 학생들의 능력(요인)이 고려되는데, 이 때의 능력을 '능력이 있다', '능력이 없다'라는 이항확률변수로 고려할 수 있다. 관측가능변수와 관측불능변수(요인)가 모두 이산변수인 경우에 관측가능변수와 요인변수들 사이의 확률모형을 개발하기 전에 반드시 점검해야 될 사항으로는 과연 몇 개의 요인변수들이 현재의 관측가능변수들에 영향을 미치느냐 즉, 요인수가 몇 개냐는 것이다. 이 절에서는 관측변수들이 이항확률변수일 때에 요인수를 찾는 방법을 제시하려 한다.

3.1 이항자료의 모형

관측가능변수와 요인이 모두 이산변수인 경우의 확률모형을 다음과 같이 고려하자. 식(2.1)에서, 일반성을 잃지 않고, 평균 μ를 0으로 하여 나타내면 다음과 같다.

$$
\begin{aligned}
X_1 &= \alpha_{11}A_1 + \alpha_{12}A_2 + \cdots + \alpha_{1m}A_m + \varepsilon_1 \\
X_2 &= \alpha_{21}A_1 + \alpha_{22}A_2 + \cdots + \alpha_{2m}A_m + \varepsilon_2 \\
&\vdots \\
&\vdots \\
X_p &= \alpha_{p1}A_1 + \alpha_{p2}A_2 + \cdots + \alpha_{pm}A_m + \varepsilon_p
\end{aligned} \qquad (3.1)
$$

단, $X_i = \begin{cases} 1, & \text{문항 i를 맞췄을 때} \\ 0, & \text{문항 i를 못 맞췄을 때} \end{cases}$

A_i = 능력i의 유(1),무(0)를 나타내는 이항확률변수

ε_i = X_i에 대한 오차

$i=1, 2, \cdots, p$

2절의 기존 방법들을 이항자료에 적용할 때의 문제점으로 아래 4가지 경우를 생각할 수 있다.

(1) 고려하려는 자료가 이항자료이므로 기존 요인모형의 전제조건에 부합되지 않는다. 즉, ε이 서로 독립이 아니며, 요인 $A_1, \cdots, A_m$ 에 의존한다.

선형결합 $X_i = \sum_{j=1}^{m} \alpha_{ij}A_j + \varepsilon_i \quad , \quad i=1, 2, \cdots, p$

에서 ε_i는 $A_1, A_2, \cdots, A_m$의 함수로 나타내진다.

(2) 고려하려는 자료의 원분포가 연속이라 가정할 때 원 분포의 분산·공분산에 비해서, 이항자료의 분산·공분산값들이 축소되어 있다.

(3) (1), (2)에 의해, 기존의 방법을 사용하면 실제보다 요인수가 적게 나타날 수 있다.

(4) 차선책으로 Tetrachoric correlation을 구하여 기존의 방법을 사용하는것을 생각할 수 있으나, 여기서의 전제조건인 다변량정규분포가 비현실적인 경우가 많다.

이와같이 이항자료인 경우, 기존의 판정기준을 사용하여 요인수를 결정하려할 때 여러 문제점이 있으므로 본논문에서는 붓스트랩 방법을 사용하여, 표본상관행렬의 고유값의 평균값을 구한 뒤, 경험적으로 요인수 결정 기준을 제시하고, 모의실험을 통하여 앞의 방법들과 비교하려 한다.

3.2 붓스트랩 방법을 이용한 요인수 결정을 위한 판정기준

여기서 제시하는 요인수 결정방법은 표본상관행렬의 고유값에 의한 기존의 Kaiser 방법을 적용하는데, 붓스트랩 방법을 이용하여 고유값의 평균을 사용한다. 이를 위한 알고리즘은 다음과 같다.

