Lecture 14-15

III Independent Samples

See Also Section 4.3 of Khattree & Naik

Suppose n units are <u>randomly</u> assigned to either treatment 1 or treatment 2.

Randomization is essential to make the inference mathematically valid. Although it does not guarantee elimination of bias, one hopes it will reduce any systematic bias due to inhomogeneity among the units with respect to uncontrolled variables.

Assume 1. $X_{11}, \cdots, X_{1n_1}$ IID (μ_1, Σ_1)

2. $X_{21}, \cdots, X_{2n_2}$ IID (μ_2, Σ_2)

3. the two samples are independent

For small n_1, n_2 further assume

4 a) the two populations are normal

b) $\Sigma_1 = \Sigma_2 \ (= \Sigma)$

A Small sample inference

Estimation

$$\hat{\mu}_1 = \bar{X}_1 \,, \quad \hat{\mu}_2 = \bar{X}_2 \,, \quad \hat{\Sigma} = S_p \overset{\text{def}}{=} \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$$

$$= \frac{1}{n_1+n_2-2} \sum_{k=1}^{2} \sum_{j=1}^{n_k} (X_{kj} - \bar{X}_k)(X_{kj} - \bar{X}_k)'$$

$$E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2$$

$$\text{Cov}[\bar{X}_1 - \bar{X}_2] = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma$$

$$\widehat{\text{Cov}}[\bar{X}_1 - \bar{X}_2] = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_p$$

Testing

$$H_0: \mu_1 - \mu_2 = \delta_0 \quad \text{vs} \quad H_1: \mu_1 - \mu_2 \neq \delta_0$$

$$T^2 = (\bar{X}_1 - \bar{X}_2 - \delta_0)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_p\right]^{-1} (\bar{X}_1 - \bar{X}_2 - \delta_0)$$

$$\sim \frac{(n_1+n_2-2)p}{n_1+n_2-1-p} F_{p,\, n_1+n_2-p-1}$$

Since when H_0 holds,

$$\begin{cases} \bar{X}_1 - \bar{X}_2 - \delta_0 \sim N_p\left(0, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma\right) \\ S_p (n_1+n_2-2) \sim W_{n_1+n_2-2}(\Sigma) \end{cases} \quad \text{indep}$$

(*)

See Result on page 33 (right after the comparison of univariate and multivariate)

If $U \sim N_p(0, \Sigma)$, $A \sim W_m(\Sigma)$ → independent

then $U'A^{-1}U \sim \frac{p}{m-p+1} F_{p,\, m-p+1}$

Confidence Region $(1-\alpha)100\%$ C.R. for $\delta = \mu_1 - \mu_2$

$$\left\{ \delta : \; [\delta - (\bar{x}_1 - \bar{x}_2)]' \left[\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right) S_p\right]^{-1} [\delta - \bar{x}_1 - \bar{x}_2] \right.$$

$$\left. \leq \frac{(n_1+n_2-2)p}{n_1+n_2-1-p} F_{p, n_1+n_2-1-p} \right\}$$

= ellipsoid with center $\bar{x}_1 - \bar{x}_2$ and axes

$$\sqrt{\frac{(n_1+n_2-2)p}{n_1+n_2-1-p}\left(\tfrac{1}{n_1}+\tfrac{1}{n_2}\right) F_c \lambda_i}\; e_i \qquad \text{where } F_c = F^{(\alpha)}_{p, n_1+n_2-1-p}$$

where $S_p = \sum_{i=1}^{p} \lambda_i e_i e_i'$ is the standard spectral decomposition

Simultaneous C.I. for $l'\delta$

$$l'(\bar{x}_1 - \bar{x}_2) \pm \sqrt{\frac{(n_1+n_2-2)p}{n_1+n_2-1-p} F^{(\alpha)}_{p, n_1+n_2-1-p}} \sqrt{\left(\tfrac{1}{n_1}+\tfrac{1}{n_2}\right) l' S_p l}$$

In particular, CI for $\mu_{1i} - \mu_{2i}$

$$(\bar{x}_{1i} - \bar{x}_{2i}) \pm \sqrt{\frac{(n_1+n_2-2)p}{n_1+n_2-1-p} F^{(\alpha)}_{p, n_1+n_2-1-p}} \sqrt{\left(\tfrac{1}{n_1}+\tfrac{1}{n_2}\right)(S_p)_{ii}}$$

Bonferroni C.I. for $\mu_{1i} - \mu_{2i}$

(*) $$(\bar{x}_{1i} - \bar{x}_{2i}) \pm t_{\cancel{n_1+n_2-1-p}\; n-2}\left(\frac{\alpha}{2p}\right) \sqrt{\left(\tfrac{1}{n_1}+\tfrac{1}{n_2}\right)(S_p)_{ii}}$$

B Large sample inference ($\Sigma_1 = \Sigma_2$)

1 Robust wrt non-normality

2 Replace $\frac{(n_1+n_2-2)p}{n_1+n_2-1-p} F_{p, n_1+n_2-1-p}$ by χ^2_p

and follow recipe A. That is,

$$T^2 = (\bar{x}_1 - \bar{x}_2 - \delta_0)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_p\right]^{-1} (\bar{x}_1 - \bar{x}_2 - \delta_0) \overset{\text{approx}}{\sim} \chi^2_p$$

$$CR \quad \left\{\delta : [\delta - (\bar{x}_1 - \bar{x}_2)]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_p\right]^{-1} [\delta - (\bar{x}_1 - \bar{x}_2)] \leq \chi^2_p(\alpha)\right\}$$

$$SCI \text{ for } l'\delta : \; l'(\bar{x}_1 - \bar{x}_2) \pm \sqrt{\chi^2_p(\alpha)} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) l' S_p l}$$

$$\text{for } \mu_{1i} - \mu_{2i} : (\bar{x}_{1i} - \bar{x}_{2i}) \pm \sqrt{\chi^2_p(\alpha)} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)(S_p)_{ii}}$$

$$\text{Bon. C.I. for } \mu_{1i} - \mu_{2i} \text{ is } (\bar{x}_{1i} - \bar{x}_{2i}) \pm t_{n_1+n_2-2}\left(\frac{\alpha}{2p}\right) \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{pii}}$$

C If $\Sigma_1 \neq \Sigma_2$ and n_1, n_2 are moderate, Try some transformation on variables that have drastically different variance estimates than others.

D Large Sample Inference $\Sigma_1 \neq \Sigma_2$

Estimation

$$\hat{\mu}_1 = \bar{x}_1, \quad \hat{\mu}_2 = \bar{x}_2, \quad \hat{\Sigma}_1 = S_1, \quad \hat{\Sigma}_2 = S_2$$

$$(\bar{x}_1 - \bar{x}_2 - \delta_0)' \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} (\bar{x}_1 - \bar{x}_2 - \delta_0) \overset{\text{approx}}{\sim} \chi^2_p$$

since $\bar{x}_1 - \bar{x}_2 \sim N\left(\delta, \frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}\right)$, under H_0.

