Multivariate Statistical Analysis

Fall 2011

C. L. Williams, Ph.D.

Lecture 17 for Applied Multivariate Analysis

Outline

The hypotheses:

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots \boldsymbol{\mu}_k$$

$$H_A: \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \quad \{i \neq j\} \quad \leq k$$

H-Hypotheses and E- Error Matrices

Sums of Squares and Cross Products Matrices SSCPs

$$\mathbf{H} = n\sum_{i=1}^{k}\left(\bar{\mathbf{y}}_{i\bullet}-\bar{\mathbf{y}}_{\bullet\bullet}\right)\left(\bar{\mathbf{y}}_{i\bullet}-\bar{\mathbf{y}}_{\bullet\bullet}\right)'$$

$$\mathbf{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{i\bullet}\right)\left(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{i\bullet}\right)'$$

$$\mathbf{H} = \begin{pmatrix} SSH_{11} & SPH_{12} & \cdots & SPH_{1p} \\ SPH_{12} & SSH_{22} & \cdots & SPH_{2p} \\ \vdots & \vdots & & \vdots \\ SPH_{1p} & SPH_{2p} & \cdots & SSH_{pp} \end{pmatrix}$$

$$SSH_{22} = n\sum_{i=1}^{k}\left(\overline{y}_{i\bullet 2} - \overline{y}_{\bullet\bullet 2}\right)^2 = \sum_{i=1}^{k}\frac{y_{i\bullet 2}}{n} - \frac{y_{\bullet\bullet 2}^2}{kn}$$

$$SPH_{12} = n\sum_{i=1}^{k}\left(\overline{y}_{i\bullet 1} - \overline{y}_{\bullet\bullet 1}\right)\left(\overline{y}_{i\bullet 2} - \overline{y}_{\bullet\bullet 2}\right) = \sum_{i=1}^{k}\frac{y_{i\bullet 1}y_{i\bullet 2}}{n} - \frac{y_{\bullet\bullet 1}y_{\bullet\bullet 2}}{kn}$$

$$\mathbf{E} = \begin{pmatrix} SSE_{11} & SPE_{12} & \cdots & SPE_{1p} \\ SPE_{12} & SSE_{22} & \cdots & SPE_{2p} \\ \vdots & \vdots & & \vdots \\ SPE_{1p} & SPE_{2p} & \cdots & SSE_{pp} \end{pmatrix}$$

$$SSE_{22} = n\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij2} - \overline{y}_{i\bullet 2}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij2} - \frac{y_{i\bullet 2}^2}{n}$$

$$SPE_{12} = n\sum_{i=1}^{k}\sum_{j=1}^{n}\left(\overline{y}_{ij1} - \overline{y}_{i\bullet 1}\right)\left(\overline{y}_{ij2} - \overline{y}_{i\bullet 2}\right)$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij1}y_{ij2} - \frac{y_{i\bullet 1}y_{i\bullet 2}}{n}$$

Wilk's Lambda

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|}$$

is a test of how significant the error variation $\mathbf{E}$ is relative to the total variation $\mathbf{E} + \mathbf{H}$. Wilk's Lambda considers the ratio of two determinants.

One measure of the "size" difference would simply be to compare the determinants.
Bearing in mind that $\mathbf{SST}_{Total} = \mathbf{SSH}_{Between} + \mathbf{SSE}_{Within}$, Wilk's Λ *can* be defined as:

$$\Lambda = \frac{|\mathbf{SSE}_{Within}|}{|\mathbf{SST}_{Total}|}$$

Roy's Test

To test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k$ based on λ_1 (the largest eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$), we use Roy's union-intersection test, also called Roy's *largest root test*. The test statistic is given by

$$\theta = \frac{\lambda_1}{1+\lambda_1}$$

Critical values for θ are given in Table A.10 We reject $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k$ if $\theta > \theta_{\alpha,s,m,N}$. The parameters s, m, and N are defined as $s = min(\nu_H, p)$, $m = \frac{1}{2}(|\nu_H - p| - 1)$, $N = \frac{1}{2}(\nu_E - p - 1)$.

Pillai and Lawley-Hotelling Tests

To test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k$ the *Pillai statistic* based on the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ is given by

$$V^{(s)} = tr\left[(\mathbf{E}+\mathbf{H})^{-1}\mathbf{H}\right]$$

We reject $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k$ for values of $V^{(s)} > V_{\alpha}^{(s)}$ given in Table A.11.

