Topic 8: Multivariate Analysis of Variance (MANOVA)

Ying Li

Stockholm University

October 15, 2012

Def.

MANOVA is used to determine if the categorical independent variable(s) with two or more levels affect the continues dependent variable**s**.

- independent variables: categorical
- dependent variables: continues

Geometry view of ANOVA

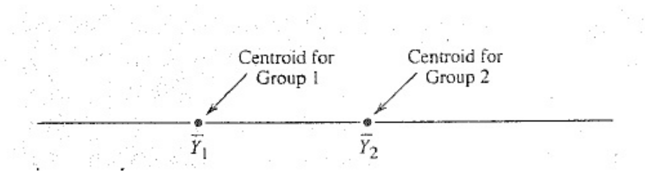

Figure: One dep. variables and one indep. variable

Geometry view of MANOVA

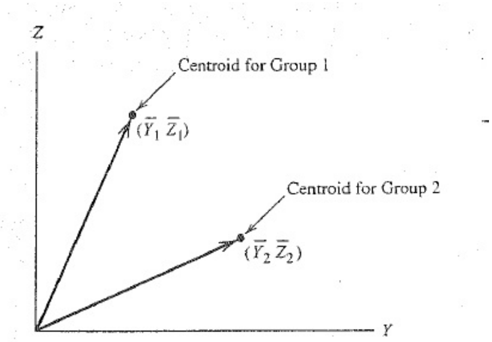

Figure: Twp dep. variables and one indep. variable

Geometry view of MANOVA

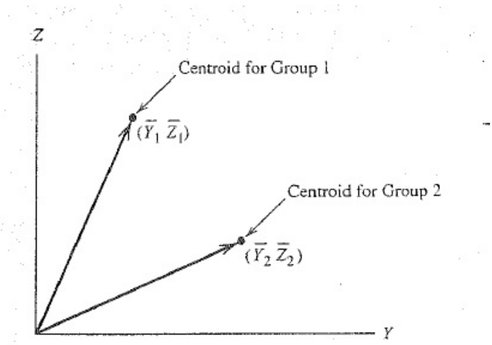

Figure: Twp dep. variables and one indep. variable

The greater distance between the two centroid, the greater the difference between the two groups.

Mahalanobis distance (MD)

For $\mathbf{a}^T : (x_1, y_1)$ and $\mathbf{b}^T : (x_2, y_2)$, the square MD distance is defined as

$$MD^2 = \frac{1}{1-r^2}[\frac{(x_1-x_2)^2}{s_x^2} + \frac{(y_1-y_2)^2}{s_y^2} - \frac{2r(x_1-x_2)(y_1-y_2)}{s_x s_y}]$$

with $s_x^2 = var(x)$, $s_y^2 = var(y)$, $r = corr(x, y)$.

Mahalanobis distance (MD)

For $\mathbf{a}^T : (x_1, y_1)$ and $\mathbf{b}^T : (x_2, y_2)$, the square MD distance is defined as

$$MD^2 = \frac{1}{1-r^2}[\frac{(x_1 - x_2)^2}{s_x^2} + \frac{(y_1 - y_2)^2}{s_y^2} - \frac{2r(x_1 - x_2)(y_1 - y_2)}{s_x s_y}]$$

with $s_x^2 = var(x)$, $s_y^2 = var(y)$, $r = corr(x, y)$. Or in matrix form:

$$MD^2 = (\textbf{a-b})S_w^{-1}(\textbf{a-b})'$$

with S_w is the pooled within group covariance between x and y.

