1.2 Lecture 17/04

Data driven forecasting

Assume to have N_{obs} scalar observations $y_{\text{obs}}(\tau_k) \in \mathbb{R}, k = 1, \dots, N_{\text{obs}}$ collected at equidistant times τ_k. To define what we understand by a *forecast* or *prediction*, we select a point in time τ_{k_*} that represents the present. Relative to τ_{k_*} we can define the past $t < \tau_{k_*}$ and the future $t > \tau_{k_*}$.

Definition 1.1. *A* forecast *(or* prediction*) is an estimate for*

$$y_{ref}(t) = h(z_{ref}(t)), \qquad t > \tau_{k_*},$$

when only observations from the past and the present are available.

Remark 1.2. *In reality, we would like to make predictions about $z_{ref}(t)$ for $t > \tau_{k_*}$ and not only about $y_{ref}(t)$. This is challenging, we will discuss that later.*

In what follows we discuss how to forecast y_{ref} by using only the observations we have and elementary mathematical tools.

Linear extrapolation Suppose we have at our disposal two observations collected at times τ_{k_*-1} and τ_{k_*}. How to predict the observation at time τ_{k_*+1} or more generally at time $t > \tau_{k_*}$?
Consider the polynomial

$$q(t) = y_{\text{obs}}(\tau_{k_*}) + (t - \tau_{k_*})\frac{(y_{\text{obs}}(\tau_{k_*}) - y_{\text{obs}}(\tau_{k_*-1}))}{\tau_{k_*} - \tau_{k_*-1}}. \tag{1.4}$$

Observe that $q(\tau_{k_*}) = y_{\text{obs}}(\tau_{k_*})$ and $q(\tau_{k_*-1}) = y_{\text{obs}}(\tau_{k_*-1})$. Hence, we can use $q(t)$ for prediction: a predicted observation at time $t > \tau_{k_*}$ is $q(t)$. Let us compute for example the prediction at time τ_{k_*+1} using (1.4). We have:

$$\begin{aligned} y_{\text{predict}}(\tau_{k_*+1}) :&= y_{\text{obs}}(\tau_{k_*}) + (\tau_{k_*+1} - \tau_{k_*})\frac{(y_{\text{obs}}(\tau_{k_*}) - y_{\text{obs}}(\tau_{k_*-1}))}{\tau_{k_*} - \tau_{k_*-1}} \\ &= 2y_{\text{obs}}(\tau_{k_*}) - y_{\text{obs}}(\tau_{k_*-1}). \end{aligned}$$

As soon as $y_{\text{obs}}(\tau_{k_{*+1}})$ becomes available:

1. use it to see if our prediction is accurate or not,
2. use it to predict $y_{\text{predict}}(\tau_{k_*+2})$, discarding τ_{k_*-1}.

How do we asses then the quality of a forecast? A simple measure is given by the following quantity.

Definition 1.2. *For a set of predictions and observations at times* $\{\tau_1, \ldots, \tau_N\}$ *the time averaged* root mean square error *(RMSE) is given by*

$$RMSE := \sqrt{\frac{1}{N}\sum_{k=1}^{N} |y_{obs}(\tau_k) - y_{predict}(\tau_k)|^2}$$

Remark 1.3. *If there are* N_{obs} *observations then* $N = N_{obs} - 2$*, since we cannot use the first two observations to make a prediction.*

Higher-order extrapolation A way to improve the accuracy of forecasts is for instance taking a linear combination of several previous data points. More precisely, suppose we have observations at times $\{\tau_{k_*-p}, \ldots \tau_{k_*}\}$ and we look for a prediction of the form:

$$y_{\text{predict}}(\tau_{k_*+1}) = \sum_{\ell=0}^{p} a_\ell \; y_{\text{obs}}(\tau_{k_*-\ell}), \tag{1.5}$$

for some coefficients $a_\ell = a_\ell(\tau_{k_*+1})$, $\ell = 0, \ldots, p$. How to choose the coefficients? As we have seen above, for the case $p = 1$ (i.e. linear extrapolation) we shall take $a_0 = 2$ and $a_1 = -1$. How to extend to the general case $p > 1$? The aim is to find a formula for the coefficients $a_\ell(t)$ such that

$$y_{\text{predict}}(t) = y_{\text{obs}}(t), \qquad \text{if } t \in \{\tau_{k_*-p}, \ldots, \tau_{k_*}\}. \tag{1.6}$$