Lawley-Hotelling statistic

Hotelling's generalized T^2-statistic

$$U^{(s)} = tr\left[\mathbf{E}^{-1}\mathbf{H}\right] = \sum_{i=1}^{s} \lambda_i$$

Table A.12 gives upper percentage points of the test statistic:

$$\frac{\nu_E}{\nu_H} U^{(s)}$$

for which we reject $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k$ for large values.

Eigenvalues

In term of the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ we can discuss the relationship between the four tests.

$$V^{(s)} = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i} : \text{ Pillai}$$

$$U^{(s)} = \sum_{i=1}^{s} \lambda_i : \text{ Lawley-Hotelling}$$

$$\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i} : \text{ Wilk's } \Lambda$$

$$\theta = \frac{\lambda_1}{1+\lambda_1} : \text{ Roy's largest root}$$

```
"rootstockdata"<-structure(.Data = list(
rootstock=as.factor(c(1,1,1,1,1,1,1,1,2,2,
2,2,2,2,2,2,3,3,3,3,3,3,3,3,4,4,
4,4,4,4,4,4,5,5,5,5,5,5,5,5,6,6,6,6,6,6,6,6)),
girth4=c(1.11,1.19,1.09...),
extension4=c(2.569,2.928,2.865,...),
girth15=c(3.58,3.75,3.93,...),
weight=c(0.76,0.821,0.928,...)),
class = "data.frame", row.names
     = c("1", "2", "3", "4", "5",...))
```

```
> summary(rootstock.fit,test='Roys')
Error in match.arg(test) :
'arg' should be one of "Pillai","Wilks","Hotelling-Lawley"
> summary(rootstock.fit,test='Roy')
          Df    Roy approx F num Df den Df    Pr(>F)
rootstock  5 1.8757   15.756      5     42 1.002e-08 ***
Residuals 42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
> summary(rootstock.fit,test='Hotelling-Lawley')
          Df Hotelling-Lawley approx F num Df den Df    Pr(
rootstock  5          2.9214   5.4776     20    150 2.568e
Residuals 42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
>
```

Wilk's lambda longhand in R

Could be calculated in R as:

```
Between <- summary(fit)$SS$rootstock
Total <- summary(fit)$SS$Residuals +
summary(fit)$SS$rootstock
Wilks <- det(Between) / det(Total)
```

But this isn't the only (or the best) way of getting it!

In performing a MANOVA, we wish to compare the p-dimensional confidence ellipses under H_0 (the means are not different) and H_1

We could ask three questions, for groups $g = 1, \ldots 3$:

- Are the profiles parallel:
  $H_{01} : \mu_{1j} - \mu_{1j-1} = \mu_{2j} - \mu_{2j-1} = \mu_{3j} - \mu_{3j-1}$; for $j = 2, \ldots p$ and
- if so, are the profiles coincident: $H_{02} : \mu_{1j} = \mu_{2j} = \mu_{3j}$ for $j = 1, \ldots, p$ and finally
- if so are the profiles level: $H_{03} : \mu_{11} = \mu_{12} = \ldots = \mu_{1p} = \mu_{21} = \mu_{22} = \ldots = \mu_{2p} = \ldots = \mu_{3p}$ for $j = 2, \ldots p$

Measures of Multivariate Association

$$R^2 = \frac{regressionsumofsquares}{totalsumofsquares}$$

In (one-way) MANOVA, we need to measure the strength of the association between several dependent variables and several independent (grouping) variables. Various measurements of multivariate association have been proposed. Wilks (1932) suggested a "generalized η^2":

$$\eta^2 = 1 - \Lambda$$

Λ is small if the spread in the means is large, then obviously η^2 is large when there is large disparity in the means.

Choice of Statistic

When H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots \boldsymbol{\mu}_k$ is true, all the mean vectors are at the same point. Therefore, all four MANOVA test statistics have the same Type I error rate, α, as noted in Section 6.1.7; that is, all have the same probability of rejection when H0 is true. However, when H_0 is false, the four tests have different probabilities of rejection. We noted in Section 6.1.7 that in a given sample the four tests need not agree, even if H_0 is true. One test could reject H_0 and the others don't reject H_0, for example. Wilks'Λ has played the dominant role in significance tests in MANOVA because it was the first to be derived and has well-known χ^2 and *F*-approximations. It can also be partitioned in certain ways we will find useful later. However, it is not always the most powerful among the four tests. The probability of rejecting H_0 when it is false is known as the power of the test.

