Multivariate Analysis of Variance (MANOVA)

XIN HUANG and YULIA R. GEL

UNIVERSITY OF TEXAS AT DALLAS, USA

September 25, 2017

Suppose that we have pottery shards collected from four sites in the British Isles (Wiesner, 2006):

- **L**: Llanedyrn
- **C**: Caldicot
- **I**: Isle Thorns
- **A**: Ashley Rails

Each pottery sample was returned to the laboratory for chemical assay. In these assays the concentrations of five different chemicals were determined:

- Al: Aluminum
- Fe: Iron
- Mg: Magnesium
- Ca: Calcium
- Na: Sodium

We focus on the question:

Does the chemical content of the pottery depend on the site from which the pottery was obtained?

If this is the case then we might be able to use the chemical content of a pottery sample of unknown origin to determine which site the sample came.

We approach this question using the Multivariate Analysis of Variance (MANOVA) – the multivariate analog of the Analysis of Variance (ANOVA) used in univariate statistics.

Suppose for a minute that we look at pottery shards at the same four sites in the British Isles but are able to measure only Calcium (Ca).

We can then organize our data into the following table

	4 British Isles sites			
	1	2	3	4
	Y_{11}	Y_{21}	Y_{31}	Y_{41}
Ca	Y_{12}	Y_{22}	Y_{32}	Y_{42}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Y_{1n_1}	Y_{2n_2}	Y_{3n_3}	Y_{4n_4}

where n_1 is the number of pottery shards in Llanedyrn, n_2 is the number of pottery shards in Caldicot and so on; Y_{11} is Calcium measurement in the first pottery shard in Llanedyrn; Y_{21} is Calcium measurement in the first pottery shard in Caldicot, and so on.

Or more generally, if we have g different sites (or different treatments or populations)

Treatments/Populations			
1	2	$\ldots$	g
Y_{11}	Y_{21}	$\ldots$	Y_{g1}
Y_{12}	Y_{22}	$\ldots$	Y_{g2}
$\vdots$	$\vdots$		$\vdots$
Y_{1n_1}	Y_{2n_2}	$\ldots$	Y_{gn_g}

Columns correspond to the responses to g different and the rows correspond to the subjects in each of these treatments or populations. Here Y_{ij} is an observation from subject j in group i, n_i is number of subjects in group i and $N = n_1 + n_2 + \ldots + n_g$ is the total sample size.

Assumptions for the Analysis of Variance are exactly the same as for a two sample t-test except they are applied to more than two groups:

- The data from group i have **a common mean** of μ_i; i.e., $E(Y_{ij}) = \mu_i$.
- **Homoskedasticity** The data from all groups have common variance of σ^2, i.e., $var(Y_{ij}) = \sigma^2$. I.e., the variability in the data does not depend on group membership.
- **Independence** The subjects are independently sampled
- **Normality** The data are normally distributed.

The hypothesis of interest is that all of the means (i.e., Ca levels in all 4 British Isles locations) are equal to one another, i.e.

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$$

vs. an alternative hypothesis

$$H_2 : \exists l, m, 1 \leq l, m \leq g \quad \mu_l \neq \mu_m,$$

i.e., pottery shards from at least 1 location have a different level of Ca.

If we use the following notations:

- Sample mean for group i is $\bar{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$
- Grand mean is is $\bar{y}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} Y_{ij}$,

then the Analysis of Variance is a partitioning of the total sum of squares defined as:

$$\begin{aligned} SS_{Total} &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{y}_{\cdot\cdot})^2 \\ &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left((Y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})\right)^2 \\ &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{g} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2. \end{aligned}$$

The results the Analysis of Variance can be summarized in an analysis of variance as a table:

ANOVA

Source	df	SS	MS	F
Treatments	$g-1$	$\sum_{i=1}^{g} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$\frac{SS_{treat}}{g-1}$	$\frac{MS_{treat}}{MS_{error}}$
Error	$N-g$	$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{y}_{i\cdot})^2$	$\frac{SS_{error}}{N-g}$	
Total	$N-1$	$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{y}_{\cdot\cdot})^2$		

Under H_0 that treatment is equal across group means, i.e. $H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$:

$$F \sim F_{g-1,N-g},$$

and we reject H_0 at level α if

$$F > F_{g-1,N-g}.$$

Now suppose that we have data on all 5 chemical variables (or even p chemical variables). Then we can arrange our data as

