The Korean Communications
in Statistics Vol. 11 No. 1, 2004
pp. 49-58

Asymptotic Test for Dimensionality in Probabilistic Principal Component Analysis with Missing Values

Chong-sun Park[1]

Abstract

In this talk we proposed an asymptotic test for dimensionality in the latent variable model for probabilistic principal component analysis with missing values at random. Proposed algorithm is a sequential likelihood ratio test for an appropriate Normal latent variable model for the principal component analysis. Modified EM-algorithm is used to find MLE for the model parameters. Results from simulations and real data sets give us promising evidences that the proposed method is useful in finding necessary number of components in the principal component analysis with missing values at random.

Keywords : Probabilistic Principal Component Analysis, Dimension Reduction, Latent Variable Model, Missing Values.

1. 서론

다변량 자료에 대한 차원축소기법으로서의 주성분분석(Jolliffe, 1986)은 자료에 포함되어 있는 정보의 손실을 최소화 하면서 변수의 수보다 상대적으로 적은 수의 선형결합을 이용하여 자료를 설명하는 기법이다. 주어진 자료에 대하여 서로 직교하면서 분산을 최대로 하는 설명변수들의 선형결합을 탐색하고(Hotelling, 1933) 적은 수의 선형결합이 전체 자료의 변동 중 많은 부분을 설명하는 경우 효과적으로 자료의 차원을 축소할 수 있다. 이러한 주성분분석은 자료축소, 형상인식, 탐색적 자료분석과 시계열분석 등에 많이 적용되고 있다.

주성분분석은 주어진 공분산행렬이나 상관행렬의 고유값과 고유벡터를 구하는 고유값문제(eigenvalue problem)를 이용하여 표현할 수 있다. 또한 인자분석에 관한 초기의 문헌들(Lawley: 1953, Anderson과 Rubin: 1956) 등에서 잠재변수모형과 인자분석, 주성분분석의 연관성에 대한 언급이 있었다. 최근 Tipping과 Bishop(1999)은 인자분석과 연관된 잠재변수모형을 수정하여 주성분분석에 해당하는 모형을 소개하고 이 모형에서 구해진 모수들이 주성분분석에서의 고유값 및 고유벡터와 같아짐을 보였다. 더불어 잠재모형의 모수들에 대한 최우추정법을 구하는 EM 알고리즘(Green, 1989)도 제시하였다.

1) Associate Professor, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, KOREA.
E-mail : cspark@skku.edu

다변량 자료의 차원축소기법 중 하나인 주성분 분석에서 자료의 변동을 적절히 설명해주는 주성분의 개수, 즉 차원을 선택하는 문제는 매우 중요하다. Anderson(1963)은 고유벡터들의 합에 대한 점근분포를 이용하여 모형에 필요한 차원을 탐색하는 검정법을 소개하였으며 일반적으로는 고유값의 크기를 고려한 방법이나 이를 크기 순으로 표시한 스크리(scree) 그래프 등을 이용하는 대략적인 방법들이 주로 사용되고 있다.

Tipping과 Bishop의 잠재변수모형은 주성분분석에 분포이론을 접목한 것으로 여러 가지 추론 등에 이용할 수 있으나 간단한 모의실험결과 모형에서 차원이 적당하지 않은 경우 주성분에 대한 추정치들이 적절하지 않게 되며 따라서 의미 있는 차원의 결정이 모형의 신뢰도에 중요한 영향을 미치게 됨을 발견하였다. 본 논문에서는 자료에 임의의 결측치가 있는 경우 확률적 주성분분석(Tipping과 Bishop, 1999)에서 제시한 잠재모형을 이용하여 의미 있는 주성분의 수, 즉 차원을 결정하는 점근 검정법을 고려하였다.

