

Lecture 3: Comparisons between several multivariate means

Key concepts:

1. Paired comparison & repeated measures
2. Comparing means of two populations
3. Comparing means of several populations: One-way multivariate analysis of variance
4. Testing for equality of covariance matrices
5. Two-way multivariate analysis of variance
6. Profile analysis
7. Growth curves

Key assumption: Normality or large sample sizes.

1 Paired comparison

A procedure to eliminate the influence of extraneous unit-to-unit variation. Measures are taken on the same or identical units for different treatments.

Recall the univariate paired t-test. Let X_{j1} and X_{j2} be the response of unit j to treatment 1 and 2, respectively, where $j = 1, \dots, n$. Let $D_j = X_{j1} - X_{j2}$ be the difference between the treatments. This is so because both responses are from the same or identical unit.

Assume $D_j \sim N(\delta, \sigma_d^2)$. Consider the testing problem with $H_o : \delta = 0$ versus $H_a : \delta \neq 0$. The paired t-test is

$$t = \frac{\bar{D} - \delta}{s_d/\sqrt{n}},$$

where $\bar{D} = \frac{1}{n} \sum_{j=1}^n D_j$ and $s_d^2 = \frac{1}{n-1} \sum_{j=1}^n (D_j - \bar{D})^2$. One rejects H_o if and only if $|t| \geq t_{n-1}(\alpha/2)$. The corresponding $100(1 - \alpha)\%$ confidence interval for the mean difference $\delta = E(X_{j1} - X_{j2})$ is

$$\bar{D} - t_{n-1}(\alpha/2) \frac{s_d}{\sqrt{n}} \leq \delta \leq \bar{D} + t_{n-1}(\alpha/2) \frac{s_d}{\sqrt{n}}.$$

Generalization: Suppose that p measurements are taken from each unit so that the responses are X_{1ji} and X_{2ji} , where X_{1ji} is the measure of the i th variable of j unit for treatment

1, and X_{2ji} is the measure of the i th variable of j unit for treatment 2. The difference is then $D_{ji} = X_{1ji} - X_{2ji}$ and $\mathbf{D}_j = (D_{j1}, \dots, D_{jp})'$.

Assume that $E(\mathbf{D}_j) = \boldsymbol{\delta} = (\delta_1, \dots, \delta_p)'$ and $\text{cov}(\mathbf{D}_j) = \boldsymbol{\Sigma}_d$. If we further assume $\mathbf{D}_j \sim N_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_d)$, then we can make inference about $\boldsymbol{\delta}$ using the Hotelling's T^2 statistic

$$T^2 = n(\bar{\mathbf{D}} - \boldsymbol{\delta})' \mathbf{S}_d^{-1} (\bar{\mathbf{D}} - \boldsymbol{\delta}),$$

where

$$\bar{\mathbf{D}} = \frac{1}{n} \sum_{j=1}^n \mathbf{D}_j \quad \text{and} \quad \mathbf{S}_d^2 = \frac{1}{n-1} \sum_{j=1}^n (\mathbf{D}_j - \bar{\mathbf{D}})(\mathbf{D}_j - \bar{\mathbf{D}})$$

Result 6.1. Let the differences $\mathbf{D}_1, \dots, \mathbf{D}_n$ be a random sample from an $N_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_d)$ population. Then,

$$T^2 = n(\bar{\mathbf{D}} - \boldsymbol{\delta})' \mathbf{S}_d^{-1} (\bar{\mathbf{D}} - \boldsymbol{\delta})$$

is distributed as an $[(n-1)p/(n-p)]F_{p,n-p}$ random variable. If n and $n-p$ are both large, T^2 is approximately distributed as χ_p^2 random variable.

Inference: Suppose the random sample consists of $\mathbf{d}_1, \dots, \mathbf{d}_n$ and the population is $N_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_d)$. Then, reject $H_o : \boldsymbol{\delta} = \mathbf{0}$ in favor of $H_a : \boldsymbol{\delta} \neq \mathbf{0}$ if

$$T^2 = n\bar{\mathbf{d}}' \mathbf{S}_d^{-1} \bar{\mathbf{d}} > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha).$$

A $100(1-\alpha)\%$ confidence region for $\boldsymbol{\delta}$ is

$$(\bar{\mathbf{d}} - \boldsymbol{\delta})' \mathbf{S}_d^{-1} (\bar{\mathbf{d}} - \boldsymbol{\delta}) \leq \frac{(n-1)p}{n(n-p)} F_{p,n-p}(\alpha).$$

A $100(1-\alpha)\%$ simultaneous confidence intervals for the individual mean differences δ_i are

$$\bar{d}_i \pm \sqrt{\frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{d_i}^2}{n}},$$

where \bar{d}_i is the i th element of $\bar{\mathbf{d}}$ and $s_{d_i}^2$ is the (i, i) th element of \mathbf{S}_d . The Bonferroni $100(1-\alpha)\%$ simultaneous confidence intervals for the individual mean differences are

$$\bar{d}_i \pm t_{n-1}(\alpha/(2p)) \sqrt{\frac{s_{d_i}^2}{n}}.$$

Finally, if n and $n-p$ are sufficiently large, then the normality assumption can be dropped, and the simultaneous C.I.s can be obtained by replacing $[(n-1)p/(n-p)]F_{p,n-p}(\alpha)$ by $\chi_p^2(\alpha)$.

Remark: The R programs of Chapter 5 can be used to perform paired comparison.

Example: Consider the data on Table 6.1 of the text.

```

> x=read.table("T6-1.DAT")

> d1=x[,1]-x[,3]
> d2=x[,2]-x[,4]
> d=cbind(d1,d2)

> source('Hotelling.txt')
> hotelling(d,rep(0,2))
           [,1]
Hotelling-T2 13.63931214
p.value      0.02082779

> source("r-cregion.txt")
> confreg(d)
[1] "C.R. based on T^2"
           [,1]      [,2]
[1,] -22.453272  3.726000
[2,]  -5.700119 32.245574
[1] "CR based on individual t"
           [,1]      [,2]
[1,] -18.8467298  0.1194570
[2,]  -0.4725958 27.0180504
[1] "CR based on Bonferroni"
           [,1]      [,2]
[1,] -20.573107  1.845835
[2,]  -2.974903 29.520358
[1] "Asymp. simu. CR"
           [,1]      [,2]
[1,] -19.781395  1.054122
[2,]  -1.827351 28.372806

```

Constrast matrix.

Definition: A p -dimensional vector is called a *contrast vector* if its elements sum to zero. By definition, contrast vectors are orthogonal to the vector of ones. A $m \times k$ matrix is called a *contrast matrix* if all its rows are contrast vectors. For example, $\mathbf{c} = (1, 0, -1, 0)'$ is a contrast vector.

The above paired comparisons can be achieved by using contrast matrix. For example, consider the effluent data of Example 6.1. Instead of computing the differenced data, we can directly use the observations in Table 6.1. The observation for Sample 1 is $\mathbf{x}_1 = (6, 27, 25.15)'$. Construct the contrast matrix

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}. \quad (1)$$

Clearly, the differenced data are $\mathbf{d}_j = \mathbf{C}\mathbf{x}_j$ for $j = 1, \dots, n$. Furthermore, $\bar{\mathbf{d}} = \mathbf{C}\bar{\mathbf{x}}$ and $\mathbf{S}_d = \mathbf{C}\mathbf{S}\mathbf{C}'$, where \mathbf{S} is the sample covariance matrix of the data. The T^2 statistic then becomes

$$T^2 = n\bar{\mathbf{x}}'\mathbf{C}'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}\mathbf{C}'\bar{\mathbf{x}}.$$

Consequently, there is no need to calculate the differenced data \mathbf{d} .