For that we may use Lagrange polynomials:

$$a_\ell(t) = L_{k_*-\ell}(t), \qquad \ell = 0, \ldots, p,$$

where

$$L_j(t) = \frac{\prod_{i \neq j}(t - \tau_i)}{\prod_{i \neq j}(\tau_j - \tau_i)}.$$

Observe that

$$a_\ell(\tau_{k_*-j}) = \begin{cases} 1 & \text{if } j = \ell, \\ 0 & \text{otherwise,} \end{cases}$$

from which we can notice that (1.6) is satisfied. In particular a solution of (1.5) is

$$a_\ell = a_\ell(\tau_{k_*+1}) = L_{\tau_{k_*-\ell}}(\tau_{k_*+1}).$$

For further details on Lagrange polynomials see Section 6.2. of Suly&Mayers.

Statistical learning Let us now see another way to determine the coefficients $(a_\ell)_{\ell=0,\dots,p}$ in (1.5). We will not use polynomial interpolation anymore but statistical learning. We divide the observations in two groups: the *training set* and the *test set*. Given $N_T < N_{\text{obs}}$ let us say for simplicity that

- $\{y_{\text{obs}}(\tau_1), \dots, y_{\text{obs}}(\tau_{N_T})\}$ is the training set,
- $\{y_{\text{obs}}(\tau_{N_T+1}), \dots, y_{\text{obs}}(\tau_{N_{\text{obs}}})\}$ is the test set.

We look for predictions of the form $y_{\text{predict}}(\tau_{j+p+1}) = \sum_{\ell=0}^{p} a_\ell y_{\text{obs}}(\tau_{j+p-\ell})$, where $0 < j \leq N_{T-p-1}$. We will choose the coefficients $(a_\ell)_\ell$ that minimize the prediction errors of the training set. Once we found them (and so we fixed them), we shall assess the performance of our extrapolation coefficients on the test set.

The aim is therefore to minimize the functional

$$L((a_0, \dots, a_p)) = \frac{1}{2} \sum_{j=1}^{N_T-p-1} \left(y_{\text{obs}}(\tau_{j+p+1}) - \sum_{\ell=0}^{p} a_\ell y_{\text{obs}}(\tau_{j+p-\ell}) \right)^2 = \frac{1}{2} \sum_{j=1}^{N_T-p-1} r_j^2,$$

using the training set. For that we may use the method of least squares. It is easy to check that the minimum of $L((a_0, \dots, a_p))$ is attained when

$$\frac{\partial L}{\partial a_\ell} = - \sum_{j=1}^{N_T-p-1} y_{\text{obs}}(\tau_{j+p-\ell}) r_j = 0, \qquad \ell = 0, \dots, p$$

i.e. we have $p+1$ linear equations forming a system whose solution will lead to find the $p+1$ unknown coefficients $(a_0, \dots, a_p)$. Suppose now we have found the coefficients a_ℓ with $\ell = 0, \dots, p$. We can use them to make predictions over the test set.

Remark 1.4. *This procedure is meaningful only if the training set and the test behave similarly, e.g. when they are realizations of a stationary time series.*

Model driven forecasting and Data Assimilation So far, we considered procedures that rely on the observed quantities alone, we did not include in our methods any knowledge about the surrogate physical process that generated the observations. The predictions were made under the assumption of a polynomial form in time or by optimising the coefficients over a training set. In this cases we talk about *empirical* or *bottom up* models. Now we introduce a method for making forecasts based on *mechanistic* or *top down* models of the physical process that are derived from *first principles* (principles well known and accepted e.g. in physics such as conservation of mass, Newton law of motion, etc.)

