Math 420, Spring 2024
Third Solo Homework: PCA and ODE Modeling

Exercise 1. Consider the dataset included in the Excel file 'ReducedTimeSeries_DCMDVA.xlsx' attached to this homework. (Your homework refers to the first spreadsheet, in case there are multiple sheets.)

Description of this data set: it is an array of size 160 x 211. Each row represents one time series of cumulative Covid-19 cases detected in a certain county in the Tri-State DC-MD-VA area. The first row is the number of cases in DC, the next 25 rows (lines 2 through 26) contain the number of cases in the 25 counties of Maryland, and the following 134 rows are the time series associated to the 134 counties in Virginia. Each time series starts on March 2, 2020 (first clumn), and ends on September 28, 2020 (last column). Note the first column is identical 0, since no cases was detected in the Tri-State area by March 2. Data has been extracted from the JHU Coronavirus Resource Center and available on GitHub repository: https://github.com/CSSEGISandData/COVID-19 . This data repository is maintained by the Center for Systems Science and Engineering (CSSE) at JHU.

(1) Preprocessing: Load this data set in your code and compute the daily rates of Covid-19 detection. To do so you need to do the following: First load this data set in a matrix, say A of size 160x211. Then transpose into matrix $B = A'$ of size 211x160. Finally, compute the matrix X of size 211 x 160 of daily rates, $X(t,k) = B(t,k) - B(t-1,k)$ for each entry except the first row that you initialize with 0 in X. Each column of X represents the 211-point time series of daily rates, for one county (or district).

(2) Implement the 4 PCA Algorithms 1, 2, 3 and 4, and apply on this data set: $x_1, \ldots, x_{160} \in \mathbb{R}^{211}$. Specifically, for each algorithm:
   (a) Plot the distribution of squared singular values σ_k^2. Find the smallest d so that the ratio of explained variance is $0.9 = 90\%$ or higher. Print d and the ratio of explained variance.
   (b) For the specific d obtained above, take the first column vector x_1, the Covid detection daily rates in DC, and project it on the best linear (or affine) subspace of dimension d. This means to find $\hat{x}_1$. Plot both x_1 and $\hat{x}_1$ in the same figure.
   (c) Plot the first 6 normalized eigenvectors (or singular vectors) in separate figures.
   (d) Compute and print $\frac{\|x_1 - \hat{x}_1\|^2}{\|x_1\|^2}$. All norms are Euclidean (i.e., 2-norm).

Exercise 2 on the next page.

Exercise 2. Consider the following initial value problem (IVP):

$$\begin{aligned} \frac{dx_1}{dt} &= -\frac{1}{2}x_1 \ , \ x_1(0) = 2000 \\ \frac{dx_2}{dt} &= \frac{1}{2}x_1 \ , \ x_2(0) = 23 \end{aligned}$$

a. Determine the exact solution $(x_1(t), x_2(t))$ of this problem, i.e., solve the IVP.
b. Implement the Euler scheme to find numerically $(x_1(T_{max}), x_2(T_{max}))$ at $T_{max} = 2$. Use the following values for the step size h: 0.1 , 0.01, 0.001, 0.0001, 0.00001. For each value of h compute the numerical error $E(h)$, i.e. the 2-norm of the difference between the exact solution at time T_{max} determined at part a) and the numerical estimate your code returns.
c. Estimate the rate $p > 0$ for the power law $E(h) = Ch^p$ by linear interpolation in the log-log scale. This means: let $t = log(h)$ and $y = log(E(h))$. Find the least-squares linear fit $\hat{y} = \beta_1 t + \beta_0$ for the data points estimated at part b), that is of the points $(log(h), log(|E(h)|))$, for $h \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. Then $C = e^{\beta_0}$ and $p = \beta_1$.
d. For a time-step $T_0 = 0.01$ implement an agent-based simulator to solve this IVP and obtain the solution at T_{max}. Run 100 times this simulator and compute the mean and variance of the estimated solution at T_{max}. That is, compute $\mu_k = \mathbb{E}[x_k(T_{max})]$ and $\sigma_k^2 = Var(x_k(T_{max}))$, for $k = 1, 2$.