Inverse problems occurring in uncertainty analysis

Shuai Fu

N° d'ordre :

Université Paris-Sud

Faculté des Sciences
d'Orsay

THÈSE

présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS XI

Spécialité : Mathématiques

par

Shuai FU

Sujet :

Inversion probabiliste bayésienne en analyse d'incertitude

Soutenue le 14 décembre 2012 devant la Commission d'examen :

M.	Lucien BIRGÉ	(Président du jury)
M.	Nicolas BOUSQUET	(Encadrant industriel)
M.	Gilles CELEUX	(Directeur de thèse)
M.	Mathieu COUPLET	(Encadrant industriel)
M.	Jean-Michel MARIN	(Rapporteur)
M.	Bruno SUDRET	(Rapporteur)

Acknowledgements

First of all, I would like to thank my adviser Gilles Celeux. Over the past three years, he taught me how to conduct rigorous research with tireless patience and responsibility. Day by day, he provided me priceless technical support, guidance, and more important, the confidence. I sincerely thank you, Gilles. It was you who made me believe that I can have my own scientific ideas; it was you who encouraged me for my first speech, my first publication during this thesis and my first class as a teaching assistant; it was you who gave me so much freedom to manage my own time. I am so lucky to have you as my adviser.

I would also like to thank my industrial advisers Mathieu Couplet and Nicolas Bousquet. I cannot decide the sequencing because both of them are so important to me. This thesis would never have been possible without any of them. I sincerely thank you Mathieu, for your continued guidance, everyday kindness, daily good mood and your unbelievable original ideas ! I would like to thank you Nicolas, sincerely. Since my Master's internship, you are the first person I knew at EDF. You impressed me since the first day, not only by your natural kindness, but also by your enthusiasm in statistics. I obtained and learned so much from you.

I would like to acknowledge Jean-Michel Marin and Bruno Sudret who accepted to examine my Ph.D work. Thank you for your patient review and valuable comments, and thank you all for coming to Orsay, early in a Friday morning !

I feel so grateful for Lucien Birgé. I deeply thank you, Mr. Birgé. Without your confidence, there would not have been my thesis. I thank you also for your kind and friendly reception of my parents in Paris !

I would like to thank Pierre Del Moral who provided me a kind visit to his laboratory. Thank you Jon Wellner for your encouragement of my IMS meeting. Thank you Emmanuel Vazquez and Estelle Kuhn for the valuable technique exchanges. During these three years, I have moved between two institutions: the Paris-Sud 11 University (Orsay) and EDF. This organization permitted me to meet many friendly people and touch different attractive ambiances.

At EDF, I would like to thank all my colleagues and especially Alberto Pasanisi for our unforgettable first publication, thank you Laurent Billet, Philippe Klein and François Billy, thank you Kateryna for your patient help for the server Ivanoe, thank you Miguel, Paul, Emmanuel, Merlin, Aurélie, Fanny, Anne-Laure, Jean-Baptiste, Roman, Guillaume, Hélène, Jérome, Françoise Talbot, Françoise Massot, Antoine, Momo, Jun, Meryam, Tu-Duong, ...

At Orsay, I wish to thank numbers of my colleagues in the famous office 227. Thank you Vincent for you continued friendly help and good spirit everyday, thank you Rémy for the nice experience of our three months' collaboration, thank you Pierre for introducing me the big world of meta-modeling, thank you Patrick, Maud, Nicolas, Julie, Raphael, thank you all engineers Hector, JF, Eric, Elodie, Célia. Let me especially thank Yves Misiti for borrowing

me the super powerful computer. I would never have finished my numerical experiences without it ! I would like to thank all the professors Pascal Massart, Danielle, Yves, Marc, Elena, Christine, Patrick, Olivier. I will certainly not forget our efficient secretaries Valérie and Katia, thank you all !

I also would like to thank my Chinese friends Quan, Fang, Ping, Lei, Bing, Hui, Xia, Wenqing, Jianhan, Peng and Wenwen, not only for numbers of delicious meals shared together, but for the continued support during these years in France !

Deeply, I thank my family, especially my parents Lixin and Yanyun, who raised me up, who believe in me and encourage me, who sometimes criticize me but always tolerate me. I can never never make it through without their emotional support.

Last but not least, my "thank you" is reserved for you, Kaelig. Thank you for accompanying me such a long time in this adventure, with patience, with kindness, with love.

Shuai Fu, November 22, 2012

Contents

Notation

The definition of the following notation will be reminded for their first appearance in the thesis and will be reused without mandatory recall. In general, the random variables are noted in capital letters, their realizations in small letters.

Variables

X	unobserved variable
Y	observed variable
U	measurement error (white noise)

Probability distributions

$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and variance matrix Σ
$\mathcal{IW}(\Lambda, \nu)$	Inverse-Wishart distribution with degrees of freedom ν and inverse scale matrix Λ
$\mathcal{U}[a, b]$	Uniform distribution on $[a, b]$
$\pi(\theta)$	Prior distribution of the parameter θ which is assumed to be random
$\pi(\theta \mid \mathbf{y}, \mathbf{d})$	Posterior distribution of θ given observations $(\mathbf{y}, \mathbf{d})$
$\pi(X \mid \theta)$	Density of X knowing its parameter θ

Mathematical symbols

$\mathbf{I}_A$	Indicator function of the event A
$L(\theta; \mathbf{y}, \mathbf{d})$	Log-likelihood of θ given observations $(\mathbf{y}, \mathbf{d})$
$\mathcal{L}(\theta; \mathbf{y}, \mathbf{d})$	Likelihood of θ given observations $(\mathbf{y}, \mathbf{d})$
$\propto$	Be equal to up to a multiplicative constant
$\overset{\exp}{\propto}$	Be equal to up to an additive constant
$\overset{\mathcal{L}}{\rightsquigarrow}$	Convergence in law
$\xrightarrow{\mathbb{P}}$	Convergence in probability
$\perp\!\!\!\perp$	Be statistically independent

Parameters

θ	Characterizing parameters (mean and variance term) for the Gaussian distribution of the unobserved variable X
R	Variance matrix of the white noise U

Abbreviations

MSE	Mean squared error
DOE	Design of experiments
MCMC	Monte Carlo Markov chain
MLE	Maximum likelihood estimation
EM	Expectation Maximization algorithm
ECME	Expectation-Conditional Maximization either algorithm
SEM	Stochastic expectation maximization algorithm
MCEM	Monte Carlo expectation maximization algorithm
SAEM	Stochastic approximation expectation maximization algorithm
GP	Gaussian process
DAC	Criterion of consistency between the prior and the data as well as the choice of DOE
KL	Kullback-Leibler divergence
MMSE	Maximum Mean Square Error
W-IMSE	Weighted-Integrated Mean Square Error criterion
E-CD	Expected-Conditional Divergence criterion
pdf	Probability density function
cdf	Cumulative distribution function
i.e.	that is (id est)
e.g.	example given, for example
c.f.	bring together (confer)
i.i.d.	independent and identically distributed
a.s.	almost surely

General introduction

$\mathcal{A}$. Context and objective of thesis

Inverse problems, as an important topic widely treated in science and engineering field, are gaining fast development. Typically, an inverse problem is a problem of quantifying an influential variable that cannot be observed for technical or cost reasons, but for which there are indirect observations. The observations are assumed to be explained by a *black-box* type expensive-to-compute function. This variable is "influential" in the sense that it is considered as the major source of *uncertainty* affecting the behavior of the decision variable. This thesis focuses on providing a probabilistic solution to such inverse problems, by accounting for the variability of the model inputs. The inverse problem studied in this thesis is part of the methodology of uncertainty treatment which has been defined by a group involving different industrial and academic organizations (De Rocquigny et al., 2008, (23), pp. 233).

The following example is to consider an industrial plant Σ located near a river and submitted to flood risk. The water level at the site is denoted by Z. It can be simulated using a hydraulic code G, with the observed input D and the unobserved input X. Input D is often related to experimental conditions typically the flow of the river and X denotes the friction characterizing the riverbed. This missing data X composes a set of geomorphological and fluctuating time and space characters. which can explain the fluctuations of Z if D is known. Assuming D is known, the fluctuations of Z can be directed accounting for the "random" nature of X in a probability space Ω, with distribution denoted by $f(X)$.

In the context of structural reliability, it is important to verify if the height h_0 of the protecting dyke at Σ is sufficient such that the probability of flooding in the disadvantageous environmental conditions, i.e. D is fixed to be a "defavorable" d (e.g. a strong flow), remains below some threshold (e.g. 1%)[1]. This probability can be defined as:

$$\begin{aligned} P(Z \geq h_0) &= P(G(X,d) \geq h_0), \\ &= \int_{\Omega} \mathbb{1}_{\{G(x,d)\geq h_0\}} f(x)\, dx. \end{aligned} \tag{1}$$

The interest of simulating according to the density $f(\cdot)$ arises immediately: the estimation of

[1] The thesis work also responds to the following dual problem: determine the minimum height of the dyke to ensure the protection duty under the threshold. The quantity of interest is no longer the probability of exceeding but a quantile of Z.

$P(Z \geq h_0)$ can be produced by a Monte Carlo method based on M samples $(x_1, \ldots, x_M) \sim f(x)$:

$$\frac{1}{M} \sum_{i=}^{M} \mathbb{1}_{\{G(x_i,d) \geq h_0\}} \quad \xrightarrow[M \to \infty]{p.s.} \quad P(Z \geq h_0). \tag{2}$$

As previously presented, the code G is expensive to compute. Various methods have been proposed to produce estimators requiring a smaller number of calls to G (Rubino et Tuffin, 2009, (96)). However, they still need to be able to handle the distribution $f(x)$ and then simulate according to it. Estimating the distribution f of x would be the central point of this thesis.

Remark 1. *Apart from the hydraulic context, there exist other relevant contexts where the methodology developed in this thesis can be applied. For example, we can find similar inverse problems in the external acoustic propagation (Leroy, 2010, (58)), the mechanical vibration (De Rocquigny and Cambier, 2009, (22)), the structural mechanics (Perrin, 2008, (80)) and in thermal (De Crécy, 2001, (21)).*

In the framework of the industrial uncertainty analysis, as proposed by De Rocquigny et al. (2008, (23)), inverse problems belong to a range of problems which characterize the quantification of uncertainties. Our thesis work is motivated by the following methodology, described in Figure 1, which is widely applied in EDF. It summarizes the three main steps of treating uncertainties in industry.

- Step A: problem specification
- Step B: quantification of uncertainty sources
- Step B': quantification of sources (Inverse methods, calibration, assimilation)
- Step C: propagation of uncertainty sources
- Step C': sensitivity analysis, prioritization

and the iterative nature of the approach in various applications (see De Rocquigny et al., 2008, (23)). This thesis is located in the step B', requiring to quantify the uncertainty sources with the help of some inverse methods. In other words, we aim at analyzing its probability density f, which plays a critical role to link the modeling step A and the propagating step C.

$\mathcal{B}$. Mathematical treatment

b.1. Mathematical modeling and Bayesian framework

Statistical estimation of the density f which quantifies the main uncertainty source X is the inverse problem addressed in this thesis. It can be described as follows. Note that to better illustrate the problem, we give a sense to each variable as in the previous hydraulic example.

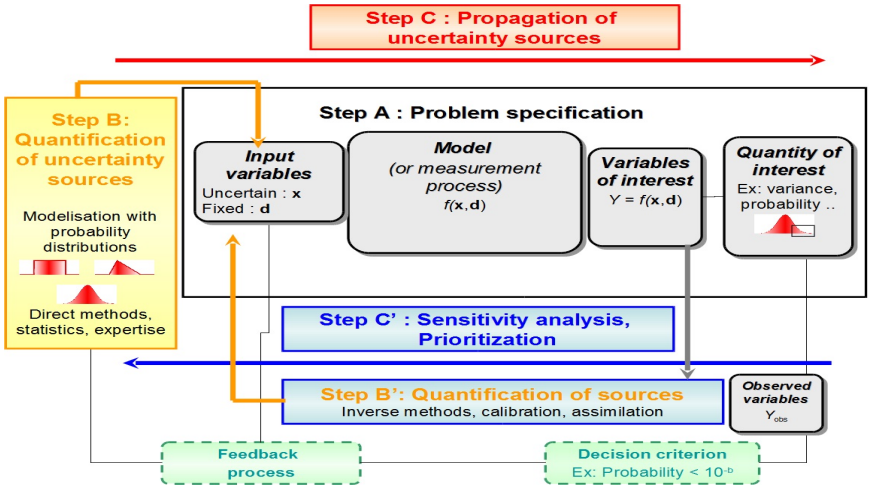


Figure 1: Methodologied framework for treating uncertainties in industry (23)

Given a set of observations of the water level $\mathbf{Y} = (Y_1, \ldots, Y_n)$ obtained in dedicated stations (not necessarily at the place of the industrial plant Σ), provided with experimental conditions $d_1, \ldots, d_n$ which are assumed to be known, and the unobserved variables $X_1, \ldots, X_n$ following the unknown distribution f, we can precise the following relationship:

$$Y_i \quad = \quad H(X_i, d_i) + U_i,$$

where U_i denotes the measurement error. Here the code H is not necessarily the same[2] as G, which has been introduced in Section $\mathcal{A}$. H can for example be the water level at the observation positions while G is related to other positions to forecast the water level. Consequently, Y_i may be different from Z. The estimation algorithms take advantage of the structure of the missing data problem by proposing iterative reconstructions of data $X_1, \ldots, X_n$ knowing $\mathbf{Y}$. Within a framework of parametric statistics, the observed data permits us to iteratively estimate the vector of parameters θ which defines the density $f(x) = f(x|\theta)$, that is assumed to be well chosen. With the estimated θ, the set of the possible values of X can be exhaustively described, that is why θ can be considered as a "hidden state of nature".

In frequentist approach of this inverse problem, several obstacles arise. The small size of the data sample $\mathbf{Y}$ and the low presence of extreme values among the $\mathbf{Y}$s can lead to poor maximum likelihood estimates of θ. Moreover, if there exists some available prior information, it would be profitable to take it into account. For example, the hydraulic literature may provide us the estimates of some characteristics of X such as the mean value in function of the nature of the ground.

For these reasons, Bayesian parametric framework has been chosen in this thesis, which considered θ as a random variable by providing a prior distribution $\pi(\theta)$ to integrate different sources of information. The posterior distribution $\pi(\theta|\mathbf{Y})$ must be estimated. As mentioned in Pasanisi et al. (2011, (77)), this posterior distribution provides a complete description of

[2]In our case study, it is worth noting that H equals G. The meta-modeling technique proposed in this thesis permits us to estimate the parameters of interest and the probability of exceeding (2) at a lower cost.

the remaining uncertainty affecting θ after the collection of all available information.

b.2. Principal tools: hybrid MCMC algorithm and meta-modeling technique

In this Bayesian framework, a multidimensional Metropolis-Hastings-within-Gibbs (hybrid MCMC) algorithm has been proposed to compute the posterior distribution of θ using a data augmentation scheme. In this algorithm, the simulator H is being called at each iteration.

But, this simulator H interpreting the physical input and output relationship is usually highly time-consuming. Meta-modeling techniques are thus necessary to approximate the original computer codes. A meta-model is an approximation of the original simulator built from its evaluations at a certain number of input values, the so-called design of experiments (DOE). In this thesis, we use a Gaussian Process (GP) meta-modeling (*kriging*) technique. The reasons for this choice are twofold: first, it is consistent with Bayesian inference as constructing a meta-model can be interpreted as providing some prior information to the original function (Rasmussen and Williams, 2006, (86)); second, the related uncertainty can be expressed and compared at every estimated point. The prediction accuracy depends on the position of the predicted point with respect to the spatial structure of the DOE.

b.3. Utilization of the thesis results in uncertainty treatment

In uncertainty treatment, the final goal is to propose an estimate $\hat{\rho}$ of the decision function $\rho(\theta)$,

$$\rho(\theta) \;=\; P_\theta(Z \geq h_0) \;=\; \int_\Omega \mathbb{1}_{\{G(x,d)\geq h_0\}} f(x|\theta)\; dx, \tag{3}$$

by minimizing a discrepancy or a cost $\mathcal{D}\{\rho(\theta), \hat{\rho}\}$. In the Bayesian framework, this cost is known through the posterior distribution $\pi(\theta|\mathbf{Y})$. For a given choice $\mathcal{D}$, the Bayes estimator ρ^* of $\rho(\theta)$ is thus:

$$\hat{\rho}^* \;=\; \arg\min_{\hat{\rho}} \int_\Theta \mathcal{D}\{\rho(\theta), \hat{\rho}\}\pi(\theta|\mathbf{Y})\; d\theta. \tag{4}$$

A practical choice for $\mathcal{D}$ is a quadratic cost. It is worth noting that for such a choice, the Bayes estimator $\hat{\rho}^*$ is the posterior mean of $\rho(\theta)$, which can be calculated as[3]:

$$\hat{\rho}^* \;=\; \int_\Theta \int_\Omega \mathbb{1}_{\{G(x,d)\geq h_0\}} f(x|\theta)\pi(\theta|\mathbf{Y})\; dx d\theta.$$

With $\pi(\theta|\mathbf{Y})$ simulated through the hybrid MCMC algorithm and ($f(x|\theta)$ assumed to be well chosen, an estimate of $\hat{\rho}^*$ can be produced by the following Monte Carlo algorithm:

1. `simulate` $(\theta_1, \ldots, \theta_M) \sim \pi(\theta|\mathbf{Y})$,

[3]It happens to be the predictive posterior probability of a flood, which means the posterior mean of the function $\mathbb{1}_{\{G(x,d)\geq h_0\}}$ (see Chapter II for more details).

