

Multiway ANOVA and MANOVA

Session 7

MATH 80667A: Experimental Design and Statistical Methods
HEC Montréal

Outline

Multiway ANOVA

MANOVA

Multifactorial designs

Beyond two factors

We can consider multiple factors A, B, C, \dots with respectively n_a, n_b, n_c, \dots levels and with n_r replications for each.

The total number of treatment combinations is

$$n_a \times n_b \times n_c \times \dots$$

Curse of dimensionality

Full three-way ANOVA model

Each cell of the cube is allowed to have a different mean

$$Y_{ijk r} = \mu_{ijk} + \varepsilon_{ijk r}$$

response cell mean error

with $\varepsilon_{ijk r}$ are independent error term for

- row i
- column j
- depth k
- replication r

Parametrization of a three-way ANOVA model

With the **sum-to-zero** parametrization with factors A , B and C , write the response as

$$\begin{aligned} \mathbf{E}(Y_{ijklr}) &= \mu \\ \text{theoretical average} & \quad \text{global mean} \\ & + \alpha_i + \beta_j + \gamma_k \\ & \quad \text{main effects} \\ & + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} \\ & \quad \text{two-way interactions} \\ & + (\alpha\beta\gamma)_{ijk} \\ & \quad \text{three-way interaction} \end{aligned}$$



global mean, row, column and depth main effects



row/col, row/depth and col/depth interactions and three-way interaction.

Example of three-way design

Petty, Cacioppo and Heesacker (1981). Effects of rhetorical questions on persuasion: A cognitive response analysis. *Journal of Personality and Social Psychology*.

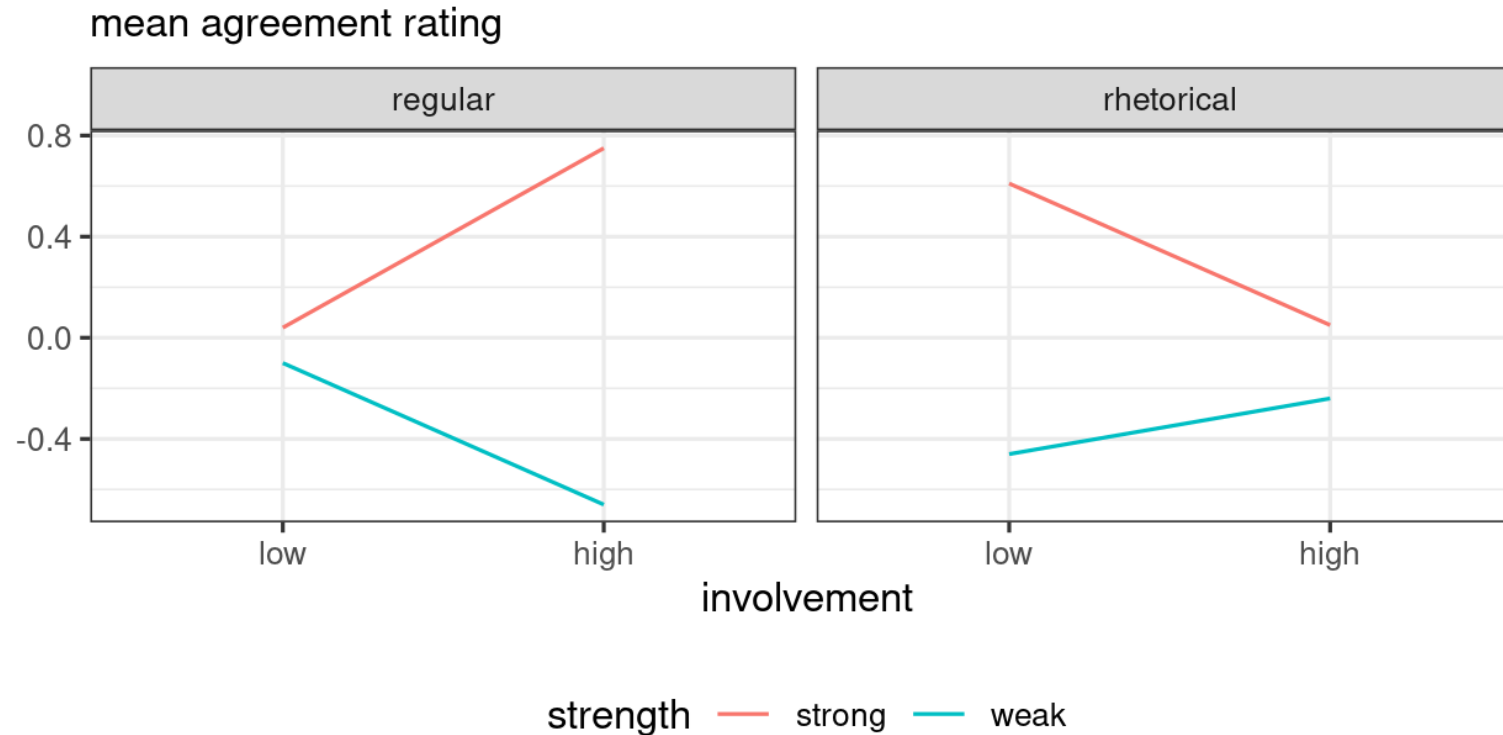
A $2 \times 2 \times 2$ factorial design with 8 treatments groups and $n = 160$ undergraduates.

