FMMS050 – Project description in statistical and machine learning methods

Wengang Mao

Department of Mechanics and Maritime Sciences, Division of Marine Technology, Chalmers University of Technology, SE-412 96, Gothenburg, Sweden

---

Project 1: Roll decay damping

A parameters identification technique (PIT) is often used to obtain the damping coefficients from the roll decay tests. In this technique, parameters in a mathematical model are determined in order to get the best fit to a roll decay time signal. A derivation of a mathematical model suitable for this study is described below together with a description of how the parameters: damping A_{44}, stiffness B_{44} and inertia coefficients C_{44} as in Eq.1 are determined. The roll decay motion can be expressed in general form but with nonlinear stiffness:

$$A_{44}\ddot{\phi} + \mathrm{B}_{44}\left(\dot{\phi}\right) + \mathrm{C}_{44}\left(\phi\right) = 0, \tag{1}$$

where $B_{44}(\dot{\phi})$ and $C_{44}(\phi)$ are the damping and stiffness models. A cubic model can be obtained by using cubic damping:

$$\mathrm{B}_{44}\left(\dot{\phi}\right) = B_1\dot{\phi} + B_2\left|\dot{\phi}\right|\dot{\phi} + B_3\dot{\phi}^3 \tag{2}$$

And a higher order stiffness model:

$$\mathrm{C}_{44}\left(\phi\right) = C_1\phi + C_3\phi^3 + C_5\phi^5 \tag{3}$$

The total equation is then written:

$$A_{44}\ddot{\phi} + \left(B_1 + B_2\left|\dot{\phi}\right| + B_3\dot{\phi}^2\right)\dot{\phi} + \left(C_1 + C_3\phi^2 + C_5\phi^4\right)\phi = 0 \tag{4}$$

NB: this equation does not have one unique solution however. If all parameters would be multiplied by a factor k these parameters would also yield as a solution to the equation.

The parameters of this equation can be identified using least square fit if the time signals $\phi(t)$, $\dot{\phi}(t)$ and $\ddot{\phi}(t)$ are all known. For model tests, where only the roll signal $\phi(t)$ is known, the other time derivatives can be estimated using numerical differentiation of a low-pass filtered roll signal or Kalman filtered roll signal. The filtering will however introduce some errors in itself. So instead of using this "differentiation approach", it has been found that solving the differential equation numerically for estimated parameter values determined

using optimization. One problem with this "Integration approach" is that in order to converge, the optimization needs a reasonable first guess of the parameters. The Differentiation approach has therefore been used as a pre-step to obtain a very good first guess of the parameters that can be passed on to the Integration approach.

Roll damping estimation for KVLCC

Your task is to use different statistical regression methods to obtain different parameters in the Roll decay Eq.(4) , especially how to get the roll damping, and discuss how sensitive of the estimation of the roll damping.

Test at 0 knots

Two roll decay model tests were conducted at zero speed referred to as Run 1 and 2. These tests where analyzed by fitting a cubic model to the model test data. The two models were very similar in terms of roll damping and stiffness, suggesting good repeatability in both the model tests and in the parameter identification technique (PIT) used. It can be seen that the dampings, from each individual oscillation obtained with the logarithmic decrement method, are very scattered. This scatter does not seem to influence the two models for the 0 speed case, which are very similar.

Test at 15.5 knots

One roll decay model tests, referred to as Run 3, was conducted at a ship speed corresponding to 15.5 knots full scale ship speed. The ship got a small yaw rate at the end of test, giving a small steady roll angle due to the centrifugal force. Since this effect is not included in the mathematical model used, the steady roll angle was instead removed by removing the linear trend in the roll angle signal.

NB:You can download data from "https://data.mendeley.com/datasets/2stvkyngj9/2", or "https://chalmersuniversity.box.com/s/wsnew9tz6bj9yzbs86i967p6l93im6zm".

Project 2: Ship power prediction by ML methods

The development and evaluation of energy efficiency measures to reduce air emissions from shipping strongly depends on reliable description of a ship's performance when sailing at sea. Normally, model tests and semi-empirical formulas are used to model a ship's performance but they are either expensive or lack accuracy. Today, a lot of ship performance-related parameters have been recorded during a ship's sailing, and different data driven machine learning methods have been applied for the ship speed-power modelling. This project will use different statistical learning and machine learning methods to get a ship's power/fuel consumption. Then we will compare the cons and pros of different methods for the power performance model building.

Machine learning model establishment

When using machine learning methods to estimate a ship's propulsion power based on full-scale measurement data, the relationship between measured ship power and all possible input feature parameters should be established first. Different machine learning algorithms are used and compared for the relationship establishment.

