ORIGIN ≡ 1

Profile Analysis

Profile analysis offers a way to go beyond multivariate comparison of mean vectors between populations. A "profile", graphed below, consists of a comparison of mean values for *each variable* within and between datasets. This typically only makes sense when the variables have the same units, or are derived from questions collected on the same scale. Comparison of profiles naturally lead to three questions:

1) Are the profiles parallel (i.e., the pattern of "up" versus "down" for the line segments the same)?
2) If parallel, are profiles coincident (overlapping)?
3) If parallel & coincident (or even if not) are profiles level (all variable means of profiles the same)?

For two populations, statistical hypotheses can be formulated for each question, typically in order, using variants of Hotelling's T^2. Sometimes embedded in this approach is the use of a "contrast matrix"C, and T^2 calculated in terms of linear combinations (using identities described in worksheet MTB 030). Although common to introductory texts in Multivariate Statistics, and useful, I have yet to find an implementation in R. So included is my own function in the associated R script. The example is drawn from AC. Rencher (AR) *Methods of Multivariate Analysis* 1995. Example 5.9.2 p. 163, using Table 5.1. Further clarification, came from Example 6.15 in RA Johnson & DW Wichern (JW) *Applied Multivariate Statistical Analysis 4th Edition* 1998. It is interesting to note a slight discrepency between these authors in formulation of the third test for combined data. However, the difference is probably not important.

Read Data:

Psychological Test scores 1=males, 2=females AR Table 5.1..

M := READPRN("c:\DATA\Multivariate\AR-T5-1-longform.txt")

X := submatrix(M, 1, 32, 3, 6)

Y := submatrix(M, 33, 64, 3, 6)

n_1 := rows(X) $n_1 = 32$ n_2 := rows(Y) $n_2 = 32$

p := cols(X) p = 4

M =

	1	2	3	4	5	6
1	1	1	15	17	24	14
2	2	1	17	15	32	26
3	3	1	15	14	29	23
4	4	1	13	12	10	16
5	5	1	20	17	26	28
6	6	1	15	21	26	21
7	7	1	15	13	26	22
8	8	1	13	5	22	22
9	9	1	14	7	30	17
10	10	1	17	15	30	27
11	11	1	17	17	26	20
12	12	1	17	20	28	24
13	13	1	15	15	29	24
14	14	1	18	19	32	28
15	15	1	18	18	31	27
16	16	1	15	14	26	21
17	17	1	18	17	33	26

Summary Statistics:

i := 1 .. n_1 ii := 1 .. n_2 j := 1 .. p

I_1 := identity(n_1) I_2 := identity(n_2) l_{n1_i} := 1 $l_{n2_{ii}}$:= 1

Mean Vectors:

$$X_{bar} := \frac{1}{n_1} \cdot X^T \cdot l_{n1} \qquad X_{bar} = \begin{pmatrix} 15.969 \\ 15.906 \\ 27.188 \\ 22.75 \end{pmatrix}$$

$$Y_{bar} := \frac{1}{n_2} \cdot Y^T \cdot l_{n2} \qquad Y_{bar} = \begin{pmatrix} 12.344 \\ 13.906 \\ 16.656 \\ 21.938 \end{pmatrix}$$

Covariance Matrices:

$$S_1 := \frac{1}{n_1 - 1} \cdot X^T \cdot \left(I_1 - \frac{1}{n_1} \cdot 1_{n1} \cdot 1_{n1}^T \right) \cdot X$$

$$S_1 = \begin{pmatrix} 5.19254032 & 4.5453629 & 6.52217742 & 5.25 \\ 4.5453629 & 13.18447581 & 6.76008065 & 6.26612903 \\ 6.52217742 & 6.76008065 & 28.6733871 & 14.46774194 \\ 5.25 & 6.26612903 & 14.46774194 & 16.64516129 \end{pmatrix}$$

$$S_2 := \frac{1}{n_2 - 1} \cdot Y^T \cdot \left(I_2 - \frac{1}{n_2} \cdot 1_{n2} \cdot 1_{n2}^T \right) \cdot Y$$

$$S_2 = \begin{pmatrix} 9.13608871 & 7.54939516 & 4.86391129 & 4.15120968 \\ 7.54939516 & 18.60383065 & 10.22479839 & 5.44556452 \\ 4.86391129 & 10.22479839 & 30.03931452 & 13.49395161 \\ 4.15120968 & 5.44556452 & 13.49395161 & 27.99596774 \end{pmatrix}$$