본 논문의 구성은 다음과 같다. 먼저 제 2 장에서는 Tipping과 Bishop의 잠재변수모형을 설명하였으며, 제 3 장에서 모형의 차원을 결정하는 검정법을 다루었고 제 4 장에서 결측치가 있는 경우 기존의 EM 알고리즘을 수정하여 최우추정치를 구하는 내용을 포함하였다. 제 5 장에서는 모의자료와 실제자료를 통하여 새로운 기법의 활용가능성을 보였으며 마지막으로 결론을 제 6 장에 포함하였다.

2. 확률적 주성분분석

주성분분석은 잘 알려진 바와 같이 다변량 자료에 대한 차원축소를 목적으로 하는 기법이다. 이 기법은 인자분석(Young, 1940; Whittle, 1952; Anderson, 1963)과 밀접한 관계가 있으며 인자분석은 Lawley(1953)와 Anderson 등(1956)에 의하여 잠재변수모형으로 표현될 수 있음이 알려져 있다. 최근 Tipping과 Bishop(1999)은 기존의 잠재변수 모형을 더 깊게 고찰하여 정규모형의 가정하에서 이 모형의 최우추정치가 주성분분석의 고유값 및 고유벡터들과 같아짐을 보이고 이들 추정치를 구하는 EM 알고리즘을 제시하였다.

표본수가 N이며 p차원의 자료 벡터 $\boldsymbol{x}_n, n \in \{1,\cdots,N\}$이 있고 이에 대한 공분산행렬을 S라 하면 주성분분석은 다음의 고유값 문제와 같아진다.

$$S\boldsymbol{w}_j = \delta_j \boldsymbol{w}_j,\ j = 1,\cdots,q.$$

자료벡터 $\boldsymbol{x}_n$의 q개 주성분은

$$\boldsymbol{c}_n = \boldsymbol{W}^T(\boldsymbol{x}_n - \bar{\boldsymbol{x}}),\quad \text{여기서}\quad \boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_q)$$

과 같으며 $\boldsymbol{w}_j$는 j번째로 큰 고유값과 연관된 직교정규 주축이 된다. 따라서 주성분 $\boldsymbol{c}_n$들은 서로

직교하게 되며 공분산행렬 $\Sigma_n c_n c_n^T/N$은 대각원소가 δ_j인 대각 행렬이 된다.

위의 주성분분석은 앞에서 언급하였듯이 p차원의 자료벡터와 이에 연관된 q차원의 잠재변수 c로 이루어진 다음과 같은 잠재변수모형으로 나타낼 수 있다.

$$x = Wc + \mu + \epsilon, \tag{2.1}$$

여기서 분포가정 $c \sim N(0, I)$와 오차 ϵ에 대한 추가적인 등방 정규분포 $N(0, \sigma^2 I)$를 가정하면 위의 모형 (2.1)과 함께 c가 주어졌을 때의 x의 조건부 분포는

$$x|c \sim N(Wc + \mu, \sigma^2 I)$$

이 된다.

잠재변수를 적분하여 얻어지는 x의 주변분포 또한 다음과 같은 정규분포가 된다.

$$x \sim N(\mu, \Psi),$$

이때 $\Psi = WW^T + \sigma^2 I$이다.

x의 주변분포에 대한 공분산행렬 중 $\sigma^2 I$를 일반적인 대각행렬로 바꾸면 위의 모형은 잘 알려진 대로 인자분석에 대한 잠재변수모형이 된다. 마지막으로 W와 σ^2에 대한 대수가능도함수는

$$l(W, \sigma^2) = -\frac{N}{2}\{p\ln(2\pi) + \ln|\Psi| + \mathrm{tr}(\Psi^{-1}\Sigma)\} \tag{2.2}$$

이 된다. 가능도함수에서

$$\Sigma = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)(x_n - \mu)^T$$

이다.

Tipping과 Bishop은 W와 σ^2에 대한 최우추정치를 구하는 방법으로 EM알고리즘을 제시하였다. 이 방법은 자료에 임의의 결측치가 있는 경우에도 사용할 수 있으며 또한 잠재변수 c에 대한 조건부분포는 베이즈 정리를 이용하면 다음과 같아진다.

$$c|x \sim N(M^{-1}W^T(x - \mu), \sigma^2 M^{-1}),$$

분산항에서 $M = W^T W + \sigma^2 I$이다.