$(1-\alpha)100\%$ CR for $\delta = \mu_1 - \mu_2$ is

$$\left\{\delta: [\delta - (\bar{x}_1 - \bar{x}_2)]' \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} [\delta - (\bar{x}_1 - \bar{x}_2)] \leq \chi^2_p(\alpha)\right\}$$

= ellipsoid with center $\bar{x}_1 - \bar{x}_2$ and axes

$$\sqrt{\chi^2_p(\alpha)}\, e_i \quad \text{where} \quad \frac{S_1}{n_1} + \frac{S_2}{n_2} = \sum_{i=1}^{p} \eta_i e_i e_i'$$

$(1-\alpha)100\%$ SCI for $l'\delta$ is

$$l'(\bar{x}_1 - \bar{x}_2) \pm \sqrt{\chi^2_p(\alpha)} \sqrt{\frac{l' S_1 l}{n_1} + \frac{l' S_2 l}{n_2}}$$

In particular, SCI for $\mu_{1i} - \mu_{2i}$ is

$$(\bar{x}_{1i} - \bar{x}_{2i}) \pm \sqrt{\chi^2_p(\alpha)} \sqrt{\frac{S_{1ii}}{n_1} + \frac{S_{2ii}}{n_2}}$$

Comparison of g multivariate population means.

Suppose ① $\underline{X}_{l1}, \cdots, \underline{X}_{ln_l}$ is a r.s from $N_p(\mu_l, \Sigma)$

$l = 1, \cdots, g$

② The g samples are independent

Note: The population covariance matrices are same.

$H_0: \mu_1 = \cdots = \mu_g$ or equivalently

$H_0: \tau_1 = \cdots = \tau_g$ where $\tau_l = \mu_l - \mu$ additional effect of treatment l

(overall mean, lth treatment effect)

One way MANOVA model

$$\underline{X}_{lj} = \underline{\mu} + \underline{\tau}_l + \underline{\varepsilon}_{lj}$$

$\sum_l \mu + \mu_l - \mu = \sum_l \mu_l$

$\overset{IID}{\sim} N_p(0, \Sigma)$

Surely, $\sum_{l=1}^{g} \underline{\tau}_l = 0$

Source	df	SS
treatment [Between] (Treatment) sum of squares and cross products	$g-1$	$B = SSTr = \sum_{l=1}^{g} n_l (\bar{x}_l - \bar{x})(\bar{x}_l - \bar{x})'$
residual [Within] (Residual/Error) sum of squares and cross products	$n-g$	$W = SSRes = \sum_{l=1}^{g} \sum_{j=1}^{n_l} (x_{lj} - \bar{x}_l)(x_{lj} - \bar{x}_l)'$
Total (corrected)	$n-1$	$T = SSTotal = \sum_{l=1}^{g} \sum_{j=1}^{n_l} (x_{lj} - \bar{x})(x_{lj} - \bar{x})'$

$n = \sum_{l=1}^{g} n_l$, $T = B + W$ $\quad x_{lj} - \bar{x} = \underbrace{x_{lj} - \bar{x}_l}_{\text{within treatment}} + \underbrace{\bar{x}_l - \bar{x}}_{\text{between treatment}}$

Wilks Lambda (ratio of mles of generalized variance under full model and under H_0)

$$\Lambda^* = \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_0|} = \frac{|W|}{|B+W|} = \prod_{i=1}^{s} \left(\frac{1}{1+\lambda_i}\right)$$

mle

Likelihood ratio $(\Lambda^*)^{\frac{n}{2}}$ is the

λ_i are eigenvalues of $W^{-1}B$

note $W \sim W_{n-g}(\Sigma)$

$B \sim W_{g-1}(\Sigma)$ > independent

provided H_0 is true

Distribution of Λ^* under H_0

Large n, $\quad -\left(n-1-\frac{p+g}{2}\right) \ln \Lambda^* \overset{\text{approx}}{\sim} \chi^2_{p(g-1)}$

Since $\nu - \nu_0 = \#$ parameters specified by $H_0 = p(g-1)$

Moderate n (special cases)

$p=1, 2$ $\quad \frac{1-\Lambda^{*\frac{1}{p}}}{\Lambda^{*\frac{1}{p}}} \sim \frac{p(g-1)}{p(n-g)-2(p-1)} F_{p(g-1),\, p(n-g)-2(p-1)}$

$g = 2,3$ $\quad \frac{1-\Lambda^{*\frac{1}{g-1}}}{\Lambda^{*\frac{1}{g-1}}} \sim \frac{p}{(n-g)-(p-1)} F_{p(g-1),\, [n-g-p+1](g-1)}$

Read 6.7 for two-Way Multivariate analysis of Variance

Example

			Sample average
8	10	9	$\begin{pmatrix}9\\8\end{pmatrix} = \bar{X}_1$
5	9	10	
3	5		$\begin{pmatrix}4\\4\end{pmatrix} = \bar{X}_2$
3	5		
4	9	8	$\begin{pmatrix}7\\8\end{pmatrix} = \bar{X}_3$
7	8	9	
			$\begin{pmatrix}7\\7\end{pmatrix} = \bar{X}$

$g = 3, \quad n = 8$

$$B = \begin{bmatrix} 12+18 & 6+18 \\ 6+18 & 3+18+3 \end{bmatrix} = \begin{bmatrix} 30 & 24 \\ 24 & 24 \end{bmatrix}$$

$$W = \begin{bmatrix} 1+1+1 & 3-1 \\ 1+1+1 & 1-1 \\ 9+4+1 & 3-1 \\ 3-1 & 9+4 \\ 1-1 & 1-1 \\ 3-1 & 1-1 \end{bmatrix} = \begin{bmatrix} 18 & 10 \\ 10 & 18 \end{bmatrix}$$

$$B+W = \begin{bmatrix} & \\ & \end{bmatrix} = \begin{bmatrix} 48 & 34 \\ 34 & 42 \end{bmatrix}$$

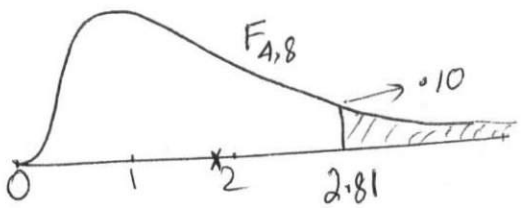
$$\Lambda^* = \frac{|W|}{|B+W|} = \frac{224}{860} = .2604651$$

$$\frac{1-(\Lambda^*)^{\frac{1}{b}}}{(\Lambda^*)^{\frac{1}{b}}} \qquad \frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}} \sim \frac{p}{n-g-p+1} F_{p(g-1),\,(n-g-p+1)(g-1)}$$

$$= \frac{2}{4} F_{4,8}$$

$$F_{obs} = \frac{4}{2} \frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}} = 2(.9594095) = 1.9188$$

