응용통계연구 제 **12**권 *1*호, *1999*년, *pp. 165–173*

인자분석에 의한 부분수량화

서혜선[1)]

요 약

사회과학분야에서 인자분석을 실시하는 경우 제1 인자의 사전정의는 연구대상분야에서의 확립된 지식 또는 전제에 의해 나온다. 예컨대 고등학교 학생들을 대상으로 국어, 영어, 수학, 사회, 과학 등의 과목성적을 측정하여 인자분석을 하는 경우 사전적으로 제1 인자를 이 과목들의 단순평균에 의해 생성되는 일반지능으로 정의하는 것이 자연스럽다. 즉 다양한 사회과학 분야에서 인자분석을 행함에 있어 선행 연구나 연구자의 사전지식 또는 가설 등에 의해 제1 인자를 전제하는 경우 후속되는 나머지 인자를 자료에서 찾아냄으로써 인자분석을 완성하고자 하는 것이 본 연구의 기본 내용이다.

1. 소개

인자분석(factor analysis)은 정보의 손실을 최소화하는 자료의 차원축약 기법으로써 원변량들의 공분산행렬이나 분산행렬을 통해 소수 몇 개의 공통인자를 찾는 것으로 Karl Pearson과 Charles Spearman 등에 의해 소개되었다 (Johnson 과 Wichern, 1982). Mardia, Kent와 Bibby (1979), Anderson (1984) 등도 이와 비슷하게 관찰되지 않는 인자들의 선형결합을 찾는 방법으로 인자분석을 설명하고 있다.

영국과 미국에서의 인자분석과는 달리 프랑스의 Lebart (1984)는 유클리디안 공간상에서 다변량 자료의 기술적인 분석의 틀로서 인자분석을 사용하고 있다.

본 연구에서는 선행 연구나 연구자의 사전 지식 등에 의해 주어진 자료에 선행적으로 제1 인자를 전제하는 경우 나머지 인자들을 어떻게 자료에서 찾을 수 있는가를 생각해 보고자 하는 것이다. 그렇게 함으로써 배경지식과 자료탐색의 조화를 꾀할 수 있을 것이다.

여기서 "부분수량화"란 다변량자료에서 사전적으로 주어진 제1 인자를 제외하고 나머지 인자들을 찾는 수량화 절차를 의미하는 것으로 쓰기로 한다.

제 2절에서 부분수량화 절차를 제안하고 이에 따른 행렬분해 및 관련된 통계적 성질을 제 3절에서, 그리고 행렬도 표현을 제 4절에서 보이고자 한다. 인자분석의 부분수량화에 대한 수치 예는 제 5절에서 다루어 질 것이다.

본 연구에서의 표기방식은 다음과 같다.

$$X : n \times p \text{ 대상행렬}, \quad X = \begin{pmatrix} \vec{x}_1^t \\ \vdots \\ \vec{x}_n^t \end{pmatrix} = (\mathbf{x}_1, \cdots, \mathbf{x}_p).$$

1) (136-701) 서울시 성북구 안암동 5가 1; 고려대학교 통계연구소, 연구원

· 따라서 $\vec{x}_1, \cdots, \vec{x}_n \in R^p$; $\mathbf{x}_1, \cdots, \mathbf{x}_p \in R^n$.

여기서 대상행렬 X는 중심화가 된 것이며 자료의 성격에 따라 척도화가 필요하다 하겠다.

2. 부분수량화

단위 길이를 같는 상수벡터 $\vec{c}_1 = (c_1, \cdots, c_p)^t$ 가 주어진 경우 제1 인자점수벡터는 다음과 같다.

$$\mathbf{a}_1 = \frac{c_1\mathbf{x}_1 + \cdots + c_p\mathbf{x}_p}{\| c_1\mathbf{x}_1 + \cdots + c_p\mathbf{x}_p \|}$$

이제 다변량 자료에서 제1 인자를 제외하고 나머지 부분에 대한 인자분석을 하기 위하여 적당한 상수 k_j에 대해 $\mathbf{x}_j$를 다음과 같이 두 부분으로 분해한다.

$$\mathbf{x}_j = k_j\mathbf{a}_1 + (\mathbf{x}_j - k_j\mathbf{a}_1), \quad j = 1, \cdots, p.$$

여기서 k_j는

$$k_j = \mathbf{a}_1^t\mathbf{x}_j, \quad j = 1, \cdots, p.$$

따라서

$$\mathbf{x}_j - k_j\mathbf{a}_1, \quad j = 1, \cdots, p$$

에 대한 일반적인 인자분석을 재 실시할 수 있다. 이러한 인자분석의 부분수량화 절차는 다음과 같다.

