教育資料
89-03-015

多 變 量 分 析

經濟企劃院 調査統計局

目　　次

PART A. 多變量分析(Multivariate Analysis) 및 多變量正規分布(Multivariate Normal Distribution)

一般的으로 統計解析을 하는 경우에 資料의 變化가 한 種類의 特性에 의해서만 影響을 받는 경우가 있다. 이 경우에는 1變量에 대한 통계해석이 필요하다. 그러나 일반적인 社會現象은 여러종류의 원인으로 나누어 이 현상의 특성을 규명하기 위한 分析이 必要하게 되는데, 이와 같은 統計的 分析을 多變量分析(multivariate analysis)이라고 한다. 즉, 多變量分析이란 分析對象 또는, 현상의 특성을 나타내는 세 개 이상의 變數들에 대해 觀測한 資料들을 바탕으로 이들의 상호 연관관계를 把握하는 統計的인 제반 節次를 말한다.

多變量分析의 技法은 지난 50년 동안 많은 統計學者들의 努力으로 발전되어 왔는데 이 方法은 자연과학, 사회과학, 인문과학등 모든 分野에 適用되고 있으며, 그 발전에는 컴퓨터가 큰 貢獻을 하였다.

1. 多變量 分析技法

多變量 分析技法은 資料의 構造나 分析目的에 알맞게 여러가지가 開發되어 왔다. 이들은 크게 從屬關係 分析方法(dependence method)과 級間從屬關係 分析方法(interdependence method)의 두 가지 범주로 區分된다.

A. 從屬關係 分析方法

從屬關係 分析方法이란 獨立變數群의 資料를 바탕으로 從屬變數群의 性質을 설명 또는 豫測하기 위해 開發된 多變量 分析技法들을 總稱하는 것이다. 이 分析方法 中 많이 使用되는 技法들은 다음과 같다.

1) 중회귀분석 (multiple regression)

2) 판별분석 (discriminant analysis)

3) 정준분석 (canonical correlation analysis)

4) Logit 분석

5) MANOVA

B. 級間從屬關係 分析技法

이 方法은 變數群을 從屬變數와 獨立變數群으로 분리시키지 않고 모든 變數들이 동등한 役割을 한다는 假定下에서 이들로 부터 얻은 多變量 資料에 包含된 情報를 간단한 형태로 바꾸어 해석하는 技法들을 말한다. 그러므로, 이 方法의 主 目的은, 從屬關係方法이 追求하는게 從屬變數群의 豫測이라면, 多變量 資料가 가진 情報를 간단히 把握할 수 있도록 어떤 체계를 導出하는데 있다. 이 分析方法들 중 특히 많이 쓰는 技法은 다음과 같다.

1) 주성분분석 (principal component analysis)

2) 인자분석 (factor analysis)

3) 집락분석 (cluster analysis)

4) MDS 법 (multidimensional scaling)

5) 대수선형모형분석 (loplinear model)

2. 多變量 資料의 形態

가령, 어떤 社會現象을 分析하려면 分析者는 分析의 目標를 세운후 그 目標에 알맞는 分析技法을 定해야 한다. 먼저 分析技法을 선택할때 考慮해야 할 事項은 컴퓨터 프로그램의 有無이다. 가능한한 컴퓨터 프로그램을 쉽게 구할 수 있는 技法을 택하여야 한다. 다음 절차는 社會現象의 특성을 資料化하여 蒐集하는 것이다. 資料 蒐集에 있어 유의해야 할 事項은 分析目標에 맞게 어미 선택된 多變量 分析技法을 念頭에 두고 이 技法에 알맞는 형태의 資料를 蒐集할 수 있도록 標本設計하는 것이다. 현상의 특성을 資料化하여 蒐集하는 方法은 여러가지 형태로 이루어진다. 資料의 형태는 다음과 같이 세 가지 형태로 分類된다.

① 명목형資料 (Nominal scale data)

資料가 상호, 성별등의 자연적인 分類와 합격, 불합격 또는 高所得, 中所得, 低所得등 인위적인 分類로 명목화된 형태를 가진 것.

② 순서형資料 (ordinal scale date)

현상의 특성을 크기, 선호도 등의 순서로 순위 또는 순서화 시킨 資料.

③ 기수형資料 (Cardinal scale date)

가장 일반적인 資料形態로 구간, 차이 및 比率을 나타낼 수 있게 만들어진 資料

3. 多變量 正規分布

多變量正規分布(multivariate normal distribution)은 多變量分析技法의 기초가 되는 分布로써 대부분의 技法에서 使用되는 유의성검정(significance test)들은 모든 資料들이 이 分布를 따른다는 假定下에서 이루어진다. 多變量分析에서 이 分布의 役割을 열거하면 다음과 같다.

① 임의의 分布를 가진 多變量 資料들의 표준화합(standardized sum)은 標本의 크기가 클때 그 資料의 分布에 관계없이 多變量正規分布를 한다.(중심극한 정리).

② 임의의 分布하에서 導出해 낼 수 없는 분석절차를 정규분포의 假定下에서는 쉽게 導出해 낼 수 있는 경우가 많다.

A. 多變量正規分布의 特性

결합확률분포 $f(x)=f(\chi_1,\chi_2,\cdots\cdots,\chi_p)$ 를 가진 確率變數벡터 $X:_{p\times1}$ 의 결합확률분포가 다음의 형태를 가지면 確率벡터 X는 平均벡터가 $\theta=(\theta_1,\theta_2,\cdots,\theta_p)'$ 이고 分散－共分散 行列(covariance matrix) 이 $\Sigma=\{\sigma_{ij}\}, j,i=1,\cdots,p$ 인 多變量正規分布를 한다고 하고, $X\sim N(\theta,\Sigma)$ 로 표시한다.

〈다 음〉

$$f(X)=\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(X-\theta)'\Sigma^{-1}(X-\theta)\},\ \text{for}\ \Sigma\rangle \qquad o(1-1)$$

위 분포의 중요한 성질들은 다음과 같다.

① 만약 $X \sim N(\theta, \Sigma)$이면, X의 모든 부분 벡터(subvector) 들도 모두 正規分布를 한다. $Y \sim N(\theta_y, \Sigma_{11})$, $Z \sim N(\theta_z, \Sigma_{22})$

단, $X = \begin{pmatrix} Y \\ Z \end{pmatrix}$, $\theta = \begin{pmatrix} \theta_y \\ \theta_z \end{pmatrix}$, $\Sigma = \begin{bmatrix} \Sigma_1, \Sigma_{12} \\ \Sigma_2, \Sigma_{22} \end{bmatrix}$

② 만약 $X \sim N(\theta, \Sigma)$이면, 벡터 X의 p개 원소 $\{X_1, X_2, \cdots\cdots, X_p\}$들은 각각 單變量正規分布를 한다.

즉, $X_i \sim N(\theta_i, \sigma_{ii})$.

③ 만약 $X \sim N(\theta, \Sigma)$ 이고, 벡터 X가 ①의 경우와 같이 $X = (Y', Z')'$로 分割되었으면, Y와 Z의 條件附確率分布는 역시 多變量正規分布를 따른다.

$Y \mid Z = z \sim N[\theta_y + \Sigma_{12}\Sigma_{22}^{-1}(z-\theta_z), \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]$,

$Z \mid Y = y \sim N[\theta_z + \Sigma_{21}\Sigma_{11}^{-1}(y-\theta_y), \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}]$.

④ 만약 $X \sim N(o, \Sigma)$이고, Σ가 대각행렬일 때(즉 $\Sigma = \mathrm{diag}(\sigma_{11}, \sigma_{22}, \cdots\cdots, \sigma_{pp})$), 確率벡터 $X = (x_1, x_2, \cdots\cdots, x_p)'$의 각 원소는 서로 獨立이다.

B. 多變量正規分布下의 統計量

統計量(statistic)이란 모집단으로 부터 추출된 標本變量들의 함수를 의미한다.

만약 $X_1, X_2, \cdots\cdots, X_N$이 多變量正規分布 $N(\theta, \Sigma)$를 따르는 모집단 (모집단의 變數는 X로 나타냄)으로 부터 랜덤추출된 標本(random sample)이면, 이들의 分布는 서로 獨立이고 모집단과 같은 正規分布를

따른다. 이것을

$X_1, X_2, \cdots\cdots, X_N \overset{iid}{\sim} N(\theta, \Sigma)$

로 표기한다.

① 모집단의 母數推定

正規모집단의 母數는 平均벡터 θ와 共分散行列 Σ이다. 이것의 값이 사전에 알려져 있지 않을 경우 標本을 통하여 이들 값을 推定하는데, 推定法에는 최우추정법 (maximum likelihood estimation), 積率法(method of moments)등이 있다.

최우추정법이란 N개 標本 $X_1, X_2, \cdots\cdots, X_N$의 결합확률분포(또는 우도함수)

$$L(\theta, \Sigma) = P(X_1, \cdots, X_N) = \frac{1}{(2\pi)^{\frac{NP}{2}} |\Sigma|^{\frac{N}{2}}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{N}(X_j-\theta)'\Sigma^{-1}(X_j-\theta)\right\} \quad \cdots\cdots (1-2)$$

를 최대화시키는 θ와 Σ의 解, 즉

$$\frac{\partial L(\theta,\Sigma)}{\partial \theta} = 0 \text{ 와 } \frac{\partial L(\theta,\Sigma)}{\partial \Sigma} = 0$$

를 연립시켜 얻은 θ와 Σ의 解를 이들의 點推定量으로 택하는 方法으로, 벡터미분을 통해 얻은 이들의 解는 다음과 같다.

$$\hat{\theta} = \frac{1}{N}\sum_{j=1}^{N} X_j = \bar{X}, \qquad \cdots\cdots (1-3)$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{j=1}^{N}(X_j-\bar{X})(X_j-\bar{X})'.$$

이들 중 $\hat{\Sigma}$은 불편추정량(unbiased estimator)이 아니므로, 통상 이것 대신 모수 Σ의 불편추정량인

$$S = \frac{1}{N-1} \sum_{j=1}^{N} (X_j - \bar{X})(X_j - \bar{X})' \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (1-4)$$

을 多變量分析에서 많이 使用한다. 여기서 Σ의 불편추정량이란 Σ의 추정량(推定에 使用되는 統計量)인 S의 기대값(expected value)의 Σ와 같은 값을 가지는 성질을 말한다. 즉,

$$E[S] = \Sigma$$

이 성립함을 의미한다.

② 獨立性檢定(Test for independence or Sphericity test)

앞에서 言及한 것과 같이 $X \sim N(\theta, \Sigma)$ 이고, 共分散行列이 對角行列(diaponal matrix)의 형태를 가지면 $X = (x_1, x_2, \cdots\cdots, x_p)'$의 각 원소 $\{x_1, x_2, \cdots\cdots, x_p\}$인 單變量變數(Univariate variable)들은 서로 獨立의 성질을 가진다.

共分散行列 Σ는 相關係數行列(correlation matrix)인 R과 다음과 같은 關係인

$$\mathrm{Var}(X) = \Sigma = \begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & \cdots\cdots & x_p \end{array} \\ \begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1,x_2) & \cdots\cdots & \mathrm{Cov}(x_1,x_p) \\ \mathrm{Cov}(x_2,x_1) & \mathrm{Var}(x_2) & \cdots\cdots & \mathrm{Cov}(x_2,x_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_p,x_1) & \cdots\cdots & & \mathrm{Var}(x_p) \end{bmatrix} \end{array},$$

$$\mathrm{Corr}(X) = R = \begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & \cdots\cdots & x_p \end{array} \\ \begin{bmatrix} 1 & \mathrm{Corr}(x_1,x_2) & \cdots\cdots & \mathrm{Corr}(x_1,x_p) \\ \mathrm{Corr}(x_2,x_1) & 1 & \cdots\cdots & \mathrm{Corr}(x_2,x_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Corr}(x_p,x_1) & \cdots\cdots & & 1 \end{bmatrix} \end{array}$$

단, $Corr(x_i, x_j) = \frac{Cov(x_i, x_j)}{\sqrt{Var(x_i)\ Var(x_j)}}$ 이므로

$$R = D^{-\frac{1}{2}}\ \Sigma D^{-\frac{1}{2}} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (1-4)$$

이 된다. 여기서, $D = diap(var(x_1), \cdots\cdots, var(x_p))$이다. 그러므로, 正規確率벡터X가 獨立性을 가졌다는 의미는 X의 相關係數行列 R이 單位行列 (identity matrix) 형태임을 의미한다. 이것은, 式(1 - 4)의 關係式으로부터 알 수 있다.

獨立性檢定이란 정규모집단 $N(\theta, \Sigma)$로 부터 추출된 資料 $X_1, X_2\cdots\cdots, X_N$을 가지고 추정한 標本相關係數行列 $\hat{R} = \hat{D}^{-\frac{1}{2}}\ \hat{\Sigma}\ \hat{D}^{-\frac{1}{2}}$ 을 토대로 모집단의 P개 變數 $\{ x_1, x_2, \cdots\cdots, x_p \}$가 서로 獨立의 성질을 가지고 있는지의 여부를 檢定하는 절차를 말하며, 이때 設定되는 가설은

$$Ho : R = 1 \quad V.S. \quad H_1 : R \neq I \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (1-5)$$

이다. 이때 사용되는 檢定統計量(test statistic)은 Bartlett에 의해 제안된

$$\lim_{N\to\infty}\left[-(N-1 - \frac{2p+5}{6}) \log |\hat{R}| \right] \sim \chi^2 \{\frac{P(P-1)}{2}\} \quad \cdots\cdots(1-6)$$

인 Chi - square 統計量이다.

③ K개 모집단 共分散行列의 同一性檢定

앞으로 설명할 多變量分析技法들 중에는 여러 모집단의 共分散行列이 同一하다는 假定下에서 資料分析을 하는 것이 있다. 이것은 共分散行列이 서로 다르면, 이 分析法을 사용할 수 없음을 의미한다. 그러므로 이런 假定을 가진 分析技法을 사용하려면 資料를 利用하여 共分散行列들이 同一

한 지를 당연히 檢定해 보아야 한다. 이것을 共分散行列의 同一性檢定 (testing for equality of covariance matrices)이라 한다. 檢定節次는 다음과 같다.

$X_1(_j), X_2(_i)$, ……, $X_{Ni}(_i)$를 $N(\theta_i, \Sigma_i)$, $i = 1, 2, \cdots\cdots, k$, 의 분포를 가진 i-번째 정규모집단으로 부터 얻은 N_i개 $p \times 1$ 觀測값벡터라 하자. 同一性 檢定에 使用될 가설(hypothesis)은

$$Ho : \Sigma_1 = \Sigma_2 = \cdots\cdots\cdots = \Sigma_k \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (1-7)$$

이다. 이때 使用되는 檢定統計量은 Box (1949)에 의해 제안된

$$-aG \overset{\cdot}{\sim} X^2_{\left\{\frac{P(P+1)(K-1)}{2}\right\}} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (1-8)$$

이다. 단, $G = c + \sum_{i=1}^{k}(N_i - 1) \log |Vi| - (N-K) \log |V_1 + \cdots\cdots + V_k|$

$$C = p(N-K) \log (N-K) - \sum_{i=1}^{k} p(N_i - 1) \log (N_i - 1),$$

$$a = 1 - \left(\sum_{i=1}^{K} \frac{1}{N_i - 1} - \frac{1}{N-K} \right) \frac{2p^2 + 3p - 1}{6(p+1)(K-1)}$$

$$b = \frac{p(p+1)}{48\,a^2} \left[(p-1)(p+2) \left(\sum_{i=1}^{K} \frac{1}{(N_i - 1)^2} - \frac{1}{(N-K)^2} \right) - 6(K-1)(1-a)^2 \right],$$

$$N = \sum_{i=1}^{K} N_i \quad , \quad V_i = (N_i - 1) S_i \ .$$

PART B. 相關分析

Ⅰ. 二變量 相關分析

SAS에는 두 變數간의 相關分析用으로 3가지 副프로그램인 PEARSON CORR, SPEARMAN CORR과 KENDALL CORR이 있다. PEARSON CORR은 定量的 變數(間隔尺度와 比率尺度 : cardinal scale)로 되어있는 두 變數에 대하여 標本相關係數(sample correlation coefficient)를 구하여 준다. SPEARMAN CORR과 KENDALL CORR는 非母數的 方法에 의해 順位尺度(ordinal scale)인 두 變數에 대해 順位相關係數(rank-order correlation coefficient)를 計算해 준다.

1. 피어슨 相關理論

1-1. 피어슨 相關係數

피어슨相關은 처음 피어슨(K.Pearson)에 의해 고안된 相關係數의 測度라는 점에서 후일 피어슨相關 또는 單純相關(simple correlation)이라 부른다. 이것은 앞으로 說明하게 될 順位相關, 部分相關, 重相關등과 區分된다. 피어슨相關이란 定量的 變數로 되어 있는 두 變數 x 및 y간의 線型的 强度의 測度로서, 이 相關係數는 다음과 같이 定義된다.

$$\rho_{xy} = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{v}(x)\ \mathrm{v}(y)}} \quad \cdots\cdots (1-1)$$

위의 係數를 定義에 따라 式으로 表現하면 x와 y의 標本相關係數 r_{xy}는 다음과 같이 얻어진다.

$$r_{xy}=\frac{n\Sigma xiyi-\Sigma xi\,\Sigma yi}{\sqrt{n\Sigma xi^2-(\Sigma xi)^2}\ \sqrt{n\Sigma yi^2-(\Sigma yi)^2}} \quad \cdots\cdots\cdots\cdots(1\text{-}2)$$

式(1-2)에서 n은 觀測값의 數를 가르킨다.

表(1•1)에 의해서 線型的 强度인 피어슨相關係數를 計算해 보기로 한다. 特別한 目的은 없으나 1次 成績을 x로 2次 成績을 y로 한다. 그러면 表(1•1)의 連算表와 式(1-2)에 의해서 피어슨相關係數인 r_{xy} 값은 다음과 같이 얻어진다.

〈表 1•1〉 피어슨相關係數 計算 例

事 例	x	y	xy	x^2	y^2
1	63	68	4,284	3,969	4,624
2	65	75	4,875	4,225	5,625
3	72	70	5,040	5,184	4,900
4	73	76	5,548	5,329	5,776
5	80	81	6,480	6,400	6,561
6	85	78	6,630	7,225	6,084
7	86	89	7,654	7,396	7,921
8	93	84	7,812	8,649	7,056
計	617	621	48,323	48,377	48,547

그림 (1•1)에서 보는 바와 같이 두 變數간의 函數關係가 增加函數이면 相關係數는 陽의 값을 갖는다. 이와는 반대로 두 變數간의 關係가 減少函數면 그것은 陰의 값을 갖게된다. 一般的으로 相關係數 ρ가 취하는 값의 範圍는

$$-1 \leq \rho \leq 1 \quad \cdots\cdots (1-3)$$

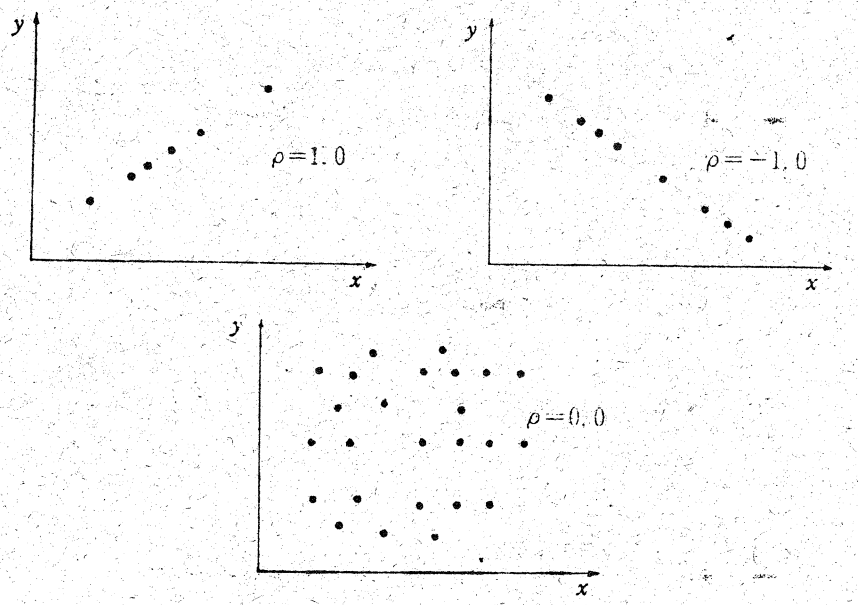
이다. 그리고, 相關係數 ρ가 취하는 값의 範圍에 따라 그 名稱이 다르다.

陰相關 : $-1 < \rho < 0$

陽相關 : $0 < \rho < 1$

零(또는 無)相關 : $\rho = 0$

完全相關 : $|\rho| = 1$



(그림 1•1) 점선도와 相關係數

1－2. 相關係數의 有意性 檢定

위와 같이 피어슨 相關係數는 大部分의 경우 標本資料에 基礎되어서 얻어진다. 그런데 두 變數간에 相關係數가 없음에도($\rho = 0$) 標本에 의해서 얻어진 標本相關係數 r은 標本抽出 變動에 의해 0보다 크게 혹은 작게 計算될 수도 있고, 반대로 相關係數가 있음($\rho \neq 0$)에도 불구하고 標本相關係數 r이 0에 가깝게 나타나는 경우가 있다. 이러한 理由로 標本相關係數 r을 基礎로 하여 母相關係數 ρ가 0과 有意的인 差異를 보이는지의 與否를 一定한 留意水準(α)下에서 檢定하여야 할 것이다. 즉

$$\text{Ho} : \rho = 0 \quad \text{v.s.} \quad \text{Ha} : \rho \neq 0$$

를 檢定統計量

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2) \cdots\cdots (1-4)$$

를 利用하여 檢定한다.

例를 들면, 〈表 1•1 〉의 標本相關係數를 留意水準 $\alpha = 0.05$로 假說檢定하면 檢定統計量(式 1－4)은 $t(\sigma = n-2)$의 分布를 하고 檢定統計값은

$$t\text{ 값} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.8238\sqrt{8-2}}{\sqrt{1-0.8238^2}} \fallingdotseq 3.56$$

이므로 t 값 $= 3.56 > t(\sigma : 0.975) = 2.447$에 의해 歸無假說(Ho)을 棄却하게 된다. 즉 두 變數간에는 陽相關이 存在한다고 본다.

(SAS 명령문)

PROC CORR ;	
VAR X Y ;	r_{xy} 계산
또는	
PROC CORR ;	r_{xy} r_{xz} r_{yz} 계산
VAR X Y Z ;	
또는	
PROC CORR ;	r_{xa} r_{yb} r_{xc}
VAR X Y Z ;	r_{ya} r_{yb} r_{yc} 계산
WITH A B C ;	r_{za} r_{zb} r_{zc}

2. 順位相關係數 理論

Spearman의 順位相關係數(r_s)와 Kendall의 順位相關係數(τ)는 두 變數의 母集團의 確率分布와 無關하므로 非母數的(nonparametic) 方法이다. 順位相關係數는 順位尺度의 變數간에 關聯性을 나타내는 計數로, r_s 와 τ 모두 (-1, 1) 사이에 있는 값을 취하기 때문에 Pearson의 相關係數 r 과 解析上의 意味는 같다.

2-1. Spearman의 順位相關係數

두 개의 變數 x, y의 n개 順位觀測값을 각각 xi, yi로 붙여서 xi와 yi의 順位差를 $di = xi - yi$라 하면 Spearman의 順位相關係數는

$$r_s = 1 - \frac{6\sum_{i=1}^{n} di^2}{n^3 - n} \quad \cdots\cdots (1-5)$$

이다.

그러나 x, y의 順位觀測값중에 각각 同順位(tie)가 나올 경우 r_s는 다음 公式으로 얻어진다.

$$r_s = \frac{Tx + Ty - \sum_{i=1}^{n} di^2}{2\sqrt{Tx\,Ty}} \quad \cdots\cdots (1-6)$$

여기서 Tx와 Ty는

$$Tx = \frac{n(n^2-1) - \Sigma Rx(Rx^2-1)}{12}$$

$$Ty = \frac{n(n^2-1) - \Sigma Ry(Ry^2-1)}{12}$$

으로 計算된 값이고, Rx와 Ry는 x, y에 대하여 同順位(tie)가 붙어진 갯수이다.

〈例〉

12名 學生의 1學期 成績 順位와 2學期 成績 順位를 調査한 結果 다음과 같다.

학 생	A	B	C	D	E	F	G	H	I	J	K	L
1學期 成績順位(x_i)	1	2	3	4	5	6	7	8	9	10	11	12
2學期 成績順位(y_i)	7	1	8	3	4	11	11	2	5.5	9	11	5.5

이 結果에서 2學期 成績 Y에는 順位 5.5를 붙인 갯수가 2, 順位 11을 붙인 갯수가 3개이다. 1學期 成績과 2學期 成績간의 順位相關係數를 구하라.

(풀이) x에는 同順位가 없으므로 $\Sigma Rx(Rx^2-1)=0$이고, y에는 2回의 同順位가 $Ry=2$, $Ry=3$으로

$$\Sigma Ry\,(\,Ry^2-1\,)=2\,(\,2^2-1\,)+3\,(\,3^2-1\,)=30$$

이다. 따라서

$$Tx=\frac{12\,(\,12^2-1\,)-0}{12}=143$$

$$Ty=\frac{12\,(\,12^2-1\,)-30}{12}=141.5$$

이고, 順位相關係數는

$$r_s=\frac{143+141.5-196.5}{2\sqrt{(143)\,(141.5)}}=0.309$$

이다. 여기서 $\Sigma di^2=\Sigma(\,xi-yi\,)^2=(\,1-7\,)^2+(\,2-1\,)^2+\cdots\cdots+(\,12-5.5\,)^2=196.5$로 얻어진 값이다.

〈 r_s에 대한 有意性檢定〉

두 變數는 x, y의 相關關係 有無에 관한 檢定은 檢定統計量으로 〈式 1－4〉에서와 같이

$$\frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}\sim t\,(\,n-2\,)\quad\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(\,1-7\,)$$

이 使用된다. 따라서 양쪽檢定에서는 留意水準 α에서

$$|\,t\text{ 값}\,|>t\,(\,n-2,\,\frac{\alpha}{2}\,)$$

이면 相關關係가 있다고 말할 수 있다.

〈SAS 명령문〉

```
PROC CORR SPEAMAN ;
 VAR    X   Y ;
```

2－2． Kendall의 타우(τ)

kendall의 τ는 다음 公式으로 計算된다.

$$\tau = \frac{S}{\sqrt{\frac{1}{2}n(n-1)-Tx}\ \sqrt{\frac{1}{2}n(n-1)-Ty}} \quad (1-8)$$

여기서 $S=P-Q$로 P는 xi를 크기 순으로 나열하였을때 (xi, yi)의 짝 중에서 yi가 크기 순으로 나열되어 있는 짝의 數(pairs in natural order)이고, Q는 크기의 逆順으로 나열되어 있는 짝의 數(pairs in reverse order)이다. Tx와 Ty는 각각 $\frac{1}{2}\Sigma t(t-1)$로 t는 두 變數 x, y의 각각에서 同順位를 이루는 갯수이다.

만약 同順位가 하나도 없다면

$$\tau = \frac{S}{\frac{1}{2}n(n-1)} \quad \cdots\cdots (1-9)$$

으로 S가 취할 수 있는 最大의 數는 $\frac{1}{2}n(n-1)$이 된다. 이 τ의 값의 範圍도 $-1 \leq \hat{\tau} \leq 1$이다.

〈例〉

두 變數 X, Y에 대하여 다음의 데이타가 얻어졌다. 여기서 X는 이미 크기 순으로 나열되어 있다.

Xi	2	5	7	9	23
Yi	4	5	20	2	17

kendall의 τ의 값을 구하라.

(풀이) 同順位가 X, Y에 모두 없으므로 式(1－9)를 使用한다. (Xi, Yi)=(2, 4)에 대하여 그 다음에 있는 4개 짝(pair)을 볼 때

(9, 2)를 除外한 3개 짝은 Y의 값이 4보다 크므로 크기순으로 되어 있다. 크기의 逆順으로 보면 (9, 2)의 1개 짝만이 逆順이다. 이처럼 따져나가면 다음 표를 作成할 수 있다.

Xi	Yi	크기順의 짝의 數	크기逆順의 짝의 數
2	4	3	1
5	5	2	1
7	20	0	2
9	2	1	0
23	17	0	0
		$P = 6$	$Q = 4$

따라서

$$S = P - Q = 6 - 4 = 2$$

이고, τ의 값은

$$\tau = \frac{S}{\frac{1}{2}n(n-1)} = \frac{2}{\frac{1}{2}(5)(4)} = 0.2$$

가 된다.

〈 τ에 의한 相關性 有無檢定〉

τ의 分布가 相關性이 없는 경우

$$\tau \sim N\left(0, \frac{4n+10}{9n(n-1)}\right)$$

이므로 檢定統計量

$$Z = \frac{\tau}{\sqrt{\frac{4n+10}{9n(n-1)}}} \sim N(0, 1) \quad \cdots\cdots\cdots\cdots\cdots\cdots (1 \sim 10)$$

을 利用하여 檢定한다. 만약 留意水準 α에서

$|Z$값$| > Z(\frac{\alpha}{2})$

이면 두 變數 X와 Y간에 相關關係가 있다고 말할 수 있다.

〈SAS 명령문〉

```
PROC CORR KENDALL ;
 VAR X  Y ;
```

3. SAS PROC CORR에 의한 例題

다음의 例는 PROC CORR을 利用하여 SKINFOLD資料의 Pearson, Spearman 및 kendall 相關係數를 구한 것이다. SKINFOLD資料는 50名을 對象으로 測定한 가슴(CHEST), 배(ABDOMEN), 팔(ARM)의 길이들로 構成되어 있다. 이들 간에 相關係數는 다음과 같다.