This idea is particularly useful in analyzing *repeated measures* in which different treatments are applied the each unit once over successive periods of time. Suppose there are q treatments, then the observation for the j th unit is

$$\mathbf{X}_j = (X_{j1}, X_{j2}, \dots, X_{jq})', \quad j = 1, \dots, n.$$

Let $\boldsymbol{\mu} = \mathbf{X}$. To test the hypothesis that all treatments have the same effect is equivalent to test all elements of $\boldsymbol{\mu}$ are equal. To this end, we can construct the contrast matrix \mathbf{C}_1 as

$$\begin{bmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_q \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} \equiv \mathbf{C}_1\boldsymbol{\mu},$$

or \mathbf{C}_2 as

$$\begin{bmatrix} \mu_1 - \mu_2 \\ \mu_2 - \mu_3 \\ \vdots \\ \mu_{q-1} - \mu_q \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} \equiv \mathbf{C}_2\boldsymbol{\mu}.$$

The problem then is to test $\mathbf{C}_1\boldsymbol{\mu} = \mathbf{0}$ or $\mathbf{C}_2\boldsymbol{\mu} = \mathbf{0}$. [Other contrast matrices are available.] This results in using the T^2 test statistic as

$$T^2 = n(\mathbf{C}\bar{\mathbf{X}})'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}\mathbf{C}\bar{\mathbf{X}}.$$

Remark. The T^2 statistic does not depend on the choice of contrast matrix \mathbf{C} . This is because $\text{rank}(\mathbf{C}_1) = \text{rank}(\mathbf{C}_2) = q - 1$ and each row of \mathbf{C}_i are orthogonal to the vector of ones. Consequently, the rows of \mathbf{C}_1 and the rows of \mathbf{C}_2 span the $(q - 1)$ -dimensional subspace that is orthogonal to $\mathbf{1}_q$. Thus, there exists a non-singular matrix $\mathbf{B}_{(q-1) \times (q-1)}$ such that $\mathbf{C}_1 = \mathbf{B}\mathbf{C}_2$. For instance, for the \mathbf{C}_1 and \mathbf{C}_2 matrices given above, we have

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 1 & 1 & \cdots & 1 \end{bmatrix}.$$

It is then easy to show that \mathbf{C}_1 and \mathbf{C}_2 give the same T^2 statistic.

Based on the prior discussion, we can test the equality of treatments in a repeated measures case by using the result below.

Consider an $N_p(\boldsymbol{\mu}\boldsymbol{\Sigma})$ population. Let \mathbf{C} be a contrast matrix. An α -level test of $H_o : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$ versus $H_a : \mathbf{C}\boldsymbol{\mu} \neq \mathbf{0}$ is to reject H_o if

$$T^2 = n(\mathbf{C}\bar{\mathbf{x}})'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}\mathbf{C}\bar{\mathbf{x}} > \frac{(n-1)(q-1)}{n-q+1}F_{q-1, n-q+1}(\alpha),$$

where $\bar{\mathbf{x}}$ and \mathbf{S} are the sample mean and covariance matrix. A confidence region for the contrasts $\mathbf{C}\boldsymbol{\mu}$ is

$$n(\mathbf{C}\bar{\mathbf{x}} - \mathbf{C}\boldsymbol{\mu})'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}(\mathbf{C}\bar{\mathbf{x}} - \mathbf{C}\boldsymbol{\mu}) \leq \frac{(n-1)(q-1)}{n-q+1}F_{q-1, n-q+1}.$$

Consequently, simultaneous $100(1-\alpha)\%$ confidence intervals for the single contrasts $\mathbf{c}'\mathbf{C}$ for any contrast vectors of interest are

$$\mathbf{c}'\bar{\mathbf{x}} \pm \sqrt{\frac{(n-1)(q-1)}{n-q+1}F_{q-1, n-q+1}(\alpha)}\sqrt{\frac{\mathbf{c}'\mathbf{S}\mathbf{c}}{n}}.$$

Example. Consider the Sleeping-dog data in Table 6.2. There are 19 observations and four treatments. To analyze the data, a R program called **r-contrast.txt** is developed. The analysis is given below:

```
> x=read.table("T6-2.DAT")
> dim(x)
[1] 19 4
> x
      V1  V2  V3  V4
1  426 609 556 600
2  253 236 392 395
3  359 433 349 357
4  432 431 522 600
5  405 426 513 513
6  324 438 507 539
7  310 312 410 456
8  326 326 350 504
9  375 447 547 548
10 286 286 403 422
11 349 382 473 497
12 429 410 488 547
13 348 377 447 514
14 412 473 472 446
```

```

15 347 326 455 468
16 434 458 637 524
17 364 367 432 469
18 420 395 508 531
19 397 556 645 625
> source("r-contrast.txt")
> cmtx=matrix(c(-1,1,1,-1,-1,-1,1,1,-1,1,-1,1),3,4)
> cmtx
      [,1] [,2] [,3] [,4]
[1,]  -1  -1   1   1
[2,]   1  -1   1  -1
[3,]   1  -1  -1   1
> contrast(x,cmtx)
[1] "Hotelling Tsq statistics & p-value"
[1] 1.160163e+02 3.317767e-07
[1] "Simultaneous C.I. for each contrast"
      [,1]      [,2]
[1,] 135.65030 282.98128
[2,] -114.72708 -5.37818
[3,] -78.72858  53.14964

```

2 Comparing mean vectors of two populations

The setup

1. $\mathbf{X}_{11}, \mathbf{X}_{12}, \dots, \mathbf{X}_{1,n_1}$ are p -dimensional random sample of size n_1 from a population with mean $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$.
2. $\mathbf{X}_{21}, \mathbf{X}_{22}, \dots, \mathbf{X}_{2,n_2}$ are p -dimensional random sample of size n_2 from a population with mean $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$.
3. The two random samples are independent.

If n_1 and n_2 are small, some additional assumptions are needed. They are

1. both populations are normal,
2. $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$.

Problem of interest: $H_o : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_o$ versus $H_a : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\delta}_o$.

Denote the sample mean and covariance of the random samples by $\bar{\mathbf{x}}_1$ and \mathbf{S}_1 and $\bar{\mathbf{x}}_2$ and \mathbf{S}_2 , respectively. Under the assumption that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, we can obtain a pooled estimate of the covariance matrix

$$\mathbf{S}_{pool} = \frac{n_1 - 1}{n_1 + n_2 - 2} \mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} \mathbf{S}_2.$$

This pooled estimate is consistent as $E(\mathbf{S}_{pool}) = \boldsymbol{\Sigma}$. Note that $E(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\text{Cov}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = \text{Cov}(\bar{\mathbf{X}}_1) + \text{Cov}(\bar{\mathbf{X}}_2) = \frac{1}{n_1}\boldsymbol{\Sigma} + \frac{1}{n_2}\boldsymbol{\Sigma}$, which can be estimated by $(\frac{1}{n_1} + \frac{1}{n_2})\mathbf{S}_{pool}$.

The following result holds.