Setup: should a comprehensive exam be administered to bachelor students in their final year?

- **Response** Likert scale on -5 (do not agree at all) to 5 (completely agree)
- **Factors**
- A : strength of the argument (strong or weak)
- B : involvement of students low (far away, in a long time) or high (next year, at their university)
- C : style of argument, either regular form or rhetorical (Don't you think?, ...)

Interaction plot

Interaction plot for a $2 \times 2 \times 2$ factorial design from Petty, Cacioppo and Heesacker (1981)



The microwave popcorn experiment

What is the best brand of microwave popcorn?

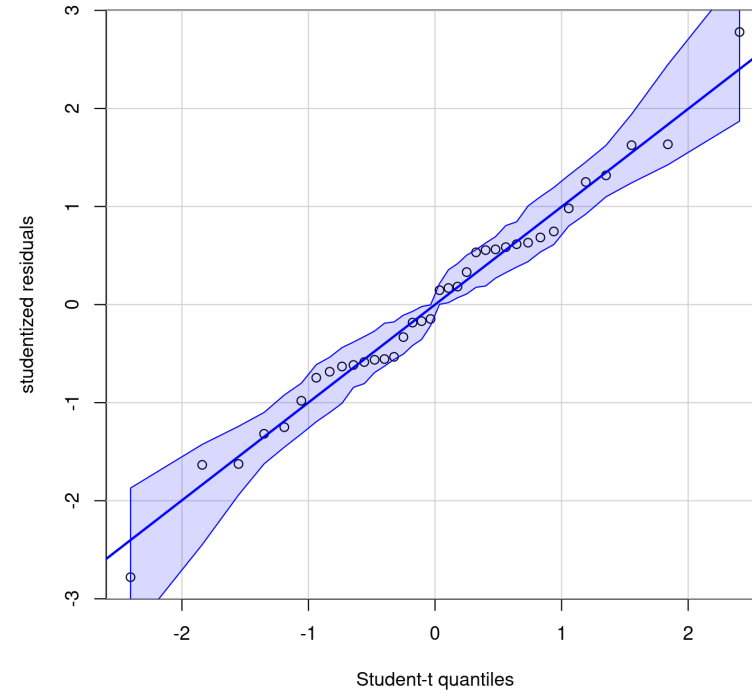
- **Factors**
- brand (two national, one local)
- power: 500W and 600W
- time: 4, 4.5 and 5 minutes
- **Response:** ~~weight, volume, number~~, percentage of popped kernels.
- Pilot study showed average of 70% overall popped kernels (10% standard dev), timing values reasonable
- Power calculation suggested at least $r = 4$ replicates, but researchers proceeded with $r = 2$...