(붓스트랩 알고리즘)

1. 자료 $X_1, X_2, \cdots, X_p$ (X_j:nx1벡터, j=1, 2, $\cdots$, p)이 선택된다.(n=자료의 크기, p=문항의 수)
2. 1의 자료에서 붓스트랩 자료 $X_1^*, X_2^*, \cdots, X_p^*$를 선택한다.
3. 2의 표본상관행렬 R_1^*을 구한다

4. R_1^*의 고유값들 $\lambda_{11}^*, \lambda_{12}^*, \cdots, \lambda_{1p}^*$을 결정한다.(단, $\lambda_{11}^* \geq \lambda_{12}^* \geq \cdots \geq \lambda_{1p}^*$)
5. 2 ~ 4과정을 반복하여 다음과 같은 고유값들을 결정한다.(b=반복횟수)

$$\begin{array}{l} \lambda_{21}^*, \lambda_{22}^*, \cdots, \lambda_{2p}^* \\ \lambda_{31}^*, \lambda_{32}^*, \cdots, \lambda_{2p}^* \\ \quad\vdots \quad\quad\vdots \quad\quad\quad\vdots \\ \lambda_{b1}^*, \lambda_{b2}^*, \cdots, \lambda_{bp}^* \quad (\text{단, } \lambda_{k1}^* \geq \lambda_{k2}^* \geq \cdots \geq \lambda_{kp}^*,\ k=2, 3, \cdots, b) \end{array}$$

6. 평균치 $\bar{\lambda_j}^* = (\sum_{k=1}^{b} \lambda_{kj}^*)/b$

표준편차 $S_{\lambda_j}^* = [\{\sum_{k=1}^{b} (\lambda_{kj}^* - \bar{\lambda_j}^*)^2\}/(b-1)]^{1/2}$ 을 구한다.

7. 고유값(λ_j^*)의 평균이 0.9이상인 고유값 수효를 요인수로 하는 판정기준을 경험적으로 제시한다.

여기서 7번째 단계에서 사용한 0.9라는 기준치는 4절에 소개된 모의실험을 포함한 여러가지 경우에 대한 실험결과를 바탕으로 얻어진 값임을 밝힌다.

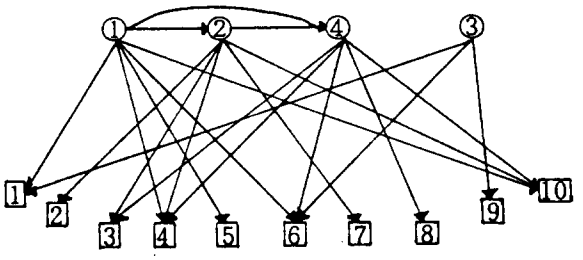
4. 모의실험결과

본 절의 목적은 3.2소절에서 기술한 알고리즘이 기존의 요인수 결정방법들과 어떻게 비교되는가를 경험적으로 보이는 데있다. 1절에서 언급하였듯이, 교육평가관련 연구에서 중요한 문제중의 하나로, 하나의 검사에 몇가지의 능력이 관련되어 있는가라는 문제가 있다. 이점을 부각시키기 위해서, 본 절에 소개된 모든 예는 문항점수와 문항이 묻는 능력과 관련지어서 서술하였다. 문항이 적을 때는 10개, 많을 때는 30개인 경우까지, 그리고 능력은 4개인 경우 부터 11개인 경우까지 고려하였다.

예 1

<그림 1>에서 ◎은 능력을 □는 문항을 표시하는데, 본절의 나머지 예에서도 같은 방법으로 표시하겠다. 능력간의 화살표는 일반적으로 선수관계(prerequisite relation)를, 능력과 문항간의 화살표는 인과관계를 나타낸다. 예컨데, 능력 A_1은 능력 A_2의 선수관계, 능력 A_1, A_2는 능력 A_4의 선수관계에 있음을 나타내고, 문항점수 X_1은 능력 A_1과 A_3의 영향을, 문항점수 X_2는 능력 A_2의 영향을 받는다는 것을 나타낸다. 또한, 화살표가 없는 경우는 확률적인 독립을 나타내는 것으로, 능력 A_1, A_2, A_4는 능력 A_3와 독립관계이다.

