CHAPTER 1

INVERSE PROBLEM AND CALIBRATION

Contents

à garder/écrire ?

gardé un sommaire très détaillé temporairement

TODO LIST

1.1 Introduction

In this chapter we will first lay the ground for developing the general ideas behind calibration, by introducing the notions of models and forward and inverse problems in Section 1.2. This implies also a short review of notions of probability theory. Calibration will be defined in Section 1.3 as the optimization of a certain objective function: Maximum likelihood estimation in a frequentist setting, or posterior maximization using Bayes' theorem. In practice, for large-scale applications, the optimization is performed using gradient-descent, and the computational cost of gradient computation can be overcome by adjoint method, as described in Section 1.4. Finally, we are going to discuss two aspects related to calibration, namely model selection in Section 1.5 and the influence of nuisance parameters and model misspecification in calibration in Section 1.6.

1.2 Forward, inverse problems and probability theory

general theory on UQ ?

Running the model using numerical tools is useful to grasp a better understanding of the physical phenomena, or to forecast them. On the other hand, observing and comparing the physical phenomena and the output of the model brings information on ?? our modelling. More specifically, models are designed with a set of parameters, that aims at representing specific physical quantities. Those quantities have to be properly known, in order have a meaningful output when evaluating the model. Model calibration or parameter estimation has been widely treated in the literature, either from a statistical and probabilistic point of view using Bayesian inference, or from a variational point of view. using optimal control

pas super bien dit

1.2.1 Model space data space and forward problem

In order to describe accurately a physical system, we have to define the notion of ~~a~~ models, and will be following [Tar05] approach to define inverse problems. 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 is that a model is not only the *forward operator*, but must also include the parameter space

> **Definition 1.2.1 – Model:** A model $\mathfrak{M}$ is defined as a pair composed of a *forward operator* $\mathcal{M}$, and a *parameter space* Θ
>
> $$\mathfrak{M} = (\mathcal{M}, \Theta) \tag{1.1}$$

> 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 choose 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 1.2.2:** 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 1.2.3:** 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.

1.2.2 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 $\mathcal{M}$.

$$\begin{array}{rccc} \mathcal{M}: & \Theta & \longrightarrow & \mathbb{Y} \\ & \theta & \longmapsto & \mathcal{M}(\theta) \end{array} \tag{1.2}$$

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 consequences, i.e. the output in the data space.

1.2.3 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.

??

?? General approaches for inverse ??

non

However, a purely deterministic approach for the inverse problem is doomed to ~~fail~~ be suboptimal 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.

!! ajouter schema !!

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 [Bil08] for a more thorough description.

1.2.4 Notions of probability theory

?? join subsections ? more brief ? ??

1.2.4.a 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 1.2.4 – 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
>
> $$\mathbb{P}[A] = \int_A \mathrm{d}\mathbb{P}(\omega) = \mathbb{P}[\{\omega; \omega \in A\}] \tag{1.3}$$
>
> 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 $A \mid B$ and its probability is
>
> $$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \tag{1.4}$$

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 $(S, \mathcal{B}(S))$. def ? notation ?

> **Definition 1.2.5 – 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*. When $S \subseteq \mathbb{R}^p$ for $p \geq 1$, X is sometimes called a random vector

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

$$\mathbb{E}[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega) \tag{1.5}$$

Using the Definition 1.2.5, the probability of an event A can be seen as the expectation of the indicator function of A:

$$\begin{aligned} \mathbb{1}_A : \quad \Omega & \longrightarrow \quad \{0,1\} \\ \omega & \longmapsto \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ if } \omega \notin A \end{cases} \end{aligned} \tag{1.6}$$

and it follows that

$$\mathbb{E}[\mathbb{1}_A] = \int_\Omega \mathbb{1}_A\,\mathrm{d}\mathbb{P}(\omega) = \int_A \mathrm{d}\mathbb{P}(\omega) = \mathbb{P}[A] \tag{1.7}$$

As we defined the notion of a r.v. in Definition 1.2.5 as a measurable function from $\Omega \to S$, we can now focus on the measurable sets through X, by using in a sense the change of variable $x = X(\omega)$.

Definition 1.2.6 – 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

$$\mathbb{P}[X \in A] = \mathbb{P}_X[A] = \int_A \mathrm{d}\mathbb{P}_X(\omega) = \int_{X^{-1}(A)} \mathrm{d}\mathbb{P}(\omega) = \mathbb{P}[X^{-1}(A)] = \mathbb{P}\left[\{\omega\,;\, X(\omega) \in A\}\right] \tag{1.8}$$

Similarly, for any measurable function h, the expectation taken with respect to a specific random variable X is

$$\mathbb{E}_X[h(X)] = \int_\Omega h(X(\omega))\,\mathrm{d}\mathbb{P}_X(\omega) \tag{1.9}$$

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.

1.2.4.b 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 1.2.7 – 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.

important ?

$$F_X(x) = \mathbb{P}\left[X \leq x\right] = \mathbb{P}_X\left[\,]-\infty, x]\,\right] \tag{1.10}$$

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:

$$Q_X(p) = \inf\{q : \, F_X(q) \geq p\} \tag{1.11}$$

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

$$\mathbb{P}[X \in A] = \int_A \mathrm{d}\mathbb{P}_X(\omega) = \int_A f(x)\,\mathrm{d}x \tag{1.12}$$

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 1.2.8: 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

$$\delta_{x_0}(x) = \begin{cases} +\infty \text{ if } x = x_0 \\ 0 \text{ elsewhere} \end{cases} \qquad \text{and} \quad \int_S \delta_{x_0}(x)\,\mathrm{d}x = 1 \tag{1.13}$$

Example 1.2.9: 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

$$F_X(x) = \begin{cases} 0 \text{ if } x < 1 \\ 0.5 \text{ if } 1 \leq x < 2 \\ 0.5 + \frac{x-2}{8} \text{ if } 2 \leq x < 4 \\ 1 \text{ if } 4 \leq x \end{cases} \tag{1.14}$$

and its pdf (as a generalized function)

$$p_X(x) = \frac{1}{2}\delta_1(x) + \frac{1}{4}\mathbb{1}_{\{2 \leq x < 4\}}(x) \tag{1.15}$$

