Notes

Victor Trappler

Directeurs de Thèse: Arthur VIDARD (Inria)
Élise ARNAUD (UGA)
Laurent DEBREU (Inria)

April 8, 2020

Forward, inverse problems and probability theory

Model space data space and forward problem

We are going to follow Tarantola's description of model and data space in Tarantola [2005]. In order to describe accurately a physical system, we have to define the notion of models. A model represents the link between some parameters and some observable quantities. A simple example is a model that takes the form of a system of ODEs or PDEs, maybe discretized, while the parameters are the initial conditions and the output is one or several time series, describing the time evolution of a quantity at one or several spatial points. An important point to make is that a model is not only the *forward operator*, but must also include the parameter space

> **Definition 1.1 – Model:** A model $\mathfrak{M}$ is defined as a pair composed of a *forward operator* $\mathcal{M}$, and a *parameter space* Θ
>
> $$\mathfrak{M} = (\mathcal{M}, \Theta)$$
>
> The forward operator is the mathematical representation of the physical system, while the parameter space is chosen here to be a subset of a finite dimensional space, so usually Θ will be a subset of $\mathbb{R}^n$.

As we will usually consider Θ as a subset of $\mathbb{R}^n$, for $n \geq 1$, we can define a kind of dimensionality of the model, based on the number of *degrees of freedom* available for the parameters to vary freely.

Remark 1.2: The dimension of a model $\mathfrak{M} = (\mathcal{M}, \Theta)$ is the number of parameters not reduced to a singleton, so if $\Theta \subset \mathbb{R}^n$, the dimension of $\mathfrak{M}$ is $d \leq n$. The dimension of a model $\mathfrak{M}$ is sometimes called the degrees of freedom of $\mathfrak{M}$.

Example 1.3: A model with parameter space $\Theta = \mathbb{R}^2 \times [0, 1]$ has dimension 3, while $\Theta = \mathbb{R}^2 \times \{1\}$ has dimension 2.

Now that we have introduced the forward operator and the parameter space, we will focus on the output of the model. The data space consists in all the physically acceptable results of the physical experiment. This set is noted $\mathbb{Y}$. Then, the forward operator $\mathcal{M}$ maps the parameter space $\Theta \subset \mathbb{R}^d$ to the data space $\mathbb{Y}$, as one can expect that all models provide physically acceptable outputs.

Forward problem

Given a model $(\mathcal{M}, \Theta)$, the *forward problem* consists in applying the forward operator to a given $\theta \in \Theta$, in order to get the *model prediction*. The forward problem is then to obtain information on the result of the experiment based on the parameters we chose as input, so deriving a satisfying forward operator $\mathcal{M}$.

$$\begin{array}{rccc} \mathcal{M}: & \Theta & \longrightarrow & \mathbb{Y} \\ & \theta & \longmapsto & \mathcal{M}(\theta) \end{array}$$

As said earlier, the forward operator can be a set of ODEs or PDEs, discretized or not. The forward problem is then the attempt to link the causes (i.e. the parameters) to the consequence, i.e. the output in the data space.

Inverse Problem

The inverse problem is the natural counterpart of the forward problem, and consists in trying to gather more information on the parameters, based on the result of the experiment or the physical process, and the knowledge of the forward operator. This kind of circular procedure: adding complexity by updating the forward operator and the parameter space by choosing a model with higher complexity for the forward problem, and reducing this complexity by comparing some observations with the output of the model, and reducing the parameter space.

However, a purely deterministic approach for the inverse problem is doomed to fail: as most physical processes are not perfectly known, some uncertainties remain in the whole modelling process. Those uncertainties are ubiquitous: the observations available may be corrupted by a random noise coming from the measurement devices and the model may not represent perfectly the reality, thus introducing a systematic bias for instance. Taking into account those uncertainties is crucial to solve the inverse problem.

In that perspective we are going to introduce briefly the usual probabilistic framework, along with common notations that we will use throughout this manuscript. Those notions are well established in the scientific literature, and one can read Billingsley [2008] for a more thorough description.

Notions of probability theory

Probability measure, and random variables

We are first going through some usual notions of probability theory. Let us consider the usual probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 1.4 – Event probability and conditioning: We call an event an element of the σ-algebra $\mathcal{F}$, and the probability of an event $A \in \mathcal{F}$ is defined as the Lebesgue integral

$$\mathbb{P}[A] = \int_A \mathrm{d}\mathbb{P}(\omega) \tag{1}$$

Observing an event $B \in \mathcal{F}$ can bring information upon another event $A \in \mathcal{F}$. In that sense, we introduce the conditional probability of A given B. Let A, $B \in \mathcal{F}$. The event A given B is written $A|B$ and its probability is

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \tag{2}$$

Formally, an event can be seen as an outcome of some uncertain experiment, and its probability is "how likely" this event will happen.

Let us now introduce a measurable state (or sample) space $(S, \mathcal{B}(S))$.

Definition 1.5 – Random Variable, Expectation: A random variable (abbreviated as r.v.) X is a measurable function from $\Omega \longrightarrow S$. A random variable will usually be written with an upper case letter. A realisation or observation x of the r.v. X is the actual image of $\omega \in \Omega$ under X: $x = X(\omega)$. If S is countable, the random variable is said to be *discrete*.

The expectation of a r.v. $X : \Omega \to S$ is defined as

$$\mathbb{E}[X] = \int_\Omega X(\omega)\, \mathrm{d}\mathbb{P}(\omega)$$

Remark 1.6: When $S = \mathbb{R}^p$ with $p > 1$, and $\mathcal{B}(\mathbb{R}^p)$ the usual borelian σ-algebra on $\mathbb{R}^p$, a random variable is called a random vector.

Remark 1.7: Using the Definition 1.5, the probability of an event A can be seen as the expectation of a well chosen random variable:

$$\begin{array}{rccc} \mathbb{1}_A : & \Omega & \longrightarrow & \{0,1\} \\ & \omega & \longmapsto & \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ if } \omega \notin A \end{cases} \end{array}$$

and

$$\begin{aligned} \mathbb{E}[\mathbb{1}_A] &= \int_\Omega \mathbb{1}_A \, \mathrm{d}\mathbb{P}(\omega) \\ &= \int_A \mathrm{d}\mathbb{P}(\omega) = \mathbb{P}[A] \end{aligned}$$

$\mathbb{1}_A$ is called the indicator function of the event A.

Definition 1.8 – Image (Pushforward) measure: Let $X : \Omega \to S$ be a random variable, and $A \subseteq S$. The image measure (also called pushforward measure) of $\mathbb{P}$ through X is denoted by $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$. This notation can differ slightly depending on the community, so one can find also $\mathbb{P}_X = \mathbb{P} \circ X^{-1} = X_\sharp \mathbb{P}$, the latter notation being used in transport theory. The probability, for the r.v. X to be in A is equal to

$$\mathbb{P}[X \in A] = \mathbb{P}_X[A] = \int_A \mathrm{d}\mathbb{P}_X(\omega) = \int_{X^{-1}(A)} \mathrm{d}\mathbb{P}(\omega) = \mathbb{P}[X^{-1}(A)] = \mathbb{P}[\{\omega \,;\, X(\omega) \in A\}]$$

Generally speaking, the sample space will be $S \subseteq \mathbb{R}^p$ for $p \geq 1$, so we are going to introduce useful tools and notations to caracterize these particular r.v.

Real-valued random variables

We are now going to focus on real-valued random variables, so measurable function from Ω to the sample space $(S, \mathcal{B}(S)) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Definition 1.9 – Distribution of a real-valued r.v.: The distribution of a r.v. can be characterized by a few functions:

- The *cumulative distribution function* (further abbreviated as cdf) of a real-valued r.v. X is defined as the probability of the right closed intervals that generate the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ of the real line.

$$F_X(x) = \mathbb{P}\left[X \leq x\right] = \mathbb{P}_X\left[\,\right] - \infty; x\left]\,\right]$$

and $\lim_{-\infty} F_X = 0$ and $\lim_{+\infty} F_X = 1$ If the cdf of a random variable is continuous, the r.v. is said to be *continuous* as well.