3. 모형의 차원결정

주성분분석에 대한 잠재변수모형에서 차원의 결정은 인자분석에서 모형에 필요한 인자의 수를 결정하는 가설검정과 비슷한 방법을 이용할 수 있다. 우선 모형에 필요한 차원을 q라 놓으면 순차적으로 1부터 $p-1$까지를 귀무가설로 놓고 대립가설은 $q=p$ 또는 공분산의 형태에 아무런 제약이 없도록 하는 가능도비 검정을 고려할 수 있다. 이 때 귀무가설과 대립가설은

$H_0 : \Psi = WW^T + \sigma^2 I$ 단, $q = 1, 2, \cdots, p-1$

$H_1 : \Psi$는 귀무가설과 다른 형태의 일반적인 양정치 행렬식

이 된다.

앞 장의 식 (2.2)에서 주어진 W와 σ^2에 대한 대수가능도함수와 이들에 대한 최우추정치를 이용하여 가능도비 검정에 대한 가능도비 검정 통계량을 표시하면 다음과 같다.

$$-2\ln\Lambda = N\left(\frac{|\widehat{W}\,\widehat{W}^T + \widehat{\sigma^2}I|}{|\widehat{\Psi}|}\right).$$

이 통계량은 N과 $N-p$가 큰 경우 자유도가 $(1/2)\{(p-q)(p-q+1)\}-1$인 χ^2분포에 근사하게 된다. 순차적인 가설검정은 모형에 필요한 차원을 q로 두고 $q=1$에서부터 시작하여 p-값이 정해진 유의수준보다 작으면 q값을 하나씩 증가시켜 최초로 p-값이 유의수준보다 커지는 경우의 q가 모형의 차원이 된다.

이와 같은 순차적 검정들을 수행하기 위해서는 모수들에 대한 최우추정치들이 필요하며 이는 수치적 방법으로 얻을 수 있다. 임의의 결측치가 있는 경우 W와 σ^2에 대한 최우추정치를 구하는 EM 알고리즘에 대한 내용은 다음 절에서 살펴보기로 한다.

4. 결측치가 있는 경우의 최우추정법

자료 $\boldsymbol{x}_n$ 중 임의의 결측이 있는 경우 W와 σ^2에 대한 최우추정치는 수정된 EM 알고리즘을 이용하여 구할 수 있다. 이 때 c_n과 $\boldsymbol{x}_n$ 중 결측이 있는 부분을 EM 알고리즘의 결측부분으로 처리하고 완전한 자료를 $(\boldsymbol{x}_n, c_n)$으로 놓는다. $\boldsymbol{x}_n$ 중 결측이 있는 부분은 EM 알고리즘 중 E-단계의 반복에서 주어진 모수값들에 따라 다중대치법(Schafer, 1997)을 이용하여 대치하는 방법을 사용하

였다.

관측 및 결측된 자료를 모두 포함하는 전체자료에 대한 대수가능도는 다음과 같다.

$$l_{x,c} = \sum_{n=1}^{N} \ln\{f(x_n, c_n)\},$$

이때 n 번 째 관측치에 대한 밀도함수는

$$f(x_n, c_n) = (2\pi\sigma^2)^{-p/2} \exp\left\{-\frac{|x_n - Wc_n - \mu|}{2\sigma^2}\right\}(2\pi)^{-q/2} \exp\left\{\frac{|c_n|^2}{2}\right\}$$

이 된다. 수정된 EM 알고리즘의 E 및 M-단계를 살펴보면 다음과 같다.

