多 變 量 分 析

(Multivariate Analysis)

－SAS Package를 利用하여－

調 査 統 計 局

目　　次

表 目 次

圖 目 次

Ⅰ. 序　　言

現代의 많은 分野에서는 여러 變數들을 同時에 觀察·分析하여 어떤 現象을 보다 包括的이고 明確하게 糾明하려는 硏究가 增加趨勢에 있다.

硏究 目的에 따라 어떤 現象을 여러 角度에서 說明할 수 있는 많은 變數들을 調査하게 되고 또한 조사된 變數들을 特性에 따라 하나의 變數群 또는 여러 개의 變數群으로 묶어 낼 수 있다.

多變量分析 (Multivariate Analysis)이란 여러 變數들로 構成된 變數群內 또는 變數群間의 相關關係를 파악하는데 있어서 必要한 情報를 유실하지 않고 간편하게 파악할 수 있는 여러가지 統計的 分析方法들을 말한다.

本稿에서는 여러가지 多變量 統計分析方法들 중에서 몇가지 유용한 方法을 소개하고자 한다. 먼저 각 方法의 理論的 背景과 數理的 背景을 略述하고 이들 각각에 대한 SAS Package 利用方法을 說明하기로 한다. 앞으로 說明하게 될 多變量分析方法들을 分析對象에 따라 나누어 보면 다음과 같다.

〈表 1 〉　　**分析對象과 分析方法**

分　析　對　象	分　析　方　法
○ 1個 變數群內의 關係分析	○ 主成分分析 (Principal Component Analysis) ○ 要因分析 (Factor Analysis)
○ 2個 變數群間의 關係分析	○ 正準相關分析 (Canonical Correlation Analysis)

Ⅱ. 理論的 背景

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

1) 正準相關分析의 理論的 背景

어떤 硏究에서는 同一한 標本에 대하여 여러가지 變數로 構成된 두 종류의 測定群(two sets of measurements)을 얻을 때가 있다. 예를 들어 보면 어떤 標本에 대하여 10가지의 能力試驗(ability tests)와 9가지의 關心度(interest inventory scales)를 測定하여 關心과 能力간의 相互關係을 알려고 하는 경우가 있다. 물론 두가지 變數群(能力과 關心)각각으로 부터 1變數 씩 抽出하여 90個(10×9)의 二變量相關關係(bivariate correlation)를 구하여 이들 모두를 同時에 (simultaneously) 觀察·分析할 수도 있지만 關心과 能力간의 相互關係라는 擴張된 槪念을 說明하기에는 아주 어렵다.

이와 같이 同一한 對象(標本群)을 同時에 測定하였지만 2個의 變數群이 槪念的으로 서로 다른 定義域(domain)을 갖는 경우와 또한 두가지 變數群이 硏究目的에 따라 獨立 혹은 從屬의 關係를 갖는 경우가 있다. 이러한 두개의 變數群사이의 相互關係(interrelation)는 正準相關分析을 利用하여 알아 볼 수 있다. 一般的으로 흔히 볼 수 있는 二變量回歸分析과 多重回歸分析(multiple regression analysis)은 正準相關分析의 特殊한 例로서 이들의 關係를 圖表化하면 다음과 같다.

〈表 2〉 分析方法과 變數의 數

分析方法	變數	
	Y 群	X 群
正準相關分析	9 個	8 個
多重回歸分析	1 個	8 個
二變量回歸 및 相關分析	1 個	1 個

多重回歸分析(multiple regression analysis)은 스칼라 確率變數 y와 q次元 벡타 確率變數 $\underline{\chi}$와의 相關關係를 最大化시키는 問題로 생각할 수 있다. 즉, 變數 y와 $\underline{\beta}'\underline{\chi}$의 相關關係를 最大化하는 回歸係數벡타 $\underline{\beta}$를 찾는 問題라 할 수 있다.

正準相關分析은 이러한 多量回歸分析의 槪念으로 부터 스칼라 確率變數 y를 P次元 벡타 確率變數 $\underline{y}$로 擴張하여 $\underline{\alpha}'\underline{y}$와 $\underline{\beta}'\underline{\chi}$의 相關係數를 最大化하는 $\underline{\alpha}'$, $\underline{\beta}'$을 찾게 된다. 이때 $Z=\underline{\alpha}'\underline{y}$, $W=\underline{\beta}'\underline{\chi}$라 하면 Z와 W를 正準變數雙(set of canonical variable), Z와 W의 相關係數(P)를 正準相關係數(canonical correlation coefficient)라 한다. 그런데 正準變數雙은 한쌍 이상 導出할 수 있다. 즉, 먼저 最大相關을 갖는 正準變數雙(第1正準變數雙)을 導出한 후 이 正準變數雙에 無相關이며 그 다음으로 最大相關을 갖는 正準變數雙(第2正準變數雙)을 구할 수 있다. 이렇게 順次的으로 두 變數群 중 次元이 낮은 群의 變數數만큼의 正準相關雙을 구할 수 있다.

2) 正準相關分析의 數理的 背景

두개의 變數群 y와 χ가 각각 p, q個의 變數들로 이루어져 있다고 하자. 즉,

$\underline{y}' = (y_1, y_2, \cdots\cdots, y_p)$

$\underline{\chi}' = (\chi_1, \chi_2, \cdots\cdots, \chi_q)$ 이고

$$\begin{pmatrix}\underline{y}\\ \underline{\chi}\end{pmatrix} = N\left(\begin{pmatrix}\underline{\mu_1}\\ \underline{\mu_2}\end{pmatrix}, \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\right) \cdots\cdots\cdots\cdots\cdots \quad (1.1)$$

이라고 하자. 그리고 이들의 線型結合을 각각

$$Z = \underline{\alpha}'\underline{y}$$

$$W = \underline{\beta}'\underline{\chi}$$

라고 하면 $\alpha' y$와 $\beta' \chi$의 相關係數 즉,

$$P = \frac{\mathrm{cov}\,[\,\underline{\alpha}'\underline{y},\ \underline{\beta}'\underline{\chi}\,]}{[\,\mathrm{var}(\underline{\alpha}'\underline{y})\ \mathrm{var}\,(\underline{\beta}'\underline{\chi})\,]^{\frac{1}{2}}} \cdots\cdots\cdots\cdots\cdots \quad (1.2)$$

을 最大化(相關係數 P의 絶對値의 最大化)하는 $\underline{\alpha}$, $\underline{\beta}$에 대하여 Z와 W를 第1正準變數雙(the first canonical variables)이라 부르며 이때의 P를 第1正準相關係數(the first canonical correlation)라고 부른다. 式(1.2)에서 $\underline{\alpha}$, $\underline{\beta}$에 대하여 P는 0次의 同次式이므로 $\underline{\alpha}_0$, $\underline{\beta}_0$가 P^2를 最大化하였으면 임의의 常數($\neq 0$) C_1, C_2에 대하여 $C_1\underline{\alpha}_0$, $C_2\underline{\beta}_0$도 역시 P^2를 最大化하고 있다. 그러므로 이를 解決하기 위하여 다음과 같은 制限條件을 준다.

$$\mathrm{var}(Z) = \mathrm{var}\,(\underline{\alpha}'\underline{y}) = \underline{\alpha}'\,\Sigma_{11}\,\underline{\alpha} = 1\ ,$$

$$\mathrm{var}(W) = \mathrm{var}\,(\underline{\beta}'\underline{\chi}) = \underline{\beta}'\,\Sigma_{22}\,\underline{\beta} = 1 \quad \cdots\cdots\cdots\cdots (1.3)$$

式(1.3)의 條件下에서 式(1.2)를 最大化하는 問題는 다음의 L을 最大化하는 式으로 表示되며(Lagrange 表示法),

$$L = \underline{\alpha}'\,\Sigma_{12}\,\underline{\beta} + \tfrac{1}{2}\lambda_1(1 - \underline{\alpha}'\,\Sigma_{11}\,\underline{\alpha}) + \tfrac{1}{2}\lambda_2(1 - \underline{\beta}'\,\Sigma_{22}\,\underline{\beta})\ ,$$

이를 $\underline{\alpha}$, $\underline{\beta}$, λ_1, λ_2에 대하여 각각 미분하여 0으로 놓으면 다음과 같다.

$$\frac{\partial L}{\partial \underline{\alpha}} = \Sigma_{12}\underline{\beta} - \lambda_1 \Sigma_{11}\underline{\alpha} = \underline{0}, \quad \cdots\cdots (1.4)$$

$$\frac{\partial L}{\partial \underline{\beta}} = \Sigma_{21}\underline{\alpha} - \lambda_2 \Sigma_{22}\underline{\beta} = 0, \quad \cdots\cdots (1.5)$$

$$\frac{\partial L}{\partial \lambda_1} = \frac{1}{2}(1 - \underline{\alpha}'\Sigma_{11}\underline{\alpha}) = 0, \quad \cdots\cdots (1.6)$$

$$\frac{\partial L}{\partial \lambda_2} = \frac{1}{2}(1 - \underline{\beta}'\Sigma_{22}\underline{\beta}) = 0 \quad \cdots\cdots (1.7)$$

式(1.6)과 (1.7)은 단지 制限條件을 表示한 것이다. 式(1.4)에 $\underline{\alpha}'$를 곱하고 式(1.5)에 $\underline{\beta}'$를 곱하면

$$\mathrm{cov}(Z, W) = \underline{\alpha}'\Sigma_{12}\underline{\beta} = \underline{\beta}'\Sigma_{21}\underline{\alpha} = \lambda_1 = \lambda_2 = \lambda \quad (\because \Sigma_{12} = \Sigma_{21}')$$

이 成立하여 다음의 연립방정식을 푸는 問題로 歸着된다.

$$\begin{pmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{pmatrix} \begin{pmatrix} \underline{\alpha} \\ \underline{\beta} \end{pmatrix} = \underline{0} \quad \cdots\cdots (1.8)$$

式(1.8)은 係數行列의 行列式이 0이 아닐 때만 $\underline{0}$벡타 이외의 解를 갖는다. 그러므로

$$\begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix} = 0 \quad \cdots\cdots (1.9)$$

의 λ에 대한 多項方程式(polynomial equation)을 갖는데 이는

P+q個의 根이 존재한다. 따라서 각 λ에 대하여 對應되는 $\underline{\alpha}$, $\underline{\beta}$의 解을 얻을 수 있다. 式(1.5)는 正確히 Σ에 대한 特性方程式(characteristic equation of Σ)은 아니지만 분할행렬에 대한 行列式의 性質에 의하여 $P \leq q$일 때 다음 式으로 表示될 수 있다.

$$(-1)^{p}\lambda^{q-p}\ \Sigma_{22}\ \lambda^{2}\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}=0 \quad \cdots\cdots (1.10)$$

즉, Σ_{22} = 이므로

$$\lambda^{2}\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}=0 \text{이다.} \quad \cdots\cdots (1.11)$$

Σ_{11}은 正値對稱行列(positive definite symmetric matrix)이므로 어떤 正則行列을 P라 하면 $\Sigma_{11}=PP'$로 表示되며 式(1.11)은 다음 式과 같다.

$$\lambda^{2}I-P^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}(P')^{-1}=0 \quad \cdots\cdots (1.12)$$

그러므로 式(1.12)는 準正値對稱行列(positive semidefinite Symmetric matrix) $P^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}(P')^{-1}$의 λ^{2}에 대한 特性方程式이 된다. 같은 原理로 $|\mu\Sigma_{11}-(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})|=0$의 根은 0보다 크며 이는 式(1.11)과 비교하면 $1-\mu=\lambda^{2}$의 관계가 있다. 그러므로 $0\leq\mu\leq 1$이 되어 $0\leq\lambda^{2}\leq 1$인 결과가 된다. 이 사실은 $-1\leq\lambda=P\leq 1$인 사실과 一致됨을 알 수 있다. 덧붙여서 式(1.11)은 다음의 方程式들과 같음을 알 수 있다.

$$\left.\begin{array}{l}|\lambda^{2}I-\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|=0 \\ \text{혹은} \\ |\lambda^{2}I-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}|=0\end{array}\right\} (1.13)$$

絶對値가 가장 큰 λ에 대한 固有벡타 ($\underline{\alpha}'$, $\underline{\beta}'$)를 한개 구한 후 式(1.3)의 조건을 얻을 수 있게 크기를 조정하면 이것이 우리가 필요로 하는 係數벡타가 된다. 第1正準變數雙의 係數를 $\underline{\alpha}_{(1)}$, $\underline{\beta}_{(1)}$이라 하고 固有値를 $\lambda^{(1)}$이라 할 때 第1正準變數雙과 無相關인 第2正準變數雙을 定義할 수 있다.

즉, $Z_1=\underline{\alpha}'_{(1)}\underline{y}$, $\omega_1=\underline{\beta}'_{(1)}\underline{\chi}$라 하고 $Z=\underline{\alpha}'\underline{y}$, $\omega=\underline{\beta}'\underline{\chi}$로 놓으면

$\mathrm{cov}(Z_1, Z) = \mathrm{cov}(\omega_1, \omega) = 0$, $\mathrm{var}(Z) = \mathrm{var}(\omega) = 1$ 을 만족하며 $\mathrm{cov}(Z, W) = \mathrm{corr}(Z, W)$ 를 最大化하는 $\underline{\alpha}$, $\underline{\beta}$ 를 定義할 수 있다. 이 $\underline{\alpha}$, $\underline{\beta}$ 에 대한 $Z = \underline{\alpha}' y$, $\omega = \underline{\beta}' \chi$ 를 第2正準變數雙이라고 부른다. 앞에서와 마찬가지로 다음 L을 最大化하는 問題로 表示되는데

$$L = \underline{\alpha}' \Sigma_{12} \underline{\beta} + \tfrac{1}{2} \lambda_1 (1 - \underline{\alpha}' \Sigma_{11} \underline{\alpha}) + \tfrac{1}{2} \lambda_2 (1 - \underline{\beta}' \Sigma_{22} \underline{\beta}) + \mu_1 \underline{\alpha}'_{(1)} \Sigma_{11} \underline{\alpha} + \mu_2 \underline{\beta}'_{(1)} \Sigma_{22} \underline{\beta}$$

이를 $\underline{\alpha}$, $\underline{\beta}$, λ_1, λ_2, μ_1, μ_2에 대하여 미분하여 0으로 놓으면 다음과 같다.

$$\frac{\partial L}{\partial \underline{\alpha}} = \Sigma_{12} \underline{\beta} - \lambda_1 \Sigma_{11} \underline{\alpha} + \mu_1 \Sigma_{11} \underline{\alpha}_{(1)} = \underline{0} \quad \cdots\cdots \text{(1. 14)}$$

$$\frac{\partial L}{\partial \underline{\beta}} = \Sigma_{21} \alpha - \lambda_2 \Sigma_{22} \underline{\beta} + \mu_2 \Sigma_{22} \underline{\beta}_{(1)} = \underline{0} \quad \cdots\cdots \text{(1. 15)}$$

$$\frac{\partial L}{\partial \lambda_1} = \tfrac{1}{2} (1 - \underline{\alpha}' \Sigma_{11} \underline{\alpha}) = 0, \quad \cdots\cdots \text{(1. 16)}$$

$$\frac{\partial L}{\partial \lambda_2} = \tfrac{1}{2} (1 - \underline{\beta}' \Sigma_{22} \underline{\beta}) = 0, \quad \cdots\cdots \text{(1. 17)}$$

$$\frac{\partial L}{\partial \mu_1} = \underline{\alpha}'_{(1)} \Sigma_{11} \underline{\alpha} = 0, \quad \cdots\cdots \text{(1. 18)}$$

$$\frac{\partial L}{\partial \mu_2} = \underline{\beta}'^{(1)} \Sigma_{22} \underline{\beta} = 0, \quad \cdots\cdots \text{(1. 19)}$$

式(1. 14)와 (1. 15)에 각각 $\underline{\alpha}'$와 $\underline{\beta}'$를 곱하면

$$\underline{\alpha}' \Sigma_{12} \underline{\beta} = \lambda_1 \underline{\alpha}' \Sigma_{11} \underline{\alpha} - \mu_1 \underline{\alpha}' \Sigma_{11} \underline{\alpha}_{(1)},$$

$$\underline{\beta}' \Sigma_{21} \underline{\alpha} = \lambda_2 \underline{\beta}' \Sigma_{22} \underline{\beta} - \mu_2 \underline{\beta}' \Sigma_{22} \underline{\beta}_{(1)}$$

을 얻게 되어 $\lambda_1 = \lambda_2 = \lambda$ 가 된다. 또한 式(1.8)로 부터

$$\lambda^{(1)} \Sigma_{11} \underline{\alpha}_{(1)} = \Sigma_{12} \underline{\beta}_{(1)}, \quad \lambda^{(1)} \Sigma_{22} \underline{\beta}_{(1)} = \Sigma_{21} \underline{\alpha}_{(1)}$$

이 되므로 式(1. 18)과 (1. 19)로 부터 다음을 알 수 있다.

$$\underline{\alpha}' \Sigma_{12} \underline{\beta}_{(1)} = \underline{\beta}' \Sigma_{21} \underline{\alpha}_{(1)} = 0 \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (1.20)$$

그러므로 式(1. 14)와 (1. 15)에서 $\mu_1 = \mu_2 = 0$ 임을 알 수 있고 第2正準變數雙을 구하는 問題는 다음의 方程式을 푸는 問題가 된다.

$$\begin{pmatrix} -\lambda \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda \Sigma_{22} \end{pmatrix} \begin{pmatrix} \underline{\alpha} \\ \underline{\beta} \end{pmatrix} = 0$$

그러나 이 式은 式(1.8)과 正確히 같은 方程式이 되므로 여기서 우리는 絕對値가 두번째로 큰 $\lambda^{(2)}$ 와 이에 對應되는 $\underline{\alpha}_{(2)}, \underline{\beta}_{(2)}$ 를 찾게 된다.

$Z_2 = \underline{\alpha}'_{(2)} y$, $\omega_2 = \underline{\beta}'_{(2)} \underline{\chi}$ 라 할 때 式(1. 20)으로 부터

$\mathrm{cov}(Z_1, W_2) = \mathrm{cov}(Z_2, W_1) = 0$

도 아울러 성립됨을 알 수 있다.

이렇게 계속하여 $P \leqq q$ 일 때 第P正準變數雙까지 구할 수 있으며 이를 종합하면 다음과 같이 쓸 수 있다.

$$\mathrm{cov}(Z_i, Z_j) = \mathrm{cov}(W_i, W_j) = \mathrm{cov}(Z_i, W_j) = 0$$

$i \neq j = 1, \cdots\cdots, P$

$$\mathrm{var}(Z_i) = \mathrm{var}(W_i) = 1 \quad i = 1, 2 \cdots\cdots, P \quad \cdots\cdots\cdots\cdots (1.21)$$

$$\rho_i = \mathrm{cov}(Z_i, W_i) = \mathrm{corr}(Z_i, W_i) \quad i = 1, 2, \cdots\cdots, P$$

여기서 P_i 는 式(1. 11)혹은 式(1. 13)에서 구한 λ^2 에서 $P=\pm\sqrt{\lambda^2}$ 이므로 $\underline{\alpha}$, $\underline{\beta}$의 부호와 관련하여 P개의 쌍으로 구해질 수 있다. 그러므로 式(1. 10)에서 $q-P$개의 $\lambda=0$의 根과 2P의 根이 나오므로 式(1. 9)에서 λ의 根의 갯수 $q+P=(q-P)+2P$와 一致됨을 알 수 있다.

그러면 지금부터 주어진 資料로 부터 推定値를 구하는 方法을 說明하여 보자. 먼저 이제까지의 記述에서 보듯이 正準變數들을 推定하는 것은 最大法으로 Σ를 推定함으로써 解決될 수 있다.

n개의 $(\underline{y}', \underline{\chi}')$의 임의標本値(random sample)가 주어졌을 때

$$\hat{\Sigma}=\begin{pmatrix}\frac{1}{n}Y'Y-\underline{\bar{y}}\,\underline{\bar{y}}' & \frac{1}{n}Y'X-\underline{\bar{y}}\,\underline{\bar{\chi}}' \\ \frac{1}{n}X'Y-\underline{\bar{\chi}}\,\underline{\bar{y}}' & \frac{1}{n}X'X-\underline{\bar{\chi}}\,\underline{\bar{\chi}}'\end{pmatrix}=\begin{pmatrix}\hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22}\end{pmatrix}$$

$$Y=\begin{pmatrix}y_{11} & y_{12} & \cdots\cdots & y_{1p} \\ \vdots & & & \vdots \\ y_{n1} & y_{n2} & \cdots\cdots & y_{np}\end{pmatrix}=\begin{pmatrix}\underline{y}_1' \\ \vdots \\ \underline{y}_n'\end{pmatrix}$$

$$X=\begin{pmatrix}\chi_{11} & \chi_{12} & \cdots\cdots & \chi_{1q} \\ \vdots & & & \vdots \\ \chi_{n1} & \chi_{n2} & \cdots\cdots & \chi_{nq}\end{pmatrix}=\begin{pmatrix}\underline{\chi}_1' \\ \vdots \\ \underline{\chi}_n'\end{pmatrix}$$

$\underline{\bar{y}}=(\bar{y}_1, \bar{y}_2, \cdots\cdots, \bar{y}_p)$, $\underline{\bar{\chi}}'=(\bar{\chi}_1, \cdots\cdots \bar{\chi}_q)$

로 Σ가 表示되며 다음 方程式에서 正準變數들과 正準相關係數가 推定된다.

$$\begin{bmatrix} -Y_i \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & -Y_i \hat{\Sigma}_{22} \end{bmatrix} \begin{pmatrix} \underline{a}_{(i)} \\ \underline{b}_{(i)} \end{pmatrix} = 0$$

$\hat{\lambda}_{(i)} = \gamma_i$, $\hat{Z}_i = \underline{a}'_{(1)}(\underline{y}-\bar{\underline{y}})$, $\hat{W}_i = \underline{b}'_{(i)}(\underline{\chi}-\bar{\underline{\chi}})$, $i=1, \cdots\cdots, \rho$.

一般的으로는 計算이 편리하도록 分散行列 $\hat{\Sigma}$ 대신 相關係數行列 $\hat{R}$를 使用하는 데 이는 다음 式에서 그 關係를 알 수 있다.

$$\hat{R} = S \hat{\Sigma} S$$

$$S = \mathrm{diag}(1/\ell_1, \ 1/\ell_2, \cdots\cdots, \ 1/\ell_{p+q})$$

$\ell_i = (\hat{\Sigma}$의 i 번째 對角元$)^{\frac{1}{2}}$

이 $\hat{R}$를 使用한 方法은 $\underline{y}$와 $\underline{\chi}$의 標本値를 標準化하여 使用한 結果와 同一하다.

3) 正準相關係數의 有意性 檢定

母正準相關係數를 $\rho_1 \geq \rho_2 \cdots\cdots \geq \rho_R$ (단 R는 P와 q중 작은 數와 同一)이라 하고 이들의 標本推定値를 $\gamma_1 \geq \gamma_2 \geq \cdots\cdots \geq \gamma_R$이라 하자.