Treatments/Populations

1	2	$\cdots$	g
$Y_{11} = \begin{bmatrix} Y_{111} \\ Y_{112} \\ \vdots \\ Y_{11p} \end{bmatrix}$	$Y_{21} = \begin{bmatrix} Y_{211} \\ Y_{212} \\ \vdots \\ Y_{21p} \end{bmatrix}$	$\cdots$	$Y_{g1} = \begin{bmatrix} Y_{g11} \\ Y_{g12} \\ \vdots \\ Y_{g1p} \end{bmatrix}$
$Y_{12} = \begin{bmatrix} Y_{121} \\ Y_{122} \\ \vdots \\ Y_{12p} \end{bmatrix}$	$Y_{22} = \begin{bmatrix} Y_{221} \\ Y_{222} \\ \vdots \\ Y_{22p} \end{bmatrix}$	$\cdots$	$Y_{g2} = \begin{bmatrix} Y_{g21} \\ Y_{g22} \\ \vdots \\ Y_{g2p} \end{bmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Y_{1n_1} = \begin{bmatrix} Y_{1n_11} \\ Y_{1n_12} \\ \vdots \\ Y_{1n_1p} \end{bmatrix}$	$Y_{2n_2} = \begin{bmatrix} Y_{2n_21} \\ Y_{2n_22} \\ \vdots \\ Y_{2n_2p} \end{bmatrix}$	$\cdots$	$Y_{gn_g} = \begin{bmatrix} Y_{gn_g1} \\ Y_{gn_g2} \\ \vdots \\ Y_{gn_gp} \end{bmatrix}$

Notice that while in a univariate ANOVA columns correspond to g number of treatments and rows corresponding to subjects, in MANOVA the scalar quantities, Y_{ij} are replaced by vectors having p observations.

We use similar notations as in ANOVA, i.e. Y_{ijk} is observation for variable k from subject j in group i that is represented as vector:

$$Y_{ij} = \begin{bmatrix} Y_{ij1} \\ Y_{ij2} \\ \vdots \\ Y_{ijp} \end{bmatrix};$$

and n_i is the number of subjects in i, while $N = n_1 + \ldots + n_g$ is the total sample size.

The assumptions here are essentially the same as the assumptions in ANOVA, only here they will apply to groups:

- The data from group i has **common mean** vector

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{ip} \end{bmatrix}$$

- **Common Covariance** The data from all groups have common variance-covariance matrix Σ.
- **Independence** The subjects are independently sampled.
- **Normality** The data are multivariate normally distributed.

Here we are interested in testing the general null hypothesis that group mean vectors are all equal to one another:

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

vs. the alternative

$$H_2 : \exists l, m, k, 1 \leq l, m \leq g \quad \mu_{lk} \neq \mu_{mk}.$$

This says that the H_0 is false if at least one pair of treatments is different on at least one variable.

Now we introduce similar notations as in ANOVA, i.e.

- Sample mean vector for group i is $\bar{\mathbf{y}}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{Y}_{ij} = \begin{bmatrix} \bar{y}_{i\cdot 1} \\ \bar{y}_{i\cdot 2} \\ \vdots \\ \bar{y}_{i\cdot p} \end{bmatrix}$

  where $\bar{y}_{i\cdot k} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ijk}$, i.e. sample mean vector for variable k in group i.

- Grand mean is is $\bar{\mathbf{y}}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\mathbf{Y}_{ij} = \begin{bmatrix} \bar{y}_{\cdot\cdot 1} \\ \bar{y}_{\cdot\cdot 2} \\ \vdots \\ \bar{y}_{\cdot\cdot p} \end{bmatrix}$, where

  $\bar{y}_{\cdot\cdot k} = \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} Y_{ijk}$.

In ANOVA, we define the Total Sums of Squares which is a scalar quantity. The multivariate analog is the Total Sum of Squares and Cross Products matrix, i.e. a $p \times p$ matrix:

$$T = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{Y}_{ij} - \bar{\mathbf{y}}_{..})(\mathbf{Y}_{ij} - \bar{\mathbf{y}}_{..})^T.$$

I.e., we look at the differences between the vectors of observations $\mathbf{Y}_{ij}$ and the Grand mean vector. The (k, l)-th element of T is

$$\sum_{i=1}^{g} \sum_{j=1}^{n_i} (Y_{ijk} - \bar{\mathbf{y}}_{..k})(Y_{ijl} - \bar{\mathbf{y}}_{..l})^T.$$

For $k = l$, this is the total sum of squares for variable k and measures the total variation in the k-th variable. For $k \neq l$, this measures the dependence between variables k and l across all of the observations.

We may partition the total sum of squares and cross products as follows:

$$
\begin{aligned}
T &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{Y}_{ij}-\bar{\mathbf{y}}_{\cdot\cdot})(\mathbf{Y}_{ij}-\bar{\mathbf{y}}_{\cdot\cdot})^T \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n_i}\Big\{(\mathbf{Y}_{ij}-\bar{\mathbf{y}}_{i\cdot})+(\bar{\mathbf{y}}_{i\cdot}-\bar{\mathbf{y}}_{\cdot\cdot})\Big\}\Big\{(\mathbf{Y}_{ij}-\bar{\mathbf{y}}_{i\cdot})+(\bar{\mathbf{y}}_{i\cdot}-\bar{\mathbf{y}}_{\cdot\cdot})\Big\}^T \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{Y}_{ij}-\bar{\mathbf{y}}_{i\cdot})(\mathbf{Y}_{ij}-\bar{\mathbf{y}}_{i\cdot})^T + \sum_{i=1}^{g} n_i(\bar{\mathbf{y}}_{i\cdot}-\bar{\mathbf{y}}_{\cdot\cdot})(\bar{\mathbf{y}}_{i\cdot}-\bar{\mathbf{y}}_{\cdot\cdot})^T
\end{aligned}
$$

Here we call the first term **the Error Sum of Squares and Cross Products** (E), and the second term **the Hypothesis Sum of Squares and Cross Products** (H).