2. simulate $(x_1, \ldots, x_M) \sim (f(x|\theta_1), \ldots, f(x|\theta_M))$,

3. estimate $\hat{\rho}^*$ through

$$\hat{\rho}^*_M = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}_{\{G(x_i,d) \geq h_0\}},$$

which must of course be adapted if G is itself time-consuming.

b.4. Main contributions

In Bayesian approach, the estimation problem related to the inverse problem involves many possible errors:

- *Estimation error:* Usually the sample size n is small with respect to the dimension of the problem and the variance of the estimates could be expected to be large;
- *Emulator error:* Since H is too complex, there is the need to replace it with an emulator $\widehat{H}$ and the discrepancy between H and $\widehat{H}$ could induce an important error;
- *Algorithmic error:* To proceed to statistical inference, there is the need to use complex stochastic algorithms. In the Bayesian setting, those algorithms are Monte Carlo Markov Chains (MCMC) algorithms which produce Markov chains converging to the desired posterior distributions. But, controlling the convergence of the MCMC algorithms towards their limit distributions is important to get reliable estimates.
- *Prior error:* The prior knowledge on the parameters m and C is expected to produce regularized estimates of smaller variances than maximum likelihood estimates. But, if the prior distributions are irrelevant, it could jeopardize the statistical analysis.

It is crucial to measure and reduce the possible impact of those errors. In this thesis, four chapters (Chapters II-V) have been desired to response this essential question. Especially, in this Bayesian framework, an original criterion has been proposed to assess the relevance of the numerical DOE and the prior choices from the point of view of a minimum error. To control the *emulator error*, as the evaluation budget of the complex function is severely limited, the choice of DOEs play a critical role. An adaptive kriging methodology has been constructed to improve the quality of the DOEs on a tight model evaluation budget. The uncertainty brought by the meta-model has been reduced and a better posterior distribution $\pi(\theta|\mathbf{Y})$ has been gained.

A real case of a complex hydrogeological computer code has been treated by applying the statistical tools and methodologies developed in the thesis.

C. Organization of the manuscript

Addressing the issues presented in the previous section, the present manuscript is organized as follows:

Chapter 1 provides a review of probabilistic inverse problems and presents the main statistical tools dedicated to estimation. Some important concepts of Bayesian inference are recalled. Meta-modeling technique and MCMC algorithms are among the main points of interest in this chapter.

Chapter 2 is devoted to the construction of the Bayesian model. A modified version of the uncertainty model adapted to the meta-modeling is proposed by introducing an additional type of uncertainty. The elicitation of the hyperparameters of the prior distributions is another central point of this chapter.

Chapter 3 is concerned with the management of the hybrid MCMC algorithms we use. The so-called Metropolis-Hastings-within-Gibbs algorithm is our principal tool to solve inverse problems. Two versions according to the available computational budget are presented.

Chapter 4 focuses on assessing the Bayesian treatment of inverse problems combined with the meta-modeling technique. The quality of DOEs is crucial for improving the accuracy of the meta-model, and it can be measured with different criteria. We propose an original criterion adapted to the Bayesian framework, which allows to check the consistency between the prior choices, the observed data and the choices of numerical DOE. The behavior of this criterion is illustrated on numerical experiments.

Chapter 5 deals with the problem of building the DOEs in an adaptive way, such that the prediction accuracy of the meta-model can be improved. Two Bayesian criteria have been proposed, one consists of reducing the global uncertainty and enhancing the exploration of regions of interest and the other aims at controlling the divergence between the current posterior distribution and the hypothesized posterior distribution, by sequentially enriching the current DOE.

A real case-study of uncertainty treatment in our hydraulic engineering is treated in Chapter 6, which permits to apply the methodologies previously proposed. Two important industrial codes used at EDF, MASCARET and TELEMAC-2D are considered. Finally, a conclusion and perspective chapter ends this thesis.

I

Introduction to inverse problems in uncertainty analysis

Contents

In this chapter, we define inverse problems and the notion "*uncertainty*" adapted to our case study. Algorithms such as the ECME, SEM, SAEM algorithms can be used in the frenquentist framework to estimate the parameters of interest. In this thesis, the Bayesian inference is favored and a Metropolis-within-Gibbs algorithm (or hybrid MCMC algorithm) has been carried out. Moreover, as the central difficulty in inverse problems, the computer simulator is often highly time-consuming who needs great numerical cares. Meta-models such as *kriging* are considered to approximate the original expensive-to-compute simulator.

I.1 Probabilistic inverse problems

I.1.1 General definition of inverse problems

Inverse problems are the problems where an unobserved variable $x \in \mathcal{X}$ is estimated from an observed variable $y \in \mathcal{Y}$ related through a physical model, which is usually complex and expensive to compute, the so-called "black-box" function. Mathematically, an inverse problem can be defined as follows.

Definition 1. *Let $y \in \mathcal{Y}$ and $H : \mathcal{X} \to \mathcal{Y}$ a deterministic function, an* **inverse problem** *is to find $x^* \in \mathcal{X}$ such that*

$$y \quad = \quad H(x^*). \tag{I.1}$$

Equivalently, it is to find a solution $x^* \in H^{-1}(\{y\})$, where $H^{-1}(\{y\}) = \{x \in \mathcal{X} \,|\, y = H(x)\}$.

I.1.2 Adapted model of probabilistic inverse problems

In our case study, the problem is somewhat different. Introducing the notion of "*uncertainty*", we aim at calibrating the distribution of the model input by taking into account its variability. More precisely, the distribution of the input x is to be explored instead of a possible "unique" solution x^*. Moreover, the observed output y is also considered uncertain by adding a random measurement error U. The *probabilistic framework* is as follows. For the i-th observation sample,

$$\begin{aligned} y_i &= y^R(x_i) + U_i \\ &= H(x_i) + \underbrace{\big(y^R(x_i) - H(x_i)\big)}_{\text{negligible model error}} + \underbrace{U_i}_{\text{measurement error}} \\ &= H(x_i) + U_i \end{aligned} \tag{I.2}$$

where $y^R(\cdot)$ denotes the real physical observation and H denotes the *computer simulator* which is supposed to almost perfectly represent the physical reality. In other words, the model error is assumed to be negligible. Thus, as presented in (I.2), we summarize two types of uncertainty considered in the present work: one comes from the random variable x_i and the other is offered by the measurement error U_i.

Given a vector of n observed outputs $\mathbf{y} = (y_1^T, \ldots, y_n^T)^T$ corresponding to the vector of n unobserved inputs $\mathbf{x} = (x_1^T, \ldots, x_n^T)^T$, both x_i and y_i are assumed to be the realizations of the real-valued random vectors $X_i \in \mathcal{X} \subseteq \mathbb{R}^q$ and $Y_i \in \mathcal{Y} \subseteq \mathbb{R}^p$. Moreover, another vector of

observed inputs $\mathbf{d} = (d_1^T, \ldots, d_n^T)^T$, with $d_i \in \mathcal{D}$, is introduced to take different experimental conditions into account. Our specific probabilistic inverse problem can now be defined.

Definition 2. *Given the following complex physical model H*

$$H : \left| \begin{array}{rcl} \mathcal{X} \times \mathcal{D} & \longrightarrow & \mathcal{Y} \\ (x_i, d_i) & \longmapsto & y_i = H(x_i, d_i), \end{array} \right. \tag{I.3}$$

the **probabilistic inverse problem** *is to calibrate the distribution of the unobserved variable $X_i \in \mathcal{X}$ from the observations (y_i, d_i), with the following relationship*

$$Y_i \;=\; H(X_i, d_i) + U_i, \;\; i \in \{1, \ldots, n\}. \tag{I.4}$$

Here Y_i denotes the random vector related to the observation y_i, d_i denotes an observed input related to the experimental conditions and U_i denotes the measurement error.

In this thesis, it is assumed that, in the model (I.4) the probability distribution of the unobserved random data $(X_i) \in \mathbb{R}^q$ is the product of n independent Gaussian distributions:

$$X_i \,|\, m, C \;\sim\; \mathcal{N}_q(m, C), \; (1 \leq i \leq n), \tag{I.5}$$

where the parameters m and C are to be estimated. It is worth noting that (I.5) explains the most important source of the uncertainty introduced in the inverse problem model. Moreover, the measurement errors $(U_i) \in \mathbb{R}^p$, as another source of uncertainty, are assumed to independently follow a centered Gaussian distribution:

$$U_i \;\sim\; \mathcal{N}_p(\mathbf{0}, R), \; (1 \leq i \leq n), \tag{I.6}$$

with a known diagonal matrix R. The error U_i and X_i are assumed to be independent for $i = 1, \ldots, n$, and the observations $(Y_i, i = 1, \ldots, n)$ with $Y_i \in \mathbb{R}^p$ are assumed to be independent between them. Under the limited evaluation budget of the complex function H, the purpose of our work is to calibrate the distribution of X_i. In other words, it is to provide an estimate of the parameters $\theta = (m, C) \in \Theta$, from the observed data $\mathbf{y}$ and $\mathbf{d}$.

I.1.3 A motivating example

An example in hydraulic engineering fields concerning the modeling of river inflows can be found in Parent et al. (1991, (75)). It consists of predicting the risk of dyke overflow during a flood which generally applies at a given river section.

As shown in Figure I.1, the observation y_i is a two-dimensional vector composed with the water level Z_c at the dyke position and the speed of the river V and the observed input d_i measures the observed flow of river Q. The two-dimensional missing data X_i is assumed to be made up with the value of Strickler coefficient K_s and the river bed level Z_v at the dyke. Moreover, the river bed level beyond upstream Z_m, the section length L as well as its width B are assumed to be fixed.

Assuming the following relationships

$$Z_c \;=\; Z_v + \left(\frac{\sqrt{L}}{B\sqrt{Z_m - Z_v}} \times \frac{Q}{K_s} \right)^{3/5}, \quad \text{and } V = \frac{Q}{B(Z_c - Z_v)},$$

with the help of the following notation

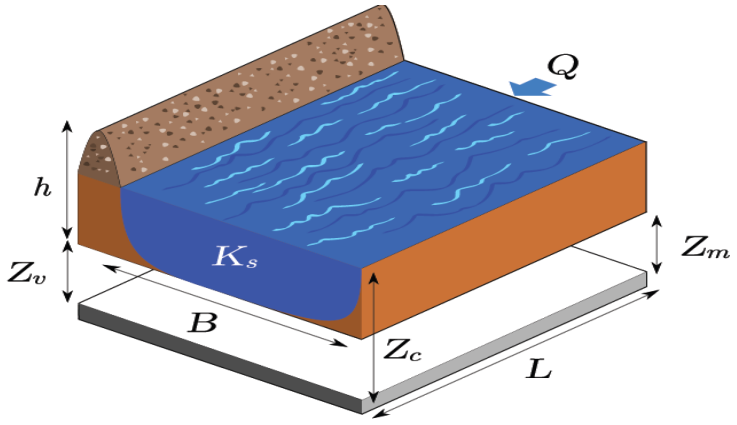


Figure I.1: Simplified hydraulic model of a section of the river

- the observed output $Y = (Y_1, Y_2)^T = (Z_c, V)^T \in \mathbb{R}^2$;
- the unobserved input $X = (X_1, X_2)^T = (K_s, Z_v)^T \in \mathbb{R}^2$;
- the observed input $d = Q \in \mathbb{R}^1$,

we recognize the standard form of the probabilistic inverse problems described in (I.4), where the expensive-to-compute function H can be derived as:

$$H(X,d) = \left(X_2 + \left(\frac{\sqrt{L}}{B\sqrt{Z_m - X_2}} \times \frac{d}{X_1} \right)^{0.6}, \frac{d^{0.4} X_1^{0.6} (Z_m - X_2)^{0.3}}{B^{0.4} \times L^{0.3}} \right). \tag{I.7}$$

I.2 Frequentist inference

The non-measurability of the X_is is usually caused by technical or cost reasons. In frequentist inference, the maximum-likelihood estimation (MLE) method aims at computing the maximizer

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; \mathbf{y}, \mathbf{d}), \tag{I.8}$$

where $L(\cdot)$ denotes the log-likelihood for the purpose of simplicity, and $(\mathbf{y}, \mathbf{d}) = (y_i, d_i, i = 1, \ldots, n)$ denotes the observations. In the present missing data context, the likelihood maximization is based on a mechanism of *data augmentation* (Tanner and Wong, 1987, (106)), which leads to the following calculation:

$$L(\theta;\, \mathbf{y}, \mathbf{d}) = \log \int \mathcal{L}(\theta;\, \mathbf{X}, \mathbf{y}, \mathbf{d})\, d\mathbf{X}, \tag{I.9}$$

where $\mathcal{L}$ denotes the standard likelihood of θ based on the completed sample $(\mathbf{X}, \mathbf{y}, \mathbf{d})$, which can be described as

$$\begin{aligned} \mathcal{L}(\theta;\, \mathbf{X}, \mathbf{y}, \mathbf{d}) &= \exp\Big[\log \pi \left[\mathbf{X}, \mathbf{y}, \mathbf{d} \,|\, \theta \right] \Big] \\ &= \exp\Big[\log \pi \left[\mathbf{y} \,|\, \mathbf{X}, \mathbf{d}, R \right] + \log \pi \left[\mathbf{X} \,|\, \theta \right] \Big]. \end{aligned} \tag{I.10}$$

In this formula, π denotes the corresponding probability density function (pdf) and the two log-densities in (I.10) are derived from the distribution assumptions (I.5) and (I.6):

$$\log \pi\left[\mathbf{y} \,|\, \mathbf{X}, \mathbf{d}, R\right] \quad \propto \quad -n \log |R| - \sum_{i=1}^{n} \Big[(y_i - H(X_i, d_i))^T R^{-1} (y_i - H(X_i, d_i)) \Big] \tag{I.11}$$

$$\log \pi\left[\mathbf{X} \,|\, \theta\right] \quad \propto \quad -n \log |C| - \sum_{i=1}^{n} \Big[(X_i - m)^T C^{-1} (X_i - m) \Big]. \tag{I.12}$$

Remark 2. *In (I.10), the term $\log \pi\left[\mathbf{d} \,|\, \theta\right]$ is omitted as it equals zero.*

Obviously, the cost of the likelihood expressed in (I.10) causes difficulty to find the maximizer $\widehat{\theta}$, due to the time-consuming function H (see (I.11)). Especially, it can even be non-integrable if H is not linear. This prevents to use the standard EM algorithm (Dempster et al., 1977, (25)), which approaches a local maximizer θ^* by maximizing the conditional expectation of the completed $L(\theta; \mathbf{X}, \mathbf{y}, \mathbf{d})$ at each iteration.

For this concern, Celeux et al. (2010, (15)) proposed a linearized version of the simulator H, where the linearization point x_0 is chosen based on prior knowledge. The approximated model is as follows:

$$Y_i \quad = \quad H(x_0, d_i) + \mathbf{J}_H(x_0, d_i)(X_i - x_0) + U_i,\ 1 \leq i \leq n, \tag{I.13}$$

with $\mathbf{J}_H(x_0, d_i) \in \mathcal{M}^{p \times q}$ denoting the Jacobian matrix of the function H at point x_0. (I.13) can be written in the following simplified form:

$$Y_i \quad = \quad H_i X_i + V_i + U_i,\ 1 \leq i \leq n, \tag{I.14}$$

where H_i is the known Jacobian matrix and V_i gathers all the remaining terms. By composing the matrix $\mathbf{H} \in \mathcal{M}^{np \times q}$ by $\{H_i, i = 1, \ldots, n\}$:

$$\mathbf{H} \quad = \quad \begin{pmatrix} H_1 \\ \vdots \\ H_n \end{pmatrix}, \tag{I.15}$$

Celeux et al. (2010, (15)) have proved the following proposition to ensure the identifiability of the linearized model (I.14). The identifiability implies that the estimation problem of θ is well defined, namely the uniqueness of the estimated θ.

Proposition 1. *(Celeux et al., 2010, (15)) Assuming $q \leq np$, Model (I.14) is identifiable if and only if rank$(\mathbf{H}) = q$, i.e. H is injective.*

Remark 3. *(Empirical Identifiability) In practice, the condition $q \leq np$ mentioned in Proposition 1 is not sufficient to ensure that enough data is available for estimation. Consider $q = p = 1$, at least $np = 2$ observations are necessary to estimate m and C. With $np = 1$ observation, only the mean can be estimated. Hence, in the Gaussian case, a supplementary condition should be added to ensure that the estimation is feasible. For instance, $n_0 q \leq np$, with n_0 greater than 2.*

I.2.1 Method based on linearization: the ECME algorithm (*Circe*)

An extension of the EM algorithm, the so-called ECME (Expectation-Conditional Maximization Either) algorithm, is presented by Liu and Rubin (1994, (60)). For the linearized model (I.13), the ECME algorithm was independently proposed by De Crecy (1996, (20)), under the name of the "*Circe*" method. Typically, ECME is maximizing the observed likelihood of some parameters and the expectation of the completed likelihood for the other parameters. Thus, it accelerates the convergence ensured by the standard EM algorithm. The $(k+1)$-th iteration of the adapted ECME method is described as follows (see De Crecy, 1996, (20)):

- **E step**: `Compute the conditional expectation of the complete log-likelihood`

$$\mathbb{E}_{\theta^{(k)}}[L(\theta;\mathbf{y},\mathbf{X}) \mid \mathbf{y},\mathbf{d}] \quad = \quad \int L(\theta;\mathbf{y},\mathbf{X})\pi(\mathbf{X} \mid \mathbf{y},\mathbf{d},\theta=\theta^{(k)})d\mathbf{X}, \tag{I.16}$$

  where $\mathbf{X}$ denotes the set of n variables $(X_i, i=1,\ldots,n)$ and $\mathbb{E}_{\theta^{(k)}}[f(\mathbf{X})|\mathbf{y},\mathbf{d}]$ indicates the conditional expectation of $f(\mathbf{X})$ knowing the current parameter $\theta=\theta^{(k)}$ and the observations $(\mathbf{y},\mathbf{d})$, with respect to the probability density:

$$\pi(\mathbf{X} \mid \mathbf{y},\mathbf{d},\theta^{(k)}) \quad = \quad \frac{\pi(\mathbf{X},\mathbf{y} \mid \mathbf{d},\theta^{(k)})}{\pi(\mathbf{y} \mid \mathbf{d},\theta^{(k)})}. \tag{I.17}$$

- **CME steps**: `Update the parameters by`

  1. `estimating` C `with` m `fixed to` $m^{(k)}$ `(same as the M step of the EM algorithm)`

$$C^{(k+1)} \quad = \quad \arg\max_{C} \mathbb{E}_{\theta^{(k)}}[L(m^{(k)},C;\mathbf{y},\mathbf{X})|\mathbf{y},\mathbf{d}]; \tag{I.18}$$

  2. `estimating` m `with` C `fixed to` $C^{(k+1)}$`, based on the incomplete-data log-likelihood`

$$m^{(k+1)} \quad = \quad \arg\max_{m} L(m,C^{(k+1)};\mathbf{y}). \tag{I.19}$$

When the function H is not highly non linear, the ECME algorithm works well in practice as shown in Celeux et al. 2010, (15), while the choice of the linearization point x_0 remains essential for good performance of this algorithm. A simple solution so-called the *iterative linearization* (applied to ECME) is described as follows:

- `Initial step: Start from a linearization point` $x_{\texttt{lin}}=x_0$ `and compute` $H(x_0,d_i)$ `and` $J_H(x_0,d_i)$`. Initiated on the point` $\theta_{\texttt{init}}=(x_0,C_0)$`, the ECME algorithm leads to the estimate` $\theta^{(1)}$`.`
- `Step` k+1`: Let the linearization point be` $x_{\texttt{lin}}=m^{(k)}$ `and compute` $H(x_{\texttt{lin}},d_i)$ `and` $J_H(x_{\texttt{lin}},d_i)$`. A new estimate` $\theta^{(k+1)}$ `is given by the ECME algorithm initiated at` $\theta_{\texttt{init}}=\theta^{(k)}=(m^{(k)},C^{(k)})$`.`

The algorithm is repeated until some stopping criterion, e.g.

$$\max_j \left(\frac{|\theta_j^{(r+1)}-\theta_j^{(r)}|}{|\theta_j^{(r)}|} \right) \quad \leq \quad \epsilon \tag{I.20}$$

with ϵ a positive small value to be specified, is satisfied.