$F_{4,8}(.10) = 2.81$ p-value $> .10$



Simultaneous C.I. for treatment effects

$\binom{g}{2}$ possible comparison (pairwise)

use $\alpha' = \frac{\alpha}{p\binom{g}{2}}$ for each test

$H_0: \tau_{ki} = \tau_{\ell i}$, $1 \le k \ne \ell \le g$, $1 \le i \le p$ fixed

$$\widehat{\tau_{ki} - \tau_{\ell i}} = \bar{x}_{ki} - \bar{x}_{\ell i} \sim N\left(0, \left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\sigma_{ii}\right)$$

$$\hat{\sigma}_{ii} = \frac{w_{ii}}{n-g} \quad \text{where} \quad W = \sum_{k=1}^{g}\sum_{j=1}^{n_k}(x_{kj} - \bar{x}_k)(x_{kj} - \bar{x}_k)'$$

$(1-\alpha)100\%$ S.C.I. for $\tau_{ki} - \tau_{\ell i}$ $(k \ne \ell)$ are

$$(\bar{x}_{ki} - \bar{x}_{\ell i}) \pm t_{n-g}\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\frac{w_{ii}}{n-g}}$$

```
options ls=80 ps=47 nodate nonumber;
title1 h=2 'Hook-billed kites';
title2 h=1 'JW: T5-12 T6-11';
data length1;
infile 'Z:\\My Documents\Teaching\Stat524\Fall 2010\Data set\T5-12.dat' firstobs=1;
input tail wing;
sex='female';
run;

data length2;
infile 'Z:\\My Documents\Teaching\Stat524\Fall 2010\Data set\T6-11.dat' firstobs=1;
input tail wing;
sex='male';
if tail >= 284 then delete;
run;

data length;
set length1 length2;
run;

proc print data=length;
run;

proc corr data=length cov;
var tail wing;
by sex;
run;
/*
---------------------------------- sex=female ----------------------------------
                          The CORR Procedure
                       2  Variables:    tail     wing

                           Covariance Matrix, DF = 44
                                      tail                wing
                     tail        120.6949495        122.3459596
                     wing        122.3459596        208.5404040

Simple Statistics
 Variable          N          Mean       Std Dev           Sum       Minimum       Maximum
 tail             45     193.62222      10.98613          8713     173.00000     216.00000
 wing             45     279.77778      14.44093         12590     245.00000     310.00000

---------------------------------- sex=male ------------------------------------
                          The CORR Procedure
                       2  Variables:    tail     wing

                           Covariance Matrix, DF = 43
                                      tail                wing
                     tail         88.3694503         90.0073996
                     wing         90.0073996        172.8350951

                              Simple Statistics
 Variable          N          Mean       Std Dev           Sum       Minimum       Maximum
 tail             44     187.15909       9.40050          8235     170.00000     212.00000
 wing             44     280.95455      13.14668         12362     254.00000     310.00000

*/
proc iml;
nf=45;
p=2;
avef={193.62222, 279.77778};
Sf={120.6949495 122.3459596,
    122.3459596 208.5404040};
```

```
nm=44;
avem={187.15909, 280.95455};
Sm={88.3694503 90.0073996,
    90.0073996 172.8350951};
diff=avef-avem;
n=nf+nm;
df1=p;
df2=n-p-1;
Sp=((nf-1)*Sf+(nm-1)*Sm)/(n-2);
Spi=inv((1/nf+1/nm)*Sp);
T2=diff`*Spi*diff;
alpha=0.05;
falpha=finv(1-alpha,df1,df2);
f=(n-2)*df1/df2*falpha;
print diff Sp, T2 f alpha df1 df2;
c1=sqrt(f);
c2=tinv(1-alpha/(2*p),n-2);
simulcl=diff-c1*sqrt((1/nf+1/nm)*vecdiag(Sp));
simuucl=diff+c1*sqrt((1/nf+1/nm)*vecdiag(Sp));
bonflcl=diff-c2*sqrt((1/nf+1/nm)*vecdiag(Sp));
bonfucl=diff+c2*sqrt((1/nf+1/nm)*vecdiag(Sp));
print c1 c2, simulcl simuucl, bonflcl bonfucl;
run;
/*
                              diff        Sp
                           6.46313 104.71798 106.36253
                          -1.17677 106.36253 190.89295

                    T2          f     alpha       df1       df2
                24.964901 6.2772565      0.05         2        86

                                       c1        c2
                                  2.5054454 2.2808563

                                    simulcl   simuucl
                                  1.0273988 11.898861
                                  -8.515861 6.1623214

                                    bonflcl   bonfucl
                                  1.5146599   11.4116
                                  -7.857982 5.5044423
*/
proc GLM data=length;
class sex;
model tail wing =sex;
manova h=sex/printe printh;
means sex/bon alpha=0.025 lines;
means sex/bon alpha=0.025 cldiff;
run;
quit;
```

Assignment #5

The day assigned: Thursday, October 14, 2010
The day due: Tuesday, October 26, 2010

1. Read Text: Chapter 6
2. Assigned problems: Exercises: 6.5, 6.8

Lecture 16-17

Chapter 8 Analysis of Covariance structure

principle Component Analysis

8.1 Introduction

(a) what is PCA?

<1> PCA is a method for re-expressing multivariate data. It allows to reorient the data so that the first few dimensions account for as much of the available information as possible.

<2> If there is substantial redundency present in the data set. Then it may be possible to account for most of the information in the original data set with a relatively small number of dimensions

(b) objectives

<1> Dimension (data) reduction: When there are many dimensions (variables), it is difficult to comprehend or even visualize the pattern of association among them

When we examine relationship among p (very large) variables, we often look at the covariance matrix.

Sometimes, it is possible to explain the structure with only k (k <= p) linear combinations of the original p variables without loss of much information.

h observation on p variables reduced to
n observation on K ($K \leq 3$) variables

(2) identifying patterns of association among variables, revealing relationship and the allowing interpretations that would not ordinarily result.

(3) serving as intermediate steps in regression, clustering factor analysis, etc.

