Lecture #20

Chapter 9 Factor Analysis

9.1 Introduction

(a) what is factor analysis?

(1) Grouped variables

Sometimes, variables from groups
Within each group variables are highly related.
Variables between groups, however, have relatively small correlations. Thus it is reasonable to represent each group by a single characteristic or factor, which is responsible for the observed correlation

(2) Factor analysis

- One of the most widely used multivariate techniques introduced by charles spearman (1904) started as a controversial difficult subject (due to the lack of powerful computing tool). It has emerged as one of the most facinating and useful tools

Its applicability to many diverse fields such as, biology, chemistry, ecology, economics, education, political science, psychology, and socciology, etc. has been demonstrated.

- The main concern is to identify the internal relationships between a set of random variables and to describe the covariance structure in terms of a few underlying unobservable random quantities called factors.
- Spearman (1904) noticed a certain systematic pattern in a correlation matrix ρ of scores obtained in scholastic exams (of 30 preparatory school boys)

(*) look at book

$$\rho = \begin{array}{c} \\ C \\ F \\ E \\ M \\ D \\ MU \end{array} \begin{bmatrix} 1.00 & 0.83 & 0.78 & 0.70 & 0.66 & 0.63 \\ 0.83 & 1.00 & 0.67 & 0.67 & 0.65 & 0.57 \\ 0.78 & 0.67 & 1.00 & 0.64 & 0.54 & 0.51 \\ 0.70 & 0.67 & 0.64 & 1.00 & 0.45 & 0.51 \\ 0.66 & 0.65 & 0.54 & 0.45 & 1.00 & 0.40 \\ 0.63 & 0.57 & 0.51 & 0.51 & 0.40 & 1.00 \end{bmatrix}$$

Ignoring the diagonal any two rows are almost proportional

$C \to M \qquad \frac{0.83}{0.67} \approx \frac{0.78}{0.64} \approx \frac{0.66}{0.45} \approx \frac{0.63}{0.51} \approx 1.2$

He argued ρ can be explained by a model

$$X_i = L_i F + \varepsilon_i \qquad i = 1, \dots, 6$$

X_i - ith variable observation

F - random variable representing general factor.

ε_i - specific factor

$$COV(X_i, X_j)$$

$$= COV(L_i F + \varepsilon_i, L_j F + \varepsilon_j)$$

$$= COV(L_i F, L_j F) + COV(\varepsilon_i, \varepsilon_j)$$

$$= (L_i \, COV(F, F) \, L_j') \; L_i L_j V(F)$$

$$\rho(X_i, X_j) = \frac{L_i L_j V(F)}{\sqrt{L_i^2 V(F) + \sigma_i^2}\sqrt{L_j^2 V(F) + \sigma_j^2}}$$

(b) PCA versus FA

<1> PCA:

- An orthogonal transformation of the coordinate axes ~~through the natural shape of the scatter plot of observations~~, that partitioning the total variance of all response into successively smaller partitions
- If the first few axes account for most of the total variance, and if their positions could be interpreted meaningfully, then the system could be described more parsimoniously by them
- Shortcomings

1) not invariant under transformation

2) no rational criteria for selecting the number of PCs to retain

(2) FA.

- construct a fundamental model ~~of~~ for covariance structure using some unobservable variable.
- Avoid the deficiency of PCA

9.2 The orthogonal factor model

a) orthognal factor model with m common factors

1. the model

. Let X be $p \times 1$ random vector with

$$E(X) = \mu \quad \text{and} \quad COV(X) = \Sigma$$

• Assume that the inter relation ships between the elements of X can be explained by

$$X_{p\times 1} = \mu_{p\times 1} + \underset{p\times m}{L}\, F_{m\times 1} + \varepsilon_{p\times 1}$$

Where $F = (F_1, \cdots, F_m)'$ are called common factors, which are random

$L = \{l_{ij}\}_{p\times m}$ are called factor loading matrix

l_{ij}: loading of ith variable on the jth factor

$\varepsilon = (\varepsilon_1, \cdots, \varepsilon_p)'$ are called specific factors and random

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1$$

$$\vdots$$

$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p$$

2. Assumptions

• We assume that

$$E(F) = 0, \quad COV(F) = I_m \quad E(\varepsilon) = 0, \quad COV(\varepsilon) = \Psi = \begin{pmatrix} \psi_1 & & 0 \\ & \ddots & \\ 0 & & \psi_p \end{pmatrix}$$

F and ε are independent $\longrightarrow$ $COV(F, \varepsilon) = 0$

(b) properties of and comments on the model

1. population of the model

- $\Sigma = COV(X) = COV(X-\mu) = E(LF+\varepsilon)(LF+\varepsilon)'$

$= E(LFF'L' + LF\varepsilon' + \varepsilon F'L' + \varepsilon\varepsilon')$

$= \underset{p\times m}{L}\underset{m\times p}{L'} + \underset{p\times p}{\Psi}$

- $COV(X, F) = COV(LF+\varepsilon, F)$

$= L\,COV(F, F) + COV(\varepsilon, F)$

$= L$

- $\sigma_{ii} = var(X_i)$

$= var(LF+\varepsilon)_{ii}$

$= (LL' + \Psi)_{ii}$

$= \sum_{j=1}^{m} l_{ij}^2 + \psi_i$

$= h_i^2 + \psi_i$

h_i^2: ith commonality; ψ_i: special variance

Statistically, $\Sigma = LL' + \Psi$ may not be solvable (when $m < p$)

(For $m = p$, $\Sigma = P \Lambda P'$)

$$\Sigma = \begin{bmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.9 & 0.4 & 1 \end{bmatrix} \quad p = 3$$

Let $m = 1$

$$= LL' + \Psi$$

$$= \begin{bmatrix} l_{11} \\ l_{21} \\ l_{31} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \end{bmatrix} + \begin{bmatrix} \psi_1 & & 0 \\ & \psi_2 & \\ 0 & & \psi_3 \end{bmatrix}$$

$$= \begin{bmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ & l_{21}^2 & l_{21}l_{31} \\ & & l_{31}^2 \end{bmatrix} + \begin{bmatrix} \psi_1 & & 0 \\ & \psi_2 & \\ 0 & & \psi_3 \end{bmatrix}$$

$1 = l_{11}^2 + \psi_1 \quad 1 = l_{21}^2 + \psi_2$

$0.9 = l_{11} l_{21} \quad 0.4 = l_{21} l_{31}$

$0.7 = l_{11} l_{31} \quad 1 = l_{31}^2 + \psi_3$

$\Downarrow$

$\frac{l_{21}}{l_{11}} = \frac{4}{7}, \quad 0.9 = \frac{4}{7} l_{11}^2 \Rightarrow l_{11} = 1.575 \Rightarrow \psi_1 = 1 - (1.575)^2 = -1.575$