Remark 4. *Each linearization step k requires n calls of H for $H(x_{lin}, d_i)$ plus $n \times q \times a$ additional calls of H for the computation of the Jacobian matrix $J_H(x_{lin}, d_i)$, with a varying from 1 to say 5 according to the roughness of H through finite differences, e.g. $a = 1$ for the first order finite difference; $a = 2$ for the second order finite difference. Thus, this iterate linearization can be quite time-consuming.*

Moreover, if H is highly non linear, the algorithm does not perform well as the linear approximation of H is not satisfactory. The sequence $(\theta^{(k)})$ does not converge or converge to a misleading estimate. For these reasons, other types of methods have been introduced to avoid the linearization step, e.g. the SEM algorithm (Celeux and Diebolt, 1985, (13)), the MCEM algorithm (Wei and Tanner, 1990, (117)) and the SAEM algorithm (Delyon et al., 1999, see also Kuhn, 2003, (55)). In what follows, the SEM algorithm is presented.

I.2.2 Method avoiding linearization : the SEM Algorithm

The SEM algorithm is regarded as the stochastic version of the EM algorithm, proposed by Celeux and Diebolt in 1985 and applied in the present framework by Barbillon (2010, (3)) to a problem similar to ours. On the $(k+1)$-th iteration, the algorithm consists of the following three steps:

- **E step**: `Calculate the conditional distribution` $\pi(\cdot \,|\, \mathbf{y}, \mathbf{d}; \theta^{(k)})$ `of` $\mathbf{X}^{(k)}$`, with` $\theta^{(k)}$ `the current estimates of the parameters` θ`;`
- **S step**: `Simulate` $\mathbf{X}^{(k)} \sim \pi(\cdot \,|\, \mathbf{y}, \mathbf{d}; \theta^{(k)})$ `and complete the sample` $\mathbf{Z}^{(k)} = (\mathbf{y}, \mathbf{X}^{(k)})$`;`
- **M step**: `Update the parameters`

$$\theta^{(k+1)} = \arg\max_{\theta \in \Theta} L(\theta;\, \mathbf{Z}^{(k)}, \mathbf{d}). \tag{I.21}$$

`That is to say,`

$$C^{(k+1)} = \arg\max_{C} L(m^{(k)}, C; \mathbf{Z}^{(k)}, \mathbf{d}); \tag{I.22}$$

$$m^{(k+1)} = \arg\max_{m} L(m, C^{(k+1)}; \mathbf{Z}^{(k)}, \mathbf{d}). \tag{I.23}$$

Usually, the conditional density $\pi(\cdot \mid \mathbf{y}, \mathbf{d}; \theta^{(k)})$ does not belong to any known family, therefore a numerical method is needed for the simulation step, typically the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953, (69)). Each **S step** consists of m iterations of the MH algorithm, described as follows:

`For each sample` $i = 1, \ldots, n$`,`

- `Initialize` $X_{i,0} = X_i^{(k)}$.
- `For` $s = 1, \ldots, m$`,`
  1. `Simulate` $\tilde{X}_{i,s}$`, using an instrumental distribution` $q(\cdot \mid X_{i,s-1}; \theta^{(k)})$.

2. Let $X_{i,s} = \tilde{X}_{i,s}$ with probability

$$\alpha(X_{i,s-1}, \tilde{X}_{i,s}) \quad = \quad \min\left(1, \frac{\pi(\tilde{X}_{i,s} \mid Y_i, d_i; \theta^{(k)})\, q(X_{i,s-1} \mid \tilde{X}_{i,s}; \theta^{(k)})}{\pi(X_{i,s-1} \mid Y_i, d_i; \theta^{(k)})\, q(\tilde{X}_{i,s} \mid X_{i,s-1}; \theta^{(k)})}\right);$$

on the contrary, $X_{i,s} = X_{i,s-1}$ with probability $1 - \alpha(X_{i,s-1}, \tilde{X}_{i,s})$.

- Take $X_i^{(k+1)} = X_{i,m}$.

After a sufficiently long *burn-in* period, the convergence of the chain $(X_{i,j})_j$ towards the target distribution $\pi(\cdot \mid Y_i, d_i; \theta^{(k)})$ (for the i-th sample) can be checked with the Brooks-Gelman (BG) statistics (Brooks and Gelman, 1998, (11)). Note that there exists a variety of convergence diagnostics apart from the BG statistic. A comparison of such criteria can be found in Cowles and Carlin (1996, (19)).

However, each iteration the MH mechanism involves high number of calls to H, which is highly CPU time consuming. Cheaper versions of H can be considered to replace the original model. Various approximation methods are described in Section I.4.

I.3 Bayesian inference

In the present work, a Bayesian viewpoint has been chosen. Bayesian inference allows to take into account the available expert knowledge by choosing an informative prior, which is favorable especially in a small sample setting. In frequentist inference, not enough observations can be quite burdensome, as the MLE may not perform well in such cases. Apart from that, choosing an informative prior may solve some identifiability problems (Paulino and Pereira, 1994, (79)), in particular when θ is of high dimension.

In Bayesian framework, the parameter θ is as random variables. The available knowledge of the model, the prior information (assembled in the chosen prior distribution $\pi(\theta)$) and the observations $(\mathbf{y}, \mathbf{d})$ are incorporated in the posterior distribution $\pi(\theta|\mathbf{y}, \mathbf{d})$, calculated according to the Bayes' rule,

$$\pi(\theta|\mathbf{y}, \mathbf{d}) \quad \propto \quad \mathcal{L}(\theta; \mathbf{y}, \mathbf{d}) \cdot \pi(\theta), \tag{I.24}$$

where $\mathcal{L}(\theta; \mathbf{y}, \mathbf{d})$ denotes the likelihood of θ based on the observations $(\mathbf{y}, \mathbf{d})$.

I.3.1 Prior choices

The chosen prior distributions of the parameters $\theta = (m, C)$ in model (I.4) are conjugate prior distributions:

- $m \,|\, C \sim \mathcal{N}_q(\mu, C/a)$;
- $C \sim \mathcal{IW}_q(\Lambda, \nu) \in \mathcal{M}^{q \times q}$,

$\mathcal{IW}_q(\Lambda, \nu)$ being the Inverse-Wishart distribution, with $\nu > q-1$ the degrees of freedom and $\Lambda \in \mathcal{M}^{q\times q}$ the positive definite inverse scale matrix. Hyperparameters μ, a, Λ and ν are to be specified. The density of C is:

$$\pi(C) \quad = \quad \frac{|\Lambda|^{\nu/2}}{2^{\frac{\nu q}{2}}\,\Gamma_q(\frac{\nu}{2})}\,|C|^{-\frac{\nu+q+1}{2}}\,\exp\Big[-\frac{1}{2}\,\mathrm{Tr}(\Lambda\cdot C^{-1})\Big], \tag{I.25}$$

where $\Gamma_q(a)$ denotes the multivariate Gamma function:

$$\Gamma_q(a) \quad = \quad \pi^{\frac{q(q-1)}{4}}\prod_{j=1}^{q}\Gamma(a+\frac{1-j}{2}) \tag{I.26}$$

with $\Gamma(a)$ the Gamma function . The restriction that $\nu > q-1$ is necessary to give sense to the function $\Gamma(\frac{\nu+1-q}{2})$.

Moreover, the mean of C exists if $\nu > q+1$, and the variance of C exists if $\nu > q+3$, since:

$$\mathbb{E}(C) \quad = \quad \frac{\Lambda}{\nu-q-1}, \tag{I.27}$$

$$\mathrm{Var}(C_{i,j}) \quad = \quad \frac{(\nu-q+1)\Lambda_{i,j}^2+(\nu-q-1)\Lambda_{i,i}\Lambda_{j,j}}{(\nu-q)(\nu-q-1)^2(\nu-q-3)},\ \forall i,j=1,\ldots,q. \tag{I.28}$$

Remark 5. *When $q=1$, the uni-variate Inverse-Wishart is the Inverse-Gamma distribution specified by:*

$$\mathcal{IW}_1(\Lambda,\nu) \quad = \quad \mathcal{IG}\left(\frac{\Lambda}{2},\frac{\nu}{2}\right). \tag{I.29}$$

In the next chapter, we will discuss the calibration of the hyperparameters μ, a, Λ and ν in detail. As it will be shown, the elicitation of hyperparameters can benefit from either conditional conjugation properties of the concept of a virtual sample, which simplify the inferential computational work. For the purpose of simplicity, the hyperparameters are grouped in $\rho = (\mu, a, \Lambda, \nu)$. Remark that the variance matrix R of error U_i is not included in ρ as it is assumed to be known.

The final aim is to estimate the posterior distribution $\pi(\theta\,|\mathbf{Y}=\mathbf{y},\mathbf{d})$. However, as calculated in (I.24), this distribution is not available in closed form, as the calculation of the log-likelihood L involves complex integration when the function H is not linear (see the discussion in Section I.2). Dedicated numerical methods, e.g. Gibbs sampling, should be carried out to approximate the posterior distribution.

I.3.2 Introduction to hybrid MCMC algorithms

The Gibbs algorithm, named after the physicist J. W. Gibbs, is first presented by Geman & Geman (1984 (34)). The main idea (adapted to our current case) is to draw alternatively each unknown quantity (the parameters m, C and the missing data $\mathbf{X}$) from its full conditional posterior distribution knowing the current simulated values of the other parameters as well as the observations $(\mathbf{y},\mathbf{d})$ and the hyperparameters ρ. With the initialized value $(m^{(0)}, C^{(0)}, \mathbf{X}^{(0)}) = (m_0, C_0, \mathbf{X}_0)$, the $(k+1)$-th iteration consists of two steps:

- `Given` $(\theta^{(k)}, \mathbf{X}^{(k)})$, `generate the parameter` $\theta^{(k+1)}$ following

$$C^{(k+1)} \sim \pi(\cdot \,|\, m^{(k)}, \mathbf{X}^{(k)}, \mathbf{Y} = \mathbf{y}, \mathbf{d}, \rho); \tag{I.30}$$
$$m^{(k+1)} \sim \pi(\cdot \,|\, C^{(k+1)}, \mathbf{X}^{(k)}, \mathbf{Y} = \mathbf{y}, \mathbf{d}, \rho). \tag{I.31}$$

- `Simulate the missing data` $\mathbf{X}^{(k+1)}$ `given the current parameter` $\theta^{(k+1)}$ `and the observations` $(\mathbf{y}, \mathbf{d})$ (*Data Augmentation Step*)

$$\mathbf{X}^{(k+1)} \sim \pi(\cdot \,|\, m^{(k+1)}, C^{(k+1)}, \mathbf{Y} = \mathbf{y}, \mathbf{d}, \rho). \tag{I.32}$$

Usually, the full conditional distribution of $\mathbf{X}$ does not belong to any known family of distributions because of the involved complex code H. A numerical method, such as the Metropolis-Hastings (MH) algorithm, is thus necessary. The so-called Metropolis-Hastings-within-Gibbs (Hybrid MCMC) algorithm (see for instance Tierney, 1995, (109)) is applied in this thesis, which is to be presented in Chapter III.

Under some regularity conditions (cf. Section III.2), the simulated Markov chain $\left(m^{(k)}, C^{(k)}, \mathbf{X}^{(k)}\right)$ is proved to converge towards its stationary distribution that is the joint posterior distribution $\pi(m, C, \mathbf{X} \,|\, \mathbf{Y} = \mathbf{y}, \mathbf{d}, \rho)$. Therefore, each marginal simulated quantity converges to its marginal posterior distribution, i.e. $\pi(m \,|\, \mathbf{Y} = \mathbf{y}, \mathbf{d})$, $\pi(C \,|\, \mathbf{Y} = \mathbf{y}, \mathbf{d})$ and $\pi(\mathbf{X} \,|\, \mathbf{Y} = \mathbf{y}, \mathbf{d})$.

I.4 Black-box function and Gaussian Process meta-modeling (*Kriging* method)

In model (I.4), the function H is usually highly time-consuming. A surrogate is thus constructed to replace the original model, which is required to carry out the iterative methods, such as the stochastic SEM, MCEM, SAEM and Gibbs algorithms. More precisely, the predictor $\widehat{H}$ is to be built from the observations gathered in a dataset. Various approximation methods can be carried out: linear models fit by least squares, local methods (such as the arithmetic mean method, k-nearest-neighbor method, distance weighted method), polynomial interpolation, spline, kernel methods (*kriging* meta-modeling techniques) and so on.

The *kriging* meta-modeling technique has been chosen in this work. Introduced by the French mathematician Georges Matheron and developed by Sacks et al. (1989b, (98)), Koehler and Owen (1996, (54)), Santner et al. (2003, (99)) and Fang et al. (2006, (28)), this approximation method consists of deriving a predictor $\widehat{H}(z)$ at any $z = (x, d) \in \Omega$, from the training set $\mathbf{H}_{D_N}$ evaluated from a *design of experiments* (DOE)

$$D_N = \left(z_{(1)}^T, \ldots, z_{(N)}^T\right)^T, \tag{I.33}$$

with each $z_{(j)} = (x_{(j)}, d_{(j)})$. It is known as a meta-model, which signifies a simplified representation or approximation of a simulator built on a training set of simulator runs. According to this approach, the function H is considered as the realization of a stationary Gaussian Process (GP) $\mathcal{H}$, described as follows:

$$\forall z \in \Omega, \qquad \mathcal{H}(z) = \sum_{i=1}^{k} \beta_i f_i(z) + \mathbf{Z}(z) = \mathbf{F}(z)^T \beta + \mathbf{Z}(z) \tag{I.34}$$

where:

i) the first term is a linear regression model based on the given basis functions $\mathbf{F}(z) = [f_1(z), \ldots, f_k(z)]^T$, which correspond to weight coefficients $\beta = [\beta_1, \ldots, \beta_k]^T$, with $k \leq N$ for the purpose of identifiability;

ii) the second term $\mathbf{Z}$ is a Gaussian Process with zero mean

$$\mathbb{E}\left[\mathbf{Z}(z)\right] = 0, \ \forall z \in \Omega, \tag{I.35}$$

and stationary autocovariance

$$\mathrm{Cov}\left[\mathbf{Z}(z), \mathbf{Z}(z')\right] = \sigma^2 K_\psi(z - z'), \ \forall (z, z') \in \Omega^2, \tag{I.36}$$

where K_ψ is a symmetric positive definite kernel, the so-called autocorrelation function. It only depends on the distance between z and z', with ψ the regularization parameter in the structure of K_ψ, such that

$$K_\psi(\mathbf{0}) = 1. \tag{I.37}$$

Remark 6. *In particular, in* (I.36)*, the assumption of homogeneity in different directions can be imposed by assuming that* K_ψ *depends on the norm* $\left\|z - z'\right\|$ *instead of the difference vector* $z - z'$*, called* ***isotropy****. Let us note that both stationarity and isotropy remain reasonable in our case study.*

Remark 7. *If* $\mathbf{F}(z)$ *is assumed to be zero,* (I.34) *is called the* ***simple*** *kriging model; if* $\mathbf{F}(z)$ *is assumed to be unit, the model is called the* ***ordinary*** *kriging; otherwise, it is called the* ***universal*** *kriging, namely that* $\mathbf{F}(z)$ *takes any general form. We apply the universal kriging model in the present work.*

The reasons for choosing the kriging method are triple.

1. The kriging modeling takes into account the spatial structure of DOE in terms of *correlation* between the points of design, by adjusting the parameters (β, σ^2, ψ). Thus it outperforms other methods like local methods, polynomial interpolation and so on.
2. The GP meta-modeling can be interpreted as providing H with some prior information (Rasmussen and Williams, 2006, (86)), which is related to the choice of the training set $\mathbf{H}_{D_N}$, the linear regression basis $\mathbf{F}(z)$ and the autocorrelation function K_ψ. It is coherent with the Bayesian perspective.
3. Kriging provides the prediction variance at every estimated point z, which can be regarded as an indicator of the accuracy of the approximation and a measure of the uncertainty introduced by the meta-model (see the definition of MSE(z) in Proposition 2).

For the choice of K_ψ, several types of autocorrelation are introduced here:

- the nugget autocorrelation function:

$$K_\psi(z - z') = \delta(z - z') = \begin{cases} 1, & \text{if } z = z'; \\ 0, & \text{otherwise.} \end{cases} \tag{I.38}$$

It is used to model the absence of any correlation between points z and z', which means that all the realizations of the GP $\mathbf{Z}(\cdot)$ are assumed to be independent and identically distributed (i.i.d.). $\mathbf{Z}(\cdot)$ is then known as a *white noise* and the trajectory of the process is discontinuous.

