응용통계연구 제 13권 2호, 2000년, pp. 383-392

주성분분석에 의한 결손 자료의 영향값 검출에 대한 연구

김현정[1] 문승호[2] 신재경[3]

요 약

1970년대 후반부터 영향력이 있는 관측값을 검출하기 위해서 회귀분석을 포함한 다양한 다변량 해석법에서의 영향분석 및 감도분석에 대한 연구가 진행되어 왔다. 결손값이 포함된 불완전한 자료에 관해서도 이러한 연구가 필요하다. 이와 관련하여 Kim et al.(1998) 등은 평균벡터와 분산공분산행렬에 대한 최우추정값에 초점을 두고 불완전한 자료에 대한 다변량 해석법에서의 감도분석에 관한 방법적 연구를 다루었다. Kim et al.(1998)에서는 Cook's의 D 통계량을 이용하였으나, 본 논문에서는 결손값이 있는 다변량 자료에 대해서 주성분을 이용하여 영향력이 있는 관측값을 검출하는 방법에 대해서 살펴보았다. 이 때, 결손값은 EM 알고리즘에 의해 대치하여 PCA 통계량을 유도하였다.

주요용어: 주성분분석, 불완전자료, EM 알고리즘.

1. 서론

다변량 자료를 해석할 때 관측값이 p변량 정규분포를 따른다고 가정할 수 있다. 이런 경우 모든 자료가 관측되었을 때에는 일반적인 다변량 해석방법을 사용할 수가 있으나, 일부분의 자료가 결손되어 있는 경우를 자주 볼 수 있다. 이러한 불완전한 자료가 있는 경우에는 i) 관측된 자료만을 사용하는 방법, ii) 모든 관측값의 쌍을 이용하여 계산한 공분산을 이용하는 방법, iii) 결손값을 구하는 방법이 있다. Little & Rubin(1987)은 결손값이 있는 경우 모든 관측값에 근거하여 EM 알고리즘을 이용하면 위에서 제시한 방법 보다 더 좋은 평균, 분산, 공분산에 대한 최우추정값을 구할 수 있음을 보였다. Tanaka(1994)는 영향력이 있는 관측값을 검출하기 위해서 다양한 다변량 방법들에 대한 영향분석 및 감도분석에 관해서 소개하였는데, 불완전한 자료에 관해서도 이러한 연구가 필요하다. Kim et al.(1998) 등은 평균벡터와 분산공분산행렬에 대한 최우추정값에 초점을 두고 불완전한 자료의 다변량 해석법에서의 감도분석에 관한 방법적 연구를 다루었다. Kim et al.(1998)에서는 Cook's의 D 통계량을 이용하였으나, 본 연구에서는 주성분을 이용하여 영향력이 큰 관측값을 검출하는 방법에 대해서 살펴보았다.

주성분 분석에서는 대부분의 다른 다변량 해석법과 같이 두 단계로 나누어 분석할 수 있다. 첫 번째 단계에서는 평균벡터와 분산공분산행렬을 추정한다. 두 번째 단계에서는 $\tilde{\mu}$와

1) (617-736) 부산시 사상구 괘법동 산1-1, 신라대학교 교양학과, 전임강사
E-mail: semikim@silla.ac.kr
2) (608-738) 부산시 남구 우암동 산55-1, 부산외국어대학 통계학과, 조교수
E-mail: shmoon@taejo.pufs.ac.kr
3) (641-773) 경남 창원시 사림동 9번지, 창원대학교 통계학과, 부교수
E-mail: jkshin@sarim.changwon.ac.kr

$\tilde{\Sigma}$을 이용하여 고유값과 고유벡터를 구하는데, 이들은 다변량 해석법에서 다양한 통계량들의 영향력을 유도하기 위해 평균벡터와 분산공분산행렬의 영향을 평가하는데 중요하다. 2장에서는 불완전한 자료로부터 $\tilde{\mu}$와 $\tilde{\Sigma}$의 최우추정에 대한 EM 알고리즘을 설명한다. 3장에서는 $\tilde{\mu}$와 $\tilde{\Sigma}$에 대한 영향함수(Influence Function, IF)를 유도한다. 4장에서는 체인 룰을 이용하여 주성분에 대한 IF를 유도하고 5장에서는 수치적 예 및 토의를 다룬다.