Example. Take again into account the model

$$\frac{\mathrm{d}z_t}{\mathrm{d}t} = f(z) + g(t).$$

We suppose that thanks to a good understanding of the SPP in addition to first principles we are able to set the following *mechanistic model*

$$\begin{cases} z^{n+1} = z^n + \delta t f(z^n), \\ z^0 = z_0, \end{cases} \tag{1.7}$$

with uniform mesh-grid $t^{n+1} = \delta t + t^n$. Denote by z_{model} the solution of (1.7) interpolated in time. In this case the resulting model error over the time interval $(t_n, t_{n+1}]$ is

$$e_{n,m} = \delta t g(t_n),$$

where m denotes the fact we are looking at the model error. The analysis based on z_{model} can lead to good predictions over related long time intervals if $z_{\text{model}}(0)$ is very close to $z_{\text{ref}}(0)$ and we are able to improve our mechanistic model making the contribution given by $g(t)$ as small as possible. Both tasks can be accomplished by clever combinations of mechanistic models with observational data.

Nonlinear method of least squares The differences between simulated and true observations is measured by

$$r_k = y_{\text{model}}(\tau_k) - y_{\text{obs}}(\tau_k) = h(z_{\text{model}}(\tau_k)) - y_{\text{obs}}(\tau_k), \qquad k = 1, \dots, N_a,$$

where N_a is the number of observations under analysis.
NB: The residual implicitly depends on the initial condition z^0 since this changes the entire model trajectory and therefore the simulated observations $y_{\text{model}}(\tau_k)$. How to choose z^0 from the data? Again, for instance, by the method of least squares. We seek the initial condition z^0 that minimizes

$$L(z^0) = \frac{1}{2} \sum_{k=1}^{N_a} \|r_k\|^2.$$

Let us denote by z^0_* the minimizer of $L(z^0)$.
NB: We stress the fact that finding a minimizer is not easy as a nonlinear dependence arises between $z_{\text{model}}(\tau_k)$ on z^0 which may result in nonexistence or nonuniqueness of minimizers.
To solve the problem one may use a generic optimization algorithm that approximates minima. Let us now suppose that z^0_* is known. In contrast

to the forecast $z_{\text{model}}(t)$, $t \geq 0$ which does not make use of $y_{\text{obs}}(\tau_k)$, for $k = 1, \ldots, N_a$, z_*^0 provided an improved (retrospective) approximation $z_{\text{model}}^a(t)$, called the *analysis*

$$z_{\text{model}}^a : \begin{cases} z^{n+1} = z^n + \delta t f(z^n), \\ z_0 = z_*^0. \end{cases}$$

This forecast-analysis can be iterated. Once observations at times τ_k with $k = N_a + 1, \ldots 2N_a$, have become available, a new DA cycle can be initiated. We obtain in this way the initial condition for the forecast over the interval $[\tau_{2N_a}, \tau_{3N_a}]$. This process of producing forecasts with the mechanistic model and correcting them by assimilating available observations over finite time intervals can be repeated as often as desired.

Chapter 2

Probability theory

2.1 Lecture 18/04

As we have already discussed, there are a lot of errors that occur when dealing with a Data Assimilation procedure. These errors are complicated and interconnected in a difficult way. So we model them as random variables. But what is a random variable?

Example. Consider the example of a coin flip. The result X can be "head" or "tail". Per se there is nothing random in flipping a coin: the outcome is determined by all the forces acting on the coin and the initial configuration of the coin. However, the dynamic is so complex that in practice we cannot be sure about the outcome, it is unpredictable, it is random. So we model X as a random variable to which we associate a measure of probability that quantifies the confidence we have about the fact that an event occurs. For instance, if we believe that the coin is not rigged then

$$\mathbb{P}(X = \text{head}) = \mathbb{P}(X = \text{tail}).$$

How to model the fact of ”choosing an object at random”? How to assign probabilities to events?

