MADR 2025 - Project tasks 5

Canonical Correlation Analysis - Problem 1

(From: Richard Wilkinson online notes, https://rich-d-wilkinson.github.io/MATH3030/) Consider a football league table from https://www.rotowire.com/soccer/league-table.php (pick standings for the 2024 Premier League season). The table contains, in particular, the variables: W, D, L are the number of matches won, drawn and lost and G and GA are the goals scored for and against, and GD is the goal difference (G−GA). We will treat W and D, the number of wins and draws, as the x-variables. The number of goals for and against, G and GA, will be treated as the y-variables. The number of losses and the goal difference, L and GD, are omitted as they provide no additional information when we know W and D.

- Read the analysis from Richard Wilkinson online notes (5.1.2 Example: Premier league football).
- Perform similar analyses for the 2024 Premier League season data.

Canonical Correlation Analysis - Problem 2

(From: Richard Wilkinson online notes, https://rich-d-wilkinson.github.io/MATH3030/) The **crabs** (R package **MASS**) data frame has 200 rows and 8 columns, describing 5 morphological measurements on 50 crabs each of two colour forms and both sexes, of the species *Leptograpsus variegatus* collected at Fremantle, W. Australia. We will focus on the 5 continuous variables, all measured in mm:

- FL = frontal lobe size,
- RW = rear width,
- CL = carapace length,
- CW = carapace width,
- BD = body depth.

Consider a canonical correlation analysis in which one set of variables, the x-set, is given by CL and CW and the other set, the y-set, is given by FL, RW and BD.

- Calculate $S_{XX}^{-1/2}$ and $S_{YY}^{-1/2}$ by first computing the spectral decomposition of S_{XX} and S_{YY}.
- Calculate the matrix Q and compute its singular value decomposition.
- Compute the first pair of CC vectors and CC variables, say η_1 and ψ_1. What is the first canonical correlation?
- Plot ψ_1 vs η_1. What does the plot tell you (if anything)?
- Repeat the above to find the second pair of CC vectors, and the second set of CC variables/scores, and plot these against each other and against the first CC scores. Is there any interesting structure in any of the plots? Which plots suggest random scatter?
- Finally, repeat the analysis above using a trusted package, that enables the CCA, in your chosen programming language.

Canonical Correlation Analysis - Problem 3

The olive oil data from the **olives** dataset (R package **classifly**) consists of the percentage composition of 8 fatty acids found in the lipid fraction of 572 Italian olive oils (see the description in the R documentation for the **classifly** package). Variable *region* takes values from $\{1, 2, 3\}$ and indicates the region of origin. The variables from *palmitic* to *eicosenoic* measure the percentage composition of 8 different fatty acids.

- Using the CCA methodology, examine the correlations between the region of origin and the fatty acid measurements. That is, take $X \in \mathbb{R}^{572\times 8}$ to contain the fatty acid measurements, and $Y \in \{0,1\}^{572\times 3}$ to be the matrix each row of which indicates the region with a 1 and otherwise has a 0. So, region 1 is coded by the row $(1, 0, 0)$ in Y, region 2 by $(0, 1, 0)$, and 3 by $(0, 0, 1)$.
- Plot the second CC variate Xa_2 as a function of the first CC variate Xa_1 (use different colors to distinguish between the points corresponding to the three regions).