ANOVA table

```
data(popcorn, package = 'hacedsm')  
# Fit model with three-way interaction  
model <- aov(percentage ~ brand*power*time,  
            data = popcorn)  
# ANOVA table - 'anova' is ONLY for balanced designs  
anova_table <- anova(model)  
# Quantile-quantile plot  
car::qqPlot(model)
```

Model assumptions: plots and tests are meaningless with $(n_r=2)$ replications per group...

Quantile-quantile plot

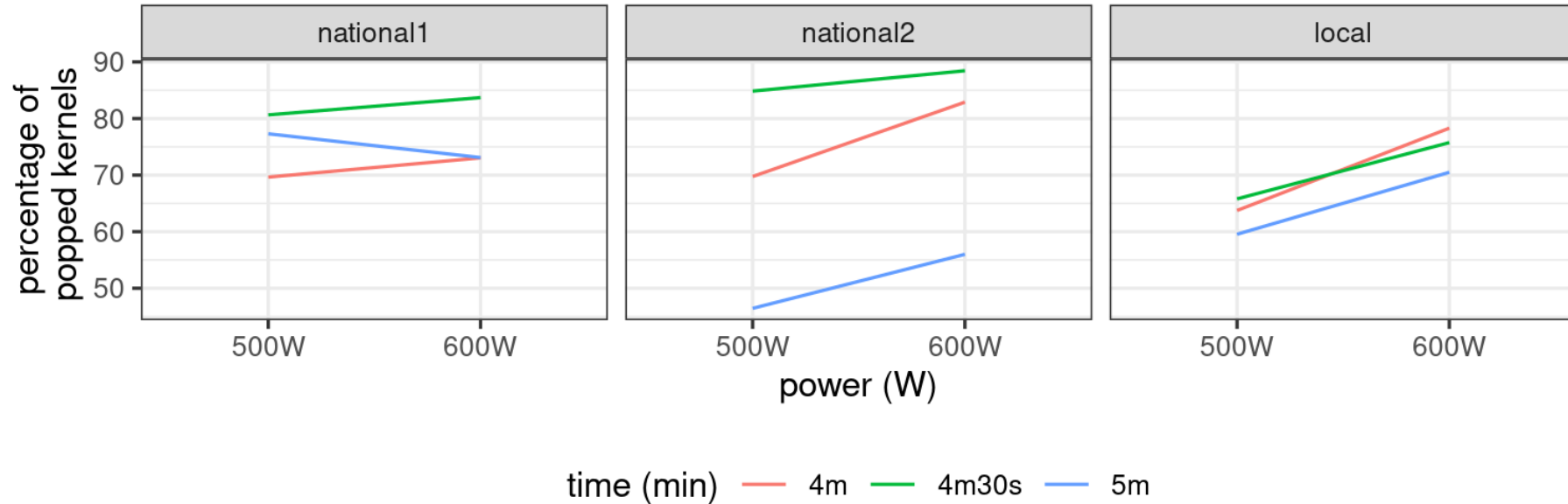


All points fall roughly on a straight line.

Code for interaction plot

```
popcorn |>
  group_by(brand, time, power) |>
  summarize(meanp = mean(percentage)) |>
  ggplot(mapping = aes(x = power,
                       y = meanp,
                       col = time,
                       group = time)) +
  geom_line() +
  facet_wrap(~brand)
```

Interaction plot



No evidence of three-way interaction (hard to tell with $r = 2$ replications).