Result 6.2. If $\mathbf{X}_{11}, \mathbf{X}_{12}, \dots, \mathbf{X}_{1,n_1}$ form a random sample from $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathbf{X}_{21}, \mathbf{X}_{22}, \dots, \mathbf{X}_{2,n_2}$ forms a random sample from $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ and the two random samples are independent, then

$$T^2 = [\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{pool} \right]^{-1} [\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]$$

is distributed as

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

Consequently,

$$Pr(T^2 \leq c^2) = 1 - \alpha,$$

where $c^2 = \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}(\alpha)$.

Proof: (1) $\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 \sim N_p(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, [(1/n_1) + (1/n_2)]\boldsymbol{\Sigma})$, (2) $(n_1 - 1)\mathbf{S}_1 \sim W_{n_1 - 1}(\boldsymbol{\Sigma})$ and $(n_2 - 1)\mathbf{S}_2 \sim W_{n_2 - 1}(\boldsymbol{\Sigma})$, and (3) $(n_1 - 1)\mathbf{S}_1$ and $(n_2 - 1)\mathbf{S}_2$ are independent so that $(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 \sim W_{n_1 + n_2 - 2}(\boldsymbol{\Sigma})$.

Result 6.3. Let $c^2 = \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}(\alpha)$. With probability $1 - \alpha$,

$$\mathbf{a}'(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \pm c \sqrt{\mathbf{a}' \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{pool} \mathbf{a}}$$

will cover $\mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for all \mathbf{a} . Thus, the simultaneous confidence intervals for $\mu_{1i} - \mu_{2i}$ is

$$(\bar{X}_{1i} - \bar{X}_{2i}) \pm c \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right) s_{ii, pool}}, \quad i = 1, \dots, p.$$

The Bonferroni $100(1 - \alpha)\%$ simultaneous C.I. for $\mu_{1i} - \mu_{2i}$ are

$$(\bar{x}_{1i} - \bar{x}_{2i}) \pm t_{n_1 + n_2 - 2}(\alpha/(2p)) \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right) s_{ii, pool}}.$$

Case: $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. In this case, there is no pooling in covariance matrix estimation and we typically require that $n_1 - p$ and $n_2 - p$ are sufficiently large. One can then replace $(\frac{1}{n_1} + \frac{1}{n_2})\mathbf{S}_{pool}$ by $\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2$ and F -distribution by χ_p^2 distribution.

When $n_1 = n_2 = n$, then

$$\begin{aligned}
 \frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2 &= \frac{1}{n}(\mathbf{S}_1 + \mathbf{S}_2) = \frac{2}{n}\left(\frac{1}{2}\mathbf{S}_1 + \frac{1}{2}\mathbf{S}_2\right) \\
 &= \left(\frac{1}{n} + \frac{1}{n}\right)\left(\frac{(n-1)\mathbf{S}_1}{(n-1) + (n-1)} + \frac{(n-1)\mathbf{S}_2}{(n-1) + (n-1)}\right) \\
 &= \left(\frac{1}{n} + \frac{1}{n}\right)\frac{(n-1)\mathbf{S}_1 + (n-1)\mathbf{S}_2}{n + n - 2} \\
 &= \left(\frac{1}{n} + \frac{1}{n}\right)\mathbf{S}_{pool}.
 \end{aligned}$$

Thus, where $n_1 = n_2$, the large sample procedure is essentially the same as the one using pooled covariance matrix. The impact of unequal covariance matrices is, therefore, least when the sample sizes are equal. The impact would be greater if either $n_1 \ll n_2$ or $n_2 \ll n_1$.

The Behrens-Fisher Problem.

Test $H_o : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ versus $H_a : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \mathbf{0}$, where the two populations are normally distributed, but have different covariance matrices, and the sample sizes are not large. [Of course, $n_1 > p$ and $n_2 > p$ are needed.]

The key issue is the distribution of

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

when $n_1 - p$ and $n_2 - p$ are small.

This problem has been widely studied in the literature; see, for instance, Krishnamoorthy and Yu (2004, *Statistics & Probability Letters*) and Nel and Van der Merwe (1986, *Communications in Statistics - Theory and Methods*). A recommended method is to approximate the distribution of T^2 as

$$T^2 = \frac{vp}{v - p + 1} F_{p, v-p+1},$$

where

$$v = \frac{p + p^2}{\sum_{i=1}^2 \frac{1}{n_i} \left\{ \text{tr} \left[\left(\frac{1}{n_i} \mathbf{S}_i \left(\frac{1}{n_1} \mathbf{S}_1 + \frac{1}{n_2} \mathbf{S}_2 \right)^{-1} \right)^2 \right] + \left(\text{tr} \left[\frac{1}{n_i} \mathbf{S}_i \left(\frac{1}{n_1} \mathbf{S}_1 + \frac{1}{n_2} \mathbf{S}_2 \right)^{-1} \right] \right)^2 \right\}},$$

where $\min(n_1, n_2) \leq v \leq n_1 + n_2$.

Remark: A R script **Behrens.txt** is available to perform the test. See course web. For illustration, consider the effluent data on Table 6.1. The paired comparison rejects the null hypothesis of equal means. The result of using Behrens-Fisher approach is given below.

```
> x=read.table("T6-1.DAT")
> dim(x)
```

```

[1] 11 4
> x1=x[,1:2]
> x2=x[,3:4]
> source("Behrens.txt")
> Behrens(x1,x2)
[1] "Estimate of v: "
[1] 18.7012
[1] "Test result:"
           [,1]
Test-T2 12.66480498
p.value 0.01028599

```

It also rejects the null hypothesis.

3 Comparing mean vectors of several populations

Setup: g populations, and n_ℓ observations for population ℓ .

1. $\{\mathbf{X}_{\ell,1}, \mathbf{X}_{\ell,2}, \dots, \mathbf{X}_{\ell,n_\ell}\}$ is a random sample of size n_ℓ from a population with mean $\boldsymbol{\mu}_\ell$, where $\ell = 1, \dots, g$. The random samples from different populations are independent.
2. All populations have a common covariance matrix $\boldsymbol{\Sigma}$, which is positive definite.
3. Each population is multivariate normal, assuming dimension p .

The condition 3 can be relaxed when the sample size is sufficiently large.

Hypothesis of interest $H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \dots = \boldsymbol{\mu}_g$ versus $H_a : \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$ for some $1 \leq i, j \leq g$ and $i \neq j$.

Univariate case: Recall the case of $p = 1$. The null hypothesis of $\mu_1 = \mu_2 = \dots = \mu_\ell$ can be written as $\tau_1 = \tau_2 = \dots = \tau_\ell = 0$, where τ_j is the deviation of μ_j from the overall mean μ , i.e., $\mu_j = \mu + \tau_j$.

The model can then be written as

$$X_{\ell,j} = \mu + \tau_\ell + e_{\ell,j}, \quad \ell = 1, \dots, g; j = 1, \dots, n_\ell,$$

where $e_{\ell,j} \sim N(0, \sigma_a^2)$. For unique identification of parameters, it is commonly assumed that $\sum_{\ell=1}^g n_\ell \tau_\ell = 0$.

For the data, an analogous decomposition is

$$x_{\ell,j} = \bar{x} + (\bar{x}_\ell - \bar{x}) + (x_{\ell,j} - \bar{x}_\ell),$$

where $\bar{x} = (\sum_{\ell=1}^g \sum_{j=1}^{n_\ell} x_{\ell,j})/n$ with $n = \sum_{\ell=1}^g n_\ell$, $\bar{x}_\ell = (\sum_{j=1}^{n_\ell} x_{\ell,j})/n_\ell$. Here \bar{x} is an estimate of μ , $(\bar{x}_\ell - \bar{x})$ is an estimate of τ_ℓ and $(x_{\ell,j} - \bar{x}_\ell)$ is an estimate of the error term $e_{\ell,j}$.