2節의 式(1. 11) 혹은 (1. 13)에서 구한 λ_i와 γ_i는

$$\lambda_i^2 = \gamma_i^2, \quad i = 1, 2, \cdots\cdots, R$$

의 關係가 있고 이들 $\gamma_i{}^2$은 一般的으로 決定係數(coefficient of determination)와 같은 意味를 갖는다. 그러므로 구하여진 $\lambda_i{}^2$을 이

용하여 母正準相關係數가 零이라는 歸無假說을 檢定하게 된다.

正準相關係數의 檢定은 Wilks의 Λ[1] 를 利用한 檢定統計量

$$V = -\left[n - 1 - \frac{1}{2}(P+q+1) \right] \ell_n \Lambda \quad \cdots\cdots\cdots\cdots\cdots\cdots (1.22)$$

$$(\text{단} \quad \Lambda = \prod_{i=1}^{R} (1-\lambda_i^2))$$

으로 하며, 이때 V는 近似的으로 自由度 $p \cdot q$ 인 χ^2 - 分布를 따른다고 알려져 있다. (1947. JRSS Ser.B)

먼저 第1正準相關係數 P_1의 有意性을 檢定하기 위해 歸無假說 H_0 : $P_1 = P_2 = \cdots = 0$을 세운다. 그러면 標本推定値 γ_1, 또는 λ_1을 구하여 式(1.22)에 따라 $V_1 (= -\left[n - 1 - \frac{1}{2}(P+q+1) \right] \ell n \Lambda_1 ; \Lambda_1 = \prod_{i=1}^{R} (1-\lambda_i^2))$으로 檢定하게 된다.

檢定結果 第1正準相關係數 P_i가 有意性이 인정되면(즉, 歸無假說이 기각) 第2正準相關係數 P_2의 有意性檢定을 하게 된다. 이렇게 正準相關係數들을 順次的으로 檢定하게 되는데 j번째 正準相關係數를 檢定할 때 檢定統計量 V_j는 다음과 같다.

$$V_j = -\left[n - 1 - \frac{1}{2}(p+q+1) \right] \ell n \Lambda_j ,$$

$$\Lambda_j = \prod_{i=j}^{R} (1-\lambda_i^2) .$$

이 때 V_j는 自由度 $(P-j+1)(q-j+1)$인 χ^2 - 分布에 近

---

1) Wilks의 Λ는 원래 多變量正規母集團의 圖心(centroid)에 관한 檢定法으로 提示된 것이다.(1932. Biometrika).

似的으로 따른다.

以上과 같이 檢定을 실시하여 최초로 歸無假說이 수락될 때 비로소 檢定을 끝내게 된다.

4) 事例硏究

低學年 兒童敎育에 利用되는 學生能力評價의 尺度로서 RIPIS (Rhode Island Pupil Identification Scale)이 있는데 이는 다음과 같은 두개의 部門으로 구성되어 있다.

〈表 3〉 **RIPIS 尺度의 構成**

部 門 I	部 門 II
○ 身體知覺 (body perception)	○ 符號再生 能力 (memory for reproduction of symbol)
○ 知覺神經作用 (sensory-motor coordination)	○ 方向, 位置感知力 (directional positional constancy)
○ 注意力 (attention)	○ 時空排列 能力 (spatial and sequential arrangement)
○ 自意識 (self-concept)	○ 認識作用의 符號記憶力 (memory for symbols for cognitive operations)
○ 事件記憶力 (memory for events)	

RIPIS尺度의 信賴度를 알아보기 위해서 603名의 兒童을 對象으로

12月에 한번 測定하고 翌年 5月에 同一한 對象에 다시 測定하였다. 이해를 돕기 위해서 部門Ⅱ 만을分析해 보기로 하자, 12月에 測定한 것을 χ群, 5月에 測定한 것을 y群이라 하자, 調査結果 얻어진 測定値간의 相關關係는 다음의 行列과 같다.

〈表 4〉 **部門Ⅱ의 相關關係行列(RIPIS)**

	χ_1	χ_2	χ_3	χ_4	y_1	y_2	y_3	y_4
χ_1	1.000							
χ_2	.5744	1.000						
χ_3	.6171	.6507	1.000					
χ_4	.6373	.5311	.6090	1.000				
y_1	.7921	.4398	.4776	.5184	1.000			
y_2	.4852	.7015	.5678	.4616	.5341	1.000		
y_3	.5794	.5682	.8244	.6037	.5832	.6818	1.000	
y_4	.5107	.4247	.5384	.7512	.5678	.5402	.6400	1.000
Mean	13.1003	5.3388	9.3564	10.5164	12.4934	4.8980	8.9128	40.2993
S·D	4.6627	2.3423	3.8148	3.7989	4.7494	1.9842	3.6127	4.0352

$\chi_1 = y_1$ = Memory for Reproduction of Symbols

$\chi_2 = y_2$ = Directional or Positional Constancy

$\chi_3 = y_3$ = Spatial and Sequential Arrangement

$\chi_4 = y_4$ = Memory for Symbols for Cognitive Operations

χ : 12月 測定値 y : 5月 測定値

〈表 5〉 **部門Ⅱ의 正準相關分析 結果(RIPIS)**

번호	固有값	正準相關係權	Will의 Λ	χ^2	DF	有意水準
1	0.72415	0.85097	0.06172	1664.11	16	0.0
2	0.50722	0.71220	0.22375	894.60	9	0.0
3	0.36910	0.60753	0.45406	471.75	4	0.0
4	0.28031	0.52944	0.71969	196.54	1	0.0

첫째群(12月 測定値)의 變數의 係數

	正準變量 1	正準變量 2	正準變量 3	正準變量 4
χ_1	0.26002	1.36187	0.07646	0.36225
χ_2	- 0.03059	- 0.10485	0.95487	- 0.99791
χ_3	0.60187	- 0.79731	0.18946	1.08702
χ_4	0.30584	- 0.35341	- 1.07108	- 0.76784

둘째群(5月 測定値)의 變數의 係數

	正準變量 1	正準變量 2	正準變量 3	正準變量 4
y_1	0.22933	1.27799	0.00980	0.22709
y_2	- 0.03233	- 0.09137	0.98713	- 1.00939
y_3	0.69533	- 0.69288	0.16367	1.20312
y_4	0.23380	- 0.29431	- 1.06824	- 0.78876

分析結果인 〈表5〉에서 보듯이 正準變量1에서 變數 X_3과 Y_3이 각 群內 가장 큰 係數를 갖고 있음은 時差(12月과 5月)에도 不拘하고 時差排列能力이라는 尺度는 일관된 性向임을 보여준다. 이러한 第1正準變量雙의 線型組合에서 正準相關係數는 $r_1 = 0.85097$로서 $r_1^2 = 0.724150$이므로 두개群의 正準變量雙은 全體의 약 2%정도를 說明하고 있다고 할 수 있다. 같은 方法으로 順次的으로 正準變量들을 살펴보면 RIPIS尺度의 部門Ⅱ는 時差에 關係없이 一貫된 尺度임을 確信할 수 있다. RIPIS의 部門Ⅰ 分析結果는 收錄하지 않았지만 部門Ⅱ와 같은 結果를 나타냈다.

이제 方向을 약간 바꾸어 RIPIS尺度의 部門Ⅱ와 새로운 兒童能力評價尺度인 CAT(認識適應試驗 : Cognitive Aptitude Test) 사이의 關係를 알아보기 위하여 同一한 學生을 對象으로 測定한 結果 얻어진 尺度別點數間 相關關係는 다음과 같다.

〈表 6〉 **RIPIS(部門Ⅱ)와 CAT의 相關關係**

	X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3
X_1	1.0000						
X_2	.4787	1.000					
X_3	.5033	.6351	1.000				
X_4	.5019	.5043	.6649	1.000			
Y_1	-.0687	-.0410	-.1328	-.1479	1.000		
Y_2	-.0341	.0722	-.2031	-.0925	.3831	1.000	
Y_3	-.0881	.0084	-.0837	-.1272	.8964	.3281	1.000
Mean	12.0931	4.6366	8.2310	9.5375	94.5493	60.5521	95.4930
S.D	4.4948	1.8732	3.3879	3.8166	29.5459	53.5368	30.4558

RIPIS의 部門Ⅱ

X_1 = Memory for Reproduction of Symbols

X_2 = Directional or Positional Constancy

X_3 = Spatial and Sequential Arrangement

X_4 = Memory for Symbols for Cognitive Operations

CAT

Y_1 = Verbal

Y_2 = Quantitative

Y_3 = Non-Verbal

〈表 7〉 **正準相關分析結果**

번 호	固有값	正準相關係數	WILK의 Λ	X^2	DF	有意水準
1	0.49427	0.70304	0.44913	280.15	12	0.0
2	0.08892	0.29819	0.88808	41.54	6	0.0
3	0.02524	0.15888	0.97476	8.95	2	0.01

첫째群 (X群)의 變數의 係數

	正準變量 1	正準變量 2	正準變量 3
X_1	- 0.26777	0.13541	1.15319
X_2	- 0.05446	1.23989	- 0.48646
X_3	- 0.80021	- 0.98076	- 0.48188
X_4	- 0.97538	0.21212	0.10526

둘째群 (Y群)의 變數의 係數

	正準變量 1	正準變量 2	正準變量 3
Y_1	0.09541	- 0.46468	1.98970
Y_2	0.15521	1.02671	0.12402
Y_3	- 0.13235	0.37687	- 2.22219

〈表7〉로 부터 RIPIS尺度와 CAT尺度를 比較 分析해 보자. 正準變量1에서 볼 수 있듯이 RIPIS尺度의 하나인 認識作用의 符號記憶力이 낮을 수록 CAT尺度의 Non-Verbal評價가 낮게 나타난다. 또 正準變量2에서는 時空排列能力이 낮을 수록 Verbal評價가 낮게 나타나며, 마지막으로 正準變量3에서 보면 方向位置感知力이 낮을 수록 Non-Verbal評價가 낮게 나타난다. 또한 正準變量2에서 方向位置感知力과 Quantitative評價는 서로 正의 相關關係임을 알 수 있다. 하지만 第1正準變量일때 正準相關係數가 가장 높으므로 正準變量1에 優位를 두고 分析하여야 할 것이다.

5) 分析結果의 解釋과 應用

正準相關分析結果 有意性이 認定된 K個의 正準相關係數 r_i와 對應되는 正準變量을 利用하여 硏究目的에 따라 解釋하게 된다.

먼저 正準相關이 가장 큰 r_1과 이에 相應하는 한쌍의 正準變量을 觀察한다. 그 다음 r_2, r_3 등 順次的으로 分析해 가는 過程에서 變數들의 線型關係를 把握하여 變數群間의 相互關係를 찾아 낸다. 이와 같이 각각의 正準相關係數와 正準變量을 觀察한 후 優先順位에 따라 意味를 부여하여 包括的인 結論에 도달하여야 한다.

이제까지 살펴 본 正準相關分析은 1節 序頭와 4節의 例에서와

같은 2個變數群間의 相互關係를 說明할 수 있을 뿐만 아니라 經濟現象의 變數들 간에도 應用될 수 있다. 例를 들면 經濟現象의 特性을 나타내는 많은 總量變數가 있고 또한 經濟活動을 나타내는 많은 指標(Indicator)가 주어졌다고 할 때 이 두 變數群간의 關係를 모두 알아내기는 힘들다. 그러므로 가능한 한두개의 새로운 指標를 만들어 두 變數群間의 關係를 알아낼 必要가 있다. 이런 경우에 指標變數들의 線型結合으로 나타나는 正準變數는 總量變數群과 가장 밀접한 關係를 갖는 새로운 指標로 說明될 수 있는 것이다.

2. 要因分析(Factor Analysis)

1) 要因分析의 理論的 背景

要因分析은 經濟學에서가 아닌 心理學部門에서 最初로 시작되었다. Spearman(1927)은 人間의 知的인 行動을 서로 獨立的이라고 假定한 2가지 要因 즉, 知能이라는 一般的 要因과 각각 個人의 特殊한 要因으로 說明할 수 있다는 知性의 二要因理論(The two-factor theory of intelligence)을 세우게 되었다. 그는 자신의 理論을 證明하기 위해 統計的 理論에서 부터 相關係數의 槪念을 이용하여 心理學的 實驗資料에 適用하였다. 그러나 分析結果 人間의 行爲에 대한 여러 測度들간의 關係를 說明하는 데는 Spearman의 二要因理論은 不適當하였다.

그 후 Thurstone(1947)은 共通要因(Common factors)이라는 槪念을 設定하게 되었다. 共通要因이란 어떤 變數群內에 있는 變數들 중 두개 이상의 變數들에 共存한다고 여겨지는 어떤 性質을 말한다. Spearman의 경우는 變數들의 關係를 각각의 特殊한 性格을 제외한 나머지 부분을 하나의 一般的 要因으로 把握하려고 하였지만 Thurstone은 特殊要因을 제외한 나머지 부분을 다시 몇개의 共通要因(common factors)으로 나누어 이들로써 全體變數群內의 相互關係를 說明하였다.

한편 要因(factors)이라는 用語의 意味는 要因分析者(factor analysts)와 數學者(mathematicians)들 사이에 差異가 있다. 要因分析者들은 要因을 어떤 理論的 또는 假想的 變數(theoretical or hypothetical Variable)로 使用하는 反面, 數學者들은 몇몇 變數들끼리 結合하여 하나의 結果를 갖는 단순한 數理的 形態로서 要因을 把握한다. 그런데 要因分析이 여러 硏究分野에 適用됨에 따라 要因을 理論的 혹은 假想的 變數로 보기에는 不適切한 點이 많아져 數學者들의 意味가 더욱 妥當性을 갖게 되었다.

이제 要因分析을 정의해 보면, 어떤 硏究目的下에 하나의 變數群內 여러 變數들의 相互關係를 把握하기 위해 몇개의 要因(또는 共通要因)이라고 불리워 지는 理論的 또는 假想的 變數로써 再配列하는 一連의 統計的 方法들을 要因分析이라고 한다.

2) 要因分析의 數理的 背景

觀測된 變數들을 $\underline{x}^1 = (x_1, x_2, \cdots\cdots, Xp)$라 하고 $\underline{x} \sim N(\underline{\mu}, \Sigma)$라고 하자. 그런데 分析에서는 共分散 또는 相關係數를 利用하게 되므로 一般性을 잃지 않고 $\underline{x} \sim N(\underline{O}, \Sigma)$라고 쓸 수 있다.

또한 각각의 變數들이 몇개의 共通要因과 特殊要因의 線型結合으로 보는 要因模型(factor model)은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} X_1 &= \lambda_{11} Y_1 + \cdots\cdots + \lambda_{1m} Y_m + e_1 \\ X_2 &= \lambda_{21} Y_1 + \cdots\cdots + \lambda_{2m} Y_m + e_2 \quad \cdots\cdots (2 \cdot 1) \\ &\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ X_p &= \lambda_{p1} Y_1 + \cdots\cdots + \lambda_{pm} Y_m + e_p \end{aligned}$$

Y_j : j 번째 共通要因

λ_{ij} : i 번째 變數에 있어서 j 번째 要因의 重要性을 나타내는 因數 (이를 j 번째 要因에 대한 i 번째 變數의 積載値 (loading)라 한다)

e_i : i 번째 變數의 特殊要因

式 (2 . 1)을 벡타로 表示한 要因模型은

$$\underline{x} = \Lambda \underline{y} + \underline{e} \quad \cdots\cdots (2 . 2)$$

이 되며

$$\underline{x}' = [\, X_1 X_2 \cdots\cdots X_p \,]$$

$$\underline{y}' = [\, Y_1 Y_2 \cdots\cdots Y_m \,]$$

$$\underline{e}' = [\, e_1 e_2 \cdots\cdots e_p \,]$$

$$= \begin{bmatrix} \lambda_{11} & \cdots\cdots & \lambda_{1m} \\ \vdots & & \vdots \\ \lambda_{p1} & \cdots\cdots & \lambda_{pm} \end{bmatrix}$$

이다.

이제 m個의 要因들이 서로 獨立이고 각각 N (0,1)이라고 하자. 또한 e_i 역시 獨立이고 각각 N (0 ; Ψ_i)를 한다고 假定하자. 이는 곧

$$\underline{Y}' \sim N (\underline{0} \quad I)$$

$$\underline{e}' \sim N (\underline{0} , \Psi)$$이며

$$\Psi = \begin{bmatrix} \Psi & \cdots\cdots & 0 \\ & \ddots & \\ 0 & \cdots\cdots & \Psi_p \end{bmatrix}$$ 이다.

i 번째 變數의 分散은

$$\mathrm{var}(x_i) = \sigma_i^2 = \sum_{k=1}^{m} \lambda^2_{ik} + \Psi_i \cdots\cdots (2.3)$$

이며 i 번째와 j 번째의 共分散은

$$\mathrm{COV}(x_i, y_j) = \sigma_{ij} = \sum_{k=1}^{m} \lambda_{ik} \lambda_{jk} \cdots\cdots (2.4)$$

이 된다. 式(2.3)과 (2.4)를 行列로 表示하면 다음과 같다.

$$\text{var}(\underline{x}) = \Sigma = \text{var}(\Lambda \underline{y} + \underline{e})$$

$$= \Lambda \ \text{var}(\underline{y})\, \Lambda' + \text{var}(\underline{e}) \quad (\because \underline{y}\text{와 } \underline{e}\text{는 獨立})\cdots\cdots(2.5)$$

$$= \Lambda \Lambda' + \Psi$$

式(2.3)에서 $\sum_{k=1}^{m} \lambda_{ik}^2$ 을 i 번째 變數의 「커뮤널러티」(communality) 라고 하며 Ψ_i 는 i 번째 變數의 特殊性(specificity) 또는 特殊分散(specific variance)이라고 한다.

이제 $\underline{x}$와 $\underline{y}$의 共分散을 구하여 보면

$$\text{cov}(\underline{x}, \underline{y}) = E[(\Lambda y + \varepsilon) y']$$

$$= \Lambda \text{이며}$$

λ_{ij} 는 i 번째 變數와 j 번째 要因의 共分散이 된다. 그런데 만약 Σ가 母相關係數行列이면 λ_{ij} 는 i 번째 變數와 j 번째 要因의 相關係數가 된다.

要因分析의 目的은 式(2.3)과 (2.4)를 만족하는 積載値行列 Λ를 決定하는 것이다. 이러한 係數行列 Λ 는 標本觀測値를 利用하여 最尤法으로 구할 수 있다. 그런데 구해진 Λ에 대해 만약 行列 T를 m×m 直交行列이라 하면

$$\Lambda T (\Lambda T)' + \Psi = \Lambda TT'\Lambda + \Psi$$

$$= \Lambda \Lambda' + \Psi$$

$$= \Sigma \cdots\cdots\cdots\cdots\cdots (2.6)$$

이 成立하게 된다. 즉 서로 다른 수 많은 T가 存在하므로 이 많은 $A''(=AT)$들 중에서 가장 意味있고 또한 解釋하기 쉬운 구조를 갖는 것을 選擇하게 된다. 바로 이러한 變換(transformation)을 要因回轉이라 한다.

이상의 過程을 다음 節에서 例를 통하여 알아 보자.

3) 事例研究

어떤 集團에 대해 6가지 試驗($x_1, x_2, x_3, x_4, x_5, x_6$)을 실시하였다. 理論的으로 6가지 試驗 중 (x_1, x_2, x_6)는 言語的 能力을, (x_3, x_4, x_5)는 數理的 能力을 알아 볼 수 있다고 한다. 이러한 理論을 실제 標本에 要因分析을 適用함으로써 確認해 볼 수 있다.

다음은 標本에서 구한 相關係數行列이다.

〈表 8〉 **相關係數行列**

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1.000	0.720	0.378	0.324	0.270	0.270
x_2	0.720	1.000	0.336	0.288	0.240	0.240
x_3	0.378	0.336	1.000	0.420	0.350	0.126
x_4	0.324	0.288	0.420	1.000	0.300	0.108
x_5	0.270	0.240	0.350	0.300	1.000	0.090
x_6	0.270	0.240	0.126	0.108	0.090	1.000

이 때 적용될 要因模型은 두 要因을 假定하였을 때 다음과 같이 表示된다.

$$\underline{x} = \Lambda \underline{y} + \underline{e} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2.7)$$

단 $\underline{x}' = (x_1 \cdots x_6)$

$\underline{y}' = (y_1, y_2)$

$\underline{e}' = (e_1, e_2 \cdots\cdots e_6)$

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \vdots & \vdots \\ \lambda_{61} & \lambda_{62} \end{bmatrix}$$

式 (2.3) 과 (2.4), 그리고 다음의 假定들

$\underline{y} \sim N(\underline{0}, I)$, $\rho(y_i, e_j) = 0$, $\rho(y_1, y_2) = 0$, $\rho(e_i, e_j) = 0$

을 만족하는 Λ 를 구해 보면 다음과 같다.

$$\Lambda = \begin{bmatrix} 0.889 & -0.138 \\ 0.791 & -0.122 \\ 0.501 & 0.489 \\ 0.429 & 0.419 \\ 0.358 & 0.349 \\ 0.296 & -0.046 \end{bmatrix}$$

과연 Λ 가 옳게 구해진 것인가를 確認해 보면 式 (2.4) 로 부터

$$\rho(x_1, x_2) = \lambda_{11}\lambda_{21} + \lambda_{12}\lambda_{22}$$

$$= 0.889 \times 0.791 + 0.138 \times 0.122$$

$$= 0.7200$$

을 얻으므로 〈表 8 〉에서 ρ_{12}와 같음을 볼 수 있다. 결국 例에 주어진 相關係數行列에 대한 共通要因 (common factors) f_1, f_2 가 存在함을 알 수 있다.

여기서 y_1과 y_2만이 模型에서 주어진 여러 條件을 만족하는 것은 아니다. 例로서

$$y_1^* = 0.988\, y_1 - 0.153\, y_2$$

$$y_2^* = 0.153\, y_1 + 0.988\, y_2$$

라 하면

$$E(y^*_1) = E(y^*_2) = 0, \text{var}(y_1^*), \text{var}(y_2) = 1$$ 이고

y_1과 y_2가 無相關이면 y_1^*과 y_2^*도 無相關이다. 이를 다시 쓰면

$$y_1 = 0.988\, y_1^* + 0.153\, y_2^*$$

$$y_2 = -0.153\, y_1^* + 0.988\, y_2^*$$

이것을 利用하여 Λ로 부터 다음과 같은 Λ^*를 얻을 수 있다.

$$\Lambda^* \begin{bmatrix} 0.90 & 0 \\ 0.80 & 0 \\ 0.42 & 0.56 \\ 0.36 & 0.48 \\ 0.30 & 0.40 \\ 0.30 & 0 \end{bmatrix}$$

이와 같이 또 다른 要因 y_1^* , y_2^* 을 얻을 수 있다. 이 때 Λ와 Λ 의 差異는 Λ 가 0인 要素가 많아 간단하다는 것을 알 수 있다. 즉 回歸方程式에서 0인 係數가 많은 것이 더 편리하다. 要因分析을 하는데 있어 이와 같이 가능한 한 0의 要素가 많은 係數行列 Λ를 갖는 要因들을 찾게 된다. 一般的으로 要因들을 相關되게 함으로써 보다 간단한 係數行列을 얻을 수 있다. 例로서 y^*_1 , y^*_2 로 부터

$$y_1^{**} = y_1^*$$
$$y_2^{**} = 0.6\, y_1^* + 0.8\, y_2^*$$

라고 하면 $\rho(y_1^{**}, y_2^{**}) = 0.6$ 으로 비록 두 共通要因이 서로 無相關이지는 않지만 보다 간단한 Λ^{**} 를 얻을 수 있다.