〈INPUT〉

```
DATA SKINFOLD;
  INPUT CHEST ABDOMEN ARM;
  CARDS;
  data lines

PROC CORR PEARSON SPEARMAN KENDALL
    OUT=SKINCORR;
  TITLE CORRELATION EXAMPLE;

PROC PRINT;
  BY _TYPE_ NOTSORTED;
  TITLE OUTPUT DATA SET CONTENTS;
```

〈OUTPUT〉

CORRELATION EXAMPLE

VARIABLE	① N	② MEAN	③ STD DEV	④ MEDIAN	⑤ MINIMUM	⑥ MAXIMUM
CHEST	50	14.41000000	5.25637611	15.00000000	4.00000000	26.00000000
ABDOMEN	50	16.01000000	7.19090270	15.00000000	3.00000000	38.00000000
ARM	50	4.15000000	1.48547386	4.00000000	2.00000000	9.00000000

⑧ PEARSON CORRELATION COEFFICIENTS / PROB > |R| UNDER H0:RHO=0 / N = 50 ⑨

	CHEST	ABDOMEN	ARM
CHEST	1.00000	0.61986	0.52973
	0.0000	0.0001	0.0001
ABDOMEN	0.61986	1.00000	0.42925
	0.0001	0.0000	0.0019
ARM	0.52973	0.42925	1.00000
	0.0001	0.0019	0.0000

⑩

SPEARMAN CORRELATION COEFFICIENTS / PROB > |R| UNDER H0:RHO=0 / N = 50

	CHEST	ABDOMEN	ARM
CHEST	1.00000	0.54825	0.51761
	0.0000	0.0001	0.0001
ABDOMEN	0.54825	1.00000	0.45687
	0.0001	0.0000	0.0009
ARM	0.51761	0.45687	1.00000
	0.0001	0.0009	0.0000

CORRELATION EXAMPLE

KENDALL TAU B CORRELATION COEFFICIENTS / PROB > |R| UNDER H0:RHO=0 / N = 50

	CHEST	ABDOMEN	ARM
CHEST	1.00000	0.41246	0.38974
	0.0000	0.0000	0.0002
ABDOMEN	0.41246	1.00000	0.33048
	0.0000	0.0000	0.0015
ARM	0.38974	0.33048	1.00000
	0.0002	0.0015	0.0000

TYPE OF OBSERVATION=MEAN

OBS	_NAME_	CHEST	ABDOMEN	ARM
1		14.41	16.01	4.15

TYPE OF OBSERVATION=STD

OBS	_NAME_	CHEST	ABDOMEN	ARM
2		5.25638	7.1909	1.48547

TYPE OF OBSERVATION=N

OBS	_NAME_	CHEST	ABDOMEN	ARM
3		50	50	50

TYPE OF OBSERVATION=CORR

OBS	_NAME_	CHEST	ABDOMEN	ARM
4	CHEST	1.00000	0.61986	0.52973
5	ABDOMEN	0.61986	1.00000	0.42925
6	ARM	0.52973	0.42925	1.00000

Ⅱ. 多變量 相關分析

지금까지 取扱한 相關은 2變量 χ, y간의 相關理論이다. 3變量 以上인 相關理論을 살펴보자. 一般的으로 k變量인 경우에 대하여 說明해 보기로 한다.

n개의 대상 각각에 k變量 $\chi_1, \chi_2, \cdots\cdots, \chi_k$에 대한 觀測값을

$(\chi_{1i}, \chi_2, \cdots\cdots, \chi_{ki})$ $i = 1, 2, \cdots\cdots, n$

라고 表示한다면 이들의 표본적률과 공분산행렬은 다음과 같이 定義된다.

1. 標本積率과 分散共分散行列

원점에 관한 標本積率 $\upsilon_{p1p2\cdots\cdots pk}$를

$$\upsilon_{p1p2\cdots\cdots p_k} = \frac{1}{n} \sum_{i=1}^{n} \chi_{1i}^{p1} \chi_{2i}^{p2} \cdots\cdots, \chi_{ki}^{pk} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2-1)$$

으로 定義하고, 平均에 관한 標本積率

$$\mu_{p1p2\cdots\cdots pk} = \frac{1}{n} \sum_{i=1}^{n} (\chi_{1i} - \overline{X_1})^{p1} (\chi_{2i} - \overline{X_2})^{p2} \cdots\cdots (\chi k_i - X_k)^{pk} \quad (2-2)$$

으로 定義한다.

지금 S_{ij}를

$$S_{ij} = \frac{1}{n} \sum_{l=1}^{n} (\chi_{il} - \overline{X_i}) (\chi_{jl} - \overline{X_j}) \quad (2-3)$$

로 두면

$i = j \Rightarrow S_{ii} = S_i^2$: χ_i의 標本分散

$i \neq j \Rightarrow S_{ij}$: χ_i와 χ_j의 標本共分散

이 된다. 그리고 S_i 를 χ_i 의 標本偏差라 부른다. $S_{ii} > 0$, $S_{jj} > 0$에 대하여

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}\, S_{jj}}} = \frac{S_{ij}}{S_i S_j} \qquad (2-4)$$

는 χ_i와 χ_j의 相關係數가 된다.

S_{ij}를 i行 j列의 元(element)으로 하는 $(k \times k)$行列을 $(\chi_1, \chi_2, \cdots\cdots, \chi_k)$의 標本共分散行列(sample variance - Covariance matrix)이라 하고 $\hat{\Sigma}$으로 나타내면

$$\hat{\Sigma} = \begin{bmatrix} s_{11} & s_{12} & \cdots\cdots & s_{1k} \\ & s_{22} & \cdots\cdots & s_{2k} \\ & \text{synon} & \ddots & \vdots \\ & & & s_{kk} \end{bmatrix}$$

와 같다. 또 $s_{11}, s_{22}, \cdots\cdots s_{kk} > 0$ 일때 式 (2-4)의 r_{ij}를 i行 j列의 元으로 하는 $(k \times k)$行列을 $(\chi_1, \chi_2, \cdots\cdots, \chi_k)$의 標本相關行列(sample Correlation matrix)이라 하고 $\hat{R}$으로 나타내면

$$\hat{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \cdots\cdots & r_{1k} \\ & r_{22} & r_{23} \cdots\cdots & r_{2k} \\ & \text{symm} & \ddots & \vdots \\ & & & r_{kk} \end{bmatrix}$$

와 같다. 앞으로는 $\hat{R}$의 行列式을 $|\hat{R}|$로 表示하고, 行列式 $|\hat{R}|$의 r_{ij}의 餘因數(cofactor)를 $\hat{R}_{ij}$로 表示하기로 한다.

2. 偏相關係數(Partial correlation coefficient)

k 變量($\chi_1, \chi_2, \cdots\cdots, \chi_k$)중 χ_i 와 χ_j 간의 k 次偏相關係數를 $r_{ij\cdot 12\cdots\cdots k}$ 로 나타내면

$$r_{ij\cdot 12\cdots\cdots k} = \frac{\bar{R}_{ij}}{\sqrt{\bar{R}_{ii}\,\bar{R}_{jj}}} \quad (i \neq j \;:\; ij = 1,2\cdots\cdots, k) \qquad (2\text{-}5)$$

와 같이 定義된다. 이 係數의 意味는, χ_i 와 χ_j 로 부터 각각 나머지 變數($\chi_1, \chi_2, \cdots\cdots, \chi_k$: χ_i, χ_j 包含하지 않음)들이 미치는 선형효과(linear effect)를 除去시키고 남은 殘差(residual)들 간의 相互關係를 나타내는 係數로 풀이할 수 있다. 또한 이것은 回歸分析에서 回歸模型의 設定節次(stepwise regression 등)에 使用된다.

〈例〉

30 年間 쌀의 단보當收穫量 χ_1 과 氣溫 χ_2 및 降雨量 χ_3에 대하여 각 變量間의 相關係數를 計算하여 각각

$$r_{12} = -0.34\,,\quad r_{13} = 0.27,\quad r_{23} = -0.20$$

을 얻었다. 偏相關係數 $r_{12\cdot 3}$ 및 $r_{13\cdot 2}$를 구하여라.

(단보 = 10 a)

〈解〉

$$r_{12\cdot 3} = -\frac{R_{12}}{\sqrt{R_{11}\cdot R_{22}}}$$

$$= \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} = -0.303$$

$$r_{13\cdot 2} = \frac{R_{13}}{\sqrt{R_{11}\cdot R_{33}}}$$

$$= \frac{\gamma_{13} - \gamma_{12} \cdot \gamma_{23}}{\sqrt{(1-\gamma_{12}^2)(1-\gamma_{23}^2)}} = 0.286$$

이 計算結果에 의하면 氣溫이 높은 해에는 쌀의 收穫量이 줄고, 降雨量이 많은 해에는 쌀의 收穫量이 增加하는 것을 알 수 있다.

〈 k 次偏相關係數의 有意性 檢定〉

歸無假說, $H_o : \rho_{ij \cdot 12 \cdots k} = o$ 를 檢定하기 위한 檢定統計量은

$$t = \frac{\sqrt{n-k-2}\ \gamma_{ij \cdot 12 \cdots k}}{\sqrt{1-\gamma^2_{ij \cdot 12 \cdots k}}} \sim t\,(n-k-2) \qquad (2-6)$$

이다. 만약 유의수준 α에서

$$|t\text{ 값}| > t\,(n-k-2 : \frac{\alpha}{2})$$

이면 양쪽檢定(two tailled test)에 의해서 두變數 χ_i, χ_j 간에 偏相關關係가 있다고 말할 수 있다.

3. 重相關係數

k變量 중 χ_i와 $\chi_1, \chi_2, \cdots, \chi_{i-1}, \chi_{i+1}, \cdots\cdots, \chi_k$ 와의 相關係數를 중상관계수(multiple correlation coefficient)이라 하고, $\gamma_{i \cdot 12 \cdots k}$ 로 表示하면, 이것은

$$\gamma_{i \cdot 12 \cdots k} = \sqrt{1 - \frac{|\hat{R}|}{\hat{R}_{ii}}}$$

로 計算된다. 또한 重相關係數는 重回歸分析(multiple regression analysis)와 밀접한 關係를 가지고 있는데, 重回歸式

$$\chi_i = \beta_o + \beta\chi_i + \cdots\cdots + \beta_{i-1}\,\chi_{i-1} + \beta_{i+1}\,\chi_{i+1} + \cdots\cdots + \beta_k\chi_k + e \qquad (2-8)$$

에서 얻어진 決定係數 R^{*2}

$$R^{*2}=\frac{SS_{reg}}{SS_{xi}}=\frac{\text{회귀방정식 (式 2 - 8)에 의해 설명된 } \chi_i \text{의 變動}}{\chi_i \text{의 總變動}}$$

과 다음 關係式을 成立시킨다.

$$\sqrt{R^{*2}} = \gamma_{i\cdot 12\cdots k} \qquad \cdots\cdots(2\text{-}9)$$

〈例〉

30年間 쌀의 단보當收穫量 χ_1과 氣溫 χ_2 및 降雨量 χ_3에 대하여 각 變量間의 相關係數를 計算하여 각각

$$\gamma_{12}=0.1,\ \gamma_{13}=0.1,\ \gamma_{23}=0.1$$

을 얻었다. 重回歸係數 $\gamma_{1\cdot 23}$를 구하여라.

〈解〉 式(2-7)로 부터

$$\gamma_{1\cdot 23}=\sqrt{1-\frac{|R|}{R_{ii}}}=\sqrt{1-\frac{0.972}{0.99}}=0.1348$$

$$|R|=\begin{vmatrix} 1 & 0.1 & 0.1 \\ 0.1 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{vmatrix}=0.972,\ R_{11}=\begin{vmatrix} 1 & 0.1 \\ 0.1 & 1 \end{vmatrix}=0.99$$

Ⅲ. 正準相關分析 (Canonical Correlation Analysis)

〈重要 用語 說明〉

① 正準變量 (canonical variates)

종속 및 獨立變數의 各 集團內에서 變數들의 선형결합을 만든 것으로, 이들 선형결합 사이에 相關關係가 높도록 만든 것을 말한다. 즉, 正準變量은 한 쌍의 선형결합으로 이루어 진다.

② 正準係數 (canonical coefficient)

正準變量과 원變數間에 선형결합係數.

③ 正準積載값 (canonical structere coefficient)

正準變量과 원變數間에 相關係數.

④ 正準相關係數 (canonical correlation coefficient)

각 쌍의 正準變量間의 相關係數를 말하며 λ_c로 表示한다.

1. 正準相關分析의 定義

正準相關分析은 한 개 以上의 變數들로 構成된 두 變數 集團들 사이의 相關係數 (canonical correlation)을 分析하는 것으로 Hotelling (1936)에 의해 提案되었다. 回歸分析에서는 한 개 또는 몇개의 獨立變數들이 한 개의 從屬變數에 어떤 影響을 미치는가 알아보기 위하여 相關關係를 檢討한다. 그런데 正準相關分析에서는 從屬變數와 獨立變數가 모두 하나보다 많을때, 從屬變數 集團과 獨立變數 集團의 相關係數를 다룬다. 일반적인 多變量 分析에서와 마찬가지로 正準分析은 모두 量的 (cardinal scale)인 變

數만 取扱한다.

2. 正準相關分析의 目的

正準相關分析은 多變量 統計分析 技法중 가장 一般化된 技法으로 그 目的들을 자세히 열거하면 아래와 같다.

i) 各 變數集團 內에서 變數들 사이에 가중값을 決定하여 그 가중값들을 該當 變數들에 곱한 후 더하여 선형결합(linear combination) 을 만든다. 즉 變數集團 內의 變數들을 綜合하여 各 變數集團을 한 개의 값으로 表現한다. 이 선형결합에서 나온 값들 사이의 相關係數가 최대가 되도록 선형결합의 가중값을 구한다. 이렇게 만들어진 선형결합을 正準變量 (canonical variates)이라 한다.

ii) 두 개의 變數集團이 서로 獨立的인지 또는 두 集團 사이에 어느 정도의 相關關係가 있는지를 把握한다. 즉 正準變量인 두 선형결합 사이의 相關係數를 求한다.

iii) 앞의 正準變量들에서 表現되지 못하고 남은 相關關係를 最大化하는 加重값을 求하여 두 번째의 선형결합을 求한다. 이때 正準變量의 數는 두 變數集團중 變數가 작은 쪽의 變數의 個數를 넘지 못한다. 一般的으로 첫번째 또는 두번째 正準變量까지만 意味를 賦與할 수 있고, 그 이상은 相關關係가 극히 작거나 說明하기 困難한 경우가 大部分이다.

iv) 두 變數集團 사이의 相關關係 또는 相對的 寄與度를 說明한다.

앞에서도 말했듯이 正準相關分析은 한개 이상의 變數集團들 사이의 複合相關關係를 다루므로 이 技法을 使用하여 두 變數集團 사이의 相關關

係를 最大化시키는 일련의 獨立的인 正準變量들을 導出할 수 있다.

3. 正準相關分析에 의한 結果

正準分析에 의한 가장 重要한 세가지 結果는 다음과 같다.

i) 正準變量(canonical variates)

ii) 變量들 사이의 正準相關關係 λ_c (canonical correlation)

iii) 正準相關關係의 統計的 留意程度

각 正準變量은 한 쌍의 變量으로 構成된다. 分析에 適用되는 各 變數集團에 대하여 한 개씩의 變量이 適用된다.

즉 各 正準變量은 두 개의 變量으로 構成되며, 그 하나는 獨立變數 集團을 나타내고 다른 하나는 從屬變數 集團을 나타낸다. 正準變量은 正準構造값(canonical structure)이라 부른다. 相關係數 集團에 의해 說明된다. 이 相關係數는 正準變量에 包含된 각 變數들의 重要性을 反映한다. 즉 어떤 變數의 係數가 크면 클수록 그 變數는 正準變量을 導出하는데 더욱 더 重要한 役割을 한다. 正準相關分析에서 얻을 수 있는 다른 두가지 情報는 正準相關係數와 正準相關係數의 統計的 유의성이다.

두 變數集團의 關係가 얼마나 깊은지는 正準相關係數에 의하여 反映된다. 또한, 正準相關係數의 제곱(λ_c^2)은 正準變量중 어느 한 變量이 다른 變量에 의하여 發生되는 分散을 말한다. 正準相關係數의 제곱을 正準고유근(canonical roots or eigen values)이라 부른다. 모든 相關係數와 마찬가지로 λ_c의 유의성을 檢定하기 위하여 여러가지 統計 方法이 使用될 수 있다.

한 Ra 統計量을 利用하면 歸無假說을 檢定할 수 있다.

$$Ra=\frac{1-\Lambda_1^{\frac{1}{s}}}{\Lambda_1^{\frac{1}{s}}}\cdot\frac{1+t\cdot s-\frac{k_1\cdot k_2}{2}}{k_1\cdot k_2}\sim F\left(k_1.k_2,\ 1+ts-\frac{k_1.k_2}{2}\right)\cdots(3-7)$$

여기서,

$$S=\sqrt{\frac{k_1^2\ k_2^2-4}{k_1^2+k_2^2-5}},\qquad t=n-1-\frac{(k_1+k_2+1)}{2}$$ 이다.

만약 Ra 統計量에 의한 檢定에서 Ho이 기각되면 (Ha : $\Sigma_{12}\neq o$) 두 變數集團間에 유의한 正準相關係數가 存在함을 의미한다. 歸無假說이 기각되면 다음 檢定을 다시 하는데 이 檢定의 歸無假說은

Ho : $\lambda_2=\lambda_3=\cdots\cdots\lambda_{k^*}=o$

이다. 檢定統計量은

$$\Lambda_2=\frac{\Lambda_1}{1-\hat{\lambda}_1^2}=\prod_{j=2}^{k^*}(1-\hat{\lambda}_j^2)$$

이고 Ra 統計量

$$Ra=\frac{1-\Lambda_2^{\frac{1}{s}}}{\Lambda_2^{\frac{1}{s}}}\cdot\frac{1+t\cdot s-\frac{(k_1-1)(k_2-1)}{2}}{(k_1-1)(k_2-1)}\sim F\{(k_1-1)(k_2-1),\ 1+t.s-\frac{(k_1-1)(k_2-1)}{2}\}$$

$$\cdots\cdots(3-8)$$

을 利用하여 歸無假說을 檢定한다. 만약 檢定에 의해 歸無假說을 (Ho : $\lambda_2=\cdots\lambda_{k^*}=o$) 採擇하면, 이는 正準相關 λ_1만 유의하다는 것을 나타내고 Ho를 기각하면 λ_1 외에 다른 正準相關係數가 存在함을 나타내므로, 또 다른 歸無假說 Ho : $\lambda_3=\lambda_4=\cdots=\lambda_{k^*}$을 Ra 統計量에 의해 檢定한다. 이런 節次를 歸無假說이 採擇될때까지 계속한다.

5. SAS PROC CANCORR을 利用한 例題

몸무게 (weight), 허리둘레 (waist)와 맥박수 (pulse)의 變數集團을 生理的 條件을 나타내는 變數들이라 하고 1일 턱걸이回數 (chins), 팔굽혀펴기 (situps)와 Jump 들로 이루어진 變數集團을 체력 단련 (exercise)을 나타내는 變數들이라 하자. 이 두 集團의 變數間에 연관성을 調査하기 위해 20名의 高等學生들을 對象으로 資料를 蒐集하여, SAS PROC CANCORR에 의해 正準分析을 하였다.

分析에 必要한 獨立變數群과 從屬變數群은 다음과 같이 定하였다.

獨立變數群 = (턱걸이回數 = chins, 팔굽혀펴기回數 = situps , jump回數 = Jumps)

從屬變數群 : (몸무게 = weights , 허리둘레 = waist , 맥박수 = pulse)

〈 **INPUT** 〉

```
DATA FIT;
  INPUT WEIGHT WAIST PULSE CHINS SITUPS JUMPS;
  CARDS;
  191 36 50  5 162  60
  189 37 52  2 110  60
  193 38 58 12 101 101
  162 35 62 12 105  37
  189 35 46 13 155  58
  182 36 56  4 101  42
  211 38 56  8 101  38
  167 34 60  6 125  40
  176 31 74 15 200  40
  154 33 56 17 251 250
  169 34 50 17 120  38
  166 33 52 13 210 115
  154 34 64 14 215 105
  247 46 50  1  50  50
  193 36 46  6  70  31
  202 37 62 12 210 120
  176 37 54  4  60  25
  157 32 52 11 230  80
  156 33 54 15 225  73
  138 33 68  2 110  43
  ;
PROC CANCORR DATA=FIT ALL
  VPREFIX=PHYS VNAME='PHYSIOLOGICAL MEASUREMENTS'
  WPREFIX=EXER WNAME='EXERCISES';
  VAR WEIGHT WAIST PULSE;
  WITH CHINS SITUPS JUMPS;
  TITLE MIDDLE-AGE MEN IN A HEALTH FITNESS CLUB;
  TITLE2 DATA COURTESY OF DR. A. C. LINNERUD, NC STATE UNIV;
```

〈OUTPUT〉

MIDDLE-AGE MEN IN A HEALTH FITNESS CLUB
DATA COURTESY OF DR. A. C. LINNERUD, NC STATE UNIV

CANONICAL CORRELATION ANALYSIS

20 OBSERVATIONS
3 PHYSIOLOGICAL MEASUREMENTS
3 EXERCISES

(continued from previous page)

(1) SIMPLE STATISTICS

VARIABLE	MEAN	ST. DEV.	SKEWNESS	KURTOSIS
WEIGHT	178.60000000	24.69055313	0.9698740166	1.8023462545
WAIST	35.40000000	3.2019730760	1.8721345109	5.6620990210
PULSE	56.10000000	7.2103726450	0.8460998408	0.6069130270
CHINS	9.45000000	5.2862701650	-.1930287524	-1.4135209790
SITUPS	145.55000000	62.5665750680	0.2236427744	-1.3291393440
JUMPS	70.30000000	51.2774701730	2.4799104223	7.6234923710

(2) CORRELATIONS AMONG THE PHYSIOLOGICAL MEASUREMENTS

	WEIGHT	WAIST	PULSE
WEIGHT	1.0000	0.8702	-.3658
WAIST	0.8702	1.0000	-.3529
PULSE	-.3658	-.3529	1.0000

CORRELATIONS AMONG THE EXERCISES

	CHINS	SITUPS	JUMPS
CHINS	1.0000	0.6957	0.4958
SITUPS	0.6957	1.0000	0.6692
JUMPS	0.4958	0.6692	1.0000

CORRELATIONS BETWEEN THE PHYSIOLOGICAL MEASUREMENTS AND THE EXERCISES

	CHINS	SITUPS	JUMPS
WEIGHT	-.3897	-.4931	-.2263
WAIST	-.5522	-.6456	-.1915
PULSE	0.1506	0.2250	0.0349

CANONICAL CORRELATION ANALYSIS

(3) (4) (5) (6) (7) (8) (9) (10)

CANONICAL CORRELATIONS AND TESTS OF HO: THE CANONICAL CORRELATION IN THE CURRENT ROW AND ALL THAT FOLLOW ARE ZERO

	CANONICAL CORRELATION	ADJUSTED CAN CORR	APPROX STD ERROR	VARIANCE RATIO	CANONICAL R-SQUARED	LIKELIHOOD RATIO	F STATISTIC	NUM DF	DEN DF	PROB>F
1	0.795608154	0.705498736	0.084197333	1.7247	0.632992335	0.350390533	2.0482	9	34.223	0.0635
2	0.200556041	-0.580144410	0.220188008	0.0419	0.040222726	0.954722659	0.1758	4	30	0.9491
3	0.072570286	-1.261851594	0.228207528	0.0053	0.005266446	0.994733554	0.0847	1	16	0.7748

MULTIVARIATE TEST STATISTICS AND F APPROXIMATIONS

STATISTIC	VALUE	F	NUM DF	DEN DF	PROB>F
(11) WILKS' LAMBDA	0.3503905	2.048234	9	34.22293	0.06353094
PILLAI'S TRACE	0.6784815	1.556707	9	48	0.1551082
HOTELLING-LAWLEY TRACE	1.771941	2.493848	9	38	0.02384017
ROY'S GREATEST ROOT	1.724739	9.198607	3	16	0.0009016772

NOTE: F STATISTIC FOR ROY'S GREATEST ROOT IS AN UPPER BOUND

RAW CANONICAL COEFFICIENTS FOR THE PHYSIOLOGICAL MEASUREMENTS

	PHYS1	PHYS2	PHYS3
WEIGHT	-.0314046879	+.0763195063	-.0077350867
WAIST	0.4932416756	0.3687229894	0.1580336471
PULSE	-.0081993154	-.0320519942	0.1457322421

RAW CANONICAL COEFFICIENTS FOR THE EXERCISES

(12)

	EXER1	EXER2	EXER3
CHINS	-.0661139864	-.0710412111	-.2452753473
SITUPS	-.0168462308	0.0019737454	0.0197676373
JUMPS	0.0139715689	0.0207141063	-.0081674724

STANDARDIZED CANONICAL COEFFICIENTS FOR THE PHYSIOLOGICAL MEASUREMENTS

	PHYS1	PHYS2	PHYS3
WEIGHT	-0.7754	-1.8844	-0.1910
WAIST	1.5793	1.1806	0.5060
PULSE	-0.0591	-0.2311	1.0508

MIDDLE-AGE MEN IN A HEALTH FITNESS CLUB
DATA COURTESY OF DR. A. C. LINNERUD, NC STATE UNIV

CANONICAL CORRELATION ANALYSIS

(12) STANDARDIZED CANONICAL COEFFICIENTS FOR THE EXERCISES

	EXER1	EXER2	EXER3
CHINS	-0.3495	-0.3755	-1.2016
SITUPS	-1.0540	0.1235	1.2368
JUMPS	0.7156	1.0622	-0.4185

MIDDLE-AGE MEN IN A HEALTH FITNESS CLUB
DATA COURTESY OF DR. A. C. LINNERUD, NC STATE UNIV

(13) CANONICAL STRUCTURE

CORRELATIONS BETWEEN THE PHYSIOLOGICAL MEASUREMENTS AND THEIR CANONICAL VARIABLES

	PHYS1	PHYS2	PHYS3
WEIGHT	0.6206	-0.7728	-0.1350
WAIST	0.9254	-0.3777	-0.0310
PULSE	-0.3328	0.0415	0.9421

CORRELATIONS BETWEEN THE EXERCISES AND THEIR CANONICAL VARIABLES

	EXER1	EXER2	EXER3
CHINS	-0.7276	0.2370	-0.6438
SITUPS	-0.8177	0.5730	0.0544
JUMPS	-0.1622	0.9586	-0.2339

CORRELATIONS BETWEEN THE PHYSIOLOGICAL MEASUREMENTS AND THE CANONICAL VARIABLES OF THE EXERCISES

	EXER1	EXER2	EXER3
WEIGHT	0.4938	-0.1589	-0.0098
WAIST	0.7363	-0.0757	-0.0022
PULSE	-0.2648	0.0083	0.0669

CORRELATIONS BETWEEN THE EXERCISES AND THE CANONICAL VARIABLES OF THE PHYSIOLOGICAL MEASUREMENTS

	PHYS1	PHYS2	PHYS3
CHINS	-0.5789	0.0475	-0.0467
SITUPS	-0.6506	0.1189	0.0080
JUMPS	-0.1290	0.1923	-0.0170

MIDDLE-AGE MEN IN A HEALTH FITNESS CLUB
DATA COURTESY OF DR. A. C. LINNERUD, NC STATE UNIV

(14) CANONICAL REDUNDANCY ANALYSIS

RAW VARIANCE OF THE PHYSIOLOGICAL MEASUREMENTS EXPLAINED BY

	THEIR OWN CANONICAL VARIABLES			THE OPPOSITE CANONICAL VARIABLES	
	PROPORTION	CUMULATIVE PROPORTION	CANONICAL R-SQUARED	PROPORTION	CUMULATIVE PROPORTION
1	0.3712	0.3712	0.6330	0.2349	0.2349
2	0.5436	0.9148	0.0402	0.0219	0.2568
3	0.0852	1.0000	0.0053	0.0004	0.2573

RAW VARIANCE OF THE EXERCISES EXPLAINED BY

	THEIR OWN CANONICAL VARIABLES			THE OPPOSITE CANONICAL VARIABLES	
	PROPORTION	CUMULATIVE PROPORTION	CANONICAL R-SQUARED	PROPORTION	CUMULATIVE PROPORTION
1	0.4111	0.4111	0.6330	0.2602	0.2602
2	0.5635	0.9746	0.0402	0.0227	0.2829
3	0.0254	1.0000	0.0053	0.0001	0.2830

(continued

PART C. 判別分析 (Discriminant Analysis)

1. 判別分析의 定義

判別分析이란 2個 以上의 그룹 또는 母集團으로부터 얻은 n개 케이스에 對한 資料를 利用하여, 個個의 케이스가 속해있는 그룹을 쉽게 判別해 줄 수 있는 特性變數들의 선형함수식을 導出하는 方法이다. 이때, 各 케이스의 特性을 나타내는 變數를 判別變數 (discriminant variable)라 부르고, 이들의 函數式을 判別函數 (discriminant function)라 한다. 一般的으로, 各 케이스의 特性을 나타내는 判別變數는 p個인 多變數이므로 이들을 가지고 케이스가 속한 그룹을 判斷하기란 어려움이 있다.

判別分析은 通常 1個 또는 2個의 判別基準을 分析者에게 提示해 주는 방법으로, 이 判別基準은 判別函數를 통해 이루어 지도록 한 技法이다. 判別函數의 役割을 例로 알아보자.

다음 그림에 나타난 對象들을 두 개의 그룹으로 判別하고자할때 變數 X_1에 의해서만 判別한다면 약 40% 程度만 區分이 可能하고, 變數 X_2에 의해서만 判別한다면 約 50% 程度 區分할 수 있을것이다. 그러나, X_1과 X_2를 선형결합한 判別函數 Y를 利用하면 거의 완벽하게 區分이 可能함을 보여준다. 세 그룹判別分析에서의 判別函數가 가진 役割도 같으며 그림 (1－2)로 說明할 수 있다.

일단 判別函數가 얻어지면 所屬이 不分明한 케이스도 이 基準에 의해 그룹의 所屬이 정해지므로, 判別分析은 患者의 診斷, 犯罪의 審査, 會社의 信用度 審査 等의 分類 (classification)의 目的에 많이 使用되는 多變量 統計技法이다.