- The *quantile function* Q_X is the generalized inverse function of the cdf:

$$Q_X(p) = \inf\{q : F_X(q) \geq p\}$$

- If there exists a function $f : S \to \mathbb{R}^+$ such that for all measurable sets A

$$\mathbb{P}[X \in A] = \int_A \mathrm{d}\mathbb{P}_X(\omega) = \int_A f(x)\,\mathrm{d}x$$

then f is called the *probability density function* (abbreviated pdf) of X and is denoted p_X. As $\mathbb{P}[X \in S] = 1$, it follows trivially that $\int_S f(x)\,\mathrm{d}x = 1$.

Remark 1.10: When restricting this search to "classical" functions, p_X may not exist. However, allowing generalized functions such as the *dirac delta function*, provides a way to consider simultaneously all types of real-valued random variables (continous, discrete, and mixture of both). Dirac's delta function can (in)formally be defined as

$$\delta_{x_0}(x) = \begin{cases} +\infty \text{ if } x = x_0 \\ 0 \text{ elsewhere} \end{cases} \qquad \text{and} \quad \int_S \delta_{x_0}(x)\,\mathrm{d}x = 1$$

Example 1.11: Let us consider the random variable X that takes the value 1 with probability 0.5, and follows a uniform distribution with probability 0.5 over $[2; 4]$. Its cdf can be expressed as

$$F_X(x) = \begin{cases} 0 \text{ if } x < 1 \\ 0.5 \text{ if } 1 \leq x < 2 \\ 0.5 + \frac{x-2}{8} \text{ if } 2 \leq x < 4 \\ 1 \text{ if } 4 \leq x \end{cases}$$

and its pdf (as a generalized function)

$$p_X(x) = \frac{1}{2}\delta_1(x) + \frac{1}{4}\mathbb{1}_{\{2 \leq x < 4\}}(x)$$

Definition 1.12 – Moments of a r.v. and L^s spaces: Let X be a random variable. The moment of order s is defined as $\mathbb{E}[X^s]$, and the centered moment of order s is

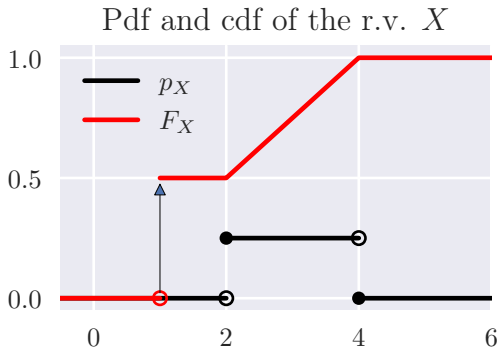

Figure 1 – Cdf and Pdf of X defined in Example 1.11. The arrow indicates a dirac delta function

defined as

$$\mathbb{E}[(X - \mathbb{E}[X])^s] = \int (X(\omega) - \mathbb{E}[X])^s \, \mathrm{d}\mathbb{P}(\omega) = \int (x - \mathbb{E}[X])^s \cdot p_X(x) \, \mathrm{d}x$$

To ensure that those moments exists, let us define $L^s(\mathbb{P})$ as the space of random variables X such that $\mathbb{E}\left[|X|^s\right] < +\infty$. If $X \in L^2(\mathbb{P})$, the centered moment of order 2 is called the variance:

$$\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{V}\mathrm{ar}[X] \geq 0$$

Extending those definitions from real-valued random variables to real-valued random vectors is pretty straightforward

Real-valued random vectors

Definition 1.13 – Joint, marginal and conditional densities: Let $X = [X_1, \cdots, X_p]$ be a random vector from $\Omega \to S \subseteq \mathbb{R}^p$

- The expected value of a random vector is the expectation taken component-wise

$$\mathbb{E}[X] = [\mathbb{E}[X_1], \ldots, \mathbb{E}[X_p]]$$

- The covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ of X is defined as

$$\Sigma = \mathrm{Cov}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])\,(X - \mathbb{E}[X])^T\right]$$

- More generally, the covariance matrix of two random vector X and Y is defined as

$$\mathrm{Cov}\,[X, Y] = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T\right]$$

- The cdf of X at the point $x = [x_1, \dots x_p]$ is

$$F_X(x) = F_{X_1,\dots,X_p}(x_1, \dots, x_p) = \mathbb{P}\left[X_1 \leq x_1, \cdots, X_p \leq x_p\right] = \mathbb{P}\left[\bigcap_{i=1}^{p} \{\omega;\, X_i(\omega) \leq x_i\}\right]$$

- Similarly as in the real-valued case, we can define the pdf of the random vector, or *joint pdf* by derivating with respect to the variables:

$$p_X(x) = p_{X_1,\dots,X_p}(x_1, \dots, x_p) = \frac{\partial^p F_X}{\partial x_1 \cdots \partial x_p}(x)$$

  and $\int_S p_{X_1,\dots,X_p}(x_1, \dots, x_p)\, \mathrm{d}(x_1, \dots, x_p) = 1$

- For notation clarity, we are going to set $X = [Y, Z]$ We can now define the *marginal densities*

$$p_Y(y) = \int_{\mathbb{R}} p_{Y,Z}(y, z)\, \mathrm{d}z \quad \text{and} \quad p_Z(z) = \int_{\mathbb{R}} p_{Y,Z}(y, z)\, \mathrm{d}y \tag{3}$$

  The random variable Y given Z, denoted by $Y \mid Z$ has the conditional density

$$p_{Y|Z}(y \mid z) = \frac{p_{Y,Z}(y, z)}{p_Z(z)}$$

  allowing us to rewrite the marginals as

$$p_Y(y) = \int_{\mathbb{R}} p_{Y|Z}(y|z) p_Z(z)\, \mathrm{d}z = \mathbb{E}_Z\left[p_{Y|Z}(y|z)\right] \tag{4}$$

$$p_Z(z) = \int_{\mathbb{R}} p_{Z|Y}(z|y) p_Y(y)\, \mathrm{d}y = \mathbb{E}_Y\left[p_{Z|Y}(z|y)\right] \tag{5}$$

Definition 1.14 – Independence: Let $A, B \in \mathcal{F}$. Those two events are deemed independent if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$. Quite similarly, two real-valued random variables Y and Z are said to be independent if $F_{Y,Z}(y, z) = F_Y(y)F_Z(z)$ or equivalently, $p_{Y,Z}(y, z) = p_Y(y)p_Z(z)$

We are now going to introduce one of the most important distribution

Example 1.15 – The Normal Distribution: One central example is the normal (or Gaussian) distribution. Let X be a r.v. from Ω to $\mathbb{R}$. X follows the normal

distribution of mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ when

$$p_X(x) = \phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

and we write $X \sim \mathcal{N}(\mu, \sigma^2)$ For the multidimensional case, so when X is a r.v. from Ω to $\mathbb{R}^p$, X follows a normal distribution of mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$, where Σ is semi-definite positive. In that case, $X \sim \mathcal{N}(\mu, \Sigma)$ the density of the random vector X can be written as

$$p_X(x) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-1} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

where $|\Sigma|$ is the determinant of the matrix Σ, and $(\cdot)^T$ is the transposition operator. As the covariance matrix appears through its inverse, another encountered parametrization is to use the precision matrix Σ^{-1}

