Data-driven Physical Modelling (SEMTM0007) Coursework – Part 1

Submission deadline: 1pm Thursday 4 December 2025

1 Coursework description

Formatting requirements The coursework consists of two parts. This is part 1, part 2 is in a separate document. For part 1 of the coursework you need to submit two files:

1. A PDF report containing a combination of Python code, numerical results and relevant figures, and insights / conclusions written as prose. In most cases, this should be your Jupyter notebook that has been converted to PDF. Ensure that irrelevant output is hidden (e.g., logging outputs from training).

2. A Jupyter notebook (.ipynb) or Python script (.py) containing all the code used to generate the results in the PDF file. Ensure that all required dependencies are documented — you can assume that NumPy, SciPy, Matplotlib, and PyTorch are available. This code will be run to ensure that the results are reproducible; to minimise any randomness in training, set the random seeds in PyTorch with torch.manual_seed(seed) and in NumPy with np.random.seed(seed), where seed is an integer of your choice. Exact reproducibility is not expected due to differences in hardware implementations. Your notebook will be automatically tested and converted to a pdf using the command
   `jupyter nbconvert --to pdf --execute <mynotebook.ipynb>`
   We will look at your pdf if the automatic conversion fails.

Whilst there is no page limit, your PDF report should be concise and to the point, avoiding extraneous code outputs. You should aim to write in a clear and professional manner, using appropriate technical language. Excessive explanations ("filler") will decrease your mark.

Assessment of your reports is not a tick-box exercise, and therefore we do not attach marks to each question. We use the university generic marking criteria for M-level units.

Intended learning outcomes From the unit catalogue:

1. Choose an appropriate data-driven modelling framework that aligns with the problem specification and provides the required accuracy and/or interpretability of the model.

2. Demonstrate mastery of a data-driven modelling technique through the analysis of a synthetic or experimental data set.

3. Use appropriate techniques to generate and prepare data for modelling purposes

4. Assess the quality of a mathematical model using metrics such as accuracy, repeatability, and generalisation

Marking criteria A pass mark (50) is achieved by demonstrating all intended learning outcomes across the two parts of the coursework. The relevant aspects of how the learning outcomes are achieved from the generic marking criteria are

1. **Content knowledge.** Whether you can make use of methods and skills explicitly taught. The precision and difficulty of skills demonstrated differentiates the marks. (20 %)

2. **Critical approach.** It is about convincing us whether you know what you are doing as opposed to following a recipe. You are able to explain why you are getting the results that the methods produce by referring to the underlying theory. For higher marks you are able to propose improvements. (15 %)

3. **Logical argument, explanation and evaluation of perspectives.** Your work follows a logical order and presents a convincing argument why you have approached the problem the way you did. It is all about the precision of your thinking. (20 %)

4. **Problem-solving.** This is strongly associated with content knowledge and demonstrated by solving the problems to various standards. (20 %)

5. **Insight.** You are able to delineate the strengths and weaknesses of your approach. You are able to come to a conclusion and summarise what you have learnt from solving the problem. (15 %)

6. **Decision-making.** When faced with a problem that has multiple solutions you are able to make a reasoned decision how to proceed. For higher marks you can make reasoned decisions about unseen problems and choose from solutions not explicitly taught. (10 %)

The weightings of the six marking criteria are approximate. As you notice we deem that content knowledge, logical explanation and evaluation and problem solving are the most important.

To achieve marks of 70 and above, you must demonstrate the use and mastery of methods and skills not explicitly taught. This does not need to be excessive and can be

- a realisation about the taught material that was not explicitly mentioned
- a method or skill that you have learnt in addition to the course material
- an insight about the problem that does not immediately follow from standard arguments.

Note that while striving for a mark of 70 and above, you should not forget about the basics. You can only get a high mark when the basics are already covered.

Performance of the machine learnt models is of secondary importance. Instead, the focus is on your understanding of the methods and your ability to critically evaluate their behaviour.