We have four tests, not one of which is uniformly most powerful. The relative powers of the four test statistics depend on the configuration of the mean vectors $\boldsymbol{\mu}_1,\ \boldsymbol{\mu}_2,\ \ldots\ \boldsymbol{\mu}_k$ in the s-dimensional space. A given test will be more powerful for one configuration of mean vectors than another. An indication of the pattern of the mean vectors is given by the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$. If there is one large eigenvalue and the others are small, the mean vectors lie close to a line. If there are two large eigenvalues, the mean vectors lie mostly in two dimensions, and so on. Because Roy's test uses only the largest eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$, it is more powerful than the others if the mean vectors are collinear. The other three tests have greater power than Roy's when the mean vectors are diffuse (spread out in several dimensions).

In conclusion, the use of Roy's θ is not recommended in any situation except the collinear case under standard assumptions. In the diffuse case its performance is inferior to that of the other three, both when the assumptions hold and when they do not. If the data come from nonnormal populations exhibiting skewness or positive kurtosis, any of the other three tests perform acceptably well. Among these three, $V^{(s)}$ is superior to the other two when there is heterogeneity of covariance matrices. Indeed $V^{(s)}$ is first in all rankings except those for the collinear case.

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

- Parallel: Test $\mathbf{C}\boldsymbol{\mu}_1 = \mathbf{C}\boldsymbol{\mu}_2 = \mathbf{C}\boldsymbol{\mu}_3$ (note this is rank $p-1$)
- Coincident: Test $\mathbf{1}\boldsymbol{\mu}_1 = \mathbf{1}\boldsymbol{\mu}_2 = \mathbf{1}\boldsymbol{\mu}_3$ (note this is univariate)
- Level: Test $\mathbf{C}\boldsymbol{\mu} = 0$ (note this is rank $p-1$)

(also note all need to use pooled estimates of means and covariances where appropriate)

Multivariate Contrasts

A multivariate contrast across all group means is defined:

$$\boldsymbol{\delta} = c_1\boldsymbol{\mu}_1 + c_2\boldsymbol{\mu}_2, \cdots + c_k\boldsymbol{\mu}_k$$

where we know that $\sum_{i=1}^{k} c_i = 0$. We also know that an unbiased estimator for this contrast is

$$\widehat{\boldsymbol{\delta}} = c_1\overline{\mathbf{y}}_1 + c_2\overline{\mathbf{y}}_2, \cdots + c_k\overline{\mathbf{y}}_k$$

Variance-covariance of contrasts

$$cov\left(\boldsymbol{\delta}\right) = \frac{\boldsymbol{\Sigma}}{n}\sum_{i=1}^{k} c_i^2$$

so that

$$\widehat{cov\left(\boldsymbol{\delta}\right)} = \frac{\mathbf{S}_{pl}}{\nu_E}\left(\frac{\sum_{i=1}^{k} c_i^2}{n}\right)$$

where $\mathbf{S}_{pl}$ is and unbiased estimate of $\frac{\boldsymbol{\Sigma}_{pl}}{\nu_E}$.

$\delta = \mathbf{0}$

or H_0: $c_1\boldsymbol{\mu}_1 + c_2\boldsymbol{\mu}_2 + \cdots c_k\boldsymbol{\mu}_k = \mathbf{0}$ is used to compare population mean vectors.

$$\begin{aligned} \boldsymbol{\mu}_1 - 2\boldsymbol{\mu}_2 + \boldsymbol{\mu}_3 &= \mathbf{0} \text{is equivalent to} \\ \boldsymbol{\mu}_2 &= \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_3) \end{aligned}$$

"says" that the mean for group 2 is equal to the average of groups 1 and 3.

$$\begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(\mu_{11} + \mu_{31}) \\ \frac{1}{2}(\mu_{12} + \mu_{32}) \\ \vdots \\ \frac{1}{2}(\mu_{1p} + \mu_{3p}) \end{pmatrix}$$

Under the appropriate multivariate normality assumptions $\boldsymbol{\delta} = \mathbf{0}$ or H_0: $c_1\boldsymbol{\mu}_1 + c_2\boldsymbol{\mu}_2 + \cdots c_k\boldsymbol{\mu}_k = \mathbf{0}$ can be tested with a one-sample T^2 test.

$$
\begin{aligned}
T^2 &= \widehat{\boldsymbol{\delta}}' \left(\frac{\mathbf{S}_{pl}}{n} \sum_{i=1}^{k} c_i^2 \right)^{-1} \widehat{\boldsymbol{\delta}} \\
&= \frac{n}{\sum_{i=1}^{k} c_i^2} \left(\sum_{i=1}^{k} c_i \mathbf{y}_{i.} \right)' \left(\frac{\mathbf{E}}{\nu_E} \right)^{-1} \left(\sum_{i=1}^{k} c_i \mathbf{y}_{i.} \right)
\end{aligned}
$$