8.2 population principal Components

(a) principal components

<1> our objective is to find linear combinations of the original variable $X=(X_1, \cdots, X_p)$ which are uncorrelated with variance as large as possible

(2) Find linear combinations of p random variables
$X_1, \cdots, X_p$

$$Y_1 = a_1' X = a_{11} X_1 + \cdots + a_{1p} X_p$$

$$Y_2 = a_2' X = a_{21} X_1 + \cdots + a_{2p} X_p$$

$$\vdots$$

$$Y_p = a_p' X = a_{p1} X_1 + \cdots + a_{pp} X_p$$

such that

$Y_1, \cdots, Y_p$

- are uncorrelated
- represent a new coordinate system obtained by rotating the original one with $x_1, \cdots, x_p$ as axes
- represent the directions with maximum variability

(3) Let $cov(X) = \Sigma$, then

- $Var(Y_i) = Var(a_i'X) = a_i'\Sigma a_i \quad i = 1, \cdots, p$

  $Cov(Y_i, Y_k) = a_i'\Sigma a_k = 0, \quad i \neq k = 1, \cdots, p$

- The 1st principal component $= a_1'x$ that maximize $Var(a'x)$ subject to $a'a = 1$

  $\max_{a} \frac{a'\Sigma a}{a'a} = \lambda_1$, $a = e_1$

  The 2nd ..... $a_2'x$ $(a_2'a_2 = 1)$

  $\max_{a \perp e_1} \frac{a'\Sigma a}{a'a} = \lambda_2$, $a = e_2$

  ..... and $Cov(a_1'x, a_2'x) = 0$

  Th i^{th} ..... $= a_i'x$ $(a_i'a_i = 1)$

  ..... and $Cov(a_i'x, a_j'x) = 0$, $\forall 1 \leq j < i$

(4) Let $\Sigma = cov(X)$ have eigenvalue - eigen vector pairs $(\lambda_1, e_1), (\lambda_2, e_2), \cdots, (\lambda_p, e_p)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ and $\|e_i\| = 1 \ \forall i \leq p$

Then

- $Y_i = e_i'X = e_{i1}X_1 + \cdots + e_{ip}X_p \quad i = 1, \cdots, p$
- $Var(Y_i) = e_i'\Sigma e_i = \lambda_i$

  $Cov(Y_i, Y_j) = 0 \quad i \neq j \quad (e_i'\Sigma e_j = e_i'\lambda_j e_j = 0)$

- $\sum_{i=1}^{p} var(Y_i) = \sum_{i=1}^{p} \lambda_i = trace(\Sigma) = \sum_{i=1}^{p} var(X_i) = \sum_{i=1}^{p} \sigma_{ii}$

↑ Total variance due to principal components

↑ Total population variance

(5) Total variation explained by principal component

- $\frac{\lambda_i}{\sum \lambda_i}$ = proportion of total variance explained by the ith principal component

Thus, we can replace p variables by the first a few principal components if they can account for most variation

(6) Importance of X_k to Y_i

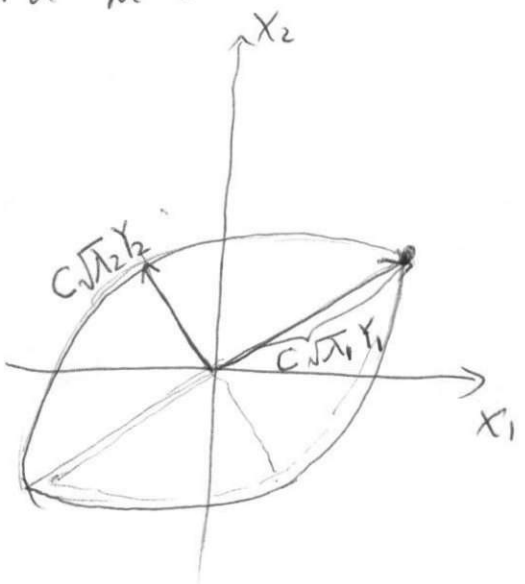
- e_{ik} in $e_i' = (e_{i1}, \cdots, e_{ik}, \cdots e_{ip})$ measures the importance of the kth variable to the ith principal component

$$Y_i = e_{i1}X_1 + \cdots + e_{ik}X_k + \cdots + e_{ip}X_p$$

- $\rho_{Y_i, X_k} = \frac{cov(Y_i, X_k)}{\sqrt{var(Y_i)}\sqrt{var(X_k)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} e_{ik}$

(7) If $X \sim N_p(\mu, \Sigma)$, then contours of constant density $(X-\mu)'\Sigma^{-1}(X-\mu) = c^2$ are ellipsoids centered at μ with $c\sqrt{\lambda_1}e_1, \dots c\sqrt{\lambda_p}e_p$ as the half axes.

For $p=2$ and $\mu=0$



Example (8.1)

Suppose $V(X) = \Sigma$, $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$, $\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

- eigenvalue - eigenvector pairs

$\lambda_1 = 5.83$ $\quad e_1' = (0.383, -0.924, 0)$

$\lambda_2 = 2.00$ $\quad e_2' = (0, 0, 1)$

$\lambda_3 = 0.17$ $\quad e_3' = (0.924, 0.383, 0)$

$(\lambda_1 + \lambda_2 + \lambda_3) = \Sigma\sigma_{ii} = 1+5+2 = 8$

- the principle componends

$Y_1 = e_1'X = 0.383X_1 - 0.924X_2$

$Y_2 = e_2'X = X_3$

$Y_3 = e_3'X = 0.924X_1 + 0.383X_2$

- proportion of total variance accounted by principal components

Y_1: $\frac{\lambda_1}{\sum \lambda_i} = \frac{5.83}{8} = 73\%$

Y_2: $\frac{\lambda_2}{\sum \lambda_i} = \frac{2}{8} = 25\%$

} 98%

X_1, X_2, X_3 can be replaced by Y_1 and Y_2

- The importance of X_i to Y_1

Y_1: $|e_{12}| = |-0.924| > |e_{11}| = |0.383|$

$$\rho_{Y_1, X_2} = \sqrt{\frac{\lambda_1}{\sigma_{22}}}\, e_{12} = -0.998$$

$$\rho_{Y_1, X_1} = \sqrt{\frac{\lambda_1}{\sigma_{11}}}\, e_{11} = \frac{\sqrt{5.83}}{1}\, 0.383 = 0.925$$

X_1 and X_2 are almost equally important to Y_1 with X_2 being slightly more important.