2. EM 알고리즘을 이용한 불완전 자료의 평균 μ와 분산 Σ의 최우추정값

관측값 $X_1, X_2, \cdots, X_p$은 p변량 정규분포의 랜덤샘플로 얻어지며, 데이터의 일부분은 랜덤하게 결손되어 있다고 가정하자. 즉, 결손확률은 변량의 결손값과는 독립적임을 말한다. 결손 데이터가 존재할 때 최우추정값 계산에 관하여 Dempster et al.(1977) 등이 일반적인 접근을 시도하였는데, 기술적인 면은 추정(estimation)과 최대화(maximization)의 두 단계를 반복하는 계산으로서, 이를 EM 알고리즘이라고 한다. 다변량 정규분포에 있어서 μ와 Σ의 최우추정값은 완전데이터의 경우에 충분통계량은 $T_1 = \Sigma X_\alpha$, $T_2 = \Sigma X_\alpha X_\alpha^T$로서, 알고리즘의 단계는 다음과 같다.

단계 2.1 초기값은 $\tilde{\mu} = \tilde{T}_1/n$, $\tilde{\Sigma} = T_2/n - \tilde{\mu}\tilde{\mu}^T$로 설정.

단계 2.2 (추정단계: Estimation step)

E 단계는 결손값을 추정하기 위하여 $\tilde{\mu}$와 $\tilde{\Sigma}$이 주어진 조건부 기대값을 계산한 다음, T_1과 T_2로부터 $\tilde{x}_\alpha^{(1)}$을 추정한다;

$$\tilde{x}_\alpha^{(1)} = \tilde{\mu}^{(1)} + \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}(x_\alpha^{(2)} - \tilde{\mu}_\alpha^{(2)}),$$

$$\widetilde{x_\alpha^{(1)}(x_\alpha^{(1)})^T} = \tilde{\Sigma}_{11} - \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}\tilde{\Sigma}_{21} + \tilde{x}_\alpha^{(1)}(\tilde{x}_\alpha^{(1)})^T,$$

$$\widetilde{x_\alpha^{(1)}(x_\alpha^{(2)})^T} = \tilde{x}_\alpha^{(1)}(x_\alpha^{(2)})^T, \quad \widetilde{x_\alpha^{(2)}(x_\alpha^{(2)})^T} = x_\alpha^{(2)}(x_\alpha^{(2)})^T, \quad \alpha = 1, 2, \cdots, n,$$

이때, 첨자 (1)과 (2)는 α번째 관측값에 대해 각각 결손값과 관측값의 변수 그룹을 나타낸다.

단계 2.3 단계 2.2의 결과를 이용하여 충분통계량을 계산한다;

$$\tilde{T}_1 = \sum_\alpha X_\alpha^+, \quad \tilde{T}_2 = \sum_\alpha (X_\alpha X_\alpha^T)^+,$$

여기서,

$$X_\alpha^+ = \begin{cases} x_\alpha, & \text{observed} \\ (\tilde{x}_\alpha^{(1)}, x_\alpha^{(2)})^T, & \text{missing} \end{cases}$$

$$(X_\alpha X_\alpha^T)^+ = \begin{cases} x_\alpha x_\alpha^T, & \text{observed} \\ \begin{bmatrix} \widetilde{x_\alpha^{(1)}(x_\alpha^{(1)})^T} & \tilde{x}_\alpha^{(1)}(x_\alpha^{(2)})^T \\ x_\alpha^{(2)}(\tilde{x}_\alpha^{(1)})^T & x_\alpha^{(2)}(x_\alpha^{(2)})^T \end{bmatrix}, & \text{missing.} \end{cases}$$

단계 2.4 (최대화 단계, Maximization step)
M 단계는 대치된 충분통계량으로 μ와 Σ의 재추정값을 계산한다;

$$\tilde{\mu} = \tilde{T}_1/n, \quad \tilde{\Sigma} = \tilde{T}_2/n - \tilde{\mu}\tilde{\mu}^T.$$

단계 2.5 수렴조건을 만족하면 단계 2.4의 값을 평균, 공분산의 추정값으로 하고, 그렇지 않으면 단계 2.2로 돌아간다.