즉,

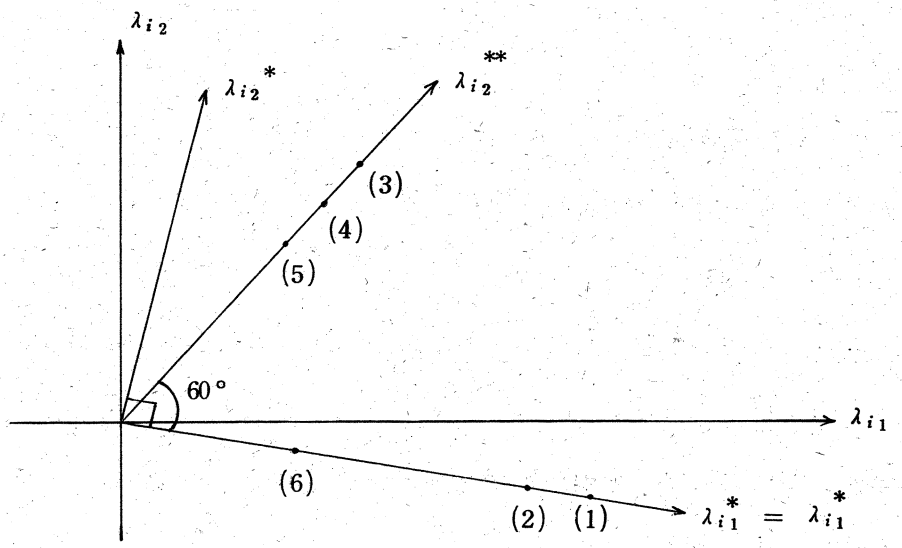
$$\Lambda^{**} = \begin{bmatrix} 0.9 & 0 \\ 0.8 & 0 \\ 0 & 0.7 \\ 0 & 0.6 \\ 0 & 0.5 \\ 0.3 & 0 \end{bmatrix}$$

이다.

行列 Λ^* 는 無相關된 要因들로서, Λ^{**} 는 相關있는 要因들로서 간단한 구조를 나타내고 있다. 이때 無相關된 要因들은 直交(orth-

ogonal), 相關된 要因들은 斜角的(oblique)이라고 하며 보통 斜角的要因(oblique factor)들이 直交要因(orthogonal factor)보다 간단한 구조를 갖는다. 이와 같이 어떤 要因群을 다른 要因群으로 變換한다는 것은 기하학적으로 軸을 回轉(rotation)하는 것이다. 위에서 본 3개의 要因群을 圖表化하면 다음과 같다.

〔 圖 1 〕 要因群의 變換 (factor rotation)



위 그림에서 보면 Λ의 各行의 要素들은 2次元 空間에서의 한 點의 座標가 된다. 또 각각의 軸들은 要因들을 나타내는데 點들은 고정되어 있는 反面, 軸인 要因들은 임의적으로 바뀌어 진다. 두 軸에 대한 사이角의 cosine은 두 要因의 相關과 같다.

이제 위의 分析結果를 살펴 보기로 한다. 먼저 Λ^{*}를 보면

y^*_1 는 6개 試驗 모두에 있어 評價할 수 있는 一般的인 要因임에 반해 y^*_2 는 y^*_1 에 관계없이 數理的能力 評價試驗(x_3 , x_4 , x_5)에 해당하는 要因임을 알 수 있다. 그러나 이러한 結論은 合理的인 것이라 할 수 없다.

Λ^{**} 에서 보면 y^{**}_1 와 y^{**}_2 는 각각 言語的, 數理的 要因이 되며 相互相關(相關係數 = 0.6)이 존재하여 言語的 能力이 良好하면 數理的 能力 역시 良好할 것이라는 結論에 도달할 수 있다. 이들 두 경우에 있어서 共分散을 살펴 보자.

直交일 때는(y^*_1 와 y^*_2)

$$\text{communality} = \lambda^2_{i1} + \lambda^2_{i2} \text{ 이고}$$

相關있는 要因일 경우는(y^{**}_1 와 y^{**}_2)

$$\text{communality} = \lambda^2_{i1} + \lambda^2_{i2} + 2\lambda_{i1}\lambda_{i2}\rho(y^{**}_1, y^{**}_2) \text{ 이다.}$$

그러나 두 경우 모두 그 結果는 同一하다. 즉 要因들이 說明하는 부분은 같다는 말이다.

다음은 위 例題의 要因分析結果로 구한 communality 와 specificity를 計算하였다.

test	communality	specificity
1	0.81	0.19
2	0.64	0.36
3	0.49	0.51
4	0.36	0.64
5	0.25	0.75
6	0.09	0.91
計	2.64	3.36

3. 主成分分析(Principal Component Analysis)

1) 主成分分析의 理論的 背景

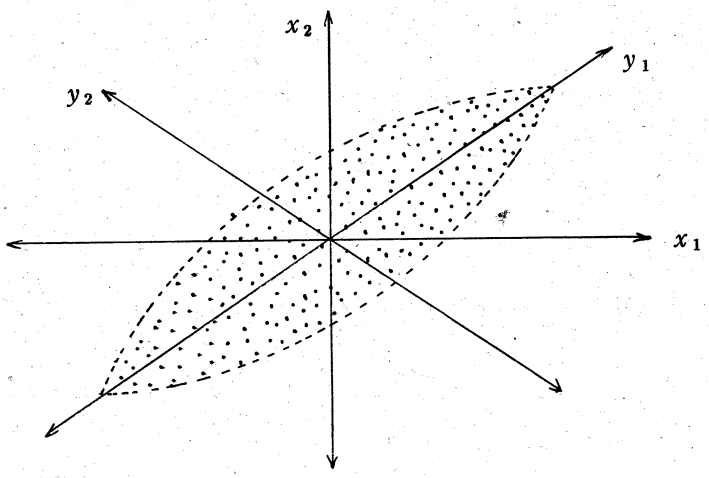
現實의 많은 變數들 간의 關係를 把握하는데 있어 次元(dimension)을 줄이고 또한 原變數들의 線型結合인 形態를 갖는 새로운 變數를 導出하는 것이 有用한 경우가 있다.

. 例를 들어, 먼저 두 確率變數 x_1 과 x_2 가 다음과 같은 分布를 한다고 하자.

$\underline{x} \sim N(\underline{\mu}, \Sigma)$ 〔단, $\underline{x}' = (x_1, x_2)$〕

이들 x_1과 x_2 의 測定値로부터 아래와 같은 形態의 分布를 얻었다.

〔圖 2〕 確率變數의 分布圖



이때 x_1 과 x_2 가 완전히 양(+)의 關係가 있다면 윗 그림에서 타원이 直線이 될 것이다. 그렇다면 이 直線은 y_1이 될 것이고 y_1은 x_1 과 x_2 의 線型結合인 $y_1 = \beta_1 x_1 + \beta_2 x_2$ 의 形態로 表現된다. 이와 같은 極端的인 경우가 아니고 윗 그림과 같은 타원의 分布일

때도 直線 y_1은 x_1과 x_2의 結合變異(joint variability)의 많은 部分을 說明할 수 있으므로 결국 y_1을 利用하여 x_1과 x_2의 分布를 把握할 수 있다.

以上의 例에서와 같은 槪念을 P次元벡타 確率變數 $x' = (x_1, x_2 \cdots\cdots x_p)$에 擴張 適用하여 確率變數 x_i들의 線型結合인 P개의 主成分을 구하게 되는데 각 主成分들은 서로 線型獨立인 관계를 維持한다.

2) 主成分分析의 數理的 背景

주어진 變數들을 $\underline{x}' = (x_1, x_2 \cdots\cdots x_p)$이고 $\underline{x}' \sim N(\underline{\mu}, \Sigma)$라고 하자. 主成分分析이란 $\underline{x}$의 線型結合의 형태를 가지면서 $\underline{x}$의 分散을 가능한 한 많이 說明하는 P개의 主成分 (principal component)을 찾는 것이다. 또한 임의의 한 主成分은 다른 모든 主成分에 線型獨立이어야 한다. 一般的으로 j번째 主成分은 다음과 같이 表現할 수 있다.

$$y_j = \underline{x}\,\underline{\beta} \cdots\cdots (3.1)$$

단, $\underline{\beta}_j' = (\beta_{1j}, \beta_{2j} \cdots\cdots, \beta_{pj})$

$\underline{x}' = (x_1, x_2, \cdots\cdots, x_p)$

만약 β_j가 주어졌다면 y_j의 分散은 다음과 같다.

$$\hat{\mathrm{var}}(y_j) = \hat{\mathrm{var}}(\underline{x}'\underline{\beta}_j) = \underline{\beta}_j' \hat{\mathrm{var}}(\underline{x})\underline{\beta}_j = \underline{\beta}_j' \Sigma \underline{\beta}_j \cdots\cdots (3.2)$$

그런데 다음과 같은 問題가 생긴다. 즉, 式(3.2)에서 $\underline{\beta}_j^* = C\underline{\beta}_j$ (C는 임의의 常數)를 $\underline{\beta}_j$ 대신에 택하면 y_j의 分散은 C에 따라 임의로 크게 될 수 있다. 이러한 不確實性을 除去하기 위해 다음과 같이 $\underline{\beta}$를 標準化한다.

$$\underline{\beta}_j'\underline{\beta}_j = \beta_{1j}^2 + \beta_{2j}^2 + \cdots\cdots + \beta_{pj}^2 = 1 \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (3.3)$$

이제 最大分散을 갖는 主成分 y_1을 찾으려면 問題는 式(3.3)의 條件下에서 y_1의 分散인 $\underline{\beta}_1' \Sigma \underline{\beta}_1$을 最大化하는 것이 된다. 이는 다시 lagrange multiplier를 利用하여

$$L = \underline{\beta}_1' \Sigma \underline{\beta}_1 - \lambda (\underline{\beta}_1'\underline{\beta}_1 - 1)$$

을 最大化하는 問題와 같다. L을 $\underline{\beta}_1$로 편미분하여 0으로 놓으면

$$\frac{\partial L}{\partial \underline{\beta}_1} = 2 \Sigma \underline{\beta}_1 - 2 \lambda \underline{\beta}_1 = 0$$

$$(\Sigma - \lambda I) \beta_1 = 0 \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (3.4)$$

$\underline{\beta}_1$가 0이 아닌 解를 갖기 위해서는

$$| \Sigma - \lambda I | = 0 \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (3.5)$$

이어야 된다. 따라서 λ는 Σ의 P개 固有根(characteristic root)중 하나가 된다.

그런데 式(3.4)에서 양변에 $\underline{\beta}'$를 곱하면

$$\underline{\beta}_1' (\Sigma - \lambda I) \underline{\beta}_1 = \underline{\beta}_1' \Sigma \underline{\beta}_1 - \lambda = 0$$

즉, $\underline{\beta}_1' \Sigma \underline{\beta}_1 = \lambda$가 된다. 結局 y_1의 分散이 λ이며 또한 最大가 되기 위해서는 λ가 $\underline{x}$의 分散 Σ의 第1固有根(the largest characteristic root)이 되어야 한다. Σ의 固有根을 $\lambda_1, \lambda_2, \cdots, \lambda_p$라고 하면 y_1의 分散은 λ_1이 되며 이때 λ_1에 對應되는 標準化된 固有벡타(characteristic vector)는 $\underline{\beta}_1$이 된다. 그러므로 第1主成分 $y_1 = \underline{x}'\underline{\beta}_1$이며 分散은 λ_1이 된다. 같은 方法으로 나머지 (P - 1)개의 主成分도 順次的으로 구할 수 있다.

각 主成分의 重要性은

$$\frac{\hat{\mathrm{var}}(y_j)}{\sum_{j=1}^{P} \hat{\mathrm{var}}(y_j)} = \frac{\lambda_j}{\sum_{j=1}^{P} \lambda_j} \quad \cdots\cdots (3.6)$$

으로 計算되며 또한

$$|\hat{\mathrm{var}}(y)| = \begin{vmatrix} \hat{\mathrm{var}}(y_1) & \bigcirc & \bigcirc & \cdots & \bigcirc \\ \bigcirc & \mathrm{var}(y_2) & \cdots & & \vdots \\ \vdots & & & & \vdots \\ \bigcirc & \cdots & & \hat{\mathrm{var}}(y_p) & \end{vmatrix}$$

$$= \prod_{j=1}^{P} \hat{\lambda}_j$$

$$= |\Sigma|$$

이다.

결국 $|\Sigma| = |\hat{\mathrm{var}}(y)|$ ····· (3.7)

임을 알 수 있다.

3) 主成分의 有意性 檢定

母共分散行列 Σ의 特性値를 $\lambda_1 \geq \lambda_2 \geq \cdots\cdots \geq \lambda_p$ 라 하고 標本共分散行列 S로부터 구한 推定値를 $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots\cdots \hat{\lambda}_p$ 라 하자.

主成分分析의 目的은 原變數들의 分散을 보다 많이 說明할 수 있는 새로운 變數의 導出이라 할 수 있는데 만약 구해진 主成分들의 重要性 즉, 全體分散 중 각각의 主成分이 說明하는 部分이 같다면 主成分을 구할 必要가 없어진다. 例로써 2變數일 때 $\lambda_1 = \lambda_2$ 라면 原變數 x_1, x_2를 主成分 y_1, y_2로 바꾸어 볼 理由가 없게 된다.

먼저 第1主成分의 重要性을 檢定하기 위해 歸無假說을 다음과 같이 세운다.

$$H_0 : \lambda_1 = \lambda_2 = \cdots\cdots = \lambda_p$$

이때 檢定統計量

$$V_1 = -(n-1)\left[\ell n \mid S \mid - P\,\ell n \frac{\mathrm{tr}\,S}{P}\right]$$

은 近似的으로 自由度 $\frac{1}{2}P(P+1)-1$인 x^2-分布에 따른다고 알려져 있다. 마찬가지로 第2, 第3의 主成分에 대한 檢定을 하게 되는데 처음 k개의 主成分으로써 充分한 部分(95%)을 說明하였다고 할 때 나머지 $P-k$개 主成分의 有意性을 檢定하게 된다. 이때 歸無假說은

$$H_0 : \lambda_{k+1} = \lambda_{k+2} = \cdots\cdots \lambda_p$$

이며 檢定統計量

$$V_{k+1} = (n-1)\left[\sum_{j=k+1}^{P} \ell n\,\hat{\lambda}_j - q\,\ell n \frac{\sum_{j=k+1}^{P} \hat{\lambda}_j}{q}\right]$$

(단, $q = P-K$)

은 近似的으로 自由度 $\frac{1}{2}q(q+1)-1$인 x^2-分布를 한다.

4) 事例研究

1900年~1910年의 美國의 3가지 生計費 指數가 다음과 같이 주어져 있다.

〈表 9〉 **Three 'Cost-of-living' Indices**

Year	(F.R.B.NY) x_1	(Hansen) x_2	(Burgress) x_3
1900	80	76	67.7
1901	82	75	70.6
1902	84	78	74.8
1903	88	81	74.8
1904	87	81	76.1
1905	87	81	76.0
1906	90	85	78.2
1907	95	90	82.0
1908	91	87	84.4
1909	91	91	88.6
1910	96	94	93.1

어떤 經濟問題의 硏究에 있어서 위와 같은 세가지 指數가 있을 때 이들 중 어떤 指數를 使用할 것인가 혹은 이 指數들 간의 線型結合의 形態를 利用할 것인가를 먼저 정하게 된다. 後者의 경우 線型結合을 組合하게 되면 가장 간단한 方法은 平均이고 또 하나는 主成分이다. 〈表 8 〉에서부터 共分散行列 S를 구한 結果는 다음과 같다.

$$S = \begin{bmatrix} 25.218 & 30.336 & 34.144 \\ 30.336 & 40.073 & 46.187 \\ 34.144 & 46.187 & 58.005 \end{bmatrix}$$

또한 標本共分散行列 S로부터 特性値와 特性벡타를 다음과 같이 구

했다.

特 性 値	特 性 벡 타		
118.44	0.4398	0.5762	0.6889
4.00	-0.7170	-0.2365	0.6557
0.85	0.5408	-0.7823	0.3091

(tr $S = \Sigma \lambda_i = 123.29$)

첫번째 主成分은

$$y_1 = 0.4398 x_1 + 0.5762 x_2 + 0.6889 x_3$$

으로 表現되며 全體分散의 96 %(118.44 / 123.29)를 說明한다. 여기의 첫번째 主成分은 주어진 세가지 生計費指數들을 하나의 指數로 묶는 意味를 갖고 사실 이들 세 系列에 있는 一般的인 分散 (generalized variance) 모두를 說明하고 있다.

5) 主成分의 解釋上 有意點

分析結果 얻어진 主成分을 線型結合 이상의 非統計的 意味를 부여하려고 할 때 많은 問題點이 야기된다.

원래 主成分 自體는 단순한 線型結合에 불과하고 硏究目的에 따른 理論的 變數는 될 수 없기 때문에 만약 意味를 주게 되면 많은 다른 硏究關聯者들간에 意見의 不一致를 초래하게 된다.

또 다른 主成分 解釋上 有意點은 資料의 原 特性과 그들의 線型獨立性이다. 또한 主成分分析을 資料로부터 假說을 낚는 낚시道具로 전락시켜서는 안될 것이다.

6) 主成分分析과 要因分析과의 差異點

어떤 變數들 사이의 關係를 把握함에 있어 變數의 數가 많아짐에 따라 그들의 實狀을 正確하게 分析하는 것은 실로 어려운 作業일 것이다. 그러므로 이 많은 變數들이 갖고 있는 情報를 유실하지 않고 보다 쉽게 把握할 수 있도록 次元(dimension)을 낮춘다면 훨씬 變數間의 關係를 理解하기가 容易할 것이다.

그런데 要因分析과 主成分分析이 혼동되는 경우가 흔히 있다. 물론 이들 두 方法은 어느 程度 類似한 點은 있지만 全的으로 다른 目的을 갖는다.

要因分析은 相關에서, 主成分分析은 分散에서부터 출발하며 要因分析은 變數들 간의 相關을 再生하는 것이 目的인데 반해 主成分分析은 全體分散을 再生하는 것이 目的이다. 또한 主成分分析에서는 몇개의 主成分으로 全體分散의 많은 部分을 說明하지만 變數들 간의 相關關係를 正確히 再生하기 위해서는 모든 主成分들이 必要하게 된다. 그러나 要因分析에서는 變數 數보다 적은 몇개의 要因으로 相關關係를 正確히 再生할 수 있다.

다른 한편으로 要因分析은 같은 갯수의 主成分에 비해 要因들이 說明할 수 있는 分散이 적은 反面, 主成分分析에서는 몇개의 主成分을 轉換(transformation)하여 똑같은 크기의 分散을 說明할 수 있는 같은 갯수의 다른 主成分을 구할 수 있다. 一般的으로 主成分分析은 分散이나 相關關係 이외의 다른 意味있는 解釋은 갖지 못하며 要因分析과의 또 다른 差異로서 主成分은 變數들의 線型結合으로 定義되지만 要因들은 變數들의 共通部分(C_1 , …… C_p)들로, 線型結合形態를 나타낼 수 있다는 點을 들 수 있다.

主成分分析은 一般的인 意味로는 模型(model)이 아니고 단순히 모든 種類의 量的인 變數(quantitative variables)들을 分析하기 위한 敍述的 分析方法(descriptive method)에 불과하지만 要因分析은 經驗的 資料에 대하여 假說檢定할 模型을 假定한다.

Ⅲ. SAS PACKAGE 利用方法 및 例題

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

SAS Package에서 이 分析을 遂行하는 Procedure는 PROC CANCORR이며 여기에서 구해지는 結果를 要約해 보면 다음과 같다.

- Canonical Correlation
- Partial Canonical Correlation
- Canonical Redundancy Analysis
- Test a Series of Hypothesis
- Canonical Coefficients
- Scores on Canonical Variables
- Plot

이 分析方法의 基本的인 段階는 正準變數들(Canonical Variables)과 그들의 相關關係를 찾는것이다. 각 變數群으로부터 變數들의 線型結合을 구하는데 이것을 正準變數双(Set of Canonical Variable)이라 하며 이때 이들의 Correlation은 最大가 되게 決定되어진다. 그러면 이 Correlation이 첫번째 正準相關이 된다. 또, 線型結合들의 係數를 正準係數(Canonical Coefficient of Canonical Weights)라 하며 보통 正準變數의 分散이 1이 되도록 係數들을 標準化시킨다. 첫번째 2개의 Canonical Variable이 만들어진후 계속해서 그 다음 正準相關을 갖는 두번째 正準變數双을 만든다. 이렇게 順次的으로 계속하여 變數數가 적은 群의 變數 갯수만큼의 正準變數双을 만든다.

각 正準變數들은 對應되는 正準變數이외의 다른 正準變數와는 無相

關(Unrelated)이며 첫번째 正準相關은 적어도 어느 群의 하나의 變數와 다른 群 變數 전부와의 多重相關(Multiple Correlation) 보다는 높아야 된다.

마지막으로 Canonical Redundancy Analysis는 正準變數(Canonical Variable)들로부터 原變數들을 얼마나 잘 豫測할 수 있는가를 알아볼 수 있다.

SAS Program Coding

PROC	CANCORR	Options	;	needed
VAR	Variable	Names	;	needed
WITH	〃	〃	;	needed
PARTIAL	〃	〃	;	optional
FREQ	〃	〃	;	optional
WEIGHT	〃	〃	;	optional
BY	〃	〃	;	optional

① PROC CANCORR options

이 Statement의 Option 部分에 나타날 수 있는 것들을 나열해 보면 다음과 같다.