Observed Difference Vector:

$$d := X_{bar} - Y_{bar} \qquad d = \begin{pmatrix} 3.625 \\ 2 \\ 10.531 \\ 0.813 \end{pmatrix}$$

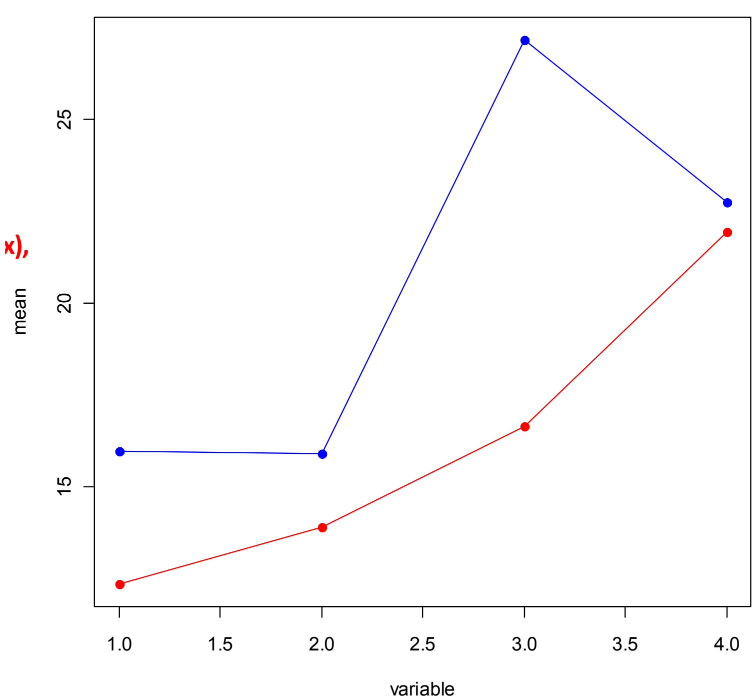
Pooled Covariance Matrix:

$$S_{pooled} := \frac{(n_1 - 1) \cdot S_1 + (n_2 - 1) \cdot S_2}{(n_1 + n_2 - 2)}$$

$$S_{pooled} = \begin{pmatrix} 7.16431452 & 6.04737903 & 5.69304435 & 4.70060484 \\ 6.04737903 & 15.89415323 & 8.49243952 & 5.85584677 \\ 5.69304435 & 8.49243952 & 29.35635081 & 13.98084677 \\ 4.70060484 & 5.85584677 & 13.98084677 & 22.32056452 \end{pmatrix}$$

Profile Plot in R:

```
#PLOT PROFILE:

minX=min(Xbar)
minY=min(Ybar)
min=min(minX,minY)
maxX=max(Xbar)
maxY=max(Ybar)
max=max(maxX,maxY)

maxmin=cbind(maxX,maxY)

index=c(0,p)

plot(index,maxmin,xlim=c(1,p),ylim=c(min,max),
type='n',xlab='variable',ylab='mean')

lines(seq(1:4),Xbar,pch=19,col='blue')
lines(seq(1:4),Ybar,pch=19,col='red')

points(seq(1:4),Xbar,pch=19,col='blue')
points(seq(1:4),Ybar,pch=19,col='red')
```



Test for Parallel Profiles:

Assumptions:

Given variance in the means, do the two profiles have the same shape?

X $[X_1, X_2, \ldots, X_p]$ rs $N_p(\mu_1, \Sigma_1)$
Y $[Y_1, Y_2, \ldots, Y_p]$ rs $N_p(\mu_2, \Sigma_2)$
X independent of Y
Because n_1 & n_2 are small $\Sigma_1 = \Sigma_2$

Specify Contrast Matrix C:

Linear combinations of mean difference vector (X_{bar}-Y_{bar}) and variance-covariance matrix S_{pooled} with C:

$$C := \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

Linear Combinations:

$$C\cdot\left(X_{bar} - Y_{bar}\right) = \begin{pmatrix} -1.625 \\ 8.531 \\ -9.719 \end{pmatrix} \qquad C\cdot S_{pooled}\cdot C^T = \begin{pmatrix} 10.96371 & -7.04738 & -1.64415 \\ -7.04738 & 28.26562 & -12.73891 \\ -1.64415 & -12.73891 & 23.71522 \end{pmatrix}$$

Hypotheses:

$H_{01} : C\mu_1 = C\mu_2$

$H_{11} : C\mu_1 <> C\mu_2$

Each segment of the profile is compared by looking at similar levels at each end. That's what the contrast matrix C tests!

