論　文

同時對角化에 의한 多變數群間 特徵抽出의 一手法

(A Feature Extraction Method by Simultaneous Diagonalization)

吳 永 煥*, 安 居 院 猛**

(Oh, Yung Hwan, and Agui, Ta Keshi)

要　約

本 論文에서는 同時對角化에 의한 一群正規化 및 混合正規化時의 座標系의 變換에 着目하여, 두 多變數群間의 特徵抽出을 行하는 手法을 提案, 그에 隨伴하는 몇몇의 性質과 그 適用例에 對해 記述했다. 또한 本手法의 有效性을 보이기 위해 因子分析結果와 比較해, 有意한 結果를 얻었다.

Abstract

Here a method is shown to extract features from two multi-variable classes by using the coordinate systems transformed by one-class and mixture normalization algorithms.

Some properties and implemented results of this technique are described. Also, comparision of these features with factor analysis results is performed.

This method is thought to be more powerful one, in feature extraction sense, than factor analysis.

1. 序　論

多變量間의 複雜하고 또한 一見 不規則하게 보이는 變動을 少數의 代表的·假說的變動에 의해 說明하려고 하는 因子分析(factor analysis)[1]이나 主成分分析(principal component analysis) 等 多變量解析(multivariate analysis)手法은 分析對象이나 分析目的等에 따라 多種類가 開發, 報告되어 왔다.

그러나 多變數群內에서의 少數의 代表的 因子나 成分의 抽出手法 뿐만 아니라, 多變數群間의 '相對的인 差異'를 少數의 代表的 因子에 의해 抽出해낼 수 있는 手法의 開發도 重要하다고 생각된다. 例를 들면, 두 學生群間의 學力檢査나 適性檢査等에서 볼 수 있는 群間의 特性抽出이나, 패턴間의 差異의 抽出等을 생각할 수 있다. 本 論文에서는 直交變換(orthogonal transformation)에 의한 同時對角化(simultaneous diagonalization)手法을 利用한 두 개의 多變數群間의 特徵抽出手法을 提案, 本手法이 지니는 몇몇의 性質에 對해 記述하고, 實際의 資料를 使用한 適用例와 因子分析 結果와의 比較에 의해 本手法의 有效性을 確認하고자 한다.

*正會員, 大田工業專門學校 (Dept. of Electronic Engineering, Member, Daejon Technical Junior College

**東京工業大學

Tokyo Instiute of Technology, Japan)

接受日字 : 1978年 4月 13日

2. 混合相關行列正規化에 의한 特徵抽出

2.1 Algorithm

두 개의 多變數群의 各變數의 變動으로부터 相關行列 R_1, R_2를 計算하여, 各各에 weight를 곱해 더한 行列을 混合相關行列(average correlation matrix) A라 定義한다. (以下 混合行列이라 줄인다)

$$A = w_1R_1 + w_2R_2 \tag{1}$$

$$\text{但,}\ w_1+w_2=1,\ 0<w_1,\ w_2<1 \tag{2}$$

$n\times n$行列 A의 固有値(eigenvalue) $\lambda_1, \cdots\cdots, \lambda_n$을 對角成分으로 하는 固有値行列을 Λ, 固有벡터(eigenvector) 行列을 Φ라 하면,

$$\Phi^T A \Phi = \Lambda \tag{3}$$

$$\Lambda^{-1/2}\Phi^T A \Phi \Lambda^{-1/2} = I \tag{4}$$

의 關係가 成立한다. 上式에서 Φ^T, I는 各各 Φ의 transpose, 單位行列(unity matrix)을 가리킨다. R_1, R에 對해서 (4)式의 whitening 變換을 行하면 다음과 같다.

$$\Lambda^{-1/2}\Phi^T R_l \Phi \Lambda^{-1/2} = K_l \quad (l=1, 2) \tag{5}$$

(1)式을 (4)式에 代入하여, (5)式을 利用해 다시 쓰면 다음 式을 얻는다.