- the exponential autocorrelation function:

$$K_\psi(z - z') = \exp\left(-\sum_{i=1}^{q+q_2} \psi_i \left|z_i - z_i'\right|^{\nu_i}\right); \tag{I.39}$$

where $\{\psi_i > 0, i = 1, \ldots, q + q_2\}$ are the scale parameters and the degrees $\nu_i \in [0, 2]$.

If all ν_is equal 1, it is called the *Exponential* kernel and the *Gaussian* kernel if they are equal to 2. Otherwise, we call it the a *Generalized* kernel.

If for $\forall i$, $0 < \nu_i < 2$, the sample paths of the GP $\mathcal{H}(\cdot)$ with such an exponential kernel are almost surely (a.s.) continuous but not differentiable for any $0 < \nu_i < 2$. However, if all ν_is equal 2, the sample paths of the GP $\mathcal{H}(\cdot)$ with a Gaussian correlation are a.s. continuous and a.s. infinitely differentiable, which will give rise to a very smooth process.

Remark 8. *If all ν_is equal* 1 *and all the ψ_is are equal, this exponential autocorrelation function corresponds to a specific GP known as the* **Ornstein-Uhlenbeck** *process.*

- the Matérn autocorrelation function:

$$K_\psi(z - z') = \prod_{i=1}^{q+q_2} \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\psi_i \left|z_i - z_i'\right|\right)^\nu \mathcal{K}_\nu \left(\psi_i \left|z_i - z_i'\right|\right), \tag{I.40}$$

where $\{\psi_i > 0, i = 1, \ldots, q + q_2\}$ are the scale parameters, $\nu > 0$ is the regularization parameter of the associated GP, Γ denotes the Gamma function and $\mathcal{K}_\nu$ denotes the modified Bessel function of the second kind and order ν. It is worth noting that with such a Matérn kernel, if $\nu > m$, the GP $\mathbf{Z}$ is m times a.s. derivable.

Remark 9. *The Gaussian autocorrelation function is the limit case of a Matérn autocorrelation function, derived from the following convergence:*

$$\frac{1}{2^{\nu-1}\Gamma(\nu)} \left(2\sqrt{\psi\nu}\left|z - z'\right|\right)^\nu \mathcal{K}_\nu \left(2\sqrt{\psi\nu}\left|z - z'\right|\right) \underset{\nu\to\infty}{\longrightarrow} \exp\left(-\psi|z - z'|^2\right). \tag{I.41}$$

Concerning the choice of the autocorrelation function, it is sometimes suggested to use the *Matérn* function which is more flexible thanks to the regularization parameter ν adjusting the power of the distance between z and z'. However, the computation is more expensive as we have more parameters to be estimated. In this thesis, we choose the smooth *Gaussian* autocorrelation function which is a.s. continuous and a.s. infinitely differentiable. Note that the choice of the autocorrelation function is not the central point of our work. We are more concerned on validating and improving the choice of DOEs, which will be presented in Chapter IV and V.

Remark 10. *As shown in Ababou et al. (1994, (1)) and in Marrel (2008, (64)), a Gaussian autocorrelation ($\nu_i = 2$) might imply an ill-conditioned variance matrix, which may lead to great numerical problems. Introducing an independent white noise $V(z)$ into the kriging model (I.34) to add a discontinuity property, the so-called nugget effet, may improve the autocorrelation condition and the robustness of the kriging approximation.*

From the Bayesian viewpoint, Santner et al. (2003, (99)) remarked that under the Gaussian assumption defined by (I.34)-(I.37), the vector gathering the process $\mathcal{H}(\cdot)$ at any point $z \notin D_N$ and at the design D_N is normally distributed:

$$\left\{ \begin{array}{c} \mathcal{H}(z) \\ \mathcal{H}_{D_N} \end{array} \right\} \sim \mathcal{N}_{1+N}\left(\left\{ \begin{array}{c} \mathbf{F}(z)^T\beta \\ \mathbf{F}_D\beta \end{array} \right\}, \sigma^2 \left[\begin{array}{cc} 1 & \mathbf{\Sigma}_{zD}^T \\ \mathbf{\Sigma}_{zD} & \mathbf{\Sigma}_{DD} \end{array} \right] \right), \tag{I.42}$$

where

- $\mathbf{F}(z) = [f_1(z), \ldots, f_k(z)]^T$ is a $k \times 1$ vector of basis functions evaluated at z. For example, the complete polynomial basis up to order k

$$\mathbf{F}(z) = [1, z, z^2, \ldots, z^k]^T; \tag{I.43}$$

- $\mathbf{F}_D = [\mathbf{F}(z_{(1)}), \ldots, \mathbf{F}(z_{(N)})]^T$ is a $N \times k$ regression matrix evaluated at the design D_N, defined as

$$\mathbf{F}_D = \left[\begin{array}{cccc} f_1(z_{(1)}) & f_2(z_{(1)}) & \ldots & f_k(z_{(1)}) \\ \vdots & \vdots & \vdots & \vdots \\ f_1(z_{(N)}) & f_2(z_{(N)}) & \ldots & f_k(z_{(N)}) \end{array} \right]. \tag{I.44}$$

  Following the choice (I.43) of the regressors, the matrix $\mathbf{F}_D$ is indeed the Vandermonde matrix:

$$\mathbf{F}_D = \left[\begin{array}{ccccc} 1 & z_{(1)} & z_{(1)}^2 & \ldots & z_{(1)}^k \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & z_{(N)} & z_{(N)}^2 & \ldots & z_{(N)}^k \end{array} \right]; \tag{I.45}$$

- $\mathbf{\Sigma}_{zD} = [K_\psi(z, z_{(1)}), \ldots, K_\psi(z, z_{(N)})]^T$ is a $N \times 1$ vector of correlations between the point of interest z and each point of the design D_N;
- $\mathbf{\Sigma}_{DD} = [K_\psi(z_{(i)}, z_{(j)})]_{1 \leq i,j \leq N}$ is a $N \times N$ correlation matrix evaluated within the design of experiments.

Remark 11. *The GP $Z(\cdot)$ is assumed to be regular, which means that regardless of the choice of the training sample D_N, the variance matrix Σ_{DD} is invertible. Under the assumption of stationarity, this is equivalent to assume that $K(\cdot)$ to be positive definite.*

Following (I.42), the posterior distribution of the GP $\mathcal{H}(z)$ given the evaluations $\mathbf{H}_{D_N}$ can be proved to be normally distributed,

$$\mathcal{H}(z) \,|\, \mathcal{H}_{D_N} = \mathbf{H}_{D_N} \sim \mathcal{N}\Big[\mu_H(z), \sigma_H^2(z)\Big]. \tag{I.46}$$

Proposition 2. *Assuming β is unknown and the autocovariance $\sigma^2 K_\psi(\cdot)$ is known, the conditional mean $\mu_H(z)$ is the best linear unbiased predictor (BLUP) of the unobserved function value $H(z)$ under the Gaussian assumptions (I.34)-(I.37), renamed $\widehat{H}(z)$ described as:*

$$\widehat{H}(z) \;=\; \mathbb{E}\Big[\mathcal{H}(z)\,|\,\mathcal{H}_{D_N} = \mathbf{H}_{D_N}\Big] \;=\; \mathbf{F}(z)^T\widehat{\beta} + \mathbf{\Sigma}_{zD}^T\mathbf{\Sigma}_{DD}^{-1}(\mathbf{H}_{D_N} - \mathbf{F}_D\widehat{\beta}). \tag{I.47}$$

The conditional variance $\sigma^2_{\mathcal{H}}(z)$ is the minimal variance, the so-called MSE (mean squared error), computed by:

$$MSE(z) = \mathbb{E}\left[\Big(\mathcal{H}(z) - \widehat{H}(z)\Big)^2\,|\,\mathcal{H}_{D_N} = \mathbf{H}_{D_N}\right] \;=\; \sigma^2\Big(1 + \gamma(z)^T(\mathbf{F}_D^T\mathbf{\Sigma}_{DD}^{-1}\mathbf{F}_D)^{-1}\gamma(z) - \mathbf{\Sigma}_{zD}^T\mathbf{\Sigma}_{DD}^{-1}\mathbf{\Sigma}_{zD}\Big), \tag{I.48}$$

where

$$\widehat{\beta} \;=\; (\mathbf{F}_D^T\Sigma_{DD}^{-1}\mathbf{F}_D)^{-1}\mathbf{F}_D^T\Sigma_{DD}^{-1}\mathbf{H}_{D_N} \tag{I.49}$$

is the generalized least-square estimate of β, and

$$\gamma(z) \;=\; \mathbf{F}(z) - \mathbf{F}_D^T\mathbf{\Sigma}_{DD}^{-1}\mathbf{\Sigma}_{zD}. \tag{I.50}$$

Sketch of the proof. By definition, the best linear unbiased predictor $\widehat{H}(z)$ of the unknown value $H(z)$ has the following properties:

- linear, which can be defined as a linear combination of the evaluations $\mathbf{H}_{D_N}$, with help of a weight vector $a_0 \in \mathbb{R}^N$:

$$\widehat{H}(z) \;=\; a_0^T\mathbf{H}_{D_N}, \tag{I.51}$$

- unbiased:

$$\mathbb{E}\left[\widehat{H}(z) - \mathcal{H}(z)\right] \;=\; 0, \tag{I.52}$$

- the best in the sense of mean squared error:

$$\widehat{H}(z) \;=\; \arg\min_{\widehat{H}\text{ linear unbiased}} \mathbb{E}\left[\Big(\widehat{H}(z) - \mathcal{H}(z)\Big)^2\right]. \tag{I.53}$$

The problem consists of finding the optimal weight vector a_0^* in the following sense:

$$a_0^* \;=\; \arg\min_{a_0\in\mathbb{R}^N} \mathbb{E}\left[(a_0^T\mathbf{H}_{D_N} - \mathcal{H}(z))^2\right] \;\text{ such that }\; \mathbb{E}\left[a_0^T\mathbf{H}_{D_N} - \mathcal{H}(z)\right] = 0. \tag{I.54}$$

The detailed proof can be found in the thesis of Dubourg, V. (2011, (26)). In Proposition 2, the BLUP $\widehat{H}$ can also be called the best MSPE (*mean squared prediction error*) predictor of H. The minimal variance MSE provides a measure of the prediction accuracy. □

Given the observations $\mathbf{H}_{D_N}$, the outputs $\mathcal{H}(z)$ and $\mathcal{H}(z')$ at the points z and z' are correlated because of the autocorrelation function K_ψ, which depends on the distance between these points, where the covariance can be calculated by

$$\mathrm{Cov}\left[\mathcal{H}(z), \mathcal{H}(z')\,|\,\mathcal{H}_{D_N} = \mathbf{H}_{D_N}\right] \;=\; \sigma^2\Big(K_\psi(z - z') + \gamma(z)^T(\mathbf{F}_D^T\mathbf{\Sigma}_{DD}^{-1}\mathbf{F}_D)^{-1}\gamma(z') - \mathbf{\Sigma}_{zD}^T\mathbf{\Sigma}_{DD}^{-1}\mathbf{\Sigma}_{z'D}\Big). \tag{I.55}$$

Remark 12. *Assuming that β is known in model (I.34), the BLUP $\widehat{H}$ can be written as*

$$\widehat{H}(z) \quad = \quad \mathbf{F}(z)^T\beta + \mathbf{\Sigma}_{zD}^T\mathbf{\Sigma}_{DD}^{-1}(\mathbf{H}_{D_N} - \mathbf{F}_D\beta), \tag{I.56}$$

with the minimal variance described as

$$MSE(z) \quad = \quad \sigma^2\Big(1 - \mathbf{\Sigma}_{zD}^T\mathbf{\Sigma}_{DD}^{-1}\mathbf{\Sigma}_{zD}\Big), \tag{I.57}$$

which is smaller than the MSE(z) defined in (I.48).

An interesting geometric interpretation (Vazquez, E., 2005, (112)) Given a Hilbert space $\mathcal{G}$ provided with the inner product

$$< h, k > \quad = \quad \mathbb{E}\left[hk\right], \tag{I.58}$$

the best linear predictor (BLP) $\widehat{H}(z)$ is in fact the orthogonal projection of the GP $\mathcal{H}(z)$ on the subspace $\mathcal{G}_s$ generated by the evaluations $\mathbf{H}_{D_N}$, i.e. the unique vector $h \in \mathcal{G}_s$ verifying:

$$< \mathcal{H}(z) - h,\, H(z_{(i)}) > \quad = \quad 0,\ \forall i \in \{1, \ldots, N\}. \tag{I.59}$$

The prediction variance MSE(z) corresponds to the distance between $\mathcal{H}(z)$ and $\widehat{H}(z)$ which is orthogonal to the observations $\mathbf{H}_{D_N}$, as illustrated in Figure I.2. It is worth noting that the Hilbert space $\mathcal{G}$ is usually the space of random variables of finite variance $\mathcal{L}^2\left(\Omega, \mathcal{A}, \mathbb{P}\right)$.

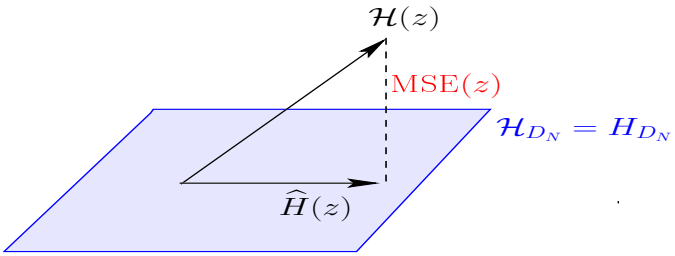


Figure I.2: A geometric interpretation of the kriging approximation

Furthermore, under the assumption that the covariance parameters (σ^2, ψ) are unknown, $\widehat{\mathcal{H}}(z)$ is called the empirical best linear unbiased predictor (EBLUP) of $H(z)$, computed by

$$\widehat{H}(z) \quad = \quad \mathbf{F}(z)^T\widehat{\beta} + \widehat{\mathbf{\Sigma}}_{zD}^T\widehat{\mathbf{\Sigma}}_{DD}^{-1}(\mathbf{H}_{D_N} - \mathbf{F}_D\widehat{\beta}), \tag{I.60}$$

where

$$\widehat{\beta} \quad = \quad (\mathbf{F}_D^T\widehat{\mathbf{\Sigma}}_{DD}^{-1}\mathbf{F}_D)^{-1}\mathbf{F}_D^T\widehat{\mathbf{\Sigma}}_{DD}^{-1}\mathbf{H}_{D_N}, \tag{I.61}$$

and $\widehat{\mathbf{\Sigma}}_{DD}$, $\widehat{\mathbf{\Sigma}}_{zD}$ are the estimators of $\mathbf{\Sigma}_{DD}$ and $\mathbf{\Sigma}_{zD}$, with help of the kernel $K_{\widehat{\psi}}$. The minimal variance MSE(z) is then given by

$$\text{MSE}(z) \quad = \quad \widehat{\sigma^2}\Big(1 + \widehat{\gamma}(z)^T(\mathbf{F}_D^T\widehat{\mathbf{\Sigma}}_{DD}^{-1}\mathbf{F}_D)^{-1}\widehat{\gamma}(z) - \widehat{\mathbf{\Sigma}}_{zD}^T\widehat{\mathbf{\Sigma}}_{DD}^{-1}\widehat{\mathbf{\Sigma}}_{zD}\Big), \tag{I.62}$$

with σ^2 replaced by its maximum likelihood estimator

$$\widehat{\sigma^2} \quad = \quad \frac{1}{N}(H_D - \mathbf{F}_D\widehat{\beta})^T\widehat{\mathbf{\Sigma}}_{DD}^{-1}(H_D - \mathbf{F}_D\widehat{\beta}). \tag{I.63}$$

Remark 13. *The name EBLUP is misleading, as it is often non-linear or even biased because of the non-linear estimator* $\widehat{\mathbf{\Sigma}}_{DD}$ *and* $\widehat{\mathbf{\Sigma}}_{zD}$.

Note that at each point $z_{(i)} \in D$, we have the following interpolation property:

$$\widehat{H}(z_{(i)}) = H(z_{(i)}), \tag{I.64}$$
$$\mathrm{MSE}(z_{(i)}) = \mathbf{0}. \tag{I.65}$$

Finding BLUP consists of solving a linear system of order of magnitude $\mathcal{O}(N^2)$, with N the number of points in the design D_N. It will be impractical and harmful for large data sets (N large). That is why various fast approximation algorithms are considered to compute the *discrete Gauss transform* to replace the Gaussian autocorrelation function. The *Improved fast Gauss transform* (IFGT) algorithm is proposed by Memarsadeghi et al. (2008, (68)), which reduces the computational cost from the quadratic $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$. Another method called *Gauss transform with nearest neighbors* (GTANN) was implemented by Raykar (2007, (87)), which is more efficient when the Gaussian models have small ranges. It is worth noting that in our case study, the budget of the time-consuming function H is always dominant.

I.4.1 Properties of the best linear unbiased predictor

- **Interpolation**

  Figure I.3 illustrates the property of the meta-model $\widehat{H}$ interpolating the function $H(x) = x\sin(x)$, based on the dataset $\mathbf{H}_{D_5} = \{x_{(1)}, \ldots, x_{(5)}\}$:

$$\begin{cases} \widehat{H}(x_{(i)}) = H(x_{(i)}); \\ \\ \mathrm{MSE}(x_{(i)}) = \mathbf{0}, \ i = 1, \ldots, 5. \end{cases} \tag{I.66}$$

  This property can be easily proved from the construction of the predictor $\widehat{H}$.