- DATA=SAS dataset ; 分析하려는 DATASET NAME
- OUT=SAS dataset ; 分析에 利用된 原變數들과 Score on Canonical Variables들을 담은 DATASET NAME
- OUTSTAT=SAS dataset ; 分析結果중 正準相關關係 및 여러 統計量을 담을 DATASET
- SIMPLE(=S) ; 平均과 標準偏差를 Print
- CORR(=C) ; 原變數들 사이의 相關을 Print

- REDUNDANCY (=RED) ; Redundancy Statistics를 Print
- ALL ; 모든 가능한 結果를 Print
- NCAN = n ; Full Output이 必要한 正準變數의 갯수
- EDF = error d.f ; 誤差自由度 (input이 residual일 때)
- RDF = regress d.f ; 回歸自由度 (input이 residual일 때)
- NOINT ; intercept가 없는 model
- VPREFIX(=VP)=name ; VAR에 나타난 變數群의 正準變數들의 name (by default, V_1, V_2, ……, Vn)
- WPREFIX(=WP)=name ; WITH에 나타난 變數群의 正準變數들의 name (by default, W_1, W_2, ……, Wn)
- VNAME(=VN)="label" ; VAR에 나타난 變數들의 label (40字以內)
- WNAME(=WN)="label" ; WITH에 나타난 變數들의 label (40字 以內)

② VAR Statement

- VAR Variables ; 2개의 變數群중 첫번째 變數群의 變數들

③ WITH Statement

- WITH Variables ; 나머지 變數群의 變數들

④ PARTIAL Statement

- PARTIAL Variables ; Partial Correlation으로 正準分析을 할 變數

⑤ WEIGHT Statement

- WEIGHT Variables ; Weighted product-moment correlation coefficient를 計算할 때 weighting할 變數. 이때

는 分散을 計算할 때 加重値의 合으로 나눈다.

⑥ FREQ Statement

• FREQ Variables ; 觀測値의 어떤 값이 얻어진 횟수를 表示하는 變數.

⑦ BY Statement

• BY Variables ; 觀測値를 grouping 하고자 하는 變數. 이때는 이 變數에 대해 觀測値들이 Sort 되어 있어야 한다.

分析結果表에 나타나는 統計量을 나열해보면 다음과 같다.

① 入力된 각 變數들의 간단한 統計量 (平均, 標準偏差, 歪度, 尖度)

② 入力된 變數들의 相關係數表

③ 正準相關

④ adjusted canonical correlation (Lawley, 1959)

⑤ 正準相關의 近似的 標準誤差

⑥ 分散比 (正準變數双에 대한 回歸分散과 誤差分散의 比), ($=1/(1-\rho_i^2)$; ρ_i 는 i 번째 正準相關)

⑦ Canonical R-squared ($=\rho_i^2$)

⑧ 歸無假說 (H_o ; $\rho_i = \rho_{i+1}, = \cdots\cdots, = \rho_k = 0$) 에 대한 尤度比 (H_o ; $\rho_1 = \rho_2 = \cdots\cdots = \rho_k = 0$ 일때는 尤度比 = Wilks' Λ)

⑨ F 統計量 (based on Rao's approximation)

⑩ NUM DF (numerator d.f),

DEN DF (denominator d.f),

PROB > F (F 統計量에 따른 P-value)

⑪ Wilks' LAMBDA, PILLAI'S TRACE, HOTELLING-LAWLEY TRACE,

ROY's GREATEST ROOT

⑫ 標準化되지 않은 正準係數와 標準化된 正準係數 (이때 正準係數의 分散은 1)

⑬ 4가지 正準構造 (canonical structure)

(原變數들과 正準變數들간의 相關係數)

⑭ Canonical Redundancy Analysis (Stewart and Love, 1968)

⑮ 각 變數와 상대편變數群의 처음 R개 正準變數들간의 Square Multiple Correlation

例 題 硏 究

어떤 크럽內의 20代 중반의 男性들을 對象으로 身體的 要素들과 運動量들의 關係를 알아보기 위하여 세가지의 身體的 變數와 세가지의 運動量 變數를 測定하여 正準相關分析을 실시하였다.

運動量變數 : 턱걸이 (Chin), 윗몸일으키기 (Situp), 높이뛰기 (Jump)

身體的變數 : 몸무게 (Weight), 허리 (Waist), 맥박 (Pulse)

(Program)

```
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;
 TITE2 DATA COURTESY OF DR.A.C. LINNERUD, NC STATE
 UNIV;
```

（分析結果）

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

①

SIMPLE STATISTICS

VARIABLE	MEAN	ST. DEV.	SKEWNESS	KURTOSIS
WEIGHT	178.60000000	24.690505313	0.9698740166	1.802346254
WAIST	35.40000000	3.201973076	1.8721345109	5.662099021
PULSE	56.10000000	7.210372645	0.8460998408	0.606913027
CHINS	9.45000000	5.286278165	-.1930287524	-1.413520979
SITUPS	145.55000000	62.566575068	0.2236427744	-1.329139344
JUMPS	70.30000000	51.277470173	2.4799104223	7.623492371

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

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

CANONICAL CORRELATION ANALYSIS

③ CANONICAL CORRELATIONS AND ④ ADJUSTED ⑤ APPROX ⑥ VARIANCE ⑦ CANONICAL ⑧ TESTS OF HO: THE CANONICAL CORRELATION IN THE CURRENT ⑨ LIKELIHOOD ROW AND ALL THAT ⑩ F FOLLOW ARE ZERO

	CANONICAL CORRELATION	ADJUSTED CAN CORR	APPROX STD ERROR	VARIANCE RATIO	CANONICAL R-SQUARED	LIKELIHOOD RATIO	F STATISTIC	NUM DF	DEN OF	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
	WILKS' LAMBDA	0.3503905	2.048234	9	34.22293	0.06353094
⑪	PILLAI'S TRACE	0.6784815	1.558707	9	48	0.1551082
	HOTELLING-LAWLEY TRACE	1.771941	2.493844	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

	PAYS 1	PHYS 2	PHYS 3
WEIGHT	-.0314046879	-.0763195063	-.0077350467
WAIST	0.4932416756	0.3687229894	0.1580336471
PULSE	-.0081993154	-.0320519942	0.1457322421

RAW CANONICAL COEFFICIENTS FOR THE EXERCISES

	EXER 1	EXER 2	EXER 3
CHINS	-.0661139864	-.0710412111	-.2452753473
⑫ SITUPS	-.0168462308	0.0019737454	0.0197676373
JUMPS	0.0139715689	0.0207141063	-.0081674724

STANDARDIZED CANONICAL COEFFICIENTS FOR THE PHYSIOLOGICAL MEASUREMENTS

	PHYS 1	PHYS 2	PHYS 3
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	-1.2966
SITUPS	-1.0540	0.1235	1.2368
JUMPS	0.7164	1.0622	-0.4188

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

	PHYS 1	PHYS2	PHYS 3
WEIGHT	0.6206	-0.7724	-0.1350
WAIST	0.9254	-0.3777	-0.0310
PULSE	-0.3328	0.0415	0.9421

CORRELATIONS BETWEEN THE EXERCISES AND THEIR CANONICAL VARIABLES

	EXER 1	EXER 2	EXER 3
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

	EXER 1	EXER 2	EXER 3
WEIGHT	0.4938	-0.1549	-0.0098
WAIST	0.7363	-0.0757	-0.0022
PULSE	-0.2648	0.0083	0.0684

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

	PHYS 1	PHYS 2	PHYS 3
CHINS	-0.5789	0.0475	-0.0467
SITUPS	-0.6506	0.1149	0.0040
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

STANDARDIZED 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	0.4508	0.4508	0.6330	2.2854	0.2854
2	0.2470	0.6978	0.0402	0.0099	0.2953
3	0.3022	1.0000	0.0053	0.0016	0.2969

STANDARDIZED VARIANCE OF THE EXERCISES
EXPLAINED BY

	THEIR OWN CANONICAL VARIABLES			THE OPPOSITE CANONICAL VARIABLES	
	PROPORTION	CUMULATIVE PROPORTION	CANONICAL R-SQUARED	PROPORTION	CUMULATIVE PROPORTION
1	0.4081	0.4081	0.6330	0.2584	0.2584
2	0.4345	0.8426	0.0402	0.0175	0.2758
3	0.1574	1.0000	0.0053	0.0008	0.2767

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

CANONICAL REDUNDANCY ANALYSIS

SQUARED MULTIPLE CORRELATIONS BETWEEN THE PHYSIOLOGICAL MEASUREMENTS AND THE FIRST "M" CANONICAL VARIABLES OF THE EXERCISES

M	1	2	3
WEIGHT	0.2438	0.2678	0.2679
WAIST	0.5421	0.5478	0.5478
PULSE	0.0701	0.0702	0.0749

⑮

SQUARED MULTIPLE CORRELATIONS BETWEEN THE EXERCISES AND THE FIRST "M" CANONICAL VARIABLES OF THE PHYSIOLOGICAL MEASUREMENTS

M	1	2	3
CHINS	0.3351	0.3374	0.3396
SITUPS	0.4233	0.4365	0.4365
JUMPS	0.0167	0.0536	0.0539

分析結果에서는 먼저 각 變數들의 간단한 統計量(平均, 標準偏差, 歪度, 尖度)를 볼 수 있다.

②번에서 變數群間 또는 變數群內의 相關係數를 보면 Weight(몸무게)와 Waist(허리)의 相關係數는 0.8702이고 Chins(턱걸이)와 Situps(윗몸일으키기)의 相關係數는 0.6957이며 變數群間에는 Waist와 Situps가 -0.6456으로 가장 높게 나타났다.

③번에는 3개의 正準相關이 나타나있고,

⑩번에는 이들 係數에 관한 歸無假說(正準相關=0)의 有意水準이 나타나있다. 첫번째 正準相關 0.7956을 제외한 나머지 2개의 正準相關은 무시할 수 있으나, 첫번째 正準相關 역시 標本(sample)이 작고 또한 有意水準이 0.0635이므로 確實한 結論을 내리기는 힘들다.

⑫번에 原正準係數와 標準化된 正準係數가 나타나 있는데, 각 變數들의 測定方法이 다르기 때문에 標準化된 正準係數를 살펴보아야 한다. 먼저 身體的 變數群의 첫번째 正準變數(PHYS1)는 Weight가 -0.7754, Waist가 1.5793, pulse(맥박)가 -0.0591인 正準係數로 表示되며 또한 運動量 變數群의 첫번째 正準變數(EXER1)는 Chins가 -0.3495, Situps가 -1.0540, Jumps(높이뛰기)가 0.7164로 表示되고 있다.

⑬번에서도 여러가지의 相關係數가 나와있는데 Waist와 PHYS1와의 相關係數는 0.9254, Weight와 PHYS1와의 相關係數는 0.6206이다. 또한 運動量 變數群과 EXER1의 關係에서는 Situps가 가장 큰 相關係數(-0.8177)를 가지며 Jumps는 -0.7276이다. 그런데 Weight變數는 PHY1과의 正準係數(-0.7754)와 相關係數(0.6206)가 서로 符號가 다르며 Jumps 變數 역시 EXER1과의 正準係數(0.7164)와 相關係

數 (-0.1622) 가 서로 符號가 다르다. 이와같이 상반된 現象을 說明하기 위해 아래와 같은 상황을 생각해보자.

먼저 Situps 變數를 豫測하기 위하여 Weight 와 Waist 의 多重回歸를 設定하자.

여윈사람보다는 뚱뚱한 사람이 Situps 를 못할 것이고, 또한 키가 큰 차이가 없다고 假定하면 Weight 와 Waist 는 높은 相關關係(0.8702)를 갖는다. 또한 Large Waist 를 가진사람은 Small Waist 를 갖는 사람보다 Situps 를 못할것임으로 Waist 와 Situps 와의 相關關係는 陰의 關係를 갖고 Weight 역시 Waist 와 마찬가지로 陰의 相關關係를 갖는다.

다른 한편으로 Weight 가 고정되었다면 Large Waist 를 가진 사람은 키가작고 뚱뚱할 것임으로 Situps 와의 多重回歸에서 Waist 의 回歸係數는 陰의 係數를 갖고, Waist 가 고정되었다면 무거운 Weight 를 가진사람은 키가 큰 반면 뚱뚱하지 않을 것임으로 Situps 와는 陽의 係數를 갖게되어 正準係數의 符號와 正準變數의 相關關係 符號가 상반됨을 알 수 있다.

結論的으로 첫번째 正準相關의 全般的인 解釋을 해보면, Waist 와 Situps 사이의 相關關係를 강조하기 위해 Weight 와 Jumps 는 부차적인 變數 (Suppressor Variable) 로 作用하고 있다.

正準變數들의 豫測力을 살펴보기 위한 ⑭번의 Canonical Redundancy Analysis 는 두 變數群으로부터 나온 한쌍의 正準變數가 각각 반대편 變數群의 全體 또는 개개 變數들에 대한 豫測力을 評價하는 것이다. 각 正準變數에 의해 說明되는 標準化된 分散表를 보면 身體的 變數群의 첫번째 正準變數가 說明할 수 있는 반대편 變數群의 分散比率

이 0.2854이고, 運動量 變數群의 分散比率은 0.2584이므로 첫번째 正準變數 한쌍의 반대편 變數에 대한 全般的인 豫測變數로서는 좋은 편이 못된다.

⑮번에서는 개개의 變數에 대한 正準變數의 豫測力이 나타나 있는데 먼저 運動量 變數群의 첫번째 正準變數의 豫測力은 Waist에 대해서는 0.5421 (양호), Weight 에 대해서는 0.2438 이고, 身體的 變數群의 경우 첫번째 正準變數의 豫測力은 Situps 에 대해서는 0.4233, Chins 에 대해서는 0.3351 이다.

2 . 要因分析 (Factor Analysis)

要因分析 (또는 因子分析) 은 어떤 하나의 變數群內 여러 變數들 사이의 相互關係로부터 共通變量을 구하고, 測定値의 重複性을 찾아 내어 몇개의 基本的인 要因들을 抽出하여 보다 明確한 說明을 가능케 하는 分析技法이다.

要因分析의 使用目的을 크게 세가지로 나누어 보면

① 探査의 目的 (Exploratory Purpose) : 變數들이 어떠한 形態 (Pattern) 로 構成되어 있는가를 알아내어 새로운 槪念을 發見해 내는 目的

② 確認의 目的 (Confirmatory Purpose) : 硏究者가 理論的 根據를 가지고 미리 設定한 要因들이 과연 예상대로 抽出되었는가를 確認하려는 目的

③ 測定의 目的 (Measurement Purpose) : 要因分析으로부터 얻어진 結果를 指數 (Index) 化하여 추후 分析에 利用하려는 目的 SAS 에서 이 分析을 遂行하는 Procedure 는 PROC FACTOR 이다.

여기에서는 여러가지의 要因分析方法을 수행할 수 있고 여러 가지의 入力形態 및 出力 (Output) 을 얻을 수 있다. 또한 이 Procedure 의 Output 으로부터 Factor Score 를 구하기 위해 연속해서 PROC SCORE 를 使用할 수 있다.

〈表 10〉 要因分析의 여러가지 方法

區 分	內 容	
入力 DATA 形 態	• 原始 DATA • Correlation matrix • Covariance matrix • Factor pattern • Matrix of scoring coefficient • Canonical coefficient (from CANDISC)	
分析方法	要因도출 方法	• Principal Component Analysis • Principal Factor Analysis • Maximum Likelihood Factor Analysis • Iterated Principal Factor Analysis • Unweighted Least Squares Factor Analysis • Alpha Factor Analysis • Image Component Analysis • Harris Component Analysis
	回轉方法	• Varimax • Quartmax • Equamax • Orthomax with user-specified Gamma • Promax with user-specified exponent

區　　分	內　　容	
		• Harris-Kaiser Case Ⅱ with user-specified exponent • Oblique Procrustean with user-specified target pattern
出　　力 （Output）	• 平均標準偏差，相關係數 • Kaisers measure of sampling adequacy • eigenvaluses , eigenvectors • scree plot • prior and final communality • unrotated factor pattern • residual and partial correlation • rotated primary factor pattern • primary factor structure • inter-factor correlations • reference structure • reference axis correlation • variance explained by each factor • plots of both rotated and unrotated factors • squared multiple correlation of each factor with the variables • scoring coefficients	

因子抽出方法 중 代表的인 것은 Principal components, Principal factor analysis, Maximum likelihood factor analysis이다. 이 중 가장 간단하고 계산이 效率的인 것은 Principal factor analysis 이며, Prior communality estimates를 준다는 點이 Principal components와의 差異이다.

Principal factor analysis에서는 一般的으로 어떤 變數와 다른 모든 變數間의 SMC (squared multiple correlation)를 Prior communality estimates로 使用하지만 相關係數行列表(correlation matrix) 가 Singular이면 SMC 대신 MAX (maximum absolute correlation)을 使用한다.

Maximum likelihood factor analysis는 推定値들이 近似的 特性(asymptotic property)을 가짐으로 큰 標本에서는 좋다. 또한 因子에 대한 假說檢定을 할 수 있다는 利點이 있으나 이 方法은 Communality를 反復的으로 推定하므로 計算時間과 經費가 많이 들며 因子의 數가 달라질 때마다 Factor procedure를 수행해야 하므로 이 때는 事前에 Principal factor analysis로써 因子數를 대강 把握한 뒤 Maximum likelihood factor analysis를 하는 것이 效果的이다.

SAS Program Coding

```
PROC   FACTOR      options     ;  ┐
PRIORS             communalities ;  ├ needed
VAR                variables   ;  ┘
PARTIAL            variables   ;  ┐
FREQ               variables   ;  ├ optional
WEIGHT             variables   ;  ┘
```

BY variables ; ┘

① PROC FACTOR options

因子分析에서는 여러가지의 選擇을 할 수 있는데 이를 性格上으로 나누어 보면 다음과 같다.

a. DATA set

b. 因子抽出方法

c. 回轉方法

d. 出力 (output)

e. 其 他

a. DATA set

- DATA=SAS dataset ; 分析하려는 DATASET NAME
- OUTSTAT=SAS dataset ; 分析結果를 담을 DATA NAME
- TARGET=SAS dataset ; 斜角的 (oblique) products 回轉에 使用될 target pattern을 담은 DATA SET

b. 因子抽出方法

- METHOD(=M)=PRINCIPAL(=P) ; principal components를 遂行. 만일 뒤에 PRIOR가 使用되면 principal factor analysis를 수행.
- M=PRINIT ; iterated principal factor analysis를 수행.
- M=U(=ULS) ; unweighted least squares factor analysis를 수행.
- M=ALPHA(=A) ; alpha factor analysis를 수행
- M=M(=ML) ; maximum likelihood factor analysis를 수행(단 Correlation matrix가 non-singular일 때).

- M = HARRIS(=H) ; Harris component analysis를 수행(단 correlation matrix가 non-singular일 때).
- M = IMAGE(=I) ; image covariance matrix의 principal component analysis를 수행.
- M = PATTERN ; DATA SET의 TYPE = FACTOR, CORR, COV일 때 factor pattern을 읽는다.
- M = SCORE ; DATA SET의 TYPE = FACTOR, CORR, COV일 때 scoring coefficients를 읽는다.
- (M =)을 생략했을 때 ; M = PRINCIPAL과 같다. 만약 DATA SET의 TYPE = FACTOR이면 M = PATTERN과 같다.
- COVARIANCE(=COV) ; 相關係數行列(correlation matrix)대신에 共分散行列(covariance matrix)로 factor를 수행한다.(단, M = P, PRIIT, ULS, IMAGE에만 使用)
- WEIGHT ; weighted correlation 또는 covariance matrix로 factor를 수행. (단, M = P, PRINIT, ULS, IMAGE에만 使用하며 DATA SET의 TYPE = CORR, COV, FACTOR이어야 한다.)
- MAXITER = n ; 反復回數를 指定(단, M = PRINIT, ULS, ALPHA, ML일 때 使用하며 default는 30)
- CONVERGE = n ; 수렴기준을 설정(default는 0.001)
- NFACTORS = n ; 抽出할 因子(factor)의 數를 指定(default는 variable의 數)
- PROPORTION(=P) = n ; 全體의 n% 만큼의 分散을 說明할 수 있는 factor의 數를 決定(단 M = PATTERN, SCORE에서는 使用不可)

- MINEIGEN(=MIN) = n ; factor의 數를 決定하는 基準으로 eigenvalue가 n보다 큰 factor를 抽出(단, M = PATTERN, SCORE에서는 使用不可)

c. 回轉方法

- ROTATE(=R) = VARIMAX(=V) ; varimax rotation
- R = QUARTIMAX(=Q) ; quartimax rotation
- R = EQUAMAX(=E) ; equamax rotation
- R = ORTHOMAX ; GAMMA = n으로 指定된 加重値를 갖는 orthomax rotation
- R = PROMAX(=P) ; promax rotation
- R = PROCRUSTES ; oblique procrustean rotation (단, TARGET = data set에서 주어진 target pattern을 利用)
- GAMMA = n ; orthomax weight를 指定
- POWER = n ; ROTATE = PROMAX에서 target pattern을 計算하는데 使用되며 power를 指定(default는 3)
- PREROTATE(=PRE) = name ; PROMAX에서의 prerotation 方法을 指定하며 만약 M = PATTERN이었다면 PRE = NONE이 되어야 한다 (default는 varimax)
- NORM= name ; 回轉을 위하여 factor pattern의 行들을 正規化시키는 方法을 指定(default는 KAISER)

d. 出力(output 혹은 print)

- SIMPLE(=S) ; 平均, 標準偏差
- CORR(=C) ; 相關係數行列(correlation matrix)

- MSA ; partial correlations, Kaiser's measure of sampling adequacy
- SCREE ; scree plot of eigenvalue
- EIGENVECTORS(=EV) ; eigenvector
- PRINT ; INPUT factor pattern 또는 scoring coefficient 그리고 관계되는 여러 統計量. oblique인 경우는 reference and factor structure (단, M=PATTERN,SCORE인 경우에만)
- RESIDUAL(=RES) ; 殘差相關係數行列(residual correlation matrix)과 partial correlation matrix
- PREPLOT ; 回轉前 factor pattern
- PLOT ; 回轉後 factor pattern
- NPLOT=n ; plot할 factor의 數를 指定
- SCORE ; factor scoring coefficients. 각 factor와 變數들 사이의 squared multiple correlation
- ALL ; plot를 除外한 모든 output
- REORDER ; 變數들의 factor에 대한 積載値(loading)의 크기 順序대로 再配列
- ROUND ; element들을 100으로 곱하여 가장 近接한 整數로 대체하여 만든 correlation 및 loading matrixs
- FLAG=n ; ROUND와 같이 使用되며 correlation 및 loading matrixs에 element의 絶對値가 n보다 크면 element에 별표(*)를 단다.
- FUZZ=n ; FLAG의 反對現象으로 n보다 적으면 결측치(missing value)로 print (partial correlation은 $\frac{n}{2}$, residual correlation은 $\frac{n}{4}$을 基準)

e. 其 他

- NOINT ; intercept 를 使用하지 않는다.
- NOCORR ; OUTSTAT= data set 에 相關係數行列을 넣지 않는다.(단 M = PATTERN,SCORE 일 때 使用)
- TOLERANCE = n ; 相關係數行列의 singuality 를 決定하는 限界(default 는 1E-7)

② PRIORS communalities ;

각 變數들의 事前 communalities 를 指定해 주는 것으로 그 順序는 VAR 에서의 variables 의 順序에 따른다.

例) PROC FACTOR ;
VAR X Y Z ;
PRIORS .7 .8 .9 ;

위와 같이 數로써 각각의 事前 communalities 를 指定하는 대신, 다음과 같은 여러가지 方法으로 일률적인 指定을 할 수 있다.