Hotelling's Test Statistic:

$$T_{sq01} := \left[C\cdot\left(X_{bar} - Y_{bar}\right)\right]^T \cdot \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\cdot\left(C\cdot S_{pooled}\cdot C^T\right)\right]^{-1} \cdot \left[C\cdot\left(X_{bar} - Y_{bar}\right)\right] \qquad T_{sq01} = (74.2404)$$

Sampling Distribution:

If Assumptions hold and H_0 is true, then $T_{sq} \sim T^2_{(n-1)} = [(n_1+n_2-2)\cdot(p-1))/(n_1+n_2-p-1)]\ F_{(p,n1+n2-p)}$

Critical Value of the Test:

$\alpha := 0.01$ **< Probability of Type I error must be explicitly set**

$$c_{01} := \frac{\left(n_1 + n_2 - 2\right)\cdot(p-1)}{\left(n_1 + n_2 - p\right)}\cdot qF\left(1 - \alpha, p - 1, n_1 + n_2 - p\right) \qquad c_{01} = 12.79026$$

^ NOTE: qF(1-α) is used in function.

Decision Rule:

Reject H_0 if $T_{sq} > c_{01}$

$T_{sq01} = (74.24)$ $c_{01} = 12.790265$

$\text{Decision} := \text{if}\left(T_{sq01_1} > c_{01}, 1, 0\right)$

Decision = 1 **< 0 = Do not reject H_0**
1 = Reject H_0

^ Therefore REJECT H_0

Probability Value:

$$P_{01} := 1 - pF\left(T_{sq01}, p - 1, n_1 + n_2 - p\right) \qquad P_{01} = (0)$$

Discriminant Function Coefficient Vector:

$$\left(C\cdot S_{pooled}\cdot C^T\right)^{-1}\cdot C\cdot d = \begin{pmatrix} -0.13561 \\ 0.10434 \\ -0.36316 \end{pmatrix}$$

Larger magnitudes here indicate greater effect of the variables. Third segment had greatest effect in discriminating a difference between populations X & Y in the presence of the other variables.

Test for Coincident Profiles:

Assumptions:

Given variance in the means, do the two profiles overlap in value?

Same as above.

Construct Unit Vector:

$i := 1 .. p \qquad 1_i := 1$

Hypotheses:

$H_{02} : 1'\mu_1 = 1'\mu_2$

$H_{12} : 1'\mu_1 <> 1'\mu_2$

The sum of means for populations X & Y are the same.

Hotelling's T^2 Test Statistic:

$$T_{sq02} := 1^T \cdot (X_{bar} - Y_{bar}) \cdot \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \cdot \left(1^T \cdot S_{pooled} \cdot 1 \right) \right]^{-1} \cdot \left[1 \cdot (X_{bar} - Y_{bar}) \right]$$

$T_{sq02} = (28.044)$

Equivalent t-statistic:

$\sqrt{T_{sq02}} = (5.2957)$

$$t_{02} := \frac{\left(1^T \cdot d\right)_1}{\sqrt{1^T \cdot S_{pooled} \cdot 1 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$t_{02} = (5.2957)$

Sampling Distribution:

If Assumptions hold and H_0 is true, then $T_{sq} \sim T^2_{(n-1)} = [(n_1+n_2-2) \cdot (p-1))/(n_1+n_2-p-1)] \, F_{(p,n1+n2-p)}$

Critical Values of the Test:

$\alpha := 0.005$ **< Probability of Type I error must be explicitly set**

$c_{02T} := qF(1 - \alpha, 1, n_1 + n_2 - 2)$ $\qquad c_{02T} = 8.4737$

^ NOTE: qF(1-α) is used in function.