- **Asymptotic consistency**

  Vazquez (2005, (112)) proved in his thesis that the universal kriging predictor $\widehat{H}$ is asymptotically consistent if the autocorrelation function $K_\psi(\cdot)$ of the GP $\mathbf{Z}$ is *continuous* on the diagonal, in the sense that

$$\mathbb{E}\left[\left(\mathcal{H}(z) - \widehat{H}(z)\right)^2 | \, \mathcal{H}_{D_N} = \mathbf{H}_{D_N}\right] \xrightarrow{\mathbb{P}} 0. \tag{I.67}$$

  The property of the kriging technique is illustrated in Figure I.4, which shows that the prediction variance MSE remains zero at the design points and converges to zero at any points $x \in \Omega$ when the number N of points increase. The convergence of the integrated MSE, the so-called IMSE, has been discussed in Le Gratiet et al. (2012, (57)). This converging property is equivalent to say the predictor $\widehat{H}$ is asymptotically consistent when the DOE becomes dense in the domain Ω.

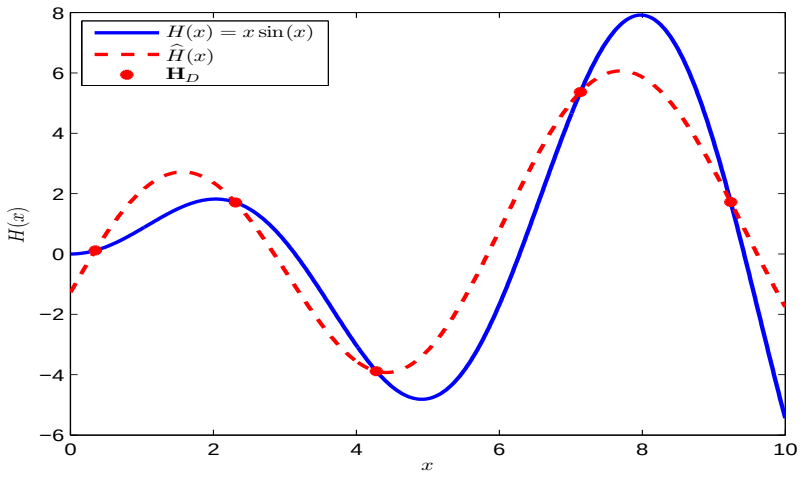


Figure I.3: Illustration of the interpolation property by the one-dimensional function $H(x) = x\sin(x)$, with squared exponential autocovariance meta-model

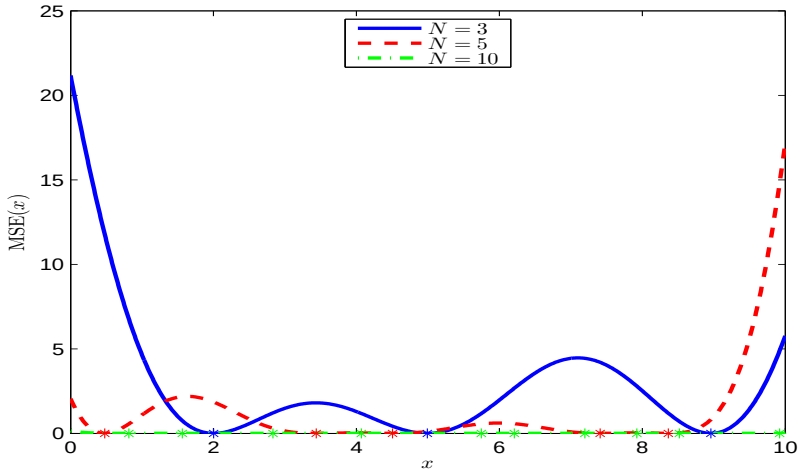


Figure I.4: Illustration of the asymptotic consistency property by the one-dimensional function $H(x) = x\sin(x)$, with squared exponential autocovariance meta-model

- **Gaussianity**

  The Gaussian assumption of the process $\mathcal{H}$ knowing the evaluations $\mathbf{H}_{D_N}$:

$$\mathcal{H}(z)\,|\,\mathbf{H}_{D_N} \quad \sim \quad \mathcal{N}\Big[\widehat{H}(z), \mathrm{MSE}(z)\Big], \tag{I.68}$$

  is equivalent to

$$\Big(\mathrm{MSE}(z)\Big)^{-1/2}\Big(\mathcal{H}(z) - \widehat{H}(z)\Big)\,|\,\mathbf{H}_{D_N} \quad \sim \quad \mathcal{N}\Big[\mathbf{0}_p, \mathbf{I}_p\Big]. \tag{I.69}$$

  The formula (I.69) is convenient to compute the confidence intervals, which can be

described as

$$\mathcal{H}(z)\,|\,\mathbf{H}_{D_N} \in \left[\widehat{H}(z) - \Big(\mathrm{MSE}(z)\Big)^{-1/2}\Phi^{-1}\left(1-\frac{\alpha}{2}\right), \widehat{H}(z) + \Big(\mathrm{MSE}(z)\Big)^{-1/2}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right], \tag{I.70}$$

for a confidence interval of probability $1-\alpha$. Φ^{-1} denotes the inverse cumulative distribution function (cdf) of the standard normal distribution. For instance, with the level α fixed to 10%, $\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ being 1.64 and the 90%-confidence interval is illustrated in Figure I.5.

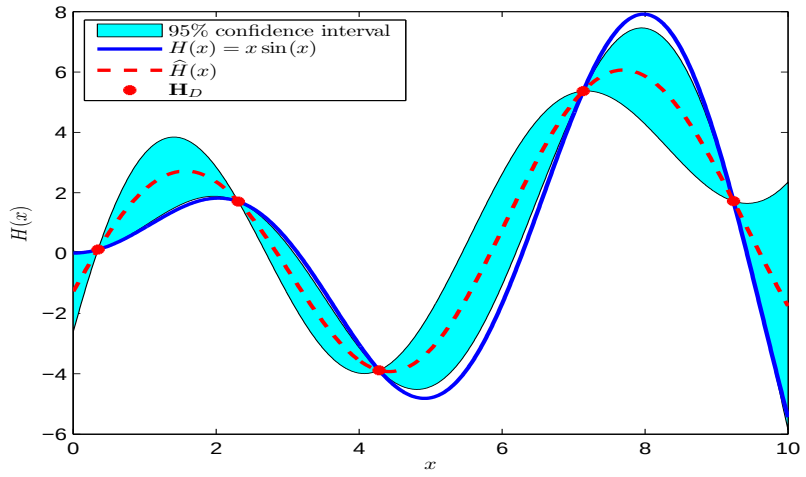


Figure I.5: Illustration of the Gaussianity property by the one-dimensional function $H(x) = x\sin(x)$, with squared exponential autocovariance meta-model

An application proposed in Jones et al. (1990, (47)) is to validate the meta-model by providing a confidence interval based on the Cross-Validation *Leave-One-Out*. More precisely, it consists of verifying if for the majority of design points, say 99.7%, we have

$$\Big(\mathrm{MSE}_{-i}(z_{(i)})\Big)^{-1/2}\left(\mathcal{H}(z_{(i)}) - \widehat{H}_{-i}(z_{(i)})\right)\,|\,\mathbf{H}_{D_{-i}} \in \Big[-3,3\Big], \tag{I.71}$$

where $\widehat{H}_{-i}(z_{(i)}$ and $\mathrm{MSE}_{-i}(z_{(i)})$ denote the predictor and the prediction variance at the point $z_{(i)}$ evaluated from the design of experiments

$$D_{-i} = \{z_{(1)},\ldots,z_{(i-1)},z_{(i+1)},\ldots,z_{(N)}\}. \tag{I.72}$$

I.4.2 Estimation of parameters (β,σ^2,ψ) for EBLUP

- **Maximum likelihood estimation (MLE)**

The MLE method consists of estimating the parameters as the maximizers of the likelihood function, or equivalently the log-likelihood function. Up to an additive constant, the log-likelihood L with respect to the observations $\mathbf{H}_{D_N}$ is

$$L\left(\beta,\sigma^2,\psi\,|\,\mathbf{H}_{D_N}\right) = -\frac{1}{2}\left[N\log\sigma^2 + \log\Big(|\mathbf{\Sigma}_{DD}(\psi)|\Big) + \frac{(\mathbf{H}_{D_N}-\mathbf{F}_D\beta)^T\mathbf{\Sigma}_{DD}^{-1}(\psi)(\mathbf{H}_{D_N}-\mathbf{F}_D\beta)}{\sigma^2}\right]. \tag{I.73}$$

The maximum likelihood estimates of β and σ^2 depend on ψ:

$$\widehat{\beta}(\psi) = (\mathbf{F}_D^T \boldsymbol{\Sigma}_{DD}^{-1}(\psi)\mathbf{F}_D)^{-1}\mathbf{F}_D^T \boldsymbol{\Sigma}_{DD}^{-1}(\psi)\mathbf{H}_{D_N}, \tag{I.74}$$

$$\widehat{\sigma^2}(\psi) = \frac{1}{N}(\mathbf{H}_{D_N} - \mathbf{F}_D\widehat{\beta})^T \boldsymbol{\Sigma}_{DD}^{-1}(\psi)(\mathbf{H}_{D_N} - \mathbf{F}_D\widehat{\beta}). \tag{I.75}$$

Plugging the two estimators in (I.73) leads to a new expression of the log-likelihood which depends only on ψ, up to an additive constant:

$$L\left(\widehat{\beta}, \widehat{\sigma^2}, \psi \,|\, \mathbf{H}_{D_N}\right) = -\frac{1}{2}\left[N \log \widehat{\sigma^2}(\psi) + \log\left(|\boldsymbol{\Sigma}_{DD}(\psi)|\right) + N\right]. \tag{I.76}$$

Thus the maximum likelihood estimate $\widehat{\psi}$ of ψ verifies:

$$\widehat{\psi} = \arg\min_{\psi}\left[N \log \widehat{\sigma^2}(\psi) + \log\left(|\boldsymbol{\Sigma}_{DD}(\psi)|\right)\right]. \tag{I.77}$$

By injecting the estimates $(\widehat{\beta}, \widehat{\sigma^2}, \widehat{\psi})$ in (I.60), we obtain then the EBLUP. It is worth noting that a calculation of order $\mathcal{O}(N^3)$ is necessary to evaluate the log-likelihood (I.73), which will be very expensive if N is large. Another remark is that the global optimization problem in (I.77) cannot be analytically solved and numerical global optimization techniques are usually required. For instance, the DACE Matlab toolbox (Lophaven et al., 2002, (63)) uses the BOXMIN algorithm which is a multivariate dichotomy algorithm, while the DiceKriging R package (Roustant et al., 2010, (95)) resorts to a gradient-based genetic algorithm (Sekhon and Mebane, 2011, (105)). The MLE algorithm applied in this thesis is already implemented in the DACE toolbox by Lophaven et al.

- **Restricted maximum likelihood estimation (RMLE)**

This RMLE (Patterson and Thompson, 1971, (78)) method aims to construct a less biased estimator of the parameters (β, σ^2, ψ) than the maximum likelihood estimator. The estimator of σ^2 is written as a function of ψ, given by

$$\widetilde{\sigma^2}(\psi) = \frac{1}{N-k}(\mathbf{H}_{D_N} - \mathbf{F}_D\widehat{\beta})^T \boldsymbol{\Sigma}_{DD}^{-1}(\psi)(\mathbf{H}_{D_N} - \mathbf{F}_D\widehat{\beta}), \tag{I.78}$$

where $\widehat{\beta}$ is the maximum likelihood estimator of β given by (I.74).

Remark 14. *We have* $\widetilde{\sigma^2} = \frac{N}{N-k}\widehat{\sigma^2}$, *with* $\widehat{\sigma^2}$ *the maximum likelihood estimate of* σ^2.

Plugging the estimator $\widetilde{\sigma^2}(\psi)$ into the "restricted" log-likelihood which depends only on ψ, up to an additive constant, gives the following result:

$$L\left(\widetilde{\sigma^2}, \psi \,|\, \mathbf{H}_{D_N}\right) = -\frac{1}{2}\left[(N-k)\log \widetilde{\sigma^2}(\psi) + \log\left(|\boldsymbol{\Sigma}_{DD}(\psi)|\right) + \log\left(|\mathbf{F}_D^T \boldsymbol{\Sigma}_{DD}^{-1}(\psi)\mathbf{F}_D|\right) + N - k\right], \tag{I.79}$$

we obtain the estimator of ψ as follows:

$$\widetilde{\psi} = \arg\min_{\psi}\left[(N-k)\log \widetilde{\sigma^2}(\psi) + \log\left(|\boldsymbol{\Sigma}_{DD}(\psi)|\right) + \log\left(|\mathbf{F}_D^T \boldsymbol{\Sigma}_{DD}^{-1}(\psi)\mathbf{F}_D|\right)\right] \tag{I.80}$$

Knowing $\widetilde{\psi}$, the RML estimator of β can be then calculated as $\widetilde{\beta} = \widehat{\beta}\left(\widetilde{\psi}\right)$. Note that the RMLE can be applied for the intrinsic kriging while the MLE is not applicable in this case.

Moreover, Li and Sudjianto (2005, (59)) proposed another approach to reduce the variance of estimation by penalizing the MLE. Fang et al. (2006, (28)) showed the advantage of this penalized method in the case where DOE contains only few points.

- **Cross-Validation (Leave-One-Out Prediction) (CV)**

The cross-validation technique is a popular tool for model selection (see Allen, 1971, (2) and Stone, 1974, (105)), which consists of regrouping the data set D_N in K mutually exclusive and collectively exhaustive subsets D_k with $k = 1, \ldots, K$, such that

$$D_i \cap D_j = \varnothing, \ \forall i, j = 1, \ldots, K \text{ and } \bigcup_{k=1}^{K} D_k = D_N. \tag{I.81}$$

The k-th cross-validated prediction is obtained by predicting the function value on the k-th fold using all the $K-1$ subsets $D_N \backslash D_k$. The parameters $\phi := (\beta, \sigma^2, \psi)$ can be estimated by minimizing the squared cross-validated error:

$$\widehat{\phi}_K = \arg\min_{\phi} \frac{1}{K} \sum_{k=1}^{K} \left(\widehat{H}_{D_N \backslash D_k}(\phi) - H_{D_k} \right)^2. \tag{I.82}$$

If $K = N$, the K-fold cross-validation method is called the *leave-one-out* procedure and each subset $D_k = \{z_{(k)}\}$ is the k-th point of the design. (I.82) becomes then

$$\widehat{\phi}_N = \arg\min_{\phi} \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{H}_{-i}(z_{(i)})(\phi) - H(z_{(i)}) \right)^2, \tag{I.83}$$

where $\widehat{H}_{-i}(z_{(i)})(\phi)$ is the predictor of $H(z_{(i)})$ obtained from the evaluations of H at all the design points D_N except the i-th point $z_{(i)}$, by using (I.60).

The choice of K could be sensitive: if $K = N$, $\widehat{\phi}_K$ is an asymptotically unbiased estimator, but may have a large variance; if K takes smaller values, $\widehat{\phi}_K$ has a smaller variance but can be biased. In practice, we often use $K = 5$ or $K = 10$ which seems to be a good compromise for the available sample size we are facing.

However, in the case of meta-model, the cross-validation technique does not estimate the variance parameter σ^2 as it is not involved in the predictor formula (I.60) and therefore not in Eq (I.83). One way to proceed is to find σ^2 such that the following ratio

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\left(\widehat{H}_{-i}(z_{(i)})(\phi) - H(z_{(i)}) \right)^2}{\mathrm{MSE}_{-i}(z_{(i)})(\phi)}, \tag{I.84}$$

is close to 1, i.e. the mean squared error should be equal to the empirical squared error (the numerator), where $\mathrm{MSE}_{-i}(z_{(i)})(\phi)$ is the variance obtained from all training data except $z_{(i)}$, using the formula (I.48).

- **Bayesian Predictors (BP)**

An alternative Bayesian inference consists of computing the posterior distribution of the unknown parameters $\phi = (\beta, \sigma^2, \psi)$, knowing the observations $\mathbf{H}_{D_N}$ and under some prior assumpions on ϕ.

A practical choice of the prior distribution (Santner et al. 2003, (99)) is as follows:

$$\pi(\phi) \quad = \quad \pi(\beta|\sigma^2)\pi(\sigma^2)\pi(\psi), \tag{I.85}$$

as it is reasonable to assume that the autocorrelation parameter ψ is a priori independent from (β, σ^2). The posterior distribution of the parameters can be derived with the help of Bayes' formula:

$$\pi(\phi|\mathbf{H}_{D_N}) \quad \propto \quad \pi(\mathbf{H}_{D_N}|\phi)\pi(\phi). \tag{I.86}$$

Proposition 2 and Remark 12 tell that the conditional distribution of the GP $\mathcal{H}(z)$ is a Gaussian distribution:

$$\left[\mathcal{H}(z)\,|\,\mathbf{H}_{D_N}, \phi\right] \quad \sim \quad \mathcal{N}_p\left[\mathbf{F}(z)^T\beta + \mathbf{\Sigma}_{zD}^T\mathbf{\Sigma}_{DD}^{-1}(\mathbf{H}_{D_N} - \mathbf{F}_D\beta), \sigma^2\Big(1 - \mathbf{\Sigma}_{zD}^T\mathbf{\Sigma}_{DD}^{-1}\mathbf{\Sigma}_{zD}\Big)\right]. \tag{I.87}$$

The joint posterior distribution of $\left[\mathcal{H}(z), \phi\,|\,\mathbf{H}_{D_N}\right]$ can then be obtained by applying Bayes' formula:

$$\pi(\mathcal{H}(z), \phi\,|\,\mathbf{H}_{D_N}) \quad \propto \quad \pi(\mathcal{H}(z)\,|\,\mathbf{H}_{D_N}, \phi)\pi(\mathbf{H}_{D_N}\,|\,\phi)\pi(\phi). \tag{I.88}$$

Finally, marginalizing the above joint posterior distribution by integrating ϕ out leads to the *posterior predictive* distribution:

$$\pi(\mathcal{H}(z)\,|\,\mathbf{H}_{D_N}) \quad = \quad \int \pi(\mathcal{H}(z), \phi\,|\,\mathbf{H}_{D_N})\,d\phi. \tag{I.89}$$

In Proposition 2, only β is assumed unknown, which corresponds to the case where no prior information for β is available, as shown in Santner et al. (2003, (99)). Thus, the corresponding non informative prior distribution is

$$\pi(\phi) \quad = \quad \pi(\beta) \,\propto\, 1, \tag{I.90}$$

which helps deriving the *posterior mode*

$$\widehat{\beta} \quad = \quad \arg\max_{\beta}\left[\mathcal{L}\left(\beta\,|\,\mathbf{H}_{D_N}\right)\pi(\beta)\right] \tag{I.91}$$

$$\quad = \quad \arg\max_{\beta}\mathcal{L}\left(\beta\,|\,\mathbf{H}_{D_N}\right), \tag{I.92}$$

and the posterior predictive distribution $\left[\mathcal{H}(z)\,|\,\mathbf{H}_{D_N}\right]$ in this case is

$$\mathcal{N}_p\left[\mathbf{F}(z)^T\widehat{\beta} + \mathbf{\Sigma}_{zD}^T\mathbf{\Sigma}_{DD}^{-1}(\mathbf{H}_{D_N} - \mathbf{F}_D\widehat{\beta}), \sigma^2\Big(1 + \gamma(z)^T(\mathbf{F}_D^T\mathbf{\Sigma}_{DD}^{-1}\mathbf{F}_D)^{-1}\gamma(z) - \mathbf{\Sigma}_{zD}^T\mathbf{\Sigma}_{DD}^{-1}\mathbf{\Sigma}_{zD}\Big)\right]. \tag{I.93}$$

We find the same expression for the predictor and the prediction variance as described in Proposition 2. Santner et al. (2003, (99)) recommend the MLE or the RMLE to obtain a good predictor from an empirical study, and MLE is chosen in our work which is already implemented by the DACE toolbox. In fact, it permits to estimate all the parameters with explicit formula and can be regarded as the non informative case with BP method. Moreover, it is known that under certain assumptions of differentiability of the likelihood, the MLE is asymptotically efficient. However, it should not be forgotten that the computation of the likelihood is expensive, especially when the number N of data is large, apart from the high cost of inverting the covariance matrix $\mathbf{\Sigma}_{DD}$ of size $N \times N$. For this reason, several approximating methods for the likelihood can be carried out (see Stein et al., 2004, (103)). A complete list of these methods can also be found in (86) (Rasmussen and Williams, 2006) .