Subtracting \bar{x} from the prior equation, taking squares, and summing, we have the identity

$$\sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} (x_{\ell,j} - \bar{x})^2 = \sum_{\ell=1}^g n_{\ell} (\bar{x}_{\ell} - \bar{x})^2 + \sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} (x_{\ell,j} - \bar{x}_{\ell})^2. \quad (2)$$

The cross-product term drops because it is zero, indicating the terms are orthogonal to each other. This identity is often thought of as

$$\left(\begin{array}{c} \text{Sum of Squares} \\ \text{of Total Variations} \end{array} \right) = \left(\begin{array}{c} \text{Sum of Squares} \\ \text{of Treatments} \end{array} \right) + \left(\begin{array}{c} \text{Sum of Squares} \\ \text{of Residuals} \end{array} \right).$$

In addition, the number of independent quantities in each term of the above identity is related by

$$\sum_{\ell=1}^g n_{\ell} - 1 = (g - 1) + \sum_{\ell=1}^g (n_{\ell} - 1).$$

This is known as the degrees of freedom for each term.

The univariate Analysis of Variance Table (ANOVA) is a summary of the above results.

Source of variation	Sum of Squares	Degrees of freedom
Treatments	$SS_{tr} = \sum_{\ell=1}^g n_{\ell} (\bar{x}_{\ell} - \bar{x})^2$	$g - 1$
Residuals	$SS_{res} = \sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} (x_{\ell,j} - \bar{x}_{\ell})^2$	$\sum_{\ell=1}^g n_{\ell} - g$
Total	$SS_{tot} = \sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} (x_{\ell,j} - \bar{x})^2$	$\sum_{\ell=1}^g n_{\ell} - 1$

The usual F -test rejects the null hypothesis $H_0 : \tau_1 = \tau_2 = \dots = \tau_g = 0$ at the α level if

$$F = \frac{SS_{tr}/(g - 1)}{SS_{res}/(\sum_{\ell=1}^g n_{\ell} - g)} > F_{g-1, \sum n_{\ell} - g}.$$

The rationale for the F -test is as follows. \bar{x}_{ℓ} is an estimate of μ_{ℓ} so that the numerator is a weighted measure of the variation of \bar{x}_{ℓ} between the g populations, where the weights depend on the sample size of each population. The issue then is to judge the magnitude of this variation. The denominator provides a reference measure of the variation because it is an estimate of the random variation (i.e., σ^2) of the data. If the variation between the populations is large with respect to the random noises, then the means are said to be different.

Remark: The R command for univariate analysis of variance is `aov`. For illustration, consider the data in Example 6.7 of the text. The R analysis corresponding to that of Example 6.8 is as follows.

```

> x=c(1,1,1,2,2,3,3,3)
> y=c(9,6,9,0,2,3,1,2)
> g1=factor(x)
> g1
[1] 1 1 1 2 2 3 3 3
Levels: 1 2 3
> help(aov)
> m1=aov(y~g1)
> m1
Call:
  aov(formula = y ~ g1)

```

Terms:

	g1	Residuals
Sum of Squares	78	10
Deg. of Freedom	2	5

Residual standard error: 1.414214
 Estimated effects may be unbalanced

```
> summary(m1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
g1	2	78	39	19.5	0.004353 **
Residuals	5	10	2		

Multivariate case. When $p > 1$, the model becomes

$$\mathbf{X}_{\ell,j} = \boldsymbol{\mu} + \boldsymbol{\tau}_\ell + \mathbf{e}_{\ell,j}, \quad j = 1, \dots, n_\ell; \quad \ell = 1, \dots, g$$

where $\mathbf{e}_{\ell,j} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$. As before, $\boldsymbol{\mu}$ is the overall mean vector, and $\boldsymbol{\tau}_\ell$ denotes the ℓ th treatment effect satisfying $\sum_{\ell=1}^g n_\ell \boldsymbol{\tau}_\ell = \mathbf{0}$.

The data can be decomposed as

$$\mathbf{x}_{\ell,j} = \bar{\mathbf{x}} + (\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}}) + (\mathbf{x}_{\ell,j} - \bar{\mathbf{x}}_\ell).$$

Subtracting $\bar{\mathbf{x}}$ from the prior equation, post-multiplying by its own transpose and summing, we obtain

$$\sum_{\ell=1}^g \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell,j} - \bar{\mathbf{x}})(\mathbf{x}_{\ell,j} - \bar{\mathbf{x}})' = \sum_{\ell=1}^g n_\ell (\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}})(\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}})' + \sum_{\ell=1}^g \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell,j} - \bar{\mathbf{x}}_\ell)(\mathbf{x}_{\ell,j} - \bar{\mathbf{x}}_\ell)',$$

where, as in the univariate case, the cross-product term sums to zero. For ease in notation, we define

$$\begin{aligned} \mathbf{W} &= \sum_{\ell=1}^g \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell,j} - \bar{\mathbf{x}}_\ell)(\mathbf{x}_{\ell,j} - \bar{\mathbf{x}}_\ell)' \\ &= (n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \dots + (n_g - 1)\mathbf{S}_g, \end{aligned}$$

to represent the *within* population sum of squares and cross products matrix, and

$$\mathbf{B} = \sum_{\ell=1}^g n_{\ell}(\bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}})'$$

to denote the *between* population sum of squares and cross-products matrix. The hypothesis of no treatment effects, $H_o : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \dots = \boldsymbol{\tau}_g = \mathbf{0}$, is tested by considering the relative sizes of the treatment and residual sums of squares and cross-products.

The multivariate analysis of variance (MANOVA) table is given by

Source of variation	Matrix of sum of squares and cross-products	Degrees of freedom
Treatment	$\mathbf{B} = \sum_{\ell=1}^g n_{\ell}(\bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}})'$	$g - 1$
Residuals	$\mathbf{W} = \sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} (\mathbf{x}_{\ell,j} - \bar{\mathbf{x}}_{\ell})(\mathbf{x}_{\ell,j} - \bar{\mathbf{x}}_{\ell})'$	$\sum_{\ell=1}^g n_{\ell} - g$
Total	$\mathbf{B} + \mathbf{W} = \sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} (\mathbf{x}_{\ell,j} - \bar{\mathbf{x}})(\mathbf{x}_{\ell,j} - \bar{\mathbf{x}})'$	$\sum_{\ell=1}^g n_{\ell} - g$

The test then involves *generalized variances*, i.e. determinant of the covariance matrix. Specifically, one rejects H_o if

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|} = \frac{\left| \sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} (\mathbf{x}_{\ell,j} - \bar{\mathbf{x}}_{\ell})(\mathbf{x}_{\ell,j} - \bar{\mathbf{x}}_{\ell})' \right|}{\left| \sum_{\ell=1}^g \sum_{j=1}^{n_{\ell}} (\mathbf{x}_{\ell,j} - \bar{\mathbf{x}})(\mathbf{x}_{\ell,j} - \bar{\mathbf{x}})' \right|}$$

is too small. This test statistics was proposed by Wilks and is commonly referred to as *Wilk's lambda*. The distribution of Λ^* is given in Table 6.3 of the text for some special cases (p. 303). For other cases and large sample sizes, a modification of Λ^* due to Bartlett (1938) can be used. Specifically, if H_o is true and $\sum_{\ell} n_{\ell} = n$ is large,

$$-\left(n - 1 - \frac{p+g}{2}\right) \ln(\Lambda^*) = -\left(n - 1 - \frac{p+g}{2}\right) \ln\left(\frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|}\right),$$

has approximately a chi-square distribution with $p(g - 1)$ degrees of freedom.

Remark. The R command for multivariate analysis of variance is **manova**. Below are some examples.