단계 1 : 단위벡터 $\mathbf{a}_1$이 제1 인자점수의 정의로 주어진 경우 후속 수량화를 위해 X를 각각 다음과 같이 대치한다.

$$X_{(1)} = (I_n - P_{\mathbf{a}_1})X. \tag{2.1}$$

여기서 $P_{\mathbf{a}} = \mathbf{a}(\mathbf{a}^t\mathbf{a})^{-1}\mathbf{a}^t$, 즉 $\mathbf{a}$로의 사영행렬

단계 2 : 임의의 $p \times 1$ 벡터 $\vec{c}_2 = (c_1, \cdots, c_p)^t$ 에 대하여

$$max \| X_{(1)}\vec{c}_2 \|^2$$

그 결과로 인자점수벡터

$$\mathbf{a}_2 = X_{(1)}\vec{c}_2 \,/\, \| X_{(1)}\vec{c}_2 \|$$

를 얻게 되고 앞 단계에서 넘어온 행렬 $X_{(1)}$ 을 다음과 같이 대치한다.

$$X_{(2)} = (I_n - P_{\mathbf{a}_2})X_{(1)}. \tag{2.2}$$

단계 3 : 이러한 과정을 단계 p까지 수행하여 인자점수벡터 $\mathbf{a}_2$, $\cdots$, $\mathbf{a}_p$를 획득한다.

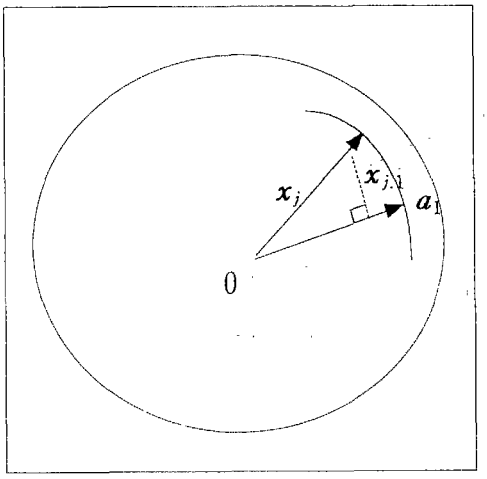


그림 2.1: 부분수량화 단계 1

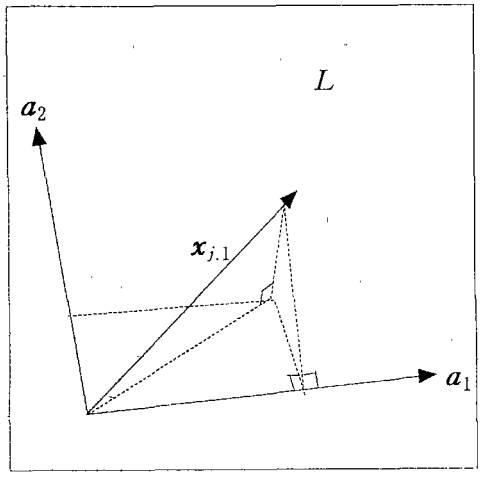


그림 2.2: 부분수량화 단계 2

이 과정을 플롯으로 표현해보면 그림 2.1, 그림 2.2와 같다. 척도화된 자료행렬 X에서 p개의 각 변수벡터들은 단일구상에 존재하게 되며 p개의 변수벡터를 대표하는 단위벡터 $\mathbf{a}_1$이 주어졌다고 하자. 그리고 $\mathbf{x}_{j.1}$을 $X_{(1)}$의 j번째 열이라 하면 그림 2.1로부터 $\mathbf{x}_{j.1}$은 $\mathbf{a}_1$과 서로 직교함을 알 수 있다. 이제 그림 2.2에서 또다른 단위벡터 $\mathbf{a}_2$를 생각한다. 그림 2.1에서 $X_{(1)}$의 모든 열이 $\mathbf{a}_1$과 직교하고 $\mathbf{a}_2$가 $X_{(1)}$의 열들을 선형결합한 것이므로 $\mathbf{a}_1^t\,\mathbf{a}_2 = 0$ 이 된다. 이를 일반화시키면 인자분석에서의 부분수량화의 경우 인자점수벡터들은 $\mathbf{a}_j^t\,\mathbf{a}_{j'} = 0\ (j \neq j',\ j,\ j' = 1\ ,\ \cdots\ ,\ p)$ 을 자동적으로 만족한다. 그리고 제 2축 인자점수들은