4.1 E-단계

E-단계에서는 전 단계에서 주어진 모수들에 대한 추정치를 이용하여 분포 $f(c_n|x_n, W, \sigma^2)$에 대한 $l_{x,c}$의 조건부 기대값을 취한다. 자료 x_n 중 임의의 결측값이 있는 경우에는 기대값을 구하기 전에 전 단계의 모수들을 바탕으로 다중대치법을 이용하여 결측치들을 대치한 후 기대값을 구한다. 결측치들의 대치는 Little과 Rubin(1987)의 EM 방법도 가능하나 Schafer(1997)의 다중대치법을 사용하였다. 다중 대치법의 기본적인 원리는 EM 알고리즘과 비슷하나 결측치를 EM 알고리즘에서 구해진 기대값 등이 아닌 가정된 분포에서 추출된 난수를 이용하여 대치하는 점이 다르다.

자료 x_n 중 임의의 결측치를 대체한 후 구해진 $l_{x,c}$의 조건부 기대값 $\langle l_{x,c}\rangle$는 다음과 같다.

$$\langle l_{x,c}\rangle = -\sum_{n=1}^{N}\left\{\frac{p}{2}\ln\sigma^2 + \frac{1}{2}\mathrm{tr}(\langle c_n c_n^T\rangle) + \frac{1}{2\sigma^2}(x_n - \mu)^T(x_n - \mu) - \frac{1}{\sigma^2}\langle c_n\rangle^T W^T(x_n - \mu)\right.$$
$$\left. + \frac{1}{2\sigma^2}\mathrm{tr}(W^T W\langle c_n c_n^T\rangle)\right\}.$$

식에서 모수들과 관계가 없는 항들은 포함하지 않았으며

$$\langle c_n\rangle = M^{-1}W^T(x_n - \mu),$$
$$\langle c_n c_n^T\rangle = \sigma^2 M^{-1} + \langle c_n\rangle\langle c_n\rangle^T,$$

이고 앞에서와 같이 $M = W^T W + \sigma^2 I$ 이다.

4.2 M-단계

M-단계에서는 E-단계에서 구해진 조건부 기대값 $\langle l_{x,c}\rangle$을 극대화하는 W와 σ^2을 구한다. 전 단계의 모수들을 이용하여 새롭게 구해진 W의 최우추정치는 다음과 같다.

$$\widehat{W} = \left\{\sum_{n=1}^{N}(x_n-\mu)\langle c_n\rangle^T\right\}\left[\sum_{n=1}^{N}\langle c_n c_n^T\rangle\right]^{-1}$$
$$= [SWM^{-T}][\sigma^2 M^{-1}+M^{-1}W^T SWM^{-T}]^{-1}.$$

추정치의 식에 포함되어 있는 σ^2과 W는 전 단계에서 구해진 추정치를 의미한다. 다음으로 σ^2에 대한 최우추정치는 다음과 같다.

$$\widehat{\sigma^2} = \frac{1}{Np}\sum\left[\,|x_n-\mu|^2 - 2\langle c_n\rangle^T\widehat{W}^T(x_n-\mu)+\mathrm{tr}\left(\langle c_n c_n^T\rangle\widehat{W}^T\widehat{W}\right)\right]$$
$$= \frac{1}{p}\mathrm{tr}(S-SWM^{-1}\widehat{W}^T).$$

위의 추정치에서 W는 이전의 M-단계에서 구해진 추정치이고 $\widehat{W}$는 현 단계에서 새롭게 구해진 W의 추정치이다. 마지막으로 위의 E 및 M-단계를 수렴할 때까지 반복하여 최우추정치들을 구할 수 있다.

전체적으로 결측치가 있는 경우의 EM 알고리즘은 기존의 EM 알고리즘의 E-단계에 자료 x_n 중 임의의 결측치를 대치하기 위한 다중대치법이 포함된 수정된 EM 알고리즘의 형태를 가진다.

5. 모의 및 실제 자료에 대한 적용

이 장에서는 우선 모의실험을 이용하여 결측치가 없는 자료를 생성하고 이 자료에서 임의로 선택된 20%의 관측치들을 결측으로 처리한 후 두 자료, 즉 결측치가 없는 경우와 결측치가 있는 경우에 대하여 적용된 차원 검정법의 결과를 비교하였다.