$= V(\varepsilon_1)$

not much sense

also $Cov(X, F) = L$

$\ell_{11} = Cov(X_1, F)$

$= Corr(X_1, F_1)\sqrt{var(X_1)} \cdot \sqrt{var(F)}$

$= corr(X_1, F_1) \Rightarrow \ell_{11} < 1$. again not make sense

- Solution may not be unique

Let $TT' = \underline{I}$ and $T'T = I$

$\Sigma = LL' + \Psi$

$= LTT'L' + \Psi$

$= (LT)(LT)' + \Psi$

$= L^* L^{*\prime} + \Psi$

$X - \mu = LF + \varepsilon$

$= LTT'F + \varepsilon$

$= L^* F^* + \varepsilon \qquad F^* = T'F$ with also $E(F^*) = 0$

and $V(F^*) = T'\underline{I}T = T'T = \underline{I}$

ε and F^* are also independent

9.3 Methods of Estimation

(a) The problem

1. Given $x_1, \cdots, x_n \overset{iid}{\sim} x_{p\times 1}$, does the model

$$x = \mu + Lf + \varepsilon, \quad L = (l_{ij})_{p\times m}, \quad f = (f_1, \cdots, f_m)', \quad var(\varepsilon) = \begin{pmatrix} \psi_1 & & 0 \\ & \ddots & \\ 0 & & \psi_p \end{pmatrix}$$

with $m<p$ adequately represent the data

2. what are l_{ij}, ψ_i and m?

(b) The principal component method (PCM)

1. population case

$$\Sigma = \sum_{i=1}^{p} \lambda_i e_i e_i'$$

$$= (\sqrt{\lambda_1} e_1, \cdots, \sqrt{\lambda_p} e_p) \begin{pmatrix} \sqrt{\lambda_1} e_1 \\ \vdots \\ \sqrt{\lambda_p} e_p \end{pmatrix}$$

$$= LL' + 0$$

$$\begin{cases} l_{ij} = \sqrt{\lambda_j}\, e_{ij} \\ \psi_i = 0 \\ m = p \end{cases}$$ is a solution

but it is not practically useful. we hope $m << p$

- $\Sigma = \lambda_1 e_1 e_1' + \cdots + \lambda_m e_m e_m'$, if $\lambda_{m+1}^2 + \cdots + \lambda_p^2$ is very small

$$\overset{\text{approx}}{=} [\sqrt{\lambda_1} e_1 \cdots \sqrt{\lambda_m} e_m] \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \vdots \\ \sqrt{\lambda_m} e_m' \end{bmatrix} \quad (m<p)$$

$$= LL'$$

$\Sigma = LL' + \Psi \Rightarrow \begin{cases} \ell_{ij} = \sqrt{\lambda_j}\, e_{ji} & i = 1, \cdots p \\ \psi_i = \sigma_{ii} - \sum_{j=1}^{m} \ell_{ij}^2 & j = 1, \cdots m \end{cases}$ e_{ij} is the ith element of e_j

(2) Sample case

- $S = \sum_{i=1}^{p} \hat{\lambda}_i \hat{e}_i \hat{e}_i'$ $\quad S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'$

  $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$

  Pick $m < p$. then let

  $\hat{L} = [\sqrt{\hat{\lambda}_1}\, \hat{e}_1, \cdots \sqrt{\hat{\lambda}_m}\, \hat{e}_m] = (\ell_{ij})_{p \times m} = (\sqrt{\hat{\lambda}_j}\, \hat{e}_{ji})_{p \times m}$

  $\hat{\Psi} = \begin{pmatrix} \hat{\psi}_1 & & 0 \\ & \ddots & \\ 0 & & \hat{\psi}_p \end{pmatrix}$ with $\hat{\psi}_i = s_{ii} - \sum_{j=1}^{m} (\sqrt{\hat{\lambda}_j}\, \hat{e}_{ji})^2$

  $= s_{ii} - \sum_{j=1}^{m} \hat{\lambda}_j \hat{e}_{ji}^2$

  $= s_{ii} - \sum_{j=1}^{m} \ell_{ij}^2$

Lecture #21

(3) How to determine m?

By prior knowledge

By proportion of total sample variance explained by factors $F_1, \cdots, F_n$

1°) $S = \sum_{i=1}^{p} \hat{\lambda}_i \hat{e}_i \hat{e}_i'$

$= \sum_{i=1}^{m} \hat{\lambda}_i \hat{e}_i \hat{e}_i' + \sum_{i=m+1}^{p} \hat{\lambda}_i e_i e_i'$

$\doteq \hat{L} \hat{L}' + \hat{\Psi}$

$(\sqrt{\hat{\lambda}_1}\hat{e}_1, \cdots, \sqrt{\hat{\lambda}_m}\hat{e}_m)$

2°) Consider residual matrix

$S - (\hat{L}\hat{L}' + \hat{\Psi})$

The smaller the sum of squared entries of $S - (\hat{L}\hat{L}' + \hat{\Psi})$, the better in the approximation

$\hat{\Psi}_i = s_{ii} - \sum_{j=1}^{m} \hat{e}_{ij}^2$

3°) Sum of squared entries of $S - (\hat{L}\hat{L}' + \hat{\Psi})$

$\doteq \sum_{i=m+1}^{p} \hat{\lambda}_i \hat{e}_i \hat{e}_i' = \hat{P}_2 \hat{\Lambda}_2 \hat{P}_2'$

is almost $tr\left(\left(\sum_{i=m+1}^{p} \hat{\lambda}_i \hat{e}_i \hat{e}_i'\right)(\quad)'\right) = tr((\quad)'(\quad))$

$= tr(\hat{P}_2 \hat{\Lambda}_2^2 \hat{P}_2')$

$= tr(\hat{\Lambda}_2^2) = \hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2$

$(\sqrt{\hat{\lambda}_{m+1}}\hat{e}_{m+1}, \cdots \sqrt{\hat{\lambda}_p}\hat{e}_p)(\quad)'(\quad)(\quad)$

4°) Thus if $\hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2$ is small, then the approximation is good. the model is adequate

We normally Remark: The RMS of residual matrix ≤ 0.05 will be ideal case.

5° Since

$s_{ii} = \sum_{j=1}^{m} \ell_{ij}^2 + \hat{\psi}_i$, the contribution of the jth factor

F_j $(1 \leq j \leq m)$ to s_{ii} is ℓ_{ij}^2

to $s_{11} + s_{22} + \cdots + s_{pp}$ = total sample variance is

$$\ell_{1j}^2 + \cdots + \ell_{pj}^2 = (\sqrt{\hat{\lambda}_j}\,\hat{e}_{1j})^2 + \cdots + (\sqrt{\hat{\lambda}_j}\,\hat{e}_{pj})^2$$
$$= \hat{\lambda}_j$$

6° $\dfrac{\hat{\lambda}_j}{\hat{\lambda}_1 + \cdots + \hat{\lambda}_p} = \dfrac{\hat{\lambda}_j}{s_{11} + \cdots + s_{pp}}$ - proportion of total sample variance explained by the jth factor F_j.

$\dfrac{\hat{\lambda}_j}{p}$ if R is used. $\underline{m = \#(\hat{\lambda}_i > 1)}$

↑ correlation matrix ↑ rule of thumb.