$$w_1K_1+w_2K_2=I \tag{6}$$

여기서, K_1의 固有 벡터行列을 P라 하면

$$w_1P^TK_1P+w_2P^TK_2P=P^TIP=I \tag{7}$$

(7)式의 第1項 및 I가 對角行列이므로 第2項도 對角行列이다. 卽, R_1 및 R_2는 다음과 같은 同一한 變換行列에 의해 對角化된다.

$$P^T\Lambda^{-1/2}\Phi^TR_l\Phi\Lambda^{-1/2}P=P^TK_lP=U^{(l)} \quad (l=1,2) \tag{7a}$$

但,

$$U^{(l)}=\begin{bmatrix}\mu_1^{(l)} & & & 0\\ & \ddots & & \\ & & \mu_i^{(l)} & \\ & & & \ddots \\ 0 & & & \mu_n^{(l)}\end{bmatrix}$$

한편, (4)式에 直交變換 P를 行해도, I는 不變이므로, 平均行列 A와 R_1 및 R_2는 同一變換行列에 의해 對角化된다.

以上의 同時對角化에 의한 座標軸의 變換은 다음과 같다.

$$Z=P^T\Lambda^{-1/2}\Phi^TX=TX \tag{8}$$

但, $T=P^T\Lambda^{-1/2}\Phi^T$이며, X는 變換前의 座標벡터, Z는 變換後의 座標벡터이다.

위의 變換에 의해, $U^{(l)}$의 큰 對角成分 $\mu_i^{(l)}$ (Z의 成分 z_i에 對應)에 對한, X의 成分 x_j의 寄與(contribution)로 부터 多變數群 l의 特徵抽出이 可能하게 된다.

2.2 平均行列正規化에 따르는 性質

(7), (7a)式으로 부터

$$w_1U^{(1)}+w_2U^{(2)}=I \tag{9}$$

가 成立한다. (9)式을 고쳐 쓰면,

$$w_1\mu_i^{(1)}+(1-w_1)\mu_i^{(2)}=I \tag{10}$$

$$\mu_i^{(2)}=\frac{1-w_1\mu_i^{(1)}}{1-w_1} \tag{11}$$

(11)式으로 부터, 固有值 $\mu_i^{(1)}$을 크기 順으로 놓으면, $\mu_i^{(2)}$는 작은 順으로 $\mu_i^{(1)}$에 1對1로 對應함을 알 수 있다. 卽,

$$\left.\begin{array}{l}\mu_1^{(1)}>\mu_2^{(1)}>\cdots\cdots>\mu_n^{(1)}\\ \mu_1^{(2)}<\mu_2^{(2)}<\cdots\cdots<\mu_n^{(2)}\end{array}\right\} \tag{12}$$

또한,

$$K_lP=U^{(l)}P \quad (l=1,2) \tag{13}$$

의 關係가 成立하므로 直交行列 P를 列벡터 p_i로 表現하면, $P=[p_1,p_2,\cdots\cdots,p_n]$이므로, (13)式으로부터

$$\left.\begin{array}{l}K_lP=[K_lp_1,K_lp_2,\cdots\cdots,K_lp_n]\\ \quad=[\mu_1^{(l)}p_1,\mu_2^{(l)}p_2,\cdots\cdots,\mu_n^{(l)}p_n]\end{array}\right\} \tag{14}$$

의 式을 얻는다.

(14)式으로 부터, 混合行列正規化 algorithm은, 同一座標系上에서 兩多變數群間의 "差"를 最大로 하는 手法임을 알 수 있다. 卽, 이 algorithm은 兩多變數群에 關해 共通性質을 特徵으로 抽出하고자 할 경우에는 不適하다.

한편, 一群正規化(one-class normalization) algorithm[3]은 混合正規化(mixture normalization) 手法의 特殊한 境遇($w_1=1$, $w_2=0$; $w_1=0$, $w_2=1$)로 생각할 수 있으므로, 統一해서 다룰 수 있다.

2.3 特徵抽出方法

本節에서는, 本論文에서 提案한 混合正規化手法에 의한 具體的인 特徵抽出方法에 關해 記述한다.