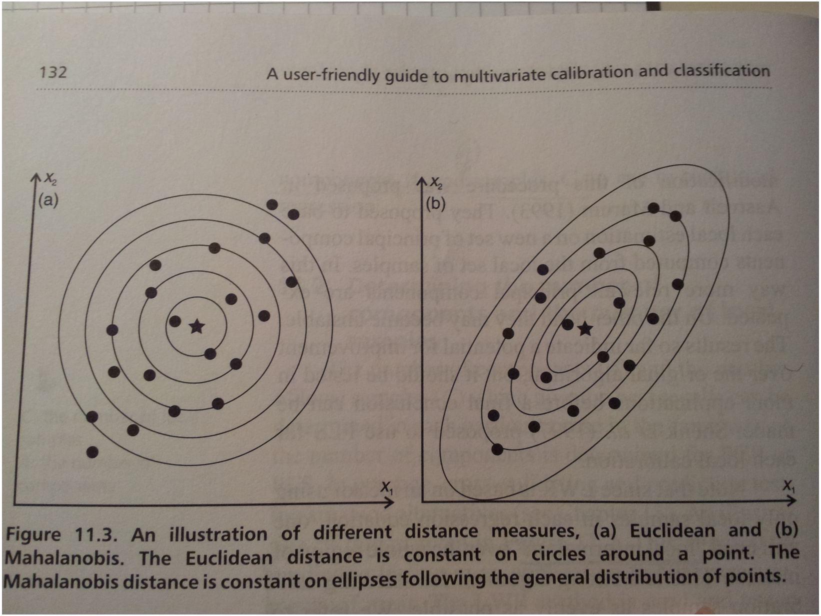

Figure 11.3. An illustration of different distance measures, (a) Euclidean and (b) Mahalanobis. The Euclidean distance is constant on circles around a point. The Mahalanobis distance is constant on ellipses following the general distribution of points.

- Statistical test are available to determine if MD between the two centroid is large.
- Geometrically, MANOVA is concerned with determining whether the MD between the group centroids is significantly greater than 0.

- $p = 2$: two dep. variables $\mathbf{y}_1$ $\mathbf{y}_2$
- $G = 2$: two groups (two levels)

- $p = 2$: two dep. variables $\mathbf{y}_1$ $\mathbf{y}_2$
- $G = 2$: two groups (two levels)

Hypothesis Test:

$$H_0 : \begin{pmatrix} \mu_{11} \\ \mu_{21} \end{pmatrix} = \begin{pmatrix} \mu_{12} \\ \mu_{22} \end{pmatrix}$$

$$H_1 : \begin{pmatrix} \mu_{11} \\ \mu_{21} \end{pmatrix} \neq \begin{pmatrix} \mu_{12} \\ \mu_{22} \end{pmatrix}$$

where μ_{ij} : ith variable for jth group.

Some statistics connect with MD distance

- Hotelling's T^2:

$$T^2 = \frac{n_1 \times n_2}{n_1 + n_2} MD^2$$

-

$$F = \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - p)p} T^2 \sim F_{p,(n_1+n_2-p-1)}$$

Some statistics connect with $SSCP_b/SSCP_w$

Some statistics connect with $SSCP_b/SSCP_w$

- Wilk's Λ: $\Lambda = \frac{|SSCP_w|}{|SSCP_t|} = \sum_{i=1}^{G} \frac{1}{1+\lambda_i}$

$$F = \left(\frac{\Lambda}{1-\Lambda}\right)\left(\frac{n_1+n_2-p-1}{p}\right) \sim F_{p,(n_1+n_2-p-1)}$$

Some statistics connect with $SSCP_b/SSCP_w$

- Wilk's Λ: $\Lambda = \frac{|SSCP_w|}{|SSCP_t|} = \sum_{i=1}^{G} \frac{1}{1+\lambda_i}$

$$F = \left(\frac{\Lambda}{1-\Lambda}\right)\left(\frac{n_1+n_2-p-1}{p}\right) \sim F_{p,(n_1+n_2-p-1)}$$

- Pillas's Trace$= \sum_{i=1}^{G} \frac{\lambda_i}{1+\lambda_i}$
- Hotelling's Trace$= \sum_{i=1}^{G} \lambda_i$
- Roy's largest Root $= \frac{\lambda_{max}}{1+\lambda_{max}}$

Some statistics connect with $SSCP_b/SSCP_w$

- Wilk's Λ: $\Lambda = \frac{|SSCP_w|}{|SSCP_t|} = \sum_{i=1}^{G} \frac{1}{1+\lambda_i}$

$$F = (\frac{\Lambda}{1-\Lambda})(\frac{n_1+n_2-p-1}{p}) \sim F_{p,(n_1+n_2-p-1)}$$

- Pillas's Trace$= \sum_{i=1}^{G} \frac{\lambda_i}{1+\lambda_i}$
- Hotelling's Trace$= \sum_{i=1}^{G} \lambda_i$
- Roy's largest Root $= \frac{\lambda_{max}}{1+\lambda_{max}}$

Note: it can be shown that for the two groups all the above measure are equivalent and can be transformed to T^2 or an F ratio

- Having determined that the means of the two groups are significantly different.

- Having determined that the means of the two groups are significantly different.
- The next obvious question is: which variables are responsible for the difference between the two groups?

- Having determined that the means of the two groups are significantly different.
- The next obvious question is: which variables are responsible for the difference between the two groups?
- To compare the means of each variable for the two groups: T-test.

Effect size can be used to assess the practical significance of the difference between the groups.

