CHAPTER 1

INVERSE PROBLEM AND CALIBRATION

Contents

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 an 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

Running a simulation using numerical tools is useful to grasp a better understanding of the physical phenomena, or to forecast them. On the other hand, when observing and comparing the measurements and the output of the numerical simulation, we can quantify the mismatch between the two and tune some parameters involved in the computations. Indeed, these parameters represent different physical quantities or processes that are unresolved at the model's scale (friction of the ocean bed, or $k - \epsilon$ turbulence models for instance). A proper estimation of these parameters has to be performed in order to guarantee 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 likelihood-based methods and Bayesian inference, or from a *variational* point of view by defining proper objective functions. To match those two approaches, we will first review the problem from a probabilistic point of view, in order to define properly some appropriate objective functions and introduce tools from optimal control theory to optimize them.

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 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.

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 $\mathbb{Y}$ consists in all the physically acceptable results of the physical experiment. 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 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 observation of the physical process and on the knowledge of the forward operator, as illustrated Fig. 1.1.

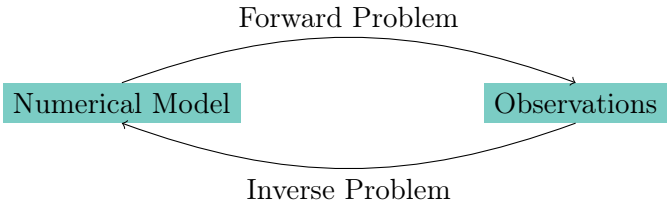

Figure 1.1: Forward and Inverse problem diagram

This is done by directly comparing the output of the forward operator, and trying to reduce the mismatch between the observed data and the model prediction.

However, a purely deterministic approach for the inverse problem is doomed to underperform: 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 [Bil08] for a more thorough description.

1.2.4 Notions of probability theory

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, that is the set of all possible events (upon which we can assign a probability).

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:

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

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

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

In most of this thesis, 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 real random variables.

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 $S = \mathbb{R}$.

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:

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

  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), or *density* of X and is denoted p_X. As $\mathbb{P}[X \in S] = 1$, it follows trivially that $\int_S f(x)\,\mathrm{d}x = 1$. One can verify that if F_X is derivable, then its derivative is the density of the r.v. :

$$\frac{\mathrm{d}F_X}{\mathrm{d}x}(x) = p_X(x) \tag{1.13}$$

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

$$\delta_{x_0}(x) = \begin{cases} +\infty \text{ if } x = x_0 \\ 0 \text{ elsewhere} \end{cases} \qquad \text{and} \quad \int_S \delta_{x_0}(x)\,\mathrm{d}x = 1 \tag{1.14}$$

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

$$F_X(x) = \begin{cases} 0 \text{ if } x < 1 \\ 0.5 \text{ if } 1 \leq x < 2 \\ 0.5 + \frac{x-2}{8} \text{ if } 2 \leq x < 4 \\ 1 \text{ if } 4 \leq x \end{cases} \tag{1.15}$$

and its pdf (as a generalized function)

$$p_X(x) = \frac{1}{2}\delta_1(x) + \frac{1}{4}\mathbb{1}_{\{2\leq x<4\}}(x) \tag{1.16}$$

The pdf and cdf are shown Fig. 1.2.

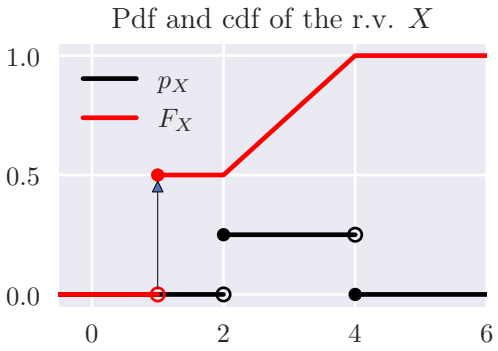

Figure 1.2: Cdf and Pdf of X defined in Example 1.2.9. The arrow indicates Dirac's delta function

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 $\mathbb{E}[X^s]$, and the centered moment of order s is defined as

$$\mathbb{E}[(X - \mathbb{E}[X])^s] = \int (X(\omega) - \mathbb{E}[X])^s \, \mathrm{d}\mathbb{P}(\omega) = \int (x - \mathbb{E}[X])^s \cdot p_X(x) \, \mathrm{d}x \tag{1.17}$$

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

$$\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{V}\mathrm{ar}[X] \geq 0 \tag{1.18}$$

These definition above hold for real-valued random variables, so 1D r.v., but can be extended for random vectors.

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

$$\mathbb{E}[X] = [\mathbb{E}[X_1], \ldots, \mathbb{E}[X_p]] \tag{1.19}$$

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

$$F_X(x) = F_{X_1,\dots,X_p}(x_1, \dots, x_p) = \mathbb{P}\left[X_1 \leq x_1, \cdots, X_p \leq x_p\right] = \mathbb{P}\left[\bigcap_{i=1}^{p} \{\omega;\ X_i(\omega) \leq x_i\}\right] \tag{1.20}$$

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:

$$p_X(x) = p_{X_1,\dots,X_p}(x_1, \dots, x_p) = \frac{\partial^p F_X}{\partial x_1 \cdots \partial x_p}(x) \tag{1.21}$$

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

For two random vectors X and Y, the (cross-)covariance matrix of X and Y is defined as

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

and based on this definition, we can extend the notion of variance to vectors. The covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ of X, is defined to be

$$\Sigma = \mathbb{C}\mathrm{ov}(X) = \mathbb{C}\mathrm{ov}[X, X] = \mathbb{E}\left[(X - \mathbb{E}[X])\,(X - \mathbb{E}[X])^T\right] = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T \tag{1.23}$$

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

$$p_Y(y) = \int_{\mathbb{R}} p_{Y,Z}(y, z)\, \mathrm{d}z \quad \text{and} \quad p_Z(z) = \int_{\mathbb{R}} p_{Y,Z}(y, z)\, \mathrm{d}y \tag{1.24}$$

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

$$p_{Y|Z}(y \mid z) = \frac{p_{Y,Z}(y, z)}{p_Z(z)} \tag{1.25}$$

allowing us to rewrite the marginals as

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

$$p_Z(z) = \int_{\mathbb{R}} p_{Z|Y}(z \mid y) p_Y(y)\, \mathrm{d}y = \mathbb{E}_Y\left[p_{Z|Y}(z \mid y)\right] \tag{1.27}$$

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, and for random variables admitting a density in Definition 1.2.11.