- ONE ; 모든 事前 communalities 가 1.0
- MAX ; 他 變數와의 maximum absolute correlation 을 事前 communalities 로 利用
- SMC ; 다른 모든 變數와의 square multiple correlation 을 利用
- ASMC ; 事前 communalities 가 SMC 에 비례하고 그 合이 MAX 와 같게 한다.
- INPUT ; DATASET(TYPE = FACTOR) 에서 事前 communalities 를 읽어 들인다.
- RANDOM ; 事前 communalities 로 0 과 1 사이의 uniformly random number 를 使用

③ VAR statement

- VAR variables ; 分析할 variables 를 나열

④ PARTIAL statement

- PARTIAL variables ; 分析에 partial correlation 이나 covariance matrix 를 利用할 때

⑤ FREQ statement

部分的으로 分析할 variables 을 나열

- FREQ variable ; 觀測値의 어떤 값이 얻어진 回數를 表示하는 變數

⑥ WEIGHT statement

- WEIGHT variable ; 入力資料에 있는 각 變數에 대해 相對的 加重値를 줄 必要가 있을 때 指定한다. 이때 分散은 加重値의 合으로 計算한다.

⑦ BY statement

- BY variables ; 觀測値 중 grouping 하고자 하는 變數를 指定. 이때 이 變數의 값에 따라 觀測値들이 sort 되어 있어야 한다.

以上에서 要因分析을 하기 위한 program 構成과 여러가지 方法을 나열하였다.

이러한 program을 施行 (RUN) 하였을 때 分析結果表에 나타나는 統計量을 나열해 보면 다음과 같다.

①; 各 變數들의 平均, 標準偏差 및 觀測値數 (Option (SIMPLE) 이 있을 때)

②; 相關係數行列 (Option(CORR) 이 있을 때)

③ ; 相關係數行列의 역행렬

(Option (ALL) 이 있을 때)

④ ; Patial Correlations (controlling all other variables)

(Option (MSA) 이 있을 때)

⑤ ; 標本적합성에 대한 KAISER'S MEASURE

(Option (MSA) 이 있을 때)

⑥ ; 事前 커뮤널리티 推定値

(Option (M = IMAGE, HARRIS, PATTERN, or SCORE) 이 없을 때)

⑦ ; Squared Multiple Correlations

(Option (M = IMAGE or HARRIS) 이 있을 때)

⑧ ; Image Coefficients (Option (M = IMAGE) 이 있을 때)

⑨ ; Image Covariance Matrix (〃)

⑩ ; Preliminary Eigenvalues (based on the prior communality)

(Option (M = PRINIT, ALPHA, ML, or ULS) 이 있을 때)

⑪ ; 要因數 (Option (M=PATTERN or SCORE) 이 없을 때)

⑫ ; A Scree Plot of Eigenvalues (Option(SCREE) 이 있을 때)

⑬ ; the iteration history (Option (M = PRINIT, ALPHA., ML, or ULS) 이 있을 때)

⑭ ; 要因數에 對한 有意性檢定 (Option (M=ML) 이 있을 때)

⑮ ; AKAIKE'S INFORMATION CRITERION (Option (M=ML) 이 있을 때)

⑯ ; SCHWARZ'S BAYESIAN CRITERION (Option (M=ML) 이 있을 때)

⑰ ; Squared Canonical Correlation (Option (M=ML) 이 있을 때)

⑱ ; Coefficient Alpha for Each Factor (Option (M=ML) 이 있을 때)

⑲ ; Eigenvectors (Option (M = PATTEN or SCORE) 이 없고 Option (EIGENVECTORS or ALL) 이 있을 때)

⑳ ; Eigenvalues of The (Weighted) (Reduced) (Image) Correlation or Covariance Matrix (Option (M = PATTERN or SCORE) 이 없을 때)

㉑ ; Factor Pattern

㉒ ; Variance (explained by each factor)

㉓ ; Final Communolity Estimates

㉔ ; Residual Correlations (with uniqueness on the diagonal) (Option (RESIDUAL or ALL) 이 있을 때)

㉕ ; Root-Mean-Square (off-diagonal residual) (Option (RESIDUAL or ALL) 이 있을 때)

㉖ ; Partial Correlations (controlling factors) (Option (RESIDUAL or ALL) 이 있을 때)

㉗ ; Root-Mean-Square (off-diagonal patials) (Option (RESIDUAL or ALL) 이 있을 때)

㉘ ; Plot of Factor Pattern (for unrotated factors) (Option (PR-EPLOT) 이 있을 때)

㉙ ; Variable Weights (for rotation) (Option (NORM = WEIGHT) 이 있을 때)

㉚ ; Factor Weights (for rotation) (Option (HRPOWER) 이 있을 때)

㉛ ; Orthogonal Transformation Matrix (直交回轉이 必要할 때)

㉜ ; Rotated Factor Pattern (直交回轉이 必要할 때)

㉝ ; Variance (explained by each factor after rotation) (直交回轉이 必要할 때)

㉞ ; Target Matrix (for Procrustean transformation) (Option (ROTATE = PROCRUSTES or PROMAX) 이 있을 때)

㉟ ; Procrustean Transformation Matrix (Option (ROTATE = PROCRUSTES or PROMAX) 이 있을 때)

㊱ ; Oblique Transformation Matrix (斜角的 回轉이 必要할 때)

㊲ ; Inter-Factor Correlations (斜角的 回轉이 必要할 때)

㊳ ; Rotated Factor Pattern (斜角的 回轉이 必要할 때)

㊴ ; Reference Axis Correlations (斜角的 回轉이 必要할 때)

㊵ ; Reference Structure (Semipatial Correlations) (斜角的 回轉이 必要할 때)

㊶ ; Variance (explained by each factor eliminating the effects of all other factors) (斜角的 回轉이 必要할 때)

㊷ ; Factor Structure (Correlations) (斜角的 回轉이 必要할 때)

㊸ ; Variance (explained by each factor ignoring the effects of all other factors) (斜角的 回轉이 必要할 때)

㊹ ; Final Communality Estimates (for the rotated factors) (Option (ROTATE=) 이 있을 때)

㊺ ; Squared Multiple Correlations (of the variables with each factor) (Option (SCORE or ALL) 이 있을 때)

㊻ ; Standardized Scoring Coefficients (Option (SCORE or ALL) 이 있을 때)

㊼ ; Plot of Factor Pattern (for rotated factors) (Option (PLOT) 이 있을 때)

㊽ ; Plot of Reference Structure (for rotated factors) (Option (PLOT) 이 있을 때)

(例題研究)

Los Angeles Standard Metropolitan Statistical Area 의 12個 調査區에서 5個 社會經濟變數를 調査하여 이를 3가지 要因分析方法으로 각각 分析하였다.

(1) Principal Components

```
DATA SOCECON;
  TITLE FIVE SOCIO-ECONOMIC VARIABLES;
  TITLE2 SEE PAGE 14 OF HARMAN:MODERN FACTOR ANALY-
  SIS,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 SIMPLE  CORR;
 TITLE3 PRINCIPAL COMPONENTS ANALYSIS;
```

（分析結果）

FIVE SOCIO-ECONOMIC VARIABLES 1
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
PRINCIPAL COMPONENTS ANALYSIS

① MEANS AND STANDARD DEVIATIONS FROM 12 OBSERVATIONS

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
MEAN	6241.67	11.4417	2333.33	120.833	17000
STD DEV	3439.99	1.78654	1241.21	114.928	6367.53

② CORRELATIONS

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
POP	1.00000	0.00975	0.97245	0.43887	0.02241
SCHOOL	0.00975	1.00000	0.15428	0.69141	0.86307
EMPLOY	0.97245	0.15428	1.00000	0.51472	0.12193
SERVICES	0.43887	0.69141	0.51472	1.00000	0.77765
HOUSE	0.02241	0.86307	0.12193	0.77765	1.00000

FIVE SOCIO-ECONOMIC VARIABLES 2
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
PRINCIPAL COMPONENTS ANALYSIS

INITIAL FACTOR METHOD: PRINCIPAL COMPONENTS

⑥ PRIOR COMMUNALITY ESTIMATES: ONE

⑳ EIGENVALUES OF THE CORRELATION MATRIX : TOTAL =5.000000 AVERAGE = 1.000000

	1	2	3	4	
EIGENVAIUE	2.873314	1.796660	0.214837	0.099934	0.015255
DIFFERENCE	1.076654	1.581823	0.114903	0.084679	
PROPORTION	0.5747	0.3593	0.0430	0.0200	0.0031
CUMULATIVE	0.5747	0.9340	0.9770	0.9969	1.0000

⑪ 2 FACTORS WILL BE RETAINED BY THE MINEIGEN CRITERION

㉑ FACTOR PATTERN

	PACTOR1	PACTOR2
POP	0.58096	0.80642
SCHOOL	0.76704	-0.54476
EMPLOY	0.67243	0.72605
SERVICES	0.93239	-0.10431
HOUSE	0.79116	-0.55818

㉒ VARIANCE EXPLAINED BY EACH FACTOR

FACTOR1	FACTOR2
2.873314	1.796660

㉓ FINAL COMMUNALITY ESTIMATES : TOTAL = 4.669974

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.987826	0.885106	0.979306	0.880236	0.937500

먼저 ①번과 ②번에는 각 變數들의 平均標準偏差, 相關關係行列表가 주어져 있다. 이 分析結果는 Principal Component Analysis이므로 事前 Communalities는 ⑥번과 같이 1이다(Default).

⑳번에는 5개의 Factor에 대한 Eigenvalue와 그들 각각이 說明할 수 있는 分散의 比率이 함께 나와 있다. 그런데 이때 Principal Component Analysis의 自動選擇에 따라 Eigenvalue가 1 以上인 Factors만 選擇하므로 처음 2개의 Factors가 抽出되고 이 2개의 Factors가 分散의 93.4%를 說明할 수 있으므로 적절하다고 말할 수 있다.

㉑번에는 變數들의 因子에 대한 Factor Pattern Matrix가 나와 있고 그들 각각의 積載値(Loading)를 보면 5개 變數 모두 Factor 1에 높은 陽의 積載値(Loading)를 갖고 특히 Services (0.93239), House(0.79116), School(0.76704)의 경우는 높은 相關을 나타내고 있다. Factor 2에서는 Pop(0.80642)과 Employ(0.72605)는 나머지 變數들과 對照的으로 높게 Loading되어 있다.

㉓번에는 Final Communality Estimates가 나타나 있는데 이것은 각 變數들의 2개의 Factor에 대한 積載値(Loading)의 自乘의 合으로 구해진다. 結果를 보면 모든 Final Communality Estimates가 0.88에서 0.987까지의 分布로, 變數들이 2개의 Component로써 잘 說明된다고 할 수 있다.

(2) Principal Factor Analysis

(Data 部分은 (1)의 Principal Components와 同一)

```
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;
```

（分析結果）

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
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	SERVICE	HOUSE
0.472079	0.551588	0.488511	0.806644	0.612814

PRIOR COMMUNALITY ESTIMATES: SMC

POP	SCHOOL	EMPLOY	SERVICE	HOUSE
0.968592	0.822285	0.969181	0.785724	0.847019

EIGENVALUES OF THE REDUCED CORRELATION MATRIX: TOTAL= 4.392801 AVERAGE = 0.878506

	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 5
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

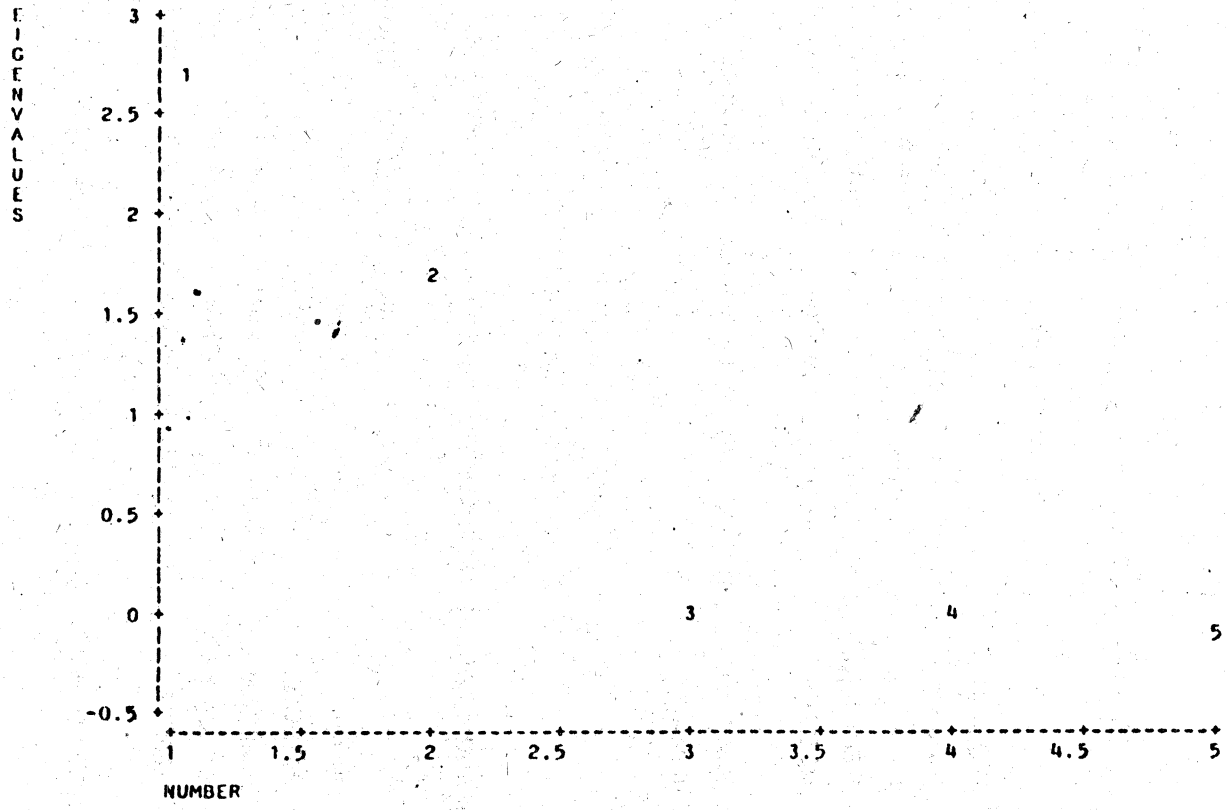
VARIANCE EXPLAINED BY EACH FACTOR

FACTOR1	FACTOR2
2.734301	1.716069

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
SCREE PLOT OF EIGENVALUES

(12)



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

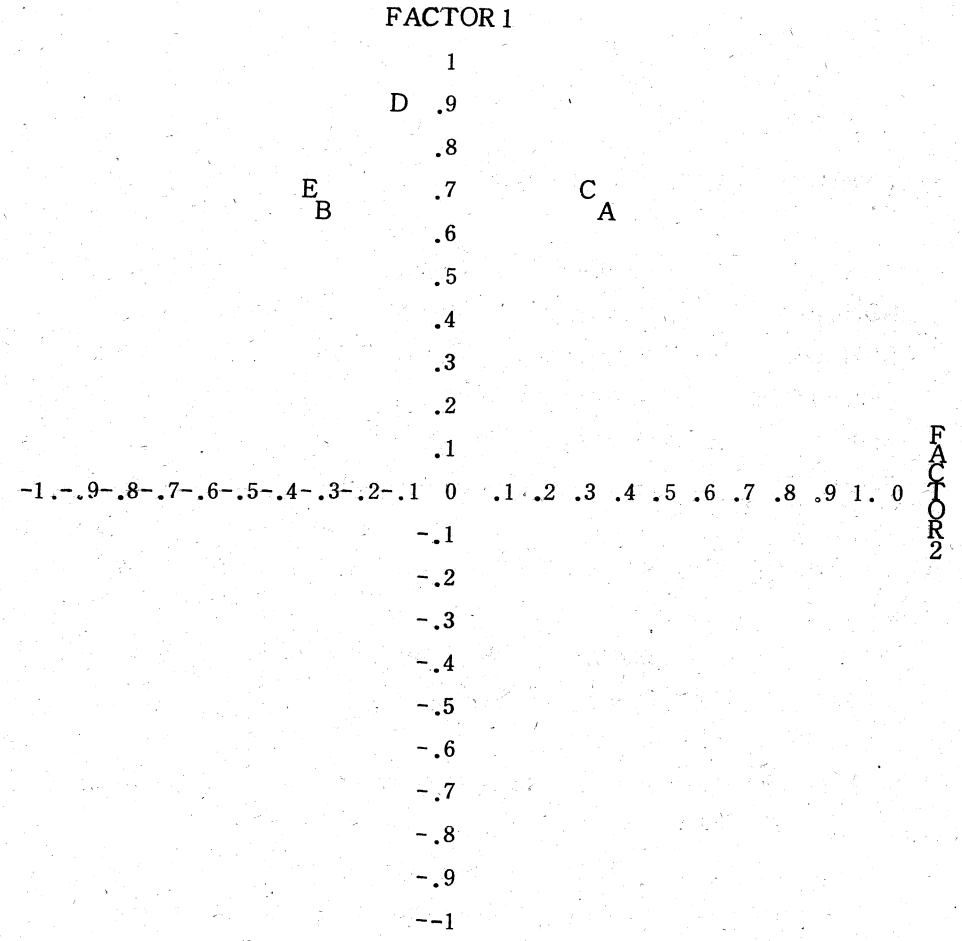
㉗ ROOT MEAN SQUARE OFF-DIAGONAL PARTIALS: OVER-ALL = 0.18550132

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.158508	0.190259	0.231818	0.154470	0.182015

FIVE SOCIO-ECONOMIC VARIABLES 6
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS. 3RD ED
PRINCIPAL FACTOR ANALYSIS WITH PROMAX ROTATION

INITIAL FACTOR METHOD: PRINCIPAL FACTORS

㉘ PLOT OF FACTOR PATTERN FOR FACTOR1 AND FACTOR2



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

PREROTATION METHOD: VARIMAX

㉛ ORTHOGONAL TRANSFORMATION MATRIX

	1	2
1	0.78895	0.61446
2	-0.61446	0.78895

㉜ ROTATED FACTOR PATTERN

	FACTOR2	FACTOR2
HOUSE	0.94072	-0.00004
SCHOOL	0.90419	0.00055
SERVICES	0.79085	0.41509
POP	0.02255	0.98874
EMPLOY	0.14625	0.97499

㉝ VARIANCE EXPLAINED BY EACH FACTOR

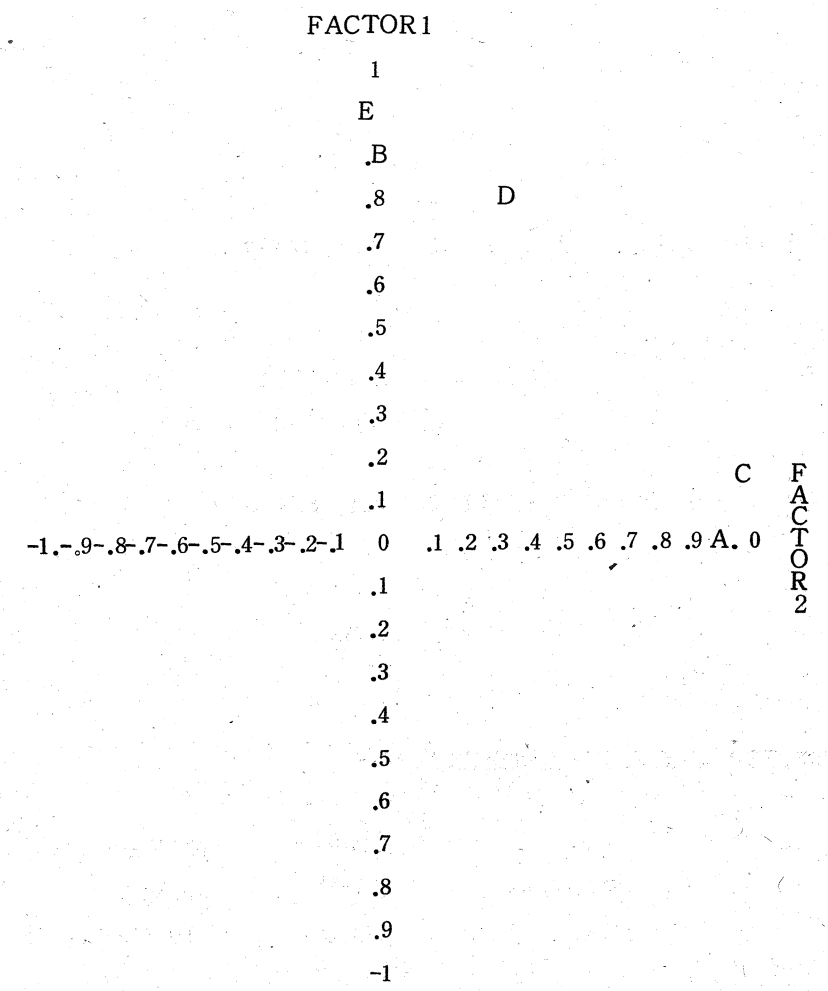
FACTOR1	FACTOR2
2.349857	2.100513

FINAL COMMUNALITY ESTIMATES: TOTAL = 4.450370

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.978113	0.817564	0.971999	0.797743	0.884950

PREROTATION METHOD: VARIMAX

PLOT OF FACTOR PATTERN FOR FACTOR1 AND FACTOR2



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

FIVE SOCIO-ECONOMIC VARIABLES 9
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
PRINCIPAL FACTOR ANALYSIS WITH PROMAX ROTATION

ROTATION METHOD: PROMAX

㉞ TARGET MATRIX FOR PROCRUSTEAN TRANSFORMATION

	FACTOR1	FACTOR 2
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

㉟ PROCRUSTEAN TRANSFORMATION MATRIX

	1	2
1	1.04117	-0.09865
2	-0.10572	0.96303

㊱ OBLIQUE TRANSFORMATION MATRIX

	1	2
1	0.73803	0.54202
2	-0.70555	0.86528

㊲ INTER-FACTOR CORRELATIONS

	FACTOR1	FACTOR 2
FACTOR 1	1.00000	0.20188
FACTOR 2	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

㊴ REFERENCE AXIS CORRELATIONS

	FACTOR1	FACTOR2
FACTOR1	1.00000	-0.20188
FACTOR2	-0.20188	1.00000

FIVE SOCIO-ECONOMIC VARIABLES 10
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
PRINCIPAL FACTOR ANALYSIS WITH PROMAX ROTATION

ROTATION METHOD: PROMAX

㊵ REFERENCE STRUCTURE (SEMIPARTIAL CORRELATIONS)

	FACTOR1	FACTOR 2
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	FACTOR 2
HOUSE	0.93582	0.09500
SCHOOL	0.89954	0.09189
SERVICES	0.82903	0.49286
POP	0.12319	0.98596
EMPLOY	0.24484	0.98478