```
> help(manova)
```

```
> help(summary.manova)
```

```
** Example 6.9 of the text on Page 304 and 305.
```

```
> x=matrix(c(1,1,1,2,2,3,3,3,9,6,9,0,2,3,1,2,3,2,7,4,0,8,9,7),8,3)
> x
```

```

      [,1] [,2] [,3]
[1,]  1    9    3
[2,]  1    6    2
[3,]  1    9    7
[4,]  2    0    4
[5,]  2    2    0
[6,]  3    3    8
[7,]  3    1    9
[8,]  3    2    7
> fac1=factor(x[,1])
> xx=x[,2:3]

> m2=manova(xx~fac1)
> m2
Call:
  manova(xx ~ fac1)

Terms:
              fac1 Residuals
resp 1             78         10
resp 2             48         24
Deg. of Freedom    2          5

Residual standard error: 1.414214 2.190890
Estimated effects may be unbalanced
---

> summary(m2)
              Df Pillai approx F num Df den Df   Pr(>F)
fac1          2 1.5408   8.3882     4    10 0.003096 **
Residuals    5
---

> summary(m2,test='Wilks')
              Df  Wilks approx F num Df den Df   Pr(>F)
fac1          2 0.0385   8.1989     4     8 0.006234 **
Residuals    5
---

** Another example **
> help(gl) % generates factors
> da=read.table("t6-9.dat")

```


$m = pg(g - 1)/2$. Consequently, with probability at least $1 - \alpha$, $\tau_{k,i} - \tau_{\ell,i}$ belongs to

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

for all $i = 1, \dots, p$ and all differences $\ell < k = 1, \dots, g$.

4 Testing for equality of covariance matrices

The setup: g populations and p variables. The covariance matrix of population j is Σ_j , which is positive definite.

$H_o : \Sigma_1 = \Sigma_2 = \dots = \Sigma_g = \Sigma$ versus $H_a : \Sigma_i \neq \Sigma_j$ for some $1 \leq i \neq j \leq g$.

The most commonly used test statistic is the Box's M test. It is a likelihood-ratio type of test. Under the normality assumption, the likelihood ratio statistic for testing equality in covariance matrix is

$$\Lambda = \prod_{\ell=1}^g \left(\frac{|\mathbf{S}_\ell|}{|\mathbf{S}_{pool}|} \right)^{(n_\ell-1)/2},$$

where n_ℓ is the sample size of the ℓ th population, \mathbf{S}_ℓ is the sample covariance matrix of ℓ th population and

$$\mathbf{S}_{pool} = \frac{1}{\sum_{\ell=1}^g (n_\ell - 1)} [(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \dots + (n_g - 1)\mathbf{S}_g],$$

is the pooled sample covariance matrix. Box's test is based on a χ^2 approximation to the sampling distribution of $-2 \ln(\Lambda)$. Specifically,

$$M \equiv -2 \ln(\Lambda) = \left[\sum_{\ell=1}^g (n_\ell - 1) \right] \ln(|\mathbf{S}_{pool}|) - \sum_{\ell=1}^g [(n_\ell - 1) \ln(|\mathbf{S}_\ell|)].$$

Under H_o , \mathbf{S}_ℓ are not expected to differ too much so that they should be close to \mathbf{S}_{pool} . In this case, the ratio of determinants should be close to 1 and the M -statistic will be small.

Box's test. Let

$$u = \left[\sum_{\ell=1}^g \frac{1}{(n_\ell - 1)} - \frac{1}{\sum_{\ell=1}^g (n_\ell - 1)} \right] \left[\frac{2p^2 + 3p - 1}{6(p+1)(g-1)} \right],$$

where p is the number of variables and g is the number of populations. Then

$$C = (1 - u)M = (1 - u) \left\{ \left[\sum_{\ell=1}^g (n_\ell - 1) \right] \ln(|\mathbf{S}_{pool}|) - \sum_{\ell=1}^g [(n_\ell - 1) \ln(|\mathbf{S}_\ell|)] \right\}$$

has an approximate χ^2 distribution with $v = \frac{1}{2}p(p+1)(g-1)$ degrees of freedom. One rejects H_o if $C > \chi_{p(p+1)(g-1)/2}^2(\alpha)$.

Remark: A simple R script, called **Box-M.txt** is written to perform the Box-M test for equal covariance matrices. For illustration, consider the data in Table 6.1. The null hypothesis cannot be rejected at the 5% level. The program requires two input variables: (a) data set and (b) a vector of (n_1, n_2, \dots, n_g) of sample sizes. The data set are arranged in population ordering that matches the sample size vector.