(b) principal components from standardized variable (correlation matrix)

<1> standardized variable and correlation matrix

- Let $EX = (\mu_1, \cdots, \mu_p)'$, $COV(X) = \Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix}$

- Let $Z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}, \cdots, Z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}$

Then $Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_p \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{\sigma_{11}}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{\sigma_{pp}}} \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ \vdots \\ x_p - \mu_p \end{pmatrix} = [\text{diag}(\Sigma)]^{-\frac{1}{2}} (X - \mu)$

$E(Z) = 0$

$COV(Z) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix} = \rho$

(2) Let $(\lambda_1, e_1), \ldots, (\lambda_p, e_p)$ be the eigenvalue - eigenvector pairs of ρ then

- $Y_i = e_i' Z = e_i' (\text{diag}\,\Sigma)^{-\frac{1}{2}} (X - \mu)$
- $\sum \text{var}(Y_i) = \sum \lambda_i = \rho = \sum \text{var}(Z_i)$
- $\rho_{Y_i, Z_k} = \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} e_{ik} = \sqrt{\lambda_i}\, e_{ik}$ $\quad i = 1, \cdots p$, $k = 1, \cdots p$
- proportion of total variance explained by Y_i : $\frac{\lambda_i}{\sum \lambda_i} = \frac{\lambda_i}{p}$

(3) Example (8.2)

Consider X_1 and X_2 with $\Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$

The $\rho = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$

- For Σ We have

$\lambda_1 = 100.16$ $\quad e_1 = [0.040, 0.999]$ $\quad \frac{\lambda_1}{\lambda_1 + \lambda_2} = 99.2\%$

$\lambda_2 = 0.84$ $\quad e_2 = [0.999, -0.040]$ $\quad \frac{\lambda_2}{\lambda_1 + \lambda_2} = 0.8\%$

$Y_i = e_i' X \quad i = 1, 2$

$e_{12} = 0.999 > e_{11} = 0.04$

$$\rho_{Y_1, X_2} = \frac{e_{12}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{0.999\sqrt{100.16}}{\sqrt{100}} = 0.9998$$

$$\rho_{Y_1, X_1} = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{0.04\sqrt{100.16}}{\sqrt{1}} \doteq 0.4$$

For ρ we have

$\lambda_1 = 1.4 \quad e_1' = (0.707, 0.707)$

$\lambda_2 = 0.6 \quad e_2' = (0.707, -0.707)$

$\frac{\lambda_1}{\lambda_1+\lambda_2} = 70\%, \quad \frac{\lambda_2}{\lambda_1+\lambda_2} = 30\%$

$e_{11} = e_{12}$

$$\rho_{Y_1, X_1} = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\rho_{11}}} = \frac{e_{12}\sqrt{\lambda_1}}{\sqrt{\rho_{22}}} = \rho_{Y_1, X_2} = 0.837$$

In this case, we should first do standardization before applying PCA.

(4) When to standardization

- standardization is not an consequential and can make difference
- standardize if
  1°) Variables are in widely different units
  2°) Variable are measured in scales with widely different ranges

- If the response are reasonably commensurable, then covariance form has a greater statistical appeal.

C) principal components for special structured covariance matrix

(1) Diagonal matrix $\Sigma = \begin{bmatrix} \sigma_{11} & & 0 \\ & \sigma_{22} & \\ 0 & & \sigma_{pp} \end{bmatrix}$

- $\lambda_i = \sigma_{ii}$ $e_i' = (0, \cdots, 0, \overset{i\text{th}}{1}, 0, \cdots, 0)$

$Y_i = e_i' X = X_i$

(No need for Principal components)

- Intraclass covariance (equicorrelation) structure

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{bmatrix} \qquad \rho = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

$$= \frac{\Sigma}{\sigma^2} (1-\rho) I + \rho \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$

For ρ.

$\lambda_1 = 1 + (p-1)\rho$ $\quad e_1' = \frac{1}{\sqrt{p}} (1, 1, \cdots, 1)$

$\lambda_2 = 1 - \rho$ $\quad e_2' = \frac{1}{\sqrt{1 \times 2}} (1, -1, 0 \cdots 0)$

$\vdots$

$\lambda_p = 1 - \rho$ $\quad e_i' = \frac{1}{\sqrt{(i-1)(i)}} (1, \cdots \overset{i-1}{1}, \overset{i}{-(i-1)}, 0, \cdots 0)$

$\vdots$

$e_p' = \frac{1}{\sqrt{(p-1)p}} (1, \cdots 1, -(p-1))$

- If $\rho > 0$, then

$$Y_1 = e_1' Z = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} Z_i$$

$$\frac{\lambda_1}{\sum \lambda_i} = \frac{1 + (p-1)\rho}{p} \approx \rho \quad \text{for large } p \text{ or } \rho \text{ close to } 1$$

$(p \to \infty \text{ or } \rho \to 1)$

d. Example

<1> Let X_1 = reading speed

X_2 = reading comprehension

X_3 = arithmetic speed

X_4 = —— comprehension

representing the scores of a seventh-grade child.

- Let the correlation matrix ρ be

$$\rho = \begin{bmatrix} 1 & 0.698 & 0.264 & 0.081 \\ 0.698 & 1 & -0.061 & 0.092 \\ 0.264 & -0.061 & 1 & 0.594 \\ 0.081 & 0.092 & 0.594 & 1 \end{bmatrix}$$

SAS code 8-1

Lecture #18

8.3 Summerizing Sample variation by Principal Components

(a) Sample principal Components

(1) Let $X_1 = \begin{pmatrix} X_{11} \\ \vdots \\ X_{1p} \end{pmatrix}, \cdots, X_n = \begin{pmatrix} X_{n1} \\ \vdots \\ X_{np} \end{pmatrix}$ be n independent observations from p-dimension space.

Data: $X = \begin{pmatrix} X_1' \\ \vdots \\ X_n' \end{pmatrix}$

population X with $E(X) = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$, $Cov(X) = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & & \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix}$

(2) Sample P.Cs

- Let $\bar{x}$, S and R be sample mean, sample covariance and correlation metrix
- ~~Sami~~ by a similar argument of obtaining the population P.Cs, the sample P.Cs are

$$\hat{Y}_i = \hat{e}_i' X, \quad i = 1, \cdots, p \quad \text{with } S(\hat{Y}_i) = \hat{\lambda}_i$$

where $(\hat{\lambda}_i, \hat{e}_i)$ being the eigen value-eigen vector pairs of S, with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$

- $\hat{Y}_1 = \hat{e}_1' X$ maximizes $S(a'x)$ subject to $a'a = 1$

$$\Leftrightarrow \max_{a \neq 0} \frac{a'Sa}{a'a} = \hat{e}_1' S \hat{e}_1 = \lambda_1 = S(\hat{Y}_1)$$

$\hat{Y}_2 = \hat{e}_2' X$ maximizes $S(a'x)$ with $cov(a'x, \hat{e}_1'x) = 0$ and $a'a = 1$. ($cov(a'x, \hat{e}'x) = a'S\hat{e}_1 = \lambda_1 a'e_1 = 0$)

$$\Leftrightarrow \max_{a \perp e_1} \frac{a'Sa}{a'a} = \hat{e}_2' S \hat{e}_2 = \lambda_2 = S(\hat{Y}_2)$$

$\hat{Y}_i = \hat{e}_i' X$ satisfies

$$\max_{a \perp \{e_1 \cdots e_{i-1}\}} \frac{a' S a}{a' a} = \hat{e}_i' S \hat{e}_i = \lambda_i = S(\hat{Y}_i)$$

- $\hat{Y}_i$'s are uncorrelated liner combination with (S.P.C$_S$)

  largest sample variance.