Variance decomposition for balanced designs

terms	degrees of freedom
A	$n_a - 1$
B	$n_b - 1$
C	$n_c - 1$
AB	$(n_a - 1)(n_b - 1)$
AC	$(n_a - 1)(n_c - 1)$
BC	$(n_b - 1)(n_c - 1)$
ABC	$(n_a - 1)(n_b - 1)(n_c - 1)$
residual	$n_a n_b n_c (R - 1)$
total	$n_a n_b n_c n_r - 1$

Analysis of variance table for the 3-way model

	Degrees of freedom	Sum of squares	Mean square	F statistic	p-value
brand	2	331.10	165.55	1.89	0.180
power	1	455.11	455.11	5.19	0.035
time	2	1554.58	777.29	8.87	0.002
brand:power	2	196.04	98.02	1.12	0.349
brand:time	4	1433.86	358.46	4.09	0.016
power:time	2	47.71	23.85	0.27	0.765
brand:power:time	4	47.33	11.83	0.13	0.967
Residuals	18	1577.87	87.66		

Omitting terms in a factorial design

The more levels and factors, the more parameters to estimate (and replications needed)

- Costly to get enough observations / power
- The assumption of normality becomes more critical when $r = 2$!

It may be useful not to consider some interactions if they are known or (strongly) suspected not to be present

- If important interactions are omitted from the model, biased estimates/output!

Guidelines for the interpretation of effects

Start with the most complicated term (top down)

- If the three-way interaction ABC is significant:
 - don't interpret main effects or two-way interactions!
 - comparison is done cell by cell within each level
- If the ABC term isn't significant:
 - can marginalize and interpret lower order terms
 - back to a series of two-way ANOVAs

What contrasts are of interest?

- Can view a three-way ANOVA as a series of one-way ANOVA or two-way ANOVAs...

Depending on the goal, could compare for variable A

- marginal contrast ψ_A (averaging over B and C)
- marginal conditional contrast for particular subgroup: ψ_A within c_1
- contrast involving two variables: ψ_{AB}
- contrast differences between treatment at $\psi_A \times B$, averaging over C .
- etc.

See helper code and chapter 22 of Keppel & Wickens (2004) for a detailed example.

Effects and contrasts for microwave-popcorn

Following preplanned comparisons

- Which combo (brand, power, time) gives highest popping rate? (pairwise comparisons of all combos)
- Best brand overall (marginal means marginalizing over power and time, assuming no interaction)
- Effect of time and power on percentage of popped kernels
- pairwise comparison of time \times power
- main effect of power
- main effect of time

Preplanned comparisons using emmeans

Let A =brand, B =power, C =time

Compare difference between percentage of popped kernels for 4.5 versus 5 minutes, for brands 1 and 2

$$\mathcal{H}_0 : (\mu_{1.2} - \mu_{1.3}) - (\mu_{2.2} - \mu_{2.3}) = 0$$

```
library(emmeans)
# marginal means
emm_popcorn_AC <- emmeans(model,
                           specs = c("brand", "time"))
contrast_list <-
  list(
    brand12with4.5vs5min = c(0, 0, 0, 1, -1, 0, -1, 1, 0))
contrast(emm_popcorn_AC, # marginal mean (no time)
         method = contrast_list) # list of contrasts
```

Preplanned comparisons

Compare all three times (4, 4.5 and 5 minutes)

At level 99% with Tukey's HSD method

- Careful! Potentially misleading because there is a brand * time interaction present.

```
# List of variables to keep go in `specs`: keep only time  
emm_popcorn_C <- emmeans(model, specs = "time")  
pairs(emm_popcorn_C,  
      adjust = "tukey",  
      level = 0.99,  
      infer = TRUE)
```

Multivariate analysis of variance

Motivational example

From Anandarajan et al. (2002), Canadian Accounting Perspective

This study questions whether the current or proposed Canadian standard of disclosing a going-concern contingency is viewed as equivalent to the standard adopted in the United States by financial statement users. We examined loan officers' perceptions across three different formats

Alternative going-concern reporting formats

Bank loan officers were selected as the appropriate financial statement users for this study.

Experiment was conducted on the user's interpretation of a going-concern contingency when it is provided in one of three disclosure formats:

1. Integrated note (Canadian standard)
2. Stand-alone note (Proposed standard)
3. Stand-alone note plus modified report with explanatory paragraph (standard adopted in US and other countries)

Multivariate response

4. Please circle the pricing you would charge on borrowings under a line of credit *as a spread over your bank's base lending rate ("Prime rate")*.