다음으로는 잘 알려진 IRIS자료(Fisher, 1936)와 Johnson과 Wichern(1992)에서 소개된 남자 육상선수들에 대한 실제자료를 가지고 모의자료 경우와 같이 결측이 있는 경우와 없는 경우에 대한 적용결과를 살펴보았다. 검정을 위한 알고리즘은 S-plus를 이용하여 구현하였다.

5.1 모의자료

모의자료는 변수의 수가 5개인 경우로 공분산 행렬의 고유벡터가

$$\text{고유벡터} = \begin{pmatrix} 0.949 & 0.000 & -0.271 & 0.110 & -0.120 \\ 0.000 & 0.949 & -0.163 & -0.184 & 0.199 \\ 0.316 & 0.000 & 0.813 & -0.331 & 0.359 \\ 0.000 & 0.316 & 0.488 & 0.551 & -0.598 \\ 0.000 & 0.000 & 0.000 & 0.735 & 0.678 \end{pmatrix}$$

이고 고유값=$(2.5, 1.5, 0.4, 0.3, 0.3)^T$인 표준정규분포에서 추출되었으며 관측치의 수는 $N=250$으로 하였다. 모의실험은 100회 반복 추출된 자료의 각 각에 대하여 임의로 20%를 결측으로 처리한 뒤 결측이 있는 경우와 없는 경우에 제시된 순차적 검정법을 적용한 결과를 살펴보았다.

세 번째 고유값부터는 상대적으로 작은 값을 가지므로 필요한 주성분의 수를 2로 생각할 수 있으며 모든 자료에 대하여 일률적으로 모형에 필요한 차원을 1, 2, 3으로 하는 귀무가설들에 대한 순차적 검정을 실시한 후 결과를 비교하였다. 결과는 p-값들에 대한 평균과 표준편차 그리고 같은 자료에서 결측이 있는 경우와 없는 경우에 대한 p-값들의 상관계수 등을 표 5.1에 정리하였으며 각 각의 차원에 대한 p-값들에 대한 상자그림을 그림 5.1에 포함시켰다. 유의수준을 0.05로 하였을 때 100번의 반복에서 p-값이 이 보다 큰 회수도 포함하였다. 필요한 차원이 2라고 가정하였을 때 $q=1$인 경우는 이 회수가 작아야 하며 2, 3차원에 대한 검정에서는 이 값이 커야할 것이다. 결측이 있는 경우 잘못된 결과를 나타내는 경우가 모든 차수에서 더 많아 결측치가 검정의 신뢰도에 영향을 미침을 알 수 있다.

표 5.1: 모의실험 결과

귀무가설	$H_0: q=1$		$H_0: q=2$		$H_0: q=3$	
결측치여부	없음	20% 결측	없음	20% 결측	없음	20% 결측
p-값 평균	0.005	0.006	0.375	0.278	0.694	0.499
p-값 표준편차	0.006	0.009	0.253	0.275	0.234	0.288
p-값 > 0.05	0	1	94	78	100	97
p-값 상관계수		0.765		0.484		0.404

필요한 차원을 1이라 놓은 경우에는 모든 반복에서 결측값의 유무에 관계없이 p-값이 대부분 0에 가깝게 나타나 필요한 차원은 2 이상이라는 가정과 부합됨을 알 수 있다. 귀무가설이 $q=1$일 때 결측이 없는 경우와 있는 경우의 검정에 대한 p-값의 상관계수는 0.765로 가장 크게 나타났으며 그 이외에는 0.5보다 작은 값을 보였다.