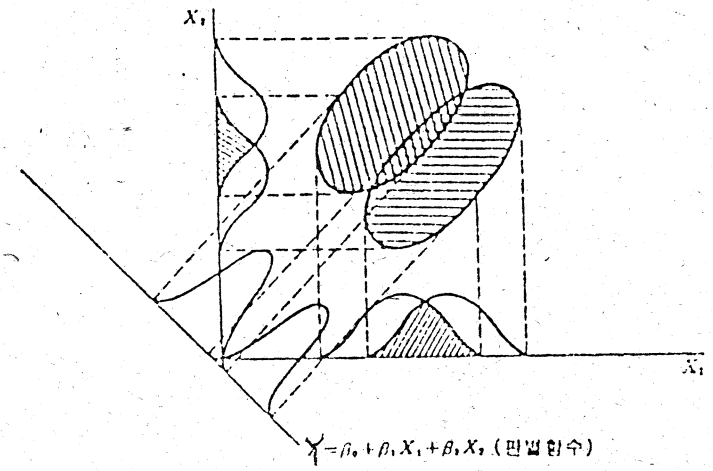


그림 (1 － 1) 두 그룹 判別分析의 例

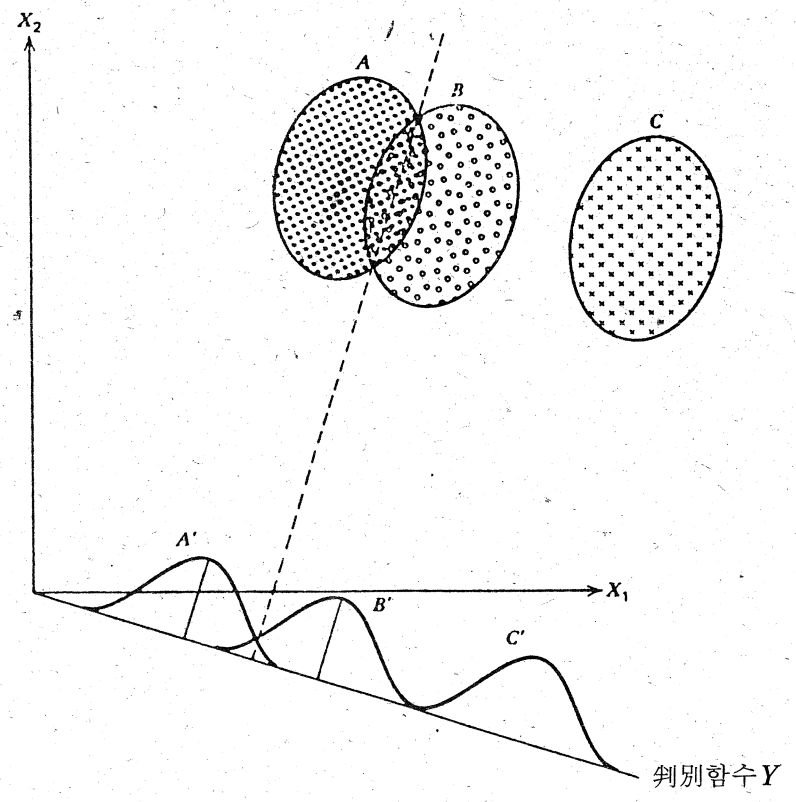


그림 (1 － 2) 세 그룹 判別分析의 例

2. 判別分析의 理論

判別分析은 所屬그룹이 이미 알려진 케이스들의 特性을 나타내는 p개 要因들인 判別變數(discriminant variable) $X_1, X_2, \cdots, X_p$를 各 케이스別로 推定하여, 判別誤謬를 最小化 시킬 수 있는 判別函數(discriminant function) $Y_1, Y_2, \cdots, Y_m$을 推定하는 것이 主目的이다.

判別分析下에서의 判別函數 $Y_1, Y_2, \cdots, Y_m$들은 다음 性質을 가진다.

(判別函數의 性質)

① 모든 判別函數는 判別變數 $X_1, X_2, \cdots, X_p$의 선형결합으로 이루어진다.

i번째 判別函數 : $Y_i = d_{i1}X_1 + d_{i2}X_2 + \cdots + d_{ip}X_p$, $i = 1, 2, \cdots, m$ ..... (1-1)

단, $d_i = (d_{i1}, d_{i2}, \cdots, d_{ip})$는 i번째 判別函數計數(discriminant function coefficient)라 한다.

② 導出可能한 判別函數의 갯수 m은 그룹의 수를 K, 判別變數의 數를 P라 할때 $m = Min(K-1, P)$이다. 그러나 最終的인 判別分析에서는 m個 判別函數의 判別力을 檢定하여, 이들 중에서 判別力이 높은 몇개의 判別函數만을 使用한다. (通常, $K > 2$일 때 두개의 判別函數를 使用)

③ 判別函數는 그룹들이 同一한 공분산 행렬을 가진 多變量 定規分布를 한다는 假定下에서 이루어 진다.

④ 判別函數 推定法은 크게 Fisher의 선형판별함수법과 정준분석을 利用하는 方法, Bayes의 理論을 導入한 意思決定的 方法이 있다.

이 PART에서는 正準分析을 利用한 判別分析法인 正準判別分析法(Canonical Discriminant Analysis)에 대해서 說明하기로 한다.

3. 正準分析을 利用한 判別分析

判別函數 $Y_1, Y_2, \cdots, Ym$을 PART B에서 說明한 正準分析技法으로 推定하는 方法으로, 이 分析技法의 節次는 다음과 같다.

節次 1. 正準分析에 使用되는 獨立變數群과 從屬變數群을 다음과 같이 設定한다.

- 獨立變數群 : 判別變數 $X_1, X_2, \cdots, X_p$를 獨立變數群으로 한다.
- 從屬變數群 : K개 各 그룹의 所屬을 나타낼 수 있는 $K-1$개의 假變數 (dummy variable) $D_1, D_2, \cdots, D_{k-1}$를 從屬變數群으로 使用한다.

從屬變數群의 各 케이스別 觀測값은 다음과 같은 Coding에 의해 얻는다.

(表 1 − 1) 假變數 coding의 例

假變數 / 假變數觀測값	D_1	D_2	D_3 ……	D_{k-1}
첫번째 그룹의 케이스	1	0	0 ……	0
둘 째 그룹의 케이스	0	1	0 ……	0
⋮				
$K-1$번째 그룹의 케이스	0	0	0 ……	1
K번째 그룹의 케이스	0	0	0 ……	0

節次 2. 各 케이스에 該當하는 獨立變數群과 從屬變數群의 관측값으로 正準分析을 하세 m개 判別函數를 구한다. 여기서 m개 判別函數란 正準分析의 正準函數 推定式에서 獨立變數群에 該當된 正準變數 推定式인

$$(\hat{\Sigma}_{XX}^{-1}\ \hat{\Sigma}_{XD}\ \hat{\Sigma}_{DD}^{-1}\ \hat{\Sigma}_{DX} - \lambda^2 I\ )\underline{\gamma} = 0 \qquad (1-2)$$

이다. 단, $\hat{\Sigma}_{XX}$ 은 獨立變數群의 標本 공분산행렬이다. 正準相關係數가 큰 값 $\lambda_1 \geq \lambda_2 \geq \cdots\cdots \geq \lambda_m$을 가진것을 選擇하여, 이에 對應되는 m개 獨立變數群의 正準函數를 判別函數로 使用한다. 여기서 正準相關係數 λ_i의 意味는 判別函數의 判別力을 나타낸다. 이 過程을 例로 들면 다음과 같다. 아래表는 4個의 그룹으로 부터 얻은 資料를 가지고 正準判別分析한 結果를 나타낸 것이다. 여기서, D_1, D_2, D_3는 그룹을 나타내는 假變數이고, X_1, X_2, X_3, X_4, X_5는 各 케이스의 特性을 나타내는 要因들이며, 獨立變數群으로 使用되었다.

(表 1－2) 正準判別分析의 結果表

	D_1	D_2	D_3	X_1	X_2	X_3	X_4	X_5	Canonical Correlation	Eigenvalue
Correlation Matrix										
D_1	1.00	−.33	−.33	.18	−.06	.20	.11	.08		
D_2	−.33	1.00	−.33	−.15	.00	.07	.01	.02		
D_3	−.33	−.33	1.00	−.16	.07	−.22	−.14	−.02		
A_1	.18	−.15	−.16	1.00	.15	.35	.36	.22		
A_2	−.06	.01	.07	.15	1.00	.39	.40	.40		
A_3	.20	.07	−.22	.35	.39	1.00	.49	.41		
A_4	.11	.01	−.14	.36	.40	.49	1.00	.47		
A_5	.08	.02	−.02	.22	.40	.41	.47	1.00		
Canonical Weights										
Variate 1	.50	−.11	−.72	.42	−.62	.73	.22	−.09		
Variate 2	.59	1.22	.49	−.94	−.24	.72	−.02	.29		
Variate 3	−.95	−.09	−.87	−.31	.11	.18	.63	−1.11		
Canonical Correlations										
Variate 1									.35	
Variate 2									.23	
Variate 3									.08	
Eigenvalue										
1										.12
2										.05
3										.01

위 表에서 얻어진 3개의 判別函數는 다음과 같이 表示된다.

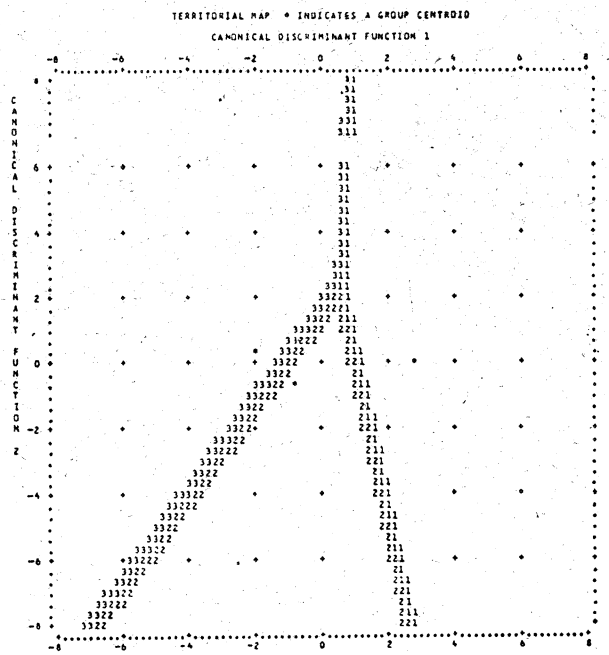
$$Y_1 = 0.42X_1 - 0.62X_2 + 0.73X_3 + 0.22X_4 - 0.09X_5$$

$$Y_2 = -0.94X_1 - 0.24X_2 + 0.72X_3 - 0.02X_4 + 0.29X_5 \quad (1-3)$$

$$Y_3 = -0.31X_1 - 0.11X_2 + 0.18X_3 + 0.63X_4 - 1.11X_5$$

그러나, 正準相關係數(canonical correlation) 값 $\lambda_1 = 0.35$, $\lambda_2 = 0.23$, $\lambda_3 = 0.08$을 比較하면 判別函數 Y_1과 Y_2의 判別力이 유의하므로, 이 두 개의 判別函數만 使用하여 分析을 한다.

節次 3. 標本으로 使用된 各 케이스들의 判別變數값을 推定된 判別函數에 대입시켜 判別點數(discriminant score)를 구한후 判別點數들의 산점도(PLOT)와 各 그룹별 平均判別點數(grouped centroid score)를 구하여 아래 그림과 같이 判別領域(discriminant region)을 導出한다. 아래 그림은 위 例에서 導出된 判別領域을 나타낸다. 여기서 X축은 Y_1의 判別點數를 나타내고 Y축은 Y_2의 判別點數를 나타낸다.



그림(1-2) 判別領域의 例

'•'表示는 各 그룹의 平均 判別點數를 나타낸다.

4. SAS의 PROC CANDISC를 利用한 例題

SAS의 PROC CANDISC는 正準判別分析을 위한 Package이며, 이것을 利用하여 다음의 例를 分析하였다.

〔例題〕 Fisher의 붓꽃(iris) 資料를 利用하여 正準判別分析을 하려고 한다.

이 資料는 세 種類의 붓꽃인 SETOSA, VERSICOLOR, VIRGINICA를 各各 50個씩 標本으로 採集하여 各各의 꽃받침(sepal)의 길이와 폭, 꽃잎의 길이와 폭을 관측한 資料이다.

正準判別分析을 하기 위해서 다음과 같이 判別變數와 從屬變數를 區分하였다.

- 從屬變數 : 假變數 D_1, D_2
- 獨立變數 : X_1 = SEPALLEN (꽃받침 길이)
  X_2 = SEPALWID (꽃받침 폭)
  X_3 = PETALLEN (꽃잎 길이)
  X_4 = PETALWID (꽃잎 폭)

SAS의 境遇 假變數 入力方法은

Group 1 : SETOSA = 1

Group 2 : VERSICOLOR = 2

Group 3 : VERSINICA = 3

으로 단순히 그룹의 番號만 入力하면 되도록 簡便化되어 있다.

〈 **INPUT** 〉

```
DATA IRIS;
  TITLE FISHER (1936) IRIS DATA;
  INPUT SEPALLEN SEPALWID PETALLEN PETALWID SPEC__NO@@;
  IF SPEC__NO=1 THEN SPECIES='SETOSA     ';
  IF SPEC__NO=2 THEN SPECIES='VERSICOLOR';
  IF SPEC__NO=3 THEN SPECIES='VIRGINICA  ';
  LABEL SEPALLEN=SEPAL LENGTH IN MM.
    SEPALWID=SEPAL WIDTH IN MM.
    PETALLEN=PETAL LENGTH IN MM.
    PETALWID=PETAL WIDTH IN MM.;
  CARDS;
50 33 14 02 1 64 28 56 22 3 65 28 46 15 2
67 31 56 24 3 63 28 51 15 3 46 34 14 03 1
69 31 51 23 3 62 22 45 15 2 59 32 48 18 2
46 36 10 02 1 61 30 46 14 2 60 27 51 16 2
65 30 52 20 3 56 25 39 11 2 65 30 55 18 3
58 27 51 19 3 68 32 59 23 3 51 33 17 05 1
57 28 45 13 2 62 34 54 23 3 77 38 67 22 3
63 33 47 16 2 67 33 57 25 3 76 30 66 21 3
49 25 45 17 3 55 35 13 02 1 67 30 52 23 3
70 32 47 14 2 64 32 45 15 2 61 28 40 13 2
48 31 16 02 1 59 30 51 18 3 55 24 38 11 2
63 25 50 19 3 64 32 53 23 3 52 34 14 02 1
49 36 14 01 1 54 30 45 15 2 79 38 64 20 3
44 32 13 02 1 67 33 57 21 3 50 35 16 06 1
58 26 40 12 2 44 30 13 02 1 77 28 67 20 3
63 27 49 18 3 47 32 16 02 1 55 26 44 12 2
50 23 33 10 2 72 32 60 18 3 48 30 14 03 1
51 38 16 02 1 61 30 49 18 3 48 34 19 02 1
50 30 16 02 1 50 32 12 02 1 61 26 56 14 3
64 28 56 21 3 43 30 11 01 1 58 40 12 02 1
51 38 19 04 1 67 31 44 14 2 62 28 48 18 3
49 30 14 02 1 51 35 14 02 1 56 30 45 15 2
58 27 41 10 2 50 34 16 04 1 46 32 14 02 1
60 29 45 15 2 57 26 35 10 2 57 44 15 04 1
50 36 14 02 1 77 30 61 23 3 63 34 56 24 3
58 27 51 19 3 57 29 42 13 2 72 30 58 16 3
54 34 15 04 1 52 41 15 01 1 71 30 59 21 3
64 31 55 18 3 60 30 48 18 3 63 29 56 18 3
49 24 33 10 2 56 27 42 13 2 57 30 42 12 2
55 42 14 02 1 49 31 15 02 1 77 26 69 23 3
60 22 50 15 3 54 39 17 04 1 66 29 46 13 2
52 27 39 14 2 60 34 45 16 2 50 34 15 02 1
44 29 14 02 1 50 20 35 10 2 55 24 37 10 2
58 27 39 12 2 47 32 13 02 1 46 31 15 02 1
69 32 57 23 3 62 29 43 13 2 74 28 61 19 3
59 30 42 15 2 51 34 15 02 1 50 35 13 03 1
56 28 49 20 3 60 22 40 10 2 73 29 63 18 3
67 25 58 18 3 49 31 15 01 1 67 31 47 15 2
63 23 44 13 2 54 37 15 02 1 56 30 41 13 2
63 25 49 15 2 61 28 47 12 2 64 29 43 13 2
51 25 30 11 2 57 28 41 13 2 65 30 58 22 3
69 31 54 21 3 54 39 13 04 1 51 35 14 03 1
72 36 61 25 3 65 32 51 20 3 61 29 47 14 2
56 29 36 13 2 69 31 49 15 2 64 27 53 19 3
68 30 55 21 3 55 25 40 13 2 48 34 16 02 1
48 30 14 01 1 45 23 13 03 1 57 25 50 20 3
57 38 17 03 1 51 38 15 03 1 55 23 40 13 2
66 30 44 14 2 68 28 48 14 2 54 34 17 02 1
51 37 15 04 1 52 35 15 02 1 58 28 51 24 3
67 30 50 17 2 63 33 60 25 3 53 37 15 02 1
;
PROC CANDISC ALL OUT=DISC;
  CLASSES SPECIES;
  VAR SEPALLEN SEPALWID PETALLEN PETALWID;
PROC PLOT;
  PLOT CAN2*CAN1=SPEC__NO;
  TITLE2 PLOT OF CANONICAL DISCRIMINANT FUNCTIONS;
```

〈OUTPUT〉

FISHER (1936) IRIS DATA

CANONICAL DISCRIMINANT ANALYSIS

150 OBSERVATIONS — 149 DF TOTAL
4 VARIABLES — 147 DF WITHIN CLASSES
3 CLASSES — 2 DF BETWEEN CLASSES

SPECIES	FREQUENCY	WEIGHT	PROPORTION
SETOSA	50	50	0.333333
VERSICOLOR	50	50	0.333333
VIRGINICA	50	50	0.333333

① ② ③ UNIVARIATE STATISTICS ④ ⑤ ⑥ ⑦

	MEAN	TOTAL STD	WITHIN STD	BETWEEN STD	R-SQUARED	VAR RATIO	F	PROB GT F
SEPALLEN	58.433333333	8.280661279[illegible]	5.147894[illegible]	8.[illegible]	.618709[illegible]	1.[illegible]	[illegible]	[illegible]
SEPALWID	30.573333333	4.358662849	3.396770[illegible]	4.[illegible]	.[illegible]	[illegible]	[illegible]	[illegible]
PETALLEN	37.580000000	17.652478402	3.[illegible]	[illegible]	.[illegible]	[illegible]	[illegible]	[illegible]
PETALWID	11.993333333	7.622238342	3.[illegible]	[illegible]	.[illegible]	[illegible]	[illegible]	[illegible]

AVERAGE R-SQUARED: UNWEIGHTED = 0.7226358 WEIGHTED BY VARIANCE = 0.8629484

⑧ CLASS MEANS

SPECIES	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SETOSA	50.060000000	34.280000000	14.620000000	2.460000000
VERSICOLOR	59.360000000	27.700000000	42.600000000	13.260000000
VIRGINICA	65.880000000	29.740000000	55.520000000	20.260000000

⑨ TOTAL-STANDARDIZED CLASS MEANS

SPECIES	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SETOSA	-1.0112	0.8504	-1.3006	-1.2507
VERSICOLOR	0.1119	-0.6592	0.2844	0.1662
VIRGINICA	0.8993	-0.1912	1.0163	1.0845

⑩ WITHIN-STANDARDIZED CLASS MEANS

SPECIES	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SETOSA	-1.6266	1.0912	-5.3354	-4.6584
VERSICOLOR	0.1800	-0.8459	1.1665	0.6139
VIRGINICA	1.8465	-0.2453	4.1689	4.0354

FISHER (1936) IRIS DATA

CANONICAL DISCRIMINANT ANALYSIS

⑪ TOTAL SAMPLE COVARIANCE MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	68.56935123	-4.24340045	127.[illegible]9362	51.62706935
SEPALWID	-4.24340045	18.59794183	-32.56563758	-12.16393736
PETALLEN	127.43154362	-32.56563758	311.62775523	129.56093960
PETALWID	51.62706935	-12.16393736	129.56093960	58.10062640

⑫ BETWEEN-CLASS COVARIANCE MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	62.[illegible]3056	-[illegible].39105145	110.90[illegible]644	47.03847875
SEPALWID	-13.39105145	7.61404922	-38.41583893	-15.39105145
PETALLEN	110.50896644	-38.41583893	293.35758389	125.35167785
PETALWID	47.03847875	-15.39105145	125.35167785	53.56868009

⑬ POOLED WITHIN-CLASS COVARIANCE MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	26.500816327	9.272108844	16.751428571	3.840136054
SEPALWID	9.272108844	11.538775510	5.524353741	3.271020408
PETALLEN	16.751428571	5.524353741	18.518775510	4.266530612
PETALWID	3.840136054	3.271020408	4.266530612	4.188163265

⑭ TOTAL SAMPLE CORRELATION MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.0000	-0.1176	0.8718	0.8179
SEPALWID	-0.1176	1.0000	-0.4284	-0.3661
PETALLEN	0.8718	-0.4284	1.0000	0.9629
PETALWID	0.8179	-0.3661	0.9629	1.0000

⑮ BETWEEN-CLASS CORRELATION MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.0000	-0.7451	0.9941	0.9998
SEPALWID	-0.7451	1.0000	-0.8128	-0.7593
PETALLEN	0.9941	-0.8128	1.0000	0.9962
PETALWID	0.9998	-0.7593	0.9962	1.0000

FISHER (1936) IRIS DATA

CANONICAL DISCRIMINANT ANALYSIS

⑯ POOLED WITHIN-CLASS CORRELATION MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.0000	0.5302	0.7562	0.3645
SEPALWID	0.5302	1.0000	0.3779	0.4705
PETALLEN	0.7562	0.3779	1.0000	0.4845
PETALWID	0.3645	0.4705	0.4845	1.0000

MAHALANOBIS DISTANCES BETWEEN CLASSES

SPECIES	SETOSA	VERSICOLOR	VIRGINICA
SETOSA			
VERSICOLOR	9.4797		13.3935
VIRGINICA	13.3935	4.1474	

⑰

PROB > MAHALANOBIS DISTANCE

SPECIES	SETOSA	VERSICOLOR	VIRGINICA
SETOSA			
VERSICOLOR	0.000000	0.000000	0.000000
VIRGINICA	0.000000	0.000000	0.000000

(continued on next page)

(18) (19) (20) (21) (22) (23) (24) (25)

CANONICAL CORRELATIONS AND TESTS OF HO: THE CANONICAL CORRELATION IN THE CURRENT ROW AND ALL THAT FOLLOW ARE ZERO

	CANONICAL CORRELATION	ADJUSTED CAN CORR	APPROX STD ERROR	VARIANCE RATIO	CANONICAL R-SQUARED	LIKELIHOOD RATIO	F STATISTIC	NUM DF	DEN DF	PROB>F
1	0.984820894	0.984097357	0.002468166	32.1919	0.969872194	0.023438631	199.1453	8	288	0.0000
2	0.471197019	0.439283495	0.063734062	0.2854	0.222026631	0.777973369	13.7939	3	145	0.0000

MULTIVARIATE TEST STATISTICS AND F APPROXIMATIONS

STATISTIC	VALUE	F	NUM DF	DEN DF	PROB>F
WILKS' LAMBDA	0.02343863	199.1453	8	288	0
(26) PILLAI'S TRACE	1.191899	53.46649	8	290	9.74216E-53
HOTELLING-LAWLEY TRACE	32.47732	580.5321	8	286	0
ROY'S GREATEST ROOT	32.19193	1166.957	4	145	0

NOTE: F STATISTIC FOR WILKS' LAMBDA IS EXACT
NOTE: F STATISTIC FOR ROY'S GREATEST ROOT IS AN UPPER BOUND

FISHER (1936) IRIS DATA

CANONICAL DISCRIMINANT ANALYSIS

(27) TOTAL CANONICAL STRUCTURE

	CAN1	CAN2	CAN3	CAN4
SEPALLEN	0.7919	0.2176	-0.4275	-0.3779
SEPALWID	-0.5308	0.7580	-0.3745	0.0590
PETALLEN	0.9850	0.0460	-0.1636	-0.0313
PETALWID	0.9728	0.2229	0.0428	-0.0460

(28) WITHIN CANONICAL STRUCTURE

	CAN1	CAN2	CAN3	CAN4
SEPALLEN	0.2226	0.3108	-0.6923	-0.6120
SEPALWID	-0.1190	0.8637	-0.4838	0.0763
PETALLEN	0.7061	0.1677	-0.6758	-0.1291
PETALWID	0.6332	0.7372	0.1606	-0.1725

(29) STANDARDIZED CANONICAL COEFFICIENTS

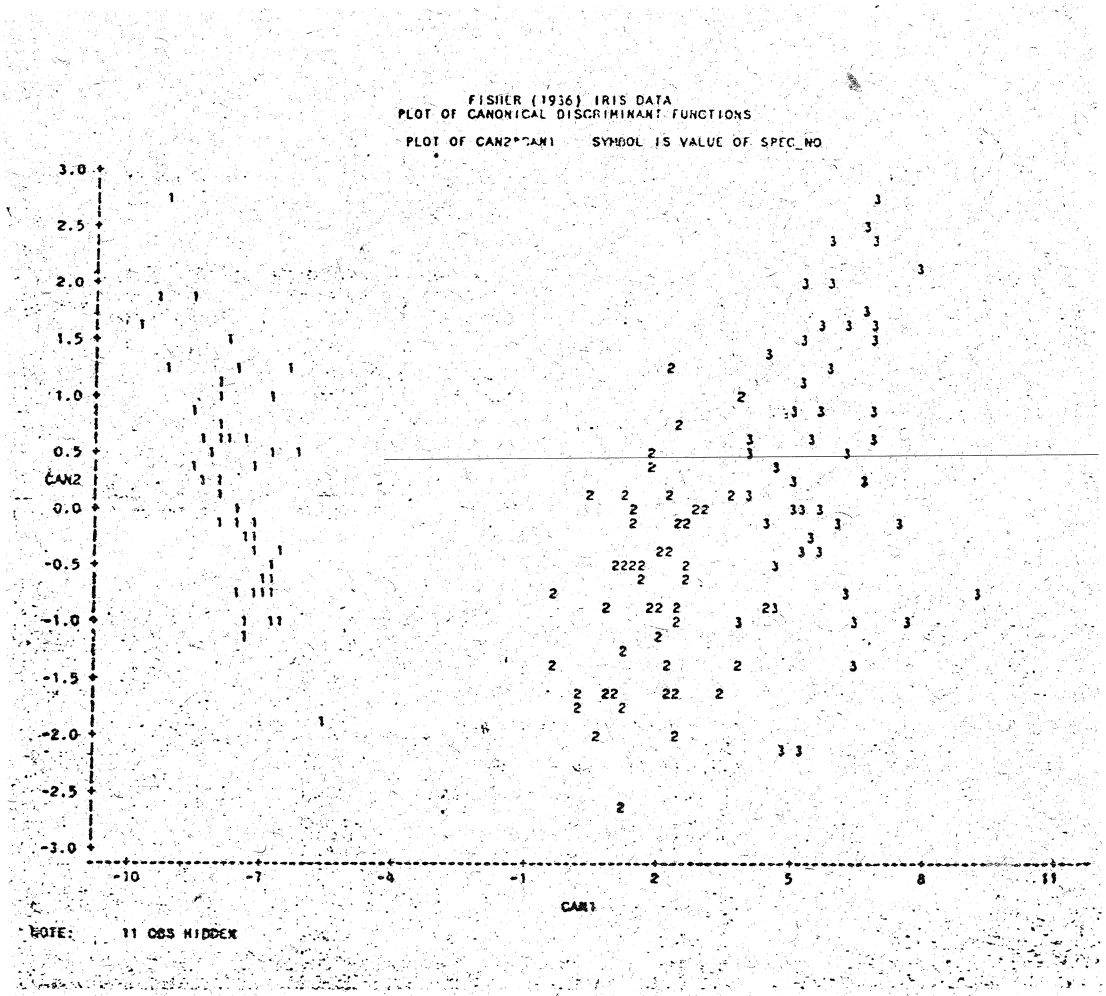
	CAN1	CAN2	CAN3	CAN4
SEPALLEN	-0.6868	0.0200	-0.2100	-2.6304
SEPALWID	-0.6688	0.9434	-0.6503	0.9715
PETALLEN	3.8858	-1.6451	-3.2263	4.2276
PETALWID	2.1422	2.1641	3.0827	-1.6091

(30) RAW CANONICAL COEFFICIENTS

	CAN1	CAN2	CAN3	CAN4
SEPALLEN	-.0829377642	0.0024102149	-.0253635563	-.3176592443
SEPALWID	-.1534473068	0.2164521235	-.1491965217	0.2228936770
PETALLEN	0.2201211656	-.0931921210	-.1827626273	0.2394822760
PETALWID	0.2810460309	0.2839187853	0.4044320622	-.2110965476

(31) CLASS MEANS ON CANONICAL VARIABLES

SPECIES	CAN1	CAN2	CAN3	CAN4
SETOSA	-7.6076	0.2151	0.0000	-0.0000
VERSICOLOR	1.8250	-0.7279	0.0000	-0.0000
VIRGINICA	5.7826	0.5128	0.0000	-0.0000


FISHER (1936) IRIS DATA
PLOT OF CANONICAL DISCRIMINANT FUNCTIONS
PLOT OF CAN2*CAN1 SYMBOL IS VALUE OF SPEC_NO
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
CAN2
-10
-7
-4
-1
2
5
8
11
CAN1
NOTE: 11 OBS HIDDEN

PART D. 主成分 分析 (Principal Components Analysis)

1. 主成分 分析의 目的

主成分 分析은 서로 相關關係가 있는 많은 變數들의 集合이 가진 복잡한 特性들을 簡單하게 說明해 줄 수 있는 구조를 가진 主成分 (principal components)들로 說明하기 위한 多變量 統計技法이다. 主成分이란 다음 열거된 性質을 가진 것이다.

① 主成分은 原變數의 선형결합 形態로써 서로 無相關의 性質을 가진다.

② 主成分 分析에서 얻을 수 있는 主成分의 갯수는 原變數의 갯수와 同一하다.