Theorem 1.2.12 – 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.

$$p_{Y|Z}(y \mid z) = \frac{p_{Z|Y}(z \mid y)p_Y(y)}{p_Z(z)} = \frac{p_{Z|Y}(z \mid y)p_Y(y)}{\int p_{Z,Y}(z, y)\,\mathrm{d}y} \propto p_{Z|Y}(z \mid y)p_Y(y) \tag{1.28}$$

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.

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, the observed one carries no information on the other. This notion of dependence (and independence) is first defined on events and extended to random variables

Definition 1.2.13 – Independence: Let $A, B \in \mathcal{F}$. Those two events are 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, $\mathbb{C}\mathrm{ov}[Y, Z] = 0$. The converse if false in general.

We discussed so far the different quantities that characterize random variables. Let us consider now two random variables which share the same sample space: $X, X' : \Omega \to S$. There exists various way to compare those two random variables, usually by quantifying some measure of distance between their pdf when they exist. One of the most used comparison tool for random variables is the Kullback-Leibler divergence.

Definition 1.2.14 – KL–divergence and entropy: The Kullback-Leibler divergence, introduced in [KL51] is a measure of dissimilarity between two distributions, based on information-theoretic considerations. Let X, X' be r.v. with the same sample space S, and p_X and $p_{X'}$ their densities, such that $\forall A$, $\int_A p_X(x)\,\mathrm{d}x = 0 \implies$

$\int_A p_{X'}(x)\,\mathrm{d}x = 0$. The KL-divergence is defined as

$$D_{\mathrm{KL}}\left(p_X \| p_{X'}\right) = \int_S p_X(x) \log \frac{p_X(x)}{p_{X'}(x)}\,\mathrm{d}x \tag{1.29}$$
$$= \mathbb{E}_X\left[-\log p_{X'}(X)\right] - \mathbb{E}_X\left[-\log p_X(X)\right] \tag{1.30}$$
$$= H\left[X', X\right] - H\left[X\right] \tag{1.31}$$

$H[X]$ is called the (differential entropy) of the random variable X, and $H[X', X]$ the cross-entropy of X' and X. Using Jensen's inequality, one can show that for all X and X' such that the KL-divergence exists, $D_{\mathrm{KL}}\left(p_X \| p_{X'}\right) \geq 0$ with equality iff they have same distribution, a desirable property when measuring dissimilarity. However, the KL–divergence is not a distance function, as it is not symmetric in general, and it does not verify the triangle inequality.

1.2.4.e Important examples of real random variables

One of the most well known distribution is the normal distribution, also called Gaussian distribution, that appears in various situations, but most notably in the central limit theorem.

Example 1.2.15 – The Normal Distribution: Let X be a r.v. from Ω to $\mathbb{R}$. If X follow the normal distribution of mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, we write $X \sim \mathcal{N}(\mu, \sigma^2)$, and its pdf is

$$p_X(x) = \phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \tag{1.32}$$

For the multidimensional case, let X be a r.v. from Ω to $\mathbb{R}^p$, that 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

$$p_X(x) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-1} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \tag{1.33}$$

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 distributions are displayed Fig. 1.3.

When adding independent squared samples of the normal distribution, the resulting random variable follows a χ^2 distribution.

Example 1.2.16 – The χ^2 distribution: Let $X_1, X_2, \ldots, X_\nu$ be ν independent random variables, such that for $1 \leq i \leq \nu$, $X_i \sim \mathcal{N}(0, 1)$ We define the random

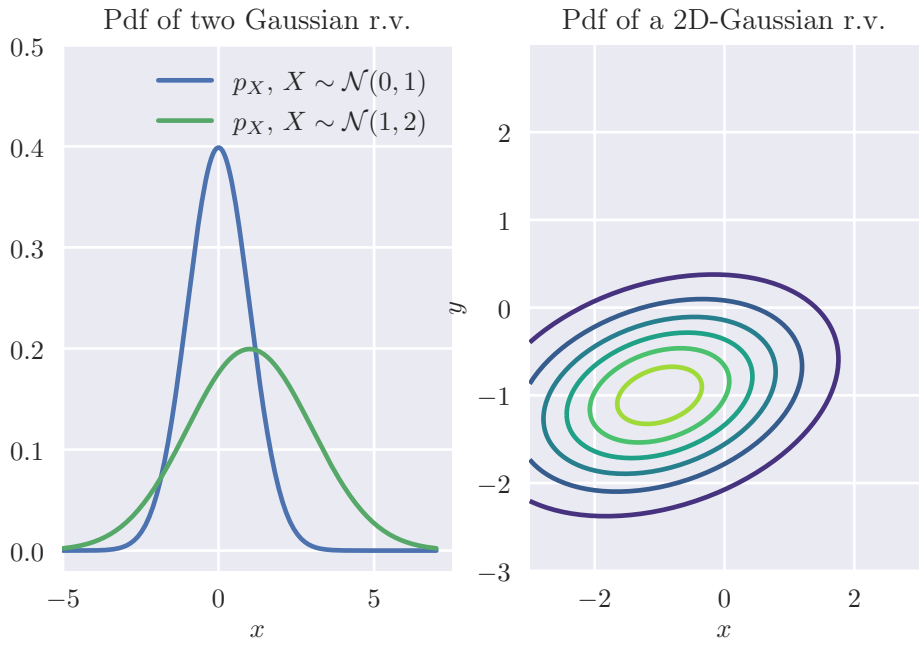

Figure 1.3: Probability Density functions of 1D Gaussian distributed r.v. (left), and density of a 2D Gaussian r.v. (right)

variable X as

$$X = \sum_{i=1}^{\nu} X_i^2 \tag{1.34}$$

By definition, the random variable X follows a χ^2 distribution with ν degrees of freedom: $X \sim \chi^2_\nu$. The quantile of order β is written $\chi^2_\nu(\beta)$ and verifies

$$\mathbb{P}[X \leq \chi^2_\nu(\beta)] = \beta \tag{1.35}$$

The pdf of such a r.v. are displayed Fig. 1.4, for different degrees of freedom.