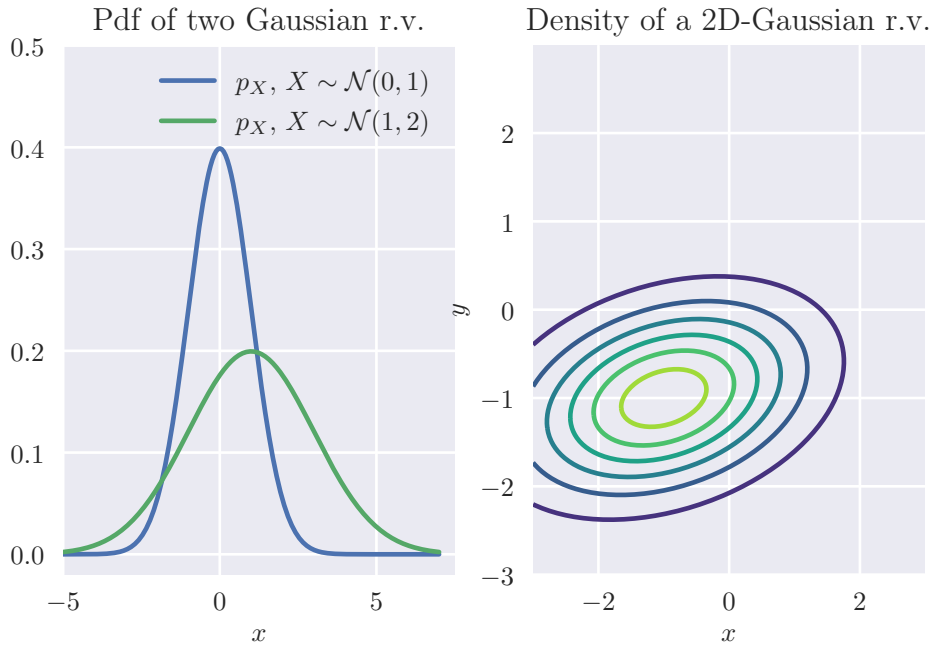

Figure 2 – Densities of 1D Gaussian distributed r.v. (left), and density of a 2D Gaussian r.v.

Bayes' Theorem

The classical Bayes' theorem is directly a consequence of the definition of the conditional probabilities in Definition 1.4, or by considering the pdf of r.v. in Definition 1.13.

Theorem 1.16 – Bayes' theorem: Let $A, B \in \mathcal{F}$. Bayes' theorem states that

$$\mathbb{P}[A \mid B] \cdot \mathbb{P}[B] = \mathbb{P}[B \mid A] \cdot \mathbb{P}[A]$$
$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[B \mid A] \cdot \mathbb{P}[A]}{\mathbb{P}[B]} \text{ if } \mathbb{P}[B] \neq 0$$

In terms of densities, the formulation is sensibly the same. Let Y and Z be two random variables. The conditional density of Y given Z can be expressed using the conditional density of Z given Y.

$$p_{Y|Z}(y \mid z) = \frac{p_{Z|Y}(z \mid y)p_Y(y)}{p_Z(z)} = \frac{p_{Z|Y}(z \mid y)p_Y(y)}{\int p_{Z,Y}(z, y)\,\mathrm{d}y} \propto p_{Z|Y}(z \mid y)p_Y(y)$$

Bayes' theorem is central as it links in a simple way conditional densities. In the inverse problem framework, if Y represents the state of information on the parameter space, while Z represents the information on the data space, $Z|Y$ can be seen as the forward problem. Bayes' theorem allow us to "swap" the conditioning, and get information on $Y|Z$, that can be seen as the inverse problem.

Parameter inference

From the physical experiment to the model

The physical system (the reality) that is observed can formally be represented by a model, so by an operator $\mathscr{M}$, applied to a set of parameters $\vartheta \in \Theta_{\text{real}}$ that is unknown:

$$\begin{array}{rccc} \mathscr{M}: & \Theta_{\text{real}} & \longrightarrow & \mathbb{Y} \\ & \vartheta & \longmapsto & \mathscr{M}(\vartheta) \end{array}$$

The physical reality yields some observations $\mathscr{M}(\vartheta)$, shortened as $\mathscr{M}(\vartheta) = y \in \mathbb{Y}$.

The main objective is to find an appropriate model $(\mathcal{M}, \Theta)$, that represents as accurately as possible the given reality.

$$\mathscr{M}(\vartheta) = \mathcal{M}(\theta) + \delta(\theta) \in \mathbb{Y} \subseteq \mathbb{R}^p$$

The difference $\delta(\theta) = \mathscr{M}(\vartheta) - \mathcal{M}(\theta)$ is the error between the physical model and the model, called sometimes the misfit, or the residuals error.

Frequentist inference, MLE

For a given choice of parameter $\theta \in \Theta$, one common assumption is that those residuals are normally distributed $\delta(\theta) \sim \mathcal{N}(0, \Sigma)$, with a given covariance matrix Σ, so the observations $\mathscr{M}(\vartheta) - \delta(\theta) = Y$ form a random variable, and as we assume that $\mathbb{Y} \subseteq \mathbb{R}^p$, Y is a random vector with the following distribution:

$$Y \sim \mathcal{N}(\mathcal{M}(\theta), \Sigma) \tag{6}$$

Its pdf will be denoted of this random variable will be $y \mapsto p_Y(y;\theta)$ to show the depedency with respect to θ. Instead of looking at this function as a pdf, we may look at it instead as a function of θ, as the observations $y \in \mathbb{Y}$ do not vary

Definition 2.1 – Likelihood function, MLE: The probability density function of the observations for a set of parameters is called the likelihood of those parameters given the observations, and is written $\mathcal{L}$:

$$\mathcal{L}(\cdot; y) : \theta \mapsto p_Y(y;\theta) = \mathcal{L}(\theta; y) \tag{7}$$

$$= (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\mathcal{M}(\theta) - y)^T \Sigma^{-1}(\mathcal{M}(\theta) - y)\right) \tag{8}$$

Based on the likelihood function, we can define the *Maximum Likelihood Estimator*, or *MLE*, that maximizes the likelihood defined above:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta; y) = \arg\min_{\theta \in \Theta} -\log \mathcal{L}(\theta; y) \tag{9}$$

In practice, instead of maximizing the likelihood, one looks for minimizing the negative log-likelihood. Given the Definition 2.1

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\theta \in \Theta} -\log \mathcal{L}(\theta; y) \tag{10}$$

where

$$-\log \mathcal{L}(\theta; y) = \frac{1}{2}(\mathcal{M}(\theta) - y)^T \Sigma^{-1}(\mathcal{M}(\theta) - y) + \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma|$$

Removing the constant terms,

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \arg\min_{\theta \in \Theta} \frac{1}{2}(\mathcal{M}(\theta) - y)^T \Sigma^{-1}(\mathcal{M}(\theta) - y) \\ &= \arg\min_{\theta \in \Theta} \frac{1}{2}\|\mathcal{M}(\theta) - y\|^2_{\Sigma^{-1}}\end{aligned}$$

Frequentist inference and Maximum Likelihood estimation boils down to Generalized non-linear least-square regression, that minimizes the squared Mahalabonis distance between $\mathcal{M}\theta$ and y. This is only true as we assumed a Gaussian form of the errors in Eq. (6). Other choices of likelihood will bring different forms of objective functions. **1.insert examples / elliptical distributions ?**

Bayesian Inference

In Bayesian inference, the uncertainty present on θ is modelled as considering it as a random variable. In that sense, we assume that we have a *prior distribution* on θ, denoted p_θ, that represents the current state of belief upon the parameter. We will

develop later on the choice of this prior distribution. The modelled likelihood of the frequentist approach can be almost be rewritten as is, just by conditioning Y with θ. Equation (6) becomes

$$Y|\theta \sim \mathcal{N}(\mathcal{M}(\theta), \Sigma) \tag{11}$$

and the likelihood is $\mathcal{L}(\theta; y) = p_{Y|\theta}(y|\theta)$. Using Bayes' theorem, the *posterior distribution* of the parameters given the observed data is

$$p_{\theta|Y}(\theta|y) = \frac{p_{Y|\theta}(y|\theta)p_\theta(\theta)}{p_Y(y)} = \frac{\mathcal{L}(\theta; y)p_\theta(\theta)}{p_Y(y)} \propto \mathcal{L}(\theta; y)p_\theta(\theta) \tag{12}$$

This posterior distribution is central in a Bayesian setting, as it represents the information we have on the parameter, given the data.