```
> source("Box-M.txt")
> mm=Box.M(y,nv)
[1] "Test result:"
      [,1]
Box.M-C 4.0572053
p.value 0.2553529
> names(mm)
[1] "Box.M"      "Test.Stat"  "p.value"
```

5 Two-way multivariate analysis of variance

Univariate case: The model is

$$X_{\ell kr} = \mu + \tau_\ell + \beta_k + \gamma_{\ell k} + e_{\ell kr}; \quad \ell = 1, \dots, g; \quad k = 1, \dots, b; \quad r = 1, \dots, n$$

where $\sum_{\ell=1}^g \tau_\ell = \sum_{k=1}^b \beta_k = \sum_{\ell=1}^g \gamma_{\ell k} = \sum_{k=1}^b \gamma_{\ell k} = 0$ and $e_{\ell kr} \sim N(0, \sigma^2)$. Here μ is the overall mean, representing the general level of response, τ_ℓ is the fixed effect of factor 1, β_k is the fixed effect of factor 2, and $\gamma_{\ell k}$ is the interaction between factor 1 and factor 2.

For the data, the corresponding decomposition is

$$x_{\ell kr} = \bar{x} + (\bar{x}_{\ell.} - \bar{x}) + (\bar{x}_{.k} - \bar{x}) + (\bar{x}_{\ell k} - \bar{x}_{\ell.} - \bar{x}_{.k} + \bar{x}) + (x_{\ell kr} - \bar{x}_{\ell k}),$$

where \bar{x} is the overall sample mean, $\bar{x}_{\ell.} = \frac{1}{bn} \sum_{k=1}^b \sum_{r=1}^n x_{\ell kr}$, $\bar{x}_{.k} = \frac{1}{gn} \sum_{\ell=1}^g \sum_{r=1}^n x_{\ell kr}$, and $\bar{x}_{\ell k} = \frac{1}{n} \sum_{r=1}^n x_{\ell kr}$. Subtracting \bar{x} , squaring and summing, we have the identity

$$\begin{aligned} \sum_{\ell=1}^g \sum_{k=1}^b \sum_{r=1}^n (x_{\ell kr} - \bar{x})^2 &= \sum_{\ell=1}^g bn(\bar{x}_{\ell.} - \bar{x})^2 + \sum_{k=1}^b gn(\bar{x}_{.k} - \bar{x})^2 \\ &+ \sum_{\ell=1}^g \sum_{k=1}^b n(\bar{x}_{\ell k} - \bar{x}_{\ell.} - \bar{x}_{.k} + \bar{x})^2 + \sum_{\ell=1}^g \sum_{k=1}^b \sum_{r=1}^n (x_{\ell kr} - \bar{x}_{\ell k})^2. \end{aligned}$$

This identity is commonly expressed as

$$SS_{tot} = SS_{fac1} + SS_{fac2} + SS_{int} + SS_{res}.$$

The corresponding degrees of freedom is

$$gbn - 1 = (g - 1) + (b - 1) + (g - 1)(b - 1) + gb(n - 1).$$

The univariate analysis of variance table is simply the summary of the prior two equations.

Univariate Two-Way Analysis of Variance Table

Source of variation	Sum of Squares	Degrees of Freedom	Mean Squares	F-ratio
Factor 1	SS_{fac1}	$g - 1$	$MS_{fac1} = \frac{SS_{fac1}}{g-1}$	$\frac{MS_{fac1}}{MSE}$
Factor 2	SS_{fac2}	$b - 1$	$MS_{fac2} = \frac{SS_{fac2}}{b-1}$	$\frac{MS_{fac2}}{MSE}$
Interaction	SS_{int}	$(g - 1)(b - 1)$	$MS_{int} = \frac{SS_{int}}{(g-1)(b-1)}$	$\frac{MS_{int}}{MSE}$
Residuals	SS_{res}	$gb(n - 1)$	$MSE = \frac{SS_{res}}{gb(n-1)}$	
Total	SS_{tot}	$gbn - 1$		

In the table, mean squares are defined as the sum of squares divided by its degrees of freedom. For instance, $MSE = \frac{1}{gb(n-1)} \sum_{\ell=1}^g \sum_{k=1}^b \sum_{r=1}^n (x_{\ell kr} - \bar{x}_{\ell k})^2$, which is an estimate of σ^2 . The hypothesis of no interaction, $H_o : \gamma_{\ell k} = 0$ for all ℓ and k versus $H_a : \gamma_{\ell k} \neq 0$ for some ℓ and k , can be tested by the F -ratio $F = \frac{MS_{int}}{MSE} \sim F_{(g-1)(b-1), gb(n-1)}$. Similar tests can be done for the factor effects.

Multivariate case. The multivariate version of the model is

$$\mathbf{X}_{\ell kr} = \boldsymbol{\mu} + \boldsymbol{\tau}_{\ell} + \boldsymbol{\beta}_k + \boldsymbol{\gamma}_{\ell k} + \mathbf{e}_{\ell kr}, \quad \ell = 1, \dots, g; \quad k = 1, \dots, b; \quad r = 1, \dots, n,$$

where $\mathbf{e}_{\ell kr} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, and $\sum_{\ell=1}^g \boldsymbol{\tau}_{\ell} = \sum_{k=1}^b \boldsymbol{\beta}_k = \sum_{\ell=1}^g \boldsymbol{\gamma}_{\ell k} = \sum_{k=1}^b \boldsymbol{\gamma}_{\ell k} = \mathbf{0}$. The corresponding decomposition for the data is

$$\mathbf{x}_{\ell kr} = \bar{\mathbf{x}} + (\bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}}) + (\bar{\mathbf{x}}_{\cdot k} - \bar{\mathbf{x}}) + (\bar{\mathbf{x}}_{\ell k} - \bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}}_{\cdot k} + \bar{\mathbf{x}}) + (\mathbf{x}_{\ell kr} - \bar{\mathbf{x}}_{\ell k}).$$

This leads to the identity

$$\begin{aligned} \sum_{\ell=1}^g \sum_{k=1}^b \sum_{r=1}^n (\mathbf{x}_{\ell kr} - \bar{\mathbf{x}})(\mathbf{x}_{\ell kr} - \bar{\mathbf{x}})' &= \sum_{\ell=1}^g bn(\bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}})' + \sum_{k=1}^b gn(\bar{\mathbf{x}}_{\cdot k} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{\cdot k} - \bar{\mathbf{x}})' \\ &+ \sum_{\ell=1}^g \sum_{k=1}^b n(\bar{\mathbf{x}}_{\ell k} - \bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}}_{\cdot k} + \bar{\mathbf{x}})(\bar{\mathbf{x}}_{\ell k} - \bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}}_{\cdot k} + \bar{\mathbf{x}})' \\ &+ \sum_{\ell=1}^g \sum_{k=1}^b \sum_{r=1}^n (\mathbf{x}_{\ell kr} - \bar{\mathbf{x}}_{\ell k})(\mathbf{x}_{\ell kr} - \bar{\mathbf{x}}_{\ell k})'. \end{aligned}$$

Denote the identity as

$$SSP_{tot} = SSP_{fac1} + SSP_{fac2} + SSP_{int} + SSP_{res},$$

where SSP stands for sum of squares and cross-products. The identity for the degrees of freedom remains the same as the univariate case. Similarly, we can construct a multivariate Two-Way Analysis of Variance table as the univariate case. However, the tests are conducted based on the *generalized variances*.

A test of no interaction,

$$H_o : \gamma_{\ell k} = \mathbf{0} \quad \text{for all } \ell, k \quad \text{vs} \quad H_a : \gamma_{\ell k} \neq \mathbf{0} \quad \text{some } \ell, k$$

is the likelihood ratio statistic

$$\Lambda^* = \frac{|SSP_{res}|}{|SSP_{int} + SSP_{res}|}.$$

Using Bartlett's approximation, one reject H_o at the α level if

$$- \left[gb(n-1) - \frac{p+1-(g-1)(b-1)}{2} \right] \ln(\Lambda^*) > \chi_{(g-1)(b-1)p}^2(\alpha).$$

The main effect of factor 1 is tested by

$$H_o : \tau_1 = \tau_2 = \dots = \tau_g = \mathbf{0} \quad \text{vs} \quad H_a : \tau_\ell \neq \mathbf{0} \quad \text{for some } \ell.$$

The test statistic is

$$\Lambda^* = \frac{|SSP_{res}|}{|SSP_{fac1} + SSP_{res}|}.$$

The corresponding Bartlett's approximation is

$$- \left[gb(n-1) - \frac{p+1-(g-1)}{2} \right] \ln(\Lambda^*) \sim \chi_{(g-1)p}^2.$$

Similarly, the main effect of factor 2 is tested by

$$H_o : \beta_1 = \beta_2 = \dots = \beta_b = \mathbf{0} \quad \text{vs} \quad H_a : \beta_k \neq \mathbf{0} \quad \text{for some } k.$$

The test statistic is

$$\Lambda^* = \frac{|SSP_{res}|}{|SSP_{fac2} + SSP_{res}|}.$$

The corresponding Bartlett's approximation is

$$- \left[gb(n-1) - \frac{p+1-(b-1)}{2} \right] \ln(\Lambda^*) \sim \chi_{(b-1)p}^2.$$

When a null hypothesis is rejected, one can consider the simultaneous confidence intervals (based on Bonferroni method) to conduct further study.

*** Third example **** Two factors

```
> da=read.table("T6-4.dat")
```

```
> da
```

```
   V1 V2 V3  V4 V5
1    0  0 6.5  9.5 4.4
```

```

2  0  0 6.2  9.9 6.4
3  0  0 5.8  9.6 3.0
4  0  0 6.5  9.6 4.1
5  0  0 6.5  9.2 0.8
6  0  1 6.9  9.1 5.7
7  0  1 7.2 10.0 2.0
8  0  1 6.9  9.9 3.9
9  0  1 6.1  9.5 1.9
10 0  1 6.3  9.4 5.7
11 1  0 6.7  9.1 2.8
12 1  0 6.6  9.3 4.1
13 1  0 7.2  8.3 3.8
14 1  0 7.1  8.4 1.6
15 1  0 6.8  8.5 3.4
16 1  1 7.1  9.2 8.4
17 1  1 7.0  8.8 5.2
18 1  1 7.2  9.7 6.9
19 1  1 7.5 10.1 2.7
20 1  1 7.6  9.2 1.9
> y=cbind(da[,3],da[,4],da[,5])
> fac1=factor(da[,1])
> fac1
 [1] 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
Levels: 0 1
> fac2=factor(da[,2])

