Multivariate Statistical Analysis

Fall 2011

C. L. Williams, Ph.D.

Lecture 13 for Applied Multivariate Analysis

Outline

1. Two sample Profile Analysis-reprise

Two sample Profile Analysis

Suppose two independent groups or samples receive the same set of p tests or measurements. If these tests are comparable, for example, all on a scale of 0 to 100, the variables will often be commensurate. Rather than testing the hypothesis that $\mu_1 = \mu_2$ we wish to be more specific in comparing the profiles obtained by connecting the points (j, μ_{1j}), $j = 1, 2, \ldots, p$, and (j, μ_{2j}), $j = 1, 2, \ldots, p$. There are three hypotheses of interest in comparing he profiles of two samples. The first of these hypotheses addresses the question, "Are the two profiles similar in appearance, or more precisely, are they parallel? We illustrate this hypothesis in Figure 5.4. If the two profiles are parallel, then one group scored uniformly better than the other group on all p tests.

$$H_0 : \begin{pmatrix} \mu_{12} & - & \mu_{11} \\ \mu_{13} & - & \mu_{12} \\ \mu_{14} & - & \mu_{13} \\ & \vdots & \\ \mu_{1p} & - & \mu_{1,p-1} \end{pmatrix} = \begin{pmatrix} \mu_{22} & - & \mu_{21} \\ \mu_{23} & - & \mu_{22} \\ \mu_{24} & - & \mu_{23} \\ & \vdots & \\ \mu_{2p} & - & \mu_{2,p-1} \end{pmatrix}$$

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

The parallelism hypothesis can be defined in terms of the slopes. The two profiles are parallel if the two slopes for each segment are the same. If the two profiles are parallel, the two increments for each segment are the same, and it is not necessary to use the actual slopes to express the hypothesis. We can simply compare the increase from one point to the next.

To Test for parallelism-Two profiles

$$H_{01}:\mathbf{C}\boldsymbol{\mu}_1 = \mathbf{C}\boldsymbol{\mu}_2$$

$$\mathbf{C} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

From two samples, $\mathbf{y}_{11}, \mathbf{y}_{12}, \ldots, \mathbf{y}_{1n_1}$ and $\mathbf{y}_{21}, \mathbf{y}_{22}, \ldots, \mathbf{y}_{2n_2}$, we obtain $\overline{\mathbf{y}}_1$, $\overline{\mathbf{y}}_2$, and $\mathbf{S}_{pl}$ as estimates of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$.

As in the two-sample T^2-test, we assume that each $\mathbf{y}_{1i}$ in the first sample is $MVN_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, and each $\mathbf{y}_{2i}$ in the second sample is $MVN_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. If $\mathbf{C}$ is a $(p-1) \times p$ contrast matrix, as before, then $\mathbf{C}\overline{\mathbf{y}}_{1i}$ and $\mathbf{C}\overline{\mathbf{y}}_{2i}$ are distributed as $MVN_{p-1}\left(\mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)$ and $MVN_{p-1}\left(\mathbf{C}\boldsymbol{\mu}_2, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)$, respectively.

Under $H_0 : \mathbf{C}\boldsymbol{\mu}_1 - \mathbf{C}\boldsymbol{\mu}_2 \;=\; \mathbf{0}$, the random vector $\mathbf{C}\overline{\mathbf{y}}_1 - \mathbf{C}\overline{\mathbf{y}}_2$ is $MVN_{p-1}\left[\mathbf{0}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]$, and

$$\begin{aligned} T^2 &= \left(\frac{n_1 n_2}{n_1 + n_2}\right)(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)'\,\mathbf{C}'\left[\mathbf{C}\mathbf{S}_{pl}\mathbf{C}'\right]^{-1}\mathbf{C}\,(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2) \\ &\sim T^2_{p-1, n_1+n_2-2} \end{aligned}$$

If we determine that there is a lack of parallelism we can examine the discriminant function coefficient vector to determine which of the segments in the profiles differ the most and hence contributes the most to the lack of parallelism. This discriminant function or form of discriminant function we will use again in constructing procedures for discriminating between groups.

$$\mathbf{a} \;=\; \left(\mathbf{C}\mathbf{S}_{pl}\mathbf{C}'\right)^{-1}\mathbf{C}\left(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2\right)$$

This discriminant function is an indication of which slope differences contributed most to rejection of H_{01} in the presence of the other components of $(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)$. There should be less need in this case to standardize the components of a, as suggested in Section 5.5, because the variables are assumed to be commensurate. The vector a is $(p-1) \times 1$, corresponding to the $(p-1)$ segments of the profile. Thus if the second component of **a**, for example, is largest in absolute value, the divergence in slopes between the two profiles on the second segment contributes most to rejection of H_{01}.