Example. Consider the experiment of throwing a dart on a dart board. What is the likelihood of hitting a specific point or landing in a particular area? To this extent we want to assign numbers to the set of possible outcomes denoting the possibility that the dart will land in that set.
At very least, to be meaningful, the numbers assigned must obey some rules :

- if we have two sets A and B of the dart board such that $A \subset B$ we will require that $\mathbb{P}(A) \leq \mathbb{P}(B)$.

- If $A \cap B = \emptyset$ we will need that $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
- if Ω models the dart board, we have $\mathbb{P}(\Omega) = 1$.
- Finally we want to include the idea of continuity, i.e. we want to be able to approximate events. This can be translated as follows. If $(A_n)_{n\in\mathbb{N}}$ is an increasing sequence of events such that $\cup_n A_n = A$, then

$$\mathbb{P}(A_n) \uparrow \mathbb{P}(A), \qquad \text{as } n \to +\infty.$$

Definition 2.1. *A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a triplet where*

- *Ω is a non-empty set,*
- *$\mathcal{F}$ is a collection of subsets of Ω closed over complementation and countable unions, i.e.*
  - *$\Omega \in \mathcal{F}$,*
  - *$A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F}$,*
  - *$(A_n)_n \subset \mathcal{F} \Longrightarrow \bigcup_n A_n \in \mathcal{F}$.*

  Such a collection is called σ-algebra or σ-field and its elements are named events.
- *$\mathbb{P} : \mathcal{F} \to [0, 1]$ is called a probability measure if it satisfies the following properties:*
  - *$\mathbb{P}(\Omega) = 1$,*
  - *if $A_1, A_2, \ldots, A_n, \ldots$ is a finite or countable collection of events in $\mathcal{F}$ and they are pairwise disjoint, then:*

$$\mathbb{P}\left(\bigcup_n A_n\right) = \sum_n \mathbb{P}(A_n)$$

Example. 1. Choosing uniformly in $\{1, \ldots, N\}$, for fixed $N \in \mathbb{N}$.

We will define the sample space as $\Omega = \{1, \ldots, N\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$ the set of all possible subsets of Ω. Moreover, for any $A \subset \Omega$ we define

$$\mathbb{P}(A) = \frac{|A|}{N},$$

where with $|A|$ we denote the number of elements in A. For instance

$$\mathbb{P}(\{2, 5\}) = \frac{2}{N}.$$

2. Choosing uniformly in $[0, 1]$.

   We set $\Omega = [0, 1]$. Intuitively, we would like to define $\mathbb{P}$ such that $\mathbb{P}([0, 0.3]) = 0.3$, $\mathbb{P}([0.1, 0.3] \cup [0.4, 0.7]) = 0.5$ and more generally $\mathbb{P}(A)=$ the "size" of A. But we cannot define the size of every subset of $[0, 1]$ (Vitali counterexample)! However, if we take the smallest σ-field that contains the intervals, i.e. the Borel σ-algebra of $[0, 1]$, then we can define the measure of any of its elements, this is the Lebesgue measure.

Conditional probability Assume that we know that the outcome of an experiment is in $B \subset \Omega$. Given that information, what is the probability that the outcome is in A?
Notation: $\mathbb{P}(A|B)$, will denote the conditional probability of A given B. The formal definition is

$$\mathbb{P}(A|B) = \begin{cases} \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} & \text{if } \mathbb{P}(B) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let us now introduce one of the most important rules when dealing with conditional probabilities, i.e. Bayes rule.
Consider $B_1, B_2 \in \mathcal{F}$ s.t. $B_1 \cap B_2 = \emptyset$ and $B_1 \cup B_2 = \Omega$. Now take $A \in \mathcal{F}$, then

$$\mathbb{P}(A) = \mathbb{P}(A \cap B_1) + \mathbb{P}(A \cap B_2) = \mathbb{P}(A|B_1)\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\mathbb{P}(B_2).$$

Thus plugging it into the definition of conditional probability of B_1 given A it holds:

$$\mathbb{P}(B_1|A) = \frac{\mathbb{P}(A \in B_1)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_1)\mathbb{P}(B_1)}{\mathbb{P}(A|B_1)\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\mathbb{P}(B_2)}.$$

This formula is known as Bayes' rule.