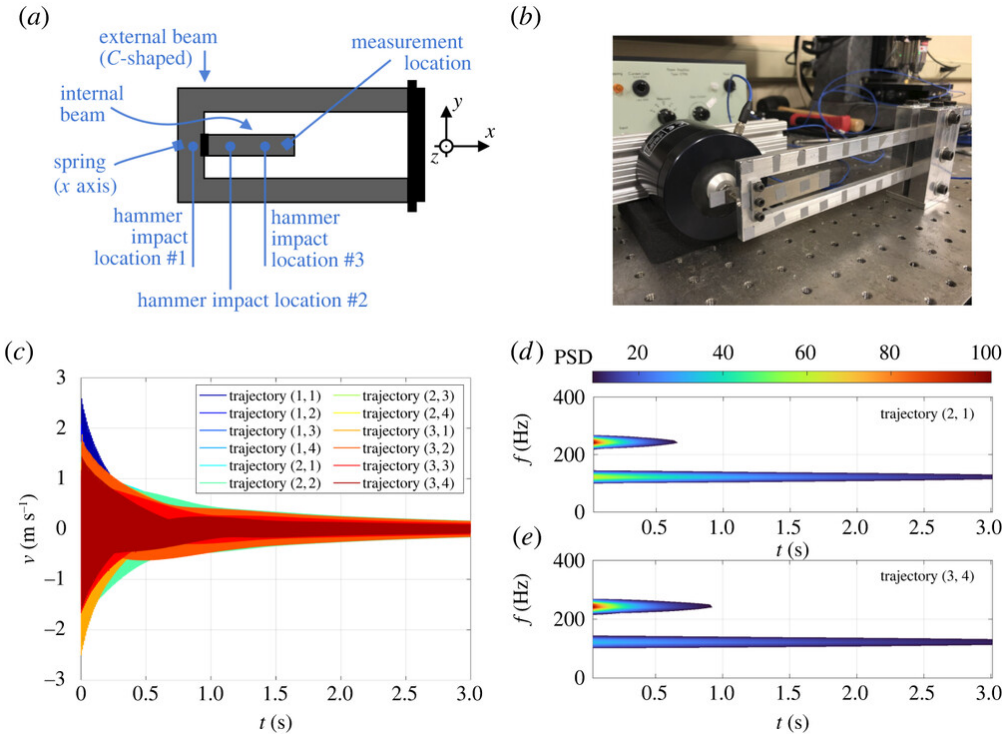

Fig. 1: a) Schematic of the double beam with impact location and measurement location. b) Experimental setup testing a double beam. c) illustration of the 12 trajectories d) power spectrum density of selected trajectories over time. The figure was reproduced from M. Cenedese, J. Axås, H. Yang, M. Eriten and G. Haller. Data-driven nonlinear model reduction to spectral submanifolds in mechanical systems, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 380 (2022) 20210194.

1.1 Problem descriptions

The data comes from impact tests of a double beam. The double beam has a near 2:1 resonance and therefore the two dominant natural frequencies cannot be separated. The data is the velocity of the measurement location in [m/s], that was captured using a laser vibrometer. Read the relevant section of the paper where the data was published to get an idea how the authors processed the data. Note that you are doing a different analysis and therefore the same delay embedding may not be appropriate.

The data is already pre-processed for you, the impact has been removed from the start of the trajectory and each trajectory is encoded as a matrix. The data is encoded as a Numpy *.npz* file. To open the file one can use the commands

```
data_dict = np.load('double_beam_data.npz')
data_all = [it[1] for it in data_dict.items()]
```

The variable *data_all* is the list of all trajectories. The sampling frequency of the data is 5120 [Hz].

Problem 1. Create a linear model using delay embedding and principal component analysis (PCA)