  $cov(\hat{Y}_i, \hat{Y}_j) = 0, \quad i \neq j \qquad V(\hat{Y}_i) = \hat{\lambda}_i$

- Total sample variance $= \sum_{i=1}^{p} s_{ii} = \sum_{i=1}^{p} \hat{\lambda}_i$

  = Total sample variance due to sample principal components

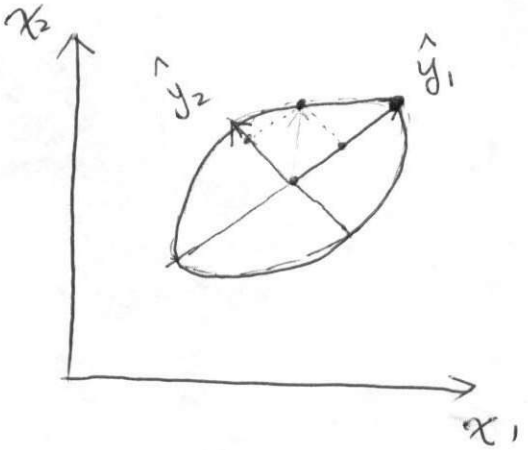
  $\frac{\hat{\lambda}_i}{\sum \hat{\lambda}_i}$ = the propotion of total sample variance explained by the sample principal components

$$\hat{r}_{\hat{y}_i, x_j} = \frac{s(\hat{y}_i, x_j)}{\sqrt{s^2(\hat{y}_i)\, s^2(x_j)}} = \frac{\hat{e}_i' S \delta_j}{\sqrt{\lambda_i} \cdot \sqrt{s_{jj}}} = \frac{\lambda_i e_{ij}}{\sqrt{\lambda_i}\sqrt{s_{jj}}} = \frac{e_{ij}\sqrt{\lambda_i}}{\sqrt{s_{jj}}}$$

- Geometric interpretaion of sample principal components

$$(X-\bar{X})' \sum_{i=1}^{p} \lambda_i e_i e_i' (X-\bar{X})$$

$$= \sum_{i=1}^{p} \lambda_i \left(e_i'(X-\bar{X})\right)^2 = c^2$$



the ellipsoid on $(X-\bar{X})' S^{-1} (X-\bar{X}) = c^2$

- $e_i'(X-\bar{X})$ is on the direction of $\hat{y}_i$

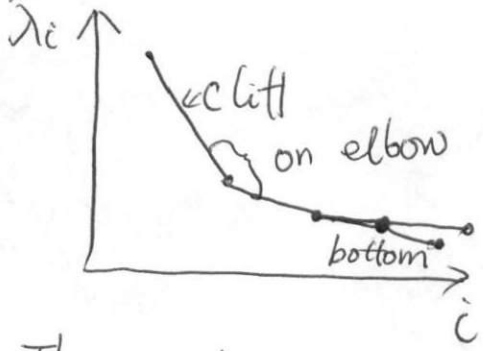
  i.e. the direction of $\hat{e}_i$

- observation $x_1, \ldots, x_n$ are often "centered" by subtracting $\bar{x}$

$$\hat{y}_i = \hat{e}_i'(x - \bar{x}) = e_{i1}(x_1 - \bar{x}_1) + \cdots + e_{ip}(x_p - \bar{x}_p)$$

(b) How many principal components are sufficient to explain sample variation?

1. No definite answer. Need to look at
   - the amount of total sample variance explained
   - The relative sizes of eigen values.
   - subject-matter interpretations of the components
2. Scree plot to help determine the number of principal components



scree plot of $\hat{\lambda}_i$ versus i

The number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size such a point is usually near an elbow.

3. Example 8.4

In study the size and shape relationship of male turtles, The length, width and height of carapases are measured. The natural log of the dimensions of 24 male turtles have been measured.

$$\hat{y}_i = \hat{e}_i'(X - \bar{X})$$

$$S = 10^{-3}\begin{bmatrix} 11.072 & 8.019 & 8.160 \\ 8.019 & 6.417 & 6.005 \\ 8.160 & 6.005 & 6.773 \end{bmatrix}$$

$$\bar{X}' = [4.725,\ 4.478,\ 3.703]$$

$\hat{\lambda}_1 = 23.3 \times 10^{-3}$ $\quad \hat{e}_1' = (0.683,\ 0.510,\ 0.523)$

$\hat{\lambda}_2 = 0.60 \times 10^{-3}$ $\quad \hat{e}_2' = (-0.159,\ -0.594,\ 0.788)$

$\hat{\lambda}_3 = 0.36 \times 10^{-3}$ $\quad \hat{e}_3' = (-0.713,\ 0.622,\ 0.324)$

$$\frac{\hat{\lambda}_1}{\sum \hat{\lambda}_i} = 0.9605 \qquad \frac{\hat{\lambda}_2}{\sum \hat{\lambda}_i} = 0.0247 \qquad \frac{\hat{\lambda}_3}{\sum \hat{\lambda}_i} = 0.0148$$

$$\hat{y}_1 = \ln(\text{length}^{0.683} \cdot \text{width}^{0.51} \cdot \text{height}^{0.523}) - 7.447624$$

measuring the ln (volume) of carapase (with adjused dimensions)

SAS code 8-2.txt

(c) Sample P. C. from standardized obervations

1. Variables measured on different scales or with widely different ranges are often standardized.

- Let $Z_j = \begin{pmatrix} \frac{x_{j1}-\bar{x}_1}{\sqrt{s_{11}}} \\ \vdots \\ \frac{x_{jp}-\bar{x}_p}{\sqrt{s_{pp}}} \end{pmatrix} = D^{-\frac{1}{2}}(X_j - \bar{X})$ $\quad D = \text{diag}(S) = \begin{pmatrix} s_{11} & & \\ & \ddots & \\ & & s_{pp} \end{pmatrix}$

$$Z = \begin{pmatrix} Z_1' \\ \vdots \\ Z_n' \end{pmatrix} = D^{-\frac{1}{2}} \begin{pmatrix} (X_1 - \bar{X})' \\ \vdots \\ (X_n - \bar{X})' \end{pmatrix} = D^{-\frac{1}{2}}\left(X - 1_n \bar{X}'\right)$$

$$S(Z) = D^{-\frac{1}{2}} S D^{-\frac{1}{2}} = R(X)$$

2. The sample P. Cs. of the standardized data Z are

$$\hat{Y}_i = \hat{e}_i' Z \qquad i = 1, \cdots, p, \qquad S^2(\hat{Y}_i) = \hat{\lambda}_i$$

where $(\hat{\lambda}_i, \hat{e}_i)$ are the eigen value - vector pairs of R, $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$