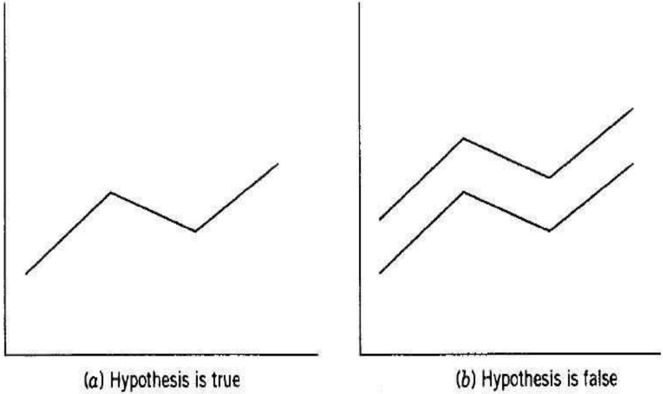
To test for equal levels -Two profiles

The second hypothesis of interest in comparing two profiles is, Are the two populations or groups at the same level? This hypothesis corresponds to a group (population) main effect in the ANOVA analogy. We can express this hypothesis in terms of the average level of group 1 compared to the average level of group 2:

$$H_{02} : \frac{\mu_{11} + \mu_{12} + \mu_{13} + \cdots + \mu_{1p}}{p} \quad = \quad \frac{\mu_{21} + \mu_{22} + \mu_{23} + \cdots + \mu_{2p}}{p}$$

To test H_{02}: $\mathbf{j}'\boldsymbol{\mu}_1 = \mathbf{j}'\boldsymbol{\mu}_2$ or H_{02}: $\mathbf{j}'\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) = \mathbf{0}$,

$$t \quad = \quad \frac{\mathbf{j}'\left(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2\right)}{\sqrt{\left(\mathbf{j}'\mathbf{S}_{pl}\mathbf{j}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$



(*a*) Hypothesis is true

(*b*) Hypothesis is false

Figure 5.5. Hypothesis H_{02} of equal group effect, assuming parallelism.

The hypothesis H_{02} can be true when H_{01} does not hold. Thus the average level of population 1 can equal the average level of population 2 without the two profiles being parallel, as illustrated in Figure 5.6.

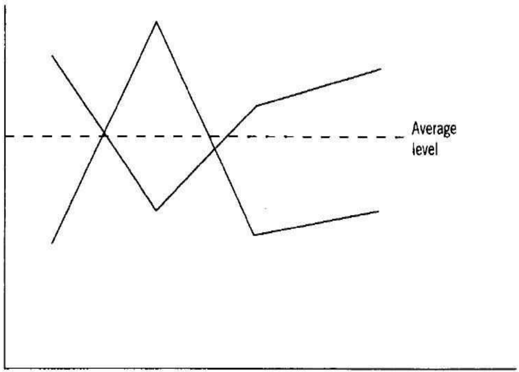


Figure 5.6. Hypothesis H_{02} of equal group effect without parallelism.

The third hypothesis of interest, corresponding to the test (or variable) main effect, is, Are the profiles flat? Assuming parallelism (assuming H_{01} is true), the "flatness" hypothesis can be pictured as in Figure 5.7. If H_{01} is not true, the test could be carried out separately for each group using the test in Section 5.9.1. If H_{02} is true, the two profiles in Figure 5.7a and Figure 5.7b will be coincident. To express the third hypothesis in a form suitable for testing, we note from Figure 5.7a that the average of the two group means is the same for each test:

To test for Flatness-Two profiles

$$H_{03} \quad = \quad \frac{1}{2}(\mu_{11}+\mu_{21}) \;=\; \frac{1}{2}(\mu_{12}+\mu_{22}) = \cdots = \frac{1}{2}(\mu_{1p}+\mu_{2p})$$