$$\mathbf{x}_{j.1}{}^t\,\mathbf{a}_2 \;=\; (\mathbf{x}_j \;-\; k_1\,\mathbf{a}_1)^t\,\mathbf{a}_2 \;=\; \mathbf{x}_j{}^t\,\mathbf{a}_2$$

를 만족하고 모든 $\mathbf{a}_j$ $(j = 1\ ,\ \cdots\ ,\ p)$가 직교하므로

$$\mathbf{x}_{j.1,\cdots,j-1}{}^t\,\mathbf{a}_j \;=\; (\mathbf{x}_j \;-\; \sum_{l=1}^{j-1} k_l\,\mathbf{a}_l)^t\,\mathbf{a}_j \;=\; \mathbf{x}_j{}^t\,\mathbf{a}_j\,,\quad j = 2\ ,\ \cdots\ ,\ p\ ,\ i = 1\ ,\ \cdots\ ,\ n$$

이 된다. 여기서 $\mathbf{a}_j$는

$$X_{(1)}\,X_{(1)}{}^t\,\mathbf{a}_j \;=\; \xi_j\,\mathbf{a}_j,\quad \xi_2 \;\geq\; \cdots \;\geq\; \xi_p$$

에 의해 획득된다.

3. 행렬분해 및 관련된 통계적 성질

2절의 조작에 따른 행렬 X의 분해를 생각해보자. 단계 1에서 식 (2.1)로부터

$$X = \mathbf{a}_1 \mathbf{a}_1^t X + X_{(1)}$$

으로 표현되며, 단계 2에서 식 (2.2)로부터

$$\begin{aligned} X_{(2)} &= (I_n - P_{\mathbf{a}_2}) X_{(1)} \\ &= X - \mathbf{a}_1 \mathbf{a}_1^t X - \mathbf{a}_2 \mathbf{a}_2^t X \end{aligned}$$

로 표현할 수 있다. 따라서 $X_{(3)}, \cdots, X_{(p)}$ 등을 고려하여 전개해 가면

$$X = \mathbf{a}_1 \mathbf{a}_1^t X + \mathbf{a}_2 \mathbf{a}_2^t X + \cdots + \mathbf{a}_{p-1} \mathbf{a}_{p-1}^t X + \mathbf{a}_p \mathbf{a}_p^t X + X_{(p)}$$

가 된다. 이때 X는 단계 p 까지의 과정을 통한 p개 성분에 의해 모두 표현되므로 남아 있는 $X_{(p)}$ 는 $\mathbf{0}$ 행렬이다. 따라서

$$X = \sum_{j=1}^{p} \mathbf{a}_j \mathbf{a}_j{}^t X.$$

이제 l_j와 인자패턴벡터 $\vec{c}_j$를 각각

$$\begin{aligned} l_j &= \mathbf{a}_j^t X X^t \mathbf{a}_j, \quad j = 1, \cdots, p, \\ \vec{c}_j &= X^t \mathbf{a}_j / \sqrt{l_j}, \quad j = 1, \cdots, p \end{aligned}$$

로 정의하자. 그러면

$$X = \sum_{j=1}^{p} \sqrt{l_j}\, \mathbf{a}_j\, \vec{c}_j{}^t = A D_{\sqrt{l}} C^t \tag{2.3}$$

가 되며 이때 $\mathbf{a}_j$는 인자점수벡터이다. 여기서

$$A = (\mathbf{a}_1, \cdots, \mathbf{a}_p), \; C = (\vec{c}_1, \cdots, \vec{c}_p), \; D_{\sqrt{l}} = diag(\sqrt{l_1}, \cdots, \sqrt{l_p}).$$

$\mathbf{a}_j{}^t \mathbf{a}_{j'} = 0 \; (j \neq j')$ 이므로 $A^t A = I_p$ 가 된다. 그리고 $j, j' = 2, \cdots, p \; (j \neq j')$ 에 대하여

$$\vec{c}_j{}^t \vec{c}_{j'} \propto \mathbf{a}_j{}^t X X^t \mathbf{a}_{j'} = \mathbf{a}_j{}^t X_{(1)} X_{(1)}^t \mathbf{a}_{j'} = \xi_j \mathbf{a}_j{}^t \mathbf{a}_{j'} = 0.$$