The full-scale measurement data is pre-porcessed and feature-selected in the data processing progress. Then the dataset is split into training set and testing set. The cross-valuation is implemented to ensure the tuned hyperarameters are actually close to the optimal model without overfitting and decrease the generalization error, in the form of k-fold for training set. **NB: the full-scale measurement that can be downloaded for this course has been post-processed, since we cannot share the actual data to people outside our project.**

The purpose of this study is to compare different machine learning methods for a ship's propulsion power prediction, based on the encountered metocean environments. It is demanded to set up a dataset containing instances with all potential input features related to the predicting target - propulsion power. The considered input features belonging to the general ship operation and weather condition, are summarized in Table 1.

Table 1: Attributes considered for the comparison study

Class	Category	Description	Attributes
Input features	Operation	Ship speed through water	V [knots]
		Ship draft	T_{fwd} & T_{aft} [m]
		Heading	HDG [°]
	Metocean	Significant wave height	H_s [m]
		Mean wave period	T_z [s]
		Mean wave direction	D_{wave} [°]
		Wind speed	U_{wind} & V_{wind} [m/s]
Output target	Operation	Propulsion power	P_b [kW]

Your tasks in the project

For the purpose of a ship's propulsion power prediction, the objective target P_b denotes the measured propulsion power. The input features consists of all the listed attributes in

Table 1, i.e., $\boldsymbol{x} = \{V, T_{fwd}, T_{aft}, HDG, H_s, T_z, D_{wave}, U_{wind}, V_{wind}\}$. The target and input features are then fed into different machine learning algorithms for model establishment.

1. Use different methods to establish the power prediction model.
2. Find which features are the most important features to describe the power performance model.
3. Compare different models.
4. Discuss the cons and pros of different methods and models.

You can download the data from the following link:
"https://chalmersuniversity.box.com/s/wsnew9tz6bj9yzbs86i967p6l93im6zm".
NB: the data has been scaled and modified to avoid the issue regarding the NDA with the company. So you might get some strange feeling of some data.

Project 3: Time series of ice condition prediction

A typical NSR route is shown by the red line in Figure 1. According to the daily navigation plan of the ship that followed this route, the NSR is divided into eight segments, and eight sub-regions (green polygons) are created, each of which covers one segment.

Safe and energy-efficient ship navigation along the Northern Sea Route (NSR) requires reliable sea ice concentration (SIC) information. The SIC is critical information to determine the route availability. Thus, it is necessary to obtain SIC data along the NSR from ERA5 grid point values. In this project, the true time series of SIC (from ERA5) at several of the sub-routes. In addition, there are climate models, e.g., Coupled Model Intercomparison Project Phase 5 (CMIP5) that have been used to predict the SIC for the upcoming 50-, 100 years. (**This data can be downloaded from ...**)

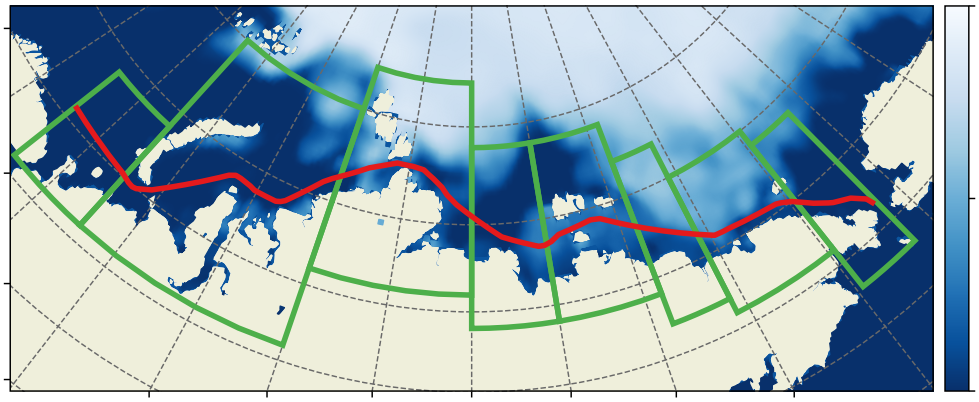

Figure 1: The NSR (red line) and representative sub-regions, as defined in the text, with sea ice conditions chosen from the ERA5 single-level monthly mean values for the date 2019-07-01.

In this project, you are supposed to develop an auto-regressive integrated moving average (ARIMA) model is developed based on ERA5 reanalysis data. The ARIMA model will be used for short-term SIC forecasts along one of the NSR sub-route. The forecast based on the ARIMA model can be compared with the climate project SIC data from the Coupled Model Intercomparison Project Phase 5 (CMIP5). Espeically, put focus on the comparison of the SIC prediction for the years from 2021 to 2025.

