Notes

Victor Trappler

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

May 21, 2020

Contents

Model selection

Frequentist approach: 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 1.1 – 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 1.2: Let us consider two models, where $\mathbb{Y} = \mathbb{R}$

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

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

Example 1.3: Now let us consider $\mathbb{Y}$ as the space of random vector of dimension n:

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

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

$$\Lambda(y) = \frac{\sup_{\theta\in\Theta_0} \mathcal{L}(\theta; y)}{\sup_{\theta\in\Theta_1} \mathcal{L}(\theta; y)} \tag{1}$$

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

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

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

$$\text{RejReg}_\alpha = \{y \mid -2\log\Lambda(y) > \chi^2_{1-\alpha,r-d}\} \tag{3}$$
$$= \{y \mid \log\Lambda(y) < -\frac{1}{2}\chi^2_{1-\alpha,r-d}\} \tag{4}$$
$$= \{y \mid (\sup_{\theta\in\Theta_0} l(\theta; y) - \sup_{\theta\in\Theta_1} l(\theta; y)) < -\frac{1}{2}\chi^2_{1-\alpha,r-d}\} \tag{5}$$
$$= \{y \mid (\sup_{\theta\in\Theta_1} l(\theta; y) - \sup_{\theta\in\Theta_0} l(\theta; y)) > \frac{1}{2}\chi^2_{1-\alpha,r-d}\} \tag{6}$$
$$\tag{7}$$

Relative Likelihood

Relative Likelihood is defined in Kalbfleisch [1985] as the ratio of the likelihood evaluated at a point θ to the maximum of the likelihood:

$$R(\theta) = \frac{\mathcal{L}(\theta; y)}{\mathcal{L}(\hat{\theta}_{\text{MLE}}; y)} = \frac{\mathcal{L}(\theta; y)}{\sup_{\theta'\in\Theta} \mathcal{L}(\theta'; y)} \tag{8}$$

In that same vein, a Likelihood interval (of level $p \in]0, 1]$) is defined as

$$\mathcal{I}_{\text{Lik}}(p) = \left\{\theta | R(\theta) = \frac{\mathcal{L}(\theta; y)}{\mathcal{L}(\hat{\theta}_{\text{MLE}}; y)} \geq p\right\} \tag{9}$$

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

$$p(\theta|\mathcal{M}, y) = \frac{p(y|\mathcal{M}, \theta)p(\theta)}{p(y|\mathcal{M})} \tag{10}$$

In Eq. (10), $p(y|\mathcal{M}) = \int_\Theta p(y|\mathcal{M}, \theta)p(\theta)\,\mathrm{d}\theta$ 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:

$$\mathrm{BF}(\mathcal{M}_1, \mathcal{M}_2) = \frac{p(y|\mathcal{M}_1)}{p(y|\mathcal{M}_1)} \tag{11}$$

Information criteria

Let $\mathcal{L}$ be the likelihood of the candidate model considered, N the number of observations available, and k its dimension

$$\mathrm{AIC} = -2\log\mathcal{L} + 2k \tag{12}$$
$$\mathrm{BIC} = -2\log\mathcal{L} + k\log(N) \tag{13}$$
$$\tag{14}$$

Model selection and robust estimation

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

$$\mathcal{L}(\theta; y) = \mathcal{L}(k, u, \phi; y) \tag{15}$$

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).

$$\arg\max_{k\in\mathbb{K}} \mathcal{L}(k, u, \phi) = k^*(u) = \arg\max_{k\in\mathbb{K}} \ell(k, u, \phi) \tag{16}$$

Now let us consider the ratio given a value u and

$$-2\log\Lambda(u, \phi') = -2\left(\ell(k, u, \phi) - \ell(k^*(u), u, \phi')\right) \tag{17}$$

Given u, let us define the following likelihoods

$$\mathcal{L}(k; u, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{J(k,u)}{2\sigma^2}\right] \tag{18}$$
$$\mathcal{L}(k = k^*(u); u, \varsigma^2) = \frac{1}{\sqrt{2\pi}\varsigma}\exp\left[-\frac{J^*(u)}{2\varsigma^2}\right] \tag{19}$$
$$\tag{20}$$

Taking the ratio yields

$$\frac{\mathcal{L}(k; u, \sigma^2)}{\mathcal{L}(k^*; u, \varsigma^2)} = \frac{\varsigma}{\sigma}\exp\left[-\frac{1}{2}\left(\frac{J(k,u)}{\sigma^2} - \frac{J^*(u)}{\varsigma^2})\right)\right] \tag{21}$$
$$= \frac{\varsigma}{\sigma}\exp\left[-\frac{1}{2\sigma^2}\left(J(k,u) - \frac{\sigma^2}{\varsigma^2}J^*(u))\right)\right] \tag{22}$$

High values of this ratio indicates that k is not that bad compared to k^*. To retrieve the asymptotic results, we compute twice the negative logarithm of the ratio of the likelihoods, we can define the log ratio ϱ:

$$\varrho(k, u, \sigma, \varsigma) = -2\log\frac{\mathcal{L}(k; u, \sigma^2)}{\mathcal{L}(k^*; u, \varsigma^2)} = \frac{1}{\sigma^2}\left(J(k,u) - \frac{\sigma^2}{\varsigma^2}J^*(u)\right) + 2\log\frac{\sigma}{\varsigma} \tag{23}$$
$$(\text{When } \sigma = 1) = \left(J(k,u) - \frac{1}{\varsigma^2}J^*(u)\right) - 2\log\varsigma \tag{24}$$

Large values of ϱ indicates that k does not give an equivalent model performance-wise , compared to $k^*(u)$.