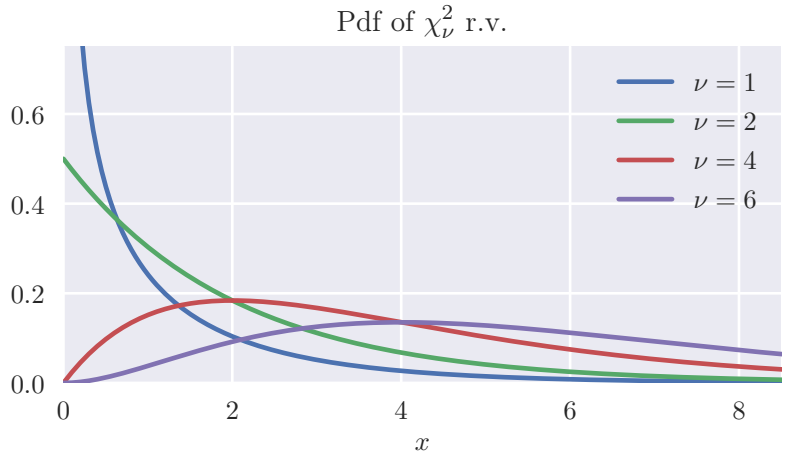

Figure 1.4: Probability density functions of χ^2_ν random variables, for different degrees of freedom

1.3 Parameter inference

1.3.1 From the physical experiment to the model

We can represent both the reality and the computer simulation as models. The physical system (the reality) that is observed can be represented by a model $\mathfrak{M} = (\mathscr{M}, \Theta_{\text{real}})$, so by a forward operator $\mathscr{M}$, and a parameter space Θ_{real}. Observing the physical system means to get access to $y \in \mathbb{Y}$ that is the image of an *unknown* parameter value $\vartheta \in \Theta_{\text{real}}$ through the forward operator, so $y = \mathscr{M}(\vartheta) \in \mathbb{Y} \subseteq \mathbb{R}^p$.

On the other hand, let us assume that a numerical model of the reality has been constructed, by successive various assumptions, discretizations and simplifications giving $(\mathcal{M}, \Theta)$. The main objective of calibration is to find $\hat{\theta}$ such that the forward operator applied to $\hat{\theta}$: $\mathcal{M}(\hat{\theta})$ represents as accurately as possible the physical system, and thus matches as closely the data $\mathscr{M}(\vartheta) = y$. This is illustrated Fig. 1.5.

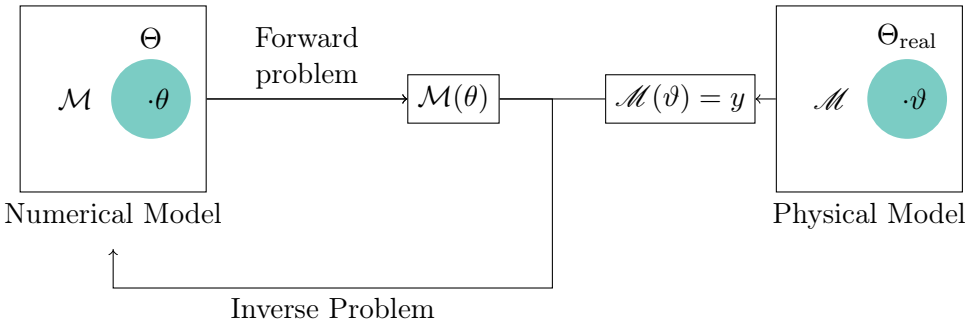

Figure 1.5: Forward and inverse problem using models as defined Definition 1.2.1

First, let us assume that $\vartheta \in \Theta \subseteq \Theta_{\text{real}}$, so we can rewrite the link between the reality and the model at this value as ([KO01, HKC+04])

$$\mathscr{M}(\vartheta) = \mathcal{M}(\vartheta) + \epsilon(\vartheta) \in \mathbb{Y} \subseteq \mathbb{R}^p \tag{1.36}$$

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 obtained data as a random sample as well.

From the diverse assumptions we can make upon this sampled random variable, we can then treat the calibration procedure as a parameter estimation problem of a random variable. The estimated parameter will be written $\hat{\theta}$, and the subscript will denote additional information on the estimator. In this thesis, we focus on extremum estimators. Those estimators are defined as the optimizer of a given objective function J, $\hat{\theta} = \arg\min J$. In the next sections, we will see the probabilistic origins of a few classical objective function.

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 (uppercase to highlight its random nature), and assume that this r.v. belong to a family of parametric distribution, whose densities are

$$\{y \mapsto p_Y(y;\theta); \theta \in \Theta\} \tag{1.37}$$

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

$$Y \sim \mathcal{N}(\mathcal{M}(\theta), \Sigma) \tag{1.38}$$

whose one sample is $y = \mathscr{M}(\vartheta)$.

