National University of Singapore
DSA5206 Advanced Topics in Data Science

Project

Instructions

- Answer all questions and number your answers accordingly. This project is **individual** work, so please complete the report and code on your own.
- Please submit your work in PDF format on canvas. The report must be typed (not hand-written). The file format is

  studentnumber_yourname_project.pdf

  Only one pdf file should be submitted, including parts 1 and 2, and source code as an appendix.
- Late policy: -20% of the total grade per day late.

Part 1 (25 points)

Let us consider a continuous-time, 2D state space model (or equivalently, a 2D control system) for a single-input single-output (SISO) system

$$\frac{d}{dt}\begin{pmatrix} h_1(t) \\ h_2(t) \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -10 \end{pmatrix}\begin{pmatrix} h_1(t) \\ h_2(t) \end{pmatrix} + \begin{pmatrix} 10^{-3} \\ 10^3 \end{pmatrix} x(t), \tag{1.1}$$

$$y(t) = \begin{pmatrix} 10^3 & 10^{-3} \end{pmatrix}\begin{pmatrix} h_1(t) \\ h_2(t) \end{pmatrix} \tag{1.2}$$

Let us apply some dimensionality reduction techniques we have learned in class to this system.

(a) First, set $x(t) = 0$ for all t. If we apply dynamic mode decomposition (DMD) to eq. (1.1), what is the leading dynamic mode? This should be a vector in $\mathbb{R}^2$. You do not need to numerically compute the DMD.

(b) Next, discretise the dynamics (1.1) in time (using the Euler method) with a time step of $\Delta t = 10^{-3}$. Sample the inputs $x(t) \sim \mathcal{N}(0, 1)$, independently and identically distributed for each t. You may set $h_1(0) = h_2(0) = 0$. Produce a total of 10^3 trajectories (with different samples of $\boldsymbol{x}$), each with 10^3 time steps so that the time interval is $t \in [0, 1]$. Perform proper orthogonal decomposition (POD) on this dataset via time-averaging. What is the leading spatial POD mode? Again, this should be a vector in $\mathbb{R}^2$.

(c) Discuss if it is appropriate to use the leading DMD or the leading (spatial) POD mode as a means to reduce the system eq. (1.1) from 2 to 1 dimension.

Now, let us consider a general continuous time state space model

$$\dot{h}(t) = Ah(t) + Bx(t) \tag{1.3}$$

$$y(t) = Ch(t) + Dx(t), \tag{1.4}$$

where $h(t) \in \mathbb{R}^m$ is the hidden state, $x(t) \in \mathbb{R}^n$ are the inputs and $y(t) \in \mathbb{R}^p$ are the outputs.

Recall that *controllability* is defined as the possibility to steer the system eq. (1.3) from 0 to any final state in finite time. Mathematically, this means that for any $h_* \in \mathbb{R}^m$, there exists a $t_* \geq 0$ and an input sequence $\boldsymbol{x}$ such that $h(t_*) = h_*$ under this input. We saw in class that for discrete SSMs, controllability is equivalent to

$$R = \begin{pmatrix} B & AB & \cdots & A^{m-1}B \end{pmatrix} \text{ has full column rank } m. \tag{1.5}$$

(d) Show that for the continuous-time dynamics in eq. (1.3), the controllability condition is also given by eq. (1.5).
[Hint: as in the discrete case, use Cayley-Hamilton theorem and the fact that the matrix exponential e^{At} can be expanded as $e^{At} = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k$.]

However, a major shortcoming of the criterion (1.5) is that it is either true or false, and does not provide a measure of how controllable the system is. This is where the *controllability Gramian* comes in. For the system (1.3), the controllability Gramian is defined as the time-dependent $m \times m$ matrix

$$W_c(t) = \int_0^t e^{As} B B^\top e^{A^\top s} ds. \tag{1.6}$$

(e) Show that if the controllability condition (1.5) is satisfied, then the controllability Gramian $W_c(t)$ defined in (1.6) is symmetric and positive definite for every $t > 0$.
[The converse is also true but we will not prove it here.]

(f) We hereafter assume that the system in eq. (1.3) is controllable. Let us fix $t_* > 0$ and $h_* \in \mathbb{R}^m$. Define

$$x_*(t) = B^\top e^{A^\top (t_* - t)} W_c(t_*)^{-1} h_* \tag{1.7}$$

Show that under the input in eq. (1.7), the state satisfies $h(t_*) = h_*$. This gives an explicit input (or control signal) to drive the system from 0 to h_* in time t_*.

(g) Now, consider the "energy" of this input, defined by

$$E = \int_0^{t_*} |x_*(s)|^2 ds. \tag{1.8}$$

Show that $E = h_*^\top W_c(t_*)^{-1} h_*$. This means that if W_c has small eigenvalues, then the "work" required to steer the system along the direction of the corresponding eigenvectors will be very large. Hence, the eigenvalues of W_c give a measure of how "easy" it is to control the system.

Let us return to the system eq. (1.1)–eq. (1.2).

(h) Compute numerically the controllability Gramian $W_c(t)$ for this system at $t = 1$, and compute its eigenvalues.

(i) Now, consider the transformation

$$\tilde{h}(t) = Th(t), \quad \text{where } T = \begin{pmatrix} 10^3 & 0 \\ 0 & 10^{-3} \end{pmatrix} \tag{1.9}$$

Compute the controllability Gramian $\tilde{W}_c(t)$ for the transformed SSM at $t = 1$ and its eigenvalues. What do you observe?

The above transformation is a special case of a general class of balanced model reductions for control systems, which aim to transform the system into a form where both the controllability and observability Gramians are diagonal and equal, making the truncated systems more representative of the input-output relationship.

Part 2 (25 points)

The second part is an open-ended project. You will pick a topic that interests you, perform a analysis of the problem, and write a report of your findings. To ensure some relevance, your topic of choice should fulfil the following criteria:

- Your problem or dataset should involve dynamical processes (they can be time-series, spatial-temporal data, sequence data, etc).
- You should apply at least two of the four types of methods we have learned in class: system identification, dimensionality reduction, prediction/generation, and control. You may also apply methods outside of those introduced in class, provided that a) they fall into one of the four categories and b) you explain the methods clearly with appropriate literature references.

Example problem statements include dimensionality reduction and prediction of time-series data, reduced order modelling for partial differential equations, generative models for image/video using diffusion models, etc.

Your report to this part should be structured as follows, each being a separate section.

1. Introduction: describe the problem and the dataset.
2. Methods: describe the methods you apply to analyse the dataset.
3. Results: present your findings with appropriate visualisations and tables. discuss any interesting results.
4. Conclusion: summarise your findings, discuss any limitations of your analysis and possible future work.
5. References: cite any relevant literature.

Your report for part 2 should not exceed 6 pages (excluding references and attached code). The report will be graded for adherence to the instructions, clarity, correctness in the application of the methods and creativity.