- $S(\hat{Y}_i, \hat{Y}_j) = 0, \quad i \neq j \qquad S^2(\hat{Y}_i) = \hat{\lambda}_i$
- $\sum_{i=1}^{p} \hat{\lambda}_i = p = \sum_{i=1}^{p} r_{ii}$
- $r_{\hat{Y}_i, Z_k} = e_{ik}\sqrt{\lambda_i} \qquad i, k = 1, \cdots, p,$

(3) proportion of sample variance due to principal components

- ith : $\frac{\hat{\lambda}_i}{p}$
- Scree plot

  Select $\hat{y}_i$ with $\hat{\lambda}_i > 1$
- extremly small $\hat{\lambda}_p$, however, may indicate linear dependency in variables

  (eg : $X_p = X_1 + \cdots + X_{p-1}$ delete X_p before analysis)

8.4 Graphing the principal component

a) objective

1. check normality assumption

```
options ls=80 ps=47 nodate nonumber;
title1 h=2 'Male turtle shape data';
title2 h=1 'JW: E8-4';

data turtle;
infile 'Z:\\My Documents\Teaching\Stat524\Fall 2010\Data set\E8-4.dat' firstobs=1;
input length width height;
lglength=log(length);
lgwidth=log(width);
lgheight=log(height);
run;

proc print data=turtle;
run;

title2 "Plot of raw data by length and height";
*%plotit(data=track,labelvar=country,plotvars=m100 m200, color=black, colors=blue);
%plotit(data=turtle, plotvars=lglength lgheight);
run;

*ods html file="sas83.html";
ods graphics on;

proc princomp data=turtle cov out=pcturtle;
var lglength lgwidth lgheight;
run;

ods graphics off;
*ods html close;

proc print data=pcturtle;
run;

title2 "Scatter plot of the first 2 principal components";

*%plotit(data=pctrack, labelvar=country, plotvars=prin2 prin1, color=black,
colors=blue);

%plotit(data=pcturtle, plotvars=prin2 prin1);
```

Lecture #19

8.4 Graphing the principal components

a) objectives

<1> check normality assumption

<2> Identify suspect observations

b) checking normality assumption

<1> Q-Q plot

- If population $X \sim N_p(\mu, \Sigma)$, then the ith P.C.

$$Y_i = e_i' X = e_{i1}X_1 + \cdots + e_{ip}X_p$$

is also normal

- For a given sample $\vec{X}_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1p} \end{pmatrix}, \cdots, \vec{X}_n = \begin{pmatrix} x_{n1} \\ \vdots \\ x_{np} \end{pmatrix}$ from $\vec{X}$

We have < For each $i = 1, \cdots p$, $\hat{Y}_i = \begin{pmatrix} \hat{y}_{1i} \\ \vdots \\ \hat{y}_{ni} \end{pmatrix}$ with

$\hat{y}_{ji} = \hat{e}_i' \hat{X}_j = \hat{e}_{i1} x_{j1} + \cdots + \hat{e}_{ip} x_{jp}$ $j = 1, \cdots n$ which should also be normal >

- We can construct Q-Q plot for each $\hat{Y}_i$ to check normality

(2) Scatter plot

- If $X_1, \cdots, X_n \sim N_p(\mu, \Sigma)$, then any pair of S.P.Cs $\hat{Y}_i, \hat{Y}_j$ should be jointly normal

then we can construct scatter plot to check the normality

4C) Identify suspect observations

<1> expression for X_j $\left(X_1, \cdots, X_n \overset{iid}{\sim} X \sim N_p(\mu, \Sigma)\right)$

- Let $\hat{e}_1, \cdots, \hat{e}_p$ be the eigen vectors of sample covariance matrix S. Let $\hat{P} = (\hat{e}_1, \cdots, \hat{e}_p)$, then

$$\hat{P}\hat{P}' = I_{n \times n}$$

$$X_{j1}\begin{pmatrix}1\\0\\0\\\vdots\end{pmatrix} + X_{j2}\begin{pmatrix}0\\1\\0\\\vdots\end{pmatrix} + \cdots + X_{jp}\begin{pmatrix}0\\\vdots\\0\\1\end{pmatrix} = \begin{pmatrix}X_{j1}\\\vdots\\X_{jp}\end{pmatrix} = X_j$$

$$= \hat{P}\hat{P}'X_j$$

$$= (\hat{e}_1 \cdots \hat{e}_p)\begin{pmatrix}\hat{e}_1'\\\vdots\\\hat{e}_p'\end{pmatrix} X_j$$

$$= (\hat{e}_1 \cdots \hat{e}_p)\begin{pmatrix}\hat{Y}_{j1}\\\vdots\\\hat{Y}_{jp}\end{pmatrix}$$

$$= \hat{Y}_{j1}\hat{e}_1 + \cdots + \hat{Y}_{jp}\hat{e}_p$$

(2) Identify suspect obs.

- If $\hat{y}_1, \cdots, \hat{y}_q$ can fit the obs very well

then $\|D\|^2 = \|X_j - (\hat{y}_{j1}\hat{e}_1 + \cdots + \hat{y}_{jq}\hat{e}_q)\|_2^2 \quad q \le p$

$$= \|\hat{y}_{jq+1}\hat{e}_{q+1} + \cdots + \hat{y}_{jp}\hat{e}_p\|_2^2$$

$$= \hat{y}_{jq+1}^2 + \cdots + \hat{y}_{jp}^2$$

should be small.