Problem: what are Θ_0 and Θ_1 ?. Again, for a given u, we test for k, so $\Theta_0 = (\{k\}, \{2\}) \subset \mathbb{K} \times \mathbb{R}^+$.

ALTERNATIVE: $J^*(u)$ adding a gaussian noise of variance ς

$$\mathcal{L}(k; u, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{J(k,u)}{2\sigma^2}\right] \tag{25}$$

$$\mathcal{L}(k = k^*(u); u) = \frac{1}{\sqrt{2\pi}\varsigma} \exp\left[-\frac{J^*(u)}{2(\sigma^2+\varsigma^2)}\right] \tag{26}$$

$$\varrho(k, u, \sigma, \varsigma) = \frac{1}{\sigma^2}\left(J(k,u) - \frac{\sigma^2}{(\sigma^2+\varsigma^2)}J^*(u)\right) + 2\log\frac{\sigma}{\varsigma+\sigma} \tag{27}$$

$$\varrho(k, u, \sigma = 1, \varsigma) = \left(J(k,u) - \frac{1}{(1+\varsigma^2)}J^*(u)\right) - 2\log(1+\varsigma) \tag{28}$$

GP, RR-based family of estimators

Random processes and Gaussian Process Regression

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

$$\begin{array}{rccc} f: & \mathbb{X} \subset \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & f(x) \end{array} \tag{29}$$

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:

$$\begin{array}{rccc} Z: \mathbb{X} \times \Omega & \longrightarrow & \mathbb{R} \\ (x, \omega) & \longmapsto & Z(x, \omega) \end{array} \tag{30}$$

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

$$\hat{Z}(x) = \left[w_1 \dots w_n\right] \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} = \mathbf{W}^T f(\mathcal{X}) = \sum_{i=1}^{n} w_i(x) f(x_i) \tag{31}$$

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. (31):

$$\mathbb{E}[\hat{Z}(x) - Z(x)] = 0 \iff m(\sum_{i=1}^{n} w_i(x) - 1) = 0 \iff \sum_{i=1}^{n} w_i(x) = 1 \iff \mathbf{1}^T \mathbf{W} = 1 \tag{32}$$

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

$$\mathbb{E}[(\hat{Z}(x) - Z(x))^2] = \mathbb{C}\mathrm{ov}\left[[\mathbf{W}^T, -1] \cdot \mathbf{Z}_n(x)\right] \tag{33}$$

$$= [\mathbf{W}^T, -1]\, \mathbb{C}\mathrm{ov}\left[\mathbf{Z}_n(x)\right] [\mathbf{W}^T, -1]^T \tag{34}$$

In addition, we have

$$\mathbb{C}\mathrm{ov}\left[\mathbf{Z}_n(x)\right] = \begin{bmatrix} \mathbb{C}\mathrm{ov}\left[\mathbf{Z}_n^T\right] & \mathbb{C}\mathrm{ov}\left[\mathbf{Z}_n^T, Z(x)\right] \\ \mathbb{C}\mathrm{ov}\left[\mathbf{Z}_n^T, Z(x)\right]^T & \mathbb{V}\mathrm{ar}\left[Z(x)\right] \end{bmatrix} \tag{35}$$

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

$$\min_{\mathbf{W}}\ \mathbf{W}^T\, \mathbb{C}\mathrm{ov}\left[\mathbf{Z}_n\right] \mathbf{W} + \mathbb{V}\mathrm{ar}\left[Z(x)\right] \tag{36}$$

$$- \mathbb{C}\mathrm{ov}\left[\mathbf{Z}_n^T, Z(x)\right]^T \mathbf{W} - \mathbf{W}^T\, \mathbb{C}\mathrm{ov}\left[\mathbf{Z}_n^T, Z(x)\right] \tag{37}$$

$$\text{s.t. } \mathbf{1}^T \mathbf{W} = 1 \tag{38}$$

This leads to

$$\begin{bmatrix} \mathbf{W} \\ m \end{bmatrix} = \begin{bmatrix} \mathbb{C}\mathrm{ov}\left[\mathbf{Z}_n\right] & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{C}\mathrm{ov}\left[\mathbf{Z}_n^T, Z(x)\right]^T \\ 1 \end{bmatrix} \tag{39}$$

$$= \begin{bmatrix} C(x_1, x_1) & \cdots & C(x_1, x_n) & 1 \\ C(x_2, x_1) & \cdots & C(x_2, x_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C(x_n, x_1) & \cdots & C(x_n, x_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} C(x_1, x) \\ C(x_2, x) \\ \vdots \\ C(x_n, x) \\ 1 \end{bmatrix} \tag{40}$$

and

$$\hat{Z}(x) = \begin{bmatrix} C(x_1, x) & C(x_2, x) & \dots & C(x_n, x) \end{bmatrix} \left(\begin{bmatrix} C(x_1, x_1) & \cdots & C(x_1, x_n) \\ C(x_2, x_1) & \cdots & C(x_2, x_n) \\ \vdots & \ddots & \vdots \\ C(x_n, x_1) & \cdots & C(x_n, x_n) \end{bmatrix}^{-1} \right)^T \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix} \tag{41}$$