Now, instead of looking at the densities of Eq. (1.37) 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 $\mathcal{L}$:

$$\mathcal{L}(\cdot; y) : \theta \mapsto p_Y(y;\theta) = \mathcal{L}(\theta; y) \tag{1.39}$$

$$= (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\mathcal{M}(\theta) - y)^T \Sigma^{-1} (\mathcal{M}(\theta) - y)\right) \tag{1.40}$$

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

$$\mathcal{L}(\theta; y) = \left(\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sigma_i}\right) \exp\left(\sum_{i=1}^{p} -\frac{(\mathcal{M}(\theta)_i - y_i)^2}{2\sigma_i^2}\right) \tag{1.41}$$

$$= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(\mathcal{M}(\theta)_i - y_i)^2}{2\sigma_i^2}\right) \tag{1.42}$$

with $y = [y_1, \dots, y_p]$ and $\mathcal{M}(\theta) = [\mathcal{M}(\theta)_1, \dots \mathcal{M}(\theta)_p]$. Based on the likelihood function, we can define the *Maximum Likelihood Estimator*, or *MLE*, that maximizes the likelihood defined above:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta; y) \tag{1.43}$$

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

$$\hat{\theta}_{\mathrm{MLE}} = \arg\min_{\theta \in \Theta} -\log \mathcal{L}(\theta; y) = \arg\min_{\theta \in \Theta} -\sum_{i=1}^{p} \log p_{Y_i|\theta}(y_i \mid \theta) \tag{1.44}$$

where

$$-\log \mathcal{L}(\theta; y) = \frac{1}{2}(\mathcal{M}(\theta) - y)^T \Sigma^{-1} (\mathcal{M}(\theta) - y) + \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| \tag{1.45}$$

As the optimization is performed on θ, we can remove the constant terms of the objective, and rewrite the cost function as a L^2 norm in Eq. (1.46).

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \arg\min_{\theta\in\Theta} \frac{1}{2}(\mathcal{M}(\theta) - y)^T \Sigma^{-1} (\mathcal{M}(\theta) - y) \\ &= \arg\min_{\theta\in\Theta} \frac{1}{2}\|\mathcal{M}(\theta) - y\|^2_{\Sigma^{-1}}\end{aligned} \tag{1.46}$$

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. (1.38). Other choices of the sampling distribution Eq. (1.38) 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 uncorrelated, thus indepedent due to their Gaussian nature as defined in Eq. (1.38). The likelihood can be rewritten as the product of densities evaluated at the different samples y_i, obtained from their true distribution Y. A direct link can be written between the KL–divergence and the MLE: The KL–divergence between the true density p_Y and the parametric sampling distribution $p_Y(\cdot;\theta)$ is

$$D_{\text{KL}}\left(p_Y \| p_Y(\cdot;\theta)\right) = \mathbb{E}_Y\left[\log p_Y(Y)\right] - \mathbb{E}_Y\left[\log p_Y(Y;\theta)\right] \tag{1.47}$$

As the first term does not depend on θ, minimizing this expression is equivalent to minimizing the second part of the equation, so

$$\arg\min_{\theta\in\Theta} D_{\text{KL}}\left(p_Y \| p_Y\left(\cdot;\theta\right)\right) = \arg\min_{\theta\in\Theta} -\mathbb{E}_Y\left[\log p_Y(Y;\theta)\right] \tag{1.48}$$

The true distribution of the observation is unknown, but samples y_i are available. Using the empirical KL-divergence denoted $D_{\text{KL}}^{\text{empirical}}$, and replacing the theoretical expectation with the empirical one, the equation above becomes:

$$\arg\min_{\theta\in\Theta} D_{\text{KL}}^{\text{empirical}}(p_Y \| p_Y\left(\cdot;\theta\right)) = \arg\min_{\theta\in\Theta} \frac{1}{p}\sum_{i=1}^{p} -\log p_Y(y_i;\theta) = \hat{\theta}_{\text{MLE}} \tag{1.49}$$

Thus, 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).

The MLE possesses desirable asymptotic properties, such as asymptotic normality when the number of observations grows large [**?**]. Those properties permit the construction of asymptotic confidence interval, and to perform hypothesis testing, especially for model selection. This aspect will be further developed in Section 1.5.

So far, the only information assumed on θ is its parameter space Θ. In the case where some belief on θ is present before the calibration, we can incorporate this information through the Bayesian framework.

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. (1.38) for the frequentist approach can be almost be rewritten as is in the Bayesian setting, just by conditioning Y with θ. Eq. (1.38) becomes

$$Y \mid \theta \sim \mathcal{N}(\mathcal{M}(\theta), \Sigma) \tag{1.50}$$

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

$$p_{\theta|Y}(\theta \mid y) = \frac{p_{Y|\theta}(y \mid \theta) p_\theta(\theta)}{p_Y(y)} = \frac{\mathcal{L}(\theta; y) p_\theta(\theta)}{p_Y(y)} \tag{1.51}$$

The denominator can be seen as normalizing constant, ensuring that $\int_\Theta p_{\theta|Y} = 1$. But it can also be seen as a measure of how well does the model explain the data obtained. This interpretation will be extended in Section 1.5

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.

$$p_Y(y) = \int_\Theta p_{Y,\theta}(y, \theta) \, \mathrm{d}\theta = \int_\Theta p_{Y|\theta}(y \mid \theta) p_\theta(\theta) \, \mathrm{d}\theta \tag{1.52}$$

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].

When the model $(\mathcal{M}, \Theta)$ and the data y is fixed, the model evidence is constant with respect to the calibration parameter θ. The posterior distribution is thus often written and evaluated up to a multiplicative constant.

$$p_{\theta|Y}(\theta \mid y) \propto \mathcal{L}(\theta; y) p_\theta(\theta) \tag{1.53}$$

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. (1.51),

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. Classical MCMC algorithms such as Metropolis-Hastings requires the use of a proposal density, and then to accept or reject the proposal based on the posterior distribution evaluated at the point.

A lot of refinement of these 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 important adaptation 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.

For time-dependent systems, Bayesian framework is particularly well-suited to treat observations sequentially, especially because Bayesian updating is done via multiplication. Bayes' theorem is the basis of many data assimilation methods, such as Kalman filter or various particle filters, that are often used for state estimation.

1.3.3.b Bayesian Point estimates

The whole posterior distribution aggregates a lot of information on the problem. However, as mentioned above, a certain work has to be done in order to get independent samples. Instead, one can try to find a point $\theta \in \Theta$ that summarizes as best this distribution. Consequently, the chosen estimate is often an indicator of the central tendency. In that sense, we wish to get a value that is quite close to all sampled values from the posterior [LC06].