<표 1>은 <그림 1>의 능력과 문항에 대한 확률값을 정리한 것이다. 예컨대, 문항 1의 경우 A_1, A_3 능력 모두가 있으면, 이 문항을 맞출 확률은 0.99로 했고, 이 두 능력중 하나만 없어도 맞출 확률을 매우 작게했다. 그러나 이 값들은 임의로 정했으므로, <표 1>의 확률값 자체에 큰 의미는 없다.



<그림 1> ◎ 요인수(능력수) : m = 4, □ 문항수 : P = 10

이 예에서는 문항이 10개이고, 이것들이 4개의 능력들에 관해서 묻는 하나의 검사를 생각했다. 각 문항의 능력들과의 관계가 <그림 1>과 <표 1>에 주어져 있듯이, 문항 2, 5, 7, 8, 9는 하나의 능력만을, 문항 1,3은 두개의 능력을, 문항 4, 6, 10은 세개의 능력을 묻고있다. 이 4개의 능력과 10개의 문항에 대한 확률모형으로부터, 능력상태와 문항점수를 Monte Carlo 방법으로 생성하였다. 생성순서는 <그림 1>의 화살표방향으로 하면 된다. 즉 A_1, A_2, A_4, A_3, X_1, X_2,···, X_{10} 순이다. 처음에 A_1을 생성한 뒤, 만약에 A_1=1이면 A_2=1이 될 확률을 0.88이 되도록 생성한다. A_2=0이라 하자, 그러면 A_4 역시 A_4=1이 될 확률을 0.1이 되도록 생성한다. 문항점수 생성도 마찬가지 방법으로 해준다. 예컨대, A_1=1, A_2=0, A_4=1이라 가정하면 X_4=1이 될 확률이 0.2가 되도록 생성해준다. 이런 식으로 X_{10}까지 값 생성이 끝나면, 한 학생에 대한 능력상태와 문항점수값이 얻어진 것이다. 이 같은 생성과정을 n번 반복해서 크기 n인 모의자료를 얻어, 문항점수 부분만을 요인수 결정문제에 자료로 사용하였다. 이렇게 얻어진, 10개의 문항에 대한 n개의 자료에서 다시 n개의 붓스트랩 자료를 얻어 상관행렬의 고유값을 구하고, 또 다시 n개의 붓스
트랩 자료를 얻어 상관행렬을 구하는 등, 이런 과정을 b번 반복하여 고유값의 평균을 구하였다.

<표 2>는 모의자료의 크기가 100, 200, 400, 1000인 경우에 대해서 붓스트랩을 200번 반복하여 본 논문에서 제시한 붓스트랩방법과 다른 요인수 결정방법들을 비교한 것이다. <표 2>의 2번째 열부터 6번째 열은 붓스트랩에서 얻은 문항점수들에 대한 상관행렬의 고유값들의 평균값들을 크기순으로 나열했을때, 큰것부터 상위 5번째까지의 값들이다. <표 2>를 보면, 자료의 크기에 따라 선택된 요인의 갯수가 약간의 차이를 보이고 있다. 붓스트랩에 의한 방법은 실제 요인과 거의 일치함을 보이고, 최대우도법에 의한 방법들은 communality가 1이상인 경우 반복을 그만두므로 유의한 결과를 유추할 수가 없었다. 물론 HEYWOOD라는 option을 사용할 수 있으나, 김기영, 전명식(1989)이 언급했듯이 논란의 여지가 있으므로 사용하지 않았다.

여기서 하나 강조하고자 하는 것은 요인수는 능력과 능력, 그리고 능력과 문항과의 관계에 더 영향을 받지 <표 1>에서와 같은 확률값 자체에는 별 영향을 받지 않는다는 것이다. 물론 확률값에 따라서 두 확률변수 사이의 관계가 독립적으로 또는 종속적으로 된다. 그러나 이 절에서 사용된 모든 예에서는 일단 그림상으로 서로 종속적 관계로 나타났으면, 그 관계가 확률값으로도 유지되도록 확률값을 주었다. 이러한 조건하에서 확률값을 변화시키면, 변수들 사이의 관계가 유지되는 한 요인수에는 큰 변화가 없었다.