**** Analyze individual response variables ****
> y1=x[,3]
> m1=aov(y1~fac1+fac2+fac1*fac2)
> summary(m1)
          Df Sum Sq Mean Sq F value Pr(>F)
fac1      1  1.74050  1.74050  15.7868 0.001092 **
fac2      1  0.76050  0.76050   6.8980 0.018330 *
fac1:fac2  1  0.00050  0.00050   0.0045 0.947143
Residuals 16  1.76400  0.11025
---

> y2=x[,4]
> m2=aov(y2~fac1+fac2+fac1*fac2)
> summary(m2)
          Df Sum Sq Mean Sq F value Pr(>F)
fac1      1  1.30050  1.30050   7.9178 0.01248 *

```

```

fac2      1 0.61250 0.61250 3.7291 0.07139 .
fac1:fac2 1 0.54450 0.54450 3.3151 0.08740 .
Residuals 16 2.62800 0.16425

```

```

> y3=x[,5]
> m3=aov(y3~fac1+fac2+fac1*fac2)
> summary(m3)

```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fac1	1	0.421	0.421	0.1036	0.7517
fac2	1	4.901	4.901	1.2077	0.2881
fac1:fac2	1	3.961	3.961	0.9760	0.3379
Residuals	16	64.924	4.058		

**** Joint analysis ****

```

> m2=manova(y~fac1+fac2+fac1*fac2)
> m2

```

Call:

```
manova(y ~ fac1 + fac2 + fac1 * fac2)
```

Terms:

	fac1	fac2	fac1:fac2	Residuals
resp 1	1.7405	0.7605	0.0005	1.7640
resp 2	1.3005	0.6125	0.5445	2.6280
resp 3	0.4205	4.9005	3.9605	64.9240
Deg. of Freedom	1	1	1	16

Residual standard error: 0.3320392 0.4052777 2.014386

Estimated effects may be unbalanced

```
> summary(m2,test="Wilks")
```

	Df	Wilks	approx F	num Df	den Df	Pr(>F)
fac1	1	0.3819	7.5543	3	14	0.003034 **
fac2	1	0.5230	4.2556	3	14	0.024745 *
fac1:fac2	1	0.7771	1.3385	3	14	0.301782
Residuals	16					

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> summary(m2,test="Pillai")
```

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
fac1	1	0.6181	7.5543	3	14	0.003034 **
fac2	1	0.4770	4.2556	3	14	0.024745 *

fac1:fac2 1 0.2229 1.3385 3 14 0.301782

Residuals 16

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

6 Profile analysis

Profile analysis pertains to situations in which a battery of p treatments are administered to two or more groups of subjects. All responses must be expressed in similar units and the responses for the different groups are assumed to be independent of one another. In profile analysis, the question of equality of mean vectors is divided into several specific possibilities. For example, consider the case of two groups. The mean vectors are $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{ip})'$, where $i = 1, 2$. The questions of interest in profile analysis are

1. Are the profile *parallel*? Equivalently, is $H_{o1} : \mu_{1j} - \mu_{1,j-1} = \mu_{2j} - \mu_{2,j-1}$ for $j = 2, 3, \dots, p$, acceptable?
2. Assuming that the profile are parallel, are the profile *coincident*? Equivalently, is $H_{o2} : \mu_{1j} = \mu_{2j}$ for $j = 1, 2, \dots, p$, acceptable?
3. Assuming that the profile are coincident, are the profiles *level*? That is, are all means equal to the same values? Equivalently, is $H_{o3} : \mu_{11} = \mu_{12} = \dots = \mu_{1p} = \mu_{21} = \mu_{22} = \dots = \mu_{2p}$ acceptable?

The null hypothesis H_{o1} can be written as

$$H_{o1} : \mathbf{C}\boldsymbol{\mu}_1 = \mathbf{C}\boldsymbol{\mu}_2,$$

where \mathbf{C} is a contrast matrix

$$\mathbf{C}_{(p-1) \times p} = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}.$$

The data can then be transformed to obtain the samples $\{\mathbf{C}\mathbf{x}_{1j}\}_{j=1}^{n_1}$ and $\{\mathbf{C}\mathbf{x}_{2j}\}_{j=1}^{n_2}$, where n_1 and n_2 are the sample sizes of the two groups, respectively. Consequently, to test for *parallel profiles* for two normal populations, one rejects $H_{o1} : \mathbf{C}\boldsymbol{\mu}_1 = \mathbf{C}\boldsymbol{\mu}_2$ at the level α if

$$T^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{C}' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{C} \mathbf{S}_{pool} \mathbf{C}' \right]^{-1} \mathbf{C} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) > c^2,$$

where

$$c^2 = \frac{(n_1 + n_2 - 2)(p - 1)}{n_1 + n_2 - p} F_{p-1, n_1 + n_2 - p}(\alpha).$$

When the profiles are parallel, then either $\mu_{1i} > \mu_{2i}$ or $\mu_{1i} < \mu_{2i}$ for all i . Under this condition, the profiles will be coincident only if $\sum_{i=1}^p \mu_{1i} = \sum_{i=1}^p \mu_{2i}$, i.e. $\mathbf{1}'\boldsymbol{\mu}_1 = utwi\mathbf{1}'\boldsymbol{\mu}_2$, where $\mathbf{1}$ is the p -dimensional vector of 1's. Therefore, the second stage of the test is $H_{o2} : \mathbf{1}'\boldsymbol{\mu}_1 = \mathbf{1}'\boldsymbol{\mu}_2$. One can transform the data and apply the usual two-sample t -test. Specifically, to test for *coincident profiles*, given that the profiles are parallel, one rejects H_{o2} at the level α if

$$\begin{aligned} T^2 &= \mathbf{1}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{1}'\mathbf{S}_{pool}\mathbf{1} \right]^{-1} \mathbf{1}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\ &= \left(\frac{\mathbf{1}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{1}'\mathbf{S}_{pool}\mathbf{1}}} \right)^2 > t_{n_1+n_2-2}^2(\alpha/2). \end{aligned}$$

The next step is to check whether all variables have the same mean so that the common profile is level. When H_{o1} and H_{o2} hold, the common mean vector $\boldsymbol{\mu}$ is estimated by

$$\bar{\mathbf{x}} = \frac{n_1}{n_1 + n_2} \bar{\mathbf{x}}_1 + \frac{n_2}{n_1 + n_2} \bar{\mathbf{x}}_2.$$

If the common profile is level, then $\mu_1 = \mu_2 = \dots = \mu_p$ and the third null hypothesis is $H_{o3} : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$, where \mathbf{C} is defined in the step 1. Thus, to test for *level profiles*, given that profiles are coincident, one rejects H_{o3} at level α if

$$(n_1 + n_2) \bar{\mathbf{x}}' \mathbf{C}' [\mathbf{C} \mathbf{S} \mathbf{C}']^{-1} \mathbf{C} \bar{\mathbf{x}} > c^2$$

where \mathbf{S} is the sample covariance matrix based on all $n_1 + n_2$ observations and

$$c^2 = \frac{(n_1 + n_2 - 1)(p - 1)}{n_1 + n_2 - p + 1} F_{p-1, n_1+n_2-p+1}(\alpha).$$