㊸ VARIANCE EXPLAINED BY EACH FACTOR IGNORING OTHER FACTORS

FACTOR 1	FACTOR 2
2.447349	2.202280

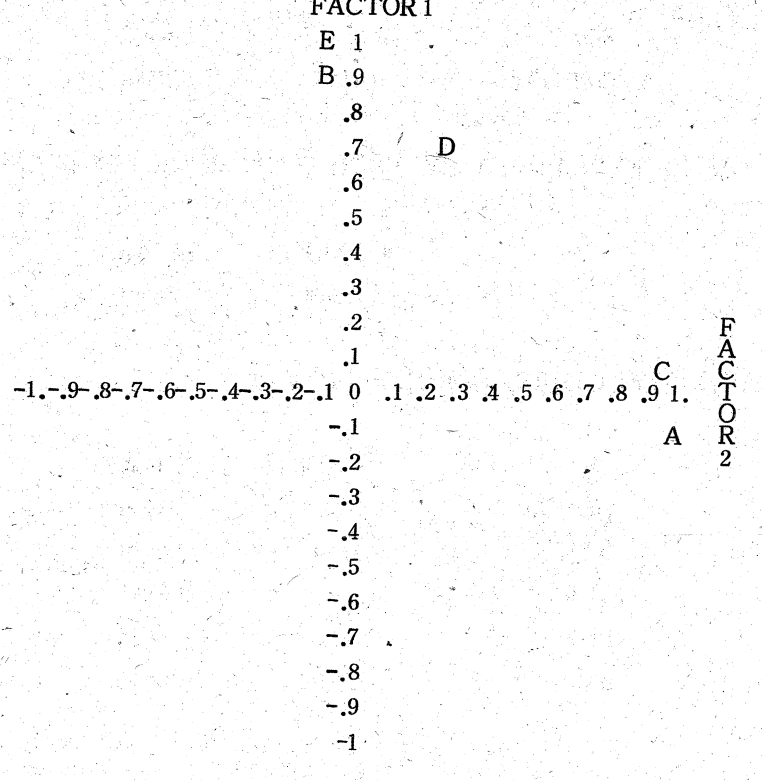
㊹ FINAL COMMUNALITY ESTIMATES: TOTAL = 4.450370

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.978113	0.817564	0.971999	0.797743	0.884950

FIVE SOCIO-ECONOMIC VARIABLES 11
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 FACTOR 1 AND FACTOR 2
REFERENCE AXIS CORRELATION = -0.2019 ANGLE = 101.65



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

FIVE SOCIO-ECONOMIC VARIABLES 12
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
FACTOR OUTPUT DATA SET

OBS	-TYPE-	NAME	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
1	MEAN		6241.67	11.4417	2333.33	120.833	17000.0
2	STD		3439.99	1.7865	1241.21	114.928	6367.5
3	N		12.00	12.0000	12.00	12.000	12.0
4	CORR	POP	1.00	0.0098	0.97	0.439	0.0
5	CORR	SCHOOL	0.01	1.0000	0.15	0.691	0.9
6	CORR	EMPLOY	0.97	0.1543	1.00	0.515	0.1
7	CORR	SERVICES	0.44	0.6914	0.51	1.000	0.8
8	CORR	HOUSE	0.02	0.8631	0.12	0.778	1.0
9	COMMUNAL		0.98	0.8176	0.97	0.798	0.9
10	PRIORS		0.97	0.8223	0.97	0.786	0.8
11	EIGENVAL		2.73	1.7161	0.04	-0.025	-0.1
12	UNROTATE	FACTOR1	0.63	0.7137	0.71	0.879	0.7
13	UNROTATE	FACTOR2	0.77	-0.5552	0.68	-0.158	-0.6
14	RESIDUAL	POP	0.02	-0.0112	0.01	0.011	0.0
15	RESIDUAL	SCHOOL	0.01	0.1824	0.02	-0.024	0.0
16	RESIDUAL	EMPLOY	0.01	0.0215	0.03	-0.006	-0.0
17	RESIDUAL	SERVICES	0.01	-0.0239	-0.01	0.202	0.0
18	RESIDUAL	HOUSE	0.00	0.0125	-0.02	0.034	0.1
19	PRETRANS	FACTOR1	0.79	-0.6145	.	.	.
20	PRETRANS	FACTOR2	0.61	0.7889	.	.	.
21	PREROTAT	FACTOR1	0.02	0.9042	0.15	0.791	0.9
22	PREROTAT	FACTOR2	0.99	0.0006	0.97	0.415	-0.0
23	TRANSFOR	FACTOR1	0.74	-0.7055	.	.	.
24	TRANSFOR	FACTOR2	0.54	0.8653	.	.	.
25	FCORR	FACTOR1	1.00	0.2019	.	.	.
26	FCORR	FACTOR2	0.20	1.0000	.	.	.
27	PATTERN	FACTOR1	0.08	0.9184	0.05	0.761	1.0
28	PATTERN	FACTOR2	1.00	-0.0935	0.98	0.339	-0.1
29	RCORR	FACTOR1	1.00	-0.2019	.	.	.
30	RCORR	FACTOR2	0.20	1.0000	.	.	.
31	REFERENC	FACTOR1	0.08	0.8995	0.05	0.745	0.9
32	REFERENC	FACTOR2	0.98	-0.0916	0.96	0.332	-0.1
33	STRUCTUR	FACTOR1	0.12	0.8995	0.24	0.829	0.9
34	STRUCTUR	FACTOR2	0.99	0.0919	0.98	0.493	0.1

이 分析에서는 몇가지 選擇을 하여 보다 많은 Output를 받아 보고 또한 回轉과 PLOT를 함으로써 보다 解釋이 容易하게 하였다.

먼저 MSA를 選擇하여 ④번과 ⑤번의 Output를 얻었다. 이는 入力資料들이 Common Factor Model에 적절한가를 보기 위해 MSA를 選擇하여 Partial Correlation Controlling all other variables (④번)와 Kaiser's measure of sampling adequacy (⑤번)를 얻은 것이다. 만약 入力資料들이 적절하다면 원래의 Correlation에 비해 Partial Correlation 이 적어야 된다. Principal Components 分析의 ②번과 Principal Factor 分析의 ④번을 比較해 보면, School과 House의 Correlation의 경우 ②번의 0.86에서 ④번의 0.64로 적어진 反面, Pop과 School의 Correlation 은 그렇지 않으므로 이 경우는 좋지 않다.

⑤번의 Kaiser's MSA는 원래의 Correlation에 비해 Partial Correlation이 얼마나 작은가를 보는 것으로 그 값이 0.8보다 크면 좋고 0.5보다 적으면 부적절하다. 결국 Pop, School, Employ는 좋지 않고, Service만이 0.806641로 좋다고 말할 수 있다. 또 全體的인 MSA는 0.57536759로 좋은 편이 못된다. 이와 같이 MSA가 좋지 않을 때에는 Common Factor를 보다 잘 定義하기 위해 부수적인 變數들을 더 包含시키든지 아니면 관계되는 變數들을 選別하여 除去시키든지 하여 Common Factor의 導出에 따른 標本의 適合性을 높일 必要가 있다. 보통 1개 Factor에 적어도 3개 以上의 變數들이 있어야(Rule of Thumb) 하나 여기의 分析에 使用된 資料들은 그렇지 못하므로 變數가 充分하지 못하다는 問題가 생기게 된다.

⑤번의 아랫쪽에는 事前 Communality 推定値가 SMC(Square Multiple Correlation)로 주어져 있는데 그 크기가 0.7857에서 0.96918로서 1에 近似한 크기이므로 要因積載値(Factor Loadings)는 Principal Components에서 구한 값과 크게 다르지 않다.

계속해서 크기 순으로 5개의 Factor 각각의 Eigenvalue와 說明할 수 있는 分散의 比率들이 나와 있고 그들 Eigenvalues의 合이 4.392801, 平均이 0.8785로 나와 있다. 또한 SMC의 合과 Eigenvalues의 合이 같음도 알 수 있다. 여기에서 앞의 2개의 Factor의 說明力이 101.3%로서 이는 反復計算(Iterating)에서 얻은 100%에 가까우므로 變數들을 說明하는데 Factor 1, 2로써 充分하며 이 역시 Principal Components의 結果와 같다.

⑫번에는 5개 Factor의 Eigenvalue 값이 圖表로 나타나 있어 구별하는데 도움이 된다.

그 다음 要因類型行列表(Factor Pattern Matrix)를 보면 Principal Components에서와 크게 다르지 않아 5개의 變數 모두 첫번째 要因에 높은 Loading을 갖고 있고 두번째 要因 역시 Pop 과 Employ가 높게 Loading되어 있는 反面 House와 School은 陰의 Loading을 갖고 있다. 變數들의 나열된 順序가 달라진 것은 RE-ORDER 選擇을 하여 먼저 첫번째 Factor에 크게 Loading된 順序, 그다음 두번째 Factor에 크게 Loading된 順序로 나열하였기 때문이다.

Final Communality Estimate는 事前 Communality들과 거의 비슷하고 House만 0.847109에서 0.884950으로 약간 개선되었다. 그러나 Final Communality는 거의 100%에 가까운 共通分散(Common variance)을 說明하므로 좋은 推定値라 할 수 있다.

㉔, ㉕, ㉖, ㉗번은 RESIDUAL을 選擇하여 구하였는데 이는 Residual Correlation에 관한 것으로 ㉔번에서 제일 큰 Residual Correlation이 0.3이다(단, 對角線은 除外), 한편 어떤 變數들은 Factor들로써 說明되는 共通變量(Common Variance)과 特殊變量(Specific Variance)으로 構成되어 있는데 Factor들이 說明하는 部分이 크고 特殊變量이 작으므로 ㉖번의 Partial Correlation은 크지 않다.

한편 ㉘번의 圖表는 PREPLOT를 選擇하여 구하였다. 이 圖表에서는 2개 Factor를 軸으로 하여 5개 變數들의 Loading 값으로 圖表化하였음을 알 수 있다. 이 圖表의 數는 Factor의 數(=n)의 ${}_nC_2$ 만큼 생기며 ㉘번의 圖表에서는 E와 B, C와 A가 近接해 있음을 볼 수 있고 D 역시 Factor 1에 높게 Loading됨을 알수 있다.

㉛번부터 ㊽번까지는 回轉을 보기 위해 ROTATE = PROMAX를 選擇하여 그 結果를 얻었다. 回轉을 하는 理由는 解釋을 보다 容易하게 하기 위하여 Factor들로 된 軸을 變數들이 軸에 보다 가깝게, 즉 1개 Factor에 많이 Loading되게 움직이는 것이다. PROMAX를 취하게 되면 2段階 作業을 하게 되는데 먼저 Varimax, 즉 Orthogonal Transformation을 하게 되고 그 다음으로 Procrustean Transformation을 하게 된다. 또한 이들 두 Transformation의 곱인 Oblique Transformation을 얻게 된다.

Transformation을 보다 쉽게 說明하자면, 例를 들어 圖表上에 2개의 變數群이 있다고 하자(이제까지의 例에서는(E와 B) 그리고 (C와 A) 또는 (E, B와 D) 그리고 (C와 A), 이 變數群

들을 가로지르는 軸을 생각할 때, 두 軸 사이가 直角이면 Orthogonal Transformation, 直角이 아닐 때는 Oblique Transformation을 使用하게 된다(단, Factor가 Correlate일 때).

㉝, ㊳번에 나타난 要因類型行列表를 보면 類似한 結果를 보이고 있으며 또한 Principal Components와도 크게 어긋나지 않는다.

㊽번 다음에 Print된 Dataset는 PROC FACTOR의 結果로 만들어진 Output Dataset를 Print한 것이다. 여기서 만들어진 Output Dataset를 利用해서 또 다른 回轉方法인 Harris-Kaiser의 Orth-oblique Rotation을 할 수 있다. Harris-Kaiser Rotation은 Output Dataset에 있는 Data중에 Promax에서 使用된 -Type- = 'Pattern'과 -Type- = 'Fcorr'은 삭제하고 그 대신에 -Type- = 'Unorotate'로 된 것은 -Type- = 'Pattern'으로 바꾸어서 이를 Input Data로 使用한다.

그 例는 다음과 같다.

```
DATA FACT2(TYPE=FACTOR);
 SET;
 IF__TYPE_='PATTERN'|_TYPE_='FCORR' THEN DELETE;
 IF_TYPE_='UNROTATE' THEN_TYPE_='PATTERN';
PROC FACTOR ROTATE=HK NORM=WEIGHT REORDER PLOT;
 TITLE3 HARRIS-KAISER ROTATION WITH CURETON-MULAIK WE-
 IGHTS;
```

（分析結果）

FIVE SOCIO-ECONOMIC VARIABLES 13

SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED

HARRIS-KAISER ROTATION WITH CURETON-MULAIK WEIGHTS

ROTATION METHOD: HARRIS-KAISER

㉙ VARIABLE WEIGHTS FOR ROTATION

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.959827	0.939454	0.997464	0.121948	0.940073

OBLIQUE TRANSFORMATION MATRIX

	1	2
1	0.73537	0.61899
2	-0.68283	0.78987

INTER-FACTOR CORRELATIONS

	FACTOR1	FACTOR2
FACTOR1	1.00000	0.08358
FACTOR2	0.08358	1.00000

ROTATED FACTOR PATTERN (STD REG COEFS)

	FACTOR1	FACTOR2
HOUSE	0.94048	0.00279
SCHOOL	0.90391	0.00327
SERVICES	0.75459	0.41892
POP	-0.06335	0.99227
EMPLOY	0.06152	0.97885

REFERENCE AXIS CORRELATIONS

	FACTOR1	FACTOR2
FACTOR1	1.00000	-0.08358
FACTOR2	-0.08358	1.00000

REFERENCE STRUCTURE(SEMIPARTIAL CORRELATIONS)

	FACTOR1	FACTOR2
HOUSE	0.93719	0.00278
SCHOOL	0.90075	0.00326
SERVICES	0.75195	0.41745
POP	-0.06312	0.98880
EMPLOY	0.06130	0.97543

VARIANCE EXPLAINED BY EACH FACTOR ELIMINATING OTHER FACTORS

FACTOR1	FACTOR2
2.262854	2.103473

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

ROTATION METHOD: HARRIS-KAISER

FACTOR STRUCTURE (CORRELATIONS)

	FACTOR1	FACTOR2
HOUSE	0.94071	0.08139
SCHOOL	0.90419	0.07882
SERVICES	0.78960	0.48198
POP	0.01958	0.98698
EMPLOY	0.14332	0.98399

VARIANCE EXPLAINED BY EACH FACTOR IGNORING OTHER FACTORS

FACTOR1	FACTOR2
2.346896	2.187516

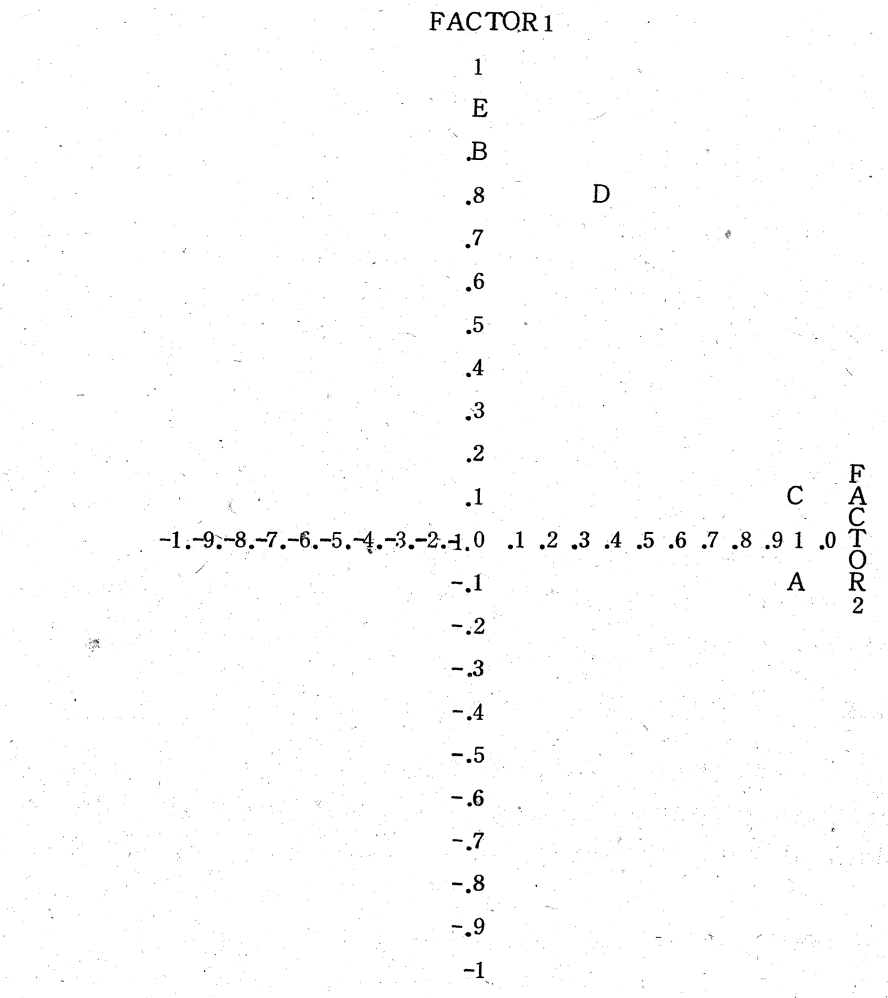
FINAL COMMUNALITY ESTIMATES: TOTAL = 4.450370

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
HARRIS-KAISER ROTATION WITH CURETON-MULAIK WEIGHTS

ROTATION METHOD: HARRIS-KAISER

PLOT OF REFERENCE STRUCTURE FOR FACTOR1 AND FACTOR2
REFERENCE AXIS CORRELATION= -0.0836 ANGLE= 94.79



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

3) Maximum Likelihood Factor Analysis

앞에서 言及한 바와 같이 ML Factor Analysis를 하기 앞서 正確한 Factor 數를 指定해 주는 것이 좋다. 즉 (2)의 Principal Factor Analysis로 Factor 數를 찾아낸 후에 MC의 特性인 바람직한 近似的(Asymptotic) 推定値를 찾는 것이다. 이 方法은 大標本인 경우에 Principal Factor Analysis 보다 좋은 推定値를 얻을 수 있다. 여기에서는 Factor 數가 1개, 2개, 3개일 경우를 각각 별도로 分析하여 比較해 보자. 入力資料는 앞의 分析에 利用한 것을 그대로 使用한다.

ⓐ Factor 數가 1인 경우

```
PROC FACTOR DATA = SOCECON METHOD = ML HEYWOOD N=1;
 TITLE3 MAXIMUM LIKELIHOOD FACTOR ANALYSIS WITH ONE
 FACTOR;
```

(分析結果): Factor 數가 1인 경우

```
FIVE SOCIO-ECONOMIC VARIABLES                    16
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
MAXIMUM LIKELIHOOD FACTOR ANALYSIS WITH ONE FACTOR

INITIAL FACTOR METHOD: MAXIMUM LIKELIHOOD

          PRIOR COMMUNALITY ESTIMATES:SMC

     POP        SCHOOL     EMPLOY     SERVICES   HOUSE
  0.968592     0.822285   0.969181   0.785724   0.847019

⑩ PRELIMINARY EIGENVALUES: TOTAL=76.116586  AVERAGE=15.223317
```

	1	2	3	4	5
EIGENVALUE	63.701009	13.054719	0.327639	-0.347281	-0.619501
DIFFERENCE	50.646289	12.727080	0.674920	0.272220	
PROPORTION	0.8369	0.1715	0.0043	-0.0046	-0.0081
CUMULATIVE	0.8369	1.0084	1.0127	1.0081	1.0000

1 FACTORS WILL BE RETAINED BY THE NFACTOR CRITERION

⑬

ITER	CRITERION	RIDGE	CHANGE	COMMUNALITIES				
1	6.54292	0.000	0.10330	0.93828	0.72227	1.00000	0.71940	0.74371
2	3.12327	0.000	0.72885	0.94566	0.02380	1.00000	0.26493	0.01487

UNABLE TO IMPROVE CRITERION
TRY A DIFFERENT 'PRIORS' STATEMENT.

SIGNIFICANCE TESTS BASED ON 12 OBSERVATIONS:

TEST OF HO: NO COMMON FACTORS.
VS HA: AT LEAST ONE COMMON FACTOR.

⑭ CHI-SQUARE = 54.252 DF = 10 PROB>CHI**2 = .0001
TEST OF HO: 1 FACTORS ARE SUFFICIENT.
VS HA: MORE FACTORS ARE NEEDED

CHI-SQUARE = 24.466 DF = 5 PROB>CHI**2 = .0002

⑮ AKAIKE'S INFORMATION CRITERION = 57.47924
⑯ SCHWARZ'S BAYESIAN CRITERION = 31.16415

⑰ SQUARED CANONICAL CORRELATIONS

FACTOR1
1.000000

EIGENVALUES OF THE WEIGHTED REDUCED CORRELATION MATRIX:
TOTAL = -0.000000 AVERAGE = -0.000000

	1	2	3	4	5
EIGENVALUE		1.927160	-0.228313	-0.792956	-0.905891
DIFFERENCE	-1.927160	2.155473	0.564643	0.112935	

FIVE SOCIO-ECONOMIC VARIABLES 17
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
MAXIMUM LIKELIHOOD FACTOR ANALYSIS WITH ONE FACTOR

INITIAL FACTOR METHOD: MAXIMUM LIKELIHOOD

FACTOR PATTERN

	FACTOR 1
POP	0.97245
SCHOOL	0.15428
EMPLOY	1.00000
SERVICES	0.51472
HOUSE	0.12193

VARIANCE EXPLAINED BY EACH FACTOR

	FACTOR 1
WEIGHTED	17.801063
UNWEIGHTED	2.249260

FINAL COMMUNALITY ESTIMATES AND VARIABLE WEIGHTS
TOTAL COMMUNALITY: WEIGHTED = 17.801603 UNWEIGHTED=2.249260

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
COMMUNALITY	0.945656	0.023803	1.000000	0.264935	0.014866
WEIGHT	18.401165	1.024384		1.360424	1.015090

Factor를 1개로 指定해 주었을 때 分析結果表의 ⑬에서와 같이 두번째 反復計算(Iteration)에서 最適値를 구하였지만 이는 수렴기준(Convergence Criterion)을 만나지 못하였기 때문에 ⑬번 밑에 "Unable to Improve Criterion", "Try a Different 'Priors' Statement"라는 것이 나타나 있다. 이때는 解를 개선하기 위해 SMC로 주어진 事前 Communality Estimates를 바꾸어 주어 다시 分析해야 되지만 같거나 또는 더 나쁜 最適値의 解를 구하게 되기 쉽다.

⑭번의 Common Factors의 存在에 대한 假說檢定에서 보듯이 1개 以上의 Factor가 있음을 알 수 있다(Chi-Square = 54.25, Pr = 0.0001). 또한 1개의 Factor로써 充分하지 못하다는 結論을 내릴 수 있다.(Chi-Square = 24.466, Pr = 0.0002).