<표 1> <그림 1>의 능력과 문항관계도에 대한 확률표

$P(A_1=1)=0.48$
$P(A_2=1|A_1=0)=0.36$, $P(A_2=1|A_1=1)=0.88$
$P(A_3=1)=0.78$
$P(A_4=1|A_1,A_2)$

(A_1, A_2)	(0,0)	(0,1)	(1,0)	(1,1)
$P(A_4=1 \mid A_1, A_2)$	0.31	0.47	0.10	0.82

$P(X_1=1|A_1,A_3)$

(A_1, A_3)	(0,0)	(0,1)	(1,0)	(1,1)
$P(X_1=1 \mid A_1, A_3)$	0.0001	0.0001	0.08	0.99

$P(X_2=1|A_2=0)=0.005$, $P(X_2=1|A_2=1)=0.83$
$P(X_3=1|A_2,A_4)$

(A_2, A_4)	(0,0)	(0,1)	(1,0)	(1,1)
$P(X_3=1 \mid A_2, A_4)$	0.004	0.21	0.02	0.77

$P(X_4=1|A_1,A_2,A_4)$

(A_1, A_2, A_4)	(0,0,0)	(0,0,1)	(0,1,0)	(0,1,1)
$P(X_4=1 \mid A_1, A_2, A_4)$	0.0001	0.0001	0.72	0.8

(A_1, A_2, A_4)	(1,0,0)	(1,0,1)	(1,1,0)	(1,1,1)
$P(X_4=1 \mid A_1, A_2, A_4)$	0.2	0.2	0.8	0.99

$P(X_5=1|A_1=0)=0.25$, $P(X_5=1|A_1=1)=0.72$
$P(X_6=1|A_1,A_3,A_4)$

(A_1, A_3, A_4)	(0,0,0)	(0,0,1)	(0,1,0)	(0,1,1)
$P(X_6=1 \mid A_1, A_3, A_4)$	0.0001	0.2	0.16	0.16

(A_1, A_3, A_4)	(1,0,0)	(1,0,1)	(1,1,0)	(1,1,1)
$P(X_6=1 \mid A_1, A_3, A_4)$	0.0001	0.29	0.53	0.7

$P(X_7=1|A_2=0)=0.02$, $P(X_7=1|A_2=1)=0.76$
$P(X_8=1|A_4=0)=0.02$, $P(X_8=1|A_4=1)=0.98$
$P(X_9=1|A_3=0)=0.0001$, $P(X_9=1|A_3=1)=0.74$
$P(X_{10}=1|A_1,A_2,A_4)$

(A_1, A_2, A_4)	(0,0,0)	(0,0,1)	(0,1,0)	(0,1,1)
$P(X_{10}=1 \mid A_1, A_2, A_4)$	0.0001	0.05	0.02	0.5

(A_1, A_2, A_4)	(1,0,0)	(1,0,1)	(1,1,0)	(1,1,1)
$P(X_{10}=1 \mid A_1, A_2, A_4)$	0.09	0.07	0.1	0.81

<표 2> <그림 1>의 확률모형에 대한 요인수 비교

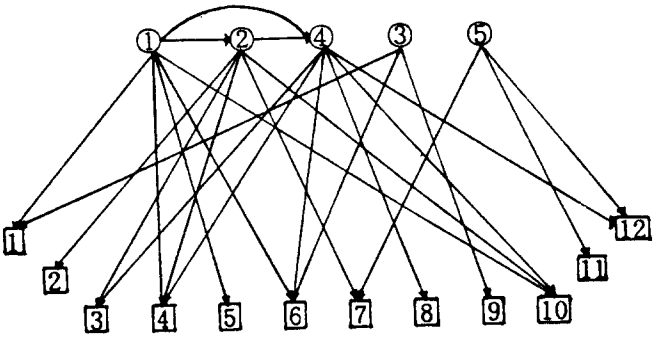
모의자료크기 (n)	고유값 평균 (괄호안은 붓스트랩 표준편차임)					판정 방법별 추천된 요인수				
	$\lambda_1 \geq$	$\lambda_2 \geq$	$\lambda_3 \geq$	$\lambda_4 \geq$	λ_5	Bootstrap (b=200)	Kaiser	유의성 검정	AIC	SBC
100	4.3028 (.3018)	1.5173 (.1631)	1.1792 (.0832)	.8562 (.0675)	.6455 (.0531)	3	3	3	3	3
200	4.1572 (.1890)	1.3753 (.0939)	1.1517 (.0849)	.9032 (.0740)	.6471 (.0504)	4	3	+	+	+
400	4.0661 (.1532)	1.3299 (.0731)	1.1775 (.0766)	.9210 (.0531)	.6890 (.0451)	4	3	+	+	+
1000	4.1260 (.0865)	1.2874 (.0523)	1.0628 (.0435)	.9589 (.0292)	.6981 (.0298)	4	3	4	4	3