모의실험은 위의 경우 이외에도 다른 몇 가지의 관측치의 수, 변수의 수에 대하여 실시하였으나 관측치가 작은 경우에 p-값 등의 변동이 더 심한 점 이외에 다른 결과들은 크게 상이하지 않아 포함시키지 않았다. 여타의 고유값과 고유벡터에 대한 결과에 있어서도 별다른 특이사항은 없는 것으로 나타나 생략하였다.

결과적으로 결측이 있는 경우 귀무가설에서의 차원과 관계없이 p-값의 변동이나 그릇된 판단을

내리는 회수 등에서 결측으로 인한 비용이 어느 정도 있음을 알 수 있다.

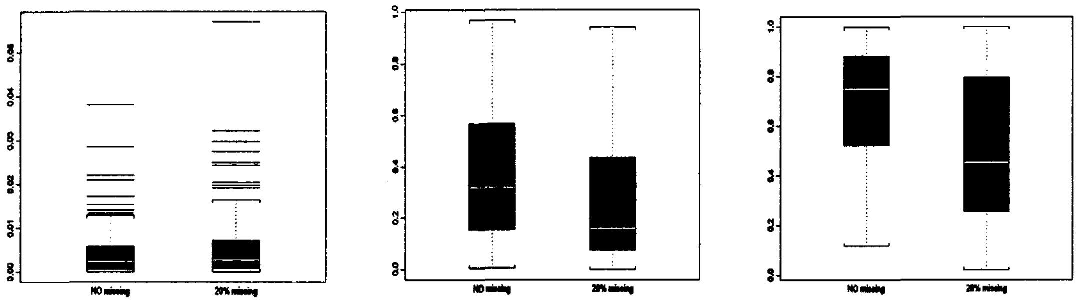

그림 5.1: 왼쪽부터 귀무가설의 차원이 1, 2, 3인 경우

5.2 실제자료 I: IRIS 자료

첫 번째 실제자료는 IRIS 자료로 잘 알려져 있으며 세 종류의 IRIS를 모두 합친 자료이다. 변수는 꽃잎길이, 꽃잎넓이, 꽃받침길이 및 꽃받침넓이 등 4개이다. 원자료에 결측치는 없기 때문에 임의로 20%를 결측치로 처리한 후 처리전과 후의 검정 결과를 비교하였으며 그 결과는 표 5.2와 같다.

표 5.2: IRIS 자료에 대한 검정 결과

귀무가설	$H_0: q=1$		$H_0: q=2$	
결측치여부	없음	20% 결측	없음	20% 결측
검정통계량	18.1	16.1	4.94	3.22
자유도	5	5	2	2
p-값	0.0028	0.0066	0.085	0.2
고유값	2.05	-	0.491	-
누적분산	0.925	-	0.978	-

결측치가 있는 경우와 없는 경우 모두 모형에 필요한 차원, 즉 주성분의 수는 2개가 적당한 것으로 보이며 결측치가 있는 경우의 p-값이 항상 조금 크게 나타났다. 참고로 주성분분석에서 나타난 고유값들은 크기 순으로 {2.05, 0.491, 0.279, 0.154}으로 첫 번째의 고유값이 상대적으로 크다. 설명된 분산은 표에서 보는 것처럼 하나인 경우 92.5%이고 두 개의 성분을 사용하는 경우 97.8%가 되어 검정법에서 2개의 주성분을 택한 결과와 상반되는 결과를 나타냈다.

5.3 실제자료 II: 육상선수 자료

이 자료는 Johnson과 Wichern(1992)에서 인용된 자료이다. 55개 국가의 육상선수들 중 남성에 대한 평균자료로 변수들은 100m, 200m, 400m, 800m, 1500m, 5000m, 10000m, 그리고 마지막으로 마라톤에 대한 기록을 포함하고 있다.