Definition 1.2.10 – Moments of a r.v. and L^s spaces: Let X be a random variable. The moment of order s is defined as $\mathbb{E}\left[X^s\right]$, and the centered moment of order s is defined as

$$\mathbb{E}[(X - \mathbb{E}[X])^s] = \int (X(\omega) - \mathbb{E}[X])^s \,\mathrm{d}\mathbb{P}(\omega) = \int (x - \mathbb{E}[X])^s \cdot p_X(x)\,\mathrm{d}x \tag{1.16}$$

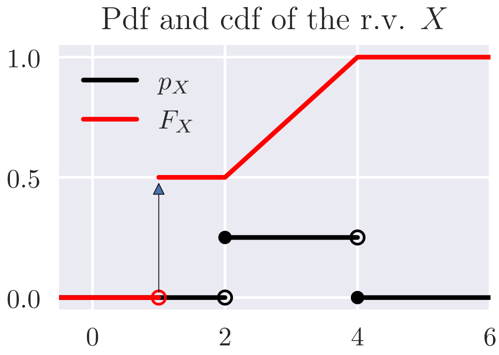

Figure 1.1: Cdf and Pdf of X defined in Example 1.2.9. The arrow indicates Dirac's delta function

> 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:
>
> $$\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{V}\mathrm{ar}[X] \geq 0 \tag{1.17}$$

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

1.2.4.c Real-valued random vectors

Most of the definitions for a random variable extends component-wise to the random vectors:

> **Definition 1.2.11 – 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 of the components
>
> $$\mathbb{E}[X] = [\mathbb{E}[X_1], \dots, \mathbb{E}[X_p]] \tag{1.18}$$
>
> The cdf of X at the point $x = [x_1, \dots x_p]$ is
>
> $$\begin{aligned} F_X(x) = F_{X_1,\dots,X_p}(x_1, \dots, x_p) &= \mathbb{P}\left[X_1 \leq x_1, \cdots, X_p \leq x_p\right] \\ &= \mathbb{P}\left[\bigcap_{i=1}^{p} \{\omega;\ X_i(\omega) \leq x_i\}\right] \end{aligned} \tag{1.19}$$
>
> 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:
>
> $$p_X(x) = p_{X_1,\dots,X_p}(x_1, \dots, x_p) = \frac{\partial^p F_X}{\partial x_1 \cdots \partial x_p}(x) \tag{1.20}$$
>
> and $\int_S p_{X_1,\dots,X_p}(x_1, \dots, x_p)\, \mathrm{d}(x_1, \dots, x_p) = 1$

c'est pas comme ça que tu l'as défini avant

Extending the variance to vectors brings the covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ of X, defined as

$$\Sigma = \mathbb{C}\text{ov}(X) = \mathbb{C}\text{ov}[X, X] = \mathbb{E}\left[(X - \mathbb{E}[X])\,(X - \mathbb{E}[X])^T\right] = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T \tag{1.21}$$

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

et la cov pour 2 v.a ?

$$\mathbb{C}\text{ov}\,[X, Y] = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T\right] = \mathbb{E}[XY^T] - \mathbb{E}[X]\mathbb{E}[Y]^T \tag{1.22}$$

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

$$p_Y(y) = \int_{\mathbb{R}} p_{Y,Z}(y, z)\,\mathrm{d}z \quad \text{and} \quad p_Z(z) = \int_{\mathbb{R}} p_{Y,Z}(y, z)\,\mathrm{d}y \tag{1.23}$$

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

$$p_{Y|Z}(y \mid z) = \frac{p_{Y,Z}(y, z)}{p_Z(z)} \tag{1.24}$$

allowing us to rewrite the marginals as

$$p_Y(y) = \int_{\mathbb{R}} p_{Y|Z}(y \mid z)p_Z(z)\,\mathrm{d}z = \mathbb{E}_Z\left[p_{Y|Z}(y \mid z)\right] \tag{1.25}$$

$$p_Z(z) = \int_{\mathbb{R}} p_{Z|Y}(z \mid y)p_Y(y)\,\mathrm{d}y = \mathbb{E}_Y\left[p_{Z|Y}(z \mid y)\right] \tag{1.26}$$

The influence of one (or a set of) random variable(s) over another can be measured with the conditional probabilities. Indeed, if the state of information on a random variable does not change when observing another one, ~~this leads to think that~~ the observed one provides no information on the other. This leads to the definition of independence. bof

Definition 1.2.12 – Independence: Let $A, B \in \mathcal{F}$. Those two events are ~~deemed~~ said 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, $p_{Y,Z}(y, z) = p_Y(y)p_Z(z)$ Speaking in terms of conditional probabilities, this can be written as $p_{Y|Z}(y, z) = p_Y(y)$. If Y and Z are independent, $\mathbb{C}\text{ov}[Y, Z] = 0$. The converse if false in general.

One of the most well known distribution is the normal distribution (or Gaussian).

Example 1.2.13 – 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

$$p_X(x) = \phi(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \tag{1.27}$$

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

$$p_X(x) = (2\pi)^{-\frac{\overset{p}{\boxed{d}}}{2}}|\Sigma|^{-1}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{1.28}$$

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}. Examples of pdf of Gaussian normal distribution are displayed on Fig. 1.2.