⑮번과 ⑯번의 AIC(Akaike's Information Criterion)나 SBC(Schwarz's Bayesian Criterion)는 ⑭번의 Chi-Square와 비슷하며 母數의 數, 여기서는 Factor 數에 대한 檢定統計量으로 그들의 값(AIC=57.48, SBC = 31.16)들은 적을수록 Factor의 數는 妥當한 것으로 알려져 있다.

⑰번의 SCC(Squared Canonical Correlation)는 變數들로서 Factor를 豫測하는데의 SMC(Squared Multiple Correlation)와 같고 그 밑에 Eigenvalue가 U와 같다. 여기서 첫번째 Eigenvalue가 Missing인 이유는, 結果表 제일 아랫쪽 Employ의 Final Communality가 1.0이므로 Weight가 無限大가 되고 이에 따라 Weight가 Missing으로 Print되며 결국은 첫번째 Eigenvalue 역시 Missing이 된다. 따라서 Total과 Average도 計算되지 않았다.

⑥ Factor의 數가 2인 경우

```
PROC FACTOR DATA=SOCECON METHOD=ML HEYWOOD N=2;
  TITLE3 MAXIMUM LIKELIHOOD FACTOR ANALYSIS WITH TWO
  FACTORS;
```

(分析結果) : Factor 數가 2인 경우

```
FIVE SOCIO-ECONOMIC VARIABLES                          18
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
MAXIMUM LIKELIHOOD FACTOR ANALYSIS WITH TWO FACTORS

INITIAL FACTOR METHOD: MAXIMUM LIKELIHOOD

                PRIOR COMMUNALITY ESTIMATES:SMC

      POP        SCHOOL     EMPLOY     SERVICES   HOUSE
    0.968592     0.822285   0.969181   0.785724   0.847019

PRELIMINARY EIGENVALUES: TOTAL = 76.116586  AVERAGE = 15.223317

                1           2           3           4           5
EIGENVALUE      63.701009   13.054719   0.327639   -0.347281   -0.619501
DIFFERENCE      50.646289   12.727080   0.674920    0.272220
PROPORTION      0.8369      0.1715      0.0043     -0.0046     -0.0081
CUMULATIVE      0.8369      1.0084      1.0127      1.0081      1.0000
```

2 FACTORS WILL BE RETAINED BY THE NFACTOR CRITERION

ITER	CRITERION	RIDGE	CHANGE	COMMUNALITIES				
1	0.343122	0.000	0.04710	1.00000	0.80672	0.95058	0.79348	0.89412
2	0.307218	0.000	0.03068	1.00000	0.80821	0.96023	0.81048	0.92480
3	0.306786	0.000	0.00629	1.00000	0.81149	0.95948	0.81677	0.92023
4	0.306737	0.000	0.00218	1.00000	0.80985	0.95963	0.81498	0.92241
5	0.306732	0.000	0.00071	1.00000	0.81019	0.95955	0.81569	0.92187

CONVERGENCE CRITERION SATISFIED.

SIGNIFICANCE TESTS BASED ON 12 OBSERVATIONS:

TEST OF HO: NO COMMON FACTORS.
VS HA: AT LEAST ONE COMMON FACTOR.

CHI-SQUARE= 54.252 DF = 10 PROB>CHI**2 = .0001

TEST OF HO: 2 FACTORS ARE SUFFICIENT.
VS HA: MORE FACTORS ARE NEEDED.

CHI-SQUARE= 2.198 DF = 1 PROB>CHI**2 = .1382

AKAIKE'S INFORMATION CRITERION= 31.68079
SCHWARZ'S BAYESIAN CRITERION = 19.23474

SQUARED CANONICAL CORRELATIONS

FACTOR1	FACTOR2
1.000000	0.951889

EIGENVALUES OF THE WEIGHTED REDUCED CORRELATION MATRIX:
TOTAL = 19.785316 AVERACE = 3.957063

	1	2	3	4	5
EIGENVALUE		19.785314	0.543185	-0.039771	-0.503412
DIFFERENCE	-19.785314	19.242129	0.582956	0.463641	
PROPORTION	0.0000	1.0000	0.0275	-0.0020	-0.0254
CUMULATIVE	0.0000	1.0000	1.0275	1.0254	1.0000

FIVE SOCIO-ECONOMIC VARIABLES 19
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
MAXIMUM LIKELIHOOD FACTOR ANALYSIS WITH TWO FACTORS

INITIAL FACTOR METHOD: MAXIMUM LIKELIHOOD

FACTOR PATTERN

	FACTOR1	FACTOR2
POP	1.00000	0.00000
SCHOOL	0.00975	0.90003
EMPLOY	0.97245	0.11797
SERVICES	0.43887	0.78930
HOUSE	0.02241	0.95989

VARIANCE EXPLAINED BY EACH FACTOR

	FACTOR1	FACTOR2
WEIGHTED	24.432971	19.785314
UNWEIGHTED	2.138861	2.368353

FINAL COMMUNALITY ESTIMATES AND VARIABLE WEIGHTS
TOTAL COMMUNALITY: WEIGHTED = 44.218285 UNWEIGHTED = 4.507214

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
COMMUNALITY	1.000000	0.810145	0.959571	0.815603	0.921894
WEIGHT		5.268294	24.724667	5.425646	12.799679

여기에서는 다섯번째 反復計算(Iteration)에서 最適値가 구해졌고 그 밑에 "Convergence Criterion Satisfied"라고 Print 되어 있다. 또 Factor 數가 2인 경우의 檢定統計量은 Chi-Square = 2.198, Pr = 0.1382로서 假說을 기각하지 못하며 AIC와 SBC 역시 앞의 Factor 數 1인 경우와 比較하여 매우 적으므로 Factor 數 2가 妥當하다고 생각할 수 있다. 그러나 이번 경우에는 變數 Pop이 Heywood Case이다.

〈Heywood Case〉

Communality는 Squared Correlation이기 때문에 1과 0 사이에 있을 것이라고 예상되지만 Final Communality는 1과 같거나 1을 超過할 수도 있다. 이런 경우 Final Communality가 1이면 Heywood Case, 1을 超過하면 Ultra-Heywood Case라고 한다.

Ultra-Heywood Case란 特殊要因(Specific Factor or Unique Factor)가 陰의 分散을 갖는다. 즉 무엇인가가 잘못되었다는 것을 意味하며 그 原因으로는 아래와 같은 경우를 들 수 있다.

- 事前 Communality 推定値가 좋지 않다.
- Common Factor가 너무 많다.
- Common Factor가 너무 적다.
- 資料가 安定된 推定値를 얻기에 不充分하다.
- 資料分析에 Factor Analysis가 적절하지 못하다.

Ultra-Heywood Case인 경우 그 要因解는 妥當하지 못하다고 볼 수 있다. 이 외의 여러가지 類似한 경우를 살펴보면

1. Quasi-Heywood Case : 變數의 Commonunality가 그 變數의 Reliability를 超過할 때

2, 最終 Communality Estimate 가 Squared Multiple Correlation 보다 적은 경우(이것은 Factor Model Fit 가 나쁘다.)

3. 하나의 要因(Factor)과 變數들간의 Squared Multiple Correlation이 1을 超過하는 경우(Alpha Factor Analysis의 경우에 많다.)

4. Squared Multiple Correlation이 陰인 경우
(너무 要因 數가 많은 경우이다.)

5. Eigenvalue가 陰인 경우(資料가 Common Factor Model에 맞지 않는 경우이다.)

6. 하나의 要因이 共通分散(Common Variance의 100% 以上을 說明하는 경우(事前 Communality estimate가 너무 낮은 경우이다)

등이 있다. 以上의 경우 또한 Ultra-Heywood Case와 같이 注意를 必要로 한다.

Principal Component Analysis에서 缺測値(Missing Values)가 없는 資料로 計算된 共分散 또는 相關關係行列을 利用하면 위와 같은 경우들은 생기지 않는다.

© Factor 의 數가 3인 경우

```
PROC FACTOR DATA=SOCECON METHOD=ML HEYWOOD N=3;
  TITLE3 MAXIMUM LIKELIHOOD FACTOR ANALYSIS WITH
  THREE FACTORS;
```

(分析結果) : Factor 의 數가 3인 경우

```
FIVE SOCIO-ECONOMIC VARIABLES                              20
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
MAXIMUM LIKELIHOOD FACTOR ANALYSIS WITH THREE FACTORS

INITIAL FACTOR METHOD: MAXIMUM LIKELIHOOD

PRIOR COMMUNALITY ESTIMATES: SMC

      POP      SCHOOL     EMPLOY     SERVICES     HOUSE
   0.968592   0.822285   0.969181   0.785724   0.847019

PRELIMINARY EIGENVALUES: TOTAL = 76.116586  AVERAGE = 15.223317

                1           2           3           4           5
EIGENVALUE    63.701009   13.054719   0.327639   -0.347281   -0.619501
DIFFERENCE    50.646289   12.727080   0.674920    0.272220
PROPORTION     0.8369      0.1715     0.0043     -0.0046     -0.0081
CUMULATIVE     0.8369      1.0084     1.0127      1.0081      1.0000

  3 FACTORS WILL BE RETAINED BY THE NFACTOR CRITERION
    WARNING: TOO MANY FACTORS FOR A UNIQUE SOLUTION.

ITER  CRITERION     RIDGE  CHANGE   COMMUNALITIES
  1   0.160126      0.031  0.05102  0.96382  0.84123  1.00000  0.80346  0.89804
  2   0.00340681    0.031  0.05878  0.98216  0.87692  1.00000  0.80295  0.95682
  3   3.000041434   0.031  0.01013  0.98334  0.88004  1.00000  0.80507  0.96695
  4   1.472E-06     0.031  0.00155  0.98316  0.88054  1.00000  0.80480  0.96850
  5   8.799E-08     0.031  0.00029  0.98311  0.88065  1.00000  0.80462  0.96879

      CONVERGED, BUT NOT TO A PROPER OPTIMUM.
      TRY A DIFFERENT 'PRIORS' STATEMENT
```

SIGNIFICANCE TESTS BASED ON 12 OBSERVATIONS:

TEST OF HO: NO COMMON FACTORS.
VS HA: AT LEAST ONE COMMON FACTOR.

CHI-SQUARE= 54.252 DF = 10 PROB>CHI**2 = .0001

TEST OF HO: 3 FACTORS ARE SUFFICIENT.
VS HA: MORE FACTORS ARE NEEDED.

CHI-SQUARE = 0.000 DF = -2 PROB>CHI**2 = .0000

AKAIKE'S INFORMATION CRITERION = 34
SCHWARZ'S BAYESIAN CRITERION = 21.12171

SQUARED CANONICAL CORRELATIONS

FACTOR1	FACTOR2	FACTOR3
1.000000	0.975846	0.699066

EIGENVALUES OF THE WEIGHTED REDUCED CORRELATION MATRIX:
TOTAL = 42.724206 AVERAGE = 8.544841

	1	2	3	4	5
EIGENVALUE		40.401173	2.322986	0.000319	-0.000273
DIFFERENCE	-40.401173	38.078186	2.322667	0.000591	
PROPORTION	0.0000	0.9456	0.0544	0.0000	-0.0000
CUMULATIVE	0.0000	0.9456	1.0000	1.0000	1.0000

FIVE SOCIO-ECONOMIC VARIABLES 21
SEE PAGE 14 OF HARMAN: MODERN FACTOR ANALYSIS, 3RD ED
MAXIMUM LIKELIHOOD FACTOR ANALYSIS WITH THREE FACTORS

INITIAL FACTOR METHOD: MAXIMUM LIKELIHOOD

FACTOR PATTERN

	FACTOR1	FACTOR2	FACTOR3
POP	0.97245	-0.11188	-0.15792
SCHOOL	0.15428	0.88881	0.25859
EMPLOY	1.00000	-0.00000	-0.00000
SERVICES	0.51472	0.72429	-0.12272
HOUSE	0.12193	0.97335	-0.08071

VARIANCE EXPLAINED BY EACH FACTOR

	FACTOR1	FACTOR2	FACTOR3
WEIGHTED	58.032408	40.401173	2.322986
UNWEIGHTED	2.249260	2.274514	0.113382

FINAL COMMUNALITY ESTIMATES AND VARIABLE WEIGHTS

TOTAL COMMUNALITY: WEIGHTED = 100.757 UNWEIGHTED = 4.637157

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
COMMUNALITY	0.983113	0.880658	1.000000	0.804594	0.968791
WEIGHT	59.218853	8.378825		5.118261	32.040674

Factor의 數가 3인 경우 5번 反復計算 후 수렴은 되지만 最適値가 되지 못하여 "Converged, But Not To a Proper Optimum"이 Print 되어 있고 反復計算結果表 윗 부분에 "Warning : Too Many Factors For a Unique Solution"이라 하여 模型(Model)의 母數(Parameter)의 數가 相關關係行列에서 推定될 要素(Element) 보다 많다는 것을 表示하고 있다.

한편 檢定統計量을 보면 3개 要因에 대한 假說檢定에서 自由度(Degree of Freedom)가 -2로 確率을 計算할 수 없고, AIC와 SBC는 Factor 數가 2개인 경우보다 크고, Factor 數가 1개인 경우와 같이 變數 Employ가 Heywood Case이다.

以上의 3가지 경우의 分析을 통해서 Factor 數가 2개인 경우가 가장 妥當하다고 結論지을 수 있다.

3. 主成分分析(Principal Component Analysis)

主成分分析(P.C.A)은 여러 變數들 간의 關係를 調査하는 多變量技法이다. 다시 말하여 資料의 情報를 要約하고 그들 간의 線型關係를 찾아내는데 利用된다. 主成分分析은 많은 變數들을 몇개의 線型結合, 즉 Principal Component인 새로운 變數로 줄여서 이를 變數間의 回歸나 集落化하는데 使用하게 된다.

P.C (Principal Component)의 特性을 要約하면 다음과 같다.

① 變數들 數 만큼의 P.C가 計算되며 각 P.C는 原變數들의 線型結合으로 구성되고 그 係數들은 原變數들의 相關關係行列에서의 Eigenvector와 같다.

② Eigenvectors는 直交하므로 P.C 역시 原變數空間에서 相互直交的 方向을 갖는다.

③ P.C Score는 Unrelated이다.

④ 첫번째 P.C는 原變數들의 線型結合들 중 가장 큰 分散을 갖고 두번째 P.C는 그 다음 큰 分散을 갖는다.

⑤ 처음 j개 P.C들은 $Y = XB + E$ 模型의 最小自乘解를 준다.
(단, Y : 原變數行列(n×p), X : 처음 j개 P.C의 Score인 (n×j)行列, E : (n×p)誤差行列)

SAS Program Coding

```
PROC PRINCOMP Options ;        ┐
VAR Variables ;                ┘ Needed

PARTIAL Variables ;            ┐
FREQ Variables ;               ┘ Optional
```

WEIGHT Variables ;

BY Variables ;

① PROC PRINCOMP Options

이 Statement의 Option 部分에 나타날 수 있는 것들 다음과 같다.

- DATA = SAS dataset : 分析할 入力 dataset name
- OUT = SAS dataset : 分析 후 原資料들과 P.C의 Scors를 담게 될 Dataset name
- OUTSTAT = SAS dataset : 入力資料들의 平均, 標準偏差, 相關係數, 共分散, eigenvalue, eigenvector 등을 담게 될 dataset name
- NOINT : 回歸模型에서 intercept를 使用하지 않을 때
- COV (= COVARIANCE) : P.C를 共分散行列 (covariance matrix) 로 부터 計算할 때
- N = n : 計算할 P.C의 갯수
- STD (= STANDARD) : output dataset에 있는 P.C score가 單位分散 (Unit Variance), 즉 分散이 1이 되도록 標準化할 때
- PREFIX = name : P.C들을 指稱할 이름 (default는 PRINT 1 , PRINT 2 , ……)
- NOPRINT : print하지 않을 때

② VAR statement

- VAR Variables : 分析할 對象인 變數들을 나열

③ PARTIAL statement

- PARTIAL variables : partial correlation 또는 共分散 (covariance)

行列을 分析할 때의 變數들

④ FREQ statement

- FREQ variables : 觀測値의 어떤 값이 일어난 횟수를 表示하는 變數(自由度를 決定할 때 總觀測値 數는 frequency variable의 合이 된다.)

⑤ WEIGHT statement

- WEIGHT variables : 얻어진 각 觀測値의 相對的 加重値를 必要로 할 때 이를 指定해 주는 變數

⑥ BY statement

- BY variables : 觀測値를 어떤 性質別로 grouping하여 分析하고자 할 때 이를 指定하는 變數, 이때는 dataset이 이 變數에 대하여 sort되어 있어야 한다.

分析結果表에 나타나는 統計量을 나열해 보면 다음과 같다.

① 各 變數들의 平均, 標準偏差(option(SIMPLE)이 있을 때)

② Correlations or Covariance Matrix

③ Total Variance(option(COV)이 있을 때)

④ Eigenvalues(of the correlation or covariance matrix)

⑤ Eigenvactors

(例題研究)

〈例題 1〉

몇개 都市의 1月과 7月의 平均溫度를 分析하고자 한다.

```
DATA TEMPERAT ;
  TITLE MEAN TEMPEATURE IN JANUARY AND JULY FOR SELECTED
  CITIES ;
```

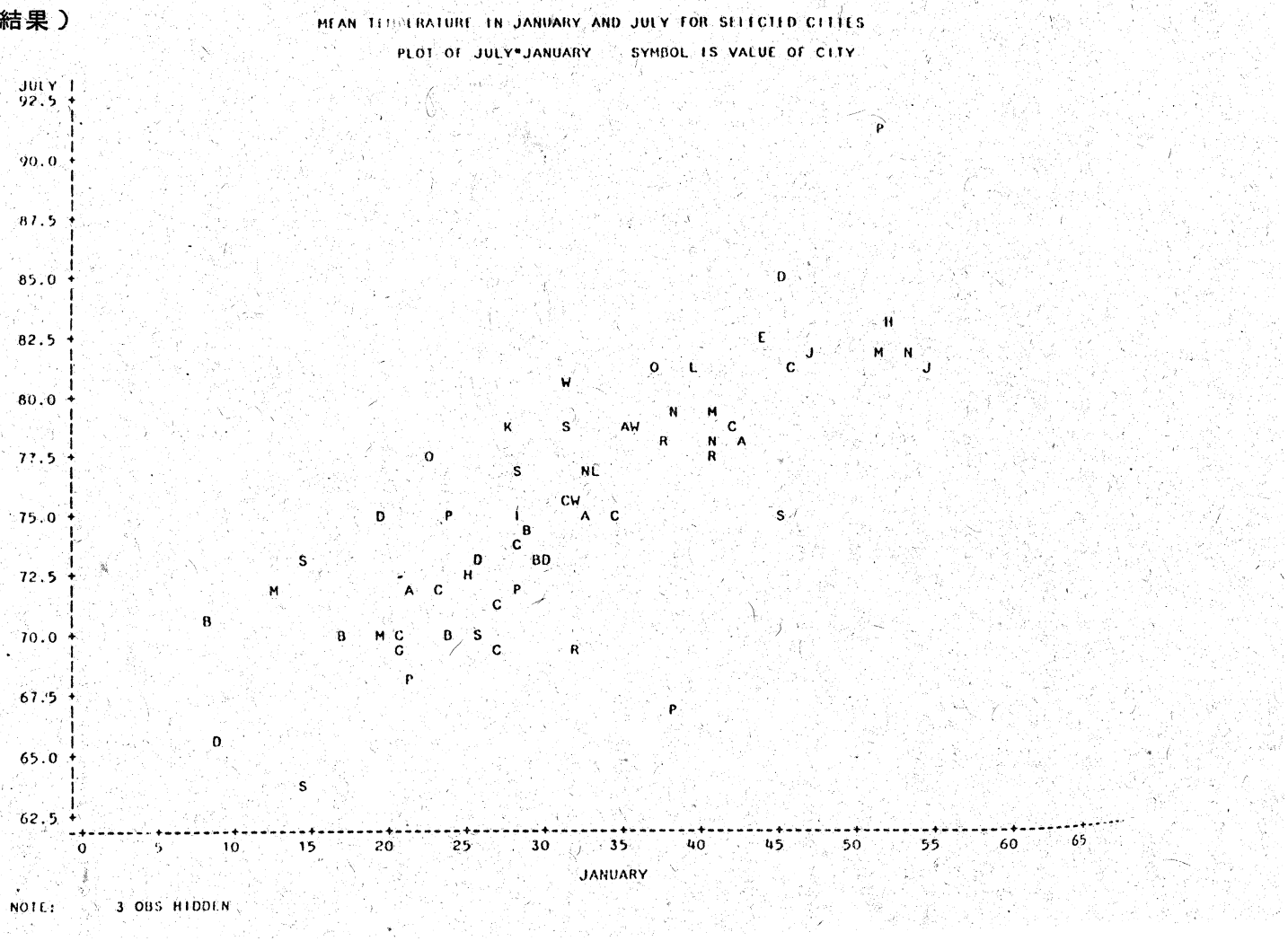
```
INPUT CITY $1-15 JANUARY JULY ;
CARDS ;
```

MOBILE	51.2	81.6
PHOENIX	51.2	91.2
LITTLE ROCK	39.5	81.4
SACRAMENTO	45.1	75.2
DENVER	29.9	73.0
HARTFORD	24.8	72.7
WILMINGTON	32.0	75.8
WASHINGTON, DC	35.6	78.7
JACKSONVILLE	54.6	81.0
MIAMI	67.2	82.3
ATLANTA	42.4	78.0
BOISE	29.0	74.5
CHICAGO	22.9	71.9
PEORIA	23.8	75.1
INDIANAPOLIS	27.9	75.0
DES MOINES	19.4	75.1
WICHITA	31.3	80.7
LOUISVILLE	33.3	76.9
NEW ORLEANS	52.9	81.9
PORTLAND, MAINE	21.5	68.0
BALTIMORE	33.4	76.6
BOSTON	29.2	73.3

DETROIT	25.5	73.3
SAULT STE MARIE	14.2	63.8
DULUTH	8.5	65.6
MINNEAPOLIS	12.2	71.9
JACKSON	47.1	81.7
KANSAS CITY	27.8	78.8
ST LOUIS	31.3	78.6
GREAT FALLS	20.5	69.3
OMAHA	22.6	77.2
RENO	31.9	69.3
CONCORD	20.6	69.7
ATLANTIC CITY	32.7	75.1
ALBUQUERQUE	35.2	78.7
ALBANY	21.5	72.0
BUFFALO	23.7	70.1
NEW YORK	32.2	76.6
CHARLOTTE	42.1	78.5
RALEIGH	40.5	77.5
BISMARCK	8.2	70.8
CINCINNATI	31.1	75.6
CLEVELAND	26.9	71.4
COLUMBUS	28.4	73.6
OKLAHOMA CITY	36.8	81.5
PORTLAND, OREG	38.1	67.1

```
PHILADELPHIA          32.3   76.8
PITTISBURGH           28.1   71.9
PROVIDENCE            28.4   72.1
COLUMBIA              45.4   81.2
SIOUX FALLS           14.2   73.3
MEMPHIS               40.5   79.6
NASHVILLE             38.3   79.6
DALLAS                44.8   84.8
EL PASO               43.6   82.3
HOUSTON               52.1   83.3
SALT LAKE CITY        28.0   76.7
BURLINGTON            16.8   69.8
NORFOLK               40.5   78.3
RICHMOND              37.5   77.9
SPOKANE               25.4   69.7
CHARLESTON, WV        34.5   75.0
MILWAUKEE             19.4   69.9
CHEYENNE              26.6   69.1
;
PROC PLOT ;
  PLOT JULY*JANUARY=CITY/ VPOS=36 ;
PROC PRINCOMP COV OUT=PRIN ;
  VAR JULY JANUARY ;
PROC PLOT ;
PLOT PRIN2*PRIN1=CITY/ VPOS=26 ;
TITLE2 PLOT OF PRINCIPAL COMPONENTS ;
```