+ : 두개 이상의 요인이 필요한 데 communality가 1이상이라 반복 과정이 중지됨.

이 절의 나머지 예들은 능력과 문항의 수, 그리고 이 확률변수들 사이의 관계를 변화시켜서, 본 논문에서 제시한 요인수 결정방법이 얼마나 효과적인가를 보였다. 아래 예들에서는, 위에서 언급한 확률값들에 대한 조건-즉, 그림상의 종속 또는 독립관계가 유지되도록 확률값이 주어진다는-하에서, 확률값들이 주어졌으며, 이 확률값들이 특별한 의미가 없기 때문에 표로 만들지 않았다. 다만, 어떤 능력에 대해서 선수관계에 있는, 또는 어떤 문항에 의해서 요구되는 능력들의 상태가 좋으면, 이 능력들에 영향을 받는 능력 또는 문항점수가 좋게 되는 확률값을 상대적으로 높게 잡아 준다는 것을 원칙으로 하였다.

예 2.

이 예에서는 능력 5개, 문항 12개의 확률모형을 고려했는데, 모의자료의 크기가 1000인 경우에 대해서 200번의 붓스트랩을 반복했다. <표 3>에 표시된 것처럼, 본 논문에서 제시한 붓스트랩방법과 Kaiser 방법만이 요인수를 제시했다. <표 3>의 제 2열부터 제 6열까지는 12개 문항점수에 대한 표본상관행렬의 고유값의 붓스트랩 평균값중에서 큰 순서로 2번째 값부터 6번째 값들이다.



<그림 2> ◎ 요인수(능력수) : m = 5, □ 문항수 : P = 12

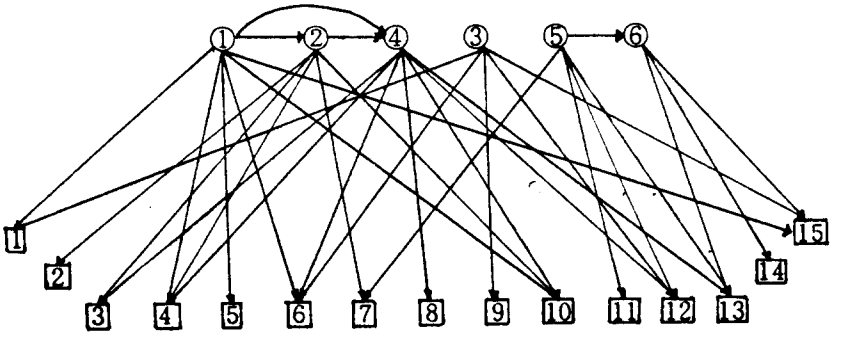
<표 3> <그림 2>의 확률모형에 대한 요인수 비교

모의자료크기 (n)	고유값 평균 (괄호안은 붓스트랩 표준편차임)					판정 방법별 추천된 요인수				
	$\lambda_2 \geq \lambda_3 \geq \lambda_4 \geq \lambda_5 \geq \lambda_6$					Bootstrap (b=200)	Kaiser	유의성 검정	AIC	SBC
1000	1.5402 (.0447)	1.3019 (.0481)	1.0849 (.0468)	.9300 (.0330)	.7874 (.0296)	5	4	+	+	+

+ : 네개 이상의 요인이 필요한 데 communality가 1이상이라 반복 과정이 중지됨.