2.3.1 多變數群의 代表特徵抽出

크기 順으로 늘어 놓은 固有值 $\mu_i^{(l)}$의 N번째까지의 합이 全固有值의 90%를 넘으면, 變換行列 T의 成分 t_{ij}에 對해,

$$v_j=\sum_{i=1}^{N}|t_{ij}| \quad (N<n) \tag{15}$$

을 變數 j의 不完全絶對 weight합 (INAWS; INcompete Absolute Weight Sum)이라 定義한다. INAWS의 값이 큰 順으로 多變數群의 代表的인 特徵으로서 抽出한다. 이렇게 抽出된 特徵을 多變數群의 "代表特徵(representative feature)"이라 부르기로 한다.

結局 '主要한 固有值에 對한 x_j變數의 寄與度가 큰 順序대로 代表特徵을 抽出해 간다. 大槪의 境遇, INAWS v_j를 큰 順으로 더해, 그 합이 全 INAWS

$$v=\sum_{j=1}^{n}\sum_{i=1}^{N}|t_{ij}| \tag{16}$$

의 50% 程度가 되도록 變數의 數를 選擇하던지, INAWS가 바로 앞의 INAWS의 半以下가 되는 곳 까지를 特徵變數로 選擇하는 것이 좋다고 생각된다.

2.3.2 多變數群의 個別特徵抽出

混合行列을 正規化했을 때, 多變數群의 '微細한' 特徵을 "個別特徵(respective feature)"이라 부르기로 한다. 變換後의 座標軸에 對한 變數의 寄與는 (8)式으로 表現된다. N個의 z_i에 對한 變數의 絶對 weight합 (absolute weight sum)

$$d_i=\sum_{j=1}^{n}|t_{ij}| \quad (i=1,2,\cdots\cdots,N) \tag{17}$$

을 計算, 各 i에 對해 크기 順으로 $|t_{ij}|$를 整理해 Q번째까지의 합 d_i'가 d_i의 50%以上이 되도록 Q를 選擇한다.

$$d_i'=\sum_{j=1}^{Q}|t_{ij}| \tag{18}$$

選擇된 Q個의 變數의 相互關係를 多變數群의 個別特徵으로 抽出한다. t_{ij}의 符號自體는 별로 意味가 없으며, 抽出한 變數는 兩多變數間의 差가 크므로, 兩群

을 比較해 變數間의 相互關係를 定해야 할 것이다.

3. 應 用 例

本 algorithm을 實際 data에 適用해서 特徵抽出實驗한 結果를 提示한다. 使用한 data는 韓國(1965年～1974年)과 日本(1964年～1973年)의 12變數에 對한 初年度比增加率이다. 變數는 人口 (x_1), GNP(x_2), 輸出(x_3), 敎育豫算(x_4), 國民所得(x_5), 大學生數(x_6), 人文高校學生數(x_7), 實業高校學生數(x_8), 大學院卒業者數(x_9), 大學進學率(x_{10}), 高校進學率(x_{11}), 特許登錄件數(x_{12})로 經濟와 敎育關係變數다. 日本과 韓國의 相關行列을 各各 R_1, R_2로 해, 因子分析의 結果와 一群正規化 및 混合正規化($w_1=w_2=0.5$) algorithm에 의한 特徵抽出結果를 보인다.

3.1 因子分析結果

相關行列 R_1, R_2에 直接배리맥스法(direct varimax method)을 適用해 主因子및 因子負荷(factor loading)을 求했다. 相關行列 $R=\{r_{ij}\}$ $(i, j=1, 2, \cdots\cdots, n)$의 對角成分 r_{ii}에는 1代身, 各行의 最大値의 絕對値 $|\max\{r \cdot j\}|$를 共通性(communality)으로 넣어 分析했다. 그 結果는 아래와 같다.