Covariance functions

1.to write

- Desired properties
  - stationarity One important property usually assumed is the stationarity of the covariance. This property implies that the covariance function for two points x and x' is a function of their difference, that is $C(x, x') = C(|x - x'|)$.
  - semi-definite positiveness
- parametric models of covariance
- examples
- usual hyperparameters estimation

Enrichment strategies for Gaussian Processes

For a unknown function f, a GP is initally constructed based on a design $\mathcal{X} = \{(x_1, f(x_1)), \dots, (x_n, f(x_n))\}$, that consists of n points of $\mathbb{X}$, and their corresponding evaluations. This GP is denoted $Y \mid \mathcal{X}$ and:

$$Y \mid \mathcal{X} \sim \mathcal{N}(m_{Y|\mathcal{X}}(x), \sigma^2_{Y|\mathcal{X}}(x)) \tag{42}$$

The main idea behind Stepwise Uncertainty Reduction is to define a criterion, say κ_n, that measures in a way the uncertainty upon a certain objective associated with the GP and f, and to maximize this criterion, in order to select the next point:

$$x_{n+1} = \arg\max_{x \in \mathbb{X}} \kappa_n(x) = \arg\max_{x \in \mathbb{X}} \kappa(x; Y \mid \mathcal{X}) \tag{43}$$

This new point is then evaluated by f, and the pair is added to the design: $\mathcal{X} \leftarrow \mathcal{X} \cup \{(x_{n+1}, f(x_{n+1})\}$.

Exploration based criteria

We are going to introduce common criteria of enrichment, that aim at exploring the input space $\mathbb{X}$

Maximum variance A measure of uncertainty on the GP is $\max_{x \in \mathbb{X}} \sigma^2_{Y|\mathcal{X}}(x)$, the maximum value of the prediction variance on the space. A simple criterion is to select and evaluate the point corresponding to this maximum of variance:

$$x_{n+1} = \arg\max_{x \in \mathbb{X}} \kappa_n(x) = \arg\max_{x \in \mathbb{X}} \sigma^2_{Y|\mathcal{X}}(x) \tag{44}$$

This criterion by its simplicity is easy to implement, as the prediction variance is cheap to compute given a GP, and does not depend directly on the evaluations of the function $f(x_i)$, uniquely on the distance between the inputs points and the covariance parameters.

Integrated Mean Square Error Sacks et al. [1989] The prediction variance is directly given by $\sigma^2_{Y|\mathcal{X}}$ and represents the uncertainty on the Gaussian regression. To summarize this uncertainty on the whole space $\mathcal{X}$, we define the Integrated Mean Square Error (IMSE) as

$$\text{IMSE}(Y \mid \mathcal{X}) = \int_{\mathbb{X}} \sigma^2_{Y|\mathcal{X}}(x) \, \mathrm{d}x \tag{45}$$

For practical reasons, we can consider to integrate the MSE only on a subset $\mathfrak{X} \subset \mathcal{X}$ that yields

$$\text{IMSE}\left(Y, \mathfrak{X} \mid \mathcal{X}\right) = \int_{\mathcal{X}} \sigma^2_{Y|\mathcal{X}}(x) \mathbb{1}_{\{x \in \mathfrak{X}\}}(x)\, \mathrm{d}x = \int_{\mathfrak{X}} \sigma^2_{Y|\mathcal{X}}(x)\, \mathrm{d}x \tag{46}$$

Unfortunately, exact evaluation of this integral is impossible, so it needs to be approximated using numerical integration, such as Monte-carlo or quadrature rules:

$$\text{IMSE}(Y, \mathfrak{X} \mid \mathcal{X}) \approx \sum_{i=1}^{n_{\text{quad}}} w_i \sigma_{Y|\mathcal{X}}(x_i) \tag{47}$$

For a given $x \in \mathbb{X}$ and an outcome $y \in \mathbb{Y}$, the augmented design is defined as the experimental design, $\mathcal{X} \cup \{(x, y)\}$, and the IMSE of the augmented design is $\text{IMSE}\left(Y \mid \mathcal{X} \cup \{(x, y)\}\right)$. Before the actual experiment though, y is unknown, but we can model it by its distribution given by the GP (per Eq. (42)). So for a given candidate x, the mean prediction error we will get when evaluating x is given by

$$\mathbb{E}_{Y(x)}\Big[\, \text{IMSE}\left(Y \mid \mathcal{X} \cup \{(x, Y(x))\}\right)\Big] \tag{48}$$

where the expectation is to be taken with respect to different realisations of $Y(x)$. As each scenario requires to fit a GP, and to compute the IMSE, a precise evaluation is quite expensive. A strategy ffound for instance in Villemonteix et al. [2006] is to take M possible outcomes for $Y(x)$, corresponding to evenly spaced quantiles of the its distribution. It is maybe important to note that the hyperparameters of the GP should not be reevaluated when augmenting the design, in order to get comparable values for the IMSE.