Bayesian Point estimates

Bayesian point estimates usually refer to point estimation of the parameter θ, using the posterior distribution $p_{\theta|Y}$. Those estimates are usually constructed to capture a central tendency of the posterior distribution. This can be done by defining Bayesian loss functions $L : \Theta \times \Theta \to \mathbb{R}^+$, and defining the associated point estimate as a minimizer of the expected value of the loss function, taken under the posterior distribution.

$$\theta_L = \arg\min_{\theta' \in \Theta} \mathbb{E}_{\theta|Y}\left[L(\theta', \theta)|y\right] \tag{13}$$

Posterior mean By taking a loss function as the squared error $L(\theta', \theta) = (\theta' - \theta)^2$, we can define the Mean Squared Error (MSE) as $\text{MSE} : \theta' \mapsto \mathbb{E}_{\theta|Y}\left[(\theta' - \theta)^2\right]$. Finally, the value corresponding to the Minimum Mean Squared Error is

$$\hat{\theta}_{\text{MMSE}} = \arg\min_{\theta' \in \Theta} \mathbb{E}_{\theta|Y}\left[(\theta' - \theta)^2|y\right] \tag{14}$$

Simple algebraic manipulations show that the minimizer is in fact the posterior mean:

$$\hat{\theta}_{\text{MMSE}} = \mathbb{E}_{\theta|Y}[\theta|y] = \int_\Theta \theta \cdot p_{\theta|Y}(\theta|y)\,\mathrm{d}\theta$$

Posterior Median Instead of a squared error, one can define $L(\theta', \theta) = |\theta' - \theta|$, and the the bayesian risk associated is called the mean absolute error. Again, one can show that the Minimum Mean Absolute Error (MMAE) is in fact the median of the posterior distribution.

$$\hat{\theta}_{\text{MMAE}} = \arg\min_{\theta' \in \Theta} \mathbb{E}_{\theta|Y}\left[|\theta' - \theta| \mid y\right] = \text{Median}(\theta|y) \tag{15}$$

Posterior Mode: the MAP Taking $L(\theta', \theta) = \mathbb{1}_{\theta' \neq \theta}$ that is 0 if $\theta' = \theta$, and 1 elsewhere, one can show that the minimizer of $\mathbb{E}_{\theta|Y}\left[\mathbb{1}_{\theta' \neq \theta}\right]$ is the mode of the posterior distribution, and is called the *Maximum A Posteriori* (MAP):

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \arg\min_{\theta' \in \Theta} \mathbb{E}_{\theta|Y}\left[\mathbb{1}_{\theta' \neq \theta}|y\right] = \arg\min_{\theta' \in \Theta} -p_{\theta|Y}(\theta'|y) \\ &= \arg\max_{\theta' \in \Theta} p_{\theta|Y}(\theta'|y)\end{aligned} \tag{16}$$

Model selection

Likelihood ratio test

The likelihood ratio test is a useful test in the case of nested models, as described in what follows:

Nested models

Definition 3.1 – Nested models: Let $\mathfrak{M}_1 = (\mathcal{M}_1, \Theta_1)$ and $\mathfrak{M}_2 = (\mathcal{M}_2, \Theta_2)$ be two models. $\mathfrak{M}_1$ is said to be nested within $\mathfrak{M}_2$ if

$$\mathcal{M}_1 = \mathcal{M}_1 \text{ and } \Theta_1 \subset \Theta_2$$

Example 3.2: Let us consider two models, where $\mathbb{Y} = \mathbb{R}$

$$\begin{aligned}\mathfrak{M}_1 &= ((a,b) \mapsto ab; \quad (a,b) \in \mathbb{R} \times [0;2]) \\ \mathfrak{M}_2 &= \left((a,b) \mapsto ab; \quad (a,b) \in \mathbb{R}^+ \times \{1/\pi\}\right)\end{aligned}$$

$\mathfrak{M}_2$ is nested within $\mathfrak{M}_1$

Example 3.3: Now let us consider $\mathbb{Y}$ as the space of random vector of dimension n:

$$\begin{aligned}&\mathfrak{M}_1 : (X, A, \sigma) \mapsto AX + \sigma\epsilon, \text{ with } (X, A, \sigma) \in \mathbb{R}^n \times \mathbb{R}^{n \times n} \times \mathbb{R}^+ \text{ and } \epsilon \sim \mathcal{N}(0, I) \\ &\mathfrak{M}_2 : (X, A, \sigma) \mapsto AX + \sigma\epsilon, \text{ with } (X, A, \sigma) \in \mathbb{R}^n \times \mathbb{R}^{n \times n} \times \{1\} \text{ and } \epsilon \sim \mathcal{N}(0, I)\end{aligned}$$

Once again in this example, $\mathfrak{M}_2$ is nested within $\mathfrak{M}_2$

Using the likelihood defined above, we can test for the following hypotheses:

- $\mathcal{H}_0$: $\theta \in \Theta_0 \subset \mathbb{R}^d$
- $\mathcal{H}_1$: $\theta \in \Theta_1 \subset \mathbb{R}^r$, and $\Theta_0 \subset \Theta_1$

Intuitively, we can see Θ_1 as the more general model. The test statistic is

$$\Lambda(y) = \frac{\sup_{\theta\in\Theta_0} \mathcal{L}(\theta; y)}{\sup_{\theta\in\Theta_1} \mathcal{L}(\theta; y)} \tag{17}$$

and under $\mathcal{H}_0$, the quantity

$$-2\log\Lambda(y) \xrightarrow{\mathrm{d}} \chi^2_{r-d} \tag{18}$$

is asymptotically distributed as a χ^2_{r-d}. Using the log-likelihood, $-2(l(\theta_0; y) - l(\theta_1; y)) \xrightarrow{\mathrm{d}} \chi^2_{r-d}$ The asymptotic rejection region of level α is then

$$\begin{aligned}
\mathrm{RejReg}_\alpha &= \{y \mid -2\log\Lambda(y) > \chi^2_{1-\alpha,r-d}\} && (19)\\
&= \{y \mid \log\Lambda(y) < -\frac{1}{2}\chi^2_{1-\alpha,r-d}\} && (20)\\
&= \{y \mid (\sup_{\theta\in\Theta_0} l(\theta; y) - \sup_{\theta\in\Theta_1} l(\theta; y)) < -\frac{1}{2}\chi^2_{1-\alpha,r-d}\} && (21)\\
&= \{y \mid (\sup_{\theta\in\Theta_1} l(\theta; y) - \sup_{\theta\in\Theta_0} l(\theta; y)) > \frac{1}{2}\chi^2_{1-\alpha,r-d}\} && (22)\\
& && (23)
\end{aligned}$$

Let us set $\theta = (k, u, \phi)$ where ϕ represents additional parameters in the likelihood

$$\mathcal{L}(\theta; y) = \mathcal{L}(k, u, \phi; y) \tag{24}$$

Let us assume furthermore that the maximizer of the likelihood depends only on u (we remove the dependence on y in the notation, to declutter).

$$\arg\max_{k\in\mathbb{K}} \mathcal{L}(k, u, \phi) = k^*(u) = \arg\max_{k\in\mathbb{K}} \ell(k, u, \phi) \tag{25}$$