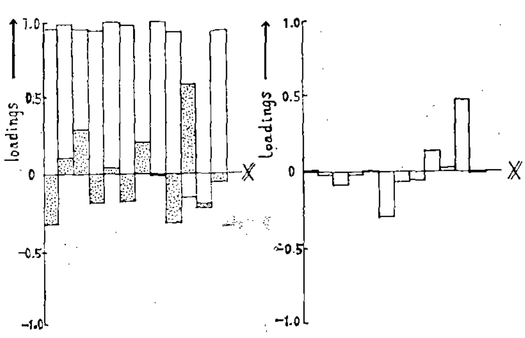
(가) 韓國(R_2)의 因子分析結果

3個의 因子가 抽出되었다. 各 因子의 寄與 및 因子에 對한 各 變數의 負荷(loading)를 그림 1에 圖示한다.

第1 因子는 x_{10}과 x_{11}을 除外한 全變數의 負荷量이 크므로, "經濟·敎育 兩面에서의 量的 高度成長에 關한 因子"라 볼 수 있을 것이다.

第2 因子는 x_{10}에 對한 因子로서, "大學進學率의 增加에 關한 因子"다.

또, 第3 因子는 x_{11}의 因子다. 卽, "高校進學率의 增加에 關한 因子"로서, 以上의 結果에서, 大學 및 高校進學率의 增加가 딴 變數와 獨立된 變動을 보이는 點이 興味롭다. 이는 他變數의 增加率에 比해 x_{10}과 x_{11}의 增加率이 적은데 起因한다고 생각된다.



(a) First factor (□) contribution: 9.46
Second factor (▦) contribution: 0.81

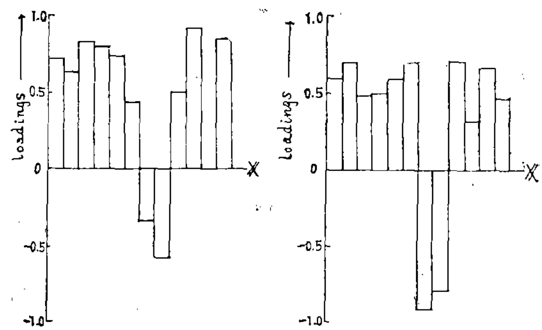
(b) Third factor contribution: 0.35

그림 1. R_2의 因子分析結果
Fig. 1. Factor analysis results for R_2.

(나) 日本(R_1)의 因子分析結果

두개의 因子가 抽出되었다. 因子의 寄與 및 因子에 對한 各 變數의 寄與를 그림 2에 보인다.

第1因子는 $x_1, x_3, x_4, x_5, x_{10}$과 x_{12}에 關한 因子로서, "敎育豫算 및 所得의 向上에 의한 大學進學率 및 硏究成果의 增加" 因子라 볼 수 있다.



(a) First factor contribution: 5.85

(b) Second factor contribution: 4.96

그림 2. R_1의 因子分析結果
Fig. 2. Factor analysis results for R_1.

한편, 第2 因子는 x_2, x_6, x_7, x_8, x_9와 x_{11}에 對한 因子다. 여기서, 눈을 끄는 것은 高校進學率과 高校學生數와의 負相關關係다. 卽, 進學率이 높아짐에도 不拘하고 學生數가 늘지 않는 (實際로는 줄고 있다) 點이다. 그러므로, 第2因子는 "高等敎育의 增加와 高等學校敎育의 義務敎育化(또는 飽和)傾向" 因子라 볼 수 있을 것이다.

3.2 混合行列의 最大固有値의 變化

위 data의 混合行列 $A=w_1R_1+w_2R_2$의 最大固有値 $\max(\lambda_i)$는 weight w_1의 變化에 따라 그림 3과 같이 變한다. 最小自乘法에 의해 近似한 結果

$$\max(\lambda_i)=9.42w_1^2-7.86w_1+9.58 \tag{19}$$

의 式을 얻었다. 여기서, 最大固有値가 全固有値中에서 占하는 比率이 最小로 되는 것은 w_1, w_2가 0.4～0.6 사이에 있을 때라는 것을 알 수 있다. 한편, $w_1=w_2=0.5$인 境遇, A는 平均相關行列(mean correlation matrix)이라 부를 수 있으며, 두개의 多變數群을 比較할 때, 大概의 境遇 weight 또는 發生確率은 兩多

