Multivariate Stats, Week 11

Noah Silbert

November 3, 2017

To do

1. Review of univariate ANOVA
2. Intro to comparison of more than two multivariate means

ANOVA

A t test compares two means. The basic ANOVA procedure generalizes this, testing the null hypothesis that 2+ means are equal: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.

For example, suppose you are interested in studying the efficacy of two different treatments, relative to each other and relative to no treatment (i.e., a control condition).

In this case, ANOVA can be used to analyze the relative magnitude of variation between and within the groups (control, treatment 1, treatment 2).

ANOVA

The standard test statistic for ANOVA is the ratio of the mean squared error (MSE) between groups to the MSE within groups:

$$
\begin{aligned}
F_{k-1,k(n-1)} &= \frac{n s_{\bar{y}}^2}{s_e^2} \\
&= \frac{n \sum_{i=1}^{k} \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2 / (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{i\cdot}\right)^2 / (k(n-1))}
\end{aligned}
$$

Here, y_{ij} is the j^{th} observation in the i^{th} group, $\bar{y}_{i\cdot}$ is the mean of the i^{th} group, and $\bar{y}_{\cdot\cdot}$ is the grand mean.

The larger the deviations between the group means relative to the within-group error, the larger the F statistic will be. Note that the F statistic does not tell us which means are different from each other.

MANOVA

MANOVA (multivariate ANOVA) is typically concerned with testing $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$, where each $\boldsymbol{\mu}_i$ is a vector.

There are a number of test statistics for MANOVA. A common test statistic is Wilk's $\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|}$, where E and H are the "within" and "between" error covariance matrices, respectively (we saw these matrices when we did LDA with more than 2 groups):

$$\mathbf{H} = \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})(\bar{y}_{i.} - \bar{y}_{..})^T$$

$$\mathbf{E} = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})(y_{ij} - \bar{y}_{i.})^T$$

Here, $\mathbf{y}_{ij}$ is the j^{th} observation in the i^{th} group, $\bar{\mathbf{y}}_{i.}$ is the mean of the i^{th} group, and $\bar{\mathbf{y}}_{..}$ is the grand mean.

MANOVA continued

Wilk's Λ can also be expressed in terms of the eigenvalues of $E^{-1}H$ (which we also saw with 3+ group LDA): $\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$, where s is the number of nonzero eigenvalues.

(Recall that λ is an eigenvalue and v is an eigenvector of the matrix A if $Av = \lambda v$.)

Note that if each group has the same mean, then $H = 0$, so $\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} = \frac{|\mathbf{E}|}{|\mathbf{E}|} = 1$. Hence, *small* values of Λ are extreme (and so lead to rejection of H_0; cf. the F statistic, which is extreme when it's large).

In order for the determinants $|\mathbf{E}|$ and $|\mathbf{E}+\mathbf{H}|$ to be positive, the number of observations minus the number of groups must be greater than the dimensionality of the data.

MANOVA continued

Some other test statistics for MANOVA:

Roy's largest root: $\theta = \dfrac{\lambda_1}{1+\lambda_1}$, where λ_1 is the largest eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$. This is based on finding the linear transformation of the data that maximizes the spread of the groups means relative to the within-group variance. Large values of θ are extreme.

Pillai's trace: $V^{(s)} = \mathrm{tr}[(\mathbf{E}+\mathbf{H})^{-1}\mathbf{H}] = \sum_{i=1}^{s} \dfrac{\lambda_i}{1+\lambda_i}$. This is an extension of Roy's largest root; large values are extreme.

Lawley-Hotelling statistic: $U^{(s)} = \mathrm{tr}\left(\mathbf{E}^{-1}\mathbf{H}\right) = \sum_{i=1}^{s} \lambda_i$. This is a generalization of Hotelling's T^2 statistic. Large values are extreme.

More MANOVA

All four test statistics (Wilk's Λ, Roy's largest root, Pillai's trace, Lawley-Hotelling $U^{(s)}$) depend on $\nu_H = k - 1$ and $\nu_E = N - k$ (i.e., the same degrees of freedom terms from univariate ANOVA).

One minus Wilk's Λ gives a generalized η^2 statistic (i.e., a measure of the multivariate relationship between the independent and the dependent variables, i.e., effect size): $\eta^2_\Lambda = 1 - \Lambda$.

Roy's largest root (θ) itself provides another measure of these relationships. In addition, $\sqrt{\theta}$ is the *canonical correlation* between the group indicator variables and the dependent variables.

We'll return to canonical correlation soon.

Still more MANOVA

These four test statistics do not have the same statistical power. If the means are collinear (left panel), these statistics are, in descending order of power, $\theta \geq U^{(s)} \geq \Lambda \geq V^{(s)}$. The order is reversed for the diffuse case(s).