예 3.

이 예에서는 능력6개, 문항13개의 확률모형을 고려했는데, 모의자료의 크기가 1000인 경우에 대해서 200번의 붓스트랩 반복을 사용했다. <표 4>의 제 2열부터 제 6열까지는 13개 문항점수에 대한 표본상관행렬의 고유값의 붓스트랩 평균값중에서 큰 순서로 2번째 값부터 6번째 값들이다. 유의성 검정, AIC는 최대우도 추정법에서 요구되는 여러 가정들 하에서 만족스런 결과를 기대할 수 있고, 유의성 검정은 요인을 더 추가하는 경향이 있는 점을 고려할때, 실제 요인수와 같은 결과가 나온 것이 매우 흥미롭다.



<그림 3> ◎ 요인수(능력수) : m = 6, □ 문항수 : P = 15

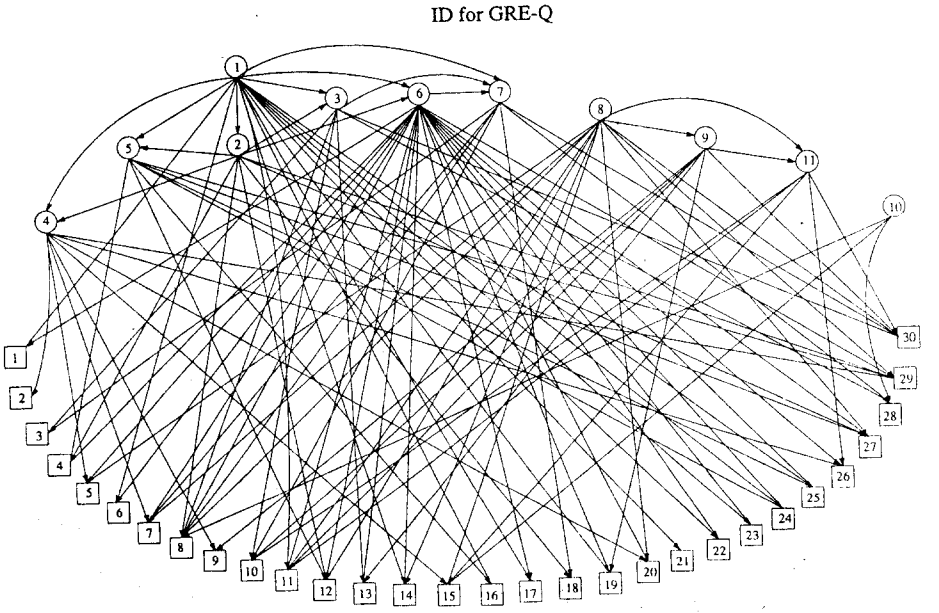
<표 4> <그림 3>의 확률모형에 대한 요인수 비교

모의자료크기 (n)	고유값 평균 (괄호안은 붓스트랩 표준편차임)					판정 방법별 추천된 요인수				
	λ_2 ≥	λ_3 ≥	λ_4 ≥	λ_5 ≥	λ_6	Bootstrap (b=200)	Kaiser	유의성 검정	AIC	SBC
1000	2.2578 (.0790)	1.4693 (.0563)	1.1716 (.0551)	.9300 (.0412)	.7612 (.0290)	5	4	6(5+)	6	5

+ : p=0.0596으로 유의수준 0.05하에서는 5가 선택됨.

예 4.

이 예는 Kim & Wang(1995)의 자료(ID for GRE-Q : 그림 4)를 적용한 것으로 능력 11개, 문항 30개의 확률모형을 고려한 것인데, 모의자료의 크기가 500인 경우에 대해서 100번의 붓스트랩 반복을 했다. 다른 예보다 모의자료의 크기와 붓스트랩 반복횟수가 작은 것은 본 연구에서 사용된 컴퓨터의 메모리 용량상 그렇게 한 것이다. 뒤의 두 예에서도 같은 이유로 자료크기와 붓스트랩 반복횟수가 제한되었다. <표 5>의 제 2열부터 제 6열까지는 30개 문항점수에 대한 표본상관행렬의 고유값의 붓스트랩 평균값중에서 큰 순서로 8번째 값부터 12번째 값들이다. 붓스트랩에 의한 방법은 실제 요인수와 같은 결과가 나왔으며, 최대우도법에 의한 방법들은 실제 요인수와 상당한 차이를 보였다.