More generally, if $\Omega = \bigcup_i B_i$, with $B_i \cap B_j = \emptyset$ if $i \neq j$.

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_j \mathbb{P}(A|B_j)\mathbb{P}(B_j)}$$

Example. A patient goes to see a doctor. The doctor performs a test with 99% reliability, i.e. 99% of people who are sick test positive and 99% of the healthy one test negative. The doctor knows that only one of every 10,000 people in the country are sick. If the patient tests positive, what are the chances the patient is sick?
We can apply the Bayes rule using

$$A = \{\text{a person is sick}\}, \qquad B = \{\text{positive test}\},$$

We have

$$\mathbb{P}(A) = \frac{1}{10000}, \qquad \mathbb{P}(B|A) = \frac{99}{100}, \qquad \mathbb{P}(B|A^c) = \frac{1}{100}.$$

Hence applying Bayes' rule we get

$$\mathbb{P}(A|B) = \frac{0,99 \times 0,0001}{0,010098} < 0,01.$$

Independence It may happen that knowing that an event occurs does not change the probability of another event. In that case we say that the events are independent. Formally, two events A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

A family of events $(A_i)_{i \in I}$, with I countable, is said to be mutually independent if for any finite subcollection $\{i_1, \dots, i_n\} \subset I$ one has

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_n}) = \prod_{j=1}^{n} \mathbb{P}(A_{i_j}).$$

Random variables Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A real random variable X is a function $X : \Omega \to \mathbb{R}$ with "good properties", i.e. a measurable.

Before giving the formal definition of measurable function let us see an example.

Example. Consider $(\Omega, \mathcal{F}, \mathbb{P})$ s.t.

- $\Omega = [0, 1]$,
- $\mathcal{F} = \{\Omega, \emptyset, [0, 0.5], (0.5, 1]\}$,
- $\mathbb{P}([0, 0.5]) = 0.7$ (observe that this uniquely defines $\mathbb{P}$).

Now take $X : \Omega \to \mathbb{R}$ defined as

$$X(\omega) = \begin{cases} 0 & \text{if } \omega \in [0, 0.3], \\ 1 & \text{if } \omega \in (0.3, 1]. \end{cases}$$

This map from Ω to $\mathbb{R}$ does not define a random variable! We cannot determine $\mathbb{P}(X = 0)$ from the information we have and the law of X is hence not defined.

What are the functions whose laws are defined by the model? These are measurable functions! Formally, we have the following.

Definition 2.2. *Fix a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *A function* $X : \Omega \to \mathbb{R}$ *is said to be* $\mathcal{F}$*-measurable if* $X^{-1}((-\infty, a]) \in \mathcal{F}$*, for all* $a \in \mathbb{R}$. *Such a function is called a real* random variable.

Remark 2.1.

$$X \text{ r.v. } \implies X^{-1}((-\infty, a]) \in \mathcal{F} \implies \mathbb{P}(X \leq a) \text{ is well defined } \forall a \in \mathbb{R}.$$

Remark 2.2. *An equivalent definition is* $X : \Omega \to \mathbb{R}$ *is* $\mathcal{F}$*-measurable if and only if* $X^{-1}(B) \in \mathcal{F}$ *for all* B *in* $\mathscr{B}(\mathbb{R})$*, which denotes the Borel* σ*-field of* $\mathbb{R}$.

Definition 2.3. *We call* law *or* distribution *of a real random variable* X *the probability measure defined as follows*

$$P_X(A) := \mathbb{P}(X^{-1}(A)), \qquad \forall A \in \mathscr{B}(\mathbb{R}).$$