③ 各 主成分의 分散은 이것이 包含하고 있는 原 變數에 대한 情報의 量을 나타낸다.

④ 主成分들은 분산의 크기에 따라 命名된다. 즉, i번째 주성분 (i-th principal component)은 분산이 i번째 큰 主成分을 나타낸다.

主成分 分析의 目的은 多變量資料들이 가진 情報를 간단히 설명하는데 있다.

이 技法에서는 觀測變數들의 分散合을 資料가 가진 情報量合으로 看做하고 이것을 最大限 說明할 수 있는 몇개의 主成分을 scree test등으로 選定한다. 이렇게 導出된 主成分들은 資料가 가진 情報를 充分히 설명해줄 수 있는 새로운 變數가 되며, 觀測變數들의 數를 壓縮하는 效果를 나타낸다. 그러므로, 이 技法은 多共線性을 가진 회귀분석, 多變量 正規性 檢定,

特異값(outlier) 決定等에 많이 應用되고 있다.

2. 主成分 分析의 理論

공분산 행렬 Σ를 가진 p개 變數, $X=(x_1, x_2, \cdots, x_p)$의 첫번째 主成分은 다음과 같이 定義된다.

$$PC_{(1)} = W_{(1)1}X_1 + W_{(1)2}X_2 + W_{(1)3}X_3 + \cdots\cdots W_{(1)p}X_p = W_{(1)}'X \cdots (1-1)$$

이미 定義된 바와 같이 첫번째 主成分은 p개 變數가 가진 分散合을 가장 많이 反映시키는 것이어야 한다. 그러므로, 첫번째 成分의 計數들인 $W_{(1)1}, W_{(1)2}, \cdots, W_{(1)p}$은 다음 條件下에서 얻어진다.

$$\max \text{ Var } (W_{(1)}'X) \quad \cdots\cdots (1-2)$$

$$\text{subiect to } W_{(1)}'W_{(1)} = 1$$

여기서, 條件 $W_{(1)}'W_{(1)}=1$은 $W_{(1)}$의 유일한 解를 구하기 위한 것이다. 式(1-2)를 Lograngian multiplier로 풀면,

$$L(W_{(1)}) = W_{(1)}'\Sigma W_{(1)} - \lambda_1 (W_{(1)}'W_{(1)} - 1)$$

$$\frac{d\,L(W_{(1)})}{d\,W_{(1)}} = 0 \Rightarrow (\Sigma - \lambda, I)W_{(1)} = 0 \quad \cdots\cdots (1-3)$$

이고, 式(1-3)으로 부터

$$W_{(1)}'\Sigma W_{(1)} = \text{Var } (W_{(1)}'X) = \lambda' \quad \cdots\cdots (1-4)$$

으로 된다. 결국, (1−3)과 (1−4)式으로 부터 구한 式(1−2)의 解는 $V(PC_{(1)}) = \lambda_1$ 으로 Σ의 最大固有根(maximum eigen value)값을 가지게 되어, 成分計數 벡터인 $(W_{(1)1}, \cdots, W_{(1)p})' = W_{(1)}$ 은 λ_1에 對應하는 고유벡터(eigen vector)가 된다.

위와 같은 導出 方法에 各 主成分의 無相關 性質($W_{(i)}' W_{(j)} = 0$, $i \neq j$)을 追加하여 p개의 主成分을 導出하면, m번째 主成分은 다음과 같다.

$$PC_{(m)} = W_{(m)1} X_1 + W_{(m)2} X_2 + \cdots + W_{(m)p} X_p, \quad m = 1, \cdots, P \quad (1-5)$$

단, $V(PC_{(m)}) = \lambda_m$은 Σ의 m번째 고유근이며 $W_{(m)} = (W_{(m)1}, \cdots, W_{(m)p})'$는 λ_m에 對應하는 Σ의 고유벡터이다.

그러므로, 主成分의 分散合은

$$\sum_{j=1}^{p} V(PC_{(j)}) = \sum_{j=1}^{p} \lambda_j = t_r(\Sigma) = \sum_{i=1}^{D} V(X_i) \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (1-6)$$

의 關係를 가지게 되어, m번째 主成分이 原變數의 分散(情報)을 설명하는 量의 比는

$$\frac{V(PC_{(m)})}{\sum_{i=1}^{p} V(X_i)} = \frac{\lambda_m}{\sum_{j=1}^{p} \lambda_j} \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (1-7)$$

로 表示된다. 또한, i번째 原變數와 m번째 主成分間의 相關計數는

$$\mathrm{corr}(PC_{(m)}, X_j) = \frac{W_{(m)j}\sqrt{\lambda_m}}{\sigma_j} \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (1-8)$$

이다. 이것을 i번째 原變數의 m번째 主成分에 대한 성분적재값(component loading)이라 부르고, 이 값들을 行列式의 形態로 나열한 것을 成分積載行列(component loading matrix)이라 부른다.

例를 들어, 다음과 같은 11個 變數(X_1 =POPCC (人口增減比), X_2 = INCADJ (所得增減比), ……, X_{11} =ExP (支出增減比)) 를 38個 都市로 부터 觀測한 資料로 부터 共分散行列 Σ를 推定하였을때

表 1－1　　共 分 散 行 列

		POPCC 1	INCADJ 2	EMP 3	PROF 4	IGPER 5	IMBAL 6	PTXRTG 7	BASE 8	DEBT 9	COMPSP 10	EXP 11
POPCC	1	.551[c]										
INCADJ	2	−.371[b]	.720[b]									
EMP	3	.434[c]	.410[a]	.437[c]								
PROF	4	.463[b]	−.245[b]	.141[b]	.950[c]							
IGPER	5	.713[c]	−.632[b]	.657[c]	.115[c]	.457[d]						
IMBAL	6	−.532[c]	.332[c]	.654[b]	.949[c]	.282[d]	.109[f]					
PTXRTG	7	.264[d]	−.199[c]	.213[d]	.301[d]	.166[d]	−.882[c]	.425[c]				
BASE	8	.367[c]	.145[c]	.340[c]	−.437[c]	−.330[c]	−.494[d]	.111[d]	.398[d]			
DEBT	9	−.238[c]	−.551[b]	−.183[c]	.368[c]	.103[c]	.468[c]	−.385[c]	−.144[d]	.207[d]		
COMPSP	10	−.518[c]	.821[c]	−.287[c]	−.134[c]	.248[b]	.247[c]	−.128[d]	−.366[c]	.719[c]	.246[d]	
EXP	11	−.478[c]	.542[b]	−.242[c]	.959[b]	−.660[b]	.194[d]	−.141[d]	−.721[c]	.889[c]	.225[d]	.352[d]

式(1－5)와 (1－7)을 利用하여 구한 各 主成分의 分散 및 分散說明比의 累積값들은 다음과 같다.

表 1－2.　　主成分分散 및 累積說明比

Factor	Variance Explained	Cumulative Proportion of Total Variance
1	10.936	64.3
2	4.323	89.7
3	0.623	93.4
4	0.464	96.1
5	0.376	98.3
6	0.118	99.0
7	0.078	99.4
8	0.061	99.8
9	0.024	99.9
10	0.007	100.0
11	0.003	100.0

위 表에 의하면 3個의 主成分($PC_{(1)}$, $PC_{(2)}$, $PC_{(3)}$)이 11個 原變數들의 分散合을 93.4% 설명하고 있다. 이는 11個의 原變數의 情報를 3個의 主成分으로 거의다(93.4%) 說明 可能하다는 意味이다. 式(1－8)을 利用하여 3個 主成分과 11個 原變數間에 相關計數行列인 成分積載行列(component loadings)을 計算하면 1－3과 같다.

表 1－3　　成 分 積 載 行 列

Variable	Component Loadings 1	2	3
POPCC	−0.0746	0.5667	−0.3723
INCADJ	0.1164	−0.1136	−0.0601
EMP	0.0043	0.5121	−0.2714
PROF	0.0925	0.4470	0.2375
IGPER	0.1311	0.1402	0.0390
IMBAL	0.9996	0.0089	−0.0091
PTXRTG	−0.0218	0.9988	0.0235
BASE	−0.2472	0.0905	−0.5753
DEBT	0.0385	−0.0517	0.6644
COMFSP	0.0206	−0.1461	0.7637
EXP	0.0106	−0.1334	0.8195

3. 主成分의 갯수 決定

대부분의 경우 p개 變數 $X_1, X_2, \cdots, X_p$의 공분산 Σ는 未知의 行列이므로 分析者는 이들의 資料로 부터 Σ를 S로부터 推定하여 使用한다.

$$S=\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})(X_i-\bar{X}') \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (1-9)$$

단, $X_i=(X_{i1}, X_{i2}, \cdots, X_{ip})'$: i 번째 觀測값벡터

$\bar{X}$ =觀測값 벡터들의 平均

그러므로 主成分의 갯수選定基準인 分散說明比(式(1-7)) 역시 S를 利用하여 推定하여야 한다. 推定값에는 항상 推定에 따른 誤差가 包含되어 있어, 이것의 推定값을 基準으로한 主成分의 갯수 決定에 問題가 發生된다.

이 問題를 解決하기 위해 提案된 主成分의 갯수 決定法에는 크게 두가지 方法이 있다.

〈主成分의 갯수 決定法〉

i) 留意性 檢定法(Bartlett (1947)檢定法)

Bartlett에 의해 提案된 檢定法으로 귀무가설

Ho : K개의 主成分만 分析에 必要하다.

다음 檢定統計量으로 檢定하는 方法이다.

$$M[-ln|S|+\sum_{j=1}^{K} ln\ \lambda_j + q\ ln\ l\ ] \sim X^2_{\{\frac{1}{2}(q-1)(q+2)\}} \quad \cdots\cdots\cdots\cdots (1-10)$$

단,

$$q=p-k,\ M=N-K-\frac{1}{b}\left(2q+1+\frac{2}{q}\right)$$

$l = \frac{1}{q}\{ t_r(s) - \sum_{j=1}^{K} \lambda_j \}$, λ_j은 λ_j의 推定값

iii) Graph的 方法(Scree 檢定法)

Scree 檢定法은 Cattell (1966)에 의해 提案된 그래프 方法으로 推定된 고유근값 λ_j을 그림 (1−1)과 같이 산점도로 그리고, 고유근값이 작은 것들을 直線에 適合 시킨후 이 直線에서 동떨어진 고유근값에 該當하는 主成分들의 갯수 만큼을 分析에 使用하는 方法

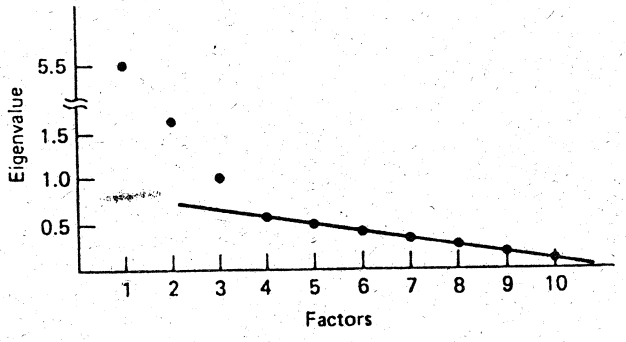


그림 (1−1) Scree 檢定法

4 . 成分點數 (Component scores)

主成分 分析의 結果에서 얻어지는 各 觀測값의 成分點數는 다공선성을 가진 중회귀분석의 解決, 標本의 多變量 定規性 檢定, 特異값 判斷등에 使用된다. i번째 觀測값 벡터 X_i의 K개 主成分 點數는 다음과 같이 구해진다.

i번째 觀測값의 첫번째 主成分點數 :

$$\hat{PC}_{i(1)} = \hat{W}'_{(1)} (X_i - \bar{X})$$

i 번째 觀測값의 두번째 主成分點數 :

$$\hat{PC}_{i(2)} = \hat{W}_{(2)}' (X_i - \bar{X})$$

$$\vdots$$

i 번째 觀測값의 K번째 主成分點數 :

$$\hat{PC}_{i(k)} = \hat{W}_{(k)}' (X_i - \bar{X})$$

5. SAS PROC PRINCOMP의 例題

아래 使用된 資料는 7가지 범주의 罪目(殺人 : Murder, 강간 : Rape, 강도 : Robbery …)에 대한 (人口 10萬名當) 犯罪發生率을 美國의 50個 주를 對象으로 調査한 것이다.

이들 資料의 性質을 쉽게 눈으로 觀察하기 위해 主成分分析을 한 結果는 다음과 같다.

```
DATA CRIME;
  TITLE CRIME RATES PER 100,000 POPULATION BY STATE;
  INPUT STATE $1-15 MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY
    AUTO;
  CARDS;

ALABAMA         14.2  25.2  96.8  278.3 1135.5 1881.9  280.7
ALASKA          10.8  51.6  96.8  284.0 1331.7 3369.8  753.3
ARIZONA          9.5  34.2 138.2  312.3 2346.1 4467.4  439.5
ARKANSAS         8.8  27.6  83.2  203.4  972.6 1862.1  183.4
CALIFORNIA      11.5  49.4 287.0  358.0 2139.4 3499.8  663.5
COLORADO         6.3  42.0 170.7  292.9 1935.2 3903.2  477.1
CONNECTICUT      4.2  16.8 129.5  131.8 1346.0 2620.7  593.2
DELAWARE         6.0  24.9 157.0  194.2 1682.6 3678.4  467.0
FLORIDA         10.2  39.6 187.9  449.1 1859.9 3840.5  351.4
GEORGIA         11.7  31.1 140.5  256.5 1351.1 2170.2  297.9
HAWAII           7.2  25.5 128.0   64.1 1911.5 3920.4  489.4
IDAHO            5.5  19.4  39.6  172.5 1050.8 2599.6  237.6
ILLINOIS         9.9  21.8 211.3  209.0 1085.0 2828.5  528.6
INDIANA          7.4  26.5 123.2  153.5 1086.2 2498.7  377.4
IOWA             2.3  10.6  41.2   89.8  812.5 2685.1  219.9
KANSAS           6.6  22.0 100.7  180.5 1270.4 2739.3  244.3
KENTUCKY        10.1  19.1  81.1  123.3  872.2 1662.1  245.4
LOUISIANA       15.5  30.9 142.9  335.5 1165.5 2469.9  337.7
MAINE            2.4  13.5  38.7  170.0 1253.1 2350.7  246.9
MARYLAND         8.0  34.8 292.1  358.9 1400.0 3177.7  428.5
MASSACHUSETTS    3.1  20.8 169.1  231.6 1532.2 2311.3 1140.1
MICHIGAN         9.3  38.9 261.9  274.6 1522.7 3159.0  545.5
MINNESOTA        2.7  19.5  85.9   85.8 1134.7 2559.3  343.1
MISSISSIPPI     14.3  19.6  65.7  189.1  915.6 1239.9  144.4
MISSOURI         9.6  28.3 189.0  233.5 1318.3 2424.2  378.4
MONTANA          5.4  16.7  39.2  156.8  804.9 2773.2  309.2
NEBRASKA         3.9  18.1  64.7  112.7  760.0 2316.1  249.1
NEVADA          15.8  49.1 323.1  355.0 2453.1 4212.6  559.2
NEW HAMPSHIRE    3.2  10.7  23.2   76.0 1041.7 2343.9  293.4
NEW JERSEY       5.6  21.0 180.4  185.1 1435.8 2774.5  511.5
NEW MEXICO       8.8  39.1 109.6  343.4 1418.7 3008.6  259.5
NEW YORK        10.7  29.4 472.6  319.1 1728.0 2782.0  745.8
NORTH CAROLINA  10.6  17.0  61.3  318.3 1154.1 2037.8  192.1
NORTH DAKOTA     0.9   9.0  13.3   43.8  446.1 1843.0  144.7
OHIO             7.8  27.3 190.5  181.1 1216.0 2696.8  400.4
OKLAHOMA         8.6  29.2  73.8  205.0 1288.2 2228.1  326.8
OREGON           4.9  39.9 124.1  286.9 1636.4 3506.1  388.9
PENNSYLVANIA     5.6  19.0 130.3  128.0  877.5 1624.1  333.2
RHODE ISLAND     3.6  10.5  86.5  201.0 1489.5 2844.1  791.4
SOUTH CAROLINA  11.9  33.0 105.9  485.3 1613.6 2342.4  245.1
SOUTH DAKOTA     2.0  13.5  17.9  155.7  570.5 1704.4  147.5
TENNESSEE       10.1  29.7 145.8  203.9 1259.7 1776.5  314.0
TEXAS           13.3  33.8 152.4  208.2 1603.1 2988.7  397.6
UTAH             3.5  20.3  68.8  147.3 1171.6 3004.6  334.5
VERMONT          1.4  15.9  30.8  101.2 1348.2 2201.0  265.2
VIRGINIA         9.0  23.3  92.1  165.7  986.2 2521.2  226.7
WASHINGTON       4.3  39.6 106.2  224.8 1605.6 3386.9  360.3
WEST VIRGINIA    6.0  13.2  42.2   90.9  597.4 1341.7  163.3
WISCONSIN        2.8  12.9  52.2   63.7  846.9 2614.2  220.7
WYOMING          5.4  21.9  39.7  173.9  811.6 2772.2  282.0
;
PROC PRINCOMP OUT=CRIMCOMP;
```

```
PROC SORT;
  BY PRIN1;
PROC PRINT;
  ID STATE;
  VAR PRIN1 PRIN2 MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY
    AUTO;
  TITLE2 STATES LISTED IN ORDER OF OVERALL CRIME RATE;
  TITLE3 AS DETERMINED BY THE FIRST PRINCIPAL COMPONENT;
PROC SORT;
  BY PRIN2;
PROC PRINT;
  ID STATE;
  VAR PRIN1 PRIN2 MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY
    AUTO;
  TITLE2 STATES LISTED IN ORDER OF PROPERTY VS. VIOLENT CRIME;
  TITLE3 AS DETERMINED BY THE SECOND PRINCIPAL COMPONENT;
  PROC PLOT;
    PLOT PRIN2*PRIN1=STATE;
    TITLE2 PLOT OF THE FIRST TWO PRINCIPAL COMPONENTS;
  PROC PLOT;
    PLOT PRIN3*PRIN1=STATE;
    TITLE2 PLOT OF THE FIRST AND THIRD PRINCIPAL COMPONENTS;
```

CRIME RATES PER 100,000 POPULATION BY STATE

PRINCIPAL COMPONENT ANALYSIS

50 OBSERVATIONS
7 VARIABLES

SIMPLE STATISTICS

	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
MEAN	7.444000	25.73400	124.0920	211.3000	1291.904	2671.288	377.5260
ST DEV	3.866769	10.75963	88.3486	100.2530	432.456	725.909	193.3944

CORRELATIONS

	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
MURDER	1.0000	0.6012	0.4837	0.6486	0.3858	0.1019	0.0688
RAPE	0.6012	1.0000	0.5919	0.7403	0.7121	0.6140	0.3489
ROBBERY	0.4837	0.5919	1.0000	0.5571	0.6372	0.4467	0.5907
ASSAULT	0.6486	0.7403	0.5571	1.0000	0.6229	0.4044	0.2758
BURGLARY	0.3858	0.7121	0.6372	0.6229	1.0000	0.7921	0.5580
LARCENY	0.1019	0.6140	0.4467	0.4044	0.7921	1.0000	0.4442
AUTO	0.0688	0.3489	0.5907	0.2758	0.5580	0.4442	1.0000

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	4.114960	2.876238	0.587851	0.587851
PRIN2	1.238722	0.512905	0.176960	0.764812
PRIN3	0.725817	0.409385	0.103688	0.868500
PRIN4	0.316432	0.058458	0.045205	0.913704
PRIN5	0.257974	0.035935	0.036853	0.950558
PRIN6	0.222039	0.097983	0.031720	0.982278
PRIN7	0.124056	.	0.017722	1.000000

EIGENVECTORS

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7
MURDER	0.300279	-.629174	0.178245	-.232114	0.538123	0.259117	0.267593
RAPE	0.431759	-.169435	-.244198	0.062216	0.188471	-.773271	-.296485
ROBBERY	0.396875	0.042247	0.495861	-.557989	-.519977	-.114385	-.003903
ASSAULT	0.396652	-.343528	-.069510	0.629804	-.506651	0.172363	0.191745
BURGLARY	0.440157	0.203341	-.209595	-.057555	0.101033	0.535987	-.648117
LARCENY	0.357360	0.402319	-.539231	-.234890	0.030099	0.039406	0.601650
AUTO	0.295177	0.502421	0.568384	0.419238	0.369753	-.057298	0.147046

CRIME RATES PER 100,000 POPULATION BY STATE
STATES LISTED IN ORDER OF PROPERTY VS. VIOLENT CRIME
AS DETERMINED BY THE SECOND PRINCIPAL COMPONENT

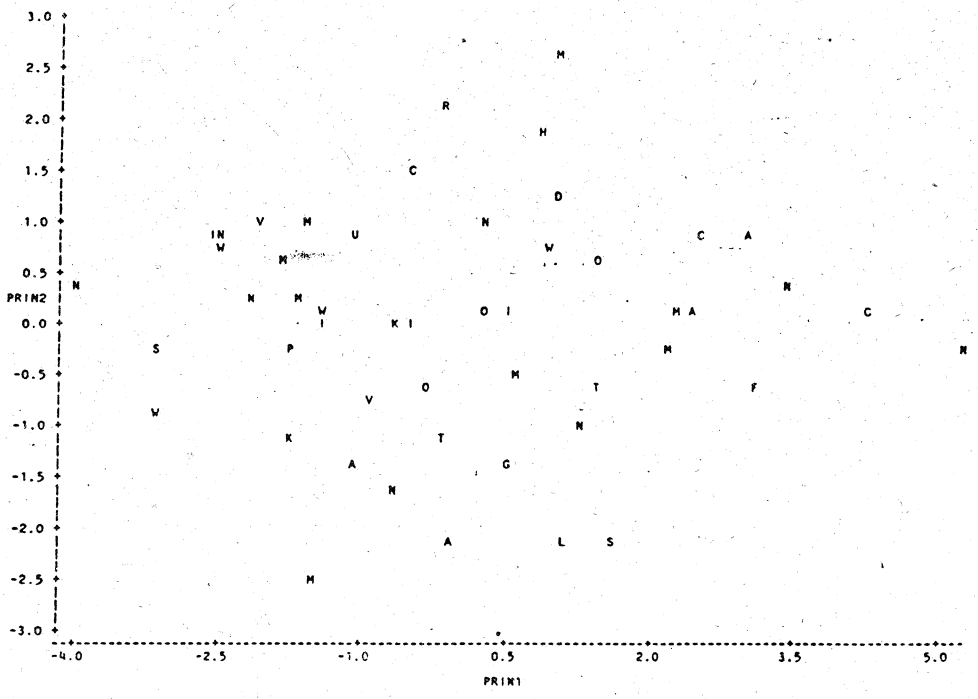
3

STATE	PRIN1	PRIN2	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
MISSISSIPPI	-1.5074	-2.5467	14.3	19.6	65.7	189.1	915.6	1239.9	144.4
SOUTH CAROLINA	1.6034	-2.1621	11.9	33.0	105.9	485.3	1613.6	2342.4	245.1
ALABAMA	-0.0499	-2.0961	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
LOUISIANA	1.1202	-2.0833	15.5	30.9	142.9	335.5	1165.5	2469.9	337.7
NORTH CAROLINA	-0.6993	-1.6703	10.6	17.0	61.3	318.3	1154.1	2037.8	192.1
GEORGIA	0.4904	-1.3808	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9
ARKANSAS	-1.0544	-1.3454	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
KENTUCKY	-1.7269	-1.1866	10.1	19.1	81.1	123.3	872.2	1662.1	245.4
TENNESSEE	-0.1366	-1.1350	10.1	29.7	145.8	203.9	1259.7	1776.5	314.0
NEW MEXICO	1.2142	-0.9508	8.8	39.1	109.6	343.4	1418.7	3008.6	259.5
WEST VIRGINIA	-3.1477	-0.8143	6.0	13.2	42.2	90.9	597.4	1341.7	163.3
VIRGINIA	-0.9162	-0.6927	9.0	23.3	92.1	165.7	986.2	2521.2	226.7
TEXAS	1.3970	-0.6813	13.3	33.8	152.4	208.2	1603.1	2988.7	397.6
OKLAHOMA	-0.3214	-0.6243	8.6	29.2	73.8	205.0	1288.2	2228.1	326.8
FLORIDA	3.1118	-0.6039	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4
MISSOURI	0.5564	-0.5585	9.6	28.3	189.0	233.5	1318.3	2424.2	378.4
SOUTH DAKOTA	-3.1720	-0.2545	2.0	13.5	17.9	155.7	570.5	1704.4	147.5
NEVADA	5.2670	-0.2526	15.8	49.1	323.1	355.0	2453.1	4212.6	559.2
PENNSYLVANIA	-1.7201	-0.1959	5.6	19.0	130.3	128.0	877.5	1624.1	333.2
MARYLAND	2.1828	-0.1947	8.0	34.8	292.1	358.9	1400.0	3177.7	428.5
KANSAS	-0.6381	-0.0280	6.6	22.0	100.7	180.5	1270.4	2739.3	244.3
IDAHO	-1.4325	-0.0080	5.5	19.4	39.6	172.5	1050.8	2599.6	237.6
INDIANA	-0.4999	0.0000	7.4	26.5	123.2	153.5	1086.2	2498.7	377.4
WYOMING	-1.4246	0.0627	5.4	21.9	39.7	173.9	811.6	2772.2	282.0
OHIO	0.2395	0.0905	7.8	27.3	190.5	181.1	1216.0	2695.8	400.4
ILLINOIS	0.5129	0.0942	9.9	21.8	211.3	209.0	1085.0	2828.5	528.6
CALIFORNIA	4.2838	0.1432	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5
MICHIGAN	2.2733	0.1549	9.3	38.9	261.9	274.6	1522.7	3159.0	545.5
ALASKA	2.4215	0.1665	10.8	51.6	96.8	284.0	1331.7	3369.8	753.3
NEBRASKA	-2.1507	0.2257	3.9	18.1	64.7	112.7	760.0	2316.1	249.1
MONTANA	-1.6680	0.2710	5.4	16.7	39.2	156.8	804.9	2773.2	309.2
NORTH DAKOTA	-3.9641	0.3877	0.9	9.0	13.3	43.8	446.1	1843.0	144.7
NEW YORK	3.8525	0.4329	10.7	29.4	472.6	319.1	1728.0	2782.0	745.8
MAINE	-1.8263	0.5788	2.4	13.5	38.7	170.0	1253.1	2350.7	246.9
OREGON	1.8490	0.5860	4.9	39.9	124.1	286.9	1636.4	3506.1	388.9
WASHINGTON	0.9306	0.7378	4.3	39.6	106.2	224.8	1605.6	3386.9	360.3
WISCONSIN	-2.5030	0.7808	2.8	12.9	52.2	63.7	846.9	2614.2	220.7
IOWA	-2.5816	0.8248	2.3	10.6	41.2	89.8	812.5	2685.1	219.9
NEW HAMPSHIRE	-2.4656	0.8250	3.2	10.7	23.2	76.0	1041.7	2343.9	293.4
ARIZONA	3.0141	0.8449	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
COLORADO	2.5093	0.9166	6.3	42.0	170.7	292.9	1935.2	3903.2	477.1
UTAH	-1.0500	0.9366	3.5	20.3	68.8	147.3	1171.6	3004.6	334.5
VERMONT	-2.0643	0.9450	1.4	15.9	30.8	101.2	1348.2	2201.0	265.2
NEW JERSEY	0.2179	0.9642	5.6	21.0	180.4	185.1	1435.8	2774.5	511.5
MINNESOTA	-1.5543	1.0564	2.7	19.5	85.9	85.8	1134.7	2559.3	343.1
DELAWARE	0.9646	1.2967	6.0	24.9	157.0	194.2	1682.6	3678.4	467.0
CONNECTICUT	-0.5413	1.5012	4.2	16.8	129.5	131.0	1346.0	2620.7	593.2
HAWAII	0.8231	1.8239	7.2	25.5	128.0	64.1	1911.5	3920.4	489.4
RHODE ISLAND	-0.2016	2.1866	3.6	10.5	86.5	201.0	1489.5	2844.1	791.4
MASSACHUSETTS	0.9784	2.6311	3.1	20.8	169.1	231.6	1532.2	2311.3	1140.1

CRIME RATES PER 100,000 POPULATION BY STATE
PLOT OF THE FIRST TWO PRINCIPAL COMPONENTS

4

PLOT OF PRIN2*PRIN1 SYMBOL IS VALUE OF STATE

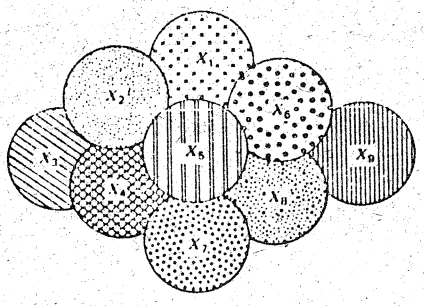


Part E 因子分析(Factor Analysis)

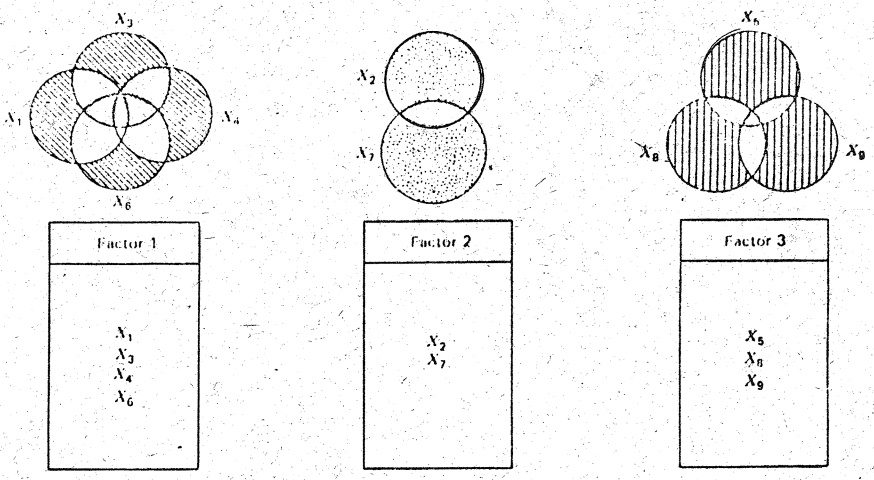
1. 因子分析의 定義

因子分析은 資料를 壓縮 要約하는 多變量 統計分析方法中의 하나로써, 分析의 目標는 여러개의 變數들 사이에 存在하는 複雜多樣한 相關關係를 몇개의 因子(factor)들로써 간단하게 說明하는 技法이다. 여기서, 因子란 서로 相關關係가 높은 變數들끼리 모아서 작은 變數群으로 나눈 것을 말한다.