SAS code 9-1.txt

Remark: $R = \hat{L}\hat{L}' + \hat{\psi} \Leftrightarrow \hat{\psi} = 0$

(C) Principal factor method

<b Population case

- Assume that $\rho = LL' + \psi$, then

1°) ρ_{ij} is determined by LL' $(i \neq j)$

2°) $\rho_{ij} = 1 + h_i^2 + \psi_i$ $(\because \psi_i = h_i^2)$

- Equivalently $\rho - \psi = LL'$ (remove ψ_i from $\rho_{ii} = 1$

(2) Sample case

- Suppose R is given and Suppose that $\tilde{\psi}_i$ can be estimated initially by ψ_i^*

Then we can replace $r_{ii} = 1$ by $\because \psi_i^* = h_i^{*2}$ in R to obtain a "reduced" sample correlation matrix

$$R_r = \begin{bmatrix} h_1^{*2} & r_{12} & \cdots & r_{1p} \\ r_{21} & h_2^{*2} & & \vdots \\ \vdots & & \ddots & \\ r_{p1} & r_{p2} & \cdots & h_p^{*2} \end{bmatrix} \approx R - \tilde{\psi}$$

- Factor R_r to get

$$R_r \doteq \tilde{L}_r^* \tilde{L}_r^{*\prime}$$

with $\hat{L}_r^* = \{\tilde{e}_{ij}^*\}_{p \times m}$ if $m = p$, $R_r = \tilde{L}_r^* \tilde{L}_r^{*\prime}$ o/w not equal

- Let $(\hat{\lambda}_i^*, \hat{e}_i^*)$ $i = 1, \cdots, m$ be the eigen pair of R_r, then the principal factor method

$$\tilde{L}^* = (\sqrt{\lambda_1^*}\,\hat{e}_1^*, \cdots, \sqrt{\lambda_m^*}\,\hat{e}_m^*)$$

that is $\tilde{\hat{e}}_{ij}^* = \sqrt{\lambda_j^*}\,\hat{e}_{ij}^*$

$$\tilde{\psi}_i^* = 1 - \sum_{j=1}^{m} \tilde{\hat{e}}_{ij}^{*2} = 1 - \tilde{h}_i^{*2}$$

- Note that we start with initial estimates h_i^{*2} (or $\psi_i^* = 1 - h_i^{*2}$)

  Then we get $\tilde{h}_i^{*2}$ (or $\tilde{\psi}_i^*$)

  Now we can start with $\tilde{h}_i^{*2}$ (or $\tilde{\psi}_i^*$)

  repeat the above step to obtain PF solutions iteratively

(3) prior values of h_i^{*2} (or ψ_i^*) (estimates)

- Squared multiple correlation coefficient (SMC)

  (prior = smc, in SAS)

  1° suppose $R^{-1} = (r^{ij})_{p\times p}$. then $r^{ii} = VIF(X_i) = (1-R^2)^{-1}$

  $R^2_{i\cdot \text{rest}} = 1 - \frac{1}{r^{ii}}$ is called the sample multiple correlation coefficient. (square of) between X_i and $(p-1)$ remaining variables

  2° Let $\psi_i^* = \frac{1}{r^{ii}}$ or $h_i^{*2} = 1 - \frac{1}{r^{ii}} = SMC_i$

- Input values for h_i^{*2} (or ψ_i^*) (priors = Input)

  $h_i^{*2} = 1 \longrightarrow$ PCM method

- selected random numbers for h_i^{*2} (or ψ_i^*)

  (priors = random)

- priors = max cm)

$$h_i^* = \max_{|i \neq j|} |r_{ij}|$$

(d) The maximum likelihood method.

<i> the method

Let $x_1, \ldots, x_n$ iid $x \sim N_p(\mu, \Sigma)$ with

$$\Sigma = LL' + \psi$$

Then the MLE of L, ψ and μ could be obtained by maximizing $L(\mu, \Sigma) = L(\mu, LL' + \psi)$

$$= (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} e^{-\frac{1}{2} \operatorname{tr}\left[\Sigma^{-1}\left(\sum_j (x_j - \bar{x})(x_j - \bar{x})' + n(\bar{x} - \mu)(\bar{x} - \mu)'\right)\right]}$$

subject to $L'\psi^{-1}L$ being diagonal (uniqueness condition)

- (MLE method is scale invariant, we can also work with ρ and

ρ) $\Sigma = LL' + \psi \quad V = \operatorname{diag}(\Sigma)$

$$\rho = V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}} = V^{-\frac{1}{2}}(LL' + \psi)V^{-\frac{1}{2}} = V^{-\frac{1}{2}} L (V^{\frac{1}{2}} L)' + V^{\frac{1}{2}} \psi V^{\frac{1}{2}}$$

$L'\psi^{-1}L$ beay diagonal

$(V^{-\frac{1}{2}}L)' \overset{''}{} V^{\frac{1}{2}} \psi^{-1} V^{\frac{1}{2}} (V^{-\frac{1}{2}} L)$

Hence,

Based on S, We have $\hat{\Sigma}$ and $\hat{\psi}$ as MLE

Based on R (the standardized data)

we have $\hat{V}^{-\frac{1}{2}}\hat{\Sigma}\hat{V}^{-\frac{1}{2}}$ as MLE.

$\hat{\Sigma}_z$ $\quad$ $\hat{\psi}_z$

$\hat{V}^{-\frac{1}{2}} = \mathrm{diag}(s_{11}^{-\frac{1}{2}}, \ldots, s_{pp}^{-\frac{1}{2}})$

See proofs in supplement 9A

- The MLE of h_i^2 is $\hat{h}_i^2 = \sum_{j=1}^{m} \hat{\ell}_{ij}^2 \quad 1 \le i \le p$

- The proportion of total sample variance explained by the jth factor

$$\begin{cases} \dfrac{\hat{\ell}_{1j}^2 + \cdots + \hat{\ell}_{pj}^2 \; (= \hat{\lambda}_j)}{s_{11} + \cdots + s_{pp}} & \text{for } S \\ \dfrac{\hat{\ell}_{1j}^2 + \cdots + \hat{\ell}_{pj}^2 \; (= \hat{\lambda}_j)}{p} & \text{for } R \end{cases}$$

SAS 9-2.SAS

(2) Determination of the number of common factors

- A large sample test (likelihood ratio test)

1° Bartellet's test

$H_0: \underset{p\times p}{\Sigma} = \underset{p\times m}{L}\underset{m\times p}{L'} + \underset{p\times p}{\Psi}$ $\quad H_1: \Sigma =$ anyother positive definite matrix

By LR method,

$$L(\mu,\Sigma) = \left(\frac{1}{2\pi}\right)^{\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)'\Sigma^{-1}(x_j-\mu)\right\}$$

$$= (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})' + n(\bar{x}-\mu)(\bar{x}-\mu)'\right)\right]\right\}$$

under H_0: $\hat{\mu}_{MLE} = \bar{X}$ and $\hat{\Sigma}_{MLE} = \hat{L}\hat{L}' + \hat{\Psi} = \hat{\Sigma}$