Let us define a function L that measures a distance in the parameter space: $L : \Theta \times \Theta$. For a candidate θ', the measured risk with respect to a sample from the posterior $\theta_{\text{sample}} \sim \theta \mid Y$ is $L(\theta', \theta_{\text{sample}})$. The *Bayesian risk* for θ' is then the expectation of this Bayesian loss functions L under the posterior distribution: $\mathbb{E}_{\theta|Y}\left[L(\theta', \theta) \mid y\right]$. A Bayesian point estimate is defined as a minimizer of the Bayesian risk:

$$\hat{\theta}_L = \underset{\theta' \in \Theta}{\arg\min}\, \mathbb{E}_{\theta|Y}\left[L(\theta', \theta) \mid y\right] \tag{1.54}$$

Obviously, different loss functions will lead to different Bayesian point estimates, and we are going evoke two of them.

Posterior mean

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

$$\hat{\theta}_{\text{MMSE}} = \underset{\theta' \in \Theta}{\arg\min}\, \mathbb{E}_{\theta|Y}\left[(\theta' - \theta)^2 \mid y\right] \tag{1.55}$$

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

$$\hat{\theta}_{\text{MMSE}} = \mathbb{E}_{\theta|Y}[\theta \mid y] = \int_{\Theta} \theta \cdot p_{\theta|Y}(\theta \mid y)\, \mathrm{d}\theta \tag{1.56}$$

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. (1.55).

Posterior Mode: the MAP

Taking $L(\theta', \theta) = -\delta_\theta(\theta')$, the dirac delta function defined in Eq. (1.14), one can show that the minimizer of $\mathbb{E}_{\theta|Y}\left[L(\theta', \theta) \mid y\right]$ is the mode of the posterior distribution, and is called the *Maximum A Posteriori* (MAP):

$$\begin{aligned} \hat{\theta}_{\text{MAP}} &= \arg\min_{\theta' \in \Theta} \mathbb{E}_{\theta|Y}\left[\delta_\theta(\theta') \mid y\right] = \arg\min_{\theta' \in \Theta} -p_{\theta|Y}(\theta' \mid y) \\ &= \arg\max_{\theta' \in \Theta} p_{\theta|Y}(\theta' \mid y) = \arg\max_{\theta' \in \Theta} \mathcal{L}(\theta'; y) p_\theta(\theta') \end{aligned} \tag{1.57}$$

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. (1.56). We can resort to classical optimization techniques for this evaluation. Similarly to the likelihood, taking the negative logarithm leads to the following minimization problem.

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta' \in \Theta} -\log \mathcal{L}(\theta'; y) - \log p_\theta(\theta') \tag{1.58}$$

1.3.3.c Choice of a prior distribution

As seen in the application of Bayes' theorem in Eq. (1.51), 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 θ: $\theta \sim \mathcal{N}(\theta_b, B)$ where B is called the background covariance error matrix and θ_b is called the *background value* that acts as a plausible reference value. Assuming a Gaussian form for the errors as well with covariance matrix Σ, the MAP can be written as

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta \in \Theta} \frac{1}{2}\|\mathcal{M}(\theta) - y\|^2_{\Sigma^{-1}} + \frac{1}{2}\|\theta - \theta_b\|^2_{B^{-1}} \tag{1.59}$$

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 lead 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 do not necessarily lead to improper posterior, allowing for the usual Bayesian analysis of the quantity. For instance, if $\Theta = \mathbb{R}^p$, the prior $p_\theta(\theta) \propto 1$ is improper, but the MAP estimation is equivalent to the MLE.

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. 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.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} J(\theta) \tag{1.60}$$

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: $\nabla_\theta J$. 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. (1.61).

$$\nabla_\theta J \approx \left[\frac{J(\theta + \epsilon e_1) - J(\theta)}{\epsilon}, \frac{J(\theta + \epsilon e_2) - J(\theta)}{\epsilon}, \ldots, \frac{J(\theta + \epsilon e_n) - J(\theta)}{\epsilon} \right] \quad \text{for } \epsilon \ll 1 \tag{1.61}$$

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. Adjoint methods are thus very popular in large-scale optimization of Computational Fluid Dynamics codes, as the additional cost of implementation is often worth the gain in the short term. This situation is common in data assimilation, as shown in [DL91, DL92, HMR+10, CMMV13], or in shape optimization of airfoils in [HB01].

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:

$$\begin{aligned} &\min_{\theta \in \Theta} J(\theta) = J(y, \theta) \\ &\text{such that } \mathcal{F}(y, \theta) = 0 \end{aligned} \tag{1.62}$$

where the constraint on $\mathcal{F}$ signifies that the model is admissible, i.e. that $y = \mathcal{M}(\theta) \in \mathbb{Y}$.

Differentiating the Eq. (1.62) with respect to θ using the chain rule gives

$$\begin{aligned} \nabla_\theta J &= \frac{\partial J}{\partial y}\frac{\partial y}{\partial \theta} + \frac{\partial J}{\partial \theta} \\ \nabla_\theta \mathcal{F} &= \frac{\partial \mathcal{F}}{\partial y}\frac{\partial y}{\partial \theta} + \frac{\partial \mathcal{F}}{\partial \theta} \end{aligned} \tag{1.63}$$

In those equations, the partial derivatives with respect to θ are quite easily obtainable, while the real challenge is to obtain the derivative with respect to the state variable: $\frac{\partial}{\partial y}$.