$c_{02t} := qt(1 - \alpha, n_1 + n_2 - 2)$ $\qquad c_{02t} = 2.65748$

Decision Rule:

Reject H_0 if $T_{sq02} > c_{02T}$ $\qquad T_{sq02} = (28.044) \qquad c_{02T} = 8.47373$

Reject H_0 if $t_{02} > c_{02t}$ $\qquad t_{02} = (5.296) \qquad c_{02t} = 2.65748$

$\text{Decision} := \text{if}\left(T_{sq02_1} > c_{02T}, 1, 0\right)$ $\qquad \text{Decision} = 1$

< 0 = Do not reject H_0
1 = Reject H_0

^ Therefore REJECT H_0

Probability Value:

$P_{02} := 1 - pF(T_{sq02}, 1, n_1 + n_2 - 2)$ $\qquad P_{02} = (0.00000166)$

Test for Combined Level Profile:

Assumptions:

Only if one assumes Parallel and Coincident Profiles:
Is the Combined Profile level?

Same as above.

Construct Unified Data Set (UD):

$UD := \text{stack}(X, Y) \qquad N := n_1 + n_2 \qquad i := 1..N \qquad 1_{B_i} := 1 \qquad I_B := \text{identity}(N)$

Mean Vector and Covariance Matrix for UD:

$$UD_{bar} := \frac{1}{N} \cdot UD^T \cdot 1_B \qquad UD_{bar} = \begin{pmatrix} 14.156 \\ 14.906 \\ 21.922 \\ 22.344 \end{pmatrix}$$

$$S_{UD} := \frac{1}{N-1} \cdot UD^T \cdot \left(I_B - \frac{1}{N} \cdot 1_B \cdot 1_B^T \right) \cdot UD \qquad S_{UD} = \begin{pmatrix} 10.388 & 7.793 & 15.298 & 5.374 \\ 7.793 & 16.658 & 13.707 & 6.176 \\ 15.298 & 13.707 & 57.057 & 15.932 \\ 5.374 & 6.176 & 15.932 & 22.134 \end{pmatrix}$$

Hypotheses:

$H_{03} : C\mu = 0$ **All means in the combined population are the same.**

$H_{03} : C\mu_1 <> 0$

Hotelling Test Statistic:

$T_{sq03} := N \cdot (C \cdot UD_{bar})^T \cdot \left(C \cdot S_{UD} \cdot C^T \right)^{-1} \cdot (C \cdot UD_{bar})$ **< JW's formula 6-64 p. 320** $T_{sq03} = (256.362)$

$T_{sq03} := N \cdot (C \cdot UD_{bar})^T \cdot \left(C \cdot S_{pooled} \cdot C^T \right)^{-1} \cdot (C \cdot UD_{bar})$ **< AR's formula 5.39 p. 163** $T_{sq03} = (254.004)$

Sampling Distribution:

If Assumptions hold and H_0 is true, then $T_{sq} \sim T^2_{(n-1)} = [(n_1+n_2-2) \cdot (p-1))/(n_1+n_2-p-1)] \, F_{(p,n1+n2-p)}$

Critical Value of the Test:

$\alpha := 0.01$ **< Probability of Type I error must be explicitly set**

$$c_{03} := \frac{(N-1) \cdot (p-1)}{(N-p+1)} \cdot qF(1-\alpha, p-1, N-p+1) \qquad c_{03} = 12.765$$

^ NOTE: qF(1-α) is used in function.

Decision Rule:

Reject H_0 if $T_{sq} > c_{03}$ $T_{sq03} = (254.004)$ $c_{03} = 12.765066$

$\text{Decision} := \text{if}\left(T_{sq03_1} > c_{03}, 1, 0\right)$ $\text{Decision} = 1$ **< 0 = Do not reject H_0**
1 = Reject H_0

^ Therefore REJECT H_0

Probability Value:

$P03 := 1 - pF(T_{sq03}, p-1, N-p+1)$ $P03 = (0)$

Test for Separate Level Profiles:

If one has rejected hypotheses H_{01} and/or H_{02} it usually makes more sense to Test populations X and Y separately for level profiles. This is done using the same format as for H_{03} above, calculating separate Hotelling test statistics:

Hotelling Test Statistics:

$$T_{sq03X} := n_1 \cdot (C \cdot X_{bar})^T \cdot (C \cdot S_1 \cdot C^T)^{-1} \cdot (C \cdot X_{bar}) \qquad T_{sq03X} = (221.126)$$

$$T_{sq03Y} := n_2 \cdot (C \cdot Y_{bar})^T \cdot (C \cdot S_2 \cdot C^T)^{-1} \cdot (C \cdot Y_{bar}) \qquad T_{sq03Y} = (103.483)$$