To enrich the design with the best point, that reduces the most the prediction error, a simple 1-step strategy is to minimize the expectation of Eq. (48).

$$x_{n+1} = \underset{x \in \mathbb{X}}{\arg\min}\ \mathbb{E}_{Y(x)}\Big[\, \text{IMSE}\left(Y \mid \mathcal{X} \cup \{(x, Y(x))\}\right)\Big] \tag{49}$$

Contour and volume estimation

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

$$p(\theta) = \mathbb{P}\left[\theta \in A\right], \theta \in \mathbb{X} \tag{50}$$

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

$$Q_\eta = \{x \in \mathbb{X} \mid p(x) \geq \eta\} \tag{51}$$

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

$$0 \leq \eta \leq \xi \leq 1 \implies Q_\xi \subseteq Q_\eta \tag{52}$$

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

$$\forall \beta < \eta^* \quad \mu(Q_\beta) \leq \mathbb{E}[\mu(A)] \leq \mu(Q_{\eta^*}) \tag{53}$$

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 $Q^C_{\frac{\eta}{2}}$ 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.

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

Estimation of relative-regret quantities

GP of the penalized cost function Δ_α

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 $\mathcal{X} = \left\{(\mathbf{k}^{(1)}, \mathbf{u}^{(1)}), \ldots, (\mathbf{k}^{(n)}, \mathbf{u}^{(n)})\right\}$, 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 $\sigma_Y^2(\mathbf{k}, \mathbf{u}) = C((\mathbf{k}, \mathbf{u}), (\mathbf{k}, \mathbf{u}))$

$$Y(\mathbf{k}, \mathbf{u}) \sim \mathcal{N}\left(m_Y(\mathbf{k}, \mathbf{u}), \sigma_Y^2(\mathbf{k}, \mathbf{u})\right) \tag{54}$$

Let us consider now the conditional minimiser:

$$J^*(\mathbf{u}) = J(\mathbf{k}^*(\mathbf{u}), \mathbf{u}) = \min_{\mathbf{k} \in \mathbb{K}} J(\mathbf{k}, \mathbf{u}) \tag{55}$$

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

$$Y^*(\mathbf{u}) \sim \mathcal{N}\left(m_Y^*(\mathbf{u}), \sigma_Y^{2,*}(\mathbf{u})\right) \tag{56}$$

where

$$m_Y^*(\mathbf{u}) = \min_{\mathbf{k} \in \mathbb{K}} m_Y(\mathbf{k}, \mathbf{u}) = m_Y(\mathbf{k}^*(\mathbf{u})) \tag{57}$$

$$\sigma_Y^{2,*}(\mathbf{u}) = \sigma_Y^{2,*}(\mathbf{k}^*(\mathbf{u})) \tag{58}$$

The surrogate conditional minimiser is used in Ginsbourger et al. [2014] for instance, but other choices could be considered, such as $m_Y(\mathbf{k}^*(\mathbf{u})) - \beta\sigma_Y^{2,*}(\mathbf{k}^*(\mathbf{u}))$. This choice for instance would lead to be more "optimistic" in the estimation of the minimum (i.e. a lower minimum), and in turn, would have a tendency to overestimate the estimated value of α.

The α-relaxed difference defined as $\Delta_\alpha = Y - \alpha Y^*$ is a linear combination of correlated Gaussian processes. Its distribution is Gaussian and can be derived by first considering the joint distribution of $Y(\mathbf{k}, \mathbf{u})$ and $Y^*(\mathbf{u}) = Y(\mathbf{k}^*(\mathbf{u}), \mathbf{u})$:

$$\begin{bmatrix} Y(\mathbf{k}, \mathbf{u}) \\ Y^*(\mathbf{u}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_Y(\mathbf{k}, \mathbf{u}) \\ m_Y^*(\mathbf{u}) \end{bmatrix}; \begin{bmatrix} C\left((\mathbf{k}, \mathbf{u}), (\mathbf{k}, \mathbf{u})\right) & C\left((\mathbf{k}, \mathbf{u}), (\mathbf{k}^*(\mathbf{u}), \mathbf{u})\right) \\ C\left((\mathbf{k}, \mathbf{u}), (\mathbf{k}^*(\mathbf{u}), \mathbf{u})\right) & C\left((\mathbf{k}^*(\mathbf{u}), \mathbf{u}), (\mathbf{k}^*(\mathbf{u}), \mathbf{u})\right) \end{bmatrix}\right) \tag{59}$$

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

$$\Delta_\alpha(\mathbf{k}, \mathbf{u}) \sim \mathcal{N}\left(m_\Delta(\mathbf{k}, \mathbf{u}); \sigma_\Delta^2(\mathbf{k}, \mathbf{u})\right) \tag{60}$$

$$m_\Delta(\mathbf{k}, \mathbf{u}) = m_Y(\mathbf{k}, \mathbf{u}) - \alpha m_Y^*(\mathbf{u}) \tag{61}$$

$$\sigma_\Delta^2(\mathbf{k}, \mathbf{u}) = \sigma_Y^2(\mathbf{k}, \mathbf{u}) + \alpha^2\sigma_{Y^*}^2(\mathbf{k}, \mathbf{u}) - 2\alpha C\left((\mathbf{k}, \mathbf{u}), (\mathbf{k}^*(\mathbf{u}), \mathbf{u})\right) \tag{62}$$