Now let us consider the ratio given a value u and

$$-2\log\Lambda(u, \phi') = -2\left(\ell(k, u, \phi) - \ell(k^*(u), u, \phi')\right) \tag{26}$$

Given u, let us define the following likelihoods

$$\begin{aligned}
\mathcal{L}(k; u, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{J(k,u)}{2\sigma^2}\right] && (27)\\
\mathcal{L}(k = k^*(u); u, \varsigma^2) &= \frac{1}{\sqrt{2\pi}\varsigma}\exp\left[-\frac{J^*(u)}{2\varsigma^2}\right] && (28)\\
& && (29)
\end{aligned}$$

Taking the ratio yields

$$\begin{aligned}
\frac{\mathcal{L}(k; u, \sigma^2)}{\mathcal{L}(k^*; u, \varsigma^2)} &= \frac{\varsigma}{\sigma}\exp\left[-\frac{1}{2}\left(\frac{J(k,u)}{\sigma^2} - \frac{J^*(u)}{\varsigma^2})\right)\right] && (30)\\
&= \frac{\varsigma}{\sigma}\exp\left[-\frac{1}{2\sigma^2}\left(J(k,u) - \frac{\sigma^2}{\varsigma^2}J^*(u))\right)\right] && (31)
\end{aligned}$$

taking twice the negative log likelihood,

$$-2\log\frac{\mathcal{L}(k;u,\sigma^2)}{\mathcal{L}(k^*;u,\varsigma^2)} = \frac{1}{\sigma^2}\left(J(k,u) - \frac{\sigma^2}{\varsigma^2}J^*(u)\right) + 2\log\frac{\sigma}{\varsigma} \tag{32}$$

The log ratio ϱ is

$$\varrho(k,u,\sigma,\varsigma) = \frac{1}{\sigma^2}\left(J(k,u) - \frac{\sigma^2}{\varsigma^2}J^*(u)\right) + 2\log\frac{\sigma}{\varsigma} \tag{33}$$

$$(\text{When } \sigma = 1) = \left(J(k,u) - \frac{1}{\varsigma^2}J^*(u)\right) - 2\log\varsigma \tag{34}$$

Relative Likelihood

Bayesian Model Selection

Let us assume that for $\mathcal{M}$ is chosen to represent the problem at stake. In this case, θ represent implicitly parameters of this model $\mathcal{M}$. Bayes' theorem gives

$$p(\theta|\mathcal{M},y) = \frac{p(y|\mathcal{M},\theta)p(\theta)}{p(y|\mathcal{M})} \tag{35}$$

In Eq. (35), $p(y|\mathcal{M}) = \int_\Theta p(y|\mathcal{M},\theta)p(\theta)\,\mathrm{d}\theta$ is called the evidence of the model $\mathcal{M}$ given the data y.

Bayes factor

When comparing two models $\mathcal{M}_1$ and $\mathcal{M}_2$, one can compute the Bayes factor, that is the ratio of the evidence of the two models:

$$\mathrm{BF}(\mathcal{M}_1,\mathcal{M}_2) = \frac{p(y|\mathcal{M}_1)}{p(y|\mathcal{M}_1)} \tag{36}$$

GP, RR-based family of estimators

Random processes

Let us assume that we have a map f from a p dimensional space to $\mathbb{R}$:

$$\begin{array}{rccc} f: & \mathbb{X} \subset \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & f(x) \end{array} \tag{37}$$

This function is assumed to have been evaluated on a design of n points, $\mathcal{X} \subset \mathbb{X}^n$. We wish to have a probabilistic modelling of this function We introduce random processes as way to have a prior distribution on function This uncertainty on f is modelled as a random process:

$$\begin{array}{rccc} Z: \mathbb{X} \times \Omega & \longrightarrow & \mathbb{R} \\ (x, \omega) & \longmapsto & Z(x, \omega) \end{array} \tag{38}$$

The ω variable will be omitted next.

Linear Estimation

A linear estimation $\hat{Z}$ of f at an unobserved point $x \notin \mathcal{X}$ can be written as

$$\hat{Z}(x) = \begin{bmatrix} w_1 \dots w_n \end{bmatrix} \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} = \mathbf{W}^T f(\mathcal{X}) = \sum_{i=1}^{n} w_i(x) f(x_i) \tag{39}$$

Using those kriging weights $\mathbf{W}$, a few additional conditions must be added, in order to obtain the Best Linear Unbiased Estimator:

- Non-biased estimation: $\mathbb{E}[\hat{Z}(x) - Z(x)] = 0$
- Minimal variance: $\min \ \mathbb{E}[(\hat{Z}(x) - Z(x))^2]$

Translating using Eq. (39):

$$\mathbb{E}[\hat{Z}(x) - Z(x)] = 0 \iff m(\sum_{i=1}^{n} w_i(x) - 1) = 0 \iff \sum_{i=1}^{n} w_i(x) = 1 \iff \mathbf{1}^T \mathbf{W} = 1 \tag{40}$$

For the minimum of variance, we introduce the augmented vector $\mathbf{Z}_n(x) = [Z(x_1), \dots Z(x_n), Z(x)]$, and the variance can be expressed as:

$$\mathbb{E}[(\hat{Z}(x) - Z(x))^2] = \mathrm{Cov}\left[[\mathbf{W}^T, -1] \cdot \mathbf{Z}_n(x)\right] \tag{41}$$

$$= [\mathbf{W}^T, -1] \, \mathrm{Cov}\left[\mathbf{Z}_n(x)\right] [\mathbf{W}^T, -1]^T \tag{42}$$

In addition, we have

$$\mathrm{Cov}\left[\mathbf{Z}_n(x)\right] = \begin{bmatrix} \mathrm{Cov}\left[\left[Z(x_1) \dots Z(x_n)\right]^T\right] & \mathrm{Cov}\left[\left[Z(x_1) \dots Z(x_n)\right]^T, Z(x)\right] \\ \mathrm{Cov}\left[\left[Z(x_1) \dots Z(x_n)\right]^T, Z(x)\right]^T & \mathbb{V}\mathrm{ar}\left[Z(x)\right] \end{bmatrix} \tag{43}$$

Once expanded, the kriging weights solve then the following optimisation problem:

$$\min_{\mathbf{W}} \mathbf{W}^T \operatorname{Cov}\left[Z(x_1)\dots Z(x_n)\right]\mathbf{W} \tag{44}$$

$$- \operatorname{Cov}\left[\left[Z(x_1)\dots Z(x_n)\right]^T, Z(x)\right]^T \mathbf{W} \tag{45}$$

$$- \mathbf{W}^T \operatorname{Cov}\left[\left[Z(x_1)\dots Z(x_n)\right]^T, Z(x)\right] \tag{46}$$

$$+ \mathbb{V}\mathrm{ar}\left[Z(x)\right] \tag{47}$$

$$\text{s.t.} \mathbf{W}^T\mathbf{1} = \mathbf{1} \tag{48}$$

This leads to

$$\begin{bmatrix}\mathbf{W}\\ m\end{bmatrix} = \begin{bmatrix}\operatorname{Cov}\left[Z(x_1)\dots Z(x_n)\right] & \mathbf{1}\\ \mathbf{1}^T & 0\end{bmatrix}^{-1} \begin{bmatrix}\operatorname{Cov}\left[\left[Z(x_1)\dots Z(x_n)\right]^T, Z(x)\right]^T\\ 1\end{bmatrix} \tag{49}$$

$$= \begin{bmatrix}C(x_1,x_1) & \cdots & C(x_1,x_n) & 1\\ C(x_2,x_1) & \cdots & C(x_2,x_n) & 1\\ \vdots & \ddots & \vdots & \vdots\\ C(x_n,x_1) & \cdots & C(x_n,x_n) & 1\\ 1 & \cdots & 1 & 0\end{bmatrix}^{-1} \begin{bmatrix}C(x_1,x)\\ C(x_2,x)\\ \vdots\\ C(x_n,x)\\ 1\end{bmatrix} \tag{50}$$

Covariance functions

- Desired properties
  - isotropy (?)
  - stationarity
  - semi-definite positiveness
- parametric models of covariance
- examples
- usual hyperparameters estimation

General SUR strategies

Generalities on SUR strategies

Exploration and Space Filling objectives

Contour Estimation

Let ξ be a random process over $\mathbb{X}$, and let us follow what has been done in Bect et al. [2012]. Let ξ_n be the GP constructed using n evaluations of the objective function.