표 5.3 남성 육상선수 자료에 대한 검정 결과

귀무가설	$H_0 : q = 1$	
결측치 여부	없음	20% 결측
검정통계량	28.6	31.2
자유도	27	27
p-값	0.383	0.263
고유값	9.40	-
누적분산	0.98	-

표 5.3의 결과를 보면 결측치가 있는 경우와 없는 경우 모두 주성분이 하나 필요한 것으로 나타났다. 첫 번째 고유값이 9.40인데 반하여 두 번 째 고유값은 1.18로 상대적으로 매우 작은 값을 가지고 있으며 분산도 1.5% 증가하는데 그치는 것으로 나타났다. 결론적으로 필요한 주성분은 하나라고 결론을 내릴 수 있겠다.

6. 결론

본 논문에서는 임의의 결측치가 있는 경우 주성분분석에서 필요한 차원을 탐색하는 점근 검정법에 대하여 살펴보았다. 검정법은 주성분분석에 대한 잠재변수모형에서의 순차적 가능도비 검정방법을 사용하였으며 모형에 필요한 주성분의 수를 하나씩 증가시켜가면서 최초로 p-값이 지정한 유의수준보다 커질 때의 귀무가설의 차원이 모형에서 필요로 하는 주성분의 수가 된다.

간단한 모의실험과 실제자료에 대한 적용 결과 제시된 방법이 효과적으로 주성분분석에서 필요한 주성분의 수를 찾아내는 것으로 판단된다. 제시된 검정법에서 정확한 유의수준과 좀 더 상세한 특성들을 살펴보기 위해서 더 많은 모의실험이 필요하다고 하겠다. 더불어 제시된 검정법이 실질적으로 분석에서 필요한 차원보다 더 큰 차원을 제시하는 경향이 있는 것으로 보이나 이를 확인하기 위해서는 기존의 가능도비 검정법과 가정된 잠재변수모형에 대한 이론적인 분석과 다양한 경우에 대한 모의실험이 필요하다고 하겠다.

제시된 방법은 특히 주성분분석을 위한 자료에 임의의 결측치가 있는 경우 필요한 주성분의 수와 이들에 대한 추론에 효과적으로 사용될 것으로 사료된다.

참고문헌

[1] Anderson, T. W. (1963). Asymptotic Theory for Principal Component Analysis, *Annals of Mathematical Statistics*, Vol. 34, 122-148.

[2] Anderson, T. W., and Rubin, H. (1956). Statistical Inference in Factor Analysis, In J. Newman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume V, U. Cal, Berkeley, 111-150.

[3] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics,* Vol. 7, 179-188.

[4] Green, P. J. (1989). On Use of the EM Algorithm for Penalized Likelihood Estimation, *Journal of the Royal Statistical Society*, Series B, Vol. 52, 443-452.

[5] Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components, *Journal of Educational Psychology*, Vol. 24, 417-441.

[6] Johnson, R. A., and Wichern, R. A. (1992). *Applied Multivariate Statistical Analysis,* New York: Prentice Hall.

[7] Jolliffe, I. T. (1986). *Principal Component Analysis*, New York: Springer-Verlag.

[8] Lawley, D. N. (1953). A Modified Method of Estimation in Factor Analysis and Some Large Sample Results, In *Uppsala Symposium on Psychological Factor Analysis,* Number 3 in Nordisk Psykologi Monograph Series, 35-42. Uppsala: Almqvist and Wiksell.

[9] Little, R., and Rubin, D. (1987). *Statistical Analysis with Missing Data,* Chichester: John Wiley.

[10] Schafer, L. (1997). *Analysis of Incomplete Multivariate Data by Simulation*, New York: Chapman & Hall.

[11] Tipping, M. E., and Bishop, C. M. (1999). Probabilistic Principal Component Analysis, *Journal of the Royal Statistical Society, Series B*, Vol. 61, 611-622.

[12] Whittle, P. (1952). On Principal Components and Least Square Methods of Factor Analysis, *Skamdinavisk Aktuarietidskrift*, Vol. 36, 223-239.

[13] Young, G. (1940). Maximum Likelihood Estimation and Factor Analysis, *Psychometrika,* Vol. 6(1), 49-53.

[2003년 12월 접수, 2004년 2월 채택]