Measures:

- MD^2.
- partial eta square (PES)

$$\Lambda = \frac{SS_b}{SS_t} = \frac{F \times df_b}{F \times df_b + df_w}$$

Example

Group1	x_11	x_21	Group2	x_12	x_22
1	0.158	0.182	13	-0.012	-0.031
2	0.210	0.206	14	0.036	0.053
3	0.207	0.188	15	0.038	0.036
4	0.280	0.236	16	-0.063	-0.074
5	0.197	0.193	17	-0.054	-0.119
6	0.227	0.173	18	0.000	-0.005
7	0.148	0.196	19	0.005	0.039
8	0.254	0.212	20	0.091	0.122
9	0.079	0.147	21	-0.036	-0.072
10	0.149	0.128	22	0.045	0.064
11	0.200	0.150	23	-0.026	-0.024
12	0.187	0.191	24	0.016	0.026

$$\bar{\mathbf{x}}_1' = (0.191, 0.184),\ n_1 = 12$$
$$\bar{\mathbf{x}}_2' = (0.003, 0.001),\ n_2 = 12$$
$$SSCP_t = \begin{pmatrix} 0.265 & 0.250 \\ 0.250 & 0.261 \end{pmatrix}$$
$$SSCP_1 = \begin{pmatrix} 0.031 & 0.012 \\ 0.012 & 0.010 \end{pmatrix} \quad SSCP_2 = \begin{pmatrix} 0.022 & 0.032 \\ 0.032 & 0.051 \end{pmatrix}$$

Multiple-Group MANOVA

- $p \geq 2$: no. of dep. variables
- $G \geq 3$: no. of groups

Example

Suppose a medial researcher hypothesizes that a treatment consisting of the simultaneous administration of two drugs is more effective than a treatment consisting of the administration of only one of the drugs.
A study is designed in which 20 subjects are randomly divided into 4 groups of 5 subjects each.

- Group1: subjects are given a placebo.
- Group2: subjects are given a combination of two drugs
- Group3: subjects are given one of two drugs
- Group4: subjects are given the other drug

The effectiveness of the drugs is measured by two response variables Y_1 and Y_2

Table 11.5 **Data for Drug Effectiveness Study**

	Treatments							
	1		2		3		4	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
	1	2	8	9	2	4	4	5
	2	1	9	8	3	2	3	3
	3	2	7	9	3	3	3	4
	2	3	8	9	3	5	5	6
	2	2	8	10	4	6	5	7
Means	2	2	8	9	3	4	4	5

Multivariate Test

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall group Effect H = Type III SSCP Matrix for group E = Error SSCP Matrix S=2 M=0 N=6.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.07252706	13.57	6	30	<.0001
Pillai's Trace	1.02715018	5.63	6	32	0.0004
Hotelling-Lawley Trace	11.41361257	27.76	6	18.326	<.0001
Roy's Greatest Root	11.29190184	60.22	3	16	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

Univariate Test

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	103.7500000	34.5833333	55.33	<.0001
Error	16	10.0000000	0.6250000		
Corrected Total	19	113.7500000			

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	130.0000000	43.3333333	28.89	<.0001
Error	16	24.0000000	1.5000000		
Corrected Total	19	154.0000000			

Contrast

A contrast is a linear combination of the group means of a given factor.

$$C_{ij} = c_{i1}\mu_{1j} + c_{i2}\mu_{2j} + \cdots + c_{iG}\mu_{Gj}$$

with C_{ij}: ith contrast, jth variable; c_{ik} : the coefficients of the contrast, μ_{kj} : the means of the kth group for the jth variable.

Contrast

A contrast is a linear combination of the group means of a given factor.

$$C_{ij} = c_{i1}\mu_{1j} + c_{i2}\mu_{2j} + \cdots + c_{iG}\mu_{Gj}$$

with C_{ij}: ith contrast, jth variable; c_{ik} : the coefficients of the contrast, μ_{kj} : the means of the kth group for the jth variable.

Note: It's a good statistical practice to perform contrast analysis determined or stated as a **prior**, rather than test all possible contrasts in search of significant test.

Example

The researcher is interested in answering the following questions:

- Is the effectiveness of the placebo different from the average effectiveness of the drugs given to the other three groups?

Example

The researcher is interested in answering the following questions:

- Is the effectiveness of the placebo different from the average effectiveness of the drugs given to the other three groups?
- Is the effectiveness of the two drugs administered to the second treatment group significantly different from the average effectiveness of the drugs administered to treatment groups 3 and 4?