Statistic interpolation of SIC

For each sub-region, let Y_t, $X_{t,sia}$, $X_{t,mean}$, $X_{t,var}$, $X_{t,std}$, $X_{t,skew}$, and $X_{t,kurtosis}$ denote the mean SIC of the route segment, area covered by sea ice, mean, variance, standard deviation, skewness, and kurtosis of SIC in the sub-region at time t, respectively. Thus, the mean SIC

of the route segment is assumed to be described by the linear model as follows:

$$\begin{cases} Y_1 &= \beta_0 + \beta_1 X_{1,sia} + \beta_2 X_{1,mean} + \beta_3 X_{1,var} + \beta_4 X_{1,std} + \beta_5 X_{1,skew} + \beta_6 X_{1,kurtosis} + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_{2,sia} + \beta_2 X_{2,mean} + \beta_3 X_{2,var} + \beta_4 X_{2,std} + \beta_5 X_{2,skew} + \beta_6 X_{2,kurtosis} + \epsilon_2 \\ \vdots & \\ Y_N &= \beta_0 + \beta_1 X_{N,sia} + \beta_2 X_{N,mean} + \beta_3 X_{N,var} + \beta_4 X_{N,std} + \beta_5 X_{N,skew} + \beta_6 X_{N,kurtosis} + \epsilon_N \end{cases} , \tag{5}$$

and in matrix notation:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \\ \vdots \\ \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{1,sia} & X_{1,mean} & X_{1,var} & X_{1,std} & X_{1,skew} & X_{1,kurtosis} \\ 1 & X_{2,sia} & X_{2,mean} & X_{2,var} & X_{2,std} & X_{2,skew} & X_{2,kurtosis} \\ & & & & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & & & & & \\ 1 & X_{N,sia} & X_{N,mean} & X_{N,var} & X_{N,std} & X_{N,skew} & X_{N,kurtosis} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \\ \vdots \\ \\ \epsilon_N \end{bmatrix} . \tag{6}$$

Finally, the linear model can be expressed in a compact form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ , \tag{7}$$

where $\boldsymbol{\beta}$ represents coefficients to be estimated, $\boldsymbol{X}$ is the design matrix, and $\boldsymbol{\varepsilon}$ denotes errors. $\boldsymbol{\beta}$ can be estimated by minimizing the sum of squared residuals (RSS),

$$\begin{aligned} \arg\min_{\beta_0,\cdots,\beta_6} \mathrm{RSS}(\boldsymbol{\beta}) &= \arg\min_{\beta_0,\cdots,\beta_6} \left(\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}\right) \\ &= \arg\min_{\beta_0,\cdots,\beta_6} \left((\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right) \\ &= \arg\min_{\beta_0,\cdots,\beta_6} \left(\boldsymbol{Y}^\top \boldsymbol{Y} - \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}\right) \ , \end{aligned} \tag{8}$$

which leads to

$$\frac{\partial RSS(\widehat{\boldsymbol{\beta}})}{\partial \widehat{\boldsymbol{\beta}}} = -2\boldsymbol{X}^\top \boldsymbol{Y} + 2\boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}} = 0 \ . \tag{9}$$

Finally, the estimator $\widehat{\boldsymbol{\beta}}$ can be expressed as

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \ , \tag{10}$$

under the assumption that $\mathrm{rank}(\boldsymbol{X}) = k + 1$ or

$$\det\left(\boldsymbol{X}^\top \boldsymbol{X}\right) \neq 0 \ . \tag{11}$$

Additionally, strict exogeneity tells that the conditional expectation of $\boldsymbol{\varepsilon}$, given the design matrix $\boldsymbol{X}$, is equal to zero, which implies that

$$\widehat{\boldsymbol{Y}} = \mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta} \, . \tag{12}$$

With Equation (12), a linear model has been developed to compute mean SICs of the NSR segments based on both the reanalysis SICs and their statistics.

Your tasks are

The historical (true) SICs along the sub-routes of the NSR will been given to you for downloading. It is forming a series of data points indexed in time order. This type of data can be described as a time series. In this project, you are supposed to use the ARIMA model to analyze the SIC series.
To be more specific your task is to use the ARIMA model to fit the SIC data and use the model for the prediction of SIC in the upcoming months/years. And discuss the accuracy and sensitive of the fitted model.

(NB: the SIC data can be downloaded through the following link:
"https://chalmersuniversity.box.com/s/wsnew9tz6bj9yzbs86i967p6l93im6zm").