3. μ와 Σ의 최우추정값에 관한 영향함수

관측점 $(\boldsymbol{x}, y)$의 통계량 T에 대한 영향을 평가하기 위해서 Hampel(1974)는 다음과 같은 IF를 도입하였다.

$$IF(\boldsymbol{x}, y; T, F) = \lim_{\epsilon \to 0}\{T[(1-\epsilon)F + \epsilon\delta(\boldsymbol{x}, y)] - T[F]\}/\epsilon \tag{3.1}$$

여기서, F는 확률변수 $(\boldsymbol{X}, Y)$의 분포함수. $\delta(\boldsymbol{x}, y)$는 어떤 특정한 $(\boldsymbol{x}, y)$를 확률 1로 취하는 분포함수이다. (3.1)식은 이론분포, 모집단 분포에 기초한 IF인데, 실제의 데이터에 적용하기 위해서 표본에 대한 영향함수 (경험영향함수(empirical influence function, EIF))가 필요하다.

지금 n개의 관측값 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ 이 얻어졌다고 하자. IF의 정의에서 분포함수 F를 $\hat{F}$로 바꾼 것에서 i번째 관측값의 EIF는

$$EIF(\boldsymbol{x}_i; \hat{T}) \equiv \lim_{\epsilon \to 0}\{T[(1-\epsilon)\hat{F} + \epsilon\delta(\boldsymbol{x})] - T[\hat{F}]\}/\epsilon \tag{3.2}$$

이다. i번째 관측값을 제거했을 때의 모수를 $\hat{T}_{(i)}$로 한다. IF를 이용하면

$$\tilde{T}_{(i)} \cong \hat{T} - (n-1)^{-1} EIF(\boldsymbol{x}_i; \hat{T}) \tag{3.3}$$

가 된다. 이 값에 의해서 i번째 관측값의 영향을 평가한다.

개체 α에 대해 $w_\alpha^* = nw_\alpha/\Sigma w_\alpha, \quad \alpha = 1, \cdots, n$과 같은 가중값을 생각하자. 그리고 관측값 x_α는 $N(\mu, w_\alpha^{*-1}\Sigma)$분포를 따르고 서로 독립이라고 가정하자(see, Kwan and Fung(1998)). 가중값이 없는 경우 모든 α에 대해 $w_\alpha = 1$ 또는 $w_\alpha^* = 1$은 2장의 모형과 같으며 이에 대해 가중값이 있는 경우에는 미세한 변동(perturbation)을 도입하여 그에 따른 변화를 살펴보자. 완전데이터의 경우 μ와 Σ의 ML추정값은 다음과 같다.

$$\hat{\mu} = \sum_\alpha w_\alpha x_\alpha / \sum_\alpha w_\alpha, \quad \hat{\Sigma} = (\sum_\alpha w_\alpha)^{-1} \sum_\alpha w_\alpha (x_\alpha - \hat{\mu})(x_\alpha - \hat{\mu})^T \tag{3.4}$$

$w_0 = (1, \cdots, 1)^T$에서 w_α에 대한 편미분은

$$\partial\hat{\mu}/\partial w_\alpha|_{w_0} = n^{-1}(x_\alpha - \hat{\mu}), \quad \partial\hat{\Sigma}/\partial w_\alpha|_{w_0} = n^{-1}\{(x_\alpha - \hat{\mu})(x_\alpha - \hat{\mu})^T - \hat{\Sigma}\}, \tag{3.5}$$

이며(see, Kim et al.(1998)), 이들 편미분은 평균벡터와 분산공분산행렬의 경험영향함수(Empirical Influence Function, EIF)와 같거나 대응하는 EIF(식(3.2))의 $1/n$배에 해당한다. $\hat{\mu}$와 $\hat{\Sigma}$에 대해 미분가능함수로 유도할 수 있는 다른 통계량들도 이와 유사하게 전개할 수 있다. 따라서, 편미분을 n배 함으로써 IF를 얻을 수 있다.

다음으로 불완전 데이터의 $\hat{\mu}$와 $\hat{\Sigma}$의 편미분을 유도해 보자.

