

Tutorial 5: Canonical Correlation Analysis

Canonical Correlation

Exercise 1: Butterfly Data Work through the butterfly example in the lecture note (where the two sets of variables are geographical and genetic). The R code is in the script accompanying the tutorial. The data is in the butterfly file `butterfly.dat`.

Exercise 2: Penguin Data Start by loading `tidyverse`. We'll use this for plots. We'll illustrate canonical correlation using the `cancor()` function in the base R stat package. The data is found in `palmer_penguins.csv` in the course data directory.

From this data set, prepare two datasets for CCA. Omit species/island (we'll see later how the canonical variables can be used for classification). Prepare two data sets for Canonical Correlation Analysis (CCA). Let X (the first data set) be `bill_depth_mm` and `bill_length_mm` and Y (the second data set) be `flipper_length_mm` and `body_mass_g`.

Perform a CCA. What are the coefficients (loadings) for the CCA? What are the correlations between the canonical variables in each pair? Remember to use scaled variables for the CCA. The relevant library is `CCA` and the function for canonical correlation is `cancor()`.

Now separate the pairs of canonical variates for each species. Verify that the correlation between each pair is the same when we take it species by species as when we take it overall.

The `mutate` command keeps the current columns of a data frame the same and adds new columns. Hence the `mutate` command used in the script adds columns we have created to the data frame.

Make a scatter plot of the first pair of canonical covariates.

Plot the first canonical variate of X against species.

Plot the first canonical variate of Y against species.

Make a scatter plot of the first pair of canonical variates, coloured by species.

Make a scatter plot of the second pair of canonical variates, coloured by species.

What are the correlations between the pairs of canonical variables?

Exercise 3: Psychological Traits and Academic Features The data in `mmreg.csv` on the course page contains 8 variables. The first three measure psychological features, while the remaining 5 relate to academic ability. We want to see how these two sets are related.

Perform a canonical correlation analysis.

Now test to see how many of the canonical correlations are significant. This does not seem to be available in any package, hence the test has to be programmed by hands. A script is available in the script to go with the tutorial. Perform and report the significance tests.

Finally, standardize the canonical coefficients for `psych` and `acad`.