Or equivalently,

$$\mathbf{H}_1 = \frac{n}{\sum_{i=1}^{k} c_i^2}\left(\sum_{i=1}^{k} c_i \mathbf{y}_{i.}\right)\frac{n}{\sum_{i=1}^{k} c_i^2}\left(\sum_{i=1}^{k} c_i \mathbf{y}_{i.}\right)'$$

so that

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}_1|}$$

Rootstock contrasts

Consider two orthogonal contrasts. First

$$2 \text{ -1 -1 -1 -1 } 2$$

which compares the means from groups 1 and 6, μ_1 and μ_6, with the other four group mean vectors, groups 2,3,4 and 5 $(\mu_2, \mu_3, \mu_4, \mu_5)$.

$$H_{01} : 2\mu_1 + 2\mu_6 = \mu_2 + \mu_3 + \mu_4 + \mu_5$$

or in terms of averages

$$H_{01} : \frac{1}{2}(\mu_1 + \mu_6) = \frac{1}{4}(\mu_2 + \mu_3 + \mu_4 + \mu_5)$$

Mean vectors for Rootstock Data

```
> apply(as.matrix(rootstockdata[,2:5]),2,mean)
    girth4 extension4    girth15     weight
  1.102708   2.758854   4.087292   1.083000
> apply(as.matrix(rootstockdata[1:8,2:5]),2,mean)
    girth4 extension4    girth15     weight
  1.137500   2.977125   3.738750   0.871125
```

Mean vectors for Rootstock Data

```
> apply(as.matrix(rootstockdata[9:16,2:5]),2,mean)
    girth4 extension4    girth15     weight
  1.157500   3.109125   4.515000   1.280500
> apply(as.matrix(rootstockdata[17:24,2:5]),2,mean)
    girth4 extension4    girth15     weight
  1.107500   2.815250   4.455000   1.391375
```

```
> apply(as.matrix(rootstockdata[25:32,2:5]),2,mean)
   girth4 extension4    girth15     weight
  1.09750    2.87975    3.90625    1.03900
> apply(as.matrix(rootstockdata[33:40,2:5]),2,mean)
   girth4 extension4    girth15     weight
  1.08000    2.55725    4.31250    1.18100
> apply(as.matrix(rootstockdata[41:48,2:5]),2,mean)
   girth4 extension4    girth15     weight
1.036250   2.214625   3.596250   0.735000
```

```
  contrast1<-as.matrix(c(2,-1,-1,-1,-1,2))
#
y1<-tapply(rootstockdata[,2],rootstock,mean)
y1contrast1<-t(contrast1)%*%y1
#
y2<-tapply(rootstockdata[,3],rootstock,mean)
y2contrast1<-t(contrast1)%*%y2
#
y3<-tapply(rootstockdata[,4],rootstock,mean)
y3contrast1<-t(contrast1)%*%y3
#
y4<-tapply(rootstockdata[,5],rootstock,mean)
y4contrast1<-t(contrast1)%*%y4
```

```
#
prod1<-rbind(y1contrast1,y2contrast1,y3contrast1,
             y4contrast1)
n<-8
constant<-(n/t(contrast1)%*%contrast1)
H<-constant[1,1]*prod1%*%t(prod1)
E<-summary(rootstock.fit)$SS$Residuals
wilks<-det(E)/det(H+E)
```

```
  contrast2<-as.matrix(c(1,0,0,0,0,-1))

y1<-tapply(rootstockdata[,2],rootstock,mean)
y1contrast2<-t(contrast2)%*%y1
#
y2<-tapply(rootstockdata[,3],rootstock,mean)
y2contrast2<-t(contrast2)%*%y2
#
y3<-tapply(rootstockdata[,4],rootstock,mean)
y3contrast2<-t(contrast2)%*%y3
#
y4<-tapply(rootstockdata[,5],rootstock,mean)
y4contrast2<-t(contrast2)%*%y4
```

```
#
prod1<-rbind(y1contrast2,y2contrast2,y3contrast2,
             y4contrast2)
n<-8
constant<-(n/t(contras2)%*%contrast2)
H<-constant[1,1]*prod1%*%t(prod1)
E<-summary(rootstock.fit)$SS$Residuals
wilks<-det(E)/det(H+E)
```