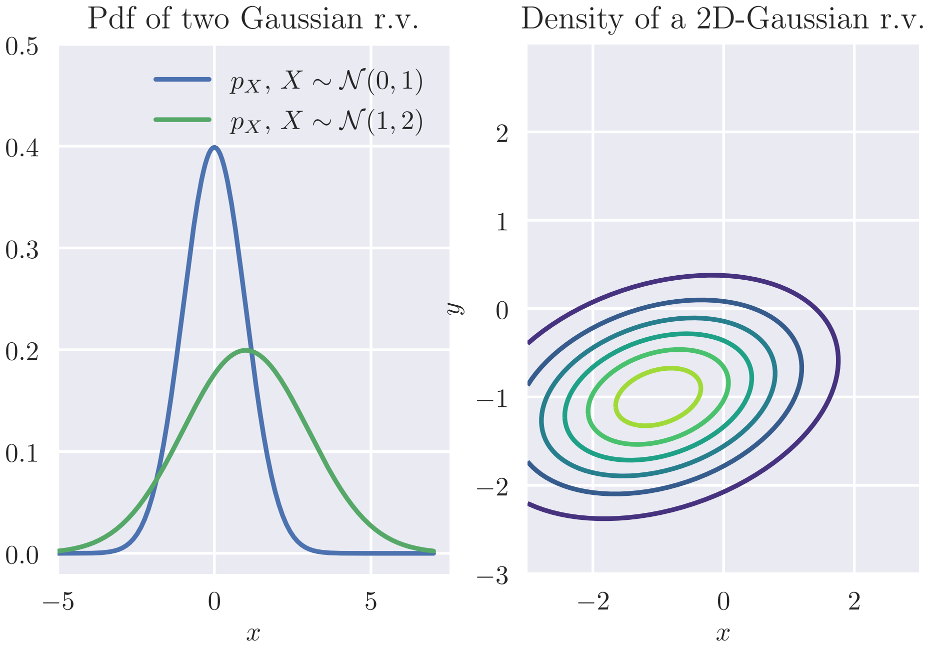

Figure 1.2: Densities of 1D Gaussian distributed r.v. (left), and density of a 2D Gaussian r.v.

↑ densities or pdf

Comparing two real valued random variables X and Y, can be done in a lot of ways, from

mieux transition?

↑ ça peut aller ☺

Definition 1.2.14 – KL–divergence: The Kullback-Leibler divergence, introduced in [KL51] is a measure of dissimilarity between two distributions, based on information-theoretic considerations. Let X, Y be r.v. and p_X and p_Y their densities. If the image measure of Y is absolutely continuous with respect to X, that is for all A,

$\mathbb{P}_X[A] = 0 \implies \mathbb{P}_Y[A] = 0$

$$D_{\mathrm{KL}}(p_X \| p_Y) = \int p_X(x) \log \frac{p_X(x)}{p_Y(x)}\,\mathrm{d}x \tag{1.29}$$

1.2.4.d Bayes' Theorem

The classical Bayes' theorem is directly a consequence of the definition of the conditional probabilities in Definition 1.2.4, or by considering the pdf of r.v. in Definition 1.2.11.

Theorem 1.2.15 – 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.

pdf

$$p_{Y|Z}(y \mid z) = \frac{p_{Z|Y}(z \mid y) p_Y(y)}{p_Z(z)} = \frac{p_{Z|Y}(z \mid y) p_Y(y)}{\int p_{Z,Y}(z, y)\,\mathrm{d}y} \propto p_{Z|Y}(z \mid y) p_Y(y) \tag{1.30}$$

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, $Z \mid Y$ can be seen as the forward problem. Bayes' theorem allow us to "swap" the conditioning, and get information on $Y \mid Z$, that can be seen as the inverse problem. ✓

1.3 Parameter inference

1.3.1 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 $\vartheta \in \Theta_{\mathrm{real}}$ that is unknown:

mais, tu changes de notations

$$\begin{aligned} \mathscr{M} : \quad \Theta_{\mathrm{real}} &\longrightarrow \mathbb{Y} \\ \vartheta &\longmapsto \mathscr{M}(\vartheta) \end{aligned} \tag{1.31}$$

The physical reality yields some observations $\mathscr{M}(\vartheta)$ that lie in $\mathbb{Y}$, a subset of $\mathbb{R}^p$. [KO01, HKC+04] Given a model $(\mathcal{M}, \Theta)$, the main objective of calibration is to find $\hat{\theta}$ such that the reduced model $(\mathcal{M}, \{\hat{\theta}\})$ represents as accurately as possible the physical system, and thus matches as closely the data $\mathscr{M}(\vartheta)$. Let us assume that $\vartheta \in \Theta$, so we can rewrite the link between the reality and the model as

c'est plutôt réduction de dimension que estimation de paramètre

$$\mathscr{M}(\vartheta) = \mathcal{M}(\vartheta) + \epsilon(\vartheta) \in \mathbb{Y} \subseteq \mathbb{R}^p \tag{1.32}$$

Hm... peut-être pas. c'est pas très clair. Insiste que $\mathscr{M}$ c'est la réalité et explique d'où vient $\mathcal{M}$

The difference $\epsilon(\vartheta) = \mathscr{M}(\vartheta) - \mathcal{M}(\vartheta)$ is the error between the physical model and the model, called sometimes the misfit, or the residual error. This error is unknown and encompasses different sources of uncertainties, such as measurement errors, or model bias (with respect to the reality). To deal with this unknown, we are going to model it as a sample of a random variable, leading us to treat the data obtained as a random sample as well

dans l'ordre :-)

From the diverse assumptions we can make upon this sampled random variable, we can then treat the calibration procedure as a problem of estimation of parameters of a ~~random variable. In this thesis,~~ we are looking for an extremum estimator, estimator defined as the optimizer of a given objective function.

why?

1.3.2 Frequentist inference, MLE

1.3.2.a Formulation of the MLE

As mentioned before, we can model the observations as a random variable, say Y, and assume that those were generated using a parametric family of distributions, whose densities are

$$\{y \mapsto p_Y(y;\theta); \theta \in \Theta\} \tag{1.33}$$

The choice has been made to keep explicit the dependency on θ. For instance, we can use the hypothesis that the residual are normally distributed with a given covariance matrix Σ. As we assume that $\mathbb{Y} \subseteq \mathbb{R}^p$, Y is a random vector distributed as

$$Y \sim \mathcal{N}(\mathcal{M}(\theta), \Sigma) \tag{1.34}$$

whose one sample is $y = \mathscr{M}(\vartheta)$. Now, instead of looking at the densities of Eq. (1.33) as functions mapping the sample space $\mathbb{Y}$ to $\mathbb{R}$, we may look at it instead as a function of θ, as the observations $y \in \mathbb{Y}$ do not vary. We can then define the likelihood function and its associated extremum estimator.