이와같은 狀況을 圖解하면 다음 그림(1－1)과 같다. 이 그림은 9개



相關된 9個의 變數



3개의 因子

의 變數 $X_1, X_2, \cdots\cdots, X_9$ 가 세개의 因子로 分解된 것을 나타낸 것이다. 各 變數群 $\{X_1, X_3, X_4, X_6\}$, $\{X_2, X_1\}$, $\{X_5, X_8, X_9\}$에 包含된 變數들은 서로 相關關係가 높은 것들로 構成되어 있다. 이때 變數群들을 9個 變數에 內在된 세개의 因子로 解析하면 變數들이 가진 複雜한 相關關係를 세개의 因子로써 간단히 說明할 수 있게 된다.

2. 因子分析의 理論

變數(普通 標準化된 變數를 使用)와 因子들의 關係를 나타내는 基本的 因子分析模型은 다음과 같이 이들의 共通部分(Common Part)과 特性部分(Unique Part)으로 이루어진다.

$$\begin{aligned} X_1 &= \lambda_{11}f_1 + \lambda_{12}f_2 + \cdots\cdots + \lambda_{1q}f_q + e_1 \\ X_2 &= \lambda_{21}f_1 + \lambda_{22}f_2 + \cdots\cdots + \lambda_{2q}f_q + e_2 \\ &\vdots \\ X_p &= \lambda_{p1}f_1 + \lambda_{p2}f_2 + \cdots\cdots + \lambda_{pq}f_q + e_p \end{aligned} \quad \cdots\cdots (1-1)$$

(共同部分 / 特性部分)

단, $f_1, f_2, \cdots\cdots, f_q$는 p개 變數 $X_1, X_2, \cdots\cdots, X_p$에 內在된 q개 ($p > q$)의 共通因子를 나타내고, λ_{ij}는 i번째 變數와 j번째 因子와의 關係를 나타내는 因子積載값(Factor loading coefficient)를 나타낸다.

2-1. 模型의 假定

式 (1-1)을 行列式으로 表示하면,

$$\underset{p\times 1}{X} = \underset{p\times q}{\Lambda}\, f_{q\times 1} + \underset{p\times 1}{e} \quad \cdots\cdots (1-2)$$

이고, 模型의 假定은 다음과 같다.

i) $\mathrm{Var}(e) = \Psi = \mathrm{diag}(\Psi_1, \Psi_2, \cdots\cdots, \Psi_p)$,

단, Ψ_i 들을 X_i 의 特性分散 (Unique Variance) 라 부른다.

ii) $\mathrm{Cov}(e, f') = 0$, 즉, 各 因子와 特性部分은 서로 獨立이다.

iii) $\mathrm{Cov}(f) = \begin{cases} I_q & : \text{直交因子模型 (Orthogonal factor model)} \\ \Phi_{q\times q} & : \text{斜角因子模型 (Oblique factor model)} \end{cases}$

단, $\Phi_{q\times q} = \begin{pmatrix} 1 & \phi_{12} & \phi_{13} & \cdots\cdots & \phi_{1q} \\ & 1 & \phi_{23} & \cdots\cdots & \phi_{2q} \\ & & \ddots & & \\ \text{Sym.} & & & & \phi_{q-1,q} \\ & & & & 1 \end{pmatrix}$

iv) 直交因子模型下에서 式 (1-1)에 包含된 未知의 母數를 唯一하게 推定할 수 있는 必要條件은

$$q < \frac{1}{2}(p-1)$$

이다.

2-2. 因子模型의 種類

因子模型의 種類는 因子들의 공분산행렬 ($\mathrm{Var}(f)$)의 形態에 따라서 直交因子模型과 斜角因子模型으로 分類된다. 直交因子模型이란, 因子들이 서로 獨立的이란 假定이 包含된 模型으로 이 模型下에서 抽出되는 因子의 軸은 90°를 維持하여 數學的으로 다루기가 간편하고 分析이 간단하다. 반면, 斜

角因子模型下에서 抽出되는 因子는 因子軸들 사이의 角度가 90°를 維持하지 않아서 直交의 경우보다 計算過程이 훨씬 複雜하며, 아직 滿足할 만한 分析方法이 開發되지 못하였다. 그러나, 因子들간에 相關關係를 認定하는 模型이어서 훨씬 現實性에 가까운 模型이라 할 수 있다.

直交 또는 斜角因子分析 中 어느 것을 選擇해야 하는지의 問題는 당면한 과제의 特殊한 要求에 根據하여 決定되어야 한다.

直交因子分析 : 硏究의 目的이 原變數의 數를 줄여서 다른 統計方法에 活用하기 위한 경우.

斜角因子分析 : 分析의 目的이 原資料로 부터 몇개의 理論的 意味를 지닌 因子나 構成을 얻는 경우.

一般的인 分析節次는 直交因子模型으로 부터 因子를 抽出한 後 分析에서 斜角因子가 必要할 경우 因子積載行列 (Factor Loading Matrix 또는 Factor pattern Matrix)를 斜角因子模型의 因子積載行列이 되도록 回轉 (Rotation)시키면 된다. 여기서 因子積載行列이란 各 因子들에 대한 모든 變數값들의 因子積載값을 行列式으로 모아놓은 表를 말하며 式 (1 - 2)에 나타난 行列 A를 意味한다.

3 . 因子分析의 目的

因子分析의 一般的인 目的은 여러개의 變數에 대한 資料속에 들어있는 情報를 最大限 反映시키는 적은 數의 因子들을 抽出함으로써 資料의 壓縮 및 要約의 效果를 얻는 것이다. 因子分析으로 達成할 수 있는 세가지 機能은 다음과 같다.

3－1．因子分析의 機能

i）R型 因子分析（R-type factor Analysis）：많은 數의 變數에 대한 資料속에 內在된 一聯의 因子를 찾아낸다.

ii）Q型 因子分析（Q-type factor Analysis）：많은 數의 對象者들을 작은 數의 集團（Group）으로 묶거나 區別한다.

iii）回歸分析이나 判別分析등 또다른 統計方法에 使用될 작은 數의 變數（因子）를 새로이 만들어 낸다.

4．因子分析의 節次

因子分析의 節次를 흐름圖로 나타내면 그림（1-2）와 같다.

因子積載값 推定法에는 主成分分析方法（principal component method），主因子分析方法（principal factor analysis method），最優推定法（maximum likelihood method） 最小自乘法등이 있으나，여기서는 가장 普遍的으로 使用되는 主因子分析方法案을 說明하기로 한다.

4－1．研究課題의 選定節次

因子分析에 앞서 分析者는 다음과 같은 事項을 確認해야 한다.

① 어떤 變數들이 包含되어야 하는가?

② 몇개의 變數가 包含되어야 하는가?

③ 어떻게 變數를 測定할 것인가?

④ 標本의 數는 充分히 큰가?

여기서 變數는 問題와 關聯하여 測定可能한 것이면 어떤것이든 包含될 수 있다. 因子分析의 變數는 一般的으로 量的인 變數를 다루지만 質的인

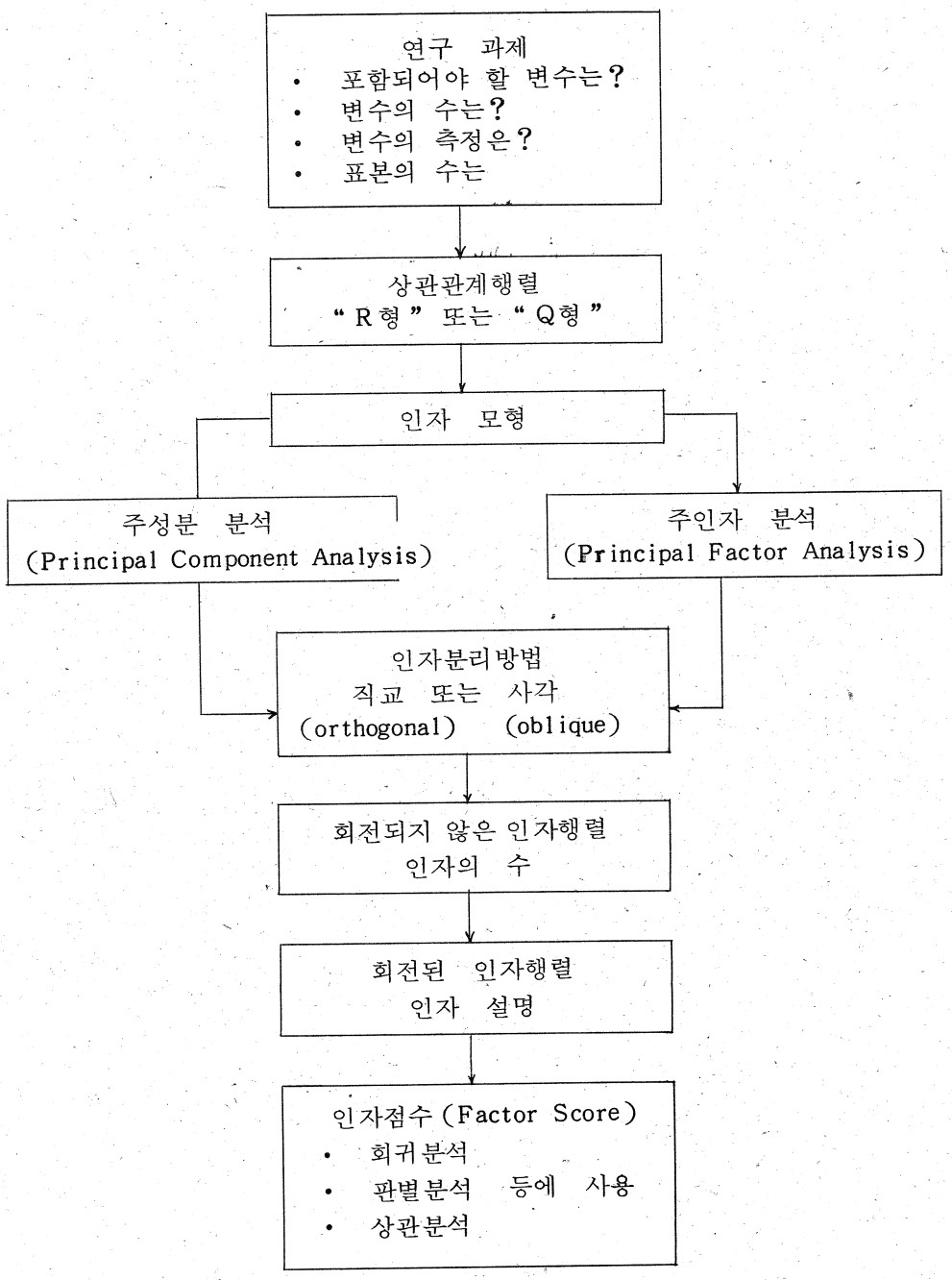


그림 (1 − 2) 因子分析의 節次

變數도 包含시킬 수 있다.

資料의 數는 一般的으로 分析하려는 變數數의 約 4～5倍 정도로 擇하나 경우에 따라서 約 2倍 정도의 觀測값(資料)들로도 分析이 可能하다.

그러나 적은 數의 資料로 分析할 경우 分析 結果의 解釋에 愼重을 기해야 한다. 아래 資料는 Rummel (1970)이 Q-Type 主因子分析에 使用한 12個國의 特性을 나타낸 資料로써 이것을 利用하여 因子分析의 節次를 說明하고자 한다.

表 1－1. 12個國의 特性(10個 變數로 構成됨)資料

	(1) GNP per Capita ($)	(2) Trade (Millions of $)	(3) Power (Rank)[b]	(4) Stability	(5) Freedom of Group Opposition[d]	(6) Foreign Conflict[e]	(7) Agreement with U.S. in U.N.[f]	(8) Defense Budget (Millions of $)	(9) GNP for Defense (%)	(10) Acceptance of International Law[g]
Brazil	91	2,729	7	0	2	0	69.1	148	2.8	0
Burma	51	407	4	0	1	0	−9.5	74	6.9	0
China	58	349	11[h]	0	0	1	−41.7[h]	3,054	8.7	0
Cuba	359	1,169	3	0	1	0	64.3	53	2.4	0
Egypt	134	923	5	1	1	1	−15.4	158	6.0	1
India	70	2,689	10	0	2	0	−28.6	410	1.9	1
Indonesia	129	1,601	8	0	1	0	−21.4	267	6.7	0
Israel	515	415	2	1	2	1	42.9	33	2.7	1
Jordan	70	83	1	0	1	1	8.3	29	25.7	0
Netherlands	707	5,395	6	1	2	0	52.3	468	6.1	1
Poland	468	1,852	9	0	0	1	−41.7	220	1.5	0
U.S.S.R.	749	6,530	13	1	0	1	−41.7	34,000	20.4	0
U.K.	998	18,677	12	1	2	1	69.0	3,934	7.8	0
U.S.	2,334	26,836	14	1	2	1	100.0	40,641	12.2	1

4－2. 調整된 相關行列의 推定節次

式(1-1)에서 定義된 (i)번째 變數(標準化된 變數)의 因子分析模型을 分散式으로 나타내면,

表 1 － 2 .　調整된　相關行列

Characteristic	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) GNP per capita	.97									
(2) Trade	.93	.97								
(3) Power	.55	.66	.89							
(4) Stability	.62	.55	.25	.63						
(5) Freedom of group opposition	.31	.40	−.10	.32	.91					
(6) Foreign conflict	.36	.30	.25	.46	−.32	.61				
(7) Agreement with U.S. in U.N.	.58	.59	−.07	.36	.75	.11	.89			
(8) Defense budget	.79	.71	.66	.49	−.07	−.38	−.18	.90		
(9) Percentage of GNP for defense	.17	.17	.06	.15	−.28	−.44	.11	.47	.73	
(10) Acceptance of international law	.34	.22	−.02	.56	.57	−.04	−.24	.14	−.24	.82

4 － 3 . 因子分離節次

調整된　相關行列　R^*에　直交分解理論(Orthogonal decomposition theorem)를　適用시켜　이것의　固有根(eigenvalue)을　求한다.

〈直交分解法〉

$$R^* = TDT'$$

$$= TD^{\frac{1}{2}}(TD^{\frac{1}{2}})' \quad \cdots\cdots (1-5)$$

, 단　$D = \mathrm{diag}(d_1, d_2, \cdots\cdots, d_p)$로써

$d_1 \geq d_2 \geq \cdots\cdots, \geq d_p$는　p개　因子의　固有根이　된다.

아래　表의　百分比는　各　因子가　變數의　情報(分散)를　얼마나　說明하는가의　量을　나타낸다. 그러므로　이들의　累積　百分比가　充分한　q개　因子만을　分離시켜　이들의　因子積載값　行列(Λ)을　式(1 - 6)으로　부터　求한다.

$$R^* = T^*D^{*\frac{1}{2}}(T^*D^{*\frac{1}{2}})' \quad \cdots\cdots (1-6)$$

$$= \Lambda\Lambda'$$

$$\mathrm{Var}(X_i) = \mathrm{Var}(\lambda_{i1}f_1 + \lambda_{i2}f_2 + \cdots\cdots + \lambda_{iq}f_q) + \mathrm{Var}(\rho_i)$$

$$\Rightarrow \quad 1 = h_i^2 + \Psi_i \quad \cdots\cdots (1-3)$$

로 表示된다. 여기서 h_i^2은 q개의 共通因子들이 變數 X_i의 情報(分散)를 얼마만큼 說明하고 있는지를 나타내는 測度로써 이것을 變數 X_i에 대한 共通因子들의 커뮤날리티(Communality)라 부른다.

調整된 相關行列(Reduced Correlation matrix)이란 p개의 變數들이 相關行列의 대각원소들을 각 變數들이 커뮤낼리티값 h_i^2으로 代替시킨 것으로 式(1-2)의 分散式에서 다음과 같이 誘導된 $\Lambda\Lambda'$의 行列을 意味한다.

$$V(X) = V(\Lambda f) + V(e)$$

$$\Rightarrow R = \Lambda\Lambda' + \Psi$$

$$\Rightarrow R - \Psi\,(\text{調整된 相關行列}) = \Lambda\Lambda' \quad \cdots\cdots (1-4)$$

調整된 相關行列을 推定하기 위해 必要한 커뮤낼리티 h_i^2, $i = 1, 2, \cdots\cdots p$는 여러 方法으로 推定可能하나 여기서는 다음 2가지 方法만 紹介하기로 한다.

方法 1 : SMC (X_i와 나머지 $p-1$개 變數 사이의 中相關係數)를 利用하여 h_i^2의 下限을 求하는 方法.

方法 2 : SMC를 初期 推定값으로 使用한 後 主因子分析을 反復的으로 施行하여 安定된 h_i^2을 구하는 方法. 이때 推定된 h_i^2의 값이 1보다 큰 數를 갖는 경우가 있는데, 이 경우를 Heywood Case라 부르며, 1보다 큰 h_i^2의 값대신 0.99 또는 1을 代入시켜 反復 推定作業을 繼續하면 된다.

이 節次에 의해 表1－1의 資料로 推定한 調整된 相關行列은 다음과 같다.

表 1－3．　表 1－2에　나타난　調整된　相關行列의　固有根　및　分散說明比

Factor	Eigenvalue	Percent	Cumulative Percentage[a]
1	4.09	49.3	49.3
2	2.26	27.2	76.5
3	0.92	11.1	87.5
4	0.77	9.3	96.9
5	0.30	3.6	100.4
6	0.18	2.2	102.6
7	0.02	0.2	102.8
8	−0.04	−0.5	102.4
9	−0.09	−1.1	101.3
10	−0.11	−1.3	100.0

단，$D^* = \mathrm{diag}(d_1, d_2, \cdots\cdots, d_q)$．

아래表는　表 1－3．의　累積百分比에　의해　4 個의　因子 $\{f_1, f_2, f_3, f_4\}$ 를　分離시키고，이들의　因子積載값行列을　求한　結果이다．特히，直交因子分析法下의　因子積載값　行列은　原變數 $(X_1, \cdots\cdots, X_p)$와　因子 $(f_1, \cdots\cdots, f_q)$間에　相關行列인　因子構造行列（factor structure matrix）의　役割도　한다．

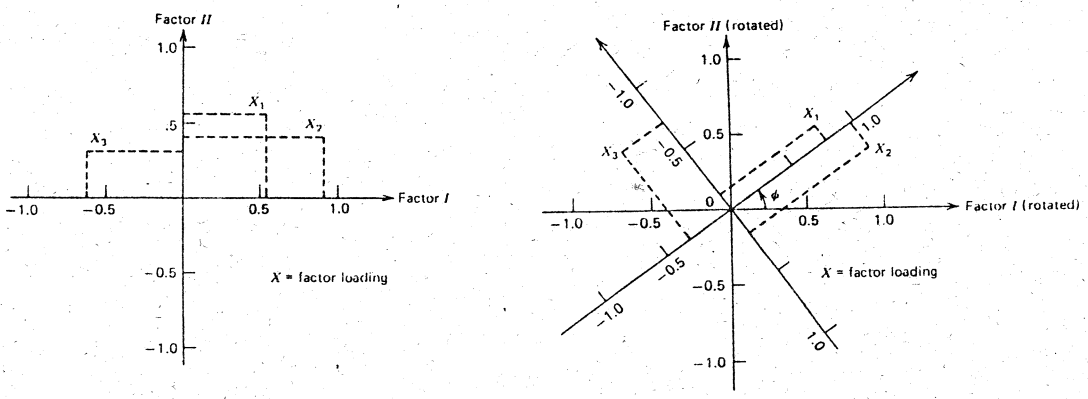
表 1 － 4 .　4 個因子의　因子行列

Variables	f_1	f_2	f_3	f_4	h_i^2	Φ_i^2
(1) GNP per capita	(.95)	－.04	－.07	.05	.92	.08
(2) Trade	(.94)	－.03	－.26	.01	.95	.05
(3) Power	(.59)	－.47	－.33	－.53	.96	.04
(4) Stability	(.71)	.07	.44	.02	.70	.30
(5) Freedom of group opposition	.39	(.81)	－.07	－.01	.82	.18
(6) Foreign conflict	.37	(－.48)	.38	.18	.54	.46
(7) Agreement with U.S. in U.N.	.57	(.62)	.29	.39	.95	.05
(8) Defense budget	.16	(－.43)	－.02	.02	.77	.23
(9) Percentage of GNP for defense	.20	(－.50)	.15	.41	.49	.51
(10) Acceptance of international law	.42	.51	(.56)	－.35	.88	.12
Total variance (%)	40.7	22.1	9.4	7.6	79.8	20.2
Common variance (%)	51.0	27.7	11.8	9.5		
Factor contribution of factor j (eigenvalue)	4.07	2.21	0.94	0.76		

4 －4 . 因子積載값行列의　回轉

因子分析은　分離된　因子積載값行列의　解釋에　의해　이루어　진다.　正確한 因子分析을　위해서는　因子積載값行列이　가장　解釋이　容易한　形態로　나타나야　한다. 이러한　形態는　因子積載값行列에　適當한　回轉行列(Rotation matrix)을　곱하여　얻을　수　있는데, 이　作業을　因子回轉(factor rotation)이라　부른다.

因子回轉의　效果를　간단히　그림으로　說明하면　다음과　같다.



(a) 回轉前　　(b) 回轉後

그림 1 ㅡ 3 . 因子回轉의 效果

위 그림 (a)에서 變數 X_1, X_2, X_3가 因子 Ⅰ과 因子 Ⅱ중 어느 것과 더 相關關係가 있는지를 判斷하기 어렵다. 이때 이것을 적당히 回轉을 시키면 그림 (b)와 같이 變數 X_1과 X_2는 因子 Ⅰ과 높은 相關關係가 있고, 變數 X_3는 因子 Ⅱ와 높은 相關關係가 있다는 것을 쉽게 알 수 있다.

Thurstone (1947)이 提案한 回轉 (Rotation)의 基準은 다음과 같다.

① 因子積載값行列의 各 列에 0에 가까운 값들이 있도록 回轉시킨다.

② 因子積載값行列의 各 列에 큰 값이 最小限의 갯수로 나타나도록 回轉시킨다.

③ 因子積載값行列의 各 列들은 서로 다른 패턴의 값들을 가지도록 한다.

위와같은 基準에 의해 因子積載값行列을 回轉시키는 方法은 回轉行列의 直交性 有無에 따라 直交回轉方法과 斜角回轉方法이 있다. 各 方法은 使用되는 回轉行列의 形態에 따라 다음과 같은 技法들이 提案되어 있다.

A. 直交回轉方法 (Orthogonal rotation method)

Varimax 法 : 因子積載값行列의 列의 패턴을 明確하게 하는 方法

Quartimax 法 : 因子積載값行列의 行의 패턴을 明確하게 하는 方法

Equimax 法 : 因子積載값行列의 行과 列의 패턴을 모두 明確하게 하는 方法

B. 斜角回轉方法 (Oblique rotation method)

回轉에 使用된 回轉行列이 直交行列이 아닌 경우로써, 이때의 回轉結果는 因子들의 軸이 直交性을 維持하지 않는다.

Oblimax 法 : 因子積載값이 큰것과 작은 것으로만 나타나게 하는 方法

Oblimin 法 : Quartmin 法과 Covarimin 法의 混合形態

Quartmin 法 : 因子積載값行列이 가진 內的의 合을 最小化시키는 方法

Cavarimin 法 : 因子積載값行列이 가진 列의 패턴을 明確하게 하는 方法

Promax 法 :

다음表는 表 1 - 5 의 因子積載값行列을 Varimax 法과 Oblique 法으로 각각 回轉시킨 結果이다.

이 表의 Varimax에 의해 分離된 因子들은 각각 다음과 같은 變數群들과 相關이 높은 것을 因子積載값行列로 부터 導出해 낼 수 있다.

첫째因子 f_1 = { GNP, Trade, Power, Defense budget }

둘째因子 f_2 = { Freedom, Agreement with U.S.A }

세째因子 f_3 = { Foreign conflict }

네째因子 f_4 = { Acceptance of international law }

表 1 − 6 . Varimax 및 Oblique 回轉結果

Variables	Unrotated Factors					Varimax Rotated Factors				Oblique Rotated Factors			
	f_1	f_2	f_3	f_4	h_i^2	f_1	f_2	f_3	f_4	f_1	f_2	f_3	f_4
(1) GNP per capita	(.95)	−.04	−.07	.05	.92	(.67)	.51	.35	.29	(.96)	.07	.10	.07
(2) Trade	(.94)	−.03	−.26	.01	.95	(.75)	.55	.23	.16	(.94)	.09	−.08	.08
(3) Power	(.59)	−.47	−.33	−.53	.96	(.97)	−.15	.02	−.00	(.60)	−.34	−.34	−.43
(4) Stability	(.71)	.07	.44	.02	.70	.27	.27	.40	(.63)	(.70)	.12	.53	.04
(5) Freedom of group opposition	.39	(.81)	−.07	−.01	.82	−.01	(.71)	−.38	.42	.34	(.87)	.12	.09
(6) Foreign conflict	.37	(−.48)	.38	.18	.54	.20	−.13	(.69)	.18	.41	(−.47)	.43	.07
(7) Agreement with U.S. in U.N.	.57	(.62)	.29	.39	.95	.09	(.96)	−.05	.11	.52	(.65)	.00	.44
(8) Defense budget	.16	−.43	−.02	.02	.77	(.68)	.15	.52	.11	(.81)	−.35	.06	.03
(9) Percentage of GNP for defense	.20	(−.50)	.15	.41	.49	.07	−.04	(.67)	−.16	.25	(−.56)	.25	.49
(10) Acceptance of international law	.42	.51	(.56)	−.35	.88	.04	.16	−.13	(.91)	.39	.51	(.59)	−.35
Total variance (%)	40.7	22.1	9.4	7.6	79.8	25.6	21.6	16.6	16.0				
Common Variance (%)	51.0	27.7	11.8	9.5		32.1	27.1	20.8	20.0				

그러므로 위 變數그룹을 잘 解析하면 因子의 性格을 把握할 수 있다. 즉, 첫째因子 f_1은 各國의 國力을 나타내는 因子이고, f_2는 各國의 自由 및 民主化 水準을 나타낸다.

直交回轉方法下에서 求한 因子積載값行列의 性質을 式 (1-2)로부터 展開하면 다음과 같다.

$$R = \Gamma\Lambda(\Gamma\Lambda)' + \Psi$$

$$= \Lambda\Lambda' + \Psi \quad \cdots\cdots (1-6)$$

단, Γ ≒直交回轉行列
$\Gamma\Lambda$ =直交回轉된 因子積載값行列

그러므로, 式(1-2)의 假定으로부터

$$\mathrm{Corr}(X, f) = \Gamma\Lambda \quad \cdots\cdots (1-7)$$

가 成立되어 回轉된 因子積載값行列은 因子構造行列(原變數 X와 因子 f의 相關係數行列)과 같다.

그러나, 因子積載값行列을 直交回轉시켜도 因子積載값行列의 패턴이 明確하지 않아 解析이 困難한 경우나 因子들이 서로 相關關係가 있다고 事前에 判斷된 경우는 斜角回轉法을 使用하여 回轉시키게 되는데, 이 方法에 의해 回轉된 因子積載값行列의 解析에는 다음과 같은 問題點이 있다.

〈斜角回轉의 問題點〉

① 回轉된 因子積載값行列은 因子構造行列(factor structure matrix)이 아니다.

② 因子들이 相關關係를 가지고 있으므로 각 變數에 대한 共通因子의 커뮤널리티(h_i^2)에 대해서 分析을 할 수 없다.

5. 因子點數

因子分析은 分析目的에 따라, 因子들에 대한 說明 뿐 아니라 因子點數를 求한 後 다른 統計方法에 의한 分析으로 連結될 수도 있다. 여기서 因子

點數의 役割은 p개 變數의 觀測값들을 q개의 因子空間에 表示할때 使用되는 座標이다. 그러므로, 因子點數란 p개 變數의 觀測값들을 壓縮, 要約시켜 q개의 새로운 變數(因子)의 觀測값들로 變型시킨 것이라 할 수 있다.

이 因子點數들은 因子分析에 이어서 使用될 回歸分析, 判別分析등 다른 統計方法의 獨立變數로 使用된다.

i번째 觀測값 $(X_{i1}, X_{i2}, \cdots\cdots, X_{ip})'$로부터 얻어진 j번째 因子의 因子點數 F_{ij}를

$$F_{ij} = B_{1j} X_{i1} + B_{2j} X_{i2} + \cdots\cdots + B_{9j} X_{ip} + e_{ij}$$

의 關係式으로 나타내어 中回歸分析으로 推定하면 因子點數들의 行列은 다음과 같이 表現된다.

$$\underset{n\times q}{\hat{F}} = \{\hat{F}_{ij}\} = \underset{n\times p}{X}\ \underset{p\times p}{R^{-1}}\ \underset{p\times q}{\Lambda} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (1-8)$$

, 단 X : 標準化된 資料行列
R : p개 變數의 相關行列
Λ : 因子積載값行列

表 1－6에서 Varimax 回轉法으로 얻은 因子積載값行列과 原資料의 標準化 값을 式 (1-8)에 代入시켜 求한 14個國의 因子點數는 다음과 같다.