0.25 0.50 1.00 1.25 1.50 1.75 2.00 2.25 2.50 2.75 3.00
3.25 3.50 3.75 4.00 Other _____

5. Please circle on the scale shown below your perception of *the ability of the company to service debt*.

LOW ABILITY

1

2

3

4

HIGH ABILITY

5

6. Please circle on the scale shown below your perception of the *likelihood that the company can improve its profitability*.

VERY UNLIKELY

1

2

3

4

VERY LIKELY

5

Why use MANOVA?

1. Control experimentwise error
 - do a single test instead of J univariate ANOVAs, thereby reducing the type I error
2. Detect differences in combination that would not be found with univariate tests
3. Increase power (context dependent)

Multivariate model

Postulate the following model:

$$\mathbf{Y}_{ij} \sim \text{Normal}_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}), \quad j = 1, \dots, J$$

Each response \mathbf{Y}_{ij} is p -dimensional.

We assume multivariate measurements are independent of one another, with

- the same multivariate normal distribution
- same covariance matrix $\boldsymbol{\Sigma}$ (each measurement can have different variance)
- same mean vector $\boldsymbol{\mu}_j$ within each $j = 1, \dots, J$ experimental groups.

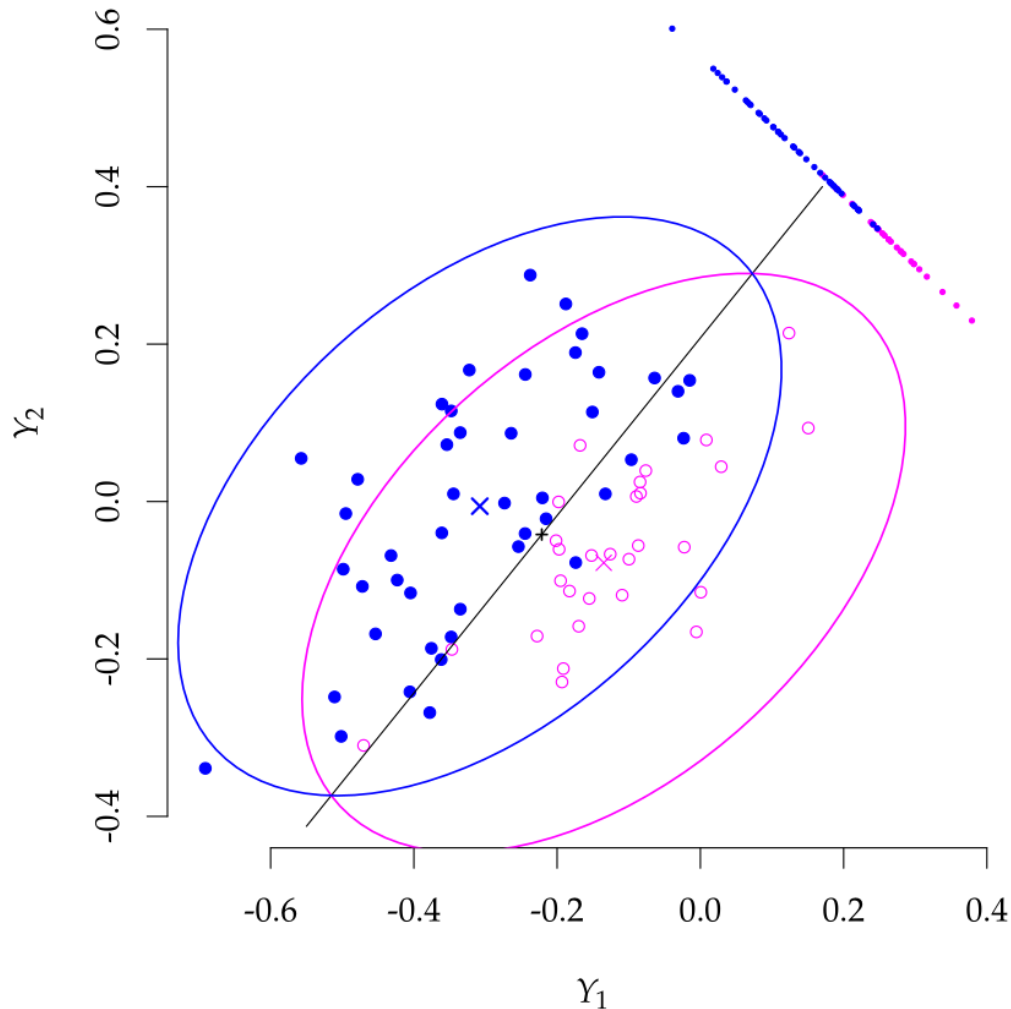
The model is fitted using multivariate linear regression.