Remark: A R script `r-profile.txt` is available on the course web to perform the three profile tests discussed. For illustration, consider the data on Table 6.14 of the textbook. The results are given below:

```
> source("r-profile.txt")
> cbind(x1,x2)
  V1 V2 V3 V4 V1 V2 V3 V4
1  2  3  5  5  4  4  5  5
2  5  5  4  4  4  5  5  5
3  4  5  5  5  4  4  5  5
4  4  3  4  4  4  5  5  5
5  3  3  5  5  4  4  5  5
6  3  3  4  5  3  3  4  4
7  3  4  4  4  4  3  5  4
8  4  4  5  5  3  4  5  5
```

```

9  4  5  5  5  4  4  5  4
10 4  4  3  3  3  4  4  4
11 4  4  5  5  4  5  5  5
12 5  5  4  4  5  5  5  5
13 4  4  4  4  4  4  5  5
14 4  3  5  5  4  4  4  4
15 4  4  5  5  4  4  5  5
16 3  3  4  5  3  4  4  4
17 4  5  4  4  5  5  5  5
18 5  5  5  5  4  5  4  4
19 5  5  4  4  3  4  4  4
20 4  4  4  4  5  3  4  4
21 4  4  4  4  5  3  4  4
22 4  4  4  4  4  5  4  4
23 3  4  5  5  2  5  5  5
24 5  3  5  5  3  4  5  5
25 5  5  3  3  4  3  5  5
26 3  3  4  4  4  4  4  4
27 4  4  4  4  4  4  5  5
28 3  3  5  5  3  4  4  4
29 4  4  3  3  4  4  5  4
30 4  4  5  5  4  4  5  5

```

```

> profile(x1,x2)
[1] "Are the profiles parallel?"
      [,1]
Test-T2 8.01617
p.value 0.06256
[1] "Are the profiles coincident?"
      [,1] [,2]
Test-T2 8.01617 1.5328
p.value 0.06256 0.2207
[1] "Are the profiles level?"
      [,1] [,2] [,3]
Test-T2 8.01617 1.5328 2.482e+01
p.value 0.06256 0.2207 1.554e-04
      [,1] [,2] [,3]
Test-T2 8.01617132 1.5327696 2.482071e+01
p.value 0.06255945 0.2206853 1.554491e-04

```

7 Growth curves

Growth curve is a special case of the repeated measure problem. Here a single treatment is applied to each subject and a single characteristic is observed over a period of time. For example, we could measure the weight of each puppy at birth and then once a month for a period of time. The weight curve of a dog that is of interest and, hence, is referred to as growth curve.

Consider the example of Potthoff-Roy model for quadratic growth. Here p measurements on all subjects are taken at times t_1, t_2, \dots, t_p , and the model is

$$E(\mathbf{X}) = E \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 t_1 + \beta_2 t_1^2 \\ \beta_0 + \beta_1 t_2 + \beta_2 t_2^2 \\ \vdots \\ \beta_0 + \beta_1 t_p + \beta_2 t_p^2 \end{bmatrix},$$

where the i th mean μ_i is the quadratic expression evaluated at t_i .

When several groups of subjects are involved, one likes to compare the growth curve among the groups. Assume that g groups of subjects are involved and for group ℓ , the random sample consists of $\mathbf{X}_{\ell 1}, \dots, \mathbf{X}_{\ell, n_\ell}$, where $n_\ell > 0$ is the sample size.

Assumption: All of the ℓj are independent and have the same covariance matrix Σ . Under the quadratic growth model, the mean vectors are

$$E[\mathbf{X}_{\ell j}] = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_p & t_p^2 \end{bmatrix} \begin{bmatrix} \beta_{\ell 0} \\ \beta_{\ell 1} \\ \beta_{\ell 2} \end{bmatrix} \equiv \mathbf{B}\boldsymbol{\beta}_\ell.$$

The model can easily be generalized to the q th-order polynomial.

Under the assumption of multivariate normality, the MLE of the $\boldsymbol{\beta}_\ell$ are

$$\hat{\boldsymbol{\beta}}_\ell = (\mathbf{B}'\mathbf{S}_{pool}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{S}_{pool}^{-1}\bar{\mathbf{X}}_\ell, \quad \ell = 1, 2, \dots, g,$$

where

$$\mathbf{S}_{pool} = \frac{1}{N-g}[(n_1-1)\mathbf{S}_1 + \dots + (n_g-1)\mathbf{S}_g] = \frac{1}{N-g}\mathbf{W}$$

with $N = \sum_{\ell=1}^g n_\ell$ is the pooled estimator of the common covariance matrix Σ . The estimate covariances of the MLE are

$$\text{Cov}(\hat{\boldsymbol{\beta}}_\ell) = \frac{k}{n_\ell}(\mathbf{B}'\mathbf{S}_{pool}^{-1}\mathbf{B})^{-1}, \quad \ell = 1, \dots, g,$$

where $k = (N-g)(N-g-1)/(N-g-p+q)(N-g-p+q+1)$. The covariance between $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{\beta}}_j$ are $\mathbf{0}$ for $i \neq j$.

To test that a q th-order polynomial is adequate, the model is fit without restrictions. That is, one fits the model separately to each group. The sum of squares and cross-product matrix then becomes

$$\mathbf{W}_q = \sum_{\ell=1}^g \sum_{j=1}^{n_\ell} (\mathbf{X}_{\ell j} - \mathbf{B}\hat{\boldsymbol{\beta}}_\ell)(\mathbf{X}_{\ell j} - \mathbf{B}\hat{\boldsymbol{\beta}}_\ell)'$$

which has $N - g + p - q - 1$ degrees of freedom. The likelihood ratio test of the null hypothesis that the q th-order polynomial is adequate can be based on the Wilk's lambda

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{W}_q|}.$$

The difference in the number of parameters between the null and alternative hypothesis is $g(p - q - 1)$ so that

$$- \left(N - \frac{1}{2}(p - q + g) \right) \ln(\Lambda^*) \sim \chi_{(p-q-1)g}^2,$$

when n_ℓ are sufficiently large.

Remark: A R script for growth curve analysis, called **r-growth.txt**, is available on the course web. For demonstration, consider the data on Tables 6.5 and 6.6.

```
> source("r-growth.txt")
> growth(x,nv,pv,2)
[1] "Growth curve model"
[1] "Order: "
[1] 2
[1] "Beta-hat: "
      [,1] [,2]
[1,] 73.070 70.139
[2,]  3.644  4.090
[3,] -2.027 -1.853
[1] "Standard errors: "
      [,1] [,2]
[1,] 2.5830 2.5010
[2,] 0.8278 0.8015
[3,] 0.2813 0.2724
[1] "W"
      V1  V2  V3  V4
V1 2762 2661 2369 2336
V2 2661 2756 2344 2328
V3 2369 2344 2302 2099
V4 2336 2328 2099 2277
[1] "Wq"
      V1  V2  V3  V4
V1 2781 2699 2363 2362
```

```
V2 2699 2832 2331 2381
V3 2363 2331 2304 2090
V4 2362 2381 2090 2314
[1] "Lambda:"
[1] 0.7627123
[1] "Test result:"
      [,1]
LR-stat 7.85536
p.value 0.01969
```