Decomposing the variance σ_Δ^2 in Eq. (62), 3 sources of uncertainty arise:

- 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.
- Assuming a stationary form of the covariance, the third term is directly dependent on the distance between $\mathbf{k}$ and $\mathbf{k}^*(\mathbf{u})$. As the covariance term can be written $C((\mathbf{k}, \mathbf{u}), (\mathbf{k}', \mathbf{u}')) = s\prod_{i \in \mathcal{I}_\mathbf{k}} \rho_{\theta_i}(\|k_i - k_i'\|) \prod_{j \in \mathcal{I}_\mathbf{u}} \rho_{\theta_j}(\|u_j - u_j'\|)$, substituting $\mathbf{k}^*(\mathbf{u})$ for $\mathbf{k}'$ gives

$$C\left((\mathbf{k}, \mathbf{u}), (\mathbf{k}^*(\mathbf{u}), \mathbf{u})\right) = s\prod_{i \in \mathcal{I}_\mathbf{k}} \rho_{\theta_i}(\|k_i - k_i^*(\mathbf{u})\|) \prod_{j \in \mathcal{I}_\mathbf{u}} \rho_{\theta_j}(0) \tag{63}$$

$$= s\prod_{i \in \mathcal{I}_\mathbf{k}} \rho_{\theta_i}(\|k_i - k_i^*(\mathbf{u})\|) \tag{64}$$

This decomposition highlights the fact that the uncertainty measured at a point $(\mathbf{k}, \mathbf{u})$ using σ_Δ^2 will not be reduced completely by evaluating the function at this point, as only the prediction variance σ_Y^2 will be significantly affected in general. In this case, reducing the uncertainty on a slice of constant $\mathbf{k}$ (candidate) will not result necessarily in an evaluation on this slice.

Approximation of the targeted probability using GP

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

$$\Gamma_\alpha(\mathbf{k}) = \mathbb{P}_U \left[J(\mathbf{k}, \mathbf{U}) \leq \alpha J^*(\mathbf{U}) \right] \tag{65}$$

$$= \mathbb{E}_U \left[\mathbb{1}_{J(\mathbf{k},\mathbf{U}) \leq \alpha J^*(\mathbf{U})} \right] \tag{66}$$

As J is not known perfectly, it devolves into a classification problem This classification problem can be approached with a plug-in approach in Eq. (67), or a probabilistic one in Eq. (68):

$$\mathbb{1}_{J(\mathbf{k},\mathbf{u}) \leq \alpha J^*(\mathbf{u})} \approx \mathbb{1}_{m_Y(\mathbf{k},\mathbf{u}) \leq \alpha m_Y^*(\mathbf{u})} \tag{67}$$

$$\mathbb{1}_{J(\mathbf{k},\mathbf{u}) \leq \alpha J^*(\mathbf{u})} \approx \mathcal{P} \left[\Delta_\alpha(\mathbf{k}, \mathbf{u}) \leq 0 \right] = \pi_\alpha(\mathbf{k}, \mathbf{u}) \tag{68}$$

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, we can define different estimations of Γ

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

$$\hat{\Gamma}_{\alpha,n}(\mathbf{k}) = \mathbb{E}_U \left[\mathcal{P} \left[\Delta_\alpha(\mathbf{k}, \mathbf{u}) \leq 0 \right] \right] = \mathbb{E}_U \left[\pi_\alpha(\mathbf{k}, \mathbf{u}) \right] \qquad \text{(Probabilistic approx)} \tag{69}$$

The probability of coverage for the set $\{Y - \alpha Y^* \leq 0\}$ is π_α, and can be computed using the CDF of the standard normal distribution Φ, because Δ_α is a GP, as defined in Eqs. (60) to (62)

$$\pi_\alpha(\mathbf{k}, \mathbf{u}) = \Phi \left(-\frac{m_{\Delta_\alpha}(\mathbf{k}, \mathbf{u})}{\sigma_{\Delta_\alpha}(\mathbf{k}, \mathbf{u})} \right) \tag{70}$$

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

$$\hat{\Gamma}_{\alpha,n}(\mathbf{k}) = \mathbb{E}_U \left[\pi_\alpha(\mathbf{k}, \mathbf{u}) \right] = \int_{\mathbb{U}} \pi_\alpha(\mathbf{k}, \mathbf{u}) p(\mathbf{u}) \, \mathrm{d}\mathbf{u} = \int_{\mathbb{U}} \Phi \left(-\frac{m_{\Delta_\alpha}(\mathbf{k}, \mathbf{u})}{\sigma_{\Delta_\alpha}(\mathbf{k}, \mathbf{u})} \right) p(\mathbf{u}) \, \mathrm{d}\mathbf{u} \tag{71}$$

The estimation of Γ is then maximised with respect to $\mathbf{k}$ to get the candidate probability

$$\max_{\mathbf{k} \in \mathbb{K}} \hat{\Gamma}_{\alpha,n}(\mathbf{k}) \tag{72}$$

By tweaking the value of α, we can get the estimate $\hat{\Gamma}$ to have its maximum equal to the targeted probability. As pointed out earlier the estimation of α_p depends on the estimation of $\hat{\Gamma}$.