II

Eliciting the prior distributions

Contents

As explained in the previous chapter, the Bayesian framework is chosen to solve inverse problems as it takes into account the prior information possibly coming from experts, and it is expected to be useful in a small sample size setting. In this chapter, we aim at calibrating prior distributions, which is required by Bayesian inference. Let us recall the Bayesian model we opt for:

$$X_i \,|\, m, C \sim \mathcal{N}_q(m, C), \tag{II.1}$$
$$U_i \sim \mathcal{N}_p(\mathbf{0}, R),\ (1 \leq i \leq n), \tag{II.2}$$

with prior assumptions

$$m \,|\, C \sim \mathcal{N}_q(\mu, C/a), \tag{II.3}$$
$$C \sim \mathcal{IW}_q(\Lambda, \nu), \tag{II.4}$$
$$\tag{II.5}$$

where μ, a, Λ, ν are the hyperparameters to be specified. This Bayesian model can be described by the following directed acyclic graph (DAG).

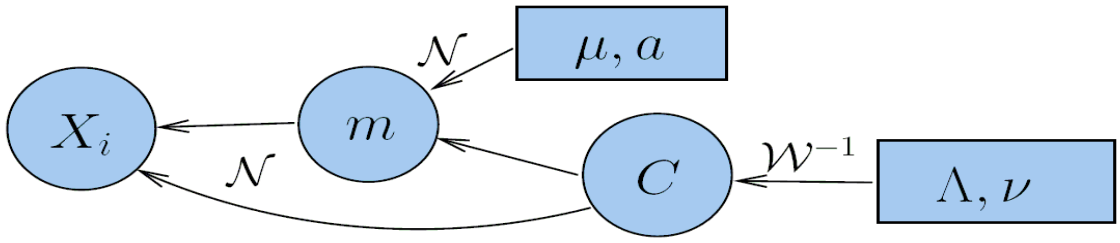


Figure II.1: DAG of the Bayesian model

To calibrate the prior distributions (II.4-II.5) , the full conditional posterior distributions of m, C and $\mathbf{X} = (X_1, \ldots, X_n)$, knowing the current simulated values, the observed data $(\mathbf{y}, \mathbf{d})$ and the hyperparameters $\rho = (\mu, a, \Lambda, \nu)$ will be useful. This calculation also provides a basis for the MCMC algorithm, presented in Chapter III. It is worth noting that the variance matrix R of the measurement error U_i is not a prior hyperparameter, which can be given by analyzing the measurement system or expertise. In the present work, R is assumed to be known.

In this chapter, we use the capital character $\mathbf{Y} = (Y_1^T, \ldots, Y_n^T)^T$ to denote the corresponding random process from which the observations $\mathbf{y} = (y_1^T, \ldots, y_n^T)^T$ arise.

II.1 Full conditional posterior distributions

For the computation of the full conditional posterior distributions of $(m, C, \mathbf{X})$, we distinguish two versions, the *rich man* version and the *poor man* version, according to our computational

budget. In the *rich man* version, which is not always realistic, it is assumed that the budget is so important that the number of calls to H is not limited. While, in the *poor man* version, the function H is replaced by a cheaper *kriging* meta-model due to a limited budget. We begin with the ideal *rich man* version.

II.1.1 Computation following the *rich man* version

Bayes' formula leads to the following equality:

$$\pi(m, C, \mathbf{X}|\mathbf{Y}, \mathbf{d}, \rho) \quad = \quad \frac{\pi(m, C, \mathbf{X}, \mathbf{Y}|\mathbf{d}, \rho)}{\pi(\mathbf{Y}|\mathbf{d}, \rho)}, \tag{II.6}$$

which is proportional to $\pi(m, C, \mathbf{X}, \mathbf{Y}|\mathbf{d}, \rho)$. We begin with treating this joint distribution. Let us recall that by injecting the prior distributions (II.4-II.5), the full conditional distribution of $\mathbf{Y}$ can be written as the product of n normal distributions:

$$Y_i \,|\, d_i, X_i, m, C, \rho \quad \sim \quad \mathcal{N}_p\Big[H(X_i, d_i), R\Big]. \tag{II.7}$$

Still applying Bayes' formula, the joint distribution $\pi(m, C, \mathbf{X}, \mathbf{Y} \,|\, \mathbf{d}, \rho)$ can be developed as follows:

$$\begin{aligned}
&\pi(m, C, \mathbf{X}, \mathbf{Y} \,|\, \mathbf{d}, \rho) \;\propto\; \pi(\mathbf{Y} \,|\, \mathbf{X}, m, C, \mathbf{d}, \rho) \cdot \pi(\mathbf{X} \,|\, m, C, \rho) \cdot \pi(m \,|\, C, \rho) \cdot \pi(C \,|\, \rho) \\
= \;& \left(\frac{1}{(2\pi)^{p/2}|R|^{1/2}}\right)^n \cdot \exp\Big[-\frac{1}{2}\sum_{i=1}^n \Big(Y_i - H(X_i, d_i)\Big)^T R^{-1}\Big(Y_i - H(X_i, d_i)\Big)\Big] \\
&\cdot \left(\frac{1}{(2\pi)^{q/2}|C|^{1/2}}\right)^n \cdot \exp\Big[-\frac{1}{2}\sum_{i=1}^n (X_i - m)^T C^{-1}(X_i - m)\Big] \\
&\cdot \left(\frac{1}{(2\pi)^{q/2}|\frac{C}{a}|^{1/2}}\right) \cdot \exp\Big[-\frac{1}{2}(m - \mu)T(\frac{C}{a})^{-1}(m - \mu)\Big] \\
&\cdot \frac{|\Lambda|^{\nu/2}\,|C|^{-\frac{\nu+q+1}{2}}\,\exp\Big[-\frac{1}{2}\operatorname{Tr}(\Lambda \cdot C^{-1})\Big]}{2^{\nu q/2}\,\Gamma_q(\nu/2)}.
\end{aligned}$$

Thus, up to an additive constant,

$$\begin{aligned}
\log \pi(m, C, \mathbf{X} \,|\, \mathbf{Y}, \mathbf{d}, \rho) \quad \overset{\exp}{\propto} \quad & -(\nu + n + q + 2)\log|C| \\
& - \sum_{i=1}^n (m - X_i)^T C^{-1}(m - X_i) \;-\; a(m - \mu)^T C^{-1}(m - \mu) \\
& - \sum_{i=1}^n \Big(Y_i - H(X_i, d_i)\Big)^T R^{-1}\Big(Y_i - H(X_i, d_i)\Big) \;-\; \operatorname{Tr}(\Lambda \cdot C^{-1}).
\end{aligned} \tag{II.8}$$

In this formula, by selecting the terms relative to m, C and $\mathbf{X}$, respectively, we obtain:

$$\log \pi(m \,|\, C, \mathbf{X}, \mathbf{Y}, \mathbf{d}, \rho) \overset{\exp}{\propto} -\sum_{i=1}^{n}(m - X_i)^T C^{-1}(m - X_i) - a(m-\mu)^T C^{-1}(m-\mu) \tag{II.9}$$

$$\begin{aligned}\log \pi(C \,|\, m, \mathbf{X}, \mathbf{Y}, \mathbf{d}, \rho) &\overset{\exp}{\propto} -(\nu + n + q + 2)\log|C| \\ &\quad - \mathrm{Tr}\Big[\Big(\sum_{i=1}^{n}(m - X_i)(m - X_i)^T + a(m-\mu)(m-\mu)^T + \Lambda\Big)C^{-1}\Big]\end{aligned} \tag{II.10}$$

$$\log \pi(\mathbf{X} \,|\, m, C, \mathbf{Y}, \mathbf{d}, \rho) \overset{\exp}{\propto} -\sum_{i=1}^{n}(m - X_i)^T C^{-1}(m - X_i) - \sum_{i=1}^{n}\Big(Y_i - H(X_i, d_i)\Big)^T R^{-1}\Big(Y_i - H(X_i, d_i)\Big). \tag{II.11}$$

The full conditional posterior distributions of m and C can then be determined as follows:

$$m \,|\, C, \mathbf{X}, \mathbf{Y}, \mathbf{d}, \rho \sim \mathcal{N}\Big(\frac{a}{n+a}\mu + \frac{n}{n+a}\overline{\mathbf{X}}_n, \frac{C}{n+a}\Big), \text{ where } \overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n} X_i; \tag{II.12}$$

$$C \,|\, m, \mathbf{X}, \mathbf{Y}, \mathbf{d}, \rho \sim \mathcal{IW}\Big(\Lambda + \sum_{i=1}^{n}(m - X_i)(m - X_i)^T + a(m-\mu)(m-\mu)^T,\ \nu + n + 1\Big). \tag{II.13}$$

While the full conditional posterior distribution of $\mathbf{X}$, as described in (II.11), cannot be formulated in a closed-form expression, due to the complex function H. Thus, numerical methods, typically Monte Carlo Markov chains (MCMC), are required to approximate this posterior distribution. This MCMC algorithm will be detailed in the next chapter.

The determination of the distribution (II.13) of C is direct. We provide the proof for m.

Proof. of (II.12)*:* We have from (II.9) that

$$
\begin{aligned}
& \log \pi(m \,|\, C, \mathbf{X}, \mathbf{Y}, \mathbf{d}, \rho) \\
= \; & -\Big[\sum_{i=1}^{n} (m - X_i)^T C^{-1} (m - X_i) + a(m-\mu)^T C^{-1} (m-\mu) \Big] \\
= \; & -\Big[\sum_{i=1}^{n} \sum_{j=1}^{q} \sum_{k=1}^{q} (m_j - X_i^j)(m_k - X_i^k) C^{-1}_{jk} + \sum_{j=1}^{q} \sum_{k=1}^{q} a(m_j - \mu_j)(m_k - \mu_k) C^{-1}_{jk} \Big] \\
= \; & -\sum_{j=1}^{q} \sum_{k=1}^{q} \Big[n m_j m_k - \sum_{i=1}^{n} m_j X_i^k - \sum_{i=1}^{n} m_k X_i^j + \sum_{i=1}^{n} X_i^j X_i^k + a m_j m_k - a m_j \mu_k \; - \; a m_k \mu_j + a \mu_j \mu_k \Big] \cdot C^{-1}_{jk} \\
= \; & -\sum_{j=1}^{q} \sum_{k=1}^{q} (n+a) \Big[m_j m_k - \frac{\sum_{i=1}^{n} X_i^k + a\mu_k}{n+a} m_j - \frac{\sum_{i=1}^{n} X_i^j + a\mu_j}{n+a} m_k \; + \; \frac{\sum_{i=1}^{n} X_i^j X_i^k + a\mu_j \mu_k}{n+a} \Big] \cdot C^{-1}_{jk} \\
= \; & -\sum_{j=1}^{q} \sum_{k=1}^{q} (n+a) \Big[\Big(m_j - \frac{\sum_{i=1}^{n} X_i^j + a\mu_j}{n+a} \Big) \Big(m_k - \frac{\sum_{i=1}^{n} X_i^k + a\mu_k}{n+a} \Big) \cdot C^{-1}_{jk} \\
+ \; & \Big(\frac{\sum_{i=1}^{n} X_i^j X_i^k + a\mu_j \mu_k}{n+a} - \frac{(\sum_{i=1}^{n} X_i^j + a\mu_j)(\sum_{i=1}^{n} X_i^k + a\mu_k)}{(n+a)^2} \Big) \cdot C^{-1}_{jk} \Big] \\
\overset{\exp}{\propto} \; & -\Big(m - \frac{n}{n+a} \overline{\mathbf{X}}_n - \frac{a}{n+a} \mu \Big)^T \Big(\frac{C}{n+a} \Big)^{-1} \Big(m - \frac{n}{n+a} \overline{\mathbf{X}}_n - \frac{a}{n+a} \mu \Big).
\end{aligned}
$$

The full conditional posterior distribution of m can then be easily deduced

$$
m \,|\, C, \mathbf{X}, \mathbf{Y}, \mathbf{d}, \rho \;\; \sim \;\; \mathcal{N}\Big(\frac{a}{n+a} \mu + \frac{n}{n+a} \overline{\mathbf{X}}_n, \; \frac{C}{n+a} \Big). \tag{II.14}
$$

□

II.1.2 Computation following the *poor man* version

In practice, the function H is usually highly computationally expensive. It is thus necessary to replace it with a cheaper *kriging* meta-model, as presented in Chapter I. As the need of this surrogate comes from a limited budget, it is called the *poor man* version.

In this version, the full conditional posterior distributions of m and C remain the same as in the *rich man* version, described in (II.12) and (II.13), since they are independent from H. The only change concerns the missing data $\mathbf{X}$, whose full conditional posterior distribution in the *rich man* version is described in (II.11).

An uncertainty model adapted to the meta-model

A direct and naive idea is to replace H by its *kriging* predictor $\widehat{H}$ defined in (I.47) such that the model (I.4) becomes

$$
Y_i \;\; = \;\; \widehat{H}(X_i, d_i) \, + \, U_i, \;\; i = 1, \ldots, n, \tag{II.15}
$$

and it leads to the following full conditional posterior distribution of $\mathbf{X}$, derived from (II.11):

$$\log \pi(\mathbf{X} \,|\, m, C, \mathbf{Y}, \mathbf{d}, \rho) \overset{\exp}{\propto} -\sum_{i=1}^{n}(m - X_i)^T C^{-1}(m - X_i) - \sum_{i=1}^{n}\Big(Y_i - \widehat{H}(X_i, d_i)\Big)^T R^{-1}\Big(Y_i - \widehat{H}(X_i, d_i)\Big). \tag{II.16}$$

However, this simple replacement ignores the uncertainty related to the meta-model. A more convenient solution is to consider Y_i as the realization of a Gaussian Process $\mathcal{Y}_i$, under assumption that H is the realization of a Gaussian process $\mathcal{H}$. The original model can be rewritten in the following way:

$$\mathcal{Y}_i = \widehat{H}(X_i, d_i) + \Big(\mathcal{H}(X_i, d_i) - \widehat{H}(X_i, d_i)\Big) + U_i, \tag{II.17}$$

$$= \widehat{H}(X_i, d_i) + \mathcal{V}_i(X_i, d_i), \quad i = 1, \ldots, n. \tag{II.18}$$

The property of the kriging modeling that it takes into account the spatial structure of the DOE permits us to model the dependence between different sample points. By defining n samples of the GP $\boldsymbol{\mathcal{Y}} = \{\mathcal{Y}_i, i = 1, \ldots, n\}$ and n samples of the input $\mathbf{Z} = \{Z_i, i = 1, \ldots, n\}$ with $Z_i = (X_i, d_i)$, the original model adapted in the *poor man* version can be written in the following form:

$$\boldsymbol{\mathcal{Y}} = \begin{pmatrix} \mathcal{Y}_{11} \\ \vdots \\ \mathcal{Y}_{1n} \\ \vdots \\ \mathcal{Y}_{p1} \\ \vdots \\ \mathcal{Y}_{pn} \end{pmatrix} = \begin{pmatrix} \widehat{H}_1(Z_1) \\ \vdots \\ \widehat{H}_1(Z_n) \\ \vdots \\ \widehat{H}_p(Z_1) \\ \vdots \\ \widehat{H}_p(Z_n) \end{pmatrix} + \underbrace{\begin{pmatrix} \mathcal{V}_{11}(Z_1) \\ \vdots \\ \mathcal{V}_{1n}(Z_n) \\ \vdots \\ \mathcal{V}_{p1}(Z_1) \\ \vdots \\ \mathcal{V}_{pn}(Z_n) \end{pmatrix}}_{\text{new uncertainty error}} = \widehat{H}(\mathbf{Z}) + \mathcal{V}(\mathbf{Z}), \tag{II.19}$$

with

$$\mathcal{V}_{ji}(Z_i) = \Big(\mathcal{H}_j(Z_i) - \widehat{H}_j(Z_i)\Big) + U_{ji}, \tag{II.20}$$

where $\mathcal{H}_j(Z_i)$ denotes the j-th component of the GP $\mathcal{H}$ at point Z_i, $\widehat{H}_j(Z_i)$ denotes the corresponding predictor and the U_{ji} denotes the j-th component of the measurement error U_i, with $i = 1, \ldots, n$ and $j = 1, \ldots, p$. It is worth noting that the new uncertainty error combines two types of uncertainty: one comes from the error term U_i (with the variance matrix R) and the other is derived from the *kriging* meta-model (i.e. the variance matrix MSE).