When under $H_0 \cup H_1$, $\hat{\mu}_{MLE} = \bar{x}$, $\hat{\Sigma}_{MLE} = S_n = \frac{(n-1)S}{n}$

$$-2\ln\Lambda = -2\ln\left[\left(\frac{|\hat{L}\hat{L}'+\hat{\Psi}|}{|S_n|}\right)^{-\frac{n}{2}} \cdot \frac{\exp\left\{-\frac{n}{2}tr\left((\hat{L}\hat{L}'+\hat{\Psi})^{-1}\cdot S_n\right)\right\}}{\exp\left\{-\frac{1}{2}np\right\}}\right]$$

$$= n\ln\left(\frac{|\hat{\Sigma}|}{|S_n|}\right) + n\left(tr(\hat{\Sigma}^{-1}S_n) - p\right)$$

$$= n\ln\left(\frac{|\hat{\Sigma}|}{|S_n|}\right)$$

$\Leftarrow$ supplement 9A indicates $tr(\hat{\Sigma}^{-1}S_n) - p = 0$ if $\hat{\Sigma} = \hat{L}\hat{L}' + \hat{\Psi}$ is the MLE of $\Sigma = LL' + \Psi$

Using Bartletts's correction,

$$\left(n-1-(2p+4m+5)/6\right)\ln\frac{|\hat{L}\hat{L}'+\hat{\Psi}|}{|S_n|} \Rightarrow \chi^2_{[(p-m)^2-p-m]/2}$$

when both n and $n-p$ are large

Hence, we reject H_0 if

$$\left(n-1-(2p+4m+5)/6\right)\ln\frac{|\hat{L}\hat{L}'+\hat{\Psi}|}{|S_n|} > \chi^2_{[(p-m)^2-p-m]/2}(\alpha)$$

$\left(m < \frac{1}{2}(2p+1-\sqrt{8p+1})\right)$

suggesting a higher-dimensional or nonlinear mechanism generating the correlation SAS (method = ML)

2°) Significant test results are common in testing the fit of a small number (m) (relative top) of factors to correlation obtained from even moderately large sample

- ~~Akaiks~~ Akaiks Information Criterion (AIC)

1°) Alternatively, we may want to use Akaiks AIC instead of a sufficient test for determining the number of factors.

Under a m factors model : $\Sigma = LL' + \Psi$

$$\ln(L(m)) = C - \frac{n}{2}\left[\ln|LL'+\Psi| + tr\left((LL'+\Psi)^{-1}S_n\right)\right]$$

With $S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})'$

Then the AIC statistic is

$$AIC(m) = -2\ln(L(m)) + [2p(m+1) - m(m-1)]$$

2°) the m-factor model with m corresponding to the smallest value of AIC(m) is considered best

SAS 9.3.txt

```
options ls=84 ps=55;
title 'SAS Factor Analysis example1';

data spearman (type=corr);
 _type_='corr';
 if _n_=1 then _type_='N';
 infile cards missover;
 input _name_$ c f e m d mu;
 datalines;
 N  33
 c 1.0
 f .83 1.0
 e .78 .67 1.0
 m .70 .67 .64 1.0
 d .66 .65 .54 .45 1.0
 mu .63 .57 .51 .51 .40 1.0
 ;

proc print data=spearman;
 run;

ods html;

proc factor data=spearman method=prin res scree;
   var c f e m d mu;
   title2 'Princial Component Method';
run;

ods html close;

proc factor data=spearman method=prin res nfact=2;
   title2 'PCM: Res matrix for nfact=2';
run;

proc factor data=spearman method=prin res nfact=3;
   title2 'PCM: Res matrix for nfact=3';
run;

proc factor data=spearman method=prin res nfact=4;
   title2 'PCM: Res matrix for nfact=4';
run;
```

```
options ls=84 ps=55;
title 'SAS Factor Analysis Example 2';

data spearman (type=factor);
 if _n_=1 then _type_='PRIORS';
 else if _n_=2 then _type_='N';
 else _type_='corr';
 infile cards missover;
 input _name_$ c f e m d mu;
 datalines;
 val .2 .3 .1 .1 .3 .2
 N 33
 c 1.0
 f .83 1.0
 e .78 .67 1.0
 m .70 .67 .64 1.0
 d .66 .65 .54 .45 1.0
 mu .63 .57 .51 .51 .40 1.0
 ;

proc print data=spearman;
run;

proc factor data=spearman method=prin priors=input;
   var c f e m d mu;
   title2 'Princial Factor Method with input priors';
run;

proc factor data=spearman method=prin res nfact=2;
   title2 'PCM: Res matrix for nfact=2';
run;

proc factor data=spearman method=prin res;* scree;
   var c f e m d mu;
   title2 'Princial Component Method';
run;

proc factor data=spearman method=prin res priors=smc;
   title2 'PFM: priors=smc';
run;

proc factor data=spearman method=prin res priors=smc nfact=3;
   title2 'PFM: priors=smc nfact=3';
run;

proc factor data=spearman method=prin res priors=max;
   title2 'PFM: priors=max';
run;

proc factor data=spearman method=prin res priors=max nfact=3;
   title2 'PFM: priors=max nfact=3';
run;
```

```
options ls=80 ps=47 nodate nonumber;
title1 h=2 'SAS Factor Analysis Example 3';
title2 h=1 'Stock-Price Data (Weekly Rate of Return) JW: T8-4';

data stock;
*infile 'Z:\\My Documents\Teaching\Stat524\Fall 2010\Data set\T8-4.dat' firstobs=1;
input JPMorgan Citibank WellsFargo Shell ExxonMobil;
datalines;
0.0130338       -0.0078431      -0.0031889      -0.0447693      0.0052151
0.0084862       0.0166886       -0.0062100      0.0119560       0.0134890
-0.0179153      -0.0086393      0.0100360       0.0000000       -0.0061428
0.0215589       -0.0034858      0.0174353       -0.0285917      -0.0069534
0.0108225       0.0037167       -0.0101345      0.0291900       0.0409751
0.0101713       -0.0121978      -0.0083768      0.0137083       0.0029895
0.0111288       0.0280044       0.0080721       0.0305433       0.0032290
    .               .               .               .               .
    .               .               .               .               .
    .               .               .               .               .
0.0033740       -0.0153061      -0.0238245      -0.0016738      -0.0172270
0.0033626       0.0029016       -0.0030507      -0.0012193      -0.0097005
0.0170147       0.0095061       0.0181994       -0.0161758      -0.0075614
0.0103929       -0.0026612      0.0044290       -0.0024818      -0.0164502
-0.0127948      -0.0143678      -0.0187402      -0.0049759      -0.0163732
;
run;

proc print data=stock;
run;


proc factor method=prin res scree;
  title2 'Principal Component Method';
run;

proc factor method=prin n=2 priors=smc res;
  title2 'Principal Factor Method nfactor=2';
run;


/*Iterative method for estimation*/

proc factor method=prinIT priors=smc res;
  title2 'ITERATIVE Principal Factor Method';
run;

proc factor method=prinIT priors=smc n=2 res;
  title2 'ITERATIVE Principal Factor Method nfactor=2';
run;

proc factor method=ml res;
run;

proc factor data=stock method=ml n=1 res;
  title2 'MLE: nfactor=1';
run;

proc factor method=ml heywood n=2 res;
  title2 'MLE: nfactor=2';
run;

proc factor method=ml heywood nfact=3 res;
  title2 'MLE: nfactor=3';
run;
```