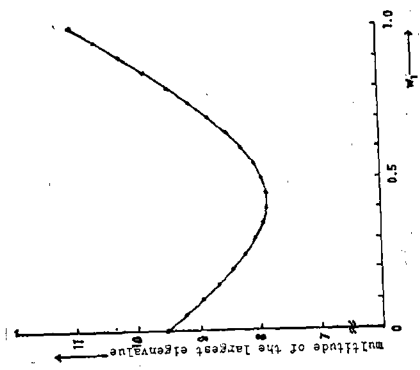


그림 3. w_1에 따른 最大固有値의 變化

Fig. 3. Variation of the largest eigenvalue with w_1.

變數群 모두 0.5라 생각해도 무방할 것이다. 또한 그림 3에서 알 수 있는 바와 같이, 最大固有値가 全固有値中의 占有率이 w_1, w_2가 0.5附近에 있을 때 最小가 되므로, 多變數群의 少數變數에 의한 代表的 特徵抽出에는 有效한 一手法이라 생각한다.

3.3 一群正規化에 의한 特徵抽出結果

(가) R_1을 正規化한 境遇($w_1=1$, $w_2=0$)

固有値中에서 크기順으로 6個를 표 1에, 最大固有値와 두번째로 큰 固有値에 對應하는 變換行列을 그림 4에 보인다. 두 固有値의 합은 全固有値의 98.5%다.

표 1. K_2의 固有値($w_1=1$, ${}_2w=0$)

Table 1. Eigenvalues of $K_2(w_1=1, w_2=0)$.

μ_1	5.30×10^4
μ_2	8.64×10^3
μ_3	6.28×10^2
μ_4	1.43×10^2
μ_5	1.03×10^2
μ_6	4.84×10^1

代表特徵은 少數의 變數로 表現하기 어려우나, 個別特徵은 最大 固有値로 부터 x_4, x_6, x_8과 x_{11} (12變數全寄與의 61%), 두번째 固有値로 부터 x_2, x_5와 x_{10} (12變數全寄與의 54%)가 抽出되었다. 卽, 路本을 基準으로 보았을 때, 韓國의 特徵은 敎育豫算, 大學生數 및 實業高等學校學生數의 增加에 比해 高校進學率의 增加가 적은 點과, 所得의 增加에 比해 GNP의 增加가 크며, 大學進學率의 增加가 작은 點이라 할 수 있다.

(나) R_2를 正規化한 境遇($w_1=0$, $w_2=1$)

固有値中에서 큰 順으로 6個를 표 2에, 最大固有

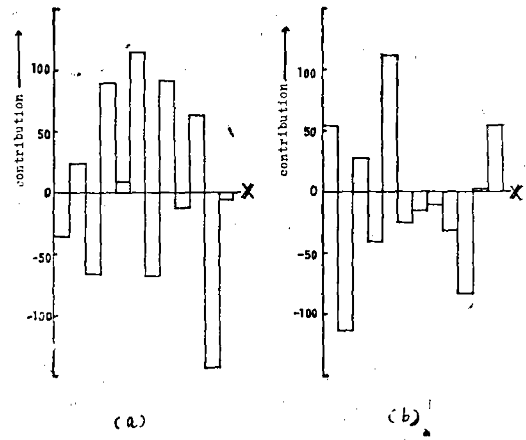


그림 4. 最大固有値(a) 및 두번째 큰 固有値(b)에 對應하는 變換行列

Fig. 4. Transform matrix associated with the largest(a) and the second largest(b) eigenvalues.

표 2. K_1의 固有値($w_1=0$, $w_2=1$)

Table 2. Eigenvalues of $K_1(w_1=0, w_2=1)$.