（分析結果）


MEAN TEMPERATURE IN JANUARY AND JULY FOR SELECTED CITIES
PLOT OF JULY*JANUARY SYMBOL IS VALUE OF CITY
JULY
92.5
90.0
87.5
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
65.0
62.5
0
5
10
15
20
25
30
35
40
45
50
55
60
65
JANUARY
NOTE: 3 OBS HIDDEN

MEAN TEMPERATURE IN JANUARY AND JULY FOR SELECTED CITIES
PRINCIPAL COMPONENT ANALYSIS

64 OBSERVATIONS
2 VARIABLES

① SIMPLE STATISTICS

	JULY	JANUARY
MEAN	75.60781	32.09531
ST DEV	5.12762	11.71243

② COVARIANCES

	JULY	JANUARY
JULY	26.292	46.828
JANUARY	46.828	137.18

③ TOTAL VARIANCE = 163.4736

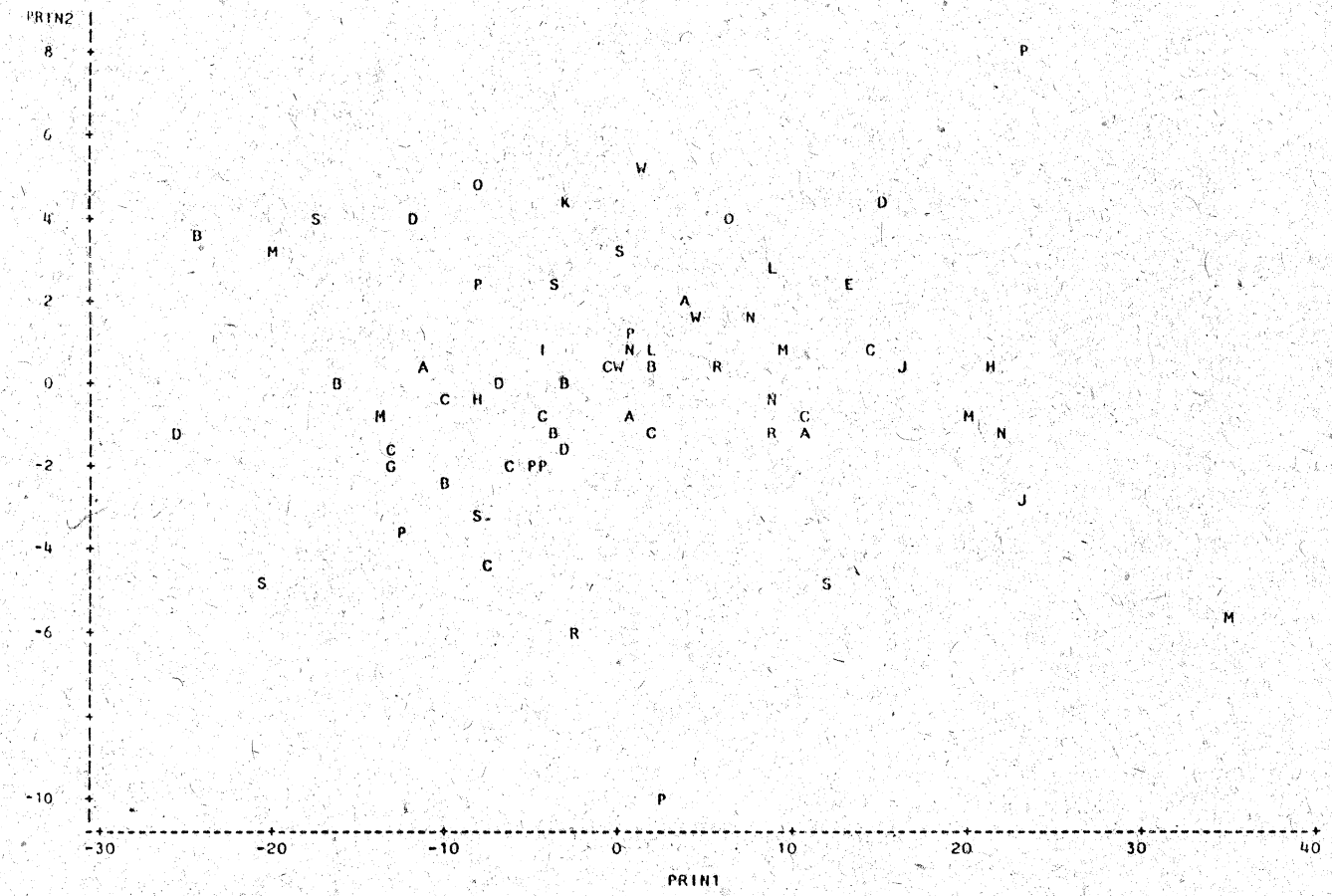
	④ EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN 1	154.3106	145.1476	0.9439	0.9439
PRIN 2	9.1630		0.0561	1.0000

⑤ EIGENVECTORS

	PRIN 1	PRIN 2
JULY	0.343532	0.939141
JANUARY	0.939141	-.343532

MEAN TEMPERATURE IN JANUARY AND JULY FOR SELECTED CITIES
PLOT OF PRINCIPAL COMPONENTS

PLOT OF PRIN2*PRIN1 SYMBOL IS VALUE OF CITY



分析結果의 ①번은 平均과 標準偏差를, ②번은 1月과 7月의 共分散行列을 表示하고 있다. ③번은 두 變數(1月과 7月의 平均溫度)의 分散合이 나와 있는데 이는 ②번의 對角成分들의 合, 즉 26.292 + 137.18 = 163.47이 된다. ④번은 主成分分析을 하여 얻은 2개의 P.C에 대한 각각의 Eigenvalue와 그들의 差(145.1476), 그리고 각각의 比率이 나타나 있다. 한편 ⑤번은 Eigenvector를 表示하고 있으며 그 밖에 2개의 圖表는 原變數와 P.C들의 圖表이며 이를 통하여 相互比較가 가능하다.

例題 2 〉

美國 50개 주에 대해서 7가지 類型의 犯罪形態로 區分하여 10萬名當 犯罪率을 調査하였다.

```
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 ;
```

（分析結果）

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
SI 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
PRIN 1	4.114960	2.876238	0.587851	0.587851
PRIN 2	1.238722	0.512905	0.176960	0.764812
PRIN 3	0.725817	0.409385	0.103688	0.868500
PRIN 4	0.316432	0.058458	0.045205	0.913704
PRIN 5	0.257974	0.035935	0.036853	0.950558
PRIN 6	0.222039	0.097983	0.031720	0.982278
PRIN 7	0.124056		0.017722	1.000000

EIGENVECTORS

	PRIN 1	PRIN 2	PRIN 3	PRIN 4	PRIN 5	PRIN 6	PRIN 7
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	-.209895	-.057555	0.101033	0.535987	-.648117
LARCENY	0.357360	0.402319	-.539231	-.234890	0.030099	0.039406	0.601690
AUTO	0.295177	0.502421	0.568384	0.419238	0.369753	-.057298	0.147046

分析結果에는 먼저 資料들에 대한 간단한 統計量(平均, 標準偏差)들과 變數들 간의 相關關係가 나와 있다. 그 밑에 7개의 P.C 가 나와 있고 각각의 Eigenvalue와 그들의 比率이 計算되어 있는데 첫번째 P.C는 4.1149 (Eigenvalue)이고 全體의 58.79 %, 두번째 P.C는 1.2387 (Eigenvalue)이고, 全體의 17.69 %, 세번째 P.C는 0.7258 (Eigenvalue)이고 全體의 10.37 %로서 이 세개의 P.C가 全體의 86.9 %를 說明할 수 있고 나머지 4개의 P.C는 각각 5% 미만의 說明力 밖에 없는 것으로 나타나 있다.

주어진 Eigenvectors의 값을 살펴보면 PRIN 1 (첫번째 P.C)에 모든 變數들이 거의 비슷한 積載値(Loading)를 가지므로 PRIN 1은 全般的인 犯罪率의 測度가 된다.

PRIN 2 (두번째 P.C)에 AUTO와 LARCENY는 陽, MURDER와 ASSAULT는 陰의 큰 積載値(Loading)를 갖고 한편으로 BURGLARY는 적은 陽의, RATE는 적은 陰의 積載値(Loading)를 갖고 있다. 이로써 暴力的 犯罪에 대한 財産的 犯罪의 優勢를 어느 程度 測定할 수 있다. 한편 PRIN 3 (세번째 P.C)는 조금 모호하다.

각 주의 犯罪率을 이들 P.C를 利用해서 把握하는 한편, 圖表를 그려서 그들의 關係를 보다 明確히 보기 위하여 Program을 연장해 보자.

```
PROC SORT ;
  BY PRIN 1 ;
PROC PRINT ;
ID STATE ;
```

```
  VAR PRIN 1 PRIN 2 MURDER RAPE ROBBERY ASSAULT BURGLARY
    LARCENY AUTO ;
  TITLE 2 STATES LISTED IN ORDER OF OVERALL CRIME RATE ;
  TITLE 3 AS DETERMINED BY THE FIRST PRINCIPAL COMPONENT ;
PROC SORT ;
  BY PRIN 2 ;
PROC PRINT ;
  ID STATE ;
  VAR PRIN 1 PRIN 2 MURDER RAPE ROBBERY ASSAULT BURGLARY
    LARCENY AUTO ;
  TITLE 2 STATES LISTED IN ORDER OF PROPERTY VS. VIOLENT
    CRIME ;
  TITLE 3 AS DETERMINED BY THE SECOND PRINCIPAL COMPO-
    NENT ;

PROC PLOT ;
  PLOT PRIN 2 PRIN 1 = STATE ;
  TITLE 2 PLOT OF THE FIRST TWO PRINCIPAL COMPONENTS ;
PROC PLOT ;
  PLOT PRIN 3 PRIN 1 = STATE ;
  TITLE 2 PLOT OF THE FIRST AND THIRD PRINCIPAL
    COMPONENTS ;
```

（分析結果）

CRIME RATES PER 100,000 POPULATION BY STATE
STATES LISTED IN ORDER OF OVERALL CRIME RATE
AS DETERMINED BY THE FIRST PRINCIPAL COMPONENT

STATE	PRIN1	PRIN2	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
NORTH DAKOTA	-3.9641	0.3877	0.9	9.0	13.3	43.8	446.1	1843.0	144.7
SOUTH DAKOTA	-3.1720	-0.2545	2.0	13.5	17.9	155.7	570.5	1704.4	147.5
WEST VIRGINIA	-3.1477	-0.8143	6.0	13.2	42.2	90.9	597.4	1341.7	163.3
IOWA	-2.5816	0.8248	2.3	10.6	41.2	89.8	812.5	2685.1	219.9
WISCONSIN	-2.5030	0.7808	2.8	12.9	52.2	63.7	846.9	2614.2	220.7
NEW HAMPSHIRE	-2.4656	0.8250	3.2	10.7	23.2	76.0	1041.7	2343.9	293.4
NEBRASKA	-2.1507	0.2257	3.9	18.1	64.7	112.7	760.0	2316.1	249.1
VERMONT	-2.0643	0.9450	1.4	15.9	30.8	101.2	1348.2	2201.0	265.2
MAINE	-1.8263	0.5788	2.4	13.5	38.7	170.0	1253.1	2350.7	246.9
KENTUCKY	-1.7269	-1.1466	10.1	19.1	81.1	123.3	872.2	1662.1	245.4
PENNSYLVANIA	-1.7201	-0.1959	5.6	19.0	130.3	128.0	877.5	1624.1	333.2
MONTANA	-1.6680	0.2710	5.4	16.7	39.2	156.8	804.9	2773.2	309.2
MINNESOTA	-1.5543	1.0564	2.7	19.5	85.9	85.8	1134.7	2559.3	343.1
MISSISSIPPI	-1.5074	-2.5467	14.3	19.6	65.7	189.1	915.6	1239.9	144.4
IDAHO	-1.4325	-0.0080	5.5	19.4	39.6	172.5	1050.8	2599.6	237.6
WYOMING	-1.4246	0.0627	5.4	21.9	39.7	173.9	811.6	2772.2	282.0

STATE	PRIN 1	PRIN 2	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
ARKANSAS	-1.0544	-1.3454	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
UTAH	-1.0500	0.9366	3.5	20.3	68.8	147.3	1171.6	3004.6	334.5
VIRGINIA	-0.9162	-0.6927	9.0	23.3	92.1	165.7	986.2	2521.2	226.7
NORTH CAROLINA	-0.6993	-1.6703	10.6	17.0	61.3	318.3	1154.1	2037.8	192.1
KANSAS	-0.6341	-0.0280	6.6	22.0	100.7	180.5	1270.4	2739.3	244.3
CONNECTICUT	-0.5413	1.5012	4.2	16.8	129.5	131.8	1346.0	2620.7	593.2
INDIANA	-0.4999	0.0000	7.4	26.5	123.2	153.5	1086.2	2498.7	377.4
OKLAHOMA	-0.3214	-0.6243	8.6	29.2	73.8	205.0	1288.2	2228.1	326.8
RHODE ISLAND	-0.2016	2.1466	3.6	10.5	86.5	201.0	1489.5	2844.1	791.4
TENNESSLE	-0.1366	-1.1350	10.1	29.7	145.8	203.9	1259.7	1776.5	314.0
ALABAMA	-0.0499	-2.0961	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
NEW JERSEY	0.2179	0.9642	5.6	21.0	180.4	185.1	1435.8	2774.5	511.5
OHIO	0.2395	0.0905	7.8	27.3	190.5	181.1	1216.0	2696.8	400.4
GEORGIA	0.4904	-1.3808	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9
ILLINOIS	0.5129	0.0942	9.9	21.8	211.3	209.0	1085.0	2828.5	528.6
MISSOURI	0.5564	-0.5585	9.6	28.3	189.0	233.5	1318.3	2424.2	378.4
HAWAII	0.8231	1.8239	7.2	25.5	128.0	64.1	1911.5	3920.4	489.4
WASHINGTON	0.9306	0.7378	4.3	39.6	106.2	224.8	1605.6	3386.9	360.3
DELAWARE	0.9646	1.2967	6.0	24.9	157.0	194.2	1682.6	3678.4	467.0
MASSACHUSETTS	0.9784	2.6311	3.1	20.8	169.1	231.6	1532.2	2311.3	1140.1
LOUISIANA	1.1202	-2.0833	15.5	30.9	142.9	335.5	1165.5	2469.9	337.7
NEW MEXICO	1.2142	-0.9508	8.8	39.1	109.6	343.4	1418.7	3008.6	259.5

STATE	PRIN 1	PRIN 2	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
TEXAS	1.3970	-0.6813	13.3	33.8	152.4	208.2	1603.1	2988.7	397.6
OREGON	1.4490	0.5860	4.9	39.9	124.1	286.9	1636.4	3506.1	388.9
SOUTH CAROLINA	1.6034	-2.1621	11.9	33.0	105.9	485.3	1613.6	2342.4	245.1
MARYLAND	2.1828	-0.1947	8.0	34.8	292.1	358.9	1400.0	3177.7	428.5
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
COLORADO	2.5093	0.9166	6.3	42.0	170.7	292.9	1935.2	3903.2	477.1
ARIZONA	3.0141	0.8449	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
FLORIDA	3.1118	-0.6039	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4
NEW YORK	3.4525	0.4329	10.7	29.4	472.6	319.1	1728.0	2782.0	745.8
CALIFORNIA	4.2838	0.1432	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5
NEVADA	5.2670	-0.2526	15.8	49.1	323.1	355.0	2453.1	4212.6	559.2

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

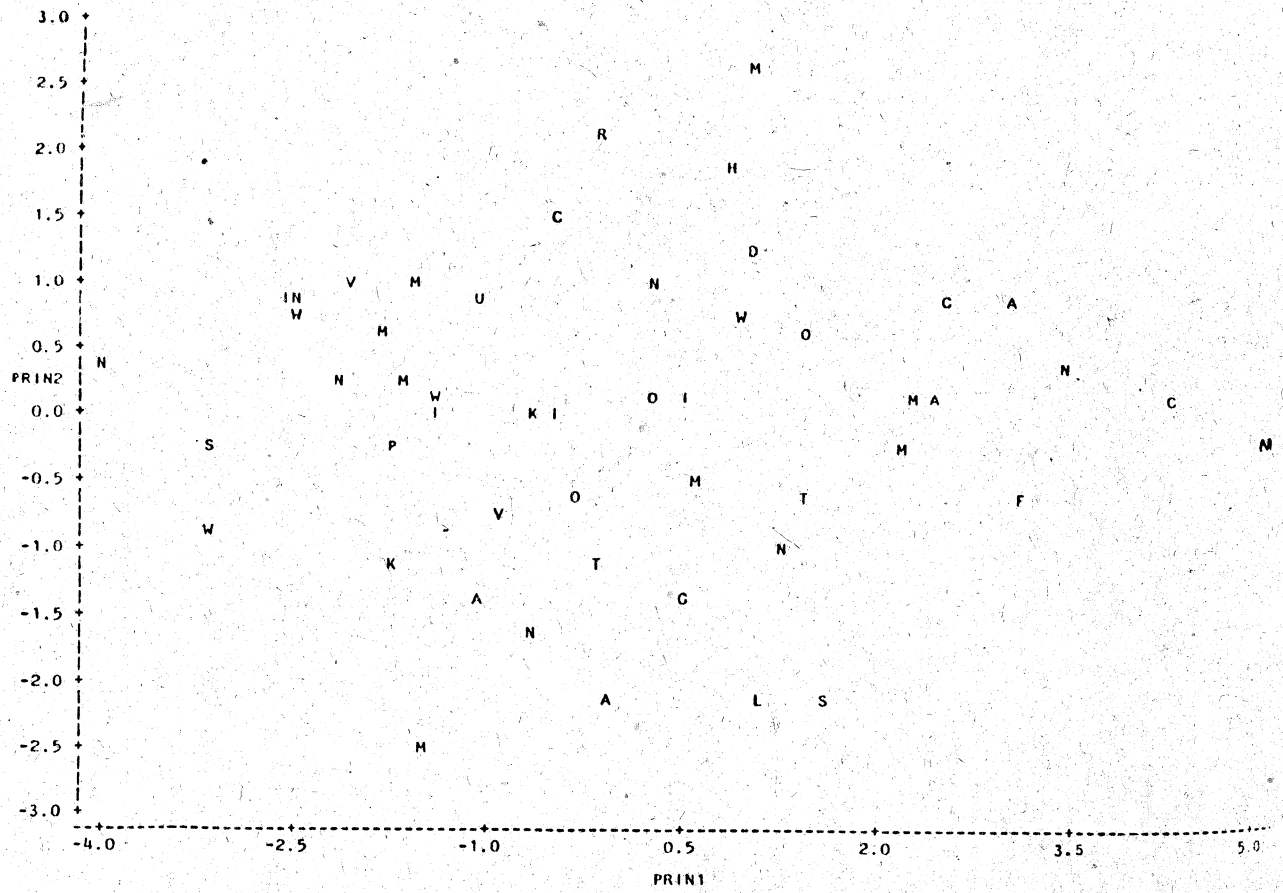
STATE	PRIN1	PRIN2	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
MISSISSIPPI	-1.5074	-2.5467	14.3	19.6	65.7	189.1	915.6	1239.9	144.4
SOUTH CAROLINA	1.6034	-2.1621	11.9	33.0	105.9	485.3	1613.6	2342.4	245.1
ALABAMA	-0.0499	-2.0961	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
LOUISIANA	1.1202	-2.0833	15.5	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	-1.1466	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 DAROIA	-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

STATE	PRIN 1	PRIN 2	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
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	-0.6341	-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	2696.8	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	738.3
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	3.4525	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
ORIGON	1.4490	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

STATE	PRIN1	PRIN2	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
COLORADO	2.5093	0.9166	6.3	42.0	170.7	292.9	1935.2	3903.2	477.1
UIAH	-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	131.8	1346.0	2620.봄	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	2.1466	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

PLOT OF PRIN2*PRIN1 SYMBOL IS VALUE OF STATE



分析結果는 앞에서 구한 P.C (PRIN 1)로 Sort 하여 全般的인 犯罪率이 낮은 順序로 주들을 Print 한 것으로 여기에서는 NORTH DAKOTA가 가장 犯罪率이 낮고 NEVADA가 가장 높다. 그 다음은 두번째 P.C (PRIN 2), 즉 暴力的 犯罪에 대한 財産的 犯罪의 比率이 낮은 順序로 Print 하였으며 여기에서는 MISSISSIPPI가 가장 낮고 MASSACHUSETTS가 가장 높다.

다음은 PRIN 1과 PRIN 2를 圖表化하여 分析해 보면 NEVADA 와 CALIFORNIA는 오른쪽으로 치우쳐 있어 全般的인 犯罪率은 높은데 반해 暴力的 犯罪에 대한 財産的 犯罪의 比率은 平均的이고 NORTH DAKOTA와 SOUTH DAKOTA는 왼쪽으로 치우쳐 全般的인 犯罪率이 낮은 것으로 要約할 수 있다. 또 남부에 있는 주들은 주로 圖表의 下段에 位置하여 財産的 犯罪에 비해 暴力的 犯罪가 優勢한 것으로 나타난 反面, New England 주들은 圖表의 上段에 位置하여 財産的 犯罪가 優勢한 것으로 나타났다. 마지막으로 PRIN 1과 PRIN 3을 보면 MASSACHUSETTS와 NEW YORK이 PRIN 3에 있어 特異値 (out-lier)로 나타났다.

Ⅳ. 參 考 文 獻

1) Anderson, T.W. (1958), An Introduction to Multivariate Statistical Analysis, New York, John Wiley & Sons.

2) Bolch, B.W. and Huang, C. J. (1974), Multivariate Statistical Methods For Business and Economics, New Jersey, Prentice.—Hall, Inc.

3) Cooley, W.W. and Lohnes, P.R. (1971), Multivariate Data Analysis, New York, John Wiley & Sons.

4) Harman, M.H. (1970), Modern Factor Analysis, The University of Chicago Press.

5) Horst, P. (1965), Factor Analysis of Data Matrices, Holt, Rinehart and Winston, Inc.

6) Kshirsagar, A.M. (1972), Multivariate Analysis, New York, Marcel Dekker.

7) Morrison, D.F. (1967), Multivariate Statistical Methods, McGraw—Hill, Inc.