그러나 일반적으로,

$$\vec{c}_1{}^t \vec{c}_j \neq 0, \quad j = 2, \cdots p.$$

결론적으로

$$A^t A = I_p \quad \text{그리고} \quad C^t C = \begin{pmatrix} 1 & \mathbf{n}^t \\ \mathbf{n} & I_{p-1} \end{pmatrix}.$$

여기서 $\mathbf{n} = (n_2, \cdots, n_p)$, $n_j = \vec{c}_1{}^t \vec{c}_j \; (j = 2, \cdots, p)$. 따라서 식 (2.3)을 행렬 X의 유사비정칙치분해라 하고 l_j 를 유사고유값이라 할 수 있겠다.

4. 행렬도 표현

1971년 Gabriel에 의해 소개된 행렬도(biplot)는 n개의 개체와 p개 변량들의 상대적인 위치를 동시에 그래프로 나타내는 기법이다. 행렬도의 타당도는 다음과 같은 자료행렬 X의 행렬분해 식에 기초한다.

$$X = GH^t. \tag{2.4}$$

여기서 식 (2.4)는 다음과 같은 행렬 X의 비정칙치분해(singular value decomposition)에 대한 몇가지 구성방법으로부터 얻어진다.

$$X = MD_{\sqrt{\xi}}N^t \tag{2.5}$$

여기서 $M^tM = N^tN = I_p$, $D_{\sqrt{\xi}} = diag(\sqrt{\xi_1}, \cdots, \sqrt{\xi_p})$, $\xi_1 \geq \cdots \geq \xi_p$. 부분수량화를 적용한 후에는 비정칙치분해인 식 (2.5)보다는 유사비정칙치분해인 식 (2.3)을 얻게 된다. $G = A$ 그리고 $H = CD_{\sqrt{l}}$을 취함으로써 식 (2.4)를 얻을 수 있다. 즉 행렬 X의 x_{ij} 는

$$x_{ij} = \vec{g}_i{}^t\vec{h}_j, \quad i = 1, \cdots, n, \; j = 1, \cdots, p$$

로 표현되어 진다. 여기서 $\vec{g}_i{}^t$는 행렬 G의 i번째 행벡터이며 $\vec{h}_j$는 행렬 H 의 j번째 행벡터이다. 벡터 $\vec{g}_i{}^t$는 '행효과' 로, $\vec{h}_j$는 '열효과' 로 고려되며 각각 행렬 X의 각 n개 행과 p개의 열에 대응한다 (Gabriel, 1971). 여기서 G는 $n \times p$ 행렬이며 H는 $p \times p$ 행렬로 각각의 계수는 p 이다. 이때

$$x_{ij(r)} \approx \vec{g}_{i(r)}{}^t\vec{h}_{j(r)}, \quad i = 1, \cdots, n, \; j = 1, \cdots, p \tag{2.6}$$

으로 $\vec{g}_{i(r)}$과 $\vec{h}_{j(r)}$은 각각 $\vec{g}_i$와 $\vec{h}_j$의 처음 r $(r < p)$ 개 원소를 갖는다. 따라서 $\vec{g}^t_{i(r)}$는 $r-$차원 부분공간에서 행렬 X의 n개 행으로 그리고 $\vec{h}_{j(r)}$은 $r-$ 차원 부분공간에서 행렬 X의 p개 열로서 표현되어진다. 결과적으로 식 (2.6)은 다음과 같이 표현될 수 있다.

$$X = G_{(r)}H_{(r)}{}^t,$$

여기서 $G_{(r)}$ 과 $H_{(r)}$은 각각 행렬 G와 H의 처음 r개 열을 갖는다.