가중모형 (3.5)에서 $\tilde{T}_1 = \sum_\alpha w_\alpha^* X_\alpha$와 $\tilde{T}_2 = \sum_\alpha w_\alpha^* X_\alpha X_\alpha^T$는 결합충분통계량으로 앞 장에서 이에 대응되는 값을 충분통계량에 대입하고, 2장의 단계 2.3에서 $\tilde{T}_1 = \sum_\alpha w_\alpha^* X_\alpha^+$와 $\tilde{T}_2 = \sum_\alpha w_\alpha^* \widetilde{(X_\alpha X_\alpha^T)^+}$를 각각 $\tilde{T}_1$과 $\tilde{T}_2$에 대입하는 과정을 이용하여 추정값 $\hat{\mu}_w$와 $\hat{\Sigma}_w$를 얻을 수 있다. $\hat{\mu}$와 $\hat{\Sigma}$에 $\tilde{\mu} + (\partial\tilde{\mu}/\partial w_j)\Delta w_j$와 $\tilde{\Sigma} + (\partial\tilde{\Sigma}/\partial w_j)\Delta w_j$를 각각 대입하고 아래의 식들을 정리하면 Δw_j의 계수를 얻을 수 있다. 수렴해 $\tilde{\mu}$와 $\tilde{\Sigma}$에서의 편미분 $\partial\tilde{\mu}_j$ $(= \partial\tilde{\mu}/\partial w_j)$와 $\partial\tilde{\Sigma}_j$ $(= \partial\tilde{\Sigma}/\partial w_j)$는 아래의 식에 의해 얻을 수 있다:

$$\begin{aligned}
n(\partial\tilde{\mu}_j) &- \sum_\alpha \{M_\alpha(\partial\tilde{\mu}_j) + \sum_\alpha M_\alpha(\partial\tilde{\Sigma}_j)\bar{M}_\alpha S_{22}^-(\alpha)\bar{M}_\alpha(X_\alpha^+ - \tilde{\mu}) \\
&- S_{12}(\alpha)S_{22}^-(\alpha)\bar{M}_\alpha(\partial\tilde{\Sigma}_j)\bar{M}_\alpha S_{22}^-(\alpha)\bar{M}_\alpha(X_\alpha^+ - \tilde{\mu}) - S_{12}(\alpha)S_{22}^-(\alpha)\bar{M}_\alpha(\partial\tilde{\mu}_j)\} \qquad (3.6) \\
&= X_j^+ - \tilde{\mu}, \\
(\partial\tilde{\Sigma}_j) &- n^{-1}\sum_\alpha \{M_\alpha(\partial\tilde{\Sigma}_j)M_\alpha - M_\alpha(\partial\tilde{\Sigma}_j)\bar{M}_\alpha S_{22}^-(\alpha)S_{21}(\alpha) \\
&- S_{12}(\alpha)S_{22}^-(\alpha)S_{21}(\alpha) - S_{12}(\alpha)S_{22}^-(\alpha)\bar{M}_\alpha(\partial\tilde{\Sigma}_j)\bar{M}_\alpha S_{22}^-(\alpha)S_{21}(\alpha)\} \\
&- n^{-1}\sum_\alpha AX_\alpha^{+T}M_\alpha - n^{-1}\sum_\alpha M_\alpha X_\alpha^+ A^T - n^{-1}\sum_\alpha \{AX_\alpha^{+T}\bar{M}_\alpha + \bar{M}_\alpha X_\alpha^- A^T\} \qquad (3.7) \\
&+ (\partial\tilde{\mu}_j)\tilde{\mu}^T + \tilde{\mu}^T(\partial\tilde{\mu}_j)^T \\
&= n^{-1}\{(X_j X_j^T)^+ - n^{-1}T_2\},
\end{aligned}$$

여기서,

$$\begin{aligned}
A = M_\alpha(\partial\tilde{\mu}_j) &+ \sum_\alpha M_\alpha(\partial\tilde{\Sigma}_j)\bar{M}_\alpha S_{22}(\alpha)\bar{M}_\alpha(X_\alpha^+ - \tilde{\mu}) \\
&- S_{12}(\alpha)S_{22}^-(\alpha)\bar{M}_\alpha(\partial\tilde{\Sigma}_j)\bar{M}_\alpha S_{22}^-(\alpha)\bar{M}_\alpha(X_\alpha^+ - \tilde{\mu}) \\
&- S_{12}(\alpha)S_{22}^-(\alpha)\bar{M}_\alpha(\partial\tilde{\mu}_j).
\end{aligned}$$

M_α와 $\bar{M}_\alpha$는 각각 $M_\alpha = \mathrm{diag}(\delta_\alpha(1), \cdots, \delta_\alpha(p))$와 $\bar{M}_\alpha = I - M_\alpha$로 정의되며 여기서,

$$\delta_\alpha(i) = \begin{cases} 1, & \text{if variate } i \text{ is missing for individual } \alpha \\ 0, & \text{otherwise} \end{cases}$$