To treat the constrained optimization in Eq. (1.62), let us introduce the Lagrange multiplier $\lambda \in \mathbb{Y}$, so that we can write the Lagrangian $\mathscr{L}$

$$\mathscr{L}(\theta, y, \lambda) = J(y, \theta) - \lambda^T \mathcal{F}(y, \theta) \tag{1.64}$$

is then

$$\min_{\theta, y, \lambda} \mathscr{L}(\theta, y, \lambda) \tag{1.65}$$

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: When differentiating with respect to the adjoint variable, we retrieve the state equation:

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

When differentiating with respect to the state variable, the equation that verifies the adjoint variable is called the adjoint equation

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

Finally, when λ verifies the adjoint equation: $\left(\frac{\partial \mathcal{F}}{\partial y}\right)^T \lambda = \left(\frac{\partial J}{\partial y}\right)^T$, the gradient of the cost function can be expressed using the partial derivative *with respect to* θ of the cost function and of the forward model, and the adjoint variable:

$$\frac{\partial \mathscr{L}}{\partial \theta} = \nabla_\theta J = \frac{\partial J}{\partial \theta} - \lambda^T \frac{\partial \mathcal{F}}{\partial \theta} = 0 \qquad \text{(Optimality condition)}$$

So, to get $\nabla_\theta J$, as the partial derivatives with respect to the control variable are relatively easy to obtain, the challenge lies in solving the adjoint equation. Albeit tedious, one can derive those equations by writing the tangent linear model of the original model, and implement a dedicated solver for the adjoint variables. A more common and way simpler approach is to derive the adjoint equations directly from the computer code implemented to solve the model, by using Automatic differentiation tools, such as TAPENADE [HP13]. Those programs directly translate the source code into a programs that solves the original model equations and the adjoint equations, and outputs the gradient along with the cost function.

1.5 Model selection

So far, we have discussed the calibration of a specific model $(\mathcal{M}, \Theta)$ given some observations, thus solving an inverse problem and finding $\hat{\theta}$ as an extremum of an objective function. But different models may be considered to explain the data. Those models may differ by their forward operator, from their parameter space, or both at the same time.

But changing models also means changing the potential "best" fit attainable. More complex models usually provide a better fit of the model but to the cost of a higher dimension in the parameter space. It is then natural to see if we can reduce the complexity of the model, without decreasing significantly its performance.

We are first going to consider the case of nested models: models that share the same forward models, but whose parameter spaces are nested. Model selection in this case is a way to reduce the dimension of the model, by reducing the parameter space.

Finally, tools introduced in this section bring up model comparison. A calibrated model $\{\mathcal{M}, \{\hat{\theta}\}\}$ is optimal given an objective function, but we can show that values close to the calibrated value $\hat{\theta}$ may also be in interest, as the decrease of performance may not be statistically detectable, giving a model $\{\mathcal{M}, \{\hat{\theta} + \varepsilon\}\}$ acceptable to describe the data.

1.5.1 Likelihood ratio test and relative likelihood

Generally speaking, more complex models have a better ability to represent the data, but all the parameters included in the model may not be relevant for the modelling. It can be interesting to test if a "simpler" model would give similar performances, or at least show a decrease in performances that is not statistically significant. One of the most 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 $r > d$. As $\mathfrak{M}_1$ is more general, one can expect better performances.

Under the null hypothesis, the two models are equivalent, that is that the "smaller" parameter space is enough to represent the inverse problem: $\theta \in \Theta_2$. The alternative 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 $\hat{\theta}_2$.

$$\Lambda(y) = \frac{\sup_{\theta\in\Theta_2} \mathcal{L}(\theta; y)}{\sup_{\theta\in\Theta_1} \mathcal{L}(\theta; y)} = \frac{\mathcal{L}(\hat{\theta}_2; y)}{\mathcal{L}(\hat{\theta}_1; y)} \leq 1 \tag{1.66}$$

Based on this quantity, we can test whether the smaller model is sufficient to explain the data as good as the larger model. The two hypothesis for this test are

- $\mathcal{H}_0$: The two models are statistically equivalent: the difference between the maximal values of the likelihood is not statistically significant. This corresponds to Λ close to 1
- $\mathcal{H}_1$: the two models are statistically different: the larger model performs better than the reduced one. This corresponds to Λ significantly smaller than 1.

Under the null hypothesis, $-2\log\Lambda$, (sometimes called the deviance) follows asymptotically (as the number of observations becomes large) a χ^2 distribution defined in Example 1.2.16, whose degrees of freedom is given by the difference of dimensionality between the two models:

$$-2\log\Lambda(y) \xrightarrow{\mathrm{d}} \chi^2_{r-d} \tag{1.67}$$

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

$$\mathrm{RejReg}_\alpha = \left\{y \mid -2\log\Lambda(y) > \chi^2_{r-d}(1-\alpha)\right\} \tag{1.68}$$

Or by reformulating using the log-likelihoods and objective functions $l(\theta; y) = \log\mathcal{L}(\theta; y) = -J(\theta)$

$$\mathrm{RejReg}_\alpha = \left\{y \mid (\sup_{\theta\in\Theta_1} l(\theta; y) - \sup_{\theta\in\Theta_2} l(\theta; y)) > \frac{1}{2}\chi^2_{r-d}(1-\alpha)\right\} \tag{1.69}$$

$$= \left\{y \mid J(\hat{\theta}_2) - J(\hat{\theta}_1) > \frac{1}{2}\chi^2_{r-d}(1-\alpha)\right\} \tag{1.70}$$

As a basis for comparison, when $\Theta \subset \mathbb{R}$, $r-d=1$ and $\chi^2_1(1-0.05) = 3.84$. When the data falls into the rejection region ($J(\hat{\theta}_2)$ significantly larger than $J(\hat{\theta}_1)$, the null hypothesis is rejected, and the model are significantly different. The rejection region defined above allows also to define some confidence interval for the parameter. By introducing the *Relative Likelihood*, as defined in [Kal85], is the ratio of the likelihood evaluated at a point θ to the maximal value of the likelihood:

$$R(\theta) = \frac{\mathcal{L}(\theta; y)}{\mathcal{L}(\hat{\theta}_{\mathrm{MLE}}; y)} = \frac{\mathcal{L}(\theta; y)}{\sup_{\theta'\in\Theta} \mathcal{L}(\theta'; y)} \tag{1.71}$$