1. Separate the data into a training and a testing set.

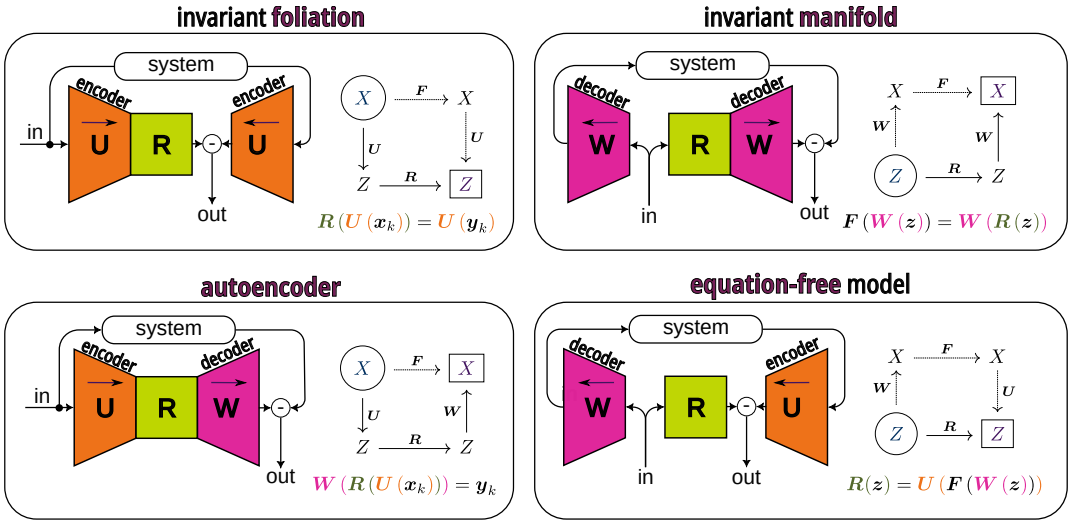

Fig. 2: Four possibilities of relating data to a model.

2. Describe at least two different ways of creating training/testing sets. Explain the differences between the methods, when are they most applicable, and if there are situations when they are inappropriate.

3. When reducing the rank of the linear model using PCA: a) explain in what coordinate system you need to test model accuracy by referring to figure 2; b) explain how to ensure that the errors calculated are comparable between any two delay lengths and any two PCA ranks.

4. Vary the delay to include up to 32 samples and the PCA rank up to 16. Find the best delay length and PCA rank combination.

5. Illustrate the procedure in a diagram displaying both the training and testing errors.

Problem 2. Create a linear model using delay embedding and dynamic mode decomposition (DMD).

1. Separate the data into a training and a testing set.

2. When reducing the rank of the linear model using DMD, explain in what coordinate system you need to test model accuracy by referring to figure 2.

3. Explain how one should order the dynamic modes from most important to least important. Explain what is measured by the residual DMD.

4. Vary the delay to include up to 32 samples and the DMD rank up to 16. Find the best delay length and DMD rank.

5. What happens if you order the dynamic modes by the energy they contain and repeat the model reduction.

Problem 3. Create nonlinear models using delay embedding, a library of monomials up to order three, and dynamic mode decomposition (DMD). Fix the delay length to 12.

1. Vary the model order from linear to cubic and the DMD rank up to 120. Find the best model order and DMD rank combination. (This assumes that you have training and testing data sets as per previous problems.)

2. DMD is defined in a function space. When iterating the resulting linear model, you are producing the coefficients of the function library as they vary over time a) How do you extract the physical coordinates from the function coefficients when your library is a set of monomials. b) How do you extract the physical coordinates when the library of functions does not include the linear monomials (such a $x_1, x_2, \ldots x_n$). c) Is there an optimal way to do the extraction given your training data?

3. Compare the first 240 samples of the testing trajectory with a simulation of your identified model in a diagram.

Problem 4. Sparse regression. Fix the delay length to 12 and use L_1 regularisation to identify a linear model.

1. Find the regularisation parameter that gives the best testing error. Illustrate the dependence of the training and testing errors on the regularisation parameter in a diagram. How many non-zero parameters are left in the model at the optimal regularisation parameter? How does the number of parameters you have found best, compares to what the theory suggest is necessary to describe the most general linear model using delay embedding.

Problem 5. Bagging (also called bootstrap aggregation) and uncertainty.

1. Fix the delay length to 12 and use a linear model the calculate the two dominant natural frequencies and damping ratios of the system. The natural frequency is given by

$$\omega_j = \frac{1}{\Delta t} \arg \lambda_j \text{ [rad/s]}$$

and the damping ratio is given by

$$\zeta_j = -\frac{\log |\lambda_j|}{\arg \lambda_j},$$

where λ_j, $j = 1, \ldots, n$ are the eigenvalues of the linear model.

2. Use bagging (bootstrap aggregation) and fit a two-dimensional Gaussian to the joint distribution of ω_j, ζ_j. What is the mean natural frequency and damping ratio and what is the covariance matrix for each of the two dominant modes?