The advantage is that the correlation between the outputs $H(Z_k)$ and $H(Z_l)$ has been taken into account. In fact, for each j-th dimension H_j of the output, the kriging meta-model error for the whole sample $\{Z_i, i = 1, \ldots, n\}$ can be written as

$$\mathcal{E}_j = \Big(\mathcal{H}_j(Z_1) - \widehat{H}_j(Z_1), \ldots, \mathcal{H}_j(Z_n) - \widehat{H}_j(Z_n)\Big)^T \tag{II.21}$$

$$= (\mathcal{E}_{j1}, \ldots, \mathcal{E}_{jn})^T. \tag{II.22}$$

For two different sample points z_k and z_l with $k \neq l$, there exists a correlation between the outputs $\mathcal{H}(z_k)$ and $\mathcal{H}(z_l)$. Derived from (I.48), the covariance of the $j-$th dimension of the output is given by:

$$\mathrm{Cov}\left[\mathcal{E}_{jk}, \mathcal{E}_{jl} \,|\, \mathcal{H}_{D_N} = \mathbf{H}_{D_N}\right] \quad = \quad \sigma^2\Big(K_\psi(z_{jk} - z_{jl}) + \gamma(z_{jk})^T(\mathbf{F}_{D_N}^T \boldsymbol{\Sigma}_{DD}^{-1} \mathbf{F}_{D_N})^{-1}\gamma(z_{jl}) - \boldsymbol{\Sigma}_{z_{jk}D}^T \boldsymbol{\Sigma}_{DD}^{-1} \boldsymbol{\Sigma}_{z_{jl}D}\Big), \tag{II.23}$$

where $K_\psi(z - z')$ denotes the autocorrelation which only depends on the distance between z and z', z_{ji} denotes the $j-$th dimension of z_i with $1 \leq i \leq n$ and $1 \leq j \leq p$, and

$$\gamma(z) \quad = \quad \mathbf{F}(z) - \mathbf{F}_{D_N}^T \boldsymbol{\Sigma}_{DD}^{-1} \boldsymbol{\Sigma}_{zD}. \tag{II.24}$$

Moreover, given the observations $\mathbf{H}_{D_N}$ of the function H on the design of experiments D_N, the vectors $\mathcal{E}_1, \ldots, \mathcal{E}_p$ are assumed to be mutually independent. It is reasonable to assume also that the vectors are independent from the random variables $U_1, \ldots, U_n$ which describe the measurement errors.

It is worth noting that the model (II.19) is ordered according to the variables' order $(\widehat{H}_1(\mathbf{Z}), \cdots, \widehat{H}_p(\mathbf{Z}))$. The advantage is that the *kriging* variance matrix can be written as a block diagonal matrix, as now detailed. Given the observations $\mathbf{H}_{D_N}$, the new error term $\mathcal{V}_i = (V_{1i}, \ldots, \mathcal{V}_{pi})^T$ described in (II.20) follows a normal distribution:

$$\mathcal{V}_i(z) \,|\, \mathcal{H}_{D_N} = \mathbf{H}_{D_N} \quad \sim \quad \mathcal{N}\Big[\mathbf{0}, R + \mathrm{MSE}(z)\Big]. \tag{II.25}$$

Following the adapted model (II.19), it can then be proved that the distribution of $\boldsymbol{\mathcal{Y}}$, knowing the variables $\mathbf{Z}$ and the observations $\mathbf{H}_{D_N}$, is also a normal distribution:

$$\boldsymbol{\mathcal{Y}} \,|\, \mathbf{Z}, \mathcal{H}_{D_N} = \mathbf{H}_{D_N} \quad \sim \quad \mathcal{N}\left[\widehat{H}(\mathbf{Z}), \mathbf{R} + \mathrm{MSE}(\mathbf{Z})\right], \tag{II.26}$$

where

$$\mathbf{R} \quad = \quad \begin{pmatrix} R_{11} & & & & & \\ & \ddots & & & \mathbf{0} & \\ & & R_{11} & & & \\ & & & \ddots & & \\ & & & & R_{pp} & \\ & \mathbf{0} & & & \ddots & \\ & & & & & R_{pp} \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \Big\} \; n \text{ lines} \\ \\ \\ \\ \Big\} \; n \text{ lines} \end{matrix},$$

with R_{ii} the $i-$th diagonal component of the diagonal variance matrix R, and $\mathrm{MSE}(\mathbf{Z})$ is the

block diagonal matrix

$$\mathrm{MSE}(\mathbf{Z}) \;=\; \left(\begin{array}{ccc} \mathrm{MSE}_1(\mathbf{Z}) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathrm{MSE}_p(\mathbf{Z}) \end{array}\right) \begin{array}{l} \Big\} \; n \text{ lines} \\ \\ \Big\} \; n \text{ lines} \end{array}$$

composed with the variance matrices $\mathrm{MSE}_j(\mathbf{Z}) \in \mathcal{M}^{n\times n}$ described as

$$\mathrm{MSE}_j(\mathbf{Z}) \;=\; \mathbb{E}\left((\mathcal{H}_j(\mathbf{Z}) - \widehat{H}_j(\mathbf{Z}))^2 \,|\, \mathcal{H}_{D_N} = \mathbf{H}_{D_N}\right), \tag{II.27}$$

for $j = 1, \ldots, p$. The precise expression of the variance and covariance is given in (I.48) and (I.55).

The distribution in (II.26) leads to the following full conditional posterior distribution of the grouped random variables $\mathbf{X}$, which is different from (II.11):

$$\begin{aligned} &\pi_{\widehat{H}}(\mathbf{X} \,|\, m, C, \boldsymbol{\mathcal{Y}}, \mathbf{d}, \rho, \mathbf{H}_{D_N}) \;\propto\; \pi_{\widehat{H}}(\boldsymbol{\mathcal{Y}} \,|\, \mathbf{X}, m, C, \mathbf{d}, \rho, \mathbf{H}_{D_N}) \cdot \pi(\mathbf{X} \,|\, m, C) \\ \propto\;& |\mathbf{R} + \mathrm{MSE}(\mathbf{Z})|^{-\frac{1}{2}} \cdot \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}(X_i - m)^T C^{-1}(X_i - m) \right. \\ &\left. -\frac{1}{2}\Big(\big(\mathcal{Y}_1 - \widehat{H}(Z_1)\big)^T, \ldots, \big(\mathcal{Y}_n - \widehat{H}(Z_n)\big)^T\Big)\Big(\mathbf{R} + \mathrm{MSE}(\mathbf{Z})\Big)^{-1} \left(\begin{array}{c} \big(\mathcal{Y}_1 - \widehat{H}(Z_1)\big) \\ \vdots \\ \big(\mathcal{Y}_n - \widehat{H}(Z_n)\big) \end{array}\right) \right\}. \end{aligned} \tag{II.28}$$

The logarithm of this conditional distribution can then be written:

$$\begin{aligned} \log \pi_{\widehat{H}}(\mathbf{X} \,|\, \ldots) \;\overset{\exp}{\propto}\; & -\log|\mathbf{R} + \mathrm{MSE}(\mathbf{Z})| - \sum_{i=1}^{n}(X_i - m)' C^{-1}(X_i - m) \\ & -\Big(\big(\mathcal{Y}_1 - \widehat{H}(Z_1)\big)', \ldots, \big(\mathcal{Y}_n - \widehat{H}(Z_n)\big)'\Big)\Big(\mathbf{R} + \mathrm{MSE}(\mathbf{Z})\Big)^{-1} \left(\begin{array}{c} \big(\mathcal{Y}_1 - \widehat{H}(Z_1)\big) \\ \vdots \\ \big(\mathcal{Y}_n - \widehat{H}(Z_n)\big) \end{array}\right), \end{aligned} \tag{II.29}$$

which is preferred in our numerical calculations as it may avoid some numerical problems.

In this way, the "block diagonal" form of the matrix $\mathbf{R} + \mathrm{MSE}(\mathbf{Z})$ makes the calculation of the inverse of the covariance matrix easier.

Remark 15. *The "block diagonal" form of the matrix* $\mathbf{R} + MSE(\mathbf{Z})$ *is validated only under the assumption that the variance matrix* R *is diagonal, which means*

$$U_i^k \perp\!\!\!\perp U_i^j, \text{ for } k \neq j, \tag{II.30}$$

for $\forall i \in \{1, \ldots, n\}$.

Once again, the full conditional posterior distribution of $\mathbf{X}$ computed in the *rich man* version in (II.11) does not belong to any closed-form family of distribution. For this reason, it is necessary to rue a MCMC algorithm, e.g. the Metropolis-Hastings algorithm which will be presented in detail in the next chapter.

II.2 Prior calibration (elicitation) of the hyperparameters

II.2.1 Initial modeling (prior predictive distribution)

To calibrate the hyperparameters $\rho = (\mu, a, \Lambda, \nu)$, it would be helpful to compute the prior predictive distribution of X_i, which means marginalizing the joint distribution of $(m, C, \mathbf{X})$ by integrating the parameters $\theta = (m, C)$ out. Integrating over values of the parameters θ is natural in the context where the expert opinion is not attached to any model (see Bousquet, 2006, (9)). The prior predictive distribution makes sense for the statistician and it is more intuitive to experts as some features of this distribution can be relatively easily assessed from them (see Kadane and Wolfson, 1998, (49), Garthwaite et al., 2005, (31)).

Given values for the hyperparameters μ, a, Λ and ν, the prior predictive probability function of X_i can be calculated as

$$\begin{aligned}
\pi_{X_i}(x) &= \iint \pi_{X_i}(x|m,C)\,\pi(m|C)\,\pi(C)\,dC\,dm \\
&\propto \iint |C|^{-1/2} \exp\Big[-\frac{1}{2}(x-m)^T C^{-1}(x-m)\Big] \\
&\quad \cdot |\frac{C}{a}|^{-1/2} \exp\Big[-\frac{1}{2}(m-\mu)^T (\frac{C}{a})^{-1}(m-\mu)\Big] \\
&\quad \cdot |C|^{-\frac{\nu+q+1}{2}} \exp\Big[-\frac{1}{2}\operatorname{Tr}(\Lambda \cdot C^{-1})\Big]\,dC\,dm \\
&\propto \iint |C|^{-\frac{\nu+q+3}{2}} \exp\Big\{-\frac{1}{2}\Big[(x-m)^T C^{-1}(x-m) + a(m-\mu)^T C^{-1}(m-\mu) + \operatorname{Tr}(\Lambda \cdot C^{-1})\Big]\Big\}\,dC\,dm.
\end{aligned} \tag{II.31}$$

Considering the inner term in the exponential function, we can prove that

$$\begin{aligned}
&(x-m)^T C^{-1}(x-m) + a(m-\mu)^T C^{-1}(m-\mu) \\
&= \left(m - \frac{x}{1+a} - \frac{a}{1+a}\mu\right)^T \left(\frac{C}{1+a}\right)^{-1} \left(m - \frac{x}{1+a} - \frac{a}{1+a}\mu\right) + (x-\mu)^T \left(\frac{1+a}{a}\,C\right)^{-1} (x-\mu).
\end{aligned} \tag{II.32}$$

Plugging this term in (II.31) leads to the following development:

$$\begin{aligned}
\pi_{X_i}(x) &\propto \int |C|^{-\frac{\nu+q+2}{2}} \exp\Big\{-\frac{1}{2}\Big[(x-\mu)^T\left(\frac{1+a}{a}C\right)^{-1}(x-\mu)+\mathrm{Tr}(\Lambda\cdot C^{-1})\Big]\Big\}\,dC \\
&\quad\cdot\int |\frac{C}{1+a}|^{-1/2}\exp\Big[-\frac{1}{2}\Big(m-\frac{x}{1+a}-\frac{a}{1+a}\mu\Big)^T(\frac{C}{1+a})^{-1}\Big(m-\frac{x}{1+a}-\frac{a}{1+a}\mu\Big)\Big]\,dm \\
&\propto |\Lambda+\frac{a}{a+1}(x-\mu)(x-\mu)^T|^{-\frac{\nu+1}{2}} \\
&\quad\underbrace{\int |\Lambda+\tfrac{a}{a+1}(x-\mu)(x-\mu)^T|^{\frac{\nu+1}{2}}|C|^{-\frac{\nu+q+2}{2}}\exp\Big[-\tfrac{1}{2}\mathrm{Tr}\Big(\Big(\Lambda+\tfrac{a}{a+1}(x-\mu)(x-\mu)^T\Big)C^{-1}\Big)\Big]\,dC}_{\propto 1}, \\
&\propto |\Lambda+\frac{a}{a+1}(x-\mu)(x-\mu)^T|^{-\frac{\nu+1}{2}} \qquad \text{(II.33)} \\
&= |\Lambda|^{\frac{-\nu+1}{2}}\left[1+(x-\mu)^T\left(\frac{a+1}{a}\Lambda\right)^{-1}(x-\mu)\right]^{\frac{-\nu+1}{2}}. \qquad \text{(II.34)}
\end{aligned}$$

Thus, the density of X_i can be written as:

$$\pi_{X_i}(x) \propto |\frac{a+1}{a(\nu+1-q)}\Lambda|^{-\frac{1}{2}}\left[1+(x-\mu)^T\left(\frac{a+1}{a(\nu+1-q)}\Lambda\right)^{-1}(x-\mu)\right]^{-\frac{(\nu+1-q)+q}{2}} \qquad \text{(II,35)}$$

which indicates the following *multivariate Student distribution*:

$$X_i \sim \mathrm{St}_q\Big(\mu,\frac{a+1}{a(\nu+1-q)}\Lambda,\nu+1-q\Big),\ (1\le i\le n), \qquad \text{(II.36)}$$

with

$$\begin{aligned}
\mathbb{E}[X_i] &= \mu, \qquad \text{(II.37)} \\
\mathrm{Var}[X_i] &= \frac{a+1}{a(\nu-1-q)}\Lambda. \qquad \text{(II.38)}
\end{aligned}$$

In (II.33), we recognize the Inverse-Wishart distribution:

$$\mathcal{IW}\left[\Lambda+\frac{a}{a+1}(x-\mu)(x-\mu)^T,\nu+1\right], \qquad \text{(II.39)}$$

using the following identity:

$$\Gamma_q\left(\frac{\nu+1}{2}\right)/\Gamma_q\left(\frac{\nu}{2}\right) = \Gamma\left(\frac{\nu+1}{2}\right)\Big/\Gamma\left(\frac{\nu+1-q}{2}\right) \qquad \text{(II.40)}$$

where $\Gamma_q(\cdot)$ denotes the multivariate gamma function of order q and $\Gamma(\cdot)$ denotes the gamma function under the following recursive relationship:

$$\Gamma_q(a) = \pi^{\frac{q(q-1)}{4}}\prod_{j=1}^{q}\Gamma\left(a+\frac{1-j}{2}\right). \qquad \text{(II.41)}$$

Remark 16. *The multivariate Student distribution described in* (II.36) *tends to a Gaussian distribution as* $\nu \to \infty$*:*

$$\sqrt{\frac{a(\nu-q-1)}{a+1}}X_i \quad \overset{\mathcal{L}}{\rightsquigarrow} \quad \mathcal{N}_q\Big(\mu, \Lambda\Big). \tag{II.42}$$

The distribution (II.36) is the prior predictive distribution of X_i, which could be quite useful in practice. As mentioned at the beginning of the section, expert information is usually expressed on intuitive variables, independently of any statistical parameterization. This information can often be assimilated to prior predictive features rather than prior parametrical features. The hyperparameters μ, a, Λ and ν can thus be more easily elicited from the expert.

II.2.2 Calibration for conjugate *priors*

Since the Gaussian and the Inverse-Wishart distributions belong to the exponential family of distributions. Conjugate prior distributions can be used for their parameters. This section addresses the issue of eliciting of the prior hyperparameters from the prior predictive representation, following the ideas promoted by Kadane and Wolfson (1998, (49)).

Analysis: How to choose the hyperparameters a and μ?

In formula (II.12) which gives the full conditional posterior distribution of m, the prior mean μ can be chosen to be m_{Exp} according to expert knowledge. Moreover, the hyperparameter a can be regarded as the size of a *virtual sample* to be adjusted with respect to our knowledge or belief, while n is the fixed size of the observed sample.

When a is close to 0, the impact of the prior distribution disappears; when a is large, the impact of the data disappears. A default choice is $a = 1$, which means that the prior information is as important as the information brought by *one* data. The advantage of this standard choice is that it does not bring an excessive importance on prior information with respect to data information.

Analysis: how to choose the Inverse-Wishart hyperparameters Λ and ν ?