H_{03}: $\frac{1}{2}\mathbf{C}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \mathbf{0}$, or H_{03}: $\mathbf{C}\boldsymbol{\mu}_1 = \mathbf{0}$ and $\mathbf{C}\boldsymbol{\mu}_2 = \mathbf{0}$ To estimate $\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$, the best estimate of the overall average is the weighted average of the two group means

$$\overline{\mathbf{y}} \quad = \quad \frac{n_1\overline{\mathbf{y}}_1 + n_2\overline{\mathbf{y}}_2}{n_1+n_2}$$

$$T^2 \quad = \quad (n_1+n_2)(\mathbf{C}\overline{\mathbf{y}})'\left[\mathbf{C}\mathbf{S}_{pl}\mathbf{C}'\right]^{-1}(\mathbf{C}\overline{\mathbf{y}})$$

Psych Data-plotting the profiles

```
plot(1:4,ybar1,main="Profiles for Psych Data",
  xlab="Variables", ylab="Scores",
  xlim=c(0, 5), ylim=c(0, 40),"l")
points(1:4,ybar2,"l")
```

Psych Data-Testing for parallelism

```
>const<-(n1*n2)/(n1+n2)
>T2parallel
<-const*t(dbar)%*%t(C)%*%solve(C%*%v%*%t(C))%*%C%*%dbar
         [,1]
[1,] 74.24037
```

$\sim \; T^2_{\alpha,p-1,n_1+n_2-2}$ This result should have been obvious from the plotted profiles. In Figure 5.8 the lack of parallelism is most notable in the second and third segments.

Psych Data-Construct the discriminant

To see which of these made the greatest statistical contribution, we can examine the discriminant function coefficient vector given in (5.35) as

```
> DisParallel<-solve(C%*%v%*%t(C))%*%C%*%dbar
> DisParallel
           [,1]
[1,]  0.1356086
[2,] -0.1043403
[3,]  0.3631646
>
```

Thus the third segment contributed most to rejection in the presence of the other two segments.

Psych Data-Test for level profiles

```
> jprime<-rep(1, 4)
> tlevel
 <-jprime%*%dbar/sqrt(t(jprime)%*%v%*%jprime*(1/n1+1/n2))
          [,1]
[1,] -5.295698
```

$\sim \quad t_{\alpha, n_1+n_2-2}$

To test for Flatness-Two profiles

$$H_{03} \;=\; \frac{1}{2}(\mu_{11}+\mu_{21}) \;=\; \frac{1}{2}(\mu_{12}+\mu_{22}) = \cdots = \frac{1}{2}(\mu_{1p}+\mu_{2p})$$

H_{03}: $\frac{1}{2}\mathbf{C}(\boldsymbol{\mu}_1 \;+\; \boldsymbol{\mu}_2) = \mathbf{0}$, or H_{03}: $\mathbf{C}\boldsymbol{\mu}_1 = \mathbf{0}$ and $\mathbf{C}\boldsymbol{\mu}_2 = \mathbf{0}$

Psych Data-Test for flatness

```
C<-matrix(c(1,0,0,-1,1,0,0,-1,1,0,0,-1),3,4)
ybarprime<-(n1*ybar1+n2*ybar2)/(n1+n2)
T2flatness
 <-(n1+n2)*t(C%*%ybarprime)%*%solve(C%*%v%*%t(C))
              %*%(C%*%ybarprime)
> T2flatness
         [,1]
[1,] 254.0038
```

$\sim \ T^2_{\alpha,p-1,n_1+n_2-2}$. The value exceeds $T^2_{.01,3,62} = 12.796$. So reject flatness.

However, since the parallelism hypothesis was rejected, a more appropriate approach would be to test each of the two groups separately for flatness using the test of Section 5.9.1.

$$\begin{aligned}
T^2 &= (n_1)\left(\mathbf{C}\overline{\mathbf{y}}_1\right)'\left[\mathbf{C}\mathbf{S}_1\mathbf{C}'\right]^{-1}\left(\mathbf{C}\overline{\mathbf{y}}_1\right) \\
&= 221.126 \\
&\sim T^2_{\alpha,p-1,n_1-1} \\
T^2 &= (n_2)\left(\mathbf{C}\overline{\mathbf{y}}_2\right)'\left[\mathbf{C}\mathbf{S}_2\mathbf{C}'\right]^{-1}\left(\mathbf{C}\overline{\mathbf{y}}_2\right) \\
&= 103.483 \\
&\sim T^2_{\alpha,p-1,n_2-1}
\end{aligned}$$

and we have significant lack of flatness.