- Suspect observations will often have at least one large coordinate among $\hat{y}_{j,q+1}, \cdots, \hat{y}_{jp}$.
- Strategy: construct scatter plots and Q-Q plots for last few principal components to identify suspect observation

$$\hat{y}_p = \hat{e}_{p1}x_1 + \cdots + \hat{e}_{pp}x_p = \hat{e}_p'x$$
$$\hat{y}_{p-1} = \hat{e}_{p-1}x_1 + \cdots + \hat{e}_{p-1,p}x_p = \hat{e}_{p-1}'x$$

(d) examples ~~8.7~~ 8.4 8.18:

SAS PCA of national track records of Women

8.5 Large sample inference

a) Large sample properties of $\hat{\lambda}_i$ and $\hat{e}_i$

<1> Large sample results for $\hat{\lambda}_i, \hat{e}_i$

- Let $X_1, \cdots X_n \overset{iid}{\sim} X \sim N_p(\mu, \Sigma)$

Let $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p > 0$ be eigen values of Σ

Let $\lambda = (\lambda_1, \cdots, \lambda_p)'$ and $\hat{\lambda} = (\hat{\lambda}_1, \cdots, \hat{\lambda}_p)'$

$$\Lambda = \text{diag}(\lambda) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

Then

(i) $\sqrt{n}(\hat{\lambda} - \lambda) \overset{d}{\Longrightarrow} N_p(0, 2\Lambda^2)$

(ii) $\sqrt{n}(\hat{e}_i - e_i) \overset{d}{\Longrightarrow} N_p(0, E_{-i})$

$$E_{-i} = \lambda_i \sum_{j \neq i} \frac{\lambda_j}{\lambda_j - \lambda_i} e_j e_j'$$

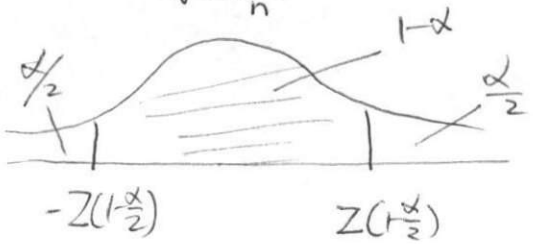
(iii) $\hat{\lambda}_i$ and $\hat{e}_{ij}$, $j = 1, \cdots, p$ are independent. $i = 1, \cdots, p$

<2> Confidence Interval for λ_i

- By the above result, we have for large n,

$$\hat{\lambda}_i \sim N(\lambda_i, \frac{2\lambda_i^2}{n})$$

Thus $$P\left(\frac{|\hat{\lambda}_i - \lambda_i|}{\sqrt{\frac{2\lambda_i^2}{n}}} \leq Z(1 - \frac{\alpha}{2}) \right) \doteq 1 - \alpha$$



Hence, $100(1-\alpha)\%$ C.I for λ_i is

$$-Z(1-\tfrac{\alpha}{2}) \le \frac{\hat{\lambda}_i - \lambda_i}{\lambda_i\sqrt{\frac{2}{n}}} \le Z(1-\tfrac{\alpha}{2})$$

$$\frac{\hat{\lambda}_i}{1+\sqrt{\frac{2}{n}}\,Z(1-\frac{\alpha}{2})} \le \lambda_i \le \frac{\hat{\lambda}_i}{1-\sqrt{\frac{2}{n}}\,Z(1-\frac{\alpha}{2})}$$

(b) Testing for the equal correlation structure

(1) Large sample result

- Let $\Sigma_0 = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & & \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix}$ and $\sigma_{ij} = \sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}\,\rho_{ij}$ with $\rho_{ij} = \rho$ $i,j = 1,\dots,p$

Thus $\rho_0 = \begin{bmatrix} 1 & \cdots & \rho \\ \rho & & 1 \end{bmatrix}$

Let $H_0: \rho = \rho_0$ v.s $H_1: \rho \neq \rho_0$

Let $R = \begin{bmatrix} r_{11} & \cdots & r_{1p} \\ \vdots & & \\ r_{p1} & \cdots & r_{pp} \end{bmatrix}$ be the sample correlation matrix

$$\bar{r}_{\cdot i} = \frac{1}{p-1}\sum_{j\neq i} r_{ij}$$

$$\bar{r}_{\cdot\cdot} = \frac{1}{p}\sum_{i=1}^{p}\bar{r}_{\cdot i}$$

$$\hat{\gamma} = \frac{(p+1)^2\left[1-(1-\bar{r}_{\cdot\cdot})^2\right]}{p-(p+2)(1-\bar{r}_{\cdot\cdot})^2}$$

Then $$T = \frac{n-1}{(1-\bar{r}_{\cdot\cdot})^2}\left[\sum_{i<j}(r_{ij}-\bar{r}_{\cdot\cdot})^2 - \hat{\gamma}\sum_{i=1}^{p}(\bar{r}_{\cdot i}-\bar{r}_{\cdot\cdot})^2\right]$$

$$\sim \chi^2\left(\frac{(p+1)(p-2)}{2}\right)$$

(2) testing

- By the large sample result of T

We reject H_0 at level α if

$$T > \chi^2_{\frac{(p+1)(p-2)}{2}}(1-\alpha)$$

When n is very large.

```
options ls=85 ps=65; *nodate nonumber;
title "SAS PCA of national track records of women";
data track;
  *infile 'Z:\\My Documents\Teaching\Stat524\Fall 2010\Data set\T1-9.dat' firstobs=1
;
  input  country$  m100 m200 m400 m800 m1500 m3000 marathon ;
  cards;
ARG     11.57   22.94   52.50   2.05    4.25    9.19    150.32
AUS     11.12   22.23   48.63   1.98    4.02    8.63    143.51
AUT     11.15   22.70   50.62   1.94    4.05    8.78    154.35
BEL     11.14   22.48   51.45   1.97    4.08    8.82    143.05
BER     11.46   23.05   53.30   2.07    4.29    9.81    174.18
BRA     11.17   22.60   50.62   1.97    4.17    9.04    147.41
.
.
.
;
run;

proc print data=track;
run;

title2 "Plot of raw data by m100 and m200";
%plotit(data=track,labelvar=country,plotvars=m100 m200, color=black, colors=blue);
run;

ods graphics on;
proc princomp data=track out=pctrack;
  *var m100 m200 m400 m800 m1500 m3000 marathon;
run;
ods graphics off;

proc sort data=pctrack;
  by prin1;
run;

proc print;
  id country;
  var prin1 prin2 m100 m200 m400 m800 m1500 m3000 marathon;
  title2 "Rankings by the 1st PC: Overall performance";
run;

proc sort data=pctrack;
  by prin2;
run;

proc print;
  id country;
  var prin1 prin2 m100 m200 m400 m800 m1500 m3000 marathon;
  title2 "Rankings by the 2nd PC: Short- vs long- range performance";
run;

title2 "Scatter plot of the first 2 principal components";
%plotit(data=pctrack, labelvar=country, plotvars=prin2 prin1, color=black,
colors=blue);

title2 "Scatter plot of the first 2 principal components";
%plotit(data=pctrack, labelvar=country, plotvars=prin7 prin6, color=black,
colors=blue);

proc univariate data=pctrack normal plot;
var prin7 prin6;
run;
```

Assignment #6

The day assigned: Tuesday, November 2, 2010
The day due: Thursday, November 11, 2010

Read Text: Chapters 8 and 9

1. Assigned problems: 8.20, 8,21, 8.22

2. Suggested additional problems: 8.10, 8.11, 8.13