따라서 $r-$차원으로 축소된 열수량화의 행렬분해 근사도는

$$GOA_{col(r)} = \frac{\| X^tA_{(r)} \|^2}{\| X^tA \|^2} = \frac{\| C_{(r)}D_{\sqrt{\lambda_{(r)}}} \|^2}{\| CD_{\sqrt{l}} \|^2} = \frac{\sum_{j=1}^{r} l_j}{\sum_{j=1}^{p} l_j}$$

로 정의할 수 있다. 왜냐하면

$$\begin{aligned} \| CD_{\sqrt{l}} \|^2 &= tr(D_{\sqrt{l}}C^tCD_{\sqrt{l}}) \\ &= tr\left(D_{\sqrt{l}}\begin{pmatrix} 1 & \mathbf{n}^t \\ \mathbf{n} & I_{p-1} \end{pmatrix}D_{\sqrt{l}}\right) = \sum_{j=1}^{p} l_j \end{aligned}$$

가 되기 때문이다. 인자분석의 부분수량화는 R^n 공간(열공간)에서의 부분수량화이므로 행공간에서의 근사도는 생각할 필요가 없다. 인자분석에서의 부분수량화가 열공간 근사에 관한 차원축소기법이기 때문이다.

5. 수치예

인자분석에서의 부분수량화 방법론을 설명하기 위하여 du Toit, et al. (1986)의 재능검사 자료를 고려하고자 한다. 남아프리카의 "재능 조사계획" 연구에서 얻어진 이 자료는 다양한 재능검사에 대한 18개 변량(25점 만점)을 2800 명의 백인 학생으로부터 측정한 것이다. 시험점수들은 교육수준 I (8 년간의 교육기간), 교육수준 II (교육 수준 I + 2 년간의 교육기간), 교육수준 III (교육수준 II + 2 년간의 교육기간) 등 3수준에서 얻어진 것이다. 본 연구에서는 28 명의 부분자료에 대하여 부분수량화를 적용해 보고자 한다. 본 연구에서 사용된 변수들은 다음과 같다.

X_1, X_7, X_{13} : 숫자배열, X_4, X_{10}, X_{16} : 단어 짝짓기,
X_2, X_8, X_{14} : 숫자계산, X_5, X_{11}, X_{17} : 말의 논리성,
X_3, X_9, X_{15} : 문단완성, X_6, X_{12}, X_{18} : 단어 추론.

제1 인자점수벡터를 인지능력으로 하여 모든 18개 변량들의 단순평균에 의해 생성된 것으로 사전정의하였을 경우 부분수량화 결과는 다음과 같다. 표 6.1의 결과 2차원으로 축소된 열수량화점의 근사도는 56.6%이며 그 중에서 제 2축에 의한 설명력은 약 20% 정도 기여함을 수 있다. 그림 6.1의 행 플롯은 제 1축을 기준으로 살펴볼 때 원점에서 우측에 위치하는 학생들의 경우 재능검사에서 높은 점수를 받은 학생들이며 좌측에 위치한 학생들은 점수가 낮음을 의미한다. 행플롯과 열플롯을 겹쳐보면 제 2축을 기준으로 볼 때 원점에서 상부에 위치하는 학생들의 경우 숫자배열($X7$, $X13$) 및 숫자계산 ($X14$) 등의 재능이 있는 학생들임을 의미한다. 특히 원점에서 좌측 상부에 위치한 학생들은 재능검사에서 점수가 낮은 학생들이지만 그중에서 수리능력 점수가 언어능력 점수보다는 높은 학생임을 의미한다. 반면 원점에서 하부에 위치한 학생들은 단어 짝짓기 ($X10$, $X4$) 및 단어추론 ($X6$) 등의 점수가 우수한 학생들임을 의미한다. 따라서 제 2축은 수리적인 능력과 언어적 능력을 구분짓는 축으로 표현된다 하겠다.

