Assignment 3 PCA, PCR, NIPALS

To be completed in groups of UP TO THREE (3)
Due: March 13, 2023 @11:59pm.
Grading: 5% of course grade (90 points available)

Submission Instructions

Please submit this assignment *electronically* before the due date. Late submissions will **not** be accepted. Submit via the A2L dropbox for the appropriate assignment. Be sure that you have the names and student numbers of all students on the front page of your submission. Submit your answers as a **single .pdf** file including all relevant figures, tables, and math. You may include relevant code embedded in the report, but you **must submit a .zip along with your report** that includes all your code for the assignment.

Please upload your files with the **naming convention** `A0X_macID1_madID2_macID3`, where the first McMaster ID is for the person uploading the submission.

Up to 10 points may be deducted from your submission for sloppy or otherwise unprofessional work. This is rare, but possible. The definition of unprofessional work may include:

- Low-resolution screenshots of figures and tables.
- Giving no context to an answer relating to the task (*i.e.* "See code" or "113.289" with no units, context, or discussion whatsoever).
- Clear changes in author denoted by format changes, blatant writing style changes, or other factors that may deduct from the cohesiveness of the report.
- Failing to provide references for work that is obviously not yours (particularly bad cases will be considered as academic dishonesty).

Problem 1 [30]

Consider the dataset `silicon-wafer-thickness.csv` from the companion for this assignment. This dataset contains the thickness of a sample silicon wafer, measured at nine locations, from 184 consecutive batches. The locations of the nine thicknesses are shown in Figure 1 below:

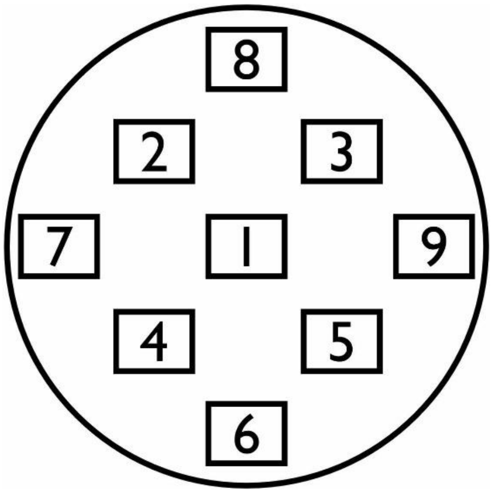

Figure 1. Locations of the nine points where the wafer thickness is measured.

It is our objective to build a "soft sensor" for a manufacturing line producing these chips. Our current data set is going to be used to build the model, which will then use modeling statistics such as Hotelling's T^2 and SPE to flag chips that are either too thick or too thin (via T^2) or have a particular measurement that is an outlier (such as measurement 1 being much different than measurement 7, which means we have an "uneven" chip). Outlier measurements will be flagged as having a high SPE.

1. Your first task is to build a PCA model for the `silicon-wafer-thickness.csv` data set. However, we would first like to know **how many components are required** to effectively represent the variance contained in the original nine columns.

   Build PCA models for this data set using 1-5 components (5 models total). Use the provided NIPALS code from assignment 2 to simplify your calculations. To ensure that the components are "justified," perform cross-validation with G = 4 groups. This should result in four training sets of 138 rows and corresponding testing sets of 46 rows. To get you started, I have provided a code called `Q1_Starter.m` to begin your assignment. It centers/scales, shuffles, and separates the data set into G = 4 sub-groups for you automatically. Feel free to continue your code from there.

   For each number of components from 1-5, report the (cumulative) R^2 and Q^2 for the model with G = 4 (remember that R^2 is calculated automatically for you by `nipalspca.p`). You **DO NOT** need to calculate the Q^2 or R^2 for each column individually, just overall. Argue based on Q^2 that the 5th component does not explain enough variance to be included in the model, and thus our model should contain A = 4 components. [**10**].