This function allows for comparing the plausibility of the value θ, compared to the MLE. The likelihood interval of level $p \in]0, 1]$ is defined as

$$\mathcal{I}_{\text{Lik}}(p) = \left\{ \theta \mid R(\theta) = \frac{\mathcal{L}(\theta; y)}{\mathcal{L}(\hat{\theta}_{\text{MLE}}; y)} \geq p \right\} \tag{1.72}$$

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. (1.68) written as a likelihood interval becomes

$$\mathcal{I}_{\text{Lik}}\left(\exp\left(-\frac{1}{2}\chi^2_{\dim(\Theta)}(1-\alpha)\right)\right) = \left\{\theta \mid R(\theta) \geq \exp\left(-\frac{1}{2}\chi^2_{\dim(\Theta)}(1-\alpha)\right)\right\} \tag{1.73}$$

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, the threshold of Eq. (1.73) is $\exp\left(-\frac{1}{2}\chi^2_{\dim(\Theta)}(1-\alpha)\right) = \exp\left(-\frac{1}{2}\chi^2_1(.95)\right) \approx 0.15$, and at a level .10, $\exp\left(-\frac{1}{2}\chi^2_1(.90)\right) \approx 0.26$.

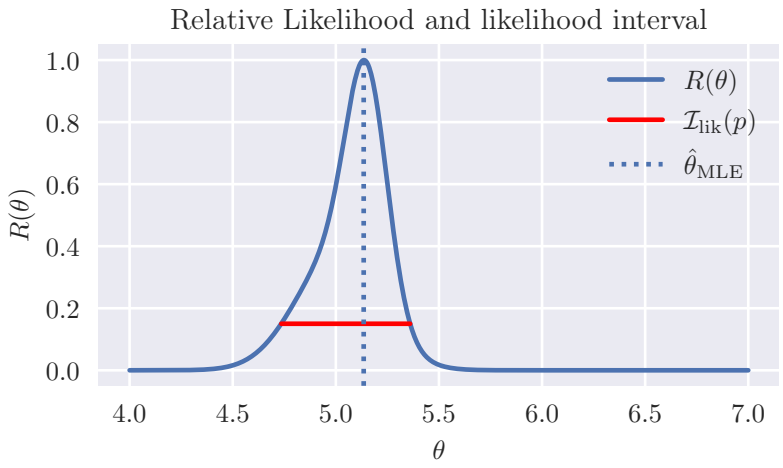

Figure 1.6: Example of relative likelihood, and associated likelihood interval

Due to the likelihood ratio test and the relative likelihood, we can see that even though $\hat{\theta}_{\text{MLE}}$ is the optimizer of the likelihood function, other values close to it may not be discarded, provided that the log-likelihood does not drop off too much.

1.5.2 Criteria for model comparison

The likelihood ratio test presented above is suited for nested models. We can also associate each model with a single numerical value, usually in the form of

$$\text{Crit}(\mathfrak{M}) = -2\log\mathcal{L}(\hat{\theta}_{\text{MLE}}) + \text{Complexity penalization} \tag{1.74}$$

where $\mathcal{L}$ is the likelihood function for the model $\mathfrak{M} = (\mathcal{M}, \Theta)$, and $\mathcal{L}(\hat{\theta}_{\text{MLE}}) = \max \mathcal{L}$. The role of the complexity penalization is to avoid overfitting, and is often directly linked to the dimension of the parameter space Θ. Two quite popular examples of criteria are the AIC (Akaike Information Criterion) and the BIC (Bayesian Information Criterion).

To compare two models $\mathfrak{M}_1$ and $\mathfrak{M}_2$, the magnitude of the difference $\text{Crit}(\mathfrak{M}_1) - \text{Crit}(\mathfrak{M}_2)$ is compared to some thresholds as shown in [BA04]. The difference shows whether a model should be preferred, or if no substantial evidence exists for either model. These criteria, as well as the likelihood ratio test, are based on the evaluation of the likelihood at its maximal value. Another approach is to marginalize the likelihood with respect to the prior distribution of the calibration parameter, giving Bayesian model comparison.

1.5.3 Bayesian model comparison

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

For two models $\mathfrak{M}_1$ and $\mathfrak{M}_2$ the Bayes' factor is defined as the ratio of the evidence of the two models:

$$\text{BF}(\mathfrak{M}_1, \mathfrak{M}_2) = \frac{p_{Y|\mathfrak{M}_1}(y \mid \mathfrak{M}_1)}{p_{Y|\mathfrak{M}_1}(y \mid \mathfrak{M}_2)} \tag{1.75}$$

where

$$p_{Y|\mathfrak{M}_i}(y \mid \mathfrak{M}_i) = \int_{\Theta_i} p_{Y|\theta,\mathfrak{M}_i}(y \mid \theta, \mathfrak{M}_i) p_{\theta|\mathfrak{M}_i}(\theta \mid \mathfrak{M}_i) \, \mathrm{d}\theta = \int_{\Theta_i} p_{Y,\theta|\mathfrak{M}_i}(y, \theta \mid \mathfrak{M}_i) \, \mathrm{d}\theta \tag{1.76}$$

Quite similarly as the criteria introduced before, the logarithm if the Bayes' factor is usually compared to specific values, allowing us to conclude roughly on how strong does the data favors $\mathfrak{M}_1$. Again, the logarithm of Bayes' factor is compared with specified threshold ([KR95, BA04]). Given a model, a criterion is derived either by maximization (for the likelihood ratio test, and criteria of Section 1.5.2) and marginalization (for Bayes' factors). In both cases, the preference toward one or another model is directly linked to the difference on the logarithms of the densities.

1.6 Parametric model misspecification

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 $(\mathscr{M}, \Theta_{\text{real}} = \Theta)$, so the parameter spaces are the same. In practice, the parameter space Θ does not contain necessarily 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 external forcings.

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 are assumed to have an inherent variability. In that sense, it may not be worth spending time and resources to infer these parameter values, as they are bound to change. Moreover, we can only get information on the environmental conditions used to generate the observations.