Lecture #22

9.4 Factor rotation

(a) Rotation

<b> nonuniqueness (might be a blessing)

- $S = \hat{L}\hat{L}' + \hat{\Psi}$

  $= \hat{L}TT'\hat{L}' + \hat{\Psi}$ $(TT' = I_m)$

  $= \hat{L}^*\hat{L}^{*\prime} + \hat{\Psi}$

- $\hat{L}$ and $\hat{L}^* = \hat{L}T$ represent the same covariance structure

  But it may be difficult to interprete using $\hat{L}$ and it may become easy to interprete factors using $\hat{L}T = L^*$. a rotation of $\hat{L}$

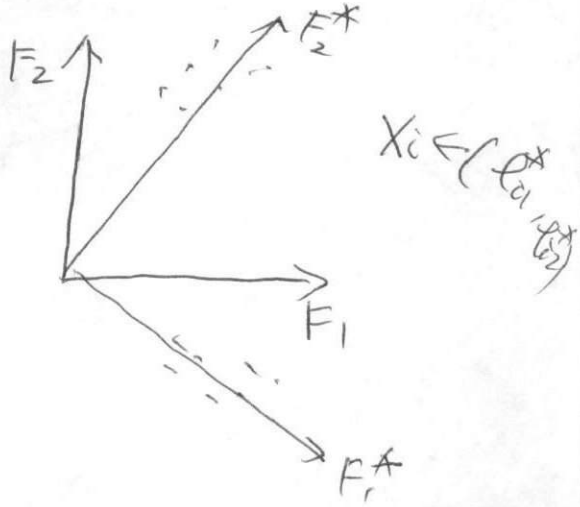
- Ideally: After rotation, each variable bads highly on a single factor and has small or moderate loading on other factors

(2) groups variables

$$X - \mu = LF + \varepsilon$$
$$= LTT'F + \varepsilon$$
$$= L^*F^* + \varepsilon$$

$X_i \Leftarrow (\ell_{i1}, \ell_{i2})$

If $m = 2$



(b) Select rotation (or T)

<1> varimax rotation (orientation)

- Let $\hat{L}^* = \hat{L}T$ and let

$$\tilde{\ell}_{ij}^{*2} = \frac{\hat{\ell}_{ij}^{*2}}{\hat{h}_i^{*2}} = \frac{\hat{\ell}_{ij}^{*2}}{\hat{h}_i^2}$$ ← (giving more weight to variables with small communality)

standardization ✓

- Let $V = \sum_{j=1}^{m} \frac{1}{p}\left[\sum_{i=1}^{p}(\tilde{\ell}_{ij}^{*2})^2 - \frac{\left(\sum_{i=1}^{p}\tilde{\ell}_{ij}^{*2}\right)^2}{p}\right]$

$$= \frac{1}{p}\sum_{i=1}^{p}\left[\tilde{\ell}_{ij}^{*2} - \left(\frac{1}{p}\sum \tilde{\ell}_{ij}^{*2}\right)\right]^2$$

↑ The variance of squares of scaled loading for jth factor

Sample

- choose T such that V is maximized that is, such that squares of scaled loadings spread out as much as possible.
- Equivalently, choose T such that among $\tilde{\ell}_{1j}^{*}, \cdots \tilde{\ell}_{pj}^{*}$ Some are large, others are negligiblely small.

(SAS: rotate = varimax)

(2) Quartimax rotation

- Let $\hat{L}^* = \hat{L}T$, choose T such that

$$\underbrace{\frac{1}{pm}\sum_{ij}\hat{\ell}_{ij}^{*4} - \left(\frac{1}{pm}\sum_{ij}\hat{\ell}_{ij}^{*2}\right)^2}_{\text{Maximized}} \text{ is maximized}$$

$$= \frac{1}{pm}\sum_{i}\hat{h}_i^2$$

(rotate = Quartimax)

(3) Non-orthogonal rotation (oblique rotation)

SAS (rotate=HK, rotate=promax)

9.5 Factor scores

(a) Estimation of factor scores

<1> The model

- $X - \mu = LF + \varepsilon$ $\quad F = (F_1, \cdots, F_m)'$

$X - \mu$: observable; F, ε: unobservable

$$S = \hat{L}\hat{L}' + \hat{\Psi}$$

(2) Factor scores

- Let f_j be the possible value F can take at jth observation $(j = 1, \cdots, n)$

  The we estimate f_j by $\hat{f}_j$, which are called factor scores

- Factor scores can be useful in diagnostic processes and used as reduced data for any subsequent analysis

- To estimate f_j's, we will treat $\hat{l}_{ij}(\hat{e}_{ij}^{*})$ and $\hat{\psi}_i$ as if they were the true values on the model.

(b) The weighted least square method

- Let $X-\mu = LF+\varepsilon$ $\quad COV(\varepsilon) = \begin{pmatrix} \psi_1 & & 0 \\ & \ddots & \\ 0 & & \psi_p \end{pmatrix} = \psi$

Assume that L, ψ and μ are known

$$\psi^{-\frac{1}{2}}(X-\mu-LF) = \psi^{-\frac{1}{2}}\varepsilon \qquad \text{with } V(\psi^{-\frac{1}{2}}\varepsilon) = I_p$$

- choose f as the estimator of F such that

$$(\psi^{-\frac{1}{2}}\varepsilon)'(\psi^{-\frac{1}{2}}\varepsilon) = \sum_{i=1}^{p}\frac{\varepsilon_i^2}{\psi_i} = (X-\mu-LF)'\psi^{-1}(X-\mu-LF)$$

is minimized

(2) The solution (factor scores)

- population case

$$\underbrace{\psi^{-\frac{1}{2}}(X-\mu)}_{Y} = \underbrace{\psi^{-\frac{1}{2}}L}_{X}F + \psi^{-\frac{1}{2}}\varepsilon$$

$$\hat{f} = (X'X)^{-1}X'Y$$

$$= (L'\psi^{-1}L)^{-1}L'\psi^{-1}(X-\mu)$$

- Sample case

1° Let $\hat{L}$ and $\hat{\psi}$ and $\hat{\mu} = \bar{x}$ be the estimator of L, ψ and μ.

$$\hat{\Sigma} = \hat{L}\hat{L}' + \hat{\psi} \doteq S$$

$$\hat{f}_j = (\hat{L}'\hat{\psi}^{-1}\hat{L})^{-1}\hat{L}'\hat{\psi}^{-1}(x_j-\bar{x})$$

2° If we start with R then

$$\hat{f}_j = (\hat{L}_z' \hat{\Psi}_z^{-1} \hat{L}_z)^{-1} \hat{L}_z' \hat{\Psi}_z^{-1} Z_i$$

Where $Z_i = (\text{diag}(S))^{-\frac{1}{2}}(X_i - \bar{X}) = D^{-\frac{1}{2}}(X_i - \bar{X})$

$$\hat{L}_z = D^{-\frac{1}{2}} \hat{L} \qquad \hat{\Psi}_z = D^{-\frac{1}{2}} \hat{\Psi} D^{-\frac{1}{2}}$$