Definition 1.3.1 – 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 $\mathcal{L}$:

$$\mathcal{L}(\cdot;y) : \theta \mapsto p_Y(y;\theta) = \mathcal{L}(\theta;y) \tag{1.35}$$

$$= (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\mathcal{M}(\theta)-y)^T\Sigma^{-1}(\mathcal{M}(\theta)-y)\right) \tag{1.36}$$

If $\Sigma = \mathrm{diag}(\sigma_1^2, \dots, \sigma_p^2)$, the likelihood can be written as the product of 1D Gaussians:

$$\mathcal{L}(\theta;y) = \left(\prod_{i=1}^{p}\frac{1}{\sqrt{2\pi}\sigma_i}\right)\exp\left(\sum_{i=1}^{p} -\frac{(\mathcal{M}(\theta)_i - y_i)^2}{2\sigma_i^2}\right) \tag{1.37}$$

$$= \prod_{i=1}^{p}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(\mathcal{M}(\theta)_i - y_i)^2}{2\sigma_i^2}\right) \tag{1.38}$$

~~pas besoin après!~~

oui ?

> Based on the likelihood function, we can define the *Maximum Likelihood Estimator*, or *MLE*, that maximizes the likelihood defined above:
>
> $$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta; y) \tag{1.39}$$

Equivalently, for practical and numerical reasons, the maximization of the likelihood is replaced by the minimization the negative log-likelihood:

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\theta \in \Theta} -\log \mathcal{L}(\theta; y) = \arg\min_{\theta \in \Theta} -\sum_{i=1}^{p} \log p_{Y_i|\theta}(y_i \mid \theta) \tag{1.40}$$

where

$$-\log \mathcal{L}(\theta; y) = \frac{1}{2}(\mathcal{M}(\theta) - y)^T \Sigma^{-1} (\mathcal{M}(\theta) - y) + \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| \tag{1.41}$$

As the optimization is performed on θ, we can remove the constant terms of the objective, and rewrite the cost function as a $L2$ norm in Eq. (1.42).

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \arg\min_{\theta \in \Theta} \frac{1}{2}(\mathcal{M}(\theta) - y)^T \Sigma^{-1} (\mathcal{M}(\theta) - y) \\ &= \arg\min_{\theta \in \Theta} \frac{1}{2}\|\mathcal{M}(\theta) - y\|^2_{\Sigma^{-1}} \end{aligned} \tag{1.42}$$

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. (1.34). Other choices of the sampling distribution Eq. (1.34) will result in different objective functions. To reduce the sensitivity on outliers, some authors such as [RSNN15] introduce Student or Laplace distributed errors, or specifically designed norm such as the Huber norm [Hub11].

If the covariance matrix is diagonal, the residual errors are then un~~c~~correlated, thus indepedent due to their Gaussian nature as defined in Eq. (1.34). The likelihood can be rewritten as the product of densities evaluated at the different samples y_i, sampled from their true distribution Y. The MLE can be rewritten as following by monotone transformation of the objective function:

f2

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\theta \in \Theta} -\frac{1}{p}\sum_{i=1}^{p} \log p_{Y_i}(y_i; \theta) \approx \arg\min_{\theta \in \Theta} -\mathbb{E}_Y\left[p_Y(y; \theta)\right] \tag{1.43}$$

$$= \arg\min_{\theta \in \Theta} D_{\text{KL}}\left(p_Y \| p_Y(\cdot; \theta)\right) \tag{1.44}$$

In this situation, the MLE minimizes the empirical KL-divergence between the true distribution of the observations, and the sampling distribution of the observation (that depends on θ).

à garder/écrire ?

3.2.b (Asymptotic) Properties of the MLE

Frequentist approaches do not take into account any information we could have on the possible values taken on θ, except for its parameter space Θ.

1.3.3 Bayesian Inference

In Bayesian inference, the uncertainty present on θ is modelled by considering it as a random variable. Instead of having a precise value for θ, albeit unknown, we assume that we have a *prior distribution* on θ, denoted p_θ, that represents the initial state of belief upon the parameter, prior to any experiment and observations.The choice of this prior distribution will be discussed later. After the experiment, whose sampling distribution is given by the likelihood, the prior distribution is updated to reflect the new state of belief upon the parameter. The Gaussian likelihood in Eq. (1.34) for the frequentist approach can be almost be rewritten as is in the Bayesian setting, just by conditioning Y with θ. Eq. (1.34) becomes

$$Y \mid \theta \sim \mathcal{N}(\mathcal{M}(\theta), \Sigma) \tag{1.45}$$

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

$$p_{\theta|Y}(\theta \mid y) = \frac{p_{Y|\theta}(y \mid \theta) p_\theta(\theta)}{p_Y(y)} = \frac{\mathcal{L}(\theta; y) p_\theta(\theta)}{p_Y(y)} \propto \mathcal{L}(\theta; y) p_\theta(\theta) \tag{1.46}$$

As the data is usually fixed, the denominator of Eq. (1.46) is constant, and the posterior is usually evaluated up to a multiplicative constant.

> **Definition 1.3.2 – Model Evidence:** The model evidence, (or marginal likelihood, integrated likelihood) is defined as the distribution of the data marginalised over the parameters.
>
> $$p_Y(y) = \int_\Theta p_{Y,\theta}(y, \theta) \,\mathrm{d}\theta = \int_\Theta p_{Y|\theta}(y \mid \theta) p_\theta(\theta) \,\mathrm{d}\theta \tag{1.47}$$
>
> This quantity depends implicitly on the underlying mathematical model $\mathfrak{M} = (\mathcal{M}, \Theta)$. Comparing evidence of different models allows for the comparison of those different models. However, computing the model evidence requires the expensive evaluation of an integral over the whole parameter space, and no analytical form is available except for trivial cases. Specific techniques for this evaluation are reviewed in [FW11].

mais tu viens de dire qu'on s'en fiche!