表 1－7．　14 個國의　因子點數

Nation	Power f_1	Agreement with U.S. f_2	Foreign Conflict f_3	Acceptance of International Law f_4
Brazil	−0.228	0.987	−1.300	−1.021
Burma	−0.970	−0.010	0.259	−0.828
China	0.670	−1.325	−0.341	−0.539
Cuba	−0.880	1.112	−0.213	−1.000
Egypt	−0.958	−0.398	0.632	1.359
India	0.431	−0.884	−1.724	0.943
Indonesia	0.134	−0.660	−0.600	−0.655
Israel	−1.332	0.301	0.270	1.399
Jordan	−1.377	0.244	1.208	−0.624
Netherlands	0.150	0.362	−0.908	1.380
Poland	0.385	−1.290	0.051	−0.450
U.S.S.R.	0.995	−1.220	1.610	−0.332
U.K.	1.031	1.130	0.146	−0.460
U.S.	1.949	1.649	0.910	0.828

위表에 얻어진 因子點數를 利用하여 因子 f_1 과 因子 f_2 空間에 나타낸 14個國의 特性은 다음과 같다.

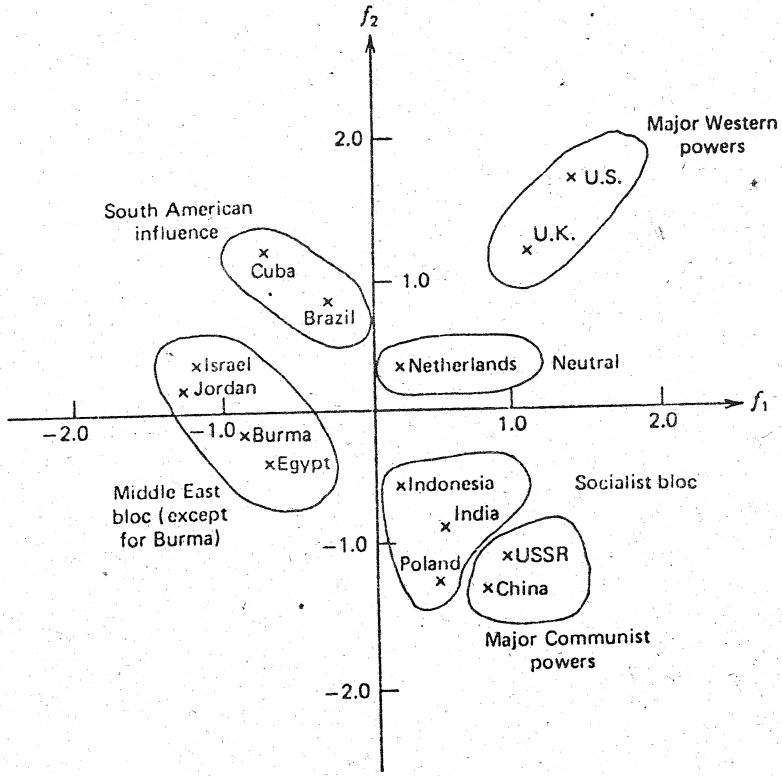


그림 1 − 4. 因子點數의 算定度
: 國力(f_1) vs. 美國과의 유대정도(f_2)

6. SAS PROC FACTOR의 例題

다음 例는 Harman(1976)이 LA統計區域에 있는 12個의 센서스地域을 對象으로 다음 다섯 개의 社會一經濟的要因을 調査한 資料를 가지고 Q-type 因子分析을 한 結果이다. 여기서 使用된 變數(要因)은 다음과 같다.

5個變數 : 人口數(POP), 先生1人當 學生數(SCHOOL),
就業人口(EMPLOY), 써비스業體數(SERVICES),
家屋數(HOUSE)

〈 Input 〉

```
DATA SOCECON;
  TITLE FIVE SOCIO-ECONOMIC VARIABLES;
  TITLE2 SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED;
  INPUT POP SCHOOL EMPLOY SERVICES HOUSE;
  CARDS;
5700 12.8 2500 270 25000
1000 10.9  600  10 10000
3400  8.8 1000  10  9000
3800 13.6 1700 140 25000
4000 12.8 1600 140 25000
8200  8.3 2600  60 12000
1200 11.4  400  10 16000
9100 11.5 3300  60 14000
9900 12.5 3400 180 18000
9600 13.7 3600 390 25000
9600  9.6 3300  80 12000
9400 11.4 4000 100 13000
;
PROC FACTOR DATA=SOCECON MSA SCREE RESIDUAL PREPLOT
            ROTATE=PROMAX REORDER PLOT
            OUT=FACT_ALL;
  PRIORS SMC;
  TITLE3 PRINCIPAL FACTOR ANALYSIS WITH PROMAX ROTATION;
PROC PRINT;
  TITLE3 FACTOR OUTPUT DATA SET;
```

FIVE SOCIO-ECONOMIC VARIABLES 3
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
PRINCIPAL FACTOR ANALYSIS WITH PROMAX ROTATION

INITIAL FACTOR METHOD: PRINCIPAL FACTORS

(4) PARTIAL CORRELATIONS CONTROLLING ALL OTHER VARIABLES

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
POP	1.00000	-0.54465	0.97083	0.09612	0.15871
SCHOOL	-0.54465	1.00000	0.54373	0.04996	0.64717
EMPLOY	0.97083	0.54373	1.00000	0.06689	-0.25572
SERVICES	0.09612	0.04996	0.06689	1.00000	0.59415
HOUSE	0.15871	0.64717	-0.25572	0.59415	1.00000

KAISER'S MEASURE OF SAMPLING ADEQUACY: OVER-ALL MSA = 0.57536759

(5)

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.472079	0.551588	0.486511	0.806644	0.612814

PRIOR COMMUNALITY ESTIMATES: SMC

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.968592	0.822285	0.969181	0.785724	0.847019

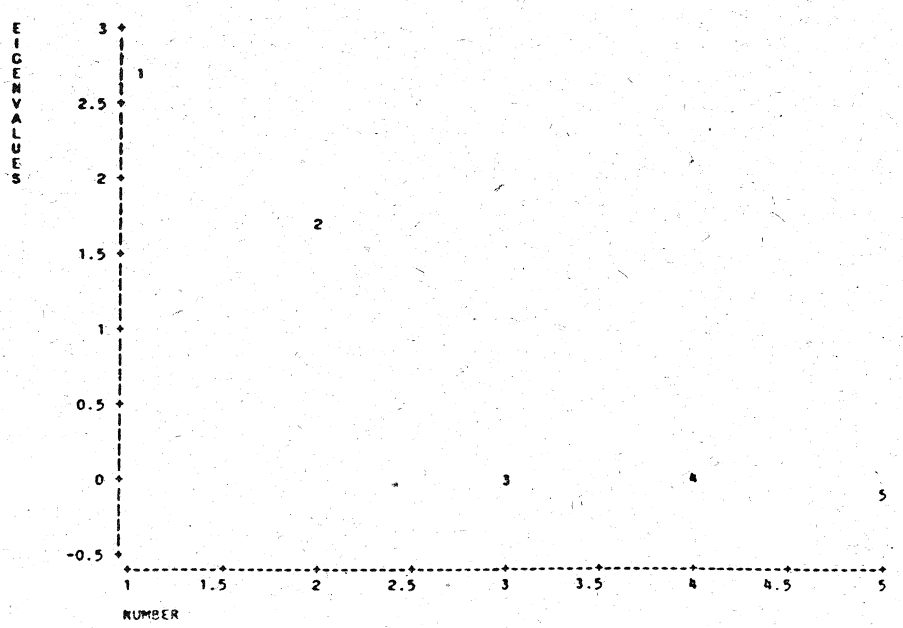
EIGENVALUES OF THE REDUCED CORRELATION MATRIX: TOTAL = 4.392801 AVERAGE = 0.878560

	1	2	3	4	5
EIGENVALUE	2.734301	1.716069	0.039563	-0.024523	-0.072608
DIFFERENCE	1.018232	1.676506	0.064086	0.048084	
PROPORTION	0.6225	0.3907	0.0090	-0.0056	-0.0165
CUMULATIVE	0.6225	1.0131	1.0221	1.0165	1.0000

2 FACTORS WILL BE RETAINED BY THE PROPORTION CRITERION

FIVE SOCIO-ECONOMIC VARIABLES 4
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
PRINCIPAL FACTOR ANALYSIS WITH PROMAX ROTATION

INITIAL FACTOR METHOD: PRINCIPAL FACTORS
SCREE PLOT OF EIGENVALUES (12)



FIVE SOCIO-ECONOMIC VARIABLES
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
PRINCIPAL FACTOR ANALYSIS WITH PROMAX ROTATION

INITIAL FACTOR METHOD: PRINCIPAL FACTORS

FACTOR PATTERN

	FACTOR1	FACTOR2
SERVICES	0.87899	-0.15847
HOUSE	0.74215	-0.57806
EMPLOY	0.71447	0.67936
SCHOOL	0.71370	-0.55515
POP	0.62533	0.76621

VARIANCE EXPLAINED BY EACH FACTOR

FACTOR1	FACTOR2
2.734301	1.716069

FINAL COMMUNALITY ESTIMATES: TOTAL = 4.450370

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.978113	0.817564	0.971999	0.797743	0.884950

(24) RESIDUAL CORRELATIONS WITH UNIQUENESS ON THE DIAGONAL

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
POP	0.02189	-0.01118	0.00514	0.01063	0.00124
SCHOOL	-0.01118	0.18244	0.02151	-0.02390	0.01248
EMPLOY	0.00514	0.02151	0.02800	-0.00565	-0.01561
SERVICES	0.01063	-0.02390	-0.00565	0.20226	0.03370
HOUSE	0.00124	0.01248	-0.01561	0.03370	0.11505

(25) ROOT MEAN SQUARE OFF-DIAGONAL RESIDUALS: OVER-ALL = 0.01693282

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.008153	0.018130	0.013828	0.021517	0.019602

(26) PARTIAL CORRELATIONS CONTROLLING FACTORS

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
POP	1.00000	-0.17693	0.20752	0.15975	0.02471
SCHOOL	-0.17693	1.00000	0.30097	-0.12443	0.08614
EMPLOY	0.20752	0.30097	1.00000	-0.07504	-0.27509
SERVICES	0.15975	-0.12443	-0.07504	1.00000	0.22093
HOUSE	0.02471	0.08614	-0.27509	0.22093	1.00000

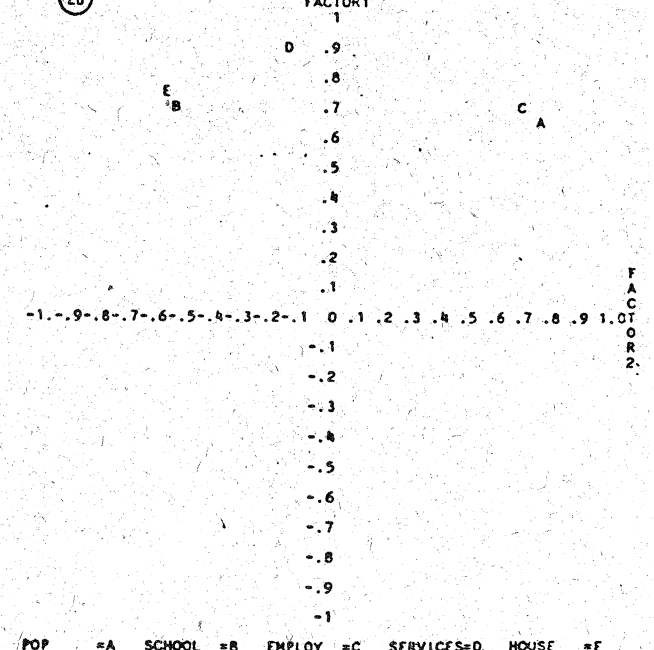
ROOT MEAN SQUARE OFF-DIAGONAL PARTIALS: OVER-ALL = 0.18550132

(27)

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.158508	0.190259	0.231818	0.154470	0.182015

PLOT OF FACTOR PATTERN FOR FACTOR1 AND FACTOR2

(28)



FIVE SOCIO-ECONOMIC VARIABLES
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
PRINCIPAL FACTOR ANALYSIS WITH PROMAX ROTATION

PREROTATION METHOD: VARIMAX

(31) ORTHOGONAL TRANSFORMATION MATRIX

	1	2
1	0.78895	0.61446
2	-0.61446	0.78895

ROTATED FACTOR PATTERN

(32)

	FACTOR1	FACTOR2
HOUSE	0.94072	-0.00004
SCHOOL	0.90419	0.00055
SERVICES	0.79085	0.41509
POP	0.02255	0.98674
EMPLOY	0.14625	0.97499

VARIANCE EXPLAINED BY EACH FACTOR

(33)

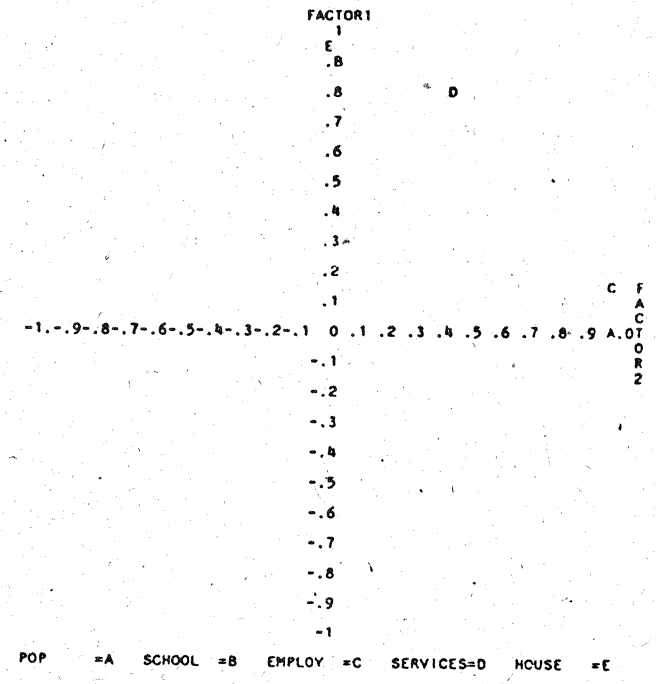
FACTOR1	FACTOR2
2.349857	2.100513

FINAL COMMUNALITY ESTIMATES: TOTAL = 4.450370

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.978113	0.817564	0.971999	0.797743	0.884950

FIVE SOCIO-ECONOMIC VARIABLES
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
PRINCIPAL FACTOR ANALYSIS WITH PROMAX ROTATION

PLOT OF FACTOR PATTERN FOR FACTOR1 AND FACTOR2



POP =A SCHOOL =B EMPLOY =C SERVICES=D HOUSE =E

ROTATION METHOD: PROMAX

(34) TARGET MATRIX FOR PROCRUSTEAN TRANSFORMATION

	FACTOR1	FACTOR2
HOUSE	1.00000	-0.00000
SCHOOL	1.00000	0.00000
SERVICES	0.69421	0.10045
POP	0.00001	1.00000
EMPLOY	0.00326	0.96793

PROCRUSTEAN TRANSFORMATION MATRIX

(35)

	1	2
1	1.04117	-0.09665
2	-0.10572	0.96303

OBLIQUE TRANSFORMATION MATRIX

(36)

	1	2
1	0.73803	0.54202
2	-0.70555	0.66528

INTER-FACTOR CORRELATIONS

(37)

	FACTOR1	FACTOR2
FACTOR1	1.00000	0.20188
FACTOR2	0.20188	1.00000

(38) ROTATED FACTOR PATTERN (STD REG COEFS)

	FACTOR1	FACTOR2
HOUSE	0.95558	-0.09792
SCHOOL	0.91842	-0.09352
SERVICES	0.76053	0.33932
POP	-0.07908	1.00192
EMPLOY	0.04799	0.97509

REFERENCE AXIS CORRELATIONS

(39)

	FACTOR1	FACTOR2
FACTOR1	1.00000	-0.20188
FACTOR2	-0.20188	1.00000

ROTATION METHOD: PROMAX

REFERENCE STRUCTURE (SEMIPARTIAL CORRELATIONS)

(40)

	FACTOR1	FACTOR2
HOUSE	0.93591	-0.09590
SCHOOL	0.89951	-0.09160
SERVICES	0.74487	0.33233
POP	-0.07745	0.98129
EMPLOY	0.04700	0.95501

(41) VARIANCE EXPLAINED BY EACH FACTOR ELIMINATING OTHER FACTORS

FACTOR1	FACTOR2
2.248089	2.003020

(42) FACTOR STRUCTURE (CORRELATIONS)

	FACTOR1	FACTOR2
HOUSE	0.93582	0.09500
SCHOOL	0.89954	0.09189
SERVICES	0.82903	0.46286
POP	0.12319	0.98596
EMPLOY	0.25484	0.98478

VARIANCE EXPLAINED BY EACH FACTOR IGNORING OTHER FACTORS

(43)

FACTOR1	FACTOR2
2.447349	2.202280

FINAL COMMUNALITY ESTIMATES: TOTAL = 4.650370

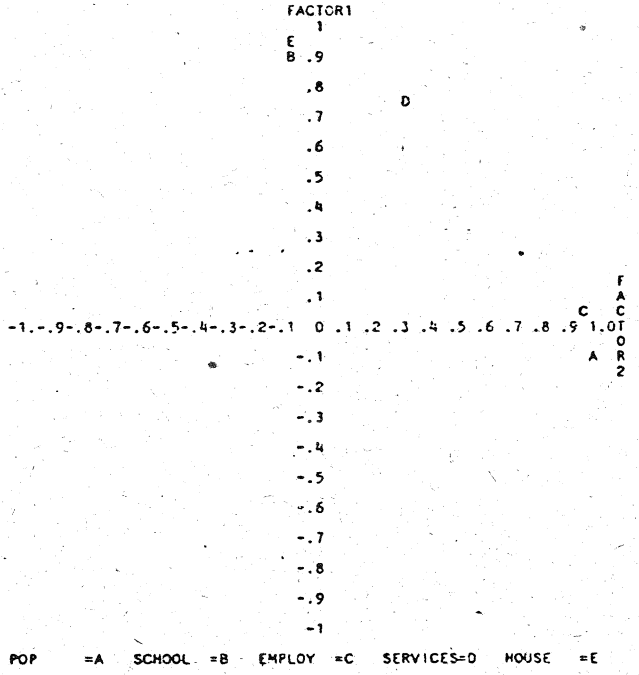
(44)

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.978113	0.817564	0.971999	0.797743	0.884950

FIVE SOCIO-ECONOMIC VARIABLES
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
PRINCIPAL FACTOR ANALYSIS WITH PROMAX ROTATION

ROTATION METHOD: PROMAX

(48) PLOT OF REFERENCE STRUCTURE FOR FACTOR1 AND FACTOR2
REFERENCE AXIS CORRELATION = -0.2019 ANGLE = 101.65 .



PART. F. MDS法(Multidimensional Scaling)

1. MDS法의 定義

多變量 資料를 壓縮, 要約하는 統計方法의 一種이다. 이 方法의 特徵은 P次元으로된 각 分析對象의 資料를 2～3次元의 導出空間下의 座標로 變換시켜 分析者가 직접 눈으로 보면서 각 分析對象들을 몇개의 그룹으로 묶거나, 各 對象의 性格을 解析할 수 있도록 한 方法이다. 여기서 導出空間(derived space)이란 MDS分析 結果로 얻어지는 작은 次元의 座標空間을 意味한다.

아래 그림은 어느 會社에서 새로 開發한 食品인 M/S candy bar의 販賣 戰略을 樹立하기 위해 이 食品의 성질을 MDS法으로 分析한 結果이다.

〈分析對象〉

Candy bar	Apple
Almonds	Yogurt
Cookie	Fruit salad
Dry roasted peanuts	Celery
Pudding	Grapefruit
Potato chips	Sucaryl
Gum	Metrecal
Ice cream	Weight Watchers diet
Milk shake	Gelatin
Coffee and danish	M/S milk shake
Cigarettes	M/S hot dog
Milk	M/S pudding
Coffee	M/S 3
Bacon	M/S 4
Sandwich	L.C. hot soup
Hot dog	L.C. chips
Soup	L.C. snacks
Codfish	L.C. cookies
Steak	M/S candy bar

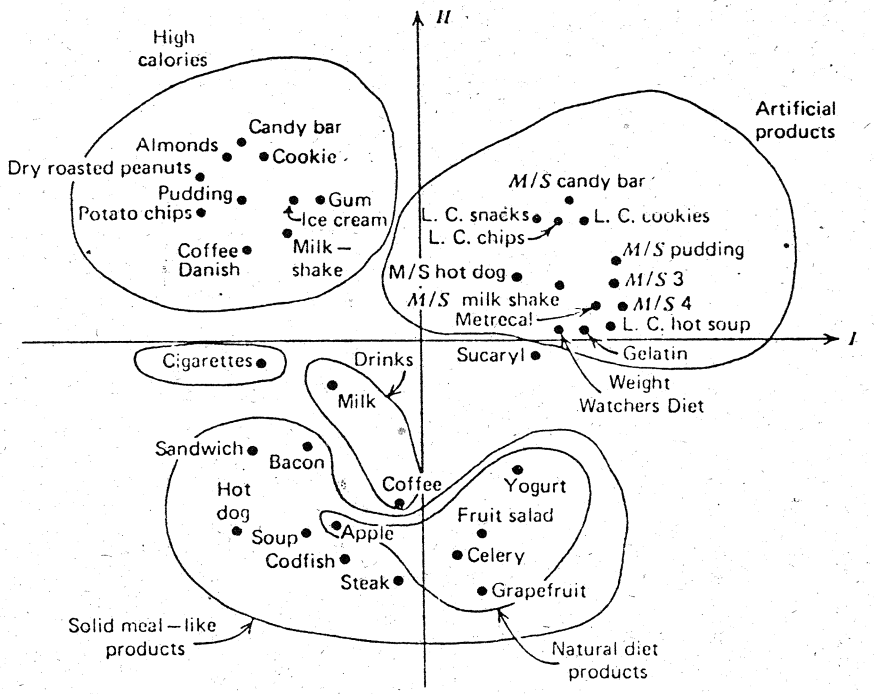


그림 (1 – 1) M/S Candy bar 및 여러食品의 多邊量 觀測값들로부터 얻은 MPS 分析結果

위의 列에서 說明한 것과 같이 MDS法은 各 分析對象에서 얻은 P次元 觀測값들을 導出空間에 座標化시켜 各 對象들의 座標로부터 分析 對象들의 유사성 정도를 把握하도록 하는 技法이다.

MDS分析 結果를 2次元 導出空間下에 나타내면 結果의 解析이 간단하지만 觀測實料에 內在된 情報構造를 正確히 把握하기 위해 不得이 높은 次元 (q > 2)의 導出空間을 使用하는 경우가 있다.

2. 資料蒐集과 近接點數의 計算

近接點數 (Proximities)란 MDS分析의 基本이 되는 入力實料로써 n개 分析對象 (object)이 이루는 ${}_nC_2$개 쌍 (pair)이 가진 類似性 或은 異質性을 數値化시킨 일종의 指數이다. 이것을 計算하기 위한 資料蒐集 方

法에는 直接算出法과 間接算出法이 있다.

A. 直接算出方法 (direct similarity method)

應答者에게 ${}_nC_2$쌍의 類似性 (similarity)을 直接 算出하게 하는 方法으로 Line Marking, Sorting, Conditional Rank order法 등이 있다.

이 중에서 가장 널리 使用되는 sorting法은 說明하기로 한다.

Sorting 法 : 이 方法의 節次는 다음과 같다.

• 節次1. n개 分析對象을 說門紙에 나열해 놓고 應答者에게 서로 비슷한 對象들을 몇개의 그룹으로 묶도록 한다. 여기서 그룹의 數는 分析者가 미리 定하는 方法과 應答者가 임의로 定하게 하는 方法이 있다.

• 節次2. 各 應答者가 만든 그룹핑을 가지고 다음과 같이 0과1의 원소로 된 정방행렬 (square matrix)을 만든다.

object	1	2	·····	n
1				
2				
⋮		0 or 1		
n				

0 : 만약 쌍의 分析對象이 서로 다른 그룹에 위치할 경우

1 : 만약 쌍의 分析對象이 같은 그룹으로 分類된 경우

• 節次3. 節次2에서 얻은 各 應答者의 정방행렬을 모두 合하여 近接點數行列 (proximity matrix)를 산출한다.

B. 間接算出方法(derived similarity method)

• 節次 1. 分析對象(object)들의 性質을 說明할 수 있는 서술적인 質疑問을 各 應答者에게 配布하여 각 질문에 1～100의 점수를 주게 한다.

例:〈담배에 대한 質問〉

	object1 솔	object2 88	object3 한라산
이 담배는 여자를 위한 것이다.	80	70	20
이 담배의 맛은 대단히 좋다.	50	80	60

• 節次 2. 節次 1에서 얻은 各 分析對象의 點數平均을 가지고 다음 公式을 利用하여 分析對象間에 近接點數行列을 算出한다.

對象 i와 對象 j間에 近接點數:

$$d_{ij} = \{ \sum_{k=1}^{P} | x_{ik} - x_{jk} |^r \}^{\frac{1}{r}} \qquad (1-1)$$

단, P:질의문의 수

x_{ik}: 對象 i가 k번째 질의문에서 얻은 應答者들의 平均點數

특히, $r=2$인 境遇 Eucledian distance가 된다.

$r=1$인 경우 City:Block metric이 된다.

위 公式에서 구한 近接點數 d_{ij}를 distance measure type이라 한다.

그리고, 이 값의 크기가 크면 對象 i와 對象 j의 類似性이 작다는 것을 나타낸다.

3. MDS 方法

MDS 方法은 資料의 性質에 따라 metric MDS法과 nonmetric MDS法으로 區分된다. metric MDS法은 近接點數行列의 算出에 使用된 資料가 量的(quantitative)인 境遇에 使用되는 것이고, 資料가 質的(qualitative)인 境遇는 nonmetric MDS法이 使用된다.

A. Metric MDS法

Metric MDS法에는 여러가지 方法이 提案되어 있으나, 여기서는 Torgerson (1952)에 의해 提案된 classical scaling法에 대해서 說明하기로 한다. 이 方法은 分析對象間에 distance measure type의 近接點數를 Eucledian 거리로 바꾸어 座標化시키는 方法으로서 그 節次는 다음과 같다.

• 節次1. n×n 近接點數行列 D의 各 원소 d_{ij}를 使用하여 다음의 행렬을 구한다.

$$B = \{ b_{ij} \}, \qquad (1-2)$$

단, $b_{ij} = -\frac{1}{2}(d_{ij}^2 - d_i^2 - d_{\cdot j}^2 + d_{\cdots}^2)$,

$$d_i^2 = \frac{1}{n}\sum_j d_{ij}^2,\ d_{\cdot j}^2 = \frac{1}{n}\sum_i d_{ij}^2,\ d_{\cdots}^2 = \frac{1}{n^2}\sum_i\sum_j d_{ij}^2$$

• 節次2. B행렬의 고유근과 고유벡터를 가지고 n개 分析對象의 座標를 구한다.

例를들어, n개 分析對象을 2次 空間下의 座標로 나타내려면, B행렬의 첫번째 고유근 $\lambda_{(1)}$과 두번째 고유근 $\lambda_{(2)}$에 對應하는 고유벡터 $X_{(1)}$과 $X_{(2)}$로 이루어진 행렬 $X = [X_{(1)}, X_{(2)}]$의 각 行을 n개 分析 對象의

導出空間(derived space) 座標로 使用한다.

$$X' = \begin{bmatrix} X_{(1)'} \\ X_{(2)'} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \cdots\cdots & X_{n1} \\ X_{12} & X_{22} & \cdots\cdots & X_{n2} \end{bmatrix}$$

첫번째 object의 座標

n번째 object의 座標

B. Nonmetric MDS法

이 方法은 質的인 資料로부터 計算된 對象間 近接點數(proximity values)가 順位의 性質을 가진것일때 使用되는 方法이다. 이 方法의 主된 分析原理는 $\binom{n}{2}$개의 對象間에 近接點數 順位와 MDS結果로 얻은 導出空間上에서 $\binom{n}{2}$개 對象을 나타내는 座標間에 거리의 順位를 同一하게 하는 것이다. 그러므로, Nonmetric MDS法에서 얻은 導出空間上에 나타난 各 對象의 座標間에 거리는 Metric MDS法과 같이 Eucledian 거리가 아니라, 각 對象間에 近接點數의 順位와 同一한 順位만 維持되도록 한 것이다.

다음 그림은 Nonmetric MDS法의 節次를 要約한 흐름도이다.

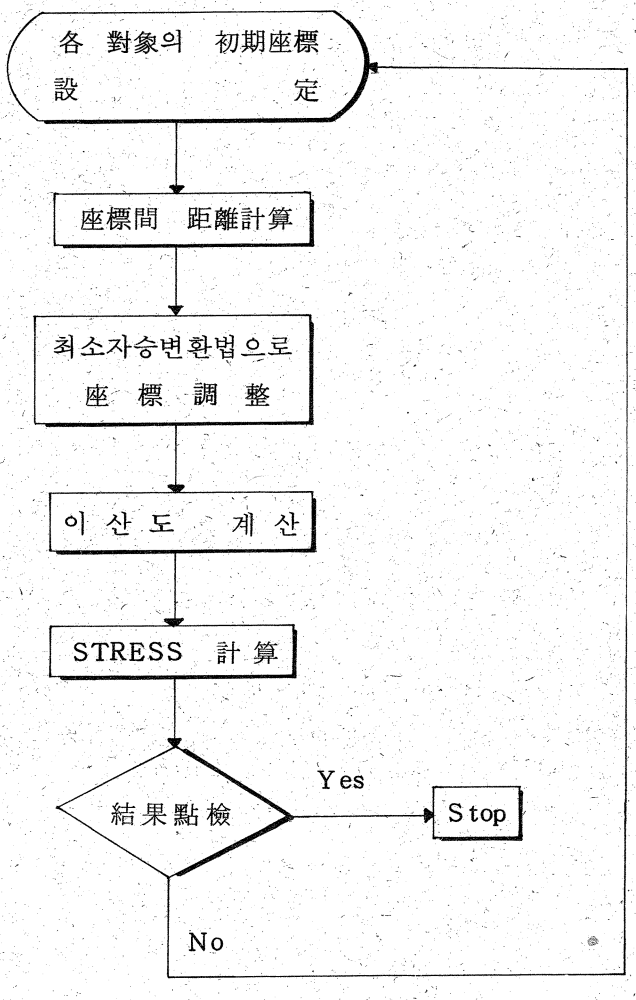


그림 (1 － 2) Nonmetric MDS 法의 흐름도

흐름도에 나타난 節次를 說明하면 다음과 같다.