In terms of models, each choice of $u \in \mathbb{U}$ gives a different model $\mathfrak{M}(u) = \{\mathcal{M}(\cdot, u), \Theta\}$. Let us assume that we can model the uncertain parameters as a random variable U. Let us consider that we chose a specific $u_0 \in \mathbb{U}$, and that we are given some observation $y = \mathscr{M}(\vartheta)$. We can formulate an inverse problem, and an objective function $J : \theta \mapsto J(\theta, u_0)$, that we wish to minimize with respect to θ. Some estimators still carry nice properties. The MLE for instance, defined Section 1.3.2 can still be written as the minimizer of the empirical KL–divergence. We assume that we can write the sampling distribution as $p_{Y|\theta,U}$, and

$$\hat{\theta}_{\mathrm{MLE}}(u_0) = \arg\min_{\theta\in\Theta} D_{\mathrm{KL}}^{\mathrm{empirical}}\left(p_Y \| p_{Y|\theta,U}(\cdot \mid \theta, U = u_0)\right) \tag{1.77}$$

and can be seen as the "best" value given $U = u_0$. However, the asymptotic properties of the MLE are slightly different as described in [Whi82]. So when the model is misspecified, minimizing the same cost function still makes sense.

However, the calibration will depend on the chosen u_0: $\hat{\theta}(u_0) = \arg\min_{\theta\in\Theta} J(\theta, u_0)$, and there is no guarantee that $\hat{\theta}(u_0)$ will minimize $J(\cdot, u_1)$ for $u_0 \neq u_1$, as illustrated Fig. 1.7.

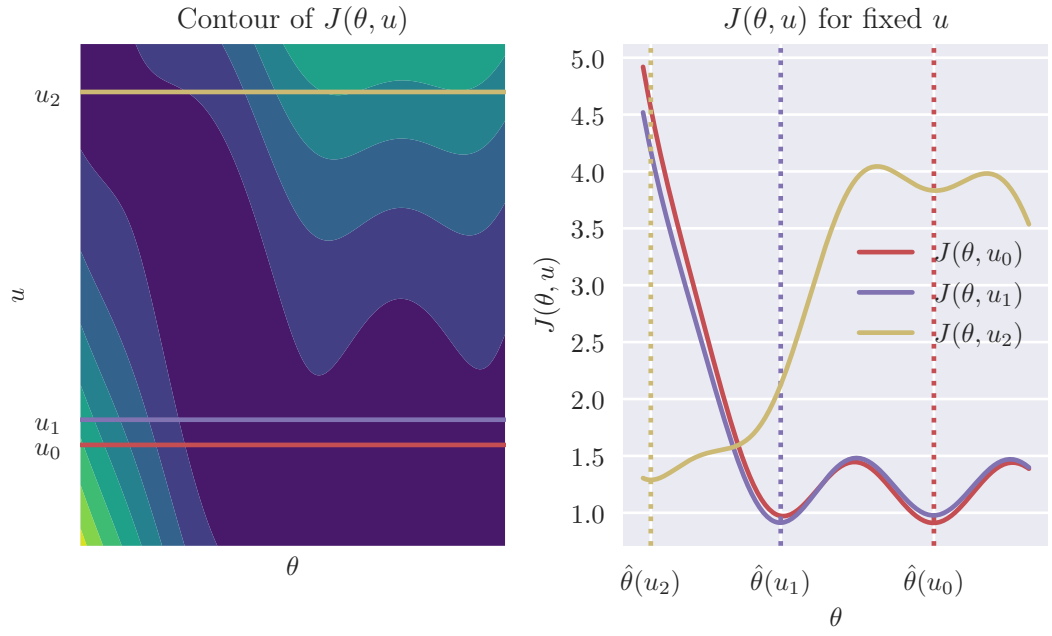

Figure 1.7: Effect of the misspecification on the minimizer.

u_0 and u_1 are close to each other, but $\hat{\theta}(u_0)$ and $\hat{\theta}(u_1)$ are not. However, as the cost function shows similar values at those points, choosing either one would lead to a satisfactory cost given $U = u_0$ or $= u_1$. A preference appears for $\hat{\theta}(u_1)$ if $U = u_2$ is considered as well.

In terms of model selection, the asymptotic distribution of the likelihood ratio statistic defined Section 1.5.1 is also slightly different. Instead of following a χ^2_r distribution, where r is the number of dimensions for the test, $-2\log\Gamma$ will, asymptotically, have the same distribution as a weighted sum of r χ^2_1 distribution, whose weights are the eigenvalues of a matrix involving the Jacobian and the Hessian of the log-likelihood [**?**].

This random misspecification leads to some issues in the calibration of the model, and it asks for a notion of robustness with respect to the environmental parameters.

1.7 Partial conclusion

In this chapter, starting from a probabilistic point of view, we established the usual tools encountered in model calibration: the misfit between the data and the numerical model is measured by a cost function J, that can be minimized using for instance gradient descent. From this optimization, we can define a "acceptable" region for the estimate. In other words, values in this set yield an misfit that is not different enough to be completely discarded.

Adding a environmental variable as a random nuisance parameter introduces a parametric misspecification of the model: each realization of this underlying random variable will yield a different estimation. In **??**, we will discuss the notion of robustness under this random misspecification, and introduce a family of robust estimators, inspired by model selection.

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.

[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.

[BMST20] Elie Bretin, Simon Masnou, Arnaud Sengers, and Garry Terii. Approximation of surface diffusion flow: A second order variational Cahn–Hilliard model with degenerate mobilities. *arXiv:2007.03793 [cs, math]*, July 2020.

[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.

[HB01] Luc Huyse and Dennis M. Bushnell. Free-form airfoil shape optimization under uncertainty using maximum expected value and second-order second-moment strategies. 2001.

[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.

[HP13] Laurent Hascoet and Valérie Pascual. The Tapenade automatic differentiation tool: Principles, model, and specification. *ACM Transactions on Mathematical Software*, 39(3):1–43, April 2013.

[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.

[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.