GP of the penalized cost function Δ_α

GP processes

Let $\Delta_\alpha(\mathbf{k},\mathbf{u}) = J(\mathbf{k},\mathbf{u}) - \alpha J^*(\mathbf{u})$. Furthermore, we assume that we constructed a GP on J on the joint space $\mathbb{K}\times\mathbb{U}$, based on a design of n points $\mathcal{X} = \left\{(\mathbf{k}^{(1)},\mathbf{u}^{(1)}),\dots,(\mathbf{k}^{(n)},\mathbf{u}^{(n)})\right\}$, denoted as $(\mathbf{k},\mathbf{u}) \mapsto Y(\mathbf{k},\mathbf{u})$.

As a GP, Y is described by its mean function m_Y and its covariance function $C(\cdot,\cdot)$, while $\sigma_Y^2(\mathbf{k},\mathbf{u}) = C((\mathbf{k},\mathbf{u}),(\mathbf{k},\mathbf{u}))$

$$Y(\mathbf{k},\mathbf{u}) \sim \mathcal{N}\left(m_Y(\mathbf{k},\mathbf{u}), \sigma_Y^2(\mathbf{k},\mathbf{u})\right) \tag{51}$$

Let us consider now the conditional minimiser:

$$J^*(\mathbf{u}) = J(\mathbf{k}^*(\mathbf{u}),\mathbf{u}) = \min_{\mathbf{k}\in\mathbb{K}} J(\mathbf{k},\mathbf{u}) \tag{52}$$

Analogous to J and J^*, we define Y^* as

$$Y^*(\mathbf{u}) \sim \mathcal{N}\left(m_Y^*(\mathbf{u}), \sigma_Y^{2,*}(\mathbf{u})\right) \tag{53}$$

where

$$m_Y^*(\mathbf{u}) = \min_{\mathbf{k}\in\mathbb{K}} m_Y(\mathbf{k},\mathbf{u}) \tag{54}$$

The surrogate conditional minimiser is used in Ginsbourger profiles etc. The α-relaxed difference Δ_α modelled as a GP can then be written as

Considering the joint distribution of $Y(\mathbf{k},\mathbf{u})$ and $Y^*(\mathbf{u}) = Y(\mathbf{k}^*(\mathbf{u}),\mathbf{u})$, we have

$$\begin{bmatrix} Y(\mathbf{k},\mathbf{u}) \\ Y^*(\mathbf{u}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_Y(\mathbf{k},\mathbf{u}) \\ m_Y^*(\mathbf{u}) \end{bmatrix}; \begin{bmatrix} C\left((\mathbf{k},\mathbf{u}),(\mathbf{k},\mathbf{u})\right) & C\left((\mathbf{k},\mathbf{u}),(\mathbf{k}^*(\mathbf{u}),\mathbf{u})\right) \\ C\left((\mathbf{k},\mathbf{u}),(\mathbf{k}^*(\mathbf{u}),\mathbf{u})\right) & C\left((\mathbf{k}^*(\mathbf{u}),\mathbf{u}),(\mathbf{k}^*(\mathbf{u}),\mathbf{u})\right) \end{bmatrix}\right) \tag{55}$$

By multiplying by the matrix $\begin{bmatrix} 1 & -\alpha \end{bmatrix}$ yields

$$\Delta_\alpha(\mathbf{k},\mathbf{u}) \sim \mathcal{N}\left(m_\Delta(\mathbf{k},\mathbf{u}); \sigma_\Delta^2(\mathbf{k},\mathbf{u})\right) \tag{56}$$

$$m_\Delta(\mathbf{k},\mathbf{u}) = m_Y(\mathbf{k},\mathbf{u}) - \alpha m_Y^*(\mathbf{u}) \tag{57}$$

$$\sigma_\Delta^2(\mathbf{k},\mathbf{u}) = \sigma_Y^2(\mathbf{k},\mathbf{u}) + \alpha^2\sigma_{Y^*}^2(\mathbf{k},\mathbf{u}) - 2\alpha C\left((\mathbf{k},\mathbf{u}),(\mathbf{k}^*(\mathbf{u}),\mathbf{u})\right) \tag{58}$$

Assuming that $C((\mathbf{k},\mathbf{u}),(\mathbf{k}',\mathbf{u}')) = s\prod_{i\in\mathcal{I}_\mathbf{k}} \rho_{\theta_i}(\|k_i - k_i'\|)\prod_{j\in\mathcal{I}_\mathbf{u}} \rho_{\theta_j}(\|u_j - u_j'\|)$

$$C\left((\mathbf{k},\mathbf{u}),(\mathbf{k}^*(\mathbf{u}),\mathbf{u})\right) = s\prod_{i\in\mathcal{I}_\mathbf{k}} \rho_{\theta_i}(\|k_i - k_i^*(\mathbf{u})\|)\prod_{j\in\mathcal{I}_\mathbf{u}} \rho_{\theta_j}(0) \tag{59}$$

$$= s\prod_{i\in\mathcal{I}_\mathbf{k}} \rho_{\theta_i}(\|k_i - k_i^*(\mathbf{u})\|) \tag{60}$$

Decomposing the variance δ in Eq. (58), 2 sources of uncertainty:

- σ_Y^2 is the prediction variance of the GP on J, that is directly reduced when additional points are evaluated
- $\sigma_{Y^*}^2$ is the variance of the predicted value of the minimizer.

Approximation of the targeted probability using GP

In order to get to $(\mathbf{k}_p, \alpha_p, p)$. We are going now to use a different notation for the probabilities, taken with respect to the GP: $\mathcal{P}$, to represent the uncertainty encompassed by the GP.

For a given $\mathbf{k} \in \mathbb{K}$, the coverage probability of α-acceptable region, i.e. the probability for $\mathbf{k}$ to be α-acceptable is

$$\Gamma_\alpha(\mathbf{k}) = \mathbb{P}_{\mathbf{U}} \left[J(\mathbf{k}, \mathbf{U}) \leq \alpha J^*(\mathbf{U}) \right] \tag{61}$$

$$= \mathbb{E}_{\mathbf{U}} \left[\mathbb{1}_{J(\mathbf{k},\mathbf{U}) \leq \alpha J^*(\mathbf{U})} \right] \tag{62}$$

As J is not known perfectly, it devolves into a classification problem This classification problem can be approached with a plug-in approach, or a probablistic one:

$$\mathbb{1}_{J(\mathbf{k},\mathbf{u}) \leq \alpha J^*(\mathbf{u})} \approx \mathbb{1}_{m_Y(\mathbf{k},\mathbf{u}) \leq \alpha m_Y^*(\mathbf{u})} \tag{63}$$

$$\mathbb{1}_{J(\mathbf{k},\mathbf{u}) \leq \alpha J^*(\mathbf{u})} \approx \mathcal{P} \left[\Delta_\alpha(\mathbf{k}, \mathbf{u}) \leq 0 \right] = \pi_\alpha(\mathbf{k}, \mathbf{u}) \tag{64}$$