• comment

1° If $\hat{L}$ and $\hat{\Psi}$ are obtained via MLE, then

$\hat{L}' \Psi^{-1} \hat{L}$ is a diagonal matrix (unique condition in MLE)

$=$

$\hat{L}_z' \hat{\Psi}_z^{-1} \hat{L}_z'$

2° If $\hat{L}$ and $\hat{\Psi}$ is obtained via PCM

and $\Psi_i = C$, $i = 1, \cdots p$ then

$\hat{f}_j$ are obtained essentially by unweighted (ordinary) least square method and

$$\hat{f}_j = (\hat{L}'\hat{L})^{-1} \hat{L}'(X_j - \bar{X})$$

$$= \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{\lambda_m}} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1}\, \hat{e}_1' \\ \vdots \\ \sqrt{\lambda_m}\, \hat{e}_m' \end{pmatrix} (X_j - \bar{X})$$

$$= \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} \hat{e}_1'(X_j - \bar{X}) \\ \vdots \\ \frac{1}{\sqrt{\lambda_m}} \hat{e}_m'(X_j - \bar{X}) \end{pmatrix} \leftarrow \text{Scaled principal component}$$

$$\hat{L} = (\sqrt{\lambda_1}\, \hat{e}_1 \; \cdots \; \sqrt{\lambda_m}\, \hat{e}_m)$$

$$\hat{L}\hat{L}' = \begin{pmatrix} \hat{\lambda}_1 & & \\ & \ddots & \\ & & \hat{\lambda}_m \end{pmatrix}$$

(C) The regression method

a) population:

Assume $(F, \Sigma)' \sim N_{m+p}\left(0, \begin{pmatrix} I_m & 0 \\ 0 & \Psi \end{pmatrix}\right)$

$X-\mu = LF + \Sigma \sim N_p(0,\ LL'+\Psi)$

$(X-\mu,\ F)' \sim N_{m+p}\left(0, \begin{pmatrix} \Sigma = LL'+\Psi & \vdots & L \\ \cdots & \cdots & \cdots \\ L' & \vdots & I_m \end{pmatrix}\right)$

$X_1 \quad X_2$

By The result 4.6 $(X_1, X_2)' \sim N\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\right) \Rightarrow E(X_2|X_1) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1-\mu_1)$

$$E(F \mid X-\mu) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1-\mu_1) \qquad \mathrm{Cov}(X_2|X_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

$X_2 \quad X_1$

$$= L'(LL'+\Psi)^{-1}(X-\mu)$$

$$\mathrm{Cov}(F \mid X-\mu) = I - L'(LL'+\Psi)^{-1}L$$

b) Sample solution $X_1, \cdots X_n \sim X \sim N_p(\mu, \Sigma)$ $\Sigma = LL'+\Psi$

$$\hat{f}_j^R = \hat{L}'(\hat{L}\hat{L}'+\hat{\Psi})^{-1}(X_j-\bar{X}) \quad (j=1,\cdots n) \qquad \text{or } \hat{f}_j^R = \hat{L}' S^{-1}(X_j-\bar{X})$$

If a correlation matrix is factored,

$$\hat{f}_j^R = \hat{L}_z'(\hat{L}_z\hat{L}_z'+\hat{\Psi}_z)^{-1}(Z_j) \qquad \text{or } \hat{f}_j^R = \hat{L}_z' R^{-1} Z_j$$

$$Z_j = D^{-\frac{1}{2}}(X_j-\bar{X}) \quad D = \mathrm{Diag}(S)$$

- Relationship between $\hat{f}_j^R$ and $\hat{f}_j^{LS}$

By $\hat{L}'(\hat{L}\hat{L}'+\hat{\Psi})^{-1} = (I+\hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1}\hat{L}'\hat{\Psi}^{-1}$ ← See Exercise 9.6

$$\hat{f}_j^{LS} = (\hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1}\hat{L}'\hat{\Psi}^{-1}(X_j-\bar{X})$$

$$= (\hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1}(I+\hat{L}'\hat{\Psi}^{-1}\hat{L})\hat{f}_j^R$$

$$= (I + (\hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1})\hat{f}_j^R$$

If $\hat{L}, \hat{\Psi}$ are MLEs, $(\hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1} = \hat{\Delta}^{-1}$ where $\hat{\Delta}$ is a diagonal matrix. when $\hat{\Delta}$ is close to zero, $\hat{f}_j^{LS}$ is close to $\hat{f}_j^{R}$

- If rotated loadings $\hat{L}^* = \hat{L}T$ are used.

$$\hat{f}_j^* = T'\hat{f}_j, \quad j = 1, 2, \cdots n$$

(SAS proc factor data = out = ...)

9.6 A strategy for factor analysis

(a) Most crucial decision in FA:
The number m of common factors

1. A large sample test (with method = ML)
   - Suitable for approximately normal data points
   - Rejection usually occurs for small n, large p.

   (AIC) Criterion
2. the proportion of the total sample variance explained by m factors
3. the RMS of residual matrix (≤ 0.05. ideal case)
4. Subject-matter knowledge_ prior knowledge
5. Interpretability

(b) Less crucial decision:

Estimation methods and rotations

1. Method = prin
   = prin priors =
   = ml
   = image
   = harris
   = ulс
   = prinit
   = alpha

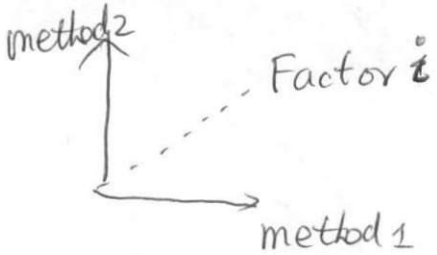
   rotate = varimax
   = quartimax
   = equamax
   = parsimax
   = orthomax
   = HK
   = promax
   = [illegible]

2. Successful FA usually leads to consistent results from different method.

(C) steps

1. PC FA
   - plot factor scores & check suspecious observations
     (calculate the standardized scores and squared distance for each observation & assessing the normality)
   - try a varimax rotation
2. MLFA and varimax rotation
3. Perhaps other methods and rotations

4. Compare results obtained
   - Compare factors (pattern of loadings)
   - plot factor scores from different methods



looking for consistency

5. repeat above steps for different m to see if get better understanding and interpretability of the data.

