Homework: Canonical Correlation Analysis

High Dimensional Data Analysis 2020

Adapted by Milan Malfait

20 Oct 2020

Canonical Correlation Analysis (CCA) is a multivariate data analysis method that aims at finding correlations between two multivariate data sets, X and Y. The method looks for the linear combination of the X-variables and the linear combination of the Y-variables that show maximal correlation. When the number of variables in X and/or Y is very large (high-dimensional), the classical CCA method needs to be adapted to deal with the high dimensionality.

The aim of this homework assignment is:

- to understand the classical CCA method (based on the literature) and a CCA method for high-dimensional data
- to implement the CCA method and its high-dimensional version (not using existing R packages or R functions for CCA)
- apply the method to a dataset

You may consult the literature to find a description of the CCA method. Here I give one possible reference (it is a paper about an R package, but remember that you may not use this R package for the implementation):

González, I., Déjean, S., Martin, P. G., & Baccini, A. (2008). CCA: An R package to extend canonical correlation analysis. Journal of Statistical Software, 23(12), 1-14. http://dx.doi.org/10.18637/jss.v023.i12

The paper also describes a *regularised CCA* method (section 2.4), which is applicable to high-dimensional data. However, there are other high-dimensional CCA methods described in the literature. You are free to choose the regularised CCA from the paper, or any other appropriate high-dimensional CCA method.

Note that in the paper a cross-validation method is proposed for selecting e.g. the tuning parameters in the regularised CCA; you should not implement this. If tuning parameters are involved, you may set them manually to an arbitrary value (or play with it when analysing the dataset and set it to a value that seems appropriate to you – no need to motivate your choice).

You must apply your implemented method to the **nutrimouse** data, which is part of the **CCA** R package. More information about the data can be found in the paper. You must only look at the first two dimensions of the CCA, which will allow you to make two-dimensional graphs.

```
# install the CCA package with
# install.packages("CCA")

library(CCA)
data("nutrimouse")
```

```r
X <- nutrimouse$gene  # the gene expression matrix
dim(X)
#> [1]  40 120
Y <- nutrimouse$lipid # the lipids matrix
dim(Y)
#> [1] 40 21
```

The assignment can be done in **groups of 2**.

You should write a report containing the following:

- a short (mathematical) description of the CCA methods (classical and high-dimensional) that you have implemented
- the application of your method to the *nutrimouse* data
  - Classical CCA on multivariate data with $p < n$. (*Hint: it will not be possible to apply the classical CCA method to the full data matrix X. You should subset the data to reflect the case of $p < n$.*)
  - High-dimensional CCA on data with $p > n$
- interpretation and conclusion of the data analysis results

The length of the written report (excl. R code, R output and graphs) should be about 2 pages.

It is recommended (but not mandatory) to prepare your report in **RMarkdown**. You can render it to either HTML (`output: html_document`) or to PDF (`output: pdf_document`). In both cases the original `.Rmd` file should be included when handing in the assignment. If you don't use RMarkdown, you should include the `.R` file(s) containing your implementation and analysis scripts.

When submitting, please use the following format:

- HW-Name1-Name2.[pdf|html]
- HW-Name1-Name2.Rmd (or HW1-Name1-Name2.R)

Submissions should be done **through UFora**.

The deadline for submission is November 8th.