Using the GPs, for a given $\mathbf{k}$, α and $\mathbf{u}$, the probability for our metamodel to verify the inequality is given by Based on those two approximation, the approximated probability Γ is

$$\hat{\Gamma}_{\alpha,n}(\mathbf{k}) = \mathbb{P}_U \left[m_Y(\mathbf{k}, \mathbf{u}) \leq \alpha m_Y^*(\mathbf{u}) \right] \qquad \text{(plug-in)}$$

$$\hat{\Gamma}_{\alpha,n}(\mathbf{k}) = \mathbb{E}_U \left[\mathcal{P} \left[\Delta_\alpha(\mathbf{k}, \mathbf{u}) \leq 0 \right] \right] = \mathbb{E}_U \left[\pi_\alpha(\mathbf{k}, \mathbf{u}) \right] \qquad \text{(Probabilistic approx)} \tag{65}$$

The probability of coverage for the set $\{Y - \alpha Y^*\}$ is π_α, and can be computed using the CDF of the standard normal distribution Φ

$$\pi_\alpha(\mathbf{k}, \mathbf{u}) = \Phi \left(-\frac{m_{\Delta_\alpha}(\mathbf{k}, \mathbf{u})}{\sigma_{\Delta_\alpha}(\mathbf{k}, \mathbf{u})} \right) \tag{66}$$

Finally, averaging over $\mathbf{u}$ yields

$$\hat{\Gamma}_{\alpha,n}(\mathbf{k}) = \mathbb{E}_U \left[\pi_\alpha(\mathbf{k}, \mathbf{u}) \right] = \int_{\mathbb{U}} \pi_\alpha(\mathbf{k}, \mathbf{u}) p(\mathbf{u}) \, \mathrm{d}\mathbf{u} = \int_{\mathbb{U}} \Phi \left(-\frac{m_{\Delta_\alpha}(\mathbf{k}, \mathbf{u})}{\sigma_{\Delta_\alpha}(\mathbf{k}, \mathbf{u})} \right) p(\mathbf{u}) \, \mathrm{d}\mathbf{u} \tag{67}$$

Sources, quantification of uncertainties, and SUR strategy ?

Formally, for a given point $(\mathbf{k}, \mathbf{u})$, the event "the point is α-acceptable" has probability $\pi_\alpha(\mathbf{k}, \mathbf{u})$ and variance $\pi_\alpha(\mathbf{k}, \mathbf{u})(1 - \pi_\alpha(\mathbf{k}, \mathbf{u}))$. Obviously, the points with the highest uncertainty have the highest variance, so have a coverage probability around 0.5.

Random sets

Let us start by introducing diverse tools based around Vorob'ev expectation of closed sets (El Amri [2019], Heinrich et al. [2012].

Let us consider A, a random closed set, such that its realizations are subsets of $\mathbb{X}$, and p is its coverage probability, that is

$$p(\theta) = \mathbb{P}\left[\theta \in A\right], \theta \in \mathbb{X} \tag{68}$$

For $\eta \in [0,1]$, we define the η-level set of p,

$$Q_\eta = \{x \in \mathbb{X} \mid p(x) \geq \eta\} \tag{69}$$

It may seem trivial, but let us still note that those sets are decreasing:

$$0 \leq \eta \leq \xi \leq 1 \implies Q_\xi \subseteq Q_\eta \tag{70}$$

Let μ be a Borel σ-finite measure on $\mathbb{X}$. We define Vorob'ev expectation, as the η^*-level set of A verifying

$$\forall \beta < \eta^* \quad \mu(Q_\beta) \leq \mathbb{E}[\mu(A)] \leq \mu(Q_{\eta^*}) \tag{71}$$

that is the level set of p, that has the volume of the mean of the volume of the random set A.

Margin of uncertainty

Using the quantiles of this level set, we can construct the η-margin of uncertainty, as Dubourg et al. [2011]. Setting the classical level $\eta = 0.05$ for instance, $Q_{1-\frac{\eta}{2}} = Q_{0.975}$ is the set of points whose probability of coverage is higher than 0.975, while $Q_{\frac{\eta}{2}} = Q_{0.025}$ is the set of points whose probability of coverage is higher than 0.025. Obviously, $Q_{1-\frac{\eta}{2}} \subset Q_{\frac{\eta}{2}}$. The complement of $Q_{\frac{\eta}{2}}$ in $\mathbb{X}$, denoted by $Q_{\frac{\eta}{2}}^C$ is the set of points whose probability of coverage is lower than 0.025. The η-margin of uncertainty $\mathbb{M}_\eta$ is defined as the sets of points whose coverage probability is between 0.025 and 0.975.

$$\mathbb{M}_\eta = \left(Q_{1-\frac{\eta}{2}} \cup Q_{\frac{\eta}{2}}^C\right)^C = Q_{1-\frac{\eta}{2}}^C \cap Q_{\frac{\eta}{2}} = Q_{\frac{\eta}{2}} \setminus Q_{1-\frac{\eta}{2}}$$

Recalling the objective, it gives upper bounds and lower bounds of the confidence interval of level η on the probability for each $\mathbf{k}$:

$$\hat{\Gamma}_\alpha^U(\mathbf{k}) = \mathbb{P}_{\mathbf{U}}\left[\theta = (\mathbf{k}, \mathbf{u}) \in Q_{1-\frac{\eta}{2}}\right] \tag{72}$$

$$\hat{\Gamma}_\alpha^L(\mathbf{k}) = \mathbb{P}_{\mathbf{U}}\left[\theta = (\mathbf{k}, \mathbf{u}) \in Q_{\frac{\eta}{2}}\right] \tag{73}$$

SUR Strategies

The main idea behind Stepwise Uncertainty Reduction is to define a criterion, say κ_n, that encapsulates the epistemic uncertainty, and to minimize this criterion, in order to select the next point:

$$x^{n+1} = \arg\max_{x \in \mathbb{X}} \kappa_n(x) \tag{74}$$

where κ_n depends on $Y \mid \mathcal{X}_n$ This approach is suitable for step by step evaluations.

Integrated Mean square criterion

Sacks et al. [1989] Let us consider that we have a kriging model over $\mathbb{X}$ based on a experimental design $\mathcal{X}$, that is denoted $Y \mid \mathcal{X}$

We define the Integrated Mean Square Error (IMSE) as

$$\text{IMSE}(Y \mid \mathcal{X}) = \int_{\mathbb{X}} \sigma^2_{Y|\mathcal{X}}(x)\,\mathrm{d}x \tag{75}$$

where

$$Y \mid \mathcal{X} \sim \mathcal{N}(m_{Y|\mathcal{X}}(x), \sigma^2_{Y|\mathcal{X}}(x)) \tag{76}$$

$$x^{n+1} = \underset{x\in\mathbb{X}}{\arg\min}\ \mathbb{E}_{y\sim Y(x)}\left[\text{IMSE}\left(Y \mid \mathcal{X} \cup \{(x,y)\}\right)\right] \tag{77}$$

So we choose th point minimizing the expected integrated mean square error.

Weighted IMSE

To include a more precise objective than the enrichment of the design, one can add a weight function to the integral, giving the $W-\text{IMSE}$:

$$w-\text{IMSE}(Y \mid \mathcal{X}) = \int_{\mathbb{X}} \sigma^2_{Y|\mathcal{X}}(x)w(x)\,\mathrm{d}x \tag{78}$$

In order to increase the accuracy of the surrogate model around some region of interest, the $w-\text{IMSE}$ can be transformed into

$$w-\text{IMSE}(Y \mid \mathcal{X}) = \int_{\mathbb{X}} \sigma^2_{Y|\mathcal{X}}(x)\mathcal{P}\left[x \in \mathbb{M}_\eta\right]\,\mathrm{d}x \tag{79}$$

where $\mathbb{M}_\eta$ is the η-margin of uncertainty.