1.3.3.a Posterior inference

This posterior distribution is central in Bayesian analysis, as it gathers all the information we have on the parameter, given the observed data. Given Eq. (1.46), evaluating the posterior density at a point requires the evaluation of the model evidence, that is an expensive integral. To bypass this evaluation, several techniques have been developed to get samples from a unnormalized arbitrary function. One of the most well-known method is based on the construction of a Markov-chain whose stationary

?

state is the searched posterior. The classical MCMC algorithm is require the use of a proposal density A lot of refinement of this methods are available in the literature in order to better tackle the high-dimensionality of the parameter space, or to improve the mixing of the sampled MC chain. One refinement important to mention is Hamiltonian Monte-Carlo [Han01, Bet17], that improves the performance of the chain by using the value of the gradient of the log-posterior distribution. Obtaining this gradient (although for a different purpose) is discussed in Section 1.4.

1.3.3.b Bayesian Point estimates

Bayesian point estimation refer to point estimation of the parameter θ, using the posterior distribution $p_{\theta|Y}$.[LC06] provides a more complete overview of this subject. This can be done by defining the *Bayesian risk*, that is the expectation of a Bayesian loss functions $L : \Theta \times \Theta \to \mathbb{R}^+$ under the posterior distribution. A Bayesian point estimate is then a minimizer of this Bayesian risk.

refondre
en partant
de L
du choix
de

$$\theta_L = \underset{\theta' \in \Theta}{\arg\min}\ \mathbb{E}_{\theta|Y}\left[L(\theta', \theta) \mid y\right] \tag{1.48}$$

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 MSE : $\theta' \mapsto \mathbb{E}_{\theta|Y}\left[(\theta' - \theta)^2\right]$. Finally, the value corresponding to the Minimum Mean Squared Error is

$$\hat{\theta}_{\text{MMSE}} = \underset{\theta' \in \Theta}{\arg\min}\ \mathbb{E}_{\theta|Y}\left[(\theta' - \theta)^2 \mid y\right] \tag{1.49}$$

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

$$\hat{\theta}_{\text{MMSE}} = \mathbb{E}_{\theta|Y}[\theta \mid y] = \int_{\Theta} \theta \cdot p_{\theta|Y}(\theta \mid y)\,\mathrm{d}\theta \tag{1.50}$$

In order to compute $\hat{\theta}_{\text{MMSE}}$, it is easier to compute directly the mean of the posterior samples obtained via posterior inference, than to solve the minimization problem in Eq. (1.49).

Posterior Mode: the MAP

Taking $L(\theta', \theta) = \delta_\theta(\theta')$, the dirac delta function defined in Eq. (1.13), one can show that the minimizer of $\mathbb{E}_{\theta|Y}[L(\theta', \theta)]$ is the mode of the posterior distribution, and is called the *Maximum A Posteriori* (MAP):

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \underset{\theta' \in \Theta}{\arg\min}\ \mathbb{E}_{\theta|Y}\left[\delta_\theta(\theta') \mid y\right] = \underset{\theta' \in \Theta}{\arg\min}\ -p_{\theta|Y}(\theta' \mid y) \\ &= \underset{\theta' \in \Theta}{\arg\max}\ p_{\theta|Y}(\theta' \mid y) = \underset{\theta' \in \Theta}{\arg\max}\ \mathcal{L}(\theta'; y)p_\theta(\theta')\end{aligned} \tag{1.51}$$

commencer
la def d'un max
que le min !

rappeler $\mathcal{L}$
def

We chose to use a generalized function as a tool to link the MAP to Bayesian point estimate, but it is sometimes introduced as the limit of $0-1$ loss functions. In [BD19], the authors shows that this claim does not always hold, unless some conditions are met. One interesting fact about the MAP, is that its evaluation does not require the full knowledge of the posterior distribution, nor samples to evaluate the integral of Eq. (1.50). We can resort to classical optimization techniques for this evaluation. Similarly to the likelihood, taking the negative logarithm leads to the following minimization problem.

$$\hat{\theta}_{\mathrm{MAP}} = \operatorname*{arg\,min}_{\theta' \in \Theta} -\ell(\theta'; y) - \log p_\theta(\theta') \tag{1.52}$$

1.3.3.c Choice of a prior distribution

As seen in the application of Bayes' theorem in Eq. (1.46), the prior has a preponderant role in the formulation of the posterior distribution. Indeed, this prior distribution represents the current state of knowledge on the value of the parameter, before any experiment. This comes usually from an expert opinion, or some reasonable assumptions about the nature of θ.

Let us assume for instance that we have a Gaussian prior for p_θ, such that $\theta \sim \mathcal{N}(\theta_b, B)$ where B is called the background covariance error matrix, and a Gaussian model for the errors as well, the MAP can be written as

$$\hat{\theta}_{\mathrm{MAP}} = \operatorname*{arg\,min}_{\theta \in \Theta} \frac{1}{2}\|\mathcal{M}(\theta) - y\|^2_{\Sigma^{-1}} + \frac{1}{2}\|\theta - \theta_b\|^2_{B^{-1}} \tag{1.53}$$

Adding a Gaussian prior for the parameter comes down to adding a L^2 regularization term to the optimization problem, also called Tikhonov regularization [TA77]. This expression is very analoguous to the state estimation in the 3D-Var method in Data assimilation Other choices of priors leads to other regularizations, such as the lasso regularization [Tib11] that is a consequence for choosing θ that follows a priori a Laplace distribution of mean 0.

The choice of a prior distribution has an influence on the inference of the parameter and its point estimation. Where there is no knowledge on the parameter beforehand, one can try to choose a non-informative prior in order to try to mitigate its effect. One can for instance choose a "flat" prior over the parameter space, but this can lead to *improper prior*, in the sense that they do not integrate to 1. However, improper priors, though mathematically questionable, do not necessarily lead to improper posterior, allowing for the usual Bayesian analysis of the quantity.

mouais...