이고, M_α와 $\bar{M}_\alpha$의 정의에 의해

$$S_{11}(\alpha) = M_\alpha S M_\alpha, \quad S_{12}(\alpha) = M_\alpha S \bar{M}_\alpha, \quad S_{22}(\alpha) = \bar{M}_\alpha S \bar{M}_\alpha$$

가 된다. 개체 α에 대해 결손이 없는 변수와 0으로 채워진 다른 부분에 대응하는 S의 한 부분을 $S_{22}(\alpha)$가 이루고 있으므로, 개체 α에 대해 모든 변수가 관측된 경우를 제외하고는 $S_{22}(\alpha)$은 정칙이다. $S_{22}^-(\alpha)$는 비정칙일 때 일반적인 역행렬과 같은 Moore-Penrose 역행렬로서 결손이 없는 부분만으로 만들어진 행렬의 역에 의해 구할 수 있고, 역행렬의 요소를 S의 결손이 없는 모든 부분의 요소로 바꾸어 준다. 위의 식에서 미지수는 $\partial\tilde{\mu}_j$에 대해서는 p개이며, $\partial\tilde{\Sigma}_j$에 대해서는 $p^* = p(p+1)/2$개이다. 이것은 곧 선형독립방정식의 수와 같다. 주의할 것은 (3.6)과 (3.7)식에서 등호의 왼쪽의 미지 계수는 개체수 j와 독립이지만 등호의 오른쪽의 j와는 종속이다. 따라서, 등호의 오른쪽은 각 j에 대해서 계산해야 되지만, 왼쪽은 단지 한번만 계산하면 된다.

4. PCA 통계량에 대한 영향함수

PCA(주성분분석), CCA(정준상관분석), FA(인자분석), CSA(공분산 구조분석) 그리고 DA(판별분석)를 포함하는 다변량 통계해석은 두 단계로 적용할 수 있다. 1장에서도 설명한 바와 같이 첫 번째 단계는 평균벡터($\tilde{\mu}$)와 분산공분산행렬($\tilde{\Sigma}$)을 추정하고, 두 번째 단계에서는 주요 통계량들을 $\tilde{\mu}$와 $\tilde{\Sigma}$의 미분가능한 형태로 얻을 수 있다. 3장에서 방정식의 풀이에 의해 불완전데이터의 $\tilde{\mu}$와 $\tilde{\Sigma}$에 대한 IF를 구할 수 있는 것을 보였다. 이들 $\tilde{\mu}$와 $\tilde{\Sigma}$에 대한 IF가 얻어지면 체인 룰을 사용하여 통계량에 대한 IF를 쉽게 구할 수 있다.

PCA의 경우를 살펴보자. PCA는 유력한 고유값 ν_s와 이에 대응하는 고유벡터 $\boldsymbol{v}_s$, q개의 주성분에 의해 확장된 부공간으로 직사영된 $P = \sum_{s=1}^{q} \boldsymbol{v}_s \boldsymbol{v}_s^T$, 그리고 크기가 큰 q개의 고유값에 대응하는 스펙트럴분해의 부분 $T = \sum_{s=1}^{q} \nu_s \boldsymbol{v}_s \boldsymbol{v}_s^T$로 구성되어 있다. $w_0 = (1, \cdots, 1)^T$에서 w_α에 대한 $\nu_s, \boldsymbol{v}_s$, P 그리고 T의 경험영향함수 EIF는 $\tilde{\Sigma}$(Tanaka, 1988)의 EIF에서 $\partial\tilde{\Sigma}(\alpha)$를 대입하면 각각 다음과 같이 유도할 수 있다.