Example

The researcher is interested in answering the following questions:

- Is the effectiveness of the placebo different from the average effectiveness of the drugs given to the other three groups?
- Is the effectiveness of the two drugs administered to the second treatment group significantly different from the average effectiveness of the drugs administered to treatment groups 3 and 4?
- Is the effectiveness of the drug given to the third treatment group significantly different from the effectiveness of the drug given to to fourth treatment group?

Orthogonal Contrast

Contrasts are said to be **orthogonal** if

- $\sum_{i=1}^{G} c_{ik} = 0$ for all i.
- $\sum_{i=1}^{G} \frac{c_{ik}c_{lk}}{n_k}$ for all $i \neq l$

where i and l are any two contrasts.

Orthogonal Contrast

Contrasts are said to be **orthogonal** if

- $\sum_{i=1}^{G} c_{ik} = 0$ for all i.
- $\sum_{i=1}^{G} \frac{c_{ik}c_{lk}}{n_k}$ for all $i \neq l$

where i and l are any two contrasts.

Note:

- The total no. of contrasts for any given factor is equal to its degree of freedom.
- There can be infinite sets of contrast with each set consisting of maximum number of allowable contrasts

Contrast	Groups 1	2	3	4
Set 1	1	−1/3	−1/3	−1/3
	0	1	−1/2	−1/2
	0	0	1	−1
Set 2	−1/3	−1/3	−1/3	1
	−1/2	−1/2	1	0
	−1	−1	0	0
Set 3	1	−1	0	0
	0	0	1	−1
	1/2	1/2	−1/2	−1/2
Set 4	1	0	0	−1
	0	1	−1	0
	1/2	−1/2	−1/2	1/2

Univariate significant test for the contrast

Hypothesis Testing:

$$H_0 : C_{ij} = 0$$

$$H_1 : C_{ij} \geq 0$$

Univariate significant test for the contrast

Hypothesis Testing:

$$H_0 : C_{ij} = 0$$

$$H_1 : C_{ij} \geq 0$$

$$var(C_{ij}) = MSE_j \sum_{k=1}^{G} c_{ik}^2 / n_k$$

Statistics

$$t = \frac{C_{ij}}{\sqrt{MSE_j \sum_{k=1}^{G} c_{ik}^2 / n_k}}$$

or

$$T^2 = t^2 = \frac{C_{ij}^2}{MSE_j \sum_{k=1}^{G} c_{ik}^2 / n_k} = (\sum_{k=1}^{G} \frac{c_{ik}^2}{n_k})^{-1} C_{ij} MSE_j^{-1} C_{ij}$$

In univariate case, T^2 is equal to F-ratio

$$F = \frac{C_{ij}^2 / (\sum_{k=1}^{G} c_{ik}^2 / n_k)}{MSE_j} \sim F(1, n - G)$$

Multivariate Significance Test for the Contrasts

Multivariate contrasts are used to simultaneously test for the effects of all the dependent variables.

Multivariate contrasts

$$\mathbf{C}_i = c_{i1}\boldsymbol{\mu}_1 + c_{i2}\boldsymbol{\mu}_2 + \cdot + c_{iG}\boldsymbol{\mu}_G$$

where $\boldsymbol{\mu}_k$ is a vector of means for the kth group and $\mathbf{C}_i$ is the ith contrast vector.

Hypothesis Testing: $H_0 : \mathbf{C}_i = 0 \; H_1 : \mathbf{C}_i \geq 0$

Test Statistics is

$$T^2 = \left(\sum_{k=1}^{G} \frac{c_{ik}^2}{n_k}\right)^{-1} \mathbf{C}_i' \hat{\Sigma}_w^{-1} \mathbf{C}_i$$

where $\hat{\Sigma}_w$ is pooled within-group covariance matrix. T^2 can be transformed into F-ratio using:

$$F = \left(\frac{df_e - p + 1}{df_e \times p}\right) T^2 \sim F(p, df_e - p + 1)$$

with df_e is the error degrees of freedom.

Note: The univariate contrasts should only be interpreted if the corresponding multivariate contrast is significant.

Example

$$\mu_1 = (2,2), \mu_2 = (8,9), \mu_3 = (3,4), \mu_4 = (4,5)$$

$$n_1 = n_2 = n_3 = n_4 = 5$$

$$\hat{\Sigma}_w = \begin{pmatrix} 0.625 & 0.438 \\ 0.438 & 1.5 \end{pmatrix}$$

Question: What's the similarities and difference between MANOVA and DA?

Question: What's the similarities and difference between MANOVA and DA?

- One of the objectives in both methods is to determine if the groups are significantly different with respect to a given set of variables.
- One of the objective of MANOVA is to test which groups are different with respect to a given set of variables, which is not the objective of DA.