• 節次 1 . 質的資料로부터 計算된 近接點數를 基礎로 n개 分析對象의 ${}_nC_2$ 쌍에 대한 近接性의 順位 (S_{ij})를 定한다.

B. 間接算出方法(derived similarity method)

• 節次1. 分析對象(object)들의 性質을 說明할 수 있는 서술적인 質疑問을 各 應答者에게 配布하여 각 질문에 1～100의 점수를 수게 한다.

例:〈담배에 대한 質問〉

	object1 솔	object2 88	object3 한라산
이 담배는 여자를 위한 것이다.	80	70	20
이 담배의 맛은 대단히 좋다.	50	80	60

• 節次2. 節次1에서 얻은 各 分析對象의 點數平均을 가지고 다음 公式을 利用하여 分析對象間에 近接點數行列을 算出한다.

對象 i 와 對象 j 間에 近接點數:

$$d_{ij} = \{ \sum_{k=1}^{P} | x_{ik} - x_{jk} |^r \}^{\frac{1}{r}} \qquad (1-1)$$

단, P:질의문의 수

x_{ik}: 對象 i 가 k번째 질의문에서 얻은 應答者들의 平均點數

특히, $r=2$인 境遇 Eucledian distance가 된다.

$r=1$인 경우 City: Block metric이 된다.

위 公式에서 구한 近接點數 d_{ij}를 distance measure type이라 한다.

그리고, 이 값의 크기가 크면 對象 i 와 對象 j 의 類似性이 작다는 것을 나타낸다.

3. MDS 方法

MDS 方法은 資料의 性質에 따라 metric MDS法과 nonmetric MDS法으로 區分된다. metric MDS法은 近接點數行列의 算出에 使用된 資料가 量的(quantitative)인 境遇에 使用되는 것이고, 資料가 質的(qualitative)인 境遇는 nonmetric MDS法이 使用된다.

A. Metric MDS法

Metric MDS法에는 여러가지 方法이 提案되어 있으나, 여기서는 Torgerson (1952)에 의해 提案된 classical scaling法에 대해서 說明하기로 한다. 이 方法은 分析對象間에 distance measure type의 近接點數를 Euclidian 거리로 바꾸어 座標化시키는 方法으로서 그 節次는 다음과 같다.

• 節次 1. n × n 近接點數行列 D의 各 원소 d_{ij}를 使用하여 다음의 행렬을 구한다.

$$B = \{ b_{ij} \}, \qquad (1-2)$$

단, $b_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i}^2 - d_{\cdot j}^2 + d_{\cdots}^2)$,

$$d_{i}^2 = \frac{1}{n}\sum_{j} d_{ij}^2, \quad d_{\cdot j}^2 = \frac{1}{n}\sum_{i} d_{ij}^2, \quad d_{\cdots}^2 = \frac{1}{n^2}\sum_{i}\sum_{j} d_{ij}^2$$

• 節次 2. B행렬의 고유근과 고유벡터를 가지고 n개 分析對象의 座標를 구한다.

例를들어, n개 分析對象을 2次 空間下의 座標로 나타내려면, B행렬의 첫번째 고유근 $\lambda_{(1)}$과 두번째 고유근 $\lambda_{(2)}$에 對應하는 고유벡터 $X_{(1)}$과 $X_{(2)}$로 이루어진 행렬 $X = [X_{(1)}, X_{(2)}]$의 각 行을 n개 分析 對象의

導出空間(derived space) 座標로 使用한다.

$$X' = \begin{bmatrix} X_{(1)}' \\ X_{(2)}' \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \cdots\cdots & X_{n1} \\ X_{12} & X_{22} & \cdots\cdots & X_{n2} \end{bmatrix}$$

첫번째 object의 座標

n번째 object의 座標

B. Nonmetric MDS法

이 方法은 質的인 資料로부터 計算된 對象間 近接點數(proximity values)가 順位의 性質을 가진것일때 使用되는 方法이다. 이 方法의 主된 分析原理는 $\binom{n}{2}$개의 對象間에 近接點數 順位와 MDS結果로 얻은 導出空間上에서 $\binom{n}{2}$개 對象을 나타내는 座標間에 거리의 順位를 同一하게 하는 것이다. 그러므로, Nonmetric MDS法에서 얻은 導出空間上에 나타난 各 對象의 座標間에 거리는 Metric MDS法과 같이 Eucledian 거리가 아니라, 각 對象間에 近接點數의 順位와 同一한 順位만 維持되도록 한 것이다.

다음 그림은 Nonmetric MDS法의 節次를 要約한 흐름도이다.

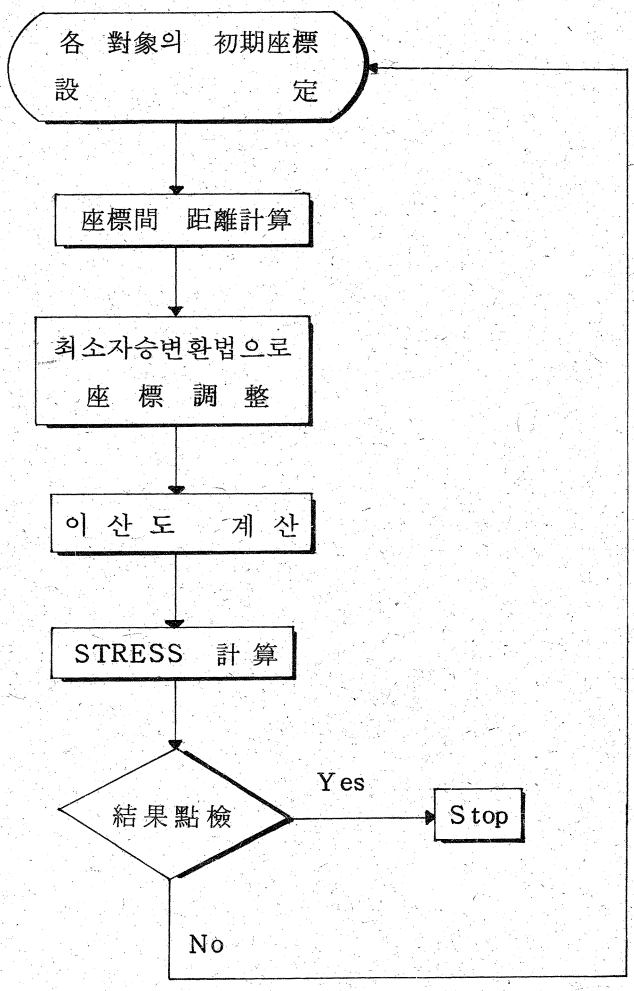


그림 (1 - 2) Nonmetric MDS 法의 흐름도

흐름도에 나타난 節次를 說明하면 다음과 같다.

• 節次 1 . 質的資料로부터 計算된 近接點數를 基礎로 n개 分析對象의 ${}_nC_2$ 쌍에 대한 近接性의 順位 (S_{ij})를 定한다.

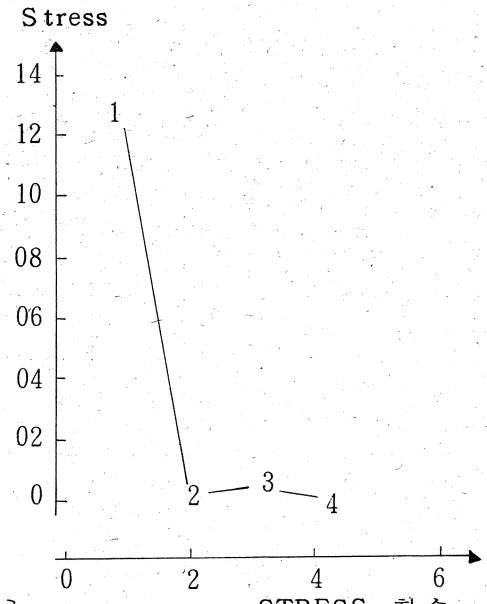
① MDS分析을 위한 컴퓨터 프로그램인 MINISSA, POLYCON, KYST, INDSCAL/SINDSCAL, ALSCAL, MULTISCALE 등에 따라 STRESS 指數값이 다르다.

② STRESS 값이 0에 가까우면(〈 0.01) n개 分析對象들이 導出空間上의 特定한 空間에 몰려 있는 狀態를 나타내므로 MDS結果가 單調性 條件을 滿足시키지 않음을 나타낸다.

③ STRESS 값은 分析對象의 數(n)와 導出空間의 차수(q)에 민감하다.

위와같은 問題點을 考慮하면서 導出空間의 차수를 選擇하기 위해서는 다음에 열거된 事項들을 綜合하여 STRESS값이 0.05以下인 차수 q 를 選擇한다.

① STRESS 함수의 檢査 : 導出空間의 차수에 따라 STRESS 指數값의 산점도(PLOT)를 그린후 산점도의 形態가 차수가 커짐에 따라 STRESS 값이 줄어드는 形態(convex形態)를 가지는지를 確認한다.



〔그림 (1-7)〕

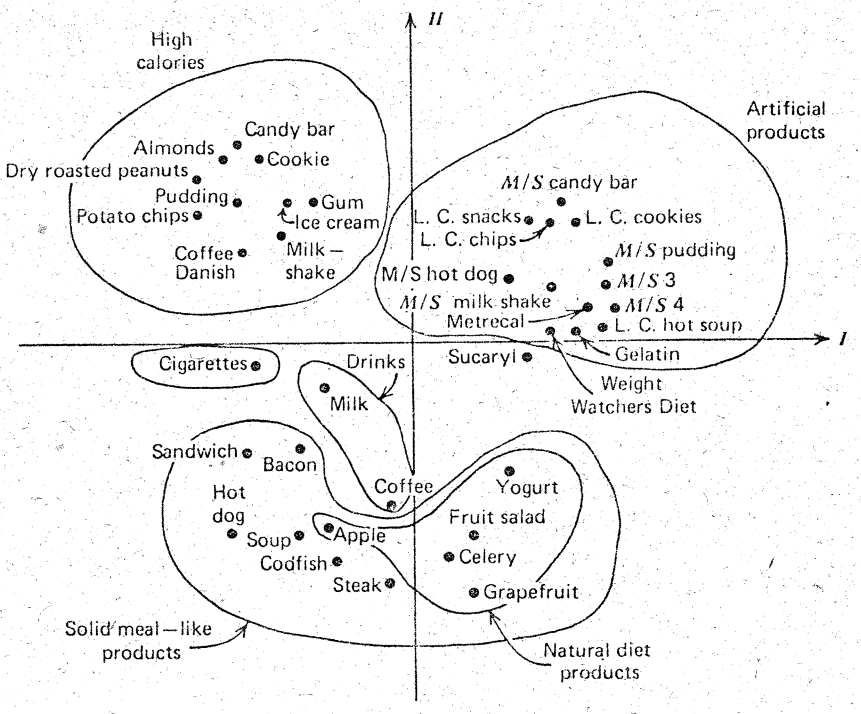
② 單調性 條件檢査：各 次元의 導出空間의 解에 해당하는 Shepard diagram 이 直線形態인지를 確認한다.

③ MDS컴퓨터 프로그램이 指定한 STRESS 指數값의 default限界를 檢査한다.

5. MDS 結果의 解析

MDS分析의 최종목표는 n개 分析對象을 몇개의 同質性 그룹(Group)으로 나누는 것과 각 分析對象 또는 그룹들의 性質을 導出空間下에서 解析하는 것이다.

첫번째 目標인 同質性 그룹化는 아래 그림과 같이 導出空間上에서 바로 얻을 수 있다.



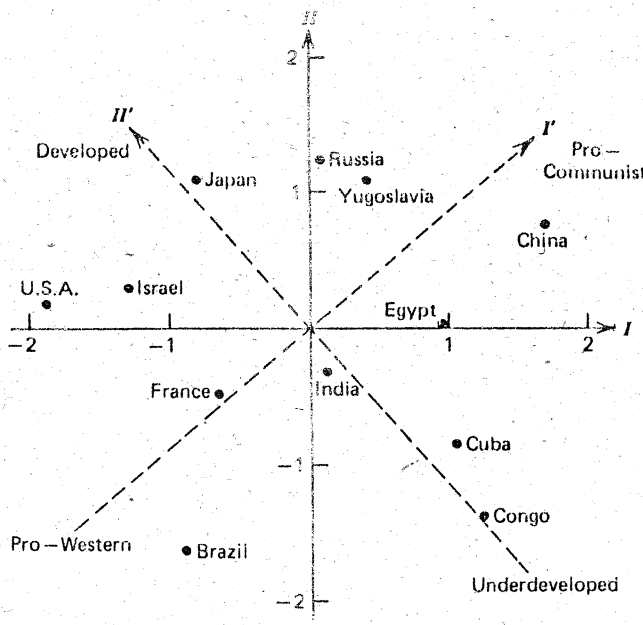
〔그림(1-8). MDS에 의한 그룹화

그러나, 導出空間의 축들은 資料의 座標化外에는 資料가 가진 情報에 連繫된 意味를 가지고 있지 않으므로 그룹의 性質 혹은 各 對象의 性質을 導出空間으로부터 直接 分析하려면 導出空間의 性質을 먼저 把握해야 한다.

導出空間의 性質을 把握하기 위해서는 補助的 統計方法을 使用하며 그 方法에는 다음의 세가지가 있다.

A. 주관적 方法 (Subjective approach)

導出空間上에 座標로 表示된 各 對象의 性格을 參考하여 導出空間上의 座標들을 解析하는 方法



〔그림 (1-9)〕 主觀的 解析法

B. 속성 벡터 (attribute vector)에 의한 解析法

속성벡터란 近接點數行列 (proximity matrix)을 計算하기 위해 만든 質疑問의 各問項 (속성 : attribute)과 導出空間과의 相關關係를 나타내는 벡

터를 意味하며 다음과 같은 節次로 구해진다.

• 節次1. 다음의 회귀모형으로부터 회귀계수벡터를 推定한다.

$$a_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots\cdots\cdots + \beta_q X_{iq} + e_i,\ i = 1, 2, \cdots, n \qquad (1-4)$$

단, a_i = i번째 分析對象이 받은 a質問의 應答點數平均,

β_j =標準化 回歸計數, $j = 1, \cdots, q$

$(X_{i1}, X_{i2}, \cdots\cdots, X_{iq})$: i번째 分析對象이 가진 導出空間上의 座標.

위의 回歸分析에서 얻어진 중상관계수(multiple correlation coefficient)는 a란 質問과 座標側사이의 相關關係를 나타낸다.

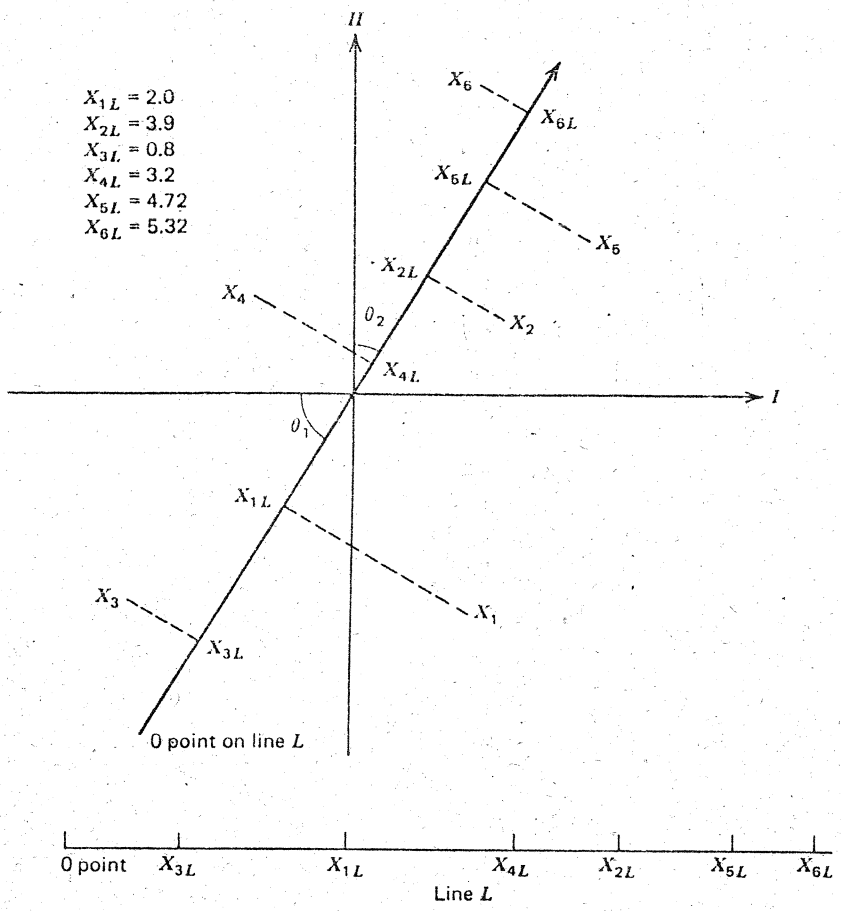
• 節次2. 回歸計數벡터의 推定값($\hat{\beta_1}, \hat{\beta_2}, \cdots, \hat{\beta_q}$)이 구해지면 이것을 利用하여 아래와 같이 導出空間上에서 속성벡터로 表示한다. 이때, 속성벡터의 길이는 重相關係數값에 비례하도록 한다.

그리고, 속성벡터의 方向角은 다음 公式으로부터 얻는다.

$$\cos\phi_i = \frac{\beta_i}{\sqrt{\sum_{j=1}^{q} \hat{\beta}_j^2}} \qquad (1-5)$$

ϕ_i = i번째 座標側과 속성벡터 사이의 角度

• 節次3. n개 分析對象을 나타내는 導出空間上의 座標들로부터 속성벡터에 수직선을 그어 얻은 속성벡터의 길이는 質問a와 各 分析對象間에 相關을 나타내므로 이를 利用하려 座標들을 解析한다.



〔그림 (1-10)〕 속 성 벡 터

C. 정준분석 (Cannonical correlation Analysis)를 利用한 解析法

座標軸을 獨立變數群으로 놓고 종속변수군인 質疑問의 各項들에 대한 정준분석을 實施하여 얻은 정준변량을 통해 座標軸들의 性質을 把握하는 方法

6. SAS PROC ALSCAL의 例題

아래表는 10個市의 類似性을 정수화시킨 近接點數 (proximity value)를 나열한 것이다. 이 資料를 SAS의 PROC ALSCAL을 利用하여 MDS分析한 結果가 아래와 같다.

〈表 10〉　　10個市의　近接點數　行列

	Atlan-ta	Chi-cago	Den-ver	Hou-ston	Los-Angel	Mi-ami	New-York	San-fran	Sea-ttle	Wash. D.C
Atlanta	-									
Chicago	4	-								
Denver	22	13	-							
Houston	8	15	12	-						
Losangel	34	31	11	24	-					
Miami	6	21	29	18	39	-				
New York	10	9	27	25	42	20	-			
Sanfran	35	32	16	28	2	44	43	-		
Seattle	36	30	19	33	17	45	40	7	-	
Wash.D.C	3	5	26	23	37	14	1	41	38	-

〈 INPUT 〉

```
DATA RANKS;
TITLE 'RANKED FLYING MILEAGES';
INPUT (ATLANTA CHICAGO DENVER HOUSTON LOSANGEL MIAMI
     NEWYORK SANFRAN SEATTLE WASHDC) (10*3.);
CARDS;

  4
 22 13
  8 15 12
 34 31 11 24
  6 21 29 18 39
 10  9 27 25 42 20
 35 32 16 28  2 44 43
 36 30 19 33 17 45 40  7
  3  5 26 23 37 14  1 41 38
;
PROC   ALSCAL
       LEVEL=RATIO
       PLOT;
TITLE2 'RATIO LEVEL, GOOD START';
PROC   ALSCAL
       LEVEL=INTERVAL
       PLOT;
TITLE2 'INTERVAL LEVEL, GOOD START';
PROC   ALSCAL
       LEVEL=ORDINAL
       PLOT;
TITLE2 'ORDINAL LEVEL, GOOD START';
DATA BADSTART;
INPUT DIM1 DIM2;
CARDS;
 1 0
 2 0
 3 0
 4 0
 5 0
 6 0
 7 0
 8 0
 9 0
10 0
;
PROC   ALSCAL DATA=RANKS
       INITIAL=BADSTART READX
       LEVEL=RATIO
       PLOT;
INVAR  DIM1 DIM2;
TITLE2 'RATIO LEVEL, BAD START';
PROC   ALSCAL DATA=RANKS
       LEVEL=INTERVAL
       INITIAL=BADSTART READX
       PLOT;
INVAR  DIM1 DIM2;
TITLE2 'INTERVAL LEVEL, BAD START';
PROC   ALSCAL DATA=RANKS
       INITIAL=BADSTART READX
       LEVEL=ORDINAL
       CONVERGE=.0001
       PLOT;
INVAR  DIM1 DIM2;
TITLE2 'ORDINAL LEVEL, BAD START';
/*
```

RANKED FLYING MILEAGES
RATIO LEVEL, GOOD START

ITERATION HISTORY FOR THE 2 DIMENSIONAL SOLUTION (IN SQUARED DISTANCES)
YOUNGS S-STRESS FORMULA 1 IS USED.

ITERATION	S-STRESS	IMPROVEMENT
1	0.06127	
2	0.05839	0.00288
3	0.05831	0.00008

ITERATIONS STOPPED BECAUSE
S-STRESS IMPROVEMENT LESS THAN 0.001000

STRESS AND SQUARED CORRELATION (RSQ) IN DISTANCES

RSQ VALUES ARE THE PROPORTION OF VARIANCE OF THE SCALED DATA (DISPARITIES) IN THE PARTITION (ROW, MATRIX, OR ENTIRE DATA) WHICH IS ACCOUNTED FOR BY THEIR CORRESPONDING DISTANCES.

STRESS VALUES ARE KRUSKAL'S STRESS FORMULA 1.

STRESS = 0.084 RSQ = 0.983

RANKED FLYING MILEAGES
RATIO LEVEL, GOOD START

CONFIGURATION DERIVED IN 2 DIMENSIONS

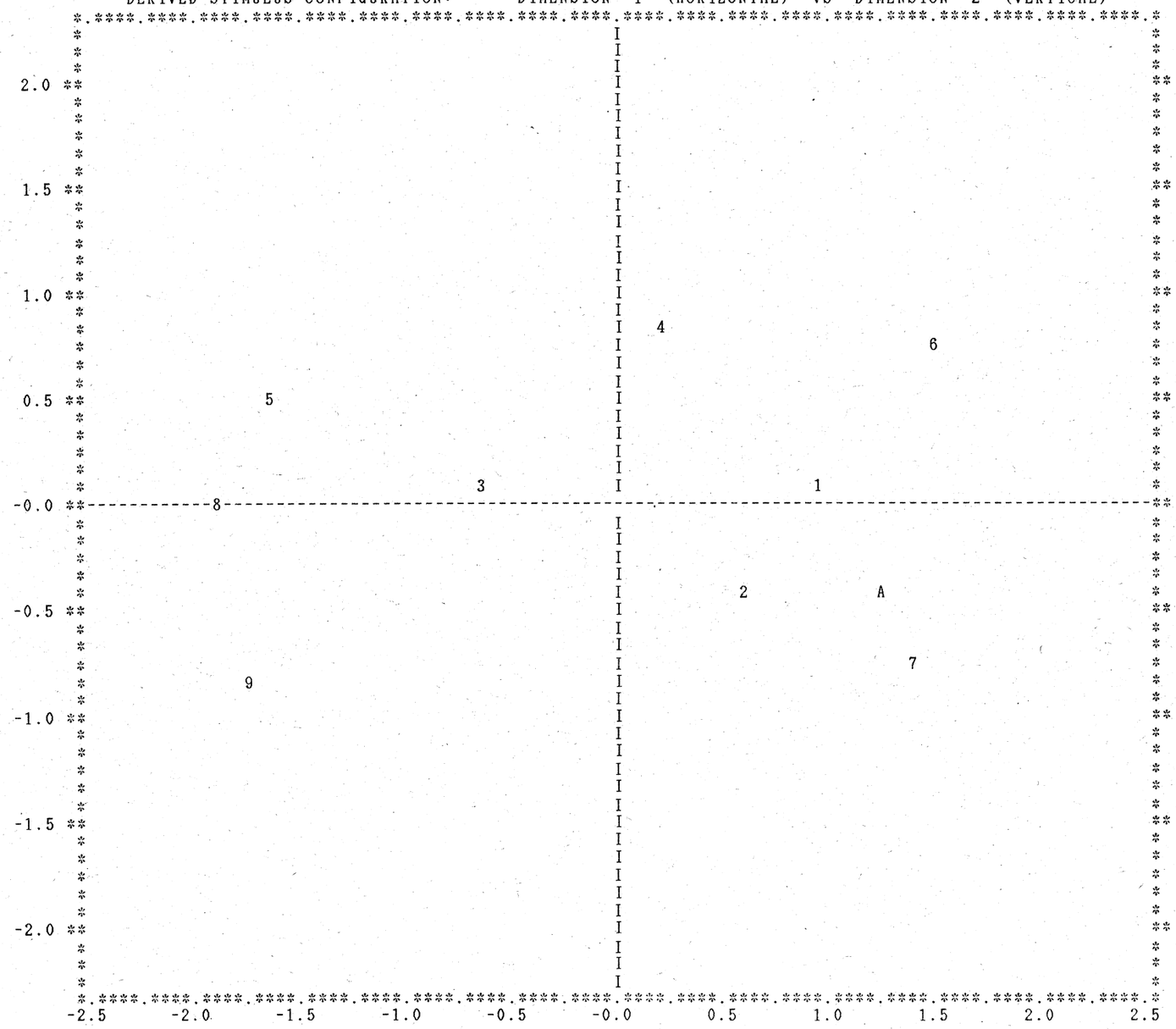
STIMULUS COORDINATES

STIMULUS NUMBER	PLOT SYMBOL	DIMENSION 1	2
1	1	0.9640	0.1217
2	2	0.5885	-0.4205
3	3	-0.6481	0.0900
4	4	0.1960	0.8458
5	5	-1.6310	0.5000
6	6	1.4867	0.7892
7	7	1.4205	-0.7210
8	8	-1.8978	0.0347
9	9	-1.7407	-0.8146
10	A	1.2619	-0.4254

RANKED FLYING MILEAGES
RATIO LEVEL, GOOD START

10:05 THURSDAY, NOVEMBER 9, 1989

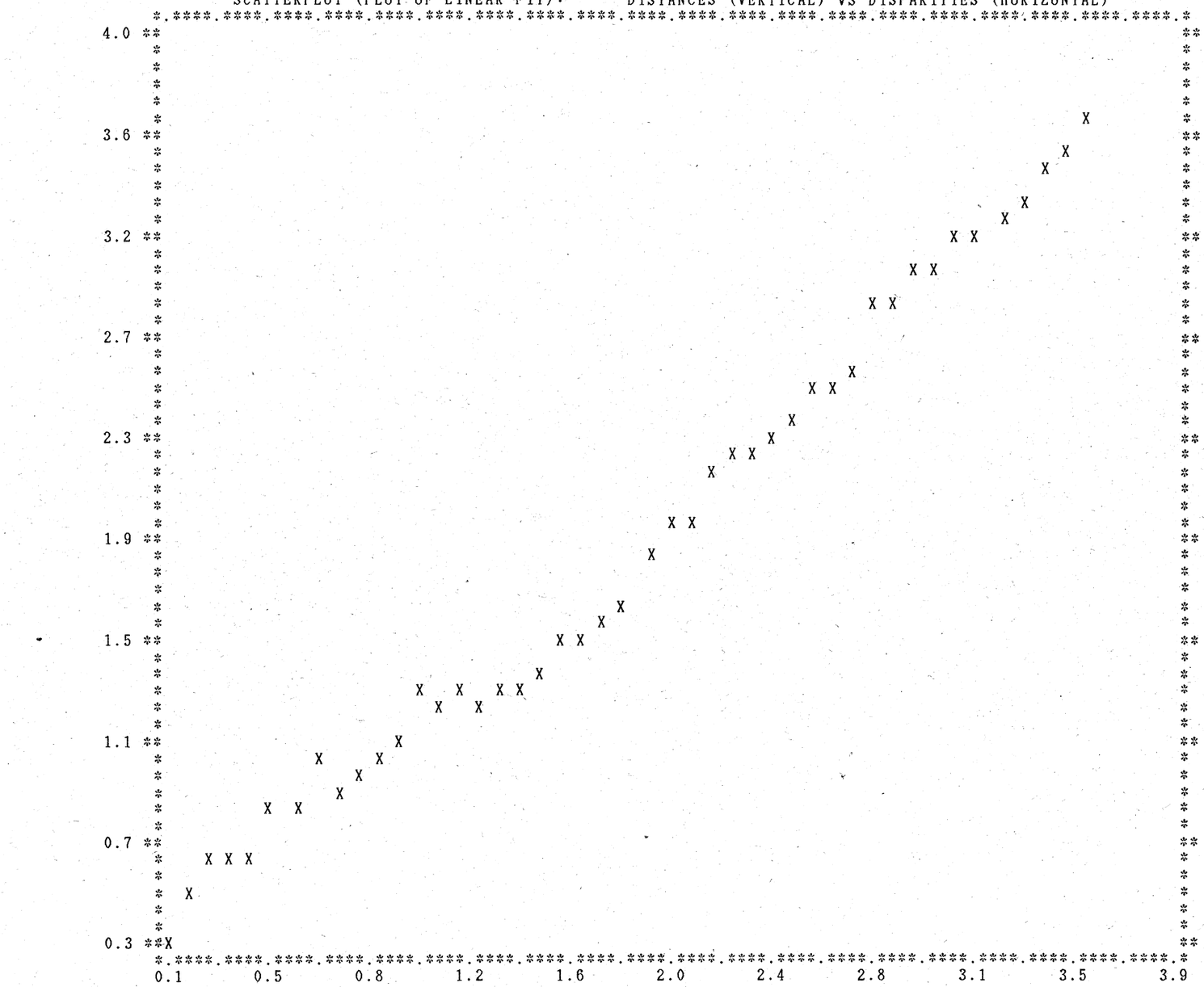
DERIVED STIMULUS CONFIGURATION: DIMENSION 1 (HORIZONTAL) VS DIMENSION 2 (VERTICAL)



RANKED FLYING MILEAGES
RATIO LEVEL, GOOD START

10:05 THURSDAY, NOVEMBER 9, 1989

SCATTERPLOT (PLOT OF LINEAR FIT): DISTANCES (VERTICAL) VS DISPARITIES (HORIZONTAL)



PART G. 集落分析 (Cluster Analysis)

1. 集落分析의 定義

集落分析이란 n개의 分析對象物이 가진 多樣한 情報를 k개의 群集(Cluster)化된 情報로 나타내는 多變量 統計資料 壓縮技法을 말한다. 卽, 多數의 對象들이 지니고 있는 多樣한 特性의 類似性을 바탕으로 이들을 몇個의 同質的인 集團으로 묶어주는 方法으로 分類學 (Taxonomy) 分野에 많이 應用된다. 例를 들어, 患者들의 症候들을 利用하여 몇개의 症候群 集團으로 分類하고 이들이 가지고 있는 공통된 特性들을 調査하거나, 動植物들이 가진 各種 生育狀態에 관한 資料를 利用하여 動食物群 (family)의 分類를 할 때도 利用될 수 있다.