$$\begin{aligned}
\partial\nu_s(\alpha) &= \boldsymbol{v}_s^T(\partial\tilde{\Sigma}(\alpha))\boldsymbol{v}_s, \\
\partial\boldsymbol{v}_s(\alpha) &= \sum_{r\neq s}^{s} (\nu_s - \nu_r)^{-1}(\boldsymbol{v}_r^T(\partial\tilde{\Sigma}(\alpha))\boldsymbol{v}_s)\boldsymbol{v}_r, \\
\partial P(\alpha) &= \sum_{s=1}^{q}\sum_{r=q+1}^{p} (\nu_s - \nu_r)^{-1}(\boldsymbol{v}_s^T(\partial\tilde{\Sigma}(\alpha))\boldsymbol{v}_r)(\boldsymbol{v}_s\boldsymbol{v}_r^T + \boldsymbol{v}_r\boldsymbol{v}_s^T), \qquad (4.1) \\
\partial T(\alpha) &= \sum_{s=1}^{q}\sum_{r=1}^{q}(\boldsymbol{v}_s^T(\partial\tilde{\Sigma}(\alpha))\boldsymbol{v}_r)\boldsymbol{v}_s\boldsymbol{v}_r^T \\
&+ \sum_{s=1}^{q}\sum_{r=q+1}^{p} \nu_s(\nu_s - \nu_r)^{-1}(\boldsymbol{v}_s^T(\partial\tilde{\Sigma}(\alpha))\boldsymbol{v}_r)(\boldsymbol{v}_s\boldsymbol{v}_r^T + \boldsymbol{v}_r\boldsymbol{v}_s^T).
\end{aligned}$$

따라서, 식(4.1)은 불완전데이터에 대한 PCA 통계량의 IF로 평가할 수 있다.

표 5.1: Sweat Data

Individual	X_1 Sweat rate	X_2 Sodium	X_3 Potassium
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
4	3.2	53.2	12.0
5	3.1*	55.5	9.7
6	4.6	36.1	7.9
7	2.4*	24.8*	14
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5*	58.8*	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

Source : Johnson and Wichern(1992), p.183

5. 수치 예 및 토의

수치적인 예로 여성 건강조사의 세 변수(X_1 : sweat rate, X_2 : sodium content, X_3 : potassium content)에 대한 20명의 자료($p = 3, n = 20$)에 적용시켜 보았다. 이 자료는 표 5.1과 같으며 결손값에 관한 연구를 위해 세 변수의 상호관계를 고려하여 두 변수에 5개의 결손값을 만들었는데, *가 붙어있는 값이 결손값을 의미한다.

우선 μ와 Σ을 추정하기 위해 EM 알고리즘을 이용한다. 수렴조건은 유클리디안 노름의 반복적인 차가 두 값 모두 $\epsilon = 0.00001$보다 작을 때, 그 값을 $\tilde{\mu}$와 $\tilde{\Sigma}$로 한다. 얻어진 추정값은

$$\tilde{\mu} = (4.692802, 45.384353, 9.965), \quad \tilde{\Sigma} = \begin{pmatrix} 1.421692 & 3.883049 & -1.839733 \\ 3.883049 & 178.486803 & -4.047695 \\ -1.839733 & -4.047695 & 3.446275 \end{pmatrix}$$

이다.

다음으로 추정값 평균벡터 $\tilde{\mu}$와 공분산행렬 $\tilde{\Sigma}$의 영향을 조사하기 위해 $\tilde{\mu}$와 $\tilde{\Sigma}$에 대한 EIF를 계산했다. 일반적으로 다변량 해석법에서는 $\tilde{\Sigma}$을 이용하므로 본 연구에서도 $\tilde{\Sigma}$에 관해 초점을 두어 PCA의 영향을 식(4.1)의 $\partial\nu_s(\alpha) = \boldsymbol{v}_s^T(\partial\tilde{\Sigma}(\alpha))\boldsymbol{v}_s$로 계산하였다. 여기서, 가장 큰 고유값에 대한 IF의 index plot은 그림 5.1과 같이 나타난다. 이 때, 고유값의 크기는 $170.2387 \gg 4.6962 > 1.0896$과 같다. 그림 5.1에서는 15번째와 17번째 관측값의 영향력이 나머지 18개 관측개체들보다 상대적으로 크다는 사실을 알 수 있다.