α through quantiles Instead of maximizing for each α the estimated $\hat{\Gamma}$, we are now going to derive an approach based on the quantiles: Let $Y(\mathbf{k}, \mathbf{u}) \sim \mathcal{N}(m_Y(\mathbf{k}, \mathbf{u}), \sigma_Y^2(\mathbf{k}, \mathbf{u}))$ be the GP fitted using $\mathcal{X}$. A realisation $y \sim Y$ is then a function from $\mathbb{K} \times \mathbb{U}$ to $\mathbb{R}_*^+$. Again, the plug-in approach is to compute the ratio $m_Y(\mathbf{k}, \mathbf{u})/m_Y^*(\mathbf{u})$ on a large grid for $\mathbf{u}$ and for each $\mathbf{k}$, look for the quantile of order p with respect to $\mathbf{U}$

$$\alpha_{m_Y}(\mathbf{k}) = Q_U \left(p; \ \frac{m_Y(\mathbf{k}, \mathbf{U})}{m_Y^*(\mathbf{U})} \right) \tag{73}$$

the estimation of the relaxation value $\hat{\alpha}_p$ is then the minimal value of the quantiles with respect to $\mathbf{k}$:

$$\hat{\alpha}_p^{\mathrm{PI}} = \min_{\mathbf{k} \in \mathbb{K}} Q_U \left(p; \ \frac{m_Y(\mathbf{k}, \mathbf{U})}{m_Y^*(\mathbf{U})} \right) \qquad \text{(plug-in)}$$

Moreover, as we can sample quite easily from the GP, we can have an idea of the uncertainty in the estimation of $\hat{\alpha}_p$. Let us say that we sampled N function from Y, namely $y^{(i)}$ for $1 \leq i \leq N$. For each of these samples, we can get $\alpha_{y^{(i)}}(\mathbf{k})$, shortened as $\alpha^{(i)}(\mathbf{k})$. Using Monte-Carlo, we can approximate the usual moments for α.

$$\mathbb{E}_Y \left[\alpha_Y(\mathbf{k}) \right] \approx \frac{1}{N} \sum_{i=1}^{N} \alpha^{(i)}(\mathbf{k}) = \alpha^{\mathrm{MC}}(\mathbf{k}) \tag{74}$$

and finally,

$$\hat{\alpha}_p^{\mathrm{MC}} = \min_{\mathbf{k} \in \mathbb{K}} \alpha^{\mathrm{MC}}(\mathbf{k}) \tag{75}$$

Iterative procedure The general Section 2.5.1

- Find a measure of uncertainty that depends as a function of $\mathbf{k} \times \mathbb{U}$.
- Find the point that reduces the most the uncertainty on this slice
- Update

At the step n:

- First, using the GP, α_p is estimated using Monte-carlo and samples from the GP, giving $\hat{\alpha}_{n,.99}$
- We choose the candidate $\mathbf{k}_{\text{candidate}}$ as the minimizer of the sampled quantiles: $\mathbf{k}_{\text{candidate}} = \arg\min_{\mathbf{k}} \alpha^{\text{MC}}(\mathbf{k})$.
- We optimize the wIMSE, defined as $\text{IMSE}(Y_n \mid \mathcal{X}, \{\mathbf{k}_{\text{candidate}}\} \times \mathbb{U})$ to get the next point to sample

On Figure Section 2.5.2 is shown the procedure applied on BHs, based on a initial design of 30 points. The wIMSE is computed by sampling a 50-point LHS on $\{\mathbf{k}_{\text{candidate}} \times \mathbb{U}\}$. The estimation using Monte-carlo seems to show convergence towards the true value, while the plug-in approach does not evolve much with the iterations

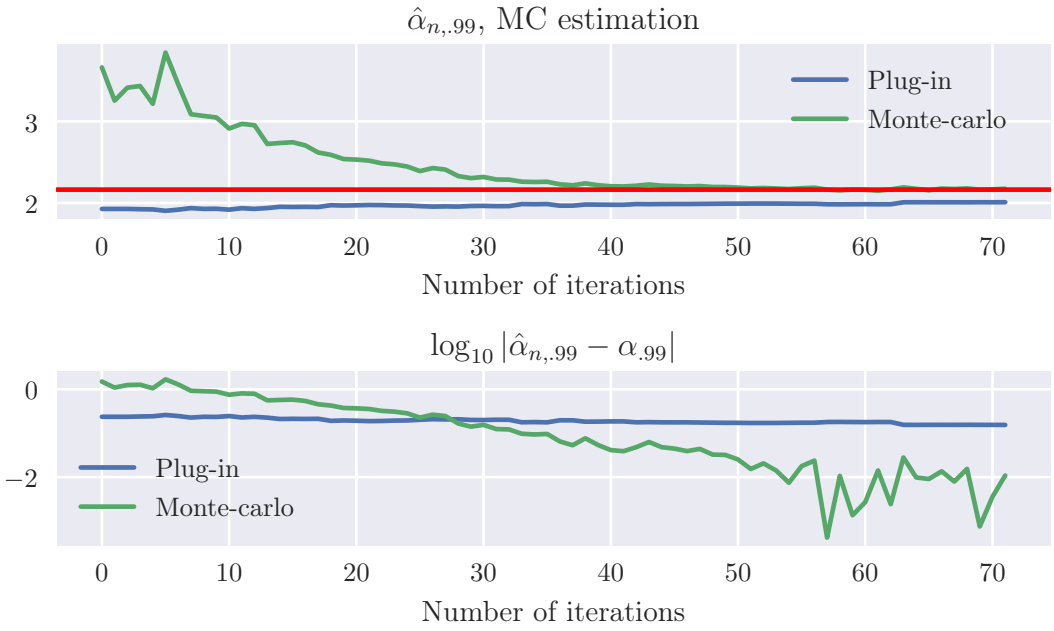

Figure 1 – $\hat{\alpha}_{n,.99}$ estimated by MC

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.