μ_1	2.75×10^5
μ_2	1.38×10^4
μ_3	1.31×10^2
μ_4	5.01
μ_5	1.34
μ_6	0.50

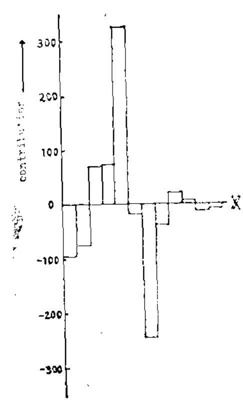


그림 5. 最大 固有値에 對應하는 變換行列

Fig. 5. Transform matrix associated with the largest eigenvalue.

値에 對應하는 變換行列을 그림 5에 보인다. 最大固有値는 全固有値의 95.2%를, 또한 그 座標軸에 對한 全寄與의 57.2%를 x_5와 x_7이 占하고 있다. 그러므로,

韓國을 基準으로 생각하면, 日本의 特徵은, 所得의 增加에도 不拘하고 人文高校學生數의 增加가 적은 點이다(實際는 줄고 있다). 卽, 高等學校敎育의 義務敎育化現象이 特徵이라 볼 수 있다.

3.4 混合行列正規化에 의한 特徵抽出結果 ($w_1=w_2=0.5$인 경우)

(가) 平均相關行列 正規化時의 R_1의 代表特徵

固有値를 표 3에, 1.5보다 큰 固有値에 對應하는

표 3. K_1의 固有値($w_1=w_2=0.5$).
Table 3. Eigenvalues of $K_1(w_1=w_2=0.5)$.

μ_1	0.79	μ_7	1.67
μ_2	2.00	μ_8	0.01
μ_3	0.02	μ_9	2.00
μ_4	0.00	μ_{10}	0.00
μ_5	2.00	μ_{11}	0.00
μ_6	−0.03	μ_{12}	1.15

표 4. R_1의 1.5보다 큰 固有値에 對應하는 變換行列
Table 4. Transform matrix associated with eigenvalues larger than 1.5 (for R_1).

z_2	−0.25	−0.79	−0.41	−0.28	2.74	0.25	1.32	0.57	−0.31	0.09	−0.03	−0.28
z_5	−0.12	−18.9	−4.76	−3.42	30.8	4.68	−8.12	8.37	−5.27	1.22	0.12	−3.78
z_7	−1.39	−4.26	−1.41	−4.22	10.3	−6.27	−2.90	1.76	4.88	0.11	−1.74	3.16
z_9	0.10	2.76	−2.21	−1.90	1.40	0.42	−1.46	1.55	−0.46	0.09	−0.02	−0.30
v_j	1.89	9.62	8.79	9.82	45.24	11.62	13.80	12.25	10.92	1.51	1.91	7.52

變換行列을 표 4에 보인다. x_5의 INAWS는 全變數의 NAWS의 34%이며, 다음으로 큰 INAWS가 x_5의 半以下이므로 x_5를 R_1의 代表特徵으로서 抽出한다. 卽, 平均相關을 基準으로 본 日本의 特徵은 딴 變數에 比해 國民所得의 增加가 큰 驗이라 볼 수 있다.

(나) 平均相關行列正規化時의 R_2의 代表特徵

固有値를 표 5에, 1.5보다 큰 固有値에 對應하는 變換行列을 표 6에 보인다. x_2와 x_5의 INAWS는 全

표 5. K_2의 固有値($w_1=w_2=0.5$)
Table 5. Eigenvalues of $K_2(w_1=w_2=0.5)$.