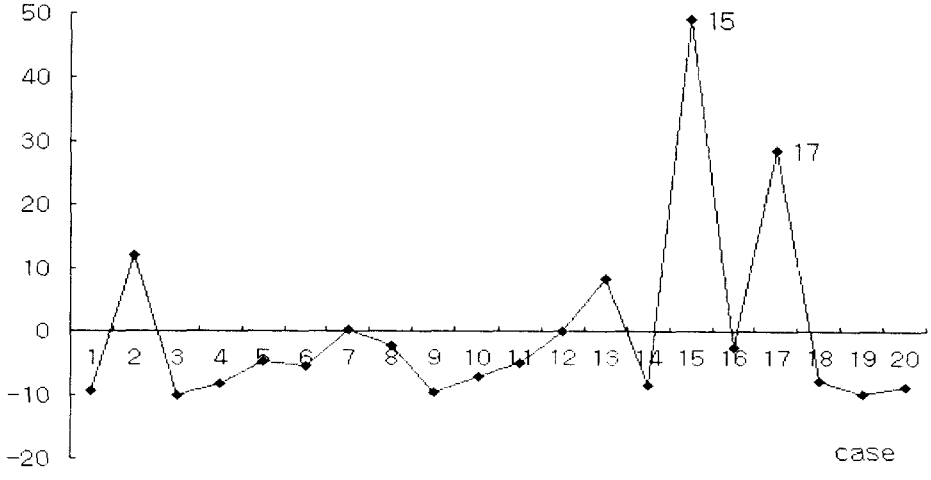


그림 5.1 PCA에서 가장 큰 고유값에 대한 IF의 index plot

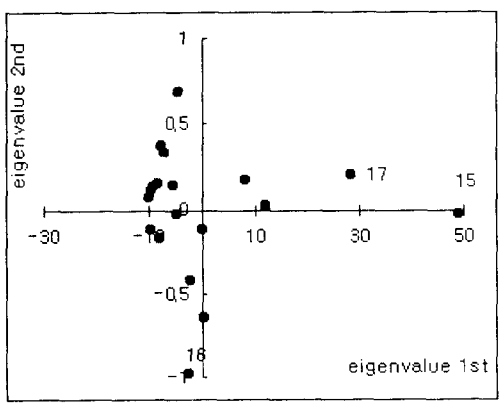
따라서 이들 두 관측값의 영향력에 관해 보다 상세히 살펴보기 위해서, PCA에서 첫번째와 두번째로 큰 두 개의 고유값에 대한 IF의 산점도를 그림 5.2에 나타내었다.

그림 5.2의 (a)는 20개의 모든 관측값에 대한 것으로 그림 5.1과 같이 15, 17번째가 중심으로부터 많이 떨어져 있는 것을 알 수 있다. 15번째를 제외한 (b)에서는 그림 5.1과 5.2의 (a)에서 드러나지 않았던 7, 18번째가 오히려 상대적으로 중심에서 많이 떨어져 있는 것을 알 수 있는데, 이것은 15번째 개체를 제외하더라도 다른 관측값의 영향력이 상대적으로 커서 불안정해짐을 보여 주고있다. (c)의 경우에는 17번째를 제외한 것으로 15, 16번째가 중심에서 많이 벗어나 있으며, 15번째의 영향력이 여전히 크다는 사실을 보여준다. (d)는 그림 5.1과 5.2에서 영향력이 아주 큰 15, 17번째를 동시에 제외한 것으로 중심에서 벗어난 관측값이 거의 보이지 않는다. 따라서, 15, 17번째의 관측값을 제외한 $\tilde{\mu}$와 $\tilde{\Sigma}$의 추정값을 살펴보면 다음과 같다.

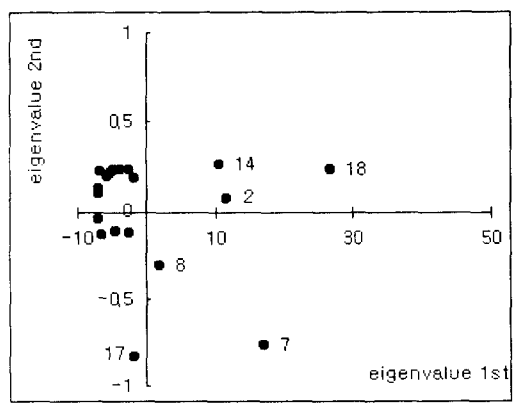
$$\tilde{\mu} = (4.880889, 44.820661, 10.055556), \quad \tilde{\Sigma} = \begin{pmatrix} 2.355619 & 6.576853 & -2.056137 \\ 6.576853 & 106.320618 & -5.184738 \\ -2.056137 & -5.184738 & 3.646914 \end{pmatrix}$$

위의 값 특히 $\tilde{\Sigma}$의 값 중에서 $\tilde{\sigma}_{22}$가 모든 자료를 이용한 경우의 값보다 상대적으로 많이 감소되어 있어서 전체적으로 안정되어 있는 것을 알 수 있었다.