Model fitting with multivariate response

In **R**, we fit a model binding the different vectors of response in a matrix with p columns

```
data(AVC02, package = "hecedsm")  
# Fit the model binding variables with cbind  
# on left of tilde (~) symbol  
modMANOVA <- manova(  
  cbind(prime, debt, profitability) ~ format,  
  data = AVC02)
```

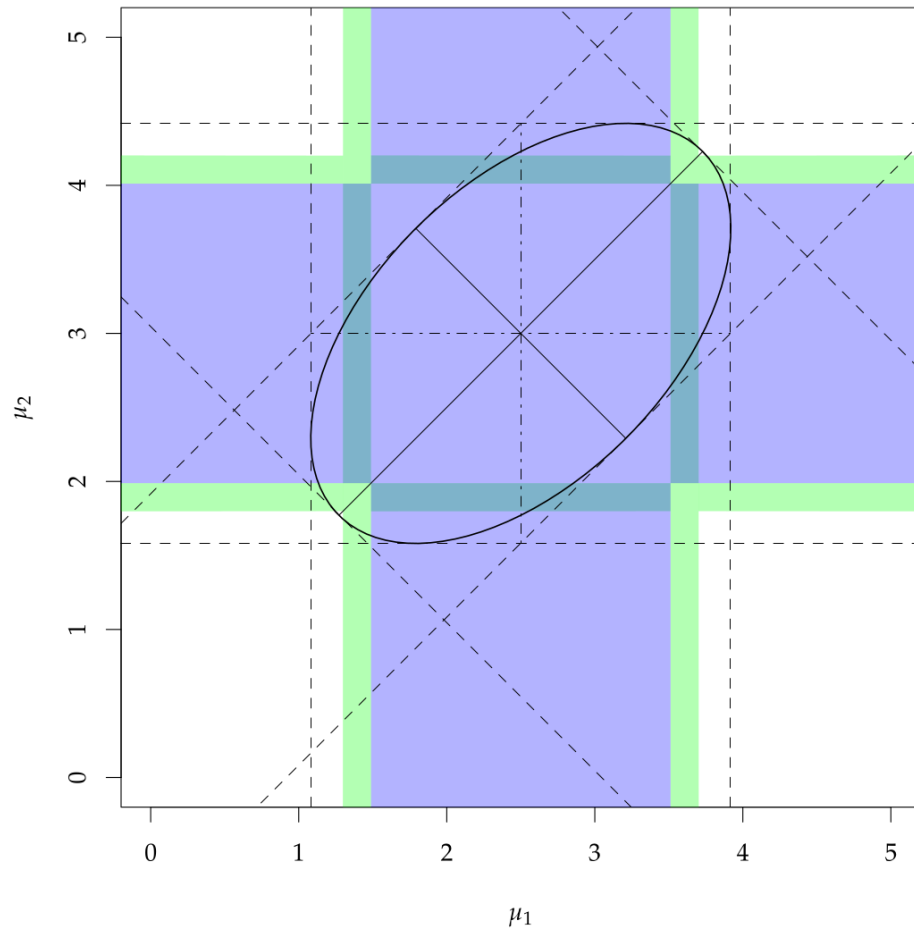
Bivariate MANOVA



Confidence ellipses for bivariate MANOVA with discriminant analysis.

We use the correlation between the p measurements to find better discriminant (the diagonal line is the best separating plane between the two variables).