<그림 4> ◎ 요인수(능력수) : m = 11, □ 문항수 : P = 30

<표 5> <그림 4>의 확률모형에 대한 요인수 비교

모의자료크기 (n)	고유값 평균 (괄호안은 붓스트랩 표준편차임)					판정 방법별 추천된 요인수				
	$\lambda_8 \geq \lambda_9 \geq \lambda_{10} \geq \lambda_{11} \geq \lambda_{12}$					Bootstrap (b=100)	Kaiser	유의성 검정	AIC	SBC
500	1.1082 (.0375)	1.0443 (.0288)	.9885 (.0259)	.9439 (.0252)	.8945 (.0246)	11	8	$9^{+}(8^{++})$	8	4

\+ : 9가 충분하나 11개까지로도 유의함
++ : p=0.042로 유의수준 0.01하에서 8개가 선택됨

예 5.

이 예에서는 <그림 4>에서 능력 A_7과 문항 X_3, X_7, X_8, X_{10}, X_{18}, X_{28}, X_{30}을 제외하여 능력 10개, 문항 23개의 확률모형을 고려했고,모의자료의 크기가 600인 경우에 대해서 100번의 붓스트랩 반복을 했다.

<표 6> <그림 4>에서 능력 A_7과 문항 X_3, X_7, X_8, X_{10}, X_{18}, X_{28}, X_{30}을 제외한 확률모형에 대한 요인수 비교

모의자료크기 (n)	고유값 평균 (괄호안은 붓스트랩 표준편차임)					판정 방법별 추천된 요인수				
	$\lambda_7 \geq \lambda_8 \geq \lambda_9 \geq \lambda_{10} \geq \lambda_{11}$					Bootstrap (b=100)	Kaiser	유의성 검정	AIC	SBC
600	1.0350 (.0282)	.9770 (.0269)	.9296 (.0238)	.8830 (.0231)	.8422 (.0211)	9	8	$8^{+}(7^{++})$	7	4

\+ : 8이 충분하나 10개까지로도 유의함
++ : p=0.0463로 유의수준 0.01하에서 7개가 충분함.

예 6.

이 예는 <그림 4>를 예 5에서 제거한 것에 다시 능력 A_{11}과 문항 X_{11}, X_{15}, X_{26}을 제외하여 9개의 능력과 20개의 문항을 고려한 것이다. 최대우도법에 의한 방법들은 실제 요인수와 상당한 차이를 보인다.

<표 7> <그림 4>에서 능력 A_7, A_{11}과 문항 X_3, X_7, X_8, X_{10}, X_{11}, X_{15}, X_{18}, X_{26}, X_{28}, X_{30}을 제외한 확률모형에 대한 요인수 비교

모의자료크기 (n)	고유값 평균 (괄호안은 붓스트랩 표준편차임)					판정 방법별 추천된 요인수				
	$\lambda_6 \geq$	$\lambda_7 \geq$	$\lambda_8 \geq$	$\lambda_9 \geq$	λ_{10}	Bootstrap (b=150)	Kaiser	유의성 검정	AIC	SBC
700	1.0540 (.0266)	.9930 (.0243)	.9419 (.0240)	.8923 (.0238)	.8463 (.0234)	8	5	6+	7	4

+ : 6가 충분하나 8개까지로도 유의함. 9개 부터는 communality가 1 이상으로 반복과정이 중단됨.

이상 6개의 예에서 붓스트랩 방법이 평균적으로 실제 요인수와 제일 가까운 값을 추천하는 것으로 나타났다. 특히 요인수가 클수록 이 붓스트랩 방법은 사용된 다른 방법보다 더 효과적인 것으로 나타났다. 이절의 모의 실험들에 사용된 프로그램 언어로는 Fortran Ver 5.0을, 고유값에 관련된 계산은 IMSL Ver 1.1, 함수 부프로그램 RNGET, RNUN, CORVC,EVLSF을, 요인수에 대한 전통적인 판정기준들로는 SAS Ver 6.04를 사용하였다.