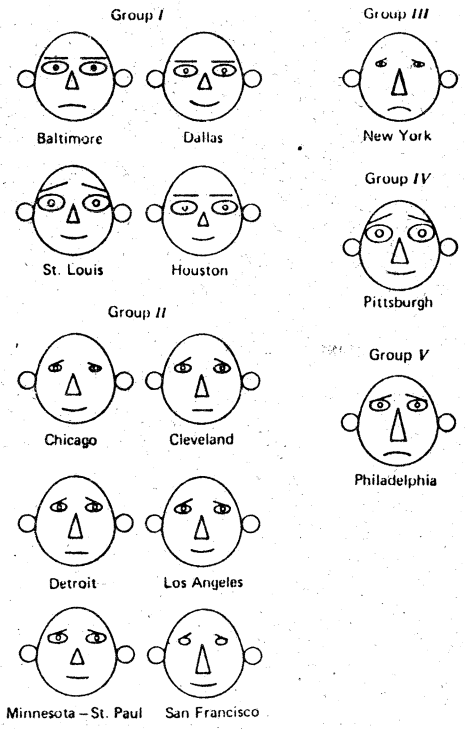


그림 (1 – 1) 集落分析의 전형적인 例 (Chernoff 얼굴)

集落分析의 節次는 다음의 세가지로 要約된다.

- 節次 1 (資料蒐集節次) : n개의 分析 對象 (object) 의 特性을 p개 變數로 觀測한다.
- 節次 2 : 蒐集된 資料로부터 n개 對象間에 $n \times n$ 類似行列 (similarity matrix) 를 求한다.
- 節次 3 : 適切히 選擇된 集落알고리즘 (cluster algorithm) 을 適用시켜 分析對象들이 가진 K개 集落 (또는 群集) 을 導出한다.

위의 節次로 構成된 集落分析法은 類似行列의 形態 및 集落 알고리즘에 따라 多樣한 方法들이 開發되어 있다.

2. 類似行列의 種類

類似行列은 n개 對象으로부터 얻은 資料의 形態에 따라 거리型 (distance type) 類似行列과 맷칭型 (maching type) 類似行列로 區分된다.

A. 거리型 類似行列

資料가 量的인 形態일 경우 使用되는 유사행렬을 말한다. n개 대상으로부터 얻은 p변량 資料벡터가 다음과 같이 量的인 形態를 가질 때

$$\begin{array}{lllll} \text{對象 1} & (X_{11}, X_{12} \cdots X_{1p})' = X_1 & & & \\ \text{對象 2} & (X_{21}, X_{22} \cdots X_{2p})' = X_2 & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{對象 } n & (X_{n1}, X_{n2} \cdots X_{np})' = X_p & & & \end{array} \qquad (1-1)$$

i번째 對象과 j번째 對象間에 類似性 (similarity) 은 두 對象으로부터

얻은 觀測값들을 하나의 距離로 換算하여 測定하게 된다.

거리의 測定方式에는 다음과 같은 것들이 있다.

① 민코우스키 메트릭(Minkowski metric);距離를 算定하는 一般式으로서 함수에 包含된 指數들을 조정해 줌으로써 多樣한 形態의 距離를 구해낼 수 있다.

i 번째 對象과 j 번째 對象間에 距離

$$d_{ij} = \{ \sum_{k=1}^{p} (X_{ik} - X_{jk})^r \}^{\frac{1}{r}} \qquad (1-2)$$

特히, $r = 2$ 인 경우 d_{ij} 를 유클리디안(Euclidean) 距離라고 한다.

② 마할라노비스 距離(Mahalanobis distance);마할라도비스에 의해 제안된 두 觀測값 사이의 距離 測定法으로 이것의 特徵은 觀測값들의 公分散을 距離測定에 連繫시킨 點이다.

i 번째 대상과 j 번째 對象의 距離

$$d_{ij} = \text{마할라노비스거리 } D_{ij} \qquad (1-3)$$

단, $D_{ij}^2 = (X_i - X_j)' S^{-1} (X_i - X_j)$

S = 統合그룹간 公分散行列

위와같이 계산된 類似型 d_{ij} 가 元素로 된 $n \times n$ 行列을 距離型 類似行列이라 한다.

〈 $n \times n$ 距離型 類似行列 〉

$$\begin{array}{c} \\ \text{對象 1} \\ \text{對象 2} \\ \vdots \\ \text{對象 } n \end{array} \begin{array}{c} \begin{array}{cccc} \text{對象 1} & \text{對象 2} & \cdots & \text{對象 } n \end{array} \\ \begin{bmatrix} d_{11} & d_{12} & & d_{1n} \\ & d_{22} & & d_{2n} \\ \text{symm} & & \ddots & \vdots \\ & & & d_{mm} \end{bmatrix} \end{array} \qquad (1-4)$$

B. 맷칭型 類似行列

觀測된 資料가 質的인 경우 使用하는 類似行列이다. 이것은 다음의 節次로 얻어진 對象間 類似性(d_{ij})을 원소로 하는 行列이다. 質的인 資料란 아래節次 1 에서 얻어진 資料形態를 말한다.

節次 1. 分析對象들의 特性을 잘 把握할 수 있도록 다음과 같이 設問紙를 만든다. 단, 設問紙의 設問은 應答者가 Yes = 1 또는 No = 0 形態로 答하도록 만든다.

〈設 問 例〉	對象 1	對象 2	……	對象 n
1. 향기가 좋다.				
2. 모양이 좋다.				
3. 색깔이 좋다.				
⋮				
p. 여자가 좋아한다.				

節次 2. N명의 應答者로 부터 얻은 資料로 부터 i번째 對象과 j번째 對象에 대한 應答의 맷칭回數 및 類型에 의해 맷칭표를 結合作成한다.

i번째 對象과 j번째 對象의 結合表(association table)는 다음의 形態를 가지고, 이 結合表에 나타난 頻度數(frequency)들은 맷칭의 類型에 따른 頻度數이다.

〈 結 合 表 〉

對象 i / 對象 j	Yes	No
Yes	a	b
No	c	d

a = 設問紙의 p개 問項中 應答者가 對象 i와 對象 j에 모두 Yes라고 對答한 問項

b = 應答者가 對象 i에는 No, 그리고 對象 j에 Yes로 對答한 問項數

c = 應答者가 對象 i에는 Yes, 그리고 對象 j에 No로 對答한 問項數

d = 모두 No로 對答한 問項數

단, $a+b+c+d=Np$.

節次 3. 節次 2에서 얻은 각 對象間에 結合表로 부터 對象 i와 對象 j의 結合係數(association coefficient)인 ϕ_{ij}를 다음의 公式으로 부터 求한다.

$$\phi_{ij}=\frac{a+d}{a+b+c+d} \quad \cdots\cdots (1-5)$$

節次 4. 結合係數 ϕ_{ij}와 맷칭係數 d_{ij}의 關係式인

對象 i와 j의 맷칭係數 :

$$d_{ij}=1-\frac{1}{\phi_{ij}} \quad \cdots\cdots (1-6)$$

을 利用하여 각 對象間에 맷칭係數(matching coefficient)를 計算하여 이를 원소(element)로 하는 맷칭型 類似行列을 式(1-4)와 같이 만든다.

3. 群集化 알고리즘

群集化 方法에는 順次的으로 群集化해 나가는 階層的 方法(hierarchical technique)과 分割的 方法(partitioning technique)이 있는데, 이 두 方法의 根本的인 差異는 階層的方法 下에서는 한번 같은 群集에 包含된 對象들은 서로다른 群集에 들어갈 수 없는 점이다. 分割的 方法에는 K-Means Clustering과 Trace法 등이 있으나 여기서는 階層的 方法만 說明하기로 한다. 階層的 方法에의 群集化過程은 距離 또 맷칭係數가 가까운 對象끼리 順次的으로 묶어가는 Agglomerative Clustering과 全體對象을 하나의 群集으로 出發하여 群集을 細分化해 나가는 Devisive Clustering이 있다. 이들에 使用되는 群集化 알고리즘의 種類는 다음과 같다.

〈알고리즘의 種類〉

- Agglomerative Clustering
  - 單一基準結合方式(single linkage)
  - 完全基準結合方式(complete linkage)
  - 平均基準結合方式(average linkage)
  - Ward의 誤差 제곱합(Ward's error sum of square method)
- Devisive clustering
  - Splinter-Average Distance Method
  - AID(automatic interaction detaction)

위와같이 多樣한 알고리즘 中에서 SAS의 PROC CLUSTER에서 使用되는 平均基準結合方式(average linkage)과 Ward法을 說明하기로 하자.

A. 平均基準結合 方式

對象間에 類似行列(similarity matrix)을 바탕으로 對象들을 群集化 시키는 過程에서 이 群集化 基準은 그림과 같이 새로운 對象이 旣存의 群集에 편입될 때 이것이 旣存의 群集內에 있는 모든 對象과의 平均距離가 가장 가까운 群集에 편입되도록 하는 方法이다.

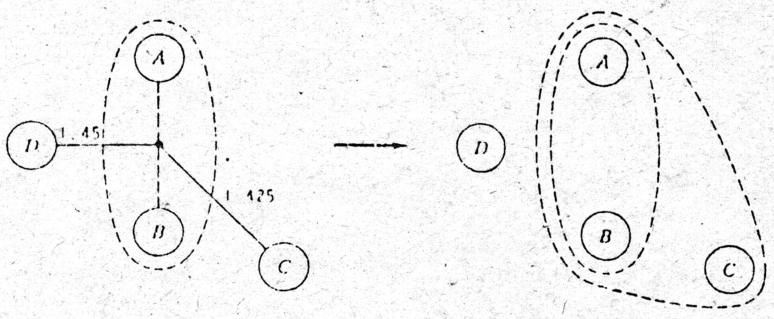


그림 (1 - 2) 平均基準 結合方式의 그림

B. Ward의 誤差제곱合法

이 方法은 量的인 資料에만 使用되며 類似行列의 計算이 必要없고, 資料로 부터 直接 群集化시키는 方法이다. 群集化基準은 旣存의 群集들에 새로운 對象이 편입될 수 있는 모든 경우에 대해서 發生하는 誤差제곱合 (E.S.S : error sum of square)들을 計算하여 이들중 最小의 誤差제곱合을 가진 群集을 골라 結合시키는 것으로 誤差제곱合은 다음과 같이 定義된다.

$$ESS = \sum_{j=1}^{k} \left(\sum_{i=1}^{nj} X_{ij}^2 - \frac{1}{n_j} \left(\sum_{i=1}^{nj} X_{ij} \right)^2 \right) \quad \cdots\cdots\cdots\cdots\cdots\cdots (1-7)$$

단, X_{ij} = j 번째 群集의 平均벡터와 i 번째 對象의 觀測벡터間에 유클리디안 距離

k = 現狀態의 群集數

n_j = j 번째 群集에 包含된 對象의 數

4. 集落分析 結果의 評價 및 留意事項

集落分析 結果가 分析對象들이 가진 性質을 잘 說明해 주는 것인지를 評價하는 作業이 最終的으로 必要하다. 結果의 評價에 많이 使用되는 評價基準으로는 Milligan(1980)과 Hubert and Levin(1976)이 각각 提案한 相關關係形態의 尺度가 있으며, 다음과 같이 定義된다.

① Milligan의 Point-biserial 一致性 尺度

$$: \frac{\bar{d}_b - \bar{d}_w}{s_d} \sqrt{f_w f_b / n_d^2} \quad \cdots\cdots (1-9)$$

단, f_w : within-cluster 距離의 數,

f_b : between-cluster 距離의 數,

$n_d = f_w + f_b$,

$\bar{d}_b$: between-cluster 距離의 平均

$\bar{d}_w$: within-cluster 距離의 平均

s_d : 모든 距離의 標準偏差

② Hubert와 Levin의 C지수

$$C = \frac{\bar{d}_w - \min d_w}{\max d_w - \min d_w} \quad \cdots\cdots (1-10)$$

단, $\min d_w$: within-cluster 距離의 最小값

$\max d_w$: within-cluster 距離의 最大값

이들 尺度는 分析者가 集落의 數(k)를 決定할 必要가 있을 때 集落의 數 決定 基準으로도 使用된다.

集落分析에서 分析者가 좋은 結果를 얻기 위해서는 다음의 事項들을 留意해야 한다.

〈集落分析의 留意事項〉

① 集落分析은 特異값(outlier)에 민감하므로 分析前에 이들을 除去시켜야 한다.

② 動植物과 같이 원래 階層的特性을 가진 對象들을 集落 分析할 경우는 階層的 方法(hierarchical method)으로 分析해야 正確한 結果를 얻을 수 있다.(例, 動植物 分類作業)

③ 對象들을 反復(m번) 測定하여 各 資料(m개)에 대해 集落分析을 實施한 後 얻은 m개의 分析結果를 比較하여 集落化(clustering)의 安定度를 檢討한다.

5. SAS PROC CLUSTER과 PROC TREE 使用例

다음 例는 PART C에서 使用한 붓꽃資料를 利用하여 Ward法으로 群集化 시킨 것이다. Ward法은 PROC CLUSTER에서 default로 內藏되어 있다. 分析結果 豫想과 같이 150개 資料가 붓꽃의 分類인 SETOSA, VERSI-COLOR, VIRGINICA의 3개 群集으로 이루어짐을 알 수 있다.

또한, PROC CLUSTER 結果를 PROC TREE를 利用하여 tree diagram (or dendrogram)으로 나타내었다.

〈 **INPUT** 〉

```
  INPUT SEPALLEN SEPALWID PETALLEN PETALWID SPEC__NO @@;
  IF SPEC__NO=1 THEN SPECIES='SETOSA';
  IF SPEC__NO=2 THEN SPECIES='VERSICOLOR';
  IF SPEC__NO=3 THEN SPECIES='VIRGINICA ';
CARDS;
50 33 14 02 1 64 28 56 22 3 65 28 46 15 2
67 31 56 24 3 63 28 51 15 3 46 34 14 03 1
69 31 51 23 3 62 22 45 15 2 59 32 48 18 2
46 36 10 02 1 61 30 46 14 2 60 27 51 16 2
65 30 52 20 3 56 25 39 11 2 65 30 55 18 3
58 27 51 19 3 68 32 59 23 3 51 33 17 05 1
57 28 45 13 2 62 34 54 23 3 77 38 67 22 3
63 33 47 16 2 67 33 57 25 3 76 30 66 21 3
49 25 45 17 3 55 35 13 02 1 67 30 52 23 3
70 32 47 14 2 64 32 45 15 2 61 28 40 13 2
48 31 16 02 1 59 30 51 18 3 55 24 38 11 2
63 25 50 19 3 64 32 53 23 3 52 34 14 02 1
49 36 14 01 1 54 30 45 15 2 79 38 64 20 3
44 32 13 02 1 67 33 57 21 3 50 35 16 06 1
58 26 40 12 2 44 30 13 02 1 77 28 67 20 3
63 27 49 18 3 47 32 16 02 1 55 26 44 12 2
50 23 33 10 2 72 32 60 18 3 48 30 14 03 1
51 38 16 02 1 61 30 49 18 3 48 34 19 02 1
50 30 16 02 1 50 32 12 02 1 61 26 56 14 3
64 28 56 21 3 43 30 11 01 1 58 40 12 02 1
51 38 19 04 1 67 31 44 14 2 62 28 48 18 3
49 30 14 02 1 51 35 14 02 1 56 30 45 15 2
58 27 41 10 2 50 34 16 04 1 46 32 14 02 1
60 29 45 15 2 57 26 35 10 2 57 44 15 04 1
50 36 14 02 1 77 30 61 23 3 63 34 56 24 3
58 27 51 19 3 57 29 42 13 2 72 30 58 16 3
54 34 15 04 1 52 41 15 01 1 71 30 59 21 3
64 31 55 18 3 60 30 48 18 3 63 29 56 18 3

49 24 33 10 2 56 27 42 13 2 57 30 42 12 2
55 42 14 02 1 49 31 15 02 1 77 26 69 23 3
60 22 50 15 3 54 39 17 04 1 66 29 46 13 2
52 27 39 14 2 60 34 45 16 2 50 34 15 02 1
44 29 14 02 1 50 20 35 10 2 55 24 37 10 2
58 27 39 12 2 47 32 13 02 1 46 31 15 02 1
69 32 57 23 3 62 29 43 13 2 74 28 61 19 3
59 30 42 15 2 51 34 15 02 1 50 35 13 03 1
56 28 49 20 3 60 22 40 10 2 73 29 63 18 3
67 25 58 18 3 49 31 15 01 1 67 31 47 15 2
63 23 44 13 2 54 37 15 02 1 56 30 41 13 2
63 25 49 15 2 61 28 47 12 2 64 29 43 13 2
51 25 30 11 2 57 28 41 13 2 65 30 58 22 3
69 31 54 21 3 54 39 13 04 1 51 35 14 03 1
72 36 61 25 3 65 32 51 20 3 61 29 47 14 2
56 29 36 13 2 69 31 49 15 2 64 27 53 19 3
68 30 55 21 3 55 25 40 13 2 48 34 16 02 1
48 30 14 01 1 45 23 13 03 1 57 25 50 20 3
57 38 17 03 1 51 38 15 03 1 55 23 40 13 2
66 30 44 14 2 68 28 48 14 2 54 34 17 02 1
51 37 15 04 1 52 35 15 02 1 58 28 51 24 3
67 30 50 17 2 63 33 60 25 3 53 37 15 02 1
;
PROC CLUSTER DATA=IRIS SIMPLE TREE=TREE;
  VAR SEPALLEN SEPALWID PETALLEN PETALWID;
PROC PRINT DATA=TREE(OBS=10);
DATA TREE;
  SET TREE;
  IF __NCL__<=15;
PROC PLOT DATA=TREE;
  PLOT __CCC__*__NCL__=__NCL__;
PROC CLUSTER DATA=IRIS;
  VAR SEPALLEN SEPALWID PETALLEN PETALWID;
  COPY SPECIES;
DATA;
  SET;
  PRERATIO=SQRT(1-__RSQ__);

PROC TREE SORT PAGES=1;
  ID SPECIES;
  HEIGHT PRERATIO;
  LABEL PRERATIO='PREDICTION RATIO';
```

〈**OUTPUT**〉

STATISTICAL ANALYSIS SYSTEM

WARD'S HIERARCHICAL CLUSTER ANALYSIS

(1) SIMPLE STATISTICS

	MEAN	STD DEV	SKEWNESS	KURTOSIS	BIMODALITY
SEPALLEN	58.43333	8.28066	0.31491	-0.55206	0.43804
SEPALWID	30.57333	4.35866	0.31897	0.22325	0.33491
PETALLEN	37.58000	17.65298	-0.27488	-1.60210	0.64622
PETALWID	11.99333	7.62238	-0.10297	-1.34060	0.58730

(2) EIGENVALUES OF THE COVARIANCE MATRIX

EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
420.0053	395.9000	0.9246	0.9246
24.1053	16.3365	0.0531	0.9777
7.7688	5.4012	0.0171	0.9948
2.3676	.	0.0052	1.0000

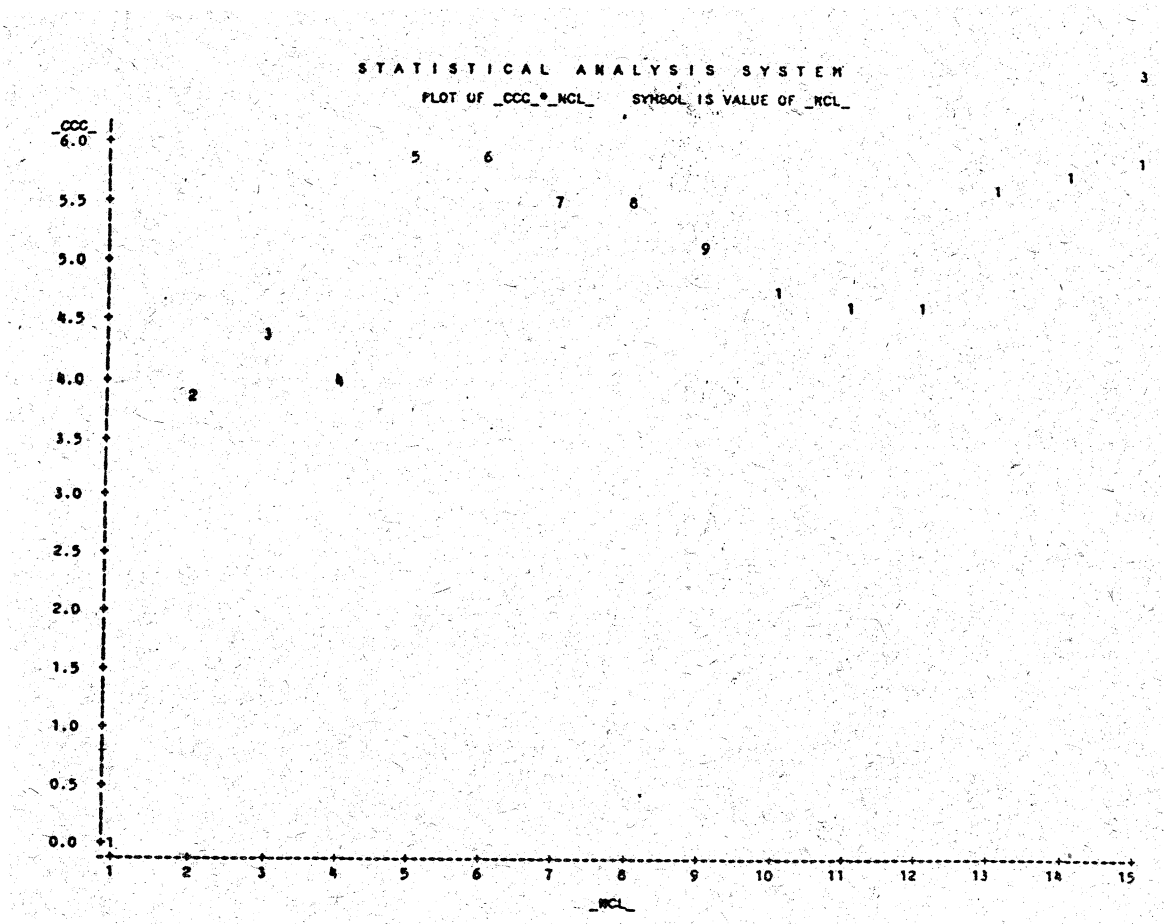
(3) ROOT-MEAN-SQUARE TOTAL-SAMPLE STANDARD DEVIATION = 10.6922

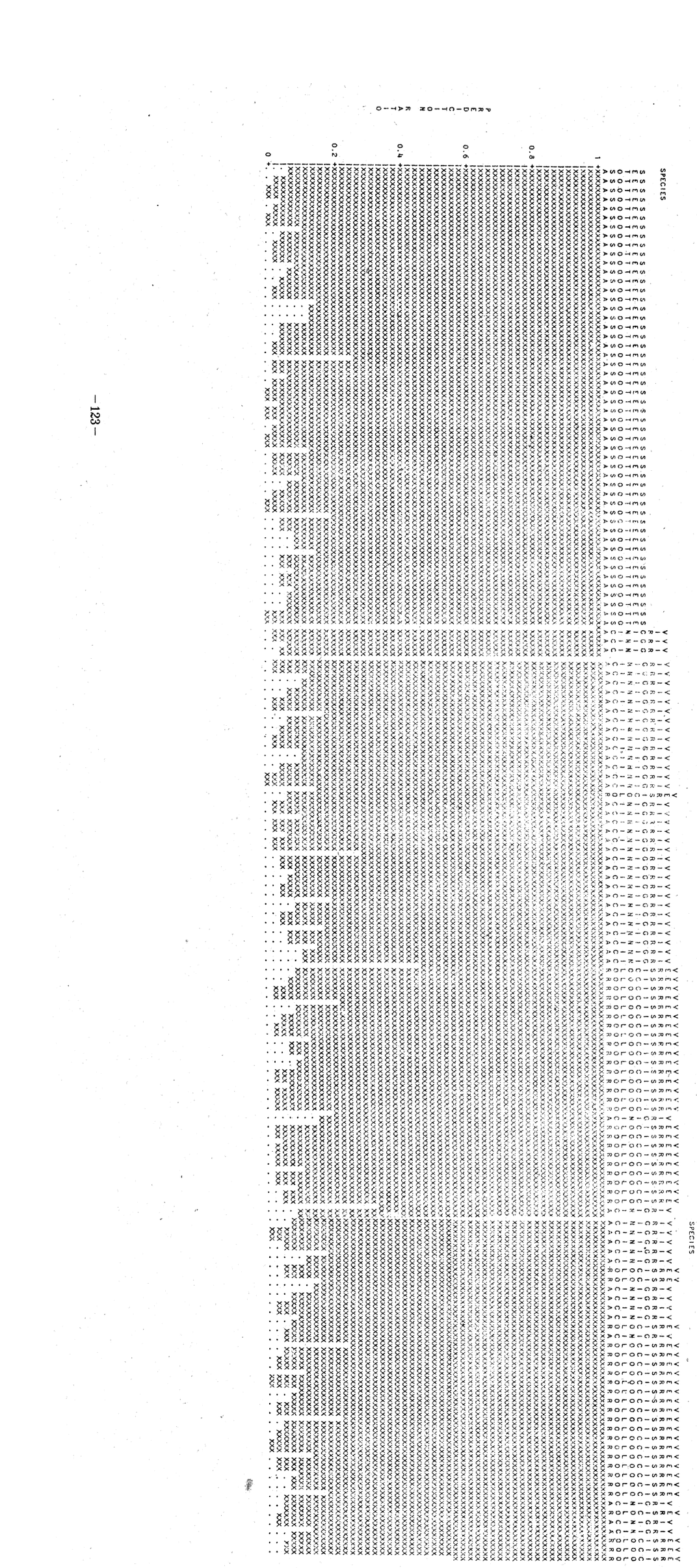
(4) ROOT-MEAN-SQUARE DISTANCE BETWEEN OBSERVATIONS = 21.3845

(5) NUMBER OF CLUSTERS	(6) FREQUENCY OF NEW CLUSTER	(7) RMS STD OF NEW CLUSTER	(8) NORMALIZED CENTROID DISTANCE	(9) NORMALIZED AVERAGE LINKAGE	(10) SEMIPARTIAL R-SQUARED	(11) R-SQUARED	(12) APPROXIMATE EXPECTED R-SQUARED	(13) CUBIC CLUSTERING CRITERION
15	15	2.15362	0.2559	0.3300	0.001641	0.970932	0.957871	5.8544
14	7	3.24771	0.4035	0.4845	0.001873	0.969059	0.955418	5.7798
13	15	2.6054	0.3186	0.3966	0.002271	0.966788	0.952570	5.6247
12	24	2.35205	0.2454	0.3501	0.002274	0.964514	0.949541	4.5819
11	12	3.25262	0.3574	0.4745	0.002500	0.962014	0.945386	4.6274
10	22	2.36245	0.2713	0.3612	0.002594	0.959320	0.941547	4.7562
9	29	2.07762	0.2392	0.3251	0.002702	0.956618	0.936296	5.0858
8	23	2.45754	0.2973	0.3780	0.003095	0.953523	0.929791	5.5068
7	25	2.99037	0.5058	0.5644	0.005811	0.947713	0.921496	5.4832
6	38	2.99017	0.3149	0.4529	0.006042	0.941671	0.910514	5.8580
5	50	2.78031	0.3627	0.4462	0.010753	0.930917	0.895232	5.8170
4	36	3.91834	0.5667	0.6726	0.017245	0.913673	0.872331	3.9657
3	64	4.11402	0.5386	0.6644	0.030051	0.883621	0.826664	4.3292
2	100	5.94155	0.8474	0.9371	0.111026	0.772595	0.696871	3.8329
1	150	10.6922	1.8584	1.9559	0.772595	0.000000	0.000000	0.0000

STATISTICAL ANALYSIS SYSTEM

OBS	NAME	PARENT	NCL	FREQ	RMSSTD	DIST	AVLINK	SPRSQ	RSQ	ERSQ	RATIO	LOGR	CCC	SEPALLEN	SEPALWID	PETALLEN	PETALWID
1	0816	CL149	150	1	0	0	0	0	1	.	.	.	.	58	27	51	19
2	0876	CL149	150	1	0	0	0	0	1	.	.	.	.	58	27	51	19
3	08116	CL148	150	1	0	0	0	0	1	.	.	.	.	54	37	15	2
4	08150	CL148	150	1	0	0	0	0	1	.	.	.	.	53	37	15	2
5	0865	CL147	150	1	0	0	0	0	1	.	.	.	.	51	35	14	2
6	08126	CL147	150	1	0	0	0	0	1	.	.	.	.	51	35	14	3
7	0889	CL146	150	1	0	0	0	0	1	.	.	.	.	49	31	15	2
8	08113	CL146	150	1	0	0	0	0	1	.	.	.	.	49	31	15	1
9	0896	CL145	150	1	0	0	0	0	1	.	.	.	.	50	34	15	2
10	08107	CL145	150	1	0	0	0	0	1	.	.	.	.	51	34	15	2


STATISTICAL ANALYSIS SYSTEM
PLOT OF _CCC_*_NCL_ SYMBOL IS VALUE OF _NCL_
CCC
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
NCL


SPECIES
PREDICTION RATIO
1
0.8
0.6
0.4
0.2
0