Confidence intervals and confidence regions



Simultaneous confidence region (ellipse), marginal confidence intervals (blue) and Bonferroni-adjusted intervals (green).

The dashed lines show the univariate projections of the confidence ellipse.

Model assumptions

The more complex the model, the more assumptions...

Same as ANOVA, with in addition

- The response follow a multivariate normal distribution
 - Shapiro–Wilk test, univariate Q-Q plots
- The covariance matrix is the same for all subjects
 - Box's M test is often used, but highly sensitive to departures from the null (other assumptions impact the test)

Normality matters more in small samples (but tests will often lead to rejection, notably because of rounded measurements or Likert scales)

When to use MANOVA?

In addition, for this model to make sense, you need just enough correlation (Goldilock principle)

- if correlation is weak, use univariate analyses
 - (no gain from multivariate approach relative to one-way ANOVAs)
 - less power due to additional covariance parameter estimation
- if correlation is too strong, redundancy
 - don't use Likert scales that measure a similar dimension (rather, consider PLS or factor analysis)

Only combine elements that theoretically or conceptually make sense together.

Testing equality of mean vectors

The null hypothesis is that the J groups have the same mean

- $\mathcal{H}_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_J$ against the alternative that at least one vector is different from the rest.
- The null imposes $(J - 1) \times p$ restrictions on the parameters.

The test statistic is Hotelling's T^2 (with associated null distribution), but we can compute using an F distribution.

Choice of test statistic

In higher dimensions, with $J \geq 3$, there are many statistics that can be used to test equality of mean.

The statistics are constructed from within/between sum covariance matrices.

These are

- Roy's largest root (most powerful provided all assumptions hold)
- Wilk's Λ : most powerful, most commonly used
- **Pillai's trace**: most robust choice for departures from normality or equality of covariance matrices

Most give similar conclusion, and they are all equivalent with $J = 2$.

Results for MANOVA

```
summary(modMANOVA) # Pilai is default
```

```
##           Df  Pillai approx F num Df den Df Pr(>F)
## format      2 0.02581  0.55782      6   256 0.7637
## Residuals 129
```

```
summary(modMANOVA, test = "Wilks")
```

```
##           Df  Wilks approx F num Df den Df Pr(>F)
## format      2 0.97424  0.5561      6   254 0.765
## Residuals 129
```

```
summary(modMANOVA, test = "Hotelling-Lawley")
```

```
##           Df Hotelling-Lawley approx F num Df den Df Pr(>F)
## format      2      0.026397  0.55434      6   252 0.7664
## Residuals 129
```

MANOVA for repeated measures

We can also use MANOVA for repeated measures to get away from the hypothesis of equal variance per group or equal correlation.

```
model <- afex::aov_ez(  
  id = "id",           # subject id  
  dv = "latency",     # response  
  within = "stimulus", # within-subject  
  data = hecedsm::AA21,  
  fun_aggregate = mean)  
model$Anova # for models fitted via 'afex'
```

Output

```
##
## Type III Repeated Measures MANOVA Tests: Pillai test statistic
##           Df test stat approx F num Df den Df      Pr(>F)
## (Intercept)  1    0.95592   238.56      1    11 8.373e-09 ***
## stimulus     1    0.09419     0.52      2    10 0.6098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Less powerful than repeated measures ANOVA because we have to estimate more parameters. Still assumes that the covariance structure is the same for each experimental group.

Follow-up analyses

Researchers often conduct *post hoc* univariate tests using univariate ANOVA. In **R**, Holm-Bonferonni's method is applied for marginal tests (you need to correct for multiple testing!)

```
# Results for univariate analysis of variance (as follow-up)  
summary.aov(modMANOVA)  
# Note the "rep.meas" as default name  
# to get means of each variable separately  
emmeans::emmeans(modMANOVA, specs = c("format", "rep.meas"))
```

A better option is to proceed with descriptive discriminant analysis, a method that tries to find the linear combinations of the vector means to discriminate between groups. Beyond the scope of the course.