UB-LB for $(p, \alpha_p, \mathbf{k}_p)$

Let us assume that we have set a probability $p \in [0,1]$. Let us recall that the triplet $(p, \alpha_p, \mathbf{k}_p)$ verifies

$$\max_{\mathbf{k}} \Gamma_{\alpha_p}(\mathbf{k}) = \Gamma_{\alpha_p}(\mathbf{k}_p) = \mathbb{P}_{\mathbf{U}}\left[J(\mathbf{k}_p, \mathbf{U}) \le \alpha_p J^*(\mathbf{U}) \mid \mathbf{U} = \mathbf{u}\right] = p \tag{80}$$

Let us say that $\bar{\Gamma}$ is the η-upper-bound, while $\underline{\Gamma}$ is the η-lower bounds, so

$$\mathcal{P}\left[\underline{\Gamma}(\mathbf{k}) \le \Gamma_n(\mathbf{k}) \le \bar{\Gamma}(\mathbf{k})\right] = \eta \tag{81}$$

- If $\underline{\Gamma}(\mathbf{k}) > p$, we are too permissive, so we should decrease α
  - by how much ?
- If $\bar{\Gamma}(\mathbf{k}) < p$, we are too conservative, so we should increase α

– by how much again ?

- If $\underline{\Gamma}(\mathbf{k}) < p < \bar{\Gamma}(\mathbf{k})$, reduce uncertainty on $\mathbf{k}_p$

Changing the value of α does not require any further evaluation of the objective function, so can be increased until $\max \hat{\Gamma} = p$? by dichotomy for instance. This $\hat{\mathbf{k}}_p$ is then the candidate.

Criterion: stepwise reduction of the variance of the estimation of $\hat{\Gamma}(\hat{\mathbf{k}}_p) = \max_{\mathbf{k}} \hat{\Gamma}(\hat{\mathbf{k}})$
For a fixed $p \in (0, 1]$, and an initial design $\mathcal{X}$. Set an initial value for $\alpha \geq 1$.

- Define Δ_α, using $Y \mid \mathcal{X}$
- Update α such that $\max \hat{\Gamma}_{\alpha,n} = p$
- Compute measure of uncertainty that we want to reduce:
  - $\bar{\Gamma}_{\alpha,n}(\mathbf{k}) - \underline{\Gamma}_{\alpha,n}(\mathbf{k})$
  - $\pi_\alpha(\mathbf{k}, \mathbf{u})(1 - \pi_\alpha(\mathbf{k}, \mathbf{u}))$

Sampling based criterion

ref Dubourg et al. [2011] Let assume that we derived a criterion κ. And let $f(x) = \frac{\kappa(x)}{\int_{\mathbb{X}} \kappa(u)\,\mathrm{d}u}$. f can be seen as a density. Using an appropriate sampler, we N samples from this criterion:

$$x_i \sim f$$

And find cluster those N samples in order to get p points to evaluate

Application to CROCO

References

J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22 (3):773–793, May 2012. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-011-9241-4.

P. Billingsley. *Probability and Measure*. John Wiley & Sons, 2008.

V. Dubourg, B. Sudret, and J.-M. Bourinet. Reliability-based design optimization using kriging surrogates and subset simulation. *Structural and Multidisciplinary Optimization*, 44(5):673–690, Nov. 2011. ISSN 1615-147X, 1615-1488. doi: 10.1007/s00158-011-0653-8.

M. El Amri. *Analyse d'incertitudes et de Robustesse Pour Les Modèles à Entrées et Sorties Fonctionnelles*. Thesis, Grenoble Alpes, Apr. 2019.

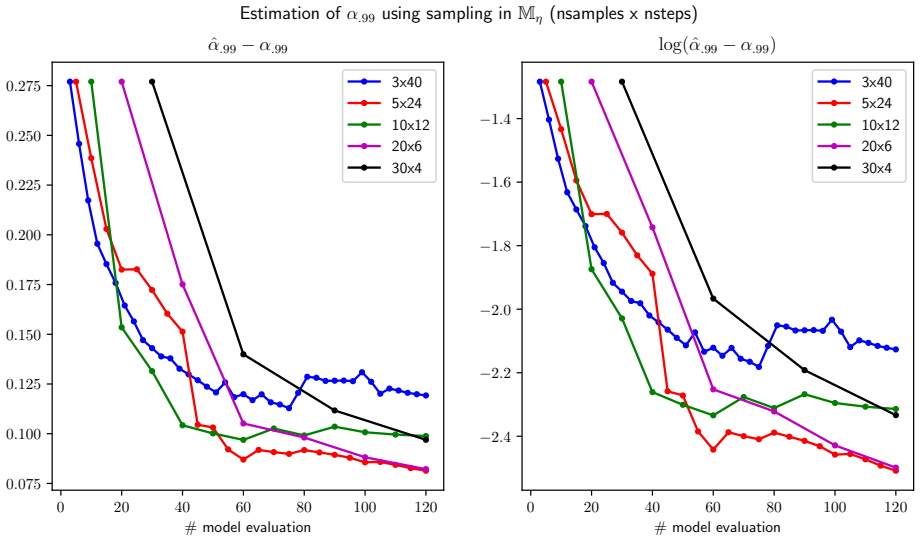

Figure 3

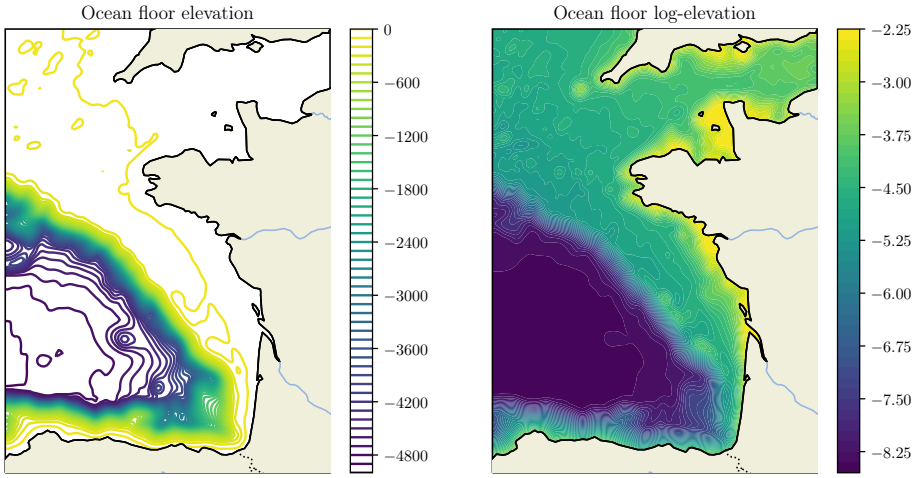

Figure 4 – Ocean floor depth

P. Heinrich, R. S. Stoica, and V. C. Tran. Level sets estimation and Vorob'ev expectation of random compact sets. *Spatial Statistics*, 2:47–61, Dec. 2012. ISSN 22116753. doi: 10.1016/j.spasta.2012.10.001.

T. K. Pogány and S. Nadarajah. On the characteristic function of the generalized normal distribution. *Comptes Rendus Mathematique*, 348(3-4):203–206, Feb. 2010. ISSN 1631073X. doi: 10.1016/j.crma.2009.12.010.

J. Sacks, S. B. Schiller, and W. J. Welch. Designs for Computer Experiments. *Technometrics*, 31(1):41–47, Feb. 1989. ISSN 0040-1706, 1537-2723. doi: 10.1080/00401706.1989.10488474.

A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, Society for Industrial and Applied Mathematics, Philadelphia, Pa, 2005. ISBN 978-0-89871-572-9. OCLC: 265659758.