PROBLEM SET 2

SUBMIT BY 6 NOVEMBER 2024

INSTRUCTIONS

The assignment should be submitted on Canvas by **5:00pm on 6 November 2024** as an `ipynb` file (i.e. a Jupyter notebook).

The ipynb should be machine readable, and it must reproduce your answers when I run it. Develop your own Python code rather than simply using existing Python modules for computing the relevant mathematics (e.g. `statsmodels`). If in doubt, send me an email. The ipynb file should contain your answers to the questions raised below in markdown (or "text") cells and not Python cells.

The use of generative AI is prohibited and will be filed, in line with the university's policy, as academic misconduct. Read more here:

https://www.universityofgalway.ie/academicintegrity/.

The homework will be graded according to a scheme in which *content* (i.e. correctness of your answers, choice of methods, python code) is weighted at 80% and *presentation* (i.e. manner in which you present your answers, methods, and code) is weighted at 20%.

Late submissions will be accepted at a 10% deduction every 24 hours past the deadline, starting with 10%. For example, submitting two hours after the deadline on Canvas will incur a 10% deduction, whereas 26 hours after the deadline will incur 20%.

PROBLEMS

Problem 1. Let M be an $n \times n$ real symmetric matrix. Prove that if U is an M-invariant subspace of $\mathbb{R}^n$, then $U^\perp$ is M-invariant.

Problem 2. Create a Python class initialised by a pandas dataframe; that is, `__init__` should have exactly two inputs. There should be a number of methods associated to this class (one for each item in the list), and they should return the following:

- The covariance matrix.
- The (ordered) principal components.
- The (ordered) eigenvalues of the covariance matrix.
- The "new" data after projecting onto the ith and the jth principal components respectively. Thus, the input should include i and j, where $i, j \in \{1, \dots, m\}$ (data points in $\mathbb{R}^m$). The data type of the output should be a pandas data frame with columns labeled "PCk" where k is replaced by the appropriate integer (e.g. "PC7" and "PC42").

You should also write `__len__` and `__repr__` methods.

You do not need to rescale data; you do not need to type check or raise errors. You may assume reasonable input. It is up to you to name things—please keep it appropriate as if it were to be published for public use. You may define attributes and additional methods, but they will not be considered when marking.

Problem 3. Perform PCA on the data set in '`UN_IRE_data.csv`'.

(1) Compute the fewest principal components, so that more than 90% of the variability is preserved. How many are needed, and what is the percentage of variability?
(2) Plot the projection of the data on the first two principal components.

Feel free to use the Python class from Problem 2.
(Data source: United Nations – https://unstats.un.org/sdgs/dataportal/countryprofiles/IRL)

Date: October 23, 2024.