If $\Theta = \mathbb{R}^p$, an improper non-informative prior is $p_\theta(\theta) \propto 1$, as it should be invariant by translation. In this case, the MAP estimation is equivalent to the MLE, as the prior in Eq. (1.52) is constant with respect to θ'. When $\Theta = \mathbb{R}_+$, the prior distribution should be invariant by multiplication by a positive constant, so $p_\theta(\theta) \propto \frac{1}{\theta}\mathbb{1}_{\theta>0}$ (i.e. flat in the log scale), and that leads to a regularization term of the form $\log(\theta)$.

pas hyper clair je trouve

All in all, when looking for the MAP or the MLE, parameter estimation boils down to the minimization of a well chosen objective function, that measures the misfit between the output of the numerical model and the observations. This cost function will be written J in the following, to match the notation of data assimilation. As mentioned before, the MAP does not require the full knowledge of the posterior distribution $p_{\theta|Y}$, as "only" an optimization is required.

In this context of calibration, we can then summarize the estimation as a minimization problem, where J represents some kind of distance between $\mathcal{M}(\theta)$ and the observations.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} J(\theta) \tag{1.54}$$

1.4 Calibration using adjoint-based optimization

Point estimates in this context take the form of extremum estimators, that is an extremum of some given objective function J. This function takes the form of the log-likelihood, or the log-posterior for the MLE and MAP, but other misfits can be considered, such as optimal transport based metrics. The formulation is then quite simple, but the problem of efficient optimization remains. For differentiable problems, most of minimization instances are solved using gradient based methods, such as gradient descent, quasi-newton methods. However, this implies to be able to compute efficiently the gradient of the cost function J with respect to the parameter: $\frac{\mathrm{d}J}{\mathrm{d}\theta}$. The straightforward way, is to compute the gradient using finite differences. Let us suppose that $\theta = (\theta_1, \cdots \theta_n)$, and e_i is 0 for all its component except the ith one which is 1. The gradient can be approximated by the usual 1st order forward finite-difference scheme, as displayed in Eq. (1.55).

$$\frac{\mathrm{d}J}{\mathrm{d}\theta} \approx \left[\frac{J(\theta+\epsilon e_1)-J(\theta)}{\epsilon}, \frac{J(\theta+\epsilon e_2)-J(\theta)}{\epsilon}, \dots, \frac{J(\theta+\epsilon e_n)-J(\theta)}{\epsilon}\right] \quad \text{for } \epsilon \ll 1 \tag{1.55}$$

In addition to the run of the model at θ, we have to evaluate the model n times, for each one of the coordinate of θ. If this is feasible in practice for low dimensional problems, this is impossible for large problems that cumulate more than hundreds of parameters. Nevertheless, different methods can be used to compute the gradient, atleast approximately for optimization purpose: for instance, [Bou15] uses Simultaneous Perturbation Stochastic Approximation to approximate the gradient using only one additional run, indepedently on the number of parameters.

In geophysical applications, parameter estimation and the subsequent optimization is usually performed by deriving the adjoint equation in order to get the exact gradient for a relatively reasonable cost. This gradient is used afterward in optimization methods such as conjuguate gradient, or BFGS for instance. This procedure is common in data assimilation, as shown in [DL91, DL92, HMR+10, CMMV13].

?? approche lagrangien ou TLM? ??

To derive the adjoint equations, we will first rewrite the cost function as a function of the forward operator and the parameter: $J(\theta) = J(\mathcal{M}(\theta), \theta)$: The estimation of the parameter can be written as the following constrained optimisation problem:

$$\begin{aligned} \min_{\theta \in \Theta} J(\theta) &= J(\mathcal{M}(\theta), \theta) \\ \text{such that } \mathcal{F}(\mathcal{M}(\theta), \theta) &= 0 \end{aligned} \tag{1.56}$$

where the constraint on $\mathcal{F}$ signifies that the model is admissible.

Introducing the Lagrange multiplier $\lambda \in \mathbb{Y}$, we can write the Lagrangian $\mathscr{L}$

$$\mathscr{L}(\theta, y, \lambda) = J(y, \theta) - \lambda^T \mathcal{F}(y, \theta) \tag{1.57}$$

The equivalent unconstrained minimization problem Eq. (1.56) is then

$$\min_{\theta, y, \lambda} \mathscr{L}(\theta, y, \lambda) \tag{1.58}$$

The first-order condition of optimality for the Lagrangian: $\frac{\partial \mathscr{L}}{\partial \theta} = \frac{\partial \mathscr{L}}{\partial y} = \frac{\partial \mathscr{L}}{\partial \lambda} = 0$ translates into the optimality condition, adjoint equation and the state equation:

$$\frac{\partial \mathscr{L}}{\partial \lambda} = -\mathcal{F}(y, \theta) = 0 \tag{State equation}$$

$$\frac{\partial \mathscr{L}}{\partial y} = \frac{\partial J}{\partial y} - \lambda^T \frac{\partial \mathcal{F}}{\partial y} = 0 \tag{Adjoint equation}$$

$$\frac{\partial \mathscr{L}}{\partial \theta} = \frac{\partial J}{\partial \theta} - \lambda^T \frac{\partial \mathcal{F}}{\partial \theta} = 0 \tag{Optimality condition}$$

1.5 Model selection

So far, we have discussed the calibratation of a specific model given some observations, thus solving an inverse problem and finding $\hat{\theta}$ as an extremum of a specified objective function.

!! Occam's razor !!

1.5.1 Likelihood ratio test

If this specific calibrated model is efficient, it can be also quite complex, and it can be interesting to test if a "simpler" model would give similar performances, or at least show a decrease in performances not statistically significant. One of the well-known test is the Likelihood-ratio test, that test if two *nested models* are equivalent: Let us consider two nested models: $\mathfrak{M}_1 = (\mathcal{M}_1, \Theta_1)$, $\mathfrak{M}_2 = (\mathcal{M}_2, \Theta_2)$, such that $\mathcal{M}_1 = \mathcal{M}_2 = \mathcal{M}$ and $\Theta_2 \subsetneq \Theta_1$. In this case, $\mathfrak{M}_2$ represents the simpler model, with a reduced parameter space, while $\mathfrak{M}_1$ is the more general model. Recalling the notion of model dimension in Remark 1.2.2, $\mathfrak{M}_1$ has dimension r, and $\mathfrak{M}_2$ has dimension d with $r > d$.