6. For large data set (n large), split it into two and perform FA on each one to check the consistency and stability of the result.

d) Example (Example 9.14)

```
options ls=80 ps=47 nodate nonumber;
title1 h=2 'SAS Factor Analysis Example 3';
title2 h=1 'Stock-Price Data (Weekly Rate of Return) JW: T8-4';

data stock;
*infile 'Z:\\My Documents\Teaching\Stat524\Fall 2010\Data set\T8-4.dat' firstobs=1;
input JPMorgan Citibank WellsFargo Shell ExxonMobil;
obs='*';
datalines;
0.0130338       -0.0078431      -0.0031889      -0.0447693      0.0052151
0.0084862       0.0166886       -0.0062100      0.0119560       0.0134890
-0.0179153      -0.0086393      0.0100360       0.0000000       -0.0061428
0.0215589       -0.0034858      0.0174353       -0.0285917      -0.0069534
0.0108225       0.0037167       -0.0101345      0.0291900       0.0409751
0.0101713       -0.0121978      -0.0083768      0.0137083       0.0029895
0.0111288       0.0280044       0.0080721       0.0305433       0.0032290
    .               .               .               .               .
    .               .               .               .               .
    .               .               .               .               .
0.0033740       -0.0153061      -0.0238245      -0.0016738      -0.0172270
0.0033626       0.0029016       -0.0030507      -0.0012193      -0.0097005
0.0170147       0.0095061       0.0181994       -0.0161758      -0.0075614
0.0103929       -0.0026612      0.0044290       -0.0024818      -0.0164502
-0.0127948      -0.0143678      -0.0187402      -0.0049759      -0.0163732
;
run;

proc print data=stock;
run;


/*Factor Ratation*/

proc factor data=stock method=ml heywood n=2 res rotate=varimax preplot plot;
  title2 'MLE: rotate=varimax';
run;

proc factor data=stock method=ml heywood n=2 res rotate=HK preplot plot;
  title2 'MLE: rotate=HK';
run;

/* Factor Scores*/

proc factor data=stock method=ml heywood n=2 res rotate=promax preplot plot
out=fscore;
  title2 'MLE: rotate=promax';
run;

proc print data=fscore;
run;

title2 "scatter plot of factor scores for factor1 and factor2";
%plotit(data=fscore,labelvar=obs, plotvars=factor2 factor1, color=black,
colors=blue);
run;

proc factor data=stock method=prin priors=asmc nfact=3 rotate=varimax out=fscore1;
run;

proc print data=fscore1;
run;

title2 "scatter plot of factor scores for factor1 and factor3";
```

code9-4

```
%plotit(data=fscore1,labelvar=obs, plotvars=factor3 factor1 ,color=black,
colors=blue);
run;

data fscorenew;
set fscore;
factor1mle=factor1;
run;

data fscorecompare;
merge fscorenew fscore1;
run;

%plotit(data=fscorecompare,labelvar=obs, plotvars=factor1mle factor1 ,color=black,
colors=blue);
run;
```

HW#7 9.19, 9.32

· due Tuesday of Nov. 23. 2010

Lecture 23

chapter 11 Discrimination and classification

11.1 objectives of discrimination and classification

(a) Discrimination (seperation)

1. To set up a method of deciding whether the observations fall into groups and if so to delineate the groups

2. Examples

- IUPUI admission office
- banking and commercial finance loan application

(b) classification (Allocation)

1. To construct a method of assigning a new observations to the correct groups (population)

(c) connection of discrimination and classification

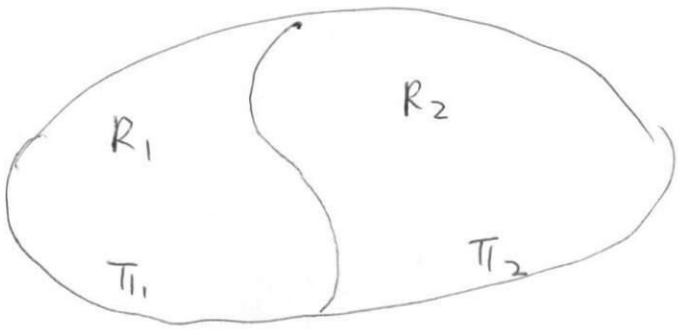
The two are interrelated, so closely, indeed that they are often confused

11.2 classification for the population

a) steps

1. rules are developed from "training" and "learning" sample
2. divide all possible sample outcomes into two groups

region → R_1 and R_2, which correspond to populations π_1 and π_2, respectively.



(3) There may not exist clear cut-off boundary (no-error free allocation methods)

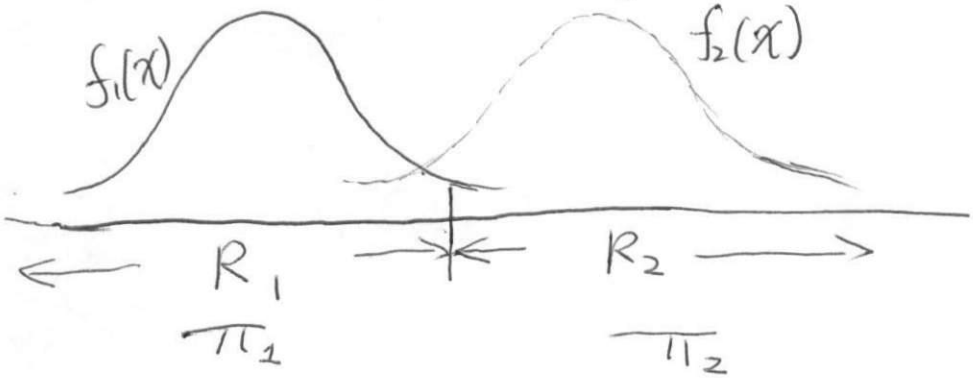
(b) goals

<1> to keep misclassification rate as small as possible by taking the following into account.

(i) Occurence frequency from individual populations

(ii) misclassification cost (penalty)

(c) misclassification probability and cost



(1) Conditional Probability

• Conditional misclassification Prob

1° $P(\text{allocate } X \text{ to } \pi_1 \mid X \text{ is from } \pi_2)$

$= P(X \in R_1 \mid X \in \pi_2) = P(X \in R_1 \mid \pi_2) = P(1/2)$

$= \int_{R_1} f_2(x)\,dx$

2°) P(allocate X to π_2 | X is from π_1)

$= P(2|1) = \int_{R_2} f_1(x)\, dx$

• misclassification probability

1°) P(X is misclassified as π_1)

$= P$(X is from π_2, but allocated to π_1)

$= P$(X is from π_2) $\cdot P(1|2)$

$= p_2 \cdot P(1|2)$

2°) P(X is misclassified as π_2)

$= p_1 P(2|1)$ $(p_1 = 1 - p_2)$

3° P(X is correctly allocated to π_1)

$= p_1 P(1|1)$

P(X is correctly allocated to π_2)

$= p_2 P(2|2)$

(2) Cost of misclassification

	CM	classified as π_1	classified as π_2
true	π_1	0	$C(2\|1)$
	π_2	$C(1\|2)$	0

(d) classification rules

<1> Minimize expected (average) cost of misclassification (MECM)

- Expected (average) cost of misclassification

$$ECM = C(2|1)\,p_1 P(2|1) + C(1|2)\,p_2 P(1|2)$$

$$= C(2|1)\,p_1 \int_{R_2} f_1(x)\,dx + C(1|2)\,p_2 \int_{R_1} f_2(x)\,dx$$

$$= C(2|1)\,p_1 \int_{R-R_1} f_1(x)\,dx + C(1|2)\,p_2 \int_{R_1} f_2(x)\,dx$$

$$= C(2|1)\,p_1 + \int_{R_1} \left(C(1|2)\,p_2 f_2(x) - C(2|1)\,p_1 f_1(x)\right) dx$$