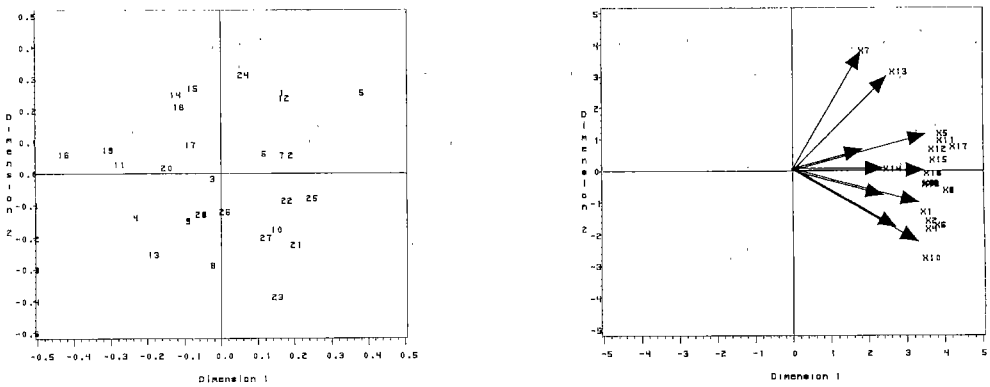
6. 결 론

지금까지 행해진 대부분의 인자분석은 주어진 자료에 의존하여 정보의 손실을 최소화하는 소수 몇 개의 인자를 찾는 데 주안점을 두어왔다. 그러나 사회과학분야 예컨대 심리학이나 교육학 등의 분야에서는 여러 다변량자료의 인자분석에 있어 연구자들이 선험적 경험이나 연구 등에 의해 제1 인자를 인지하는 경우가 많다.

따라서 본 연구의 내용은 이러한 관련분야 연구자들의 확립된 지식 및 견해를 존중하는 관점에서 제1 인자를 자료에 선행하여 전제하고 후속되는 나머지 인자를 자료에서 찾아내고자 하는 것이었다.

표 6.1: 부분수량화에 의한 인자패턴벡터와 유사고유값

변량	인자패턴벡터	
	$\vec{c}_1$	$\vec{c}_2$
X_1	0.229	-0.188
X_2	0.237	-0.229
X_3	0.233	-0.054
X_4	0.237	-0.267
X_5	0.258	0.174
X_6	0.253	-0.247
X_7	0.121	0.559
X_8	0.268	-0.088
X_9	0.242	-0.055
X_{10}	0.234	-0.401
X_{11}	0.262	0.142
X_{12}	0.247	0.098
X_{13}	0.180	0.460
X_{14}	0.168	0.007
X_{15}	0.249	0.051
X_{16}	0.235	-0.060
X_{17}	0.285	0.110
X_{18}	0.239	-0.011



1) 행 플롯 2) 열 플롯

그림 6.1: 부분수량화 플롯

제 2절에서는 인자분석에서의 부분수량화 방법론을 제안하였으며 이와 유사하게 주성분분석에서의 부분수량화를 구축할 수 있다. 인자분석은 연관성이 큰 변수들을 묶어서 그것을 인자라고 하는 잠재적 변인으로 해석한다. 따라서 주성분분석에서의 부분수량화에 있어 주 목표가 행공간(개체공간)에서의 개체간의 관련성을 보는 것이라면 인자분석의 주 목표는 열공간(변수공간)에서의 변수간의 관련성을 보는 것이 큰 차이점이라 할 수 있다. 이에 대한 자세한 내용은 서혜선(1997)및 Suh 와 Huh(1997)를 참조할 수 있다.

아울러 제1 인자뿐만 아니라 제2 인자까지 사전적으로 주어지는 경우에 있어서도 본 연구에서 제안한 부분수량화 알고리즘이 일부의 수정만으로 적용이 가능하다 하겠다.

참고문헌

[1] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis.* Second Edition. Wiley, New York.

[2] du Toit, S. H. C., Steyn, A. G. W. and Stumpf, R. H. (1986). *Graphical Exploratory Data Analysis.* Springer-Verlag, New York.

[3] Gabriel, K. R. (1971). The biplot graphical display of matrices with applications to principal component analysis. *Biometrika,* **58**, 453–467.

[4] Lebart, L., Morineau, A. and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices.* Wiley, New York.

[5] Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, London.

[6] Suh, H. S. and Huh, M. H. (1997). Partial Quantification in Principal Component Analysis. *The Korean Communications in Statistics,* **4**, 637–644.

[7] 서혜선 (1997). 〈사회연구를 위한 세가지 측면에서의 통계적 방법의 개발과 응용〉 박사학위 논문, 고려대학교.

[1998년 6월 접수, 1998년 12월 최종수정]

Partial Quantification in Factor Analysis

Hye-Sun Suh [1)]

ABSTRACT

Sometimes, the first factor may come logically from the established knowledge and premises. For example, for the high school students' test scores of Korean, English, Mathematics, Social Study, and Science, it is natural to define the first factor as the general intelligence. Thus the aim of this study is to find the remaining factor when the first factor is defined by related researchers' view and background knowledge.

1) Postdoctoral Researcher, Institute of Statistics, Korea University. Seoul 136-701, Korea.