Critical Value of the Test:

$\alpha := 0.01$ < **Probability of Type I error must be explicitly set**

$$c_{03X} := \frac{(n_1 - 1) \cdot (p - 1)}{(n_1 - p + 1)} \cdot qF(1 - \alpha, p - 1, n_1 - 1) \qquad c_{03X} = 14.379$$

$$c_{03Y} := \frac{(n_2 - 1) \cdot (p - 1)}{(n_2 - p + 1)} \cdot qF(1 - \alpha, p - 1, n_2 - 1) \qquad c_{03X} = 14.379$$

Decision Rules:

Reject H_0 if $T_{sq03X} > c_{03X}$ $\qquad T_{sq03X} = (221.126) \qquad c_{03X} = 14.3787$

Reject H_0 if $T_{sq03Y} > c_{03Y}$ $\qquad T_{sq03Y} = (103.483) \qquad c_{03Y} = 14.3787$

Probability Value:

$$P_{03X} := 1 - pF(T_{sq03X}, p - 1, n_1 - 1) \qquad P_{03X} = (0)$$

$$P_{03Y} := 1 - pF(T_{sq03Y}, p - 1, n_2 - 1) \qquad P_{03Y} = (0)$$

Prototype in R:

```
#TWO SAMPLE PROFILE ANALYSIS:
#LOAD DATA:
M=read.table("c:/DATA/Multivariate/AR-T5-1-longform.txt")
M   #DATA TABLE
X=M[(M$sex=='1'),2:5]
Y=M[(M$sex=='2'),2:5]
X
Y
Spooled=((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
Spooled
```

```
> Spooled
         y1        y2        y3        y4
y1 7.164315  6.047379  5.693044  4.700605
y2 6.047379 15.894153  8.492440  5.855847
y3 5.693044  8.492440 29.356351 13.980847
y4 4.700605  5.855847 13.980847 22.320565
```

```
#CONTRAST MATRIX:
C=matrix(c(-1,1,0,0,0,-1,1,0,0,0,-1,1),nrow=3,ncol=4,byrow=T)
C
```

```
> C
     [,1] [,2] [,3] [,4]
[1,]   -1    1    0    0
[2,]    0   -1    1    0
[3,]    0    0   -1    1
```

#FUNCTION FOR TWO SAMPLE PROFILE ANALYSIS:
#X = dataset 1
#Y = dataset 2
#alpha = alpha of the test

T2Profile.Analysis <- function(X,Y,alpha)
… body of function in R script …

RES=T2Profile.Analysis(X,Y,alpha=0.01)
RES=T2Profile.Analysis(X,Y,alpha=0.005)

> RES=T2Profile.Analysis(X,Y,alpha=0.01)

```
Hotelling's T2 Profile Analysis

Test for Parallel Profiles:

Hotelling's T2 =  74.24037   critical value =  12.79026   Probability =  0

Test for Coincident Profiles:

Hotelling's T2 =  28.04441   critical value =  7.062192   Probability =  1.657157e-06
Student's t =  5.295698   critical value =  2.388011

Test for Combined Level Profile:

Hotelling's T2 (JW) =  256.3622   critical value =  12.76507   Probability =  0
Hotelling's T2 (AR) =  254.0038   critical value =  12.76507   Probability =  0

Test for Separate Level Profiles:

Hotelling's T2 for X =  221.1263   critical value =  14.37872   Probability =  0
Hotelling's T2 for Y =  103.4827   critical value =  14.37872   Probability =  3.330669e-16
```

> RES=T2Profile.Analysis(X,Y,alpha=0.005)

```
Hotelling's T2 Profile Analysis

Test for Parallel Profiles:

Hotelling's T2 =  74.24037   critical value =  14.65987   Probability =  0

Test for Coincident Profiles:

Hotelling's T2 =  28.04441   critical value =  8.473728   Probability =  1.657157e-06
Student's t =  5.295698   critical value =  2.657479

Test for Combined Level Profile:

Hotelling's T2 (JW) =  256.3622   critical value =  14.62785   Probability =  0
Hotelling's T2 (AR) =  254.0038   critical value =  14.62785   Probability =  0

Test for Separate Level Profiles:

Hotelling's T2 for X =  221.1263   critical value =  16.68843   Probability =  0
Hotelling's T2 for Y =  103.4827   critical value =  16.68843   Probability =  3.330669e-16
```