2. OK, now that we have nailed down that 4 components are appropriate, we'll focus on that model. For your PCA model, create bar plots of the **first and second** component loadings. Comment on what you see. Does one measurement appear to have uncorrelated variance with the rest? [**3**]

3. Use the provided score-plot function to plot the scores for the first two components and the 95% / 99% confidence interval ellipses (the provided function should make the ellipses for you). Do not plot the loadings vectors (they will be messy). Briefly comment on any outliers on the score plot and how they might be influencing your model. [**3**]

4. Compute Hotelling's T^2 for each observation. Also compute the 95% and 99% confidence levels for Hotelling's T^2. Produce a line plot of T^2 vs observation number along with the 95% and 99% confidence limits as horizontal lines. Are the outliers present on this plot as well? [**5**]

5. Produce a line plot of SPE vs observation number along with the 95% and 99% confidence limits. Are there any major outliers in this dataset according to SPE? Note that you can produce your own confidence limits for SPE using the equation below. [**5**]

$$SPE_\alpha = \left(\frac{v}{2m}\right)\chi^2_{\left\{\frac{2m^2}{v},\alpha\right\}}$$

   Where m and v are the sample mean and variance of the SPE vector, respectively. $\chi^2_{\left\{\frac{2m^2}{v},\alpha\right\}}$ is the critical value from the chi-squared (χ^2) distribution with $\frac{2m^2}{v}$ degrees of freedom at confidence level α. This can be computed for you in `MATLAB` using the function `ch2inv`. It might also help to note that the given NIPALS algorithms compute the residuals for the model, although it does not return them.

6. Remove any major outliers in the dataset and rebuild the PCA model **ONCE**. Re-plot the T^2 and SPE plots along with their confidence limits at 95% and 99%. If you see additional outliers, comment on why this may be the case. [**4**]

Problem 2 [15]

Consider the dataset `peas.csv` from the data companion. This dataset contains 17 columns. The first 11 columns can be considered **X** and include laboratory measurements for 60 samples of frozen peas for:

- Tenderness (col 1)
- Dry matter before and after freezing (cols 2-3)
- Sucrose content (col 4)
- Two measures of glucose content (cols 5-6)
- The paleness of the pea skin (col 7)
- Three normalized colour measurements (cols 8-10)
- A measure of skin consistency (col 11)

The last six columns are **Y**; a set of ratings on a scale from 1-9 from a panel of judges for the peas according to the following categories:

- Overall flavour (col 12)
- Sweetness (col 13)
- Fruitiness (col 14)
- Off-flavour (col 15)
- Mealiness (col 16)
- Hardness (col 17)

Our goal for this problem is to build a model that predicts the scores for a pea sample based on its laboratory measurements (kind of like an electronic taste-tester, if you will). We will do this using PCR.

1. Create a scatterplot matrix of **X**. Based on this plot, argue why fitting a multi-linear regression to this data set to predict any column in **Y** using **X** is a bad idea. [**2**]

2. Fit a 2-component PCA model to **X**. Note that we should probably have more components, but we will use two for visualization purposes. Show the loadings plots for the first two components and comment on any trends. [**3**]

3. Using the scores **T** from your PCA model, create a PCR model to predict the FLAVOUR rating (first column in **Y**) as a function of $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$. You do not have to reverse centering and scaling for **Y**. Plot the data ($\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{y}$) as a 3-D scatterplot and display the model equation as a surface. [**5**]

4. Based on the coefficients and loadings, what qualities (in **X**) would you associate with higher rated peas? Explain how you know. [**3**]

5. Create a scatterplot matrix of **Y**. Identify any trends or interesting features. Is it necessary to create linear regressions for the other five qualities? Briefly explain your thinking. [**2**]

Problem 3 [10]

Develop a script for the NIPALS algorithm for PLS. Your function should take in X, Y, and the number of components to fit and output the scores (t, u) and loadings (w*, c, p) for X and Y along with the R^2 for each component. Confirm your function works by comparing w*, c, and R^2 for the peas data set from problem 2 to the REAL values provided in `peaspls.xlsx`. Use A = 3 components. If there is a difference in the loadings or scores, mention why you think that is. *Note that the answers are given assuming the rows in X are left in order as provided*. You do not need to perform any sort of cross-validation. You do not need to handle missing data.