Préciser l'entité avec (, et 2 !

Under the null hypothesis, the two models are equivalent, that is the "smaller" parameter space is enough to represent the inverse problem: $\theta \in \Theta_2$. The alternative

def

hypothesis is that $\theta \in \Theta_1$. The likelihood ratio is defined as the ratio of the largest values taken by the likelihood on their respective parameter space, value that is assumed to be attained as $\hat{\theta}_1$ and $\hat{\theta}_2$.

$$\Lambda(y) = \frac{\sup_{\theta\in\Theta_2} \mathcal{L}(\theta; y)}{\sup_{\theta\in\Theta_1} \mathcal{L}(\theta; y)} = \frac{\mathcal{L}(\hat{\theta}_2; y)}{\mathcal{L}(\hat{\theta}_1; y)} \tag{1.59}$$

? on n'a pas $\Lambda = 1$ où ?

Under the null hypothesis (that the models are equivalent), $-2\log\Lambda$, (sometimes called the deviance) follows asymptotically (as the number of observations becomes large) a χ^2 distribution, whose degrees of freedom is given by the difference of dimensionality between the two models:

faut la définir ou pas ?

$$-2\log\Lambda(y) \xrightarrow{\text{d}} \chi^2_{r-d} \tag{1.60}$$

By denoting $\chi^2_{r-d}(1-\alpha)$ the quantile of order $1-\alpha$ of the χ^2 distribution of $r-d$ degrees of freedom, the asymptotic rejection region of level α is:

ça veut dire quoi ?

$$\text{RejReg}_\alpha = \{y \mid -2\log\Lambda(y) > \chi^2_{r-d}(1-\alpha)\} \tag{1.61}$$

$$= \left\{y \mid (\sup_{\theta\in\Theta_1} l(\theta; y) - \sup_{\theta\in\Theta_2} l(\theta; y)) > \frac{1}{2}\chi^2_{r-d}(1-\alpha)\right\} \tag{1.62}$$

$$= \left\{y \mid J(\hat{\theta}_2) - J(\hat{\theta}_1) > \frac{1}{2}\chi^2_{r-d}(1-\alpha)\right\} \tag{1.63}$$

on comprend pas bien pourquoi tu racontes ça

As a basis for comparison, when $\Theta \subset \mathbb{R}$, $r-d = 1$ and $\chi^2_1(1-0.05) = 3.84$

1.5.2 Relative Likelihood

Relative Likelihood, as defined in [Kal85], is the ratio of the likelihood evaluated at a point θ to the maximal value of the likelihood:

$$R(\theta) = \frac{\mathcal{L}(\theta; y)}{\mathcal{L}(\hat{\theta}_{\text{MLE}}; y)} = \frac{\mathcal{L}(\theta; y)}{\sup_{\theta'\in\Theta} \mathcal{L}(\theta'; y)} \tag{1.64}$$

This function allows for comparing the plausibility of the value θ, compared to the MLE. Extending a bit the relative likelihood, we can define in the same vein the likelihood interval of level $p \in]0, 1]$, defined as

$$\mathcal{I}_{\text{Lik}}(p) = \left\{\theta \mid R(\theta) = \frac{\mathcal{L}(\theta; y)}{\mathcal{L}(\hat{\theta}_{\text{MLE}}; y)} \geq p\right\} \tag{1.65}$$

This interval can be understood as a kind of confidence interval for the MLE. p can be arbitrarily set to arbitrary threshold, but it can also be chosen specifically in order to avoid the rejection region of a likelihood ratio test with certain confidence. When comparing the models $(\mathcal{M}, \{\theta\})$ and $(\mathcal{M}, \Theta)$, $R(\theta)$ is their likelihood ratio, the complement of the rejection region of Eq. (1.61) written as a likelihood interval becomes

$$\mathcal{I}_{\text{Lik}}\left(\exp\left(-\frac{1}{2}\chi^2_{\dim(\Theta)}(1-\alpha)\right)\right) = \left\{\theta \mid R(\theta) \geq \exp\left(-\frac{1}{2}\chi^2_{\dim(\Theta)}(1-\alpha)\right)\right\} \tag{1.66}$$

The values in this set generate models that are statistically equivalent to the model comprising the MLE as its calibrated parameter. Again, for 1 dimensional models, and the confidence level of .05, $\exp\left(-\frac{1}{2}\chi^2_{\dim(\Theta)}(1-\alpha)\right) = \exp\left(-\frac{1}{2}\chi^2_1(.95)\right) \approx 0.15$

j'ai un peu lâché l'affaire sur le 1.5 je sens bien que tu listes des "métriques" mais je sais pas pourquoi, ni ce qu'elles représentent. p.ê des petites illustrations seraient utiles.

1.5.3 AIC and non-nested models

The AIC: Akaike's Criterion Information is defined as

$$\mathrm{AIC}(\mathfrak{M}) = -2\left(\sup_{\Theta} \log \mathcal{L}\right) + 2\dim\Theta = -2\log\mathcal{L}(\hat{\theta}_{\mathrm{MLE}}) + 2\dim\Theta \tag{1.67}$$

where $\mathcal{L}$ is the likelihood function for the model $\mathfrak{M} = (\mathcal{M}, \Theta)$. Small values of the AIC tend towards a better representation of the reality, whilst avoiding overfitting due to the penalization term. Contrary to the Likelihood ratio test that compares two nested models, the AIC let us compare non-nested models: the difference between the AIC can be considered as an indication toward one or the other model [BA04]. When both models have the same number of dimension, the difference of AIC is equal to the deviance.

1.5.4 Bayesian model comparison and model averaging

For a model $\mathfrak{M}$, the model evidence as introduced in Definition 1.3.2, is the likelihood marginalized over the parameter space, and will be written $p_{Y|\mathfrak{M}}$. This evidence represents how likely have the data y been generated using the statistical model $\mathfrak{M}$.