5. 결 론

본 논문에서는 이항자료에 대한 요인수 결정에 있어서, 분산·공분산들의 축소현상, 연속확률모형하에서의 제반조건들의 비현실성 등의 문제점들을 중요시 했다. 이와같은 문제점들로 인해서, 연속확률모형을 전제한 기존의 요인수 결정방법들은 부적절하다. 그중에서 Kaiser의 방법은 모든 관측변수들의 독립성을 가정한 확률모형과의 비교에 의해서 요인수를 결정하는 비교적 간단한 방법이며 또, 요인수 결정과정에서는, 다른 방법들에 비해서, 확률모형에의 의존도가 낮다. 이런 점들로 볼 때, Kaiser의 방법에서의 기본 이론을 이항자료에 적용하는데 비교적 무리가 적다고 본다. 다만 Kaiser의 방법을 이항자료에 적용했을 때, 요인수 결정을 위한 고유값의 기준이 1보다 작으리라는 것이 우리의 예상이었다. 그리하여 여러 모의실험결과, 판정기준을 고유값이 0.9이상인 것들 만큼 수를 요인수로 결정하니 실제 또는 실제에 가까운 요인수를 얻을 수 있었다. 특히 관측이항변수의 갯수가 커질수록 본 논문에서 사용한 방법이 다른 기존의 방법들보다 더 효과적이었다.

기존 방법들 중 유의성 검정방법은 유의한 요인수를 결정할 때까지, AIC와 SBC는 최적의 요인수를 결정할 때까지 계속적인 수행이 필요하며, Heywood 현상의 영향을 많이 받았다. 본 논문의 붓스트랩 방법은 변수가 많아지면(예컨대 60이상) 그 만큼 많은 시간이 필요하지만, 이 방법의 효과에 비하면, 계산시간은 별로 큰 문제가 아니라고 본다.

본 논문에서 제안한 이항자료에 대한 요인수 결정방법은 안정적인 통계치인 고유값들의 평균치의 사용을 강력히 권한다.

감사의 말

본 논문을 자상하게 심사하여 주신 두분의 심사분께 깊은 감사를 드린다. 수정과정에서 그 분들의 심사평이 매우 큰 도움이 되었다.

참 고 문 헌

[1] 김기영, 전명식 (1989). 「SAS 인자분석」, 자유아카데미.

[2] Bartlett, M.S. (1954). A Note on Multiplying Factors for Various Chi-Square Approximations, *Journal of the Royal Statistical Society, Series B,* Vol. 16, 196-198.

[3] Cartell, L. S. & Harman, A. J. (1966). The Scree Test for the Number of Factors *Multivariate Behavioral Research,* Vol. 1, 245-276.

[4] Kim, J-O & Charles W. (1978). *Mueller Factor Analysis-Statistical Methods and Practical Issues,* University of Iowa.

[5] Kaiser, H. F. & Rice, J. (1974). Little Jiffy, Mark IV *Educational and Psychological Measurement,* Vol. 34, 111-117.

[6] Richard A, Johnson & Dean W. Wichern (1982). *Applied Multivariate Statistical Analysis* Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, U.S.A.

[7] Kim, S-H. & Wang, M. (1995). Influence Diagram of Test Performance in GRE-Q, An ETS Research Report.

A Bootstrap Approach for Factor Numbers in Binary Data

SungHo Kim[3], MiSook Jeong[4]

Abstract

A method of determining the factor numbers is explored in this paper, when data and the factors are binary. We applied a bootstrap approach and proposed a criterion for the method. Simulation results suggest that the proposed method in this paper is very useful in determining the factor numbers for binary data and factors.

3) KAIST, School of Humanities and General Sciences, Division of Basic Sciences, Taejeon 305-701, KOREA.

4) Pusan National Univ, Dept. of Statistics, Pusan 609-735, KOREA.