- Regions R_1 and R_2 that minimizes ECM

1°) General case

R_1: $C(1|2)\,p_2 f_2(x) - C(2|1)\,p_1 f_1(x) \leq 0$

R_2: $C(1|2)\,p_2 f_2(x) - C(2|1)\,p_1 f_1(x) > 0$

$$\Leftrightarrow R_1 = \left\{ x : \frac{f_1(x)}{f_2(x)} \geq \frac{C(1|2)}{C(2|1)} \cdot \frac{p_2}{p_1} \right\}$$

$$R_2 = \left\{ x : \frac{f_1(x)}{f_2(x)} < \frac{C(1|2)}{C(2|1)} \cdot \frac{p_2}{p_1} \right\}$$

2°) Special cases

If $\frac{p_2}{p_1}$ unknown, then assume $\frac{p_2}{p_1} = 1$

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{C(1|2)}{C(2|1)}$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < \frac{C(1|2)}{C(2|1)}$$

If $\frac{C(1|2)}{C(2|1)}$ is unknown, then assume $\frac{C(1|2)}{C(2|1)} = 1$

$$\Rightarrow R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1}$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1}$$

If both $\frac{p_2}{p_1}$ and $\frac{c(1|2)}{c(2|1)}$ unknown. assume that

$$\frac{p_2}{p_1} \cdot \frac{c(1|2)}{c(2|1)} = 1 \Rightarrow \begin{cases} R_1: & f_1(x) \geq f_2(x) \\ R_2: & f_1(x) < f_2(x) \end{cases}$$

Question: what if $f_i(x)$ are unknown. $i = 1, 2$

(2) other classification rules:

- Minimize total prob of misclassification (MTPM)

$$\begin{aligned} TPM &= p_1 P(2|1) + p_2 P(1|2) \\ &= p_1 \int_{R_2} f_1(x)\,dx + p_2 \int_{R_1} f_2(x)\,dx \\ &= p_1 + \int_{R_1} p_2 f_2(x) - p_1 f_1(x)\,dx \end{aligned}$$

$$\Rightarrow \begin{cases} R_1: & \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1} \\ R_2: & \frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1} \end{cases}$$

Code:

11.3 classification with two multivariate normal distribution

(a) Equal covariance metrices $\Sigma_1 = \Sigma_2 = \Sigma$

(1). classification regions R_1 and R_2 based on MECM

- Desities

1° $f_i(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{(x-\mu_i)'\Sigma^{-1}(x-\mu_i)}{2}}$ $\quad i=1,2$

$$\frac{f_1(x)}{f_2(x)} = e^{\underbrace{-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)}_{\text{II}\ Q}}$$

- Regions (general case)

1°) R_1: $\frac{f_1(x)}{f_2(x)} \geq \frac{C(1|2)P_2}{C(2|1)P_1} \nwarrow r$

$\Leftrightarrow Q \geq \ln r$

$\Leftrightarrow (\mu_1-\mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) \geq \ln r$

2°) R_2: ---------- $< \ln r$

Linear (in x) classification rule

(2) classification in practice (we assume ① $\Sigma_1 = \Sigma_2 = \Sigma$, but Σ (μ_1, μ_2) unknown
② $X_{11}, \cdots X_{1n_1} \sim N_p(\mu_1, \Sigma)$
$X_{21}, \cdots X_{2n_2} \sim N_p(\mu_2, \Sigma)$

- Estimation of μ_i and Σ

1°) $\hat{\mu}_i = \bar{X}_i \quad i=1,2$

$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$, $S_i = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)(X_{ij}-\bar{X}_i)'$, $i=1,2$

2°) $\hat{\Sigma} = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2} = S_{pooled}$

- The rule

1° Allocate x_0 to

π_1 if $(\bar{x}_1-\bar{x}_2)'\hat{\Sigma}^{-1}x_0 - \frac{1}{2}(\bar{x}_1-\bar{x}_2)'\hat{\Sigma}^{-1}(\bar{x}_1+\bar{x}_2) \geq \ln r$

π_2 if $\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad < \ln r$

2° If $r = 1 \left(\frac{c(1|2)\,p_2}{c(2|1)\,p_1} = 1\right)$ then

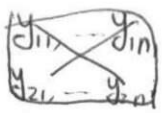
R_1: $(\hat{\Sigma}^{-1}(\bar{x}_1-\bar{x}_2))'x_0 \geq \frac{1}{2}(\hat{\Sigma}^{-1}(\bar{x}_1-\bar{x}_2))'(\bar{x}_1+\bar{x}_2)$

$\Leftrightarrow \quad \hat{a}'x_0 \geq \hat{a}'\cdot\frac{\bar{x}_1+\bar{x}_2}{2}$

R_2: $\quad \hat{a}'x_0 < \hat{a}'\cdot\frac{\bar{x}_1+\bar{x}_2}{2}$

$\hat{a} = \hat{\Sigma}^{-1}(\bar{x}_1-\bar{x}_2)$
$= S_{pooled}^{-1}(\bar{x}_1-\bar{x}_2)$ ← discriminant coefficients

If let $y = \hat{a}'x$, then $y_0 = \hat{a}'x_0 \quad \bar{y}_1 = \hat{a}'\bar{x}_1 \quad \bar{y}_2 = \hat{a}'\bar{x}_2$



$y_{ij} = \hat{a}'x_{ij}$

Then

R_1: $\quad y_0 \geq \frac{1}{2}(\bar{y}_1+\bar{y}_2)$

R_2: $\quad y_0 < \frac{1}{2}(\bar{y}_1+\bar{y}_2)$

(3) Scaling

- Uniqueness problem

Any $c\hat{a} = c\cdot\hat{\Sigma}^{-1}(\bar{x}_1-\bar{x}_2)$, $c>0$ can serve as $\hat{a}$ in above rule

- Scale $\hat{a}$. $\hat{a}^* = \frac{\hat{a}}{\|\hat{a}\|} = \frac{\hat{a}}{\sqrt{\hat{a}'\hat{a}}}$ recommended (only if x variable have been standardized)

If this is not the case, great deal of caution must be exercised in interpreting the result

(assess the relative importance (compare to $\sqrt{x_1}$) of variables $x_2, \ldots, x_p$ as discriminators)

```
options ls=85 ps=65;
title1 'SAS DISCRIM example 1';
title2 h=1 'Salmon data JW: T11-2';
data salmon;
 infile 'Z:\\My Documents\Teaching\Stat524\Fall 2010\Data set\T11-2.dat' firstobs=1;
 input species gender fresh marine;
run;

proc print data=salmon;
run;

proc plot data=salmon;
   plot fresh*marine=species;
run;

title2 "scatter plot of fresh and marine";
%plotit(data=salmon,labelvar=_blank_, symvar=species,
plotvars=fresh marine, color=black, colors=blue);
run;

proc discrim data=salmon method=normal pool=test
             wcov pcov manova;
 class species;
 var fresh marine;
 title2 'Testing the equality of normal population parameters';
run;
```