Problem 3: Daily temperature curves

The file `temperature.txt` contains hourly temperature measurements (in degrees Celsius) recorded in Milan throughout the year 2023. Each row corresponds to one of the 365 days, and each column represents one of the 25 hourly time points from 00:00 to 24:00 (`h0` to `h24`). Additionally, each day is labeled according to the season (`Winter`, `Spring`, `Summer`, or `Autumn`) through the variable `season`.

We take a functional data analysis perspective, treating the hourly temperature profiles as discretized evaluations of smooth functions defined over the domain $[0, 24]$.

a) Apply penalized smoothing to the temperature data using a basis of quadratic B-splines, with knots placed at each observed time point (i.e., every hour). Penalize the first-order derivative and use a smoothing parameter of $\lambda = 1$. Report the number of splines used and the mean generalized cross-validation (GCV) error. Provide a plot of the smoothed temperature curves. What is the approximate dimension of the space in which the smoothed curves live?

b) Conduct a functional principal component analysis (FPCA) on the smoothed functions. What proportion of the total variance is explained by the second principal component? From a dimensionality reduction perspective, how many principal components would you retain? Justify your choice.

c) Provide a plot showing the effect of the second principal component.

d) Is the representation given by the first principal component satisfying for distinguishing seasons? Support your answer with a plot.

e) Could we successfully classify seasons based on the representation given by the second principal component? Justify your answer.

Upload your results here: `https://forms.office.com/e/1zpRAZBXsF`