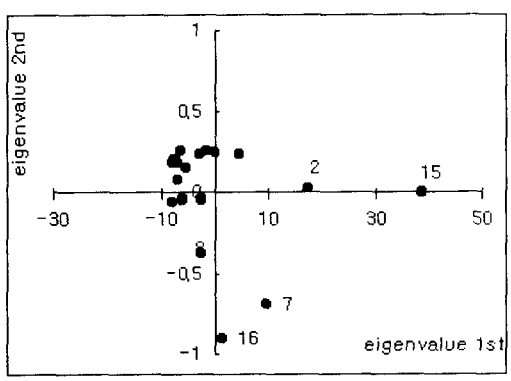
본 논문에서는 결손값을 포함하는 자료에 대해 EM 알고리즘을 이용하여 얻어진 평균벡터와 분산공분산행렬의 최우추정값에 대한 EIF를 유도하였다. 3장에서 논의한 EIF는 α번째에 가중값 $nw_\alpha/\Sigma w_\beta$을 주었을 때 가중값 w_α에 대한 편미분으로 나타난다. 따라서, 결손값이 포함된 다변량 자료에 대해 주성분 분석으로 편미분에 의해 영향력이 큰 개체를 검출하는 방법을 제시하였다.



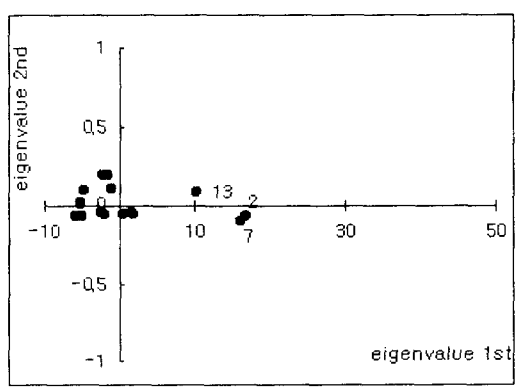
(a) 모든 관측값(full data set)



(b) 15번째를 제외한 경우



(c) 17번째를 제외한 경우



(d) 15, 17번째를 제외한 경우

그림 5.2 PCA에서 첫 번째 · 두 번째 큰 고유값에 대한 IF 의 산점도

감사의 글

본 논문에 대해 많은 조언과 배려를 해 주신 심사 위원장님과 위원님들께 감사를 드립니다.

참고문헌

[1] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society*, B39, 1-38.

[2] Hampel, F.R. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, 69, 383-93.

[3] Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*, Third Edition, Prentice Hall.

[4] Kim, H.J., Tarumi, T. and Tanaka, Y. (1998). Assessing Local Influence in Multivariate Analyses of Incomplete Data, *Journal of the Faculty of Environmental Science and Technology, Okayama University*, Vol. 3. No. 1, 37-46.

[5] Kwan, C.W. and Fung, W.K. (1998). Assessing local influence for specific restricted likelihood : Application to factor analysis, *Psychometrica*, 63, 35-46.

[6] Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

[7] Tanaka, Y. (1988). Sensitivity analysis in principal component analysis : Influence on the subspace spanned by principal components, *Communications in Statistics*, A17, 3157-75. (Corrections, A 18(1989), 4305).

[8] Tanaka, Y. (1994). Recent Advance in Sensitivity Analysis in Multivariate Statistical Methods, *Journal of the Japanese Society of Computational Statistics*, 7, 1-25.

[1999년 7월 접수, 2000년 5월 채택]

Detecting Influential Observations in Multivariate Statistical Analysis of Incomplete Data by PCA

Hyun-Jeong Kim[1] Sung-Ho Moon[2] Jae-Kyoung Shin[3]

ABSTRACT

Since late 1970, methods of influence or sensitivity analysis for detecting influential observations have been studied not only in regression and related methods but also in various multivariate methods. If results of multivariate analyses sometimes depend heavily on a small number of observations, we should be very careful to draw a conclusion. Similar phenomena may also occur in the case of incomplete data. In this research we try to study such influential observations in multivariate statistical analysis of incomplete data. Case of principal component analysis is studied with a numerical example.

Keywords: Principal Component Analysis; Missing Data; EM Algorithm.

1) Fulltime-lecturer, Department of General Education, Silla University.
E-mail: semikim@silla.ac.kr
2) Assistant Professor, Department of Pusan University of Foreign Studies.
E-mail: shmoon@taejo.pufs.ac.kr
3) Associate Professor, Department of Statistics, Changwon National University.
E-mail: jkshin@sarim.changwon.ac.kr