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

$$\hat{\Gamma}_\alpha^{UB}(\mathbf{k}) = \mathbb{P}_U \left[\theta = (\mathbf{k}, \mathbf{u}) \in Q_{1-\frac{\eta}{2}} \right] \tag{76}$$

$$\hat{\Gamma}_\alpha^{LB}(\mathbf{k}) = \mathbb{P}_U \left[\theta = (\mathbf{k}, \mathbf{u}) \in Q_{\frac{\eta}{2}} \right] \tag{77}$$

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

$$\max_{\mathbf{k}} \Gamma_{\alpha_p}(\mathbf{k}) = \Gamma_{\alpha_p}(\mathbf{k}_p) = \mathbb{P}_{\mathbf{u}}\left[J(\mathbf{k}_p, \mathbf{u}) \leq \alpha_p J^*(\mathbf{u}) \mid \mathbf{u} = \mathbf{u}\right] = p \tag{78}$$

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

$$\mathcal{P}\left[\underline{\Gamma}(\mathbf{k}) \leq \Gamma_n(\mathbf{k}) \leq \bar{\Gamma}(\mathbf{k})\right] = \eta \tag{79}$$

- 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

This technique is described in 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 can generate N iid samples from this criterion $\{x_i\}_{1 \leq i \leq N}$

However, as N should be large, there is no point in evaluating all the samples x_i. This goes by the statistical reduction of the samples: This can be done by KMeans algorithm,

Application to CROCO

By comparing Fig. 3 and Fig. 4, we can see that the optimization focuses primarly on the region of low depth, as the region north of Spain remains untouched. In the English channel, the optimization finds the true value.

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.

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.