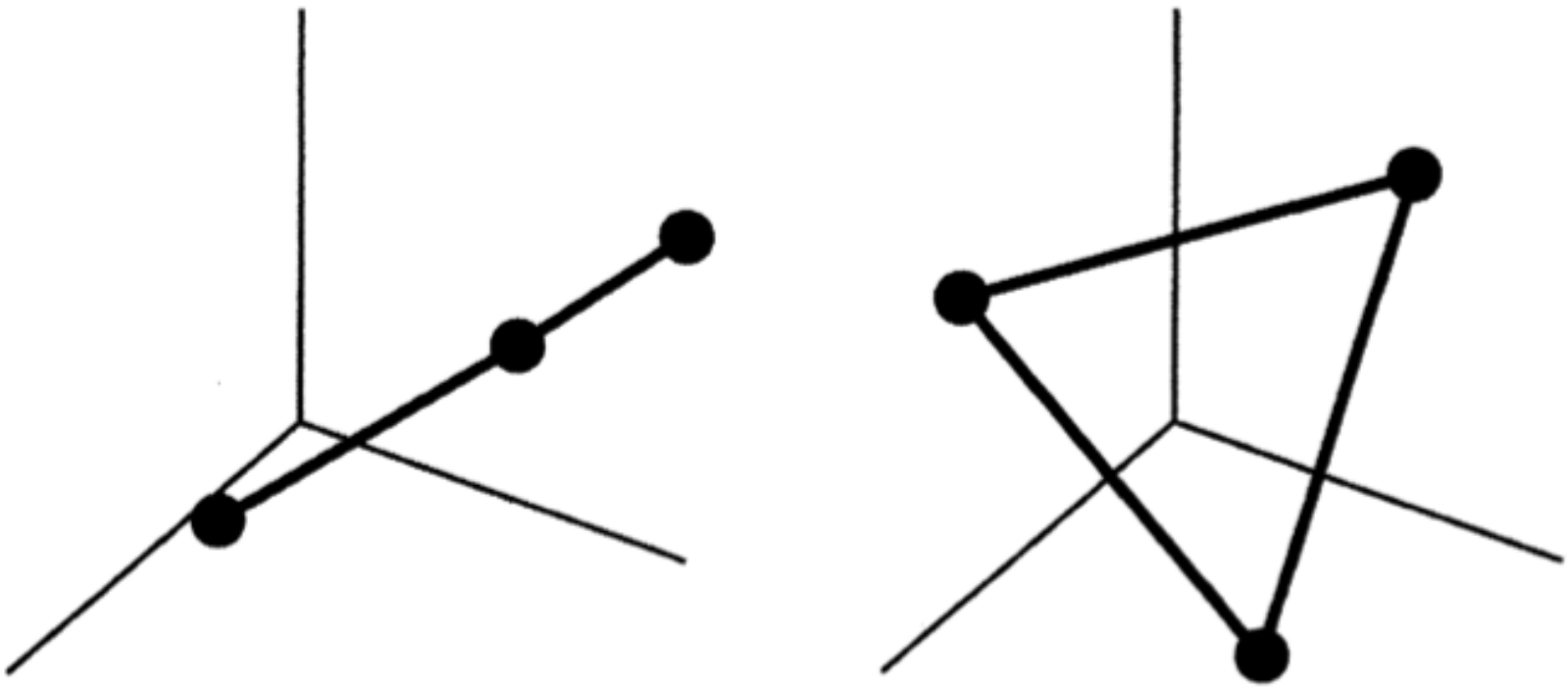

Figure 6.2 Two possible configurations for three mean vectors in 3-space.

More MANOVA already

We can get a sense of the dimensionality of the means in a data set by looking at the proportional size of the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$.

We will do this and look at the four test statistics discussed above in an R script (mvstats_week11a.R).

Factorial MANOVA

Factorial MANOVA is much like one-way MANOVA, just with a different set of **H** matrices and a different **E** matrix. These are defined in Table 6.5 (p. 201). And, just like one-way MANOVA is a multivariate generalization of one-way ANOVA, factorial MANOVA is a multivariate generalization of factorial ANOVA.

Factorial here means that the groups are defined by combinations of separate factors. For example, if you had treatment and control groups for males and females, you would have four groups: male control, female control, male treatment, female treatment.

Factorial MANOVA

This additional structure makes everything more complicated (but typically also more interesting), since there are null hypotheses for each factor on its own (main effects) and for interactions between the factors.

With the example above, the null hypothesis for sex would be $H_0 : \boldsymbol{\mu}_F = \boldsymbol{\mu}_M$; for treatment vs conrol it would be $H_0 : \boldsymbol{\mu}_t = \boldsymbol{\mu}_c$; and for the interaction of sex and treatment
$H_0 : \boldsymbol{\mu}_{Ft} - \boldsymbol{\mu}_{Fc} = \boldsymbol{\mu}_{Mt} - \boldsymbol{\mu}_{Mc}$ or, equivalently,
$H_0 : \boldsymbol{\mu}_{Ft} - \boldsymbol{\mu}_{Mt} = \boldsymbol{\mu}_{Fc} - \boldsymbol{\mu}_{Mc}$.

Factorial MANOVA in R

We can use the `manova()` function in R to do one-way or factorial MANOVA tests. Here's an example from the book (starting on p. 203) in which the torque and strain of bar steel was measured for two different rotational velocities and four lubricants:

```
# read in the data
dd = read.table('T6_6_BARSTEEL.DAT')
# name the columns
colnames(dd) = c("velocity","lubricant","torque","strain")
# make the velocity and lubricant columns into factors
dd$velocity = factor(dd$velocity)
dd$lubricant = factor(dd$lubricant)
```

Factorial MANOVA continued

```
out = manova(cbind(torque,strain) ~ velocity*lubricant,data
summary(out,test="Wilks")
```

```
##                      Df   Wilks approx F num Df den Df
## velocity              1 0.47396  12.7636      2     23 0.0
## lubricant             3 0.69158   1.5524      6     46 0.1
## velocity:lubricant    3 0.93193   0.2751      6     46 0.9
## Residuals            24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

We can do this analysis in an R script to see everything more clearly...