μ_1	1.21	μ_7	0.33
μ_2	0.00	μ_8	1.99
μ_3	1.98	μ_9	0.00
μ_4	2.00	μ_{10}	2.00
μ_5	0.00	μ_{11}	2.00
μ_6	2.03	μ_{12}	0.85

표 6. R_2의 1.5보다 큰 固有値에 對應하는 變換行列
Table 6. Transform matrix associated with eigenvalues larger than 1.5 (for R_2).

z_3	−2.51	−7.74	0.10	−0.27	6.21	6.34	−2.63	2.56	−3.19	0.25	0.52	0.56
z_4	−0.93	−1.64	−0.93	2.83	−0.09	0.96	−2.02	2.44	−0.19	0.76	0.32	−0.41
z_6	6.11	11.9	8.38	2.64	−22.7	−5.39	1.88	−2.74	1.78	−1.00	−0.76	−2.04
z_8	1.46	−4.39	3.35	−2.88	3.24	0.18	−3.90	4.75	−0.16	0.63	0.73	−0.90
z_{10}	−0.64	2.18	−0.42	0.21	−1.90	−0.29	0.93	−0.73	0.61	0.86	0.24	−0.66
z_{11}	0.49	−1.09	0.49	−0.74	0.94	−0.69	0.16	−0.42	−0.17	−0.85	0.84	0.39
v_j	12.14	28.92	13.67	9.57	35.08	13.84	11.52	13.64	6.10	4.35	3.41	4.96

變數의 INAWS의 41%이다. 한편, 本 algorithm은 같은 性質을 特徵으로 가지지 않는 特性이 있으므로, 平均相關을 基準으로 본 때의 韓國의 代表特徵은 國民所得의 增加에 比해 GNP의 增加가 큰 驗이라 볼 수 있다.

3.5 分析結果의 比較 및 檢討

以上의 分析結果로 부터, 두 多變數群間의 特徵抽出手法으로서의 一群正規化나 混合行列正規化手法은, 外部基準을 設定하지 않는 主成分分析이나 因子分析에 의해 抽出되기 어려운 相對的인 "差異"의 抽出에 有效한 一手法이라 생각된다. 卽, 本 algorithm은 比較基準인 混合行列의 正規化變換을 各各의 多變數群에 對해 施行함으로써, 基準과 同一한 尺度로 比較할 수 있는 利點을 가지고 있다. 한편, 一群正規化手法은 多變數群의 個別特徵의 抽出에 混合正規化手法은 代表特徵의 抽出에 適合하다 생각된다.

4. 結 論

本 論文에서는 直交變換을 利用한 두 多變數群의 同時對角化手法에 의한 混合行列正規化 algorithm과 그

에 隨伴되는 몇몇의 性質 및 그를 利用한 多變數群間의 特徵抽出法에 關해 記述했다. 또한, 實際데이터에 對한 特徵抽出結果 및 因子分析結果를 比較·檢討했다. Fukunaga等[(2),(3)]은 sample을 clustering하기 위해, 平均共分散行列을 正規化해, 線型分離函數를 求했으나 本 論文에서는 混合行列正規化에 따른 多次元座標系의 變換에 着目하여, 變換後의 座標軸에 對한 變換前의 變數座標軸의 寄與를 計算하여 두 多變數群間의 特徵을 抽出하는 手法을 提案, 그의 有效性을 適用實驗에 의해 檢證했다.

本 algorithm은 두 多變數群間의 "差異"를 同一座標軸上에서 되도록 크게 해 特徵으로서 抽出하므로, 因子分析等 他手法으로 抽出하기 어려운, 작은 差異를 抽出하는 境遇, 有效한 一手法이라 생각된다. 將來, 다른 時系列데이터等에의 應用에 關해서 檢討해 갈 豫定이다.

參 考 文 獻

1. 芝 祐順：因子分析法, p.1, 東京大學出版會, 1972.
2. K. Fukunaga & W.L.G. Koontz: "Application of the Karhunen-Loeve Expansion to Feature Selection and Ordering", IEEE Vol. C-19, No. 4, pp.311~318, 1970.
3. K. Fukunaga: "Introduction to Statistical Pattern Recognition", Academic Press, 1972.
4. 安居院, 中島, 吳： 經濟と敎育の相互關係について, 電子通信學會技術硏究報告, ET 76-5, pp.33~36, 1976.
5. T. Agui, M. Nakajima & Y. OH: "A Method of Features Extraction Between Two Multivariable Classes", Information Processing Society of Japan, Vol.18, No.12, pp. 1231~1235, Dec 1977.