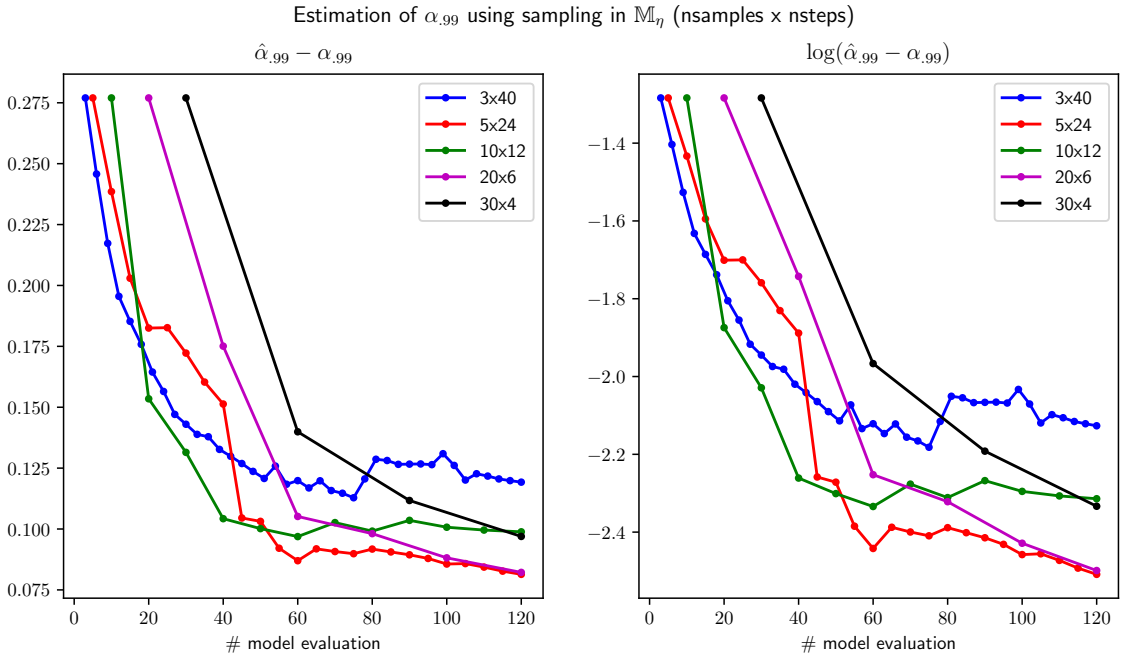

Figure 2

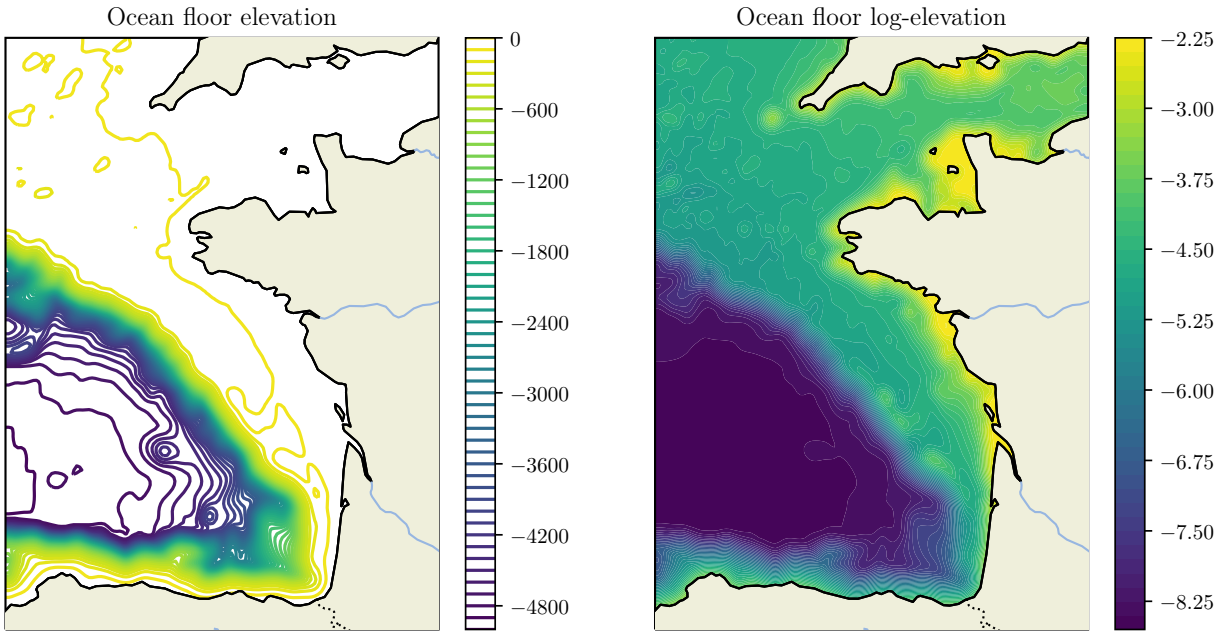

Figure 3 – Ocean floor depth for CROCO

D. Ginsbourger, J. Baccou, C. Chevalier, F. Perales, N. Garland, and Y. Monerie. Bayesian Adaptive Reconstruction of Profile Optima and Optimizers. *SIAM/ASA Journal on Uncertainty Quantification*, 2 (1):490–510, Jan. 2014. ISSN 2166-2525. doi: 10.1137/130949555.

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.

J. G. Kalbfleisch. *Probability and Statistical Inference*. Springer Texts in Statistics. Springer New York, New York, NY, 1985. ISBN 978-1-4612-7009-6 978-1-4612-1096-2. doi: 10.1007/978-1-4612-1096-2.

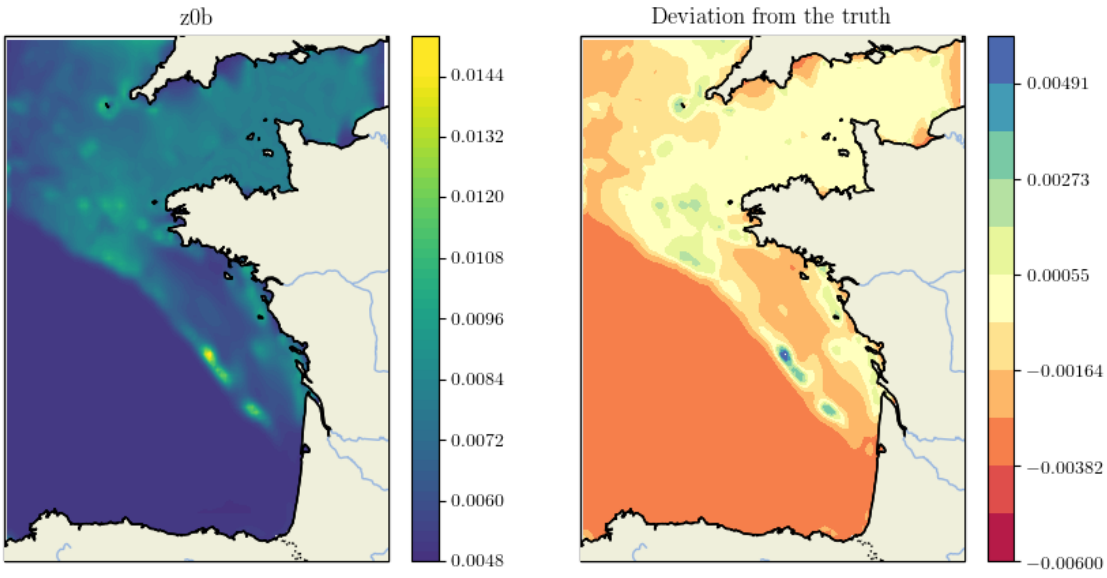

Figure 4 – Optimization result on z0b. The bottom friction has been initialized at 5×10^{-3}, and the true value (used to generate the observations) is 8×10^{-3}.

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.

J. Villemonteix, E. Vazquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *arXiv:cs/0611143*, Nov. 2006.