The Bayes' factor is defined as the ratio of the evidence of the two models.

$$\mathrm{BF}(\mathfrak{M}_1, \mathfrak{M}_2) = \frac{p_{Y|\mathfrak{M}_1}(y \mid \mathfrak{M}_1)}{p_{Y|\mathfrak{M}_1}(y \mid \mathfrak{M}_2)} \tag{1.68}$$

Quite similarly as the AIC introduced in Section 1.5.3, the Bayes' factor is usually compared to specific values, allowing us to conclude roughly on how strong does the data favors $\mathfrak{M}_1$: for [KR95], if $\log \mathrm{BF}(\mathfrak{M}_1, \mathfrak{M}_2) > 2$, there is "decisive" evidence for $\mathfrak{M}_1$ against $\mathfrak{M}_2$, while lower values will indicate "strong", "substantial" and then "not worth mentioning" difference of evidence.

?? parler de model averaging ? AIC weights et Bayesian factor as weights ??

1.6 Parametric model misspecification and nuisance parameters

We introduced earlier the mathematical model $(\mathcal{M}, \Theta)$, and based our analysis on the fact that the "target model", i.e. the reality is $(\mathscr{M}, \Theta_{\mathrm{real}} = \Theta)$, so the parameter spaces are the same. In practice, the parameter space Θ does not contain all the parameters needed to run the forward model, but represents the space of the parameters of interest, or calibration parameters. In addition to them, some other parameters are at play, that we are going to call the *environmental parameters*, or *uncertain parameters* written $u \in \mathbb{U}$. These parameters come from instance from ~~the physical~~ external forcings.

The environmental parameters introduced before, are in fact a bit different from the calibration parameters generically introduced as θ. Bayesian framework and more specifically Bayesian update of the prior by the likehihood puts the emphasis on the update of the information on the *parameter of interest*. However the environmental parameters have an inherent variability. In that sense, it may not be worth spending time and resources to infer these parameter values. Moreover, we can only get information on the environmental conditions used to generate the observations. We aim at letting this parameter stay free, and at finding a good value of the calibration parameter θ, without doing any inference on u.

1.6.1 Model misspecification

We have then a family of models, indexed by u: $\{(\mathcal{M}(\cdot, u), \Theta);\ u \in \mathbb{U}\}$ that have to be compared with the reality: $(\mathscr{M}, \Theta_{\text{real}})$, that has been "evaluated" at a value $\vartheta \notin \Theta$, so for each u, we have an inverse problem. The main issue that arises is that every choice of u for $\mathcal{M}$ will lead to another inverse problem: When looking at an extremum estimator, $\hat{\theta}(u) = \arg\min_{\theta \in \Theta} J(\theta, u)$, it becomes clear that the estimation of the parameter θ depends on u.

Similarly to the characterization of the MLE inf Section 1.3.2, for each u, $\hat{\theta}_{\text{MLE}}(u)$ minimizes the empirical KL-divergence between the true distribution of the observations and the misspecified sampling model, and can be seen as the "best" value given u. However, the asymptotic properties of the MLE are slightly different as described in [Whi82]. ah?

BIBLIOGRAPHY

[BA04] Kenneth P. Burnham and David R. Anderson. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304, November 2004.

[BD19] Robert Bassett and Julio Deride. Maximum a Posteriori Estimators as a Limit of Bayes Estimators. *Mathematical Programming*, 174(1-2):129–144, March 2019.

[Bet17] Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]*, January 2017.

[Bil08] Patrick Billingsley. *Probability and Measure.* John Wiley & Sons, 2008.

[Bou15] Martial Boutet. *Estimation Du Frottement Sur Le Fond Pour La Modélisation de La Marée Barotrope.* PhD thesis, Université d'Aix Marseille, 2015.

[CMMV13] Frédéric Couderc, Ronan Madec, Jérôme Monnier, and Jean-Paul Vila. *Dassfow-Shallow, Variational Data Assimilation for Shallow-Water Models: Numerical Schemes, User and Developer Guides.* PhD thesis, University of Toulouse, CNRS, IMT, INSA, ANR, 2013.

[DL91] S. K. Das and R. W. Lardner. On the estimation of parameters of hydraulic models by assimilation of periodic tidal data. *Journal of Geophysical Research*, 96(C8):15187, 1991.

[DL92] S. K. Das and R. W. Lardner. Variational parameter estimation for a two-dimensional numerical tidal model. *International Journal for Numerical Methods in Fluids*, 15(3):313–327, August 1992.

[FW11] Nial Friel and Jason Wyse. Estimating the evidence – a review. *arXiv:1111.1957 [stat]*, November 2011.

[Han01] K. Hanson. MARKOV CHAIN MONTE CARLO POSTERIOR SAMPLING WITH THE HAMILTONIAN METHOD. Technical Report LA-UR-01-1016, Los Alamos National Lab., NM (US), February 2001.

c'est bon, crie pas!

[HKC+04] Dave Higdon, Marc Kennedy, James C. Cavendish, John A. Cafeo, and Robert D. Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466, 2004.

[HMR+10] Marc Honnorat, Jérôme Monnier, Nicolas Rivière, Étienne Huot, and François-Xavier Le Dimet. Identification of equivalent topography in an open channel flow using Lagrangian data assimilation. *Computing and Visualization in Science*, 13(3):111–119, March 2010.

[Hub11] Peter J. Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

[Kal85] J. G. Kalbfleisch. *Probability and Statistical Inference*. Springer Texts in Statistics. Springer New York, New York, NY, 1985.

[KL51] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951.

[KO01] Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, January 2001.

[KR95] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[LC06] Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.

[RSNN15] Vishwas Rao, Adrian Sandu, Michael Ng, and Elias Nino-Ruiz. Robust data assimilation using L_1 and Huber norms. *SciRate*, November 2015.

[TA77] Andrei Tikhonov and Vasily Arsenin. *Solutions of Ill-Posed Problems*, volume 14. 1977.

[Tar05] Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, Society for Industrial and Applied Mathematics, Philadelphia, Pa, 2005. OCLC: 265659758.

[Tib11] Robert Tibshirani. Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

[Whi82] Halbert White. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1–25, 1982.