The (k, l)-th element of the error sum of squares and cross products matrix E is

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(Y_{ijk} - \bar{\mathbf{y}}_{i\cdot k}\right)\left(Y_{ijl} - \bar{\mathbf{y}}_{i\cdot l}\right).$$

For $k = l$, this is the error sum of squares for variable k and measures the within treatment variation for the k-th variable. For $k \neq l$, this measures the dependence between variables k and l after taking into account the treatment.

The (k, l)-th element of the hypothesis sum of squares and cross products matrix H is

$$\sum_{i=1}^{g} n_i(\bar{y}_{i\cdot k} - \bar{y}_{\cdot\cdot k})(\bar{y}_{i\cdot l} - \bar{y}_{\cdot\cdot l})^T.$$

For $k = l$, this is the treatment sum of squares for variable k and measures the between treatment variation for the k-th variable. For $k \neq l$, this measures dependence of variables k and l across treatments.

The partitioning of the total sum of squares and cross products matrix may be summarized in the MANOVA table:

MANOVA

Source	df	SSP
Treatments	$g-1$	H
Error	$N-g$	E
Total	$N-1$	T

We reject

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

if the hypothesis sum of squares and cross products matrix H is large relative to the error sum of squares and cross products matrix E.

We can use the following different statistics based on the MANOVA table:

- **Wilk's Lambda**

$$\Lambda^* = \frac{|E|}{|H+E|}$$

If H is large relative to E, then $|H+E|$ is large relative to $|E|$. Thus, we reject the null hypothesis if Wilk's lambda is small (close to zero).

- **Hotelling-Lawley Trace**

$$T_0^2 = trace(HE^{-1})$$

If H is large relative to E, then the Hotelling-Lawley trace takes a large value. Thus, we reject the null hypothesis if this test statistic is large.

- **Pillai Trace**

$$V = trace(H(H+E)^{-1})$$

If H is large relative to E, then the Pillai trace takes a large value. Thus, we reject the null hypothesis if this test statistic is large.

- **Roy's Maximum Root**

$$\max_{1 \leq s \leq p} \lambda_s(HE^{-1}),$$

where $\lambda(HE^{-1})$ are eigenvalues of HE^{-1}. If H is large relative to E, then the Roy's root takes a large value. Thus, we will reject the null hypothesis if this test statistic is large.

Statistical tables are not available for the above test statistics. However, each of the above test statistics has an F approximation.

Example: The British Isles pottery shards

```
> fit <- manova(pot~Site)

    #### the default statistic is Pillai ######################
> summary(fit)                  # same F statistics as single-df terms
          Df Pillai approx F num Df den Df    Pr(>F)
Site       3 1.5539   4.2984     15     60 2.413e-05 ***
Residuals 22
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> summary(fit, test = "Wilks") # ANOVA table of Wilks' lambda
          Df    Wilks approx F num Df den Df   Pr(>F)
Site       3 0.012301   13.088     15 50.091 1.84e-12 ***
Residuals 22
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Example: The results of the individual ANOVAs are presented below

```
> fit <- manova(pot~Site)
>
>
>
> summary.aov(fit)                # univariate ANOVA tables
 Response Al :
             Df  Sum Sq Mean Sq F value    Pr(>F)
Site          3 175.610  58.537  26.669 1.627e-07 ***
Residuals    22  48.288   2.195
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Example: The results of the individual ANOVAs are presented below

```
Response Fe :
            Df  Sum Sq Mean Sq F value    Pr(>F)
Site         3 134.222  44.741  89.883 1.679e-12 ***
Residuals   22  10.951   0.498
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

 Response Mg :
            Df Sum Sq Mean Sq F value    Pr(>F)
Site         3 103.35  34.450   49.12 6.452e-10 ***
Residuals   22  15.43   0.701
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Example: The results of the individual ANOVAs are presented below

```
Response Ca :
            Df   Sum Sq  Mean Sq F value    Pr(>F)
Site         3 0.204703 0.068234  29.157 7.546e-08 ***
Residuals   22 0.051486 0.002340
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

 Response Na :
            Df  Sum Sq  Mean Sq F value    Pr(>F)
Site         3 0.25825 0.086082  9.5026 0.0003209 ***
Residuals   22 0.19929 0.009059
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Conclusion: Means for all chemical elements differ significantly among the sites.

For each element, the means for that element are different for at least one pair of sites.

To ensure reliability of your conclusions, it is essential to use MANOVA for multivariate correlated data (rather than individual ANOVAs!)