The other two hyperparameters Λ and ν, known as two components of the prior Inverse-Wishart ($\mathcal{IW}$) distribution of C, are more difficult to interpret. First, we choose the inverse scale matrix Λ under the following form:

$$\Lambda \quad = \quad t \cdot C_{\text{Exp}} \tag{II.43}$$

where C_{Exp} denotes the prior variance matrix with respect to the expert opinion and t is a related hyperparameter to be specified. This formulation is natural since for any $C \sim \mathcal{IW}_q(\Lambda, \nu)$, we have

$$\mathbb{E}(C) \quad = \quad \frac{1}{\nu-q-1}\Lambda. \tag{II.44}$$

By fixing $t = \nu - q - 1$, we get from (II.43):

$$\mathbb{E}(C) \;=\; \frac{t}{\nu - q - 1} \cdot C_{\text{Exp}} \;=\; C_{\text{Exp}}. \tag{II.45}$$

Thus, a natural prior choice for ν is:

$$\nu \;=\; t + q + 1. \tag{II.46}$$

In what follows, we only need to calibrate the hyperparameter t. We choose to analyze the full conditional posterior distribution of C, which is an Inverse-Wishart distribution, given in (II.13). Note that the inverse scale matrix contains three terms :

$$\Lambda, \qquad \sum_{i=1}^{n}(m - X_i)(m - X_i)^T \;\text{ and }\; a(m-\mu)(m-\mu)^T.$$

The second and the third terms correspond to the total squared deviation within the sample $\{X_i\}$ of size n and the *virtual* sample $\{m\}$ of size a. Considering an unbiased estimator of the sample variance:

$$\widehat{C} \;=\; \frac{1}{n}\sum_{i=1}^{n}(m - X_i)(m - X_i)^T, \tag{II.47}$$

the total squared derivation in $\{X_i\}$ can be written $n\,\widehat{C}$. Moreover, assuming

$$m \,|\, \widehat{C} \;\sim\; \mathcal{N}\left(\mu, \widehat{C}/a\right) \tag{II.48}$$

leads to

$$\frac{\widehat{C}}{a} \;\simeq\; \frac{1}{a}\sum_{i=1}^{a}(m-\mu)(m-\mu)^T, \tag{II.49}$$

which provides a measure of the total squared derivation of $\{m\}$ as $a \cdot \widehat{C}/a = \widehat{C}$. The full conditional posterior distribution (II.13) of C can then be written as follows:

$$C \,|\, m, \mathbf{X}, \mathbf{Y}, \mathbf{d}, \rho \;\sim\; \mathcal{IW}\Big(t\,C_{\text{Exp}} + (n+1)\,\widehat{C},\; \nu + n + 1\Big). \tag{II.50}$$

Under the assumption that $\nu = q + t + 1$, the posterior mean of C equals finally

$$\mathbb{E}(C \,|\, m, \mathbf{X}, \mathbf{Y}, \mathbf{d}, \rho) \;=\; \frac{t\,C_{\text{Exp}} + (n+1)\,\widehat{C}}{t + n + 1} \tag{II.51}$$

$$\;=\; \frac{t}{t+n+1} \cdot C_{\text{Exp}} + \frac{n+1}{t+n+1} \cdot \widehat{C}. \tag{II.52}$$

We obtain an elegant expression which provides us the possibility to tune the importance (weight) of the prior choice by choosing a proper t, which is homogeneous to $n+1$. Recalling that we have interpreted a as the size of a *virtual* sample. Formula (II.52) gives us a good reason to take:

$$t \;=\; a + 1 \;\;\Rightarrow\;\; t = 2 \text{ is a default choice (with } a = 1). \tag{II.53}$$

II.3 An alternative view: Jeffreys non informative prior

Unfortunately, sometimes neither prior information nor expert knowledge is available. In this case, a non informative prior distribution can be chosen. In this section, we aim at computing the full conditional posterior distributions in this case.

II.3.1 General introduction

The Jeffreys non informative prior distribution $\pi^J(\theta)$ (see Kass and Wasserman, 1996, (51)), which remains invariant under reparameterization, is a standard non informative prior distribution. It is as follows:

$$\pi^J(\theta) \;=\; \pi^J(m)\,\pi^J(C) \;=\; \frac{\mathbf{I}_{\Omega_m}(m)}{\mathrm{Vol}(\Omega_m)} \cdot \frac{\Delta_C}{|C|^{\frac{q+2}{2}}} \mathbf{I}_{\Omega_C}(C), \tag{II.54}$$

with

$$\Delta_C \;=\; \left(\int_{\Omega_C} \frac{1}{|C|^{\frac{q+2}{2}}} dC \right)^{-1}. \tag{II.55}$$

In our case study, as the meta-model is defined on a compact set, the simulated missing data $\mathbf{X}$ should be in this compact set. Moreover, the parameters m and C are also restricted to be in compact sets Ω_m and Ω_C as expressed in (II.54). If Ω_m and Ω_C are strictly included in $\mathbb{R}_+^q$ and $\mathbb{R}_+^{q\times q}$, the Jeffreys non informative distribution is proper in our case.

II.3.2 Calculation of the full conditional posterior distributions

This computation is restricted to the *poor man* version as it is more complicated. According to Bayes' formula, the joint distribution $\pi^J(m, C, \mathbf{X} \,|\, \boldsymbol{\mathcal{Y}}, \mathbf{d}, \mathbf{H}_{D_N})$ can be written as:

$$\begin{aligned}
&\pi^J(m, C, \mathbf{X} \,|\, \boldsymbol{\mathcal{Y}}, \mathbf{d}, \mathbf{H}_{D_N}) \;\propto\; \pi_{\widehat{H}}(\boldsymbol{\mathcal{Y}} \,|\, \mathbf{X}, m, C, \mathbf{d}, \mathbf{H}_{D_N}) \cdot \pi(\mathbf{X} \,|\, m, C) \cdot \pi^J(m, C) \\
\propto\;& |\mathbf{R} + \mathrm{MSE}(\mathbf{Z})|^{-\frac{1}{2}} \cdot \exp\Bigg\{ -\frac{1}{2} \sum_{i=1}^{n} (X_i - m)' C^{-1} (X_i - m) \\
&-\frac{1}{2} \Big(\big(\mathcal{Y}_1 - \widehat{H}(Z_1)\big)', \ldots, \big(\mathcal{Y}_n - \widehat{H}(Z_n)\big)' \Big) \Big(\mathbf{R} + \mathrm{MSE}(\mathbf{Z})\Big)^{-1} \begin{pmatrix} \big(\mathcal{Y}_1 - \widehat{H}(Z_1)\big) \\ \vdots \\ \big(\mathcal{Y}_n - \widehat{H}(Z_n)\big) \end{pmatrix} \Bigg\} \\
&\cdot |C|^{-\frac{n}{2}} \cdot \exp\Big[-\frac{1}{2} \sum_{i=1}^{n} (X_i - m)^T C^{-1} (X_i - m) \Big] \cdot \mathbf{I}_{\Omega_m}(m) \cdot \mathbf{I}_{\Omega_C}(C) \cdot |C|^{-\frac{q+2}{2}}.
\end{aligned}$$

Up to an additive constant, we have

$$\begin{aligned}
&\pi^J(m, C, \mathbf{X} \,|\, \boldsymbol{\mathcal{Y}}, \mathbf{d}, \mathbf{H}_{D_N}) \\
\propto\ & |C|^{-\frac{n+q+2}{2}} \cdot |\mathbf{R} + \mathrm{MSE}(\mathbf{Z})|^{-\frac{1}{2}} \cdot \exp\Bigg\{ -\frac{1}{2}\sum_{i=1}^{n}(m - X_i)^T C^{-1}(m - X_i) \\
& -\frac{1}{2}\Big(\Big(\mathcal{Y}_1 - \widehat{H}(Z_1)\Big)', \ldots, \Big(\mathcal{Y}_n - \widehat{H}(Z_n)\Big)'\Big)\Big(\mathbf{R} + \mathrm{MSE}(\mathbf{Z})\Big)^{-1} \begin{pmatrix} \Big(\mathcal{Y}_1 - \widehat{H}(Z_1)\Big) \\ \vdots \\ \Big(\mathcal{Y}_n - \widehat{H}(Z_n)\Big) \end{pmatrix} \Bigg\}.
\end{aligned} \tag{II.56}$$

Denoting $\overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^n X_i$, the full conditional posterior distribution of m verifies

$$\pi^J(m \,|\, C, \mathbf{X}, \boldsymbol{\mathcal{Y}}, \mathbf{d}, \mathbf{H}_{D_N}) \ \propto\ \mathbf{I}_{\Omega_m} \exp\Big[-\frac{1}{2}(m - \overline{\mathbf{X}}_n)' \left(\frac{C}{n}\right)^{-1} (m - \overline{\mathbf{X}}_n)\Big]. \tag{II.57}$$

This is a normal distribution truncated on Ω_m:

$$m \,|\, C, \mathbf{X}, \boldsymbol{\mathcal{Y}}, \mathbf{d}, \mathbf{H}_{D_N} \ \sim\ \mathbf{I}_{\Omega_m} \cdot \mathcal{N}\left(\overline{\mathbf{X}}_n, \frac{C}{n}\right). \tag{II.58}$$

Similarly, the full conditional posterior distribution of C verifies

$$\pi^J(C \,|\, m, \mathbf{X}, \boldsymbol{\mathcal{Y}}, \mathbf{d}, \mathbf{H}_{D_N}) \ \propto\ \mathbf{I}_{\Omega_C} |C|^{-\frac{n+q+2}{2}} \exp\Big[-\frac{1}{2}\mathrm{Tr}\left(n\,(m - \overline{X}_n)(m - \overline{X}_n)' \cdot C^{-1}\right)\Big]. \tag{II.59}$$

Thus, it is an Inverse-Wishart distribution truncated on Ω_C:

$$C \,|\, m, \mathbf{X}, \boldsymbol{\mathcal{Y}}, \mathbf{d}, \mathbf{H}_{D_N} \ \sim\ \mathbf{I}_{\Omega_C} \cdot \mathcal{IW}\Big(n\,(m - \overline{X}_n)(m - \overline{X}_n)', n+1\Big). \tag{II.60}$$

Moreover, the full conditional posterior distribution of the missing data $\mathbf{X}$ knowing the current parameters θ, the observations $\boldsymbol{\mathcal{Y}}, \mathbf{d}$ and the evaluations $\mathbf{H}_D$ is given by (II.28). It requires a numerical method, e.g. the MCMC algorithm, for the simulation of those missing data.

III

MCMC method adapted to inverse problems

Contents

This chapter describes the Markov chain Monte Carlo (MCMC) method, namely the Metropolis-Hastings-within-Gibbs algorithm (or Hybrid MCMC algorithm) we used in this thesis. First, we recall the general definition of a MCMC method.

Definition 3. *((90)) A Markov chain Monte Carlo method for the simulation of a distribution f is any method producing an ergodic Markov chain $(X^{[t]})$ whose stationary distribution is f.*

In other words, a MCMC algorithm generates a Markov chain $(X^{[t]})$ with the help of a chosen transition kernel from an arbitrary starting point $x^{[0]}$, converging towards the target distribution f.

III.1 Metropolis-Hastings-within-Gibbs algorithm (Hybrid MCMC algorithm)

III.1.1 Target Gibbs sampler

Gibbs sampling is a MCMC algorithm which draws each unknown quantity (the parameters m, C and the unobserved data $\mathbf{X}$) in the present context iteratively from the full conditional posterior distributions given the current values of the other quantities. Unfortunately, the distribution of $\mathbf{X}$ knowing (m, C) does not belong to any known family of distributions, and that is why a numerical method such as the Metropolis-Hastings (MH) algorithm (see for instance Tierney, 1995, (109)) is necessary. This type of Gibbs sampler combined with a MH step is named the *Metropolis-Hastings-within-Gibbs algorithm* (or *Hybrid MCMC algorithm*).

The convergence of the simulated samples $\left(m^{[r]}, C^{[r]}, \mathbf{X}^{[r]}\right)$ towards the stationary joint distribution $\pi_H\left(m, C, \mathbf{X} \,|\, \mathbf{Y}, \mathbf{d}, \rho\right)$[1] can be verified under some regularity conditions (cf. Section III.2). Moreover, each variable will converge to its own marginal posterior distribution, i.e. $\pi(m \,|\, \mathbf{Y}, \mathbf{d})$, $\pi(C \,|\, \mathbf{Y}, \mathbf{d})$ and $\pi(\mathbf{X} \,|\, \mathbf{Y}, \mathbf{d})$. Thus, thanks to MCMC, Bayesian inference avoids some numerical difficulties related to the missing data structure arising in frequentist inference, although there is still the price of numerical simulations to be paid.

With the prior choices described in Chapter II by (II.4) and (II.5), the calculated posterior distributions (II.11), (II.12), (II.13) and (II.28) lead to the following algorithm:

Gibbs sampler (at the $(r+1)$-th iteration)

`Given` $\left(m^{[r]}, C^{[r]}, \mathbf{X}^{[r]}\right)$ `for` $r = 0, 1, 2, \ldots,$ `generate:`

1. $C^{[r+1]} | \cdots \sim \mathcal{IW}\Big(\Lambda + \sum_{i=1}^{n}(m^{[r]} - X_i^{[r]})(m^{[r]} - X_i^{[r]})' + a(m^{[r]} - \mu)(m^{[r]} - \mu)',\ \nu + n + 1\Big)$.
2. $m^{[r+1]} | \cdots \sim \mathcal{N}\Big(\frac{a}{n+a}\mu + \frac{n}{n+a}\overline{\mathbf{X}_n^{[r]}},\ \frac{C^{[r+1]}}{n+a}\Big)$ where $\overline{\mathbf{X}_n^{[r]}}$ `denotes the empirical mean of the` n `vectors` $X_i^{[r]}, i = 1, \ldots, n$.

[1] Or $\pi_{\widehat{H}}\left(m, C, \mathbf{X} \,|\, \boldsymbol{\mathcal{Y}}, \mathbf{d}, \mathbf{H}_D, \rho\right)$ for the more general *poor man* version

3. `In the` *rich man* `version,`
$\mathbf{X}^{[r+1]}|\cdots \propto \exp -\frac{1}{2}\sum_{i=1}^{n}\left[(X_i^{[r+1]}-m^{[r+1]})^T\left[C^{[r+1]}\right]^{-1}(X_i^{[r+1]}-m^{[r+1]})+\left(Y_i-H_i^{[r+1]}\right)^T R^{-1}\left(Y_i-H_i^{[r+1]}\right)\right]\},$
`where` $H_i^{[r+1]}=H(X_i^{[r+1]},d)$. `While in the` *poor man* `version,`

$$\mathbf{X}^{[r+1]}|\cdots \propto |\mathbf{R}+\mathrm{MSE}^{[r+1]}|^{-\frac{1}{2}}\cdot\exp\left\{-\tfrac{1}{2}\textstyle\sum_{i=1}^{n}(X_i^{[r+1]}-m^{[r+1]})^T\left[C^{[r+1]}\right]^{-1}(X_i^{[r+1]}-m^{[r+1]})-\right.$$
$$\left.\tfrac{1}{2}\left(\left(\mathcal{Y}_1-\widehat{H}_{N,1}^{[r+1]}\right)^T,\ldots,\left(\mathcal{Y}_n-\widehat{H}_{N,n}^{[r+1]}\right)^T\right)\left(\mathbf{R}+\mathrm{MSE}^{[r+1]}\right)^{-1}\begin{pmatrix}\mathcal{Y}_1-\widehat{H}_{N,1}^{[r+1]}\\ \vdots\\ \mathcal{Y}_n-\widehat{H}_{N,n}^{[r+1]}\end{pmatrix}\right\},$$

`where` $\widehat{H}_{N,i}^{[r+1]}=\widehat{H}_N(X_i^{[r+1]},d)$ `and` $\mathtt{MSE}^{[r+1]}=\mathtt{MSE}(\mathbf{X}^{[r+1]},d)$.

In the *poor man* version, $\mathrm{MSE}(\mathbf{X}^{[r+1]},d)$ is a block diagonal matrix of size $np\times np$ and $\mathbf{R}$ is a diagonal matrix of the same size, both of which have been described in Section II.1.2. Since the distribution of $\mathbf{X}^{[r+1]}$ is not closed-form, numerical methods, typically the Metropolis-Hastings (MH) algorithm, are required. Suppose that l iterations of the MH algorithm are applied at each iteration of the Gibbs sampler. In this thesis, l is chosen equal to 1 (referring to Section III.1.2). The two versions of the MH algorithm are now presented.

III.1.2 Inner Metropolis-Hastings algorithm (the *rich man* version)

The MH algorithm is based on an *instrumental* distribution, which causes useless simulations (*rejections*) when it is badly chosen.

Metropolis-Hastings algorithm (the *rich man* version, at the $(r+1)$-th iteration)
`Given` $\theta^{[r+1]}=(m^{[r+1]},C^{[r+1]})$ `and` $\mathbf{X}^{[r]}=\left(X_1^{[r]},\ldots,X_n^{[r]}\right)^T$, `for each sample` $X_i^{[r]}$, $1\leq i\leq n$:

1. `Let` $X_{i,0}=X_i^{[r]}$
2. `For` $s=1,\ldots,l$, `updating` $X_i^{[r]}$:
   - `Generate`
   $$\widetilde{X}_{i,s} \sim J(\cdot\mid m^{[r+1]},C^{[r+1]},X_{i,s-1}) \tag{III.1}$$
   `where` J `is the proposal (`*instrumental*`) distribution.`
   - `Let`
   $$\alpha\left(X_{i,s-1},\widetilde{X}_{i,s}\right) = \min\left\{\frac{\pi_H(\widetilde{X}_{i,s}\mid Y_i,d_i,\theta^{[r+1]},\rho)}{\pi_H(X_{i,s-1}\mid Y_i,d_i,\theta^{[r+1]},\rho)}\frac{J(X_{i,s-1}|\widetilde{X}_{i,s},\theta^{[r+1]})}{J(\widetilde{X}_{i,s}|X_{i,s-1},\theta^{[r+1]})},1\right\}, \tag{III.2}$$

take

$$X_{i,s} \;=\; \begin{cases} \widetilde{X}_{i,s}, & \text{with probability } \alpha\left(X_{i,s-1}, \widetilde{X}_{i,s}\right); \\ X_{i,s-1}, & \text{otherwise.} \end{cases}$$

More precisely, generate $u \sim \mathcal{U}_{[0,1]}$, then

$$X_{i,s} \;=\; \begin{cases} \widetilde{X}_{i,s}, & \text{if } u < \alpha\left(X_{i,s-1}, \widetilde{X}_{i,s}\right); \\ X_{i,s-1}, & \text{otherwise.} \end{cases}$$

3. Let $X_i^{[r+1]} = X_{i,l}$

A specific Hybrid MCMC algorithm, which uses one MH step within an iteration of Gibbs sampling, was proposed in Muller (1991, (73)). By assuming $l = 1$, the modified MH algorithm is as follows.

Modified Metropolis-Hastings algorithm (the *rich man* version, at the $(r+1)$-th iteration)

Given $\left(m^{[r+1]}, C^{[r+1]}, \mathbf{X}^{[r]}\right)$, for each sample $X_i^{[r]}, 1 \le i \le n$:

1. Simulate

$$\widetilde{X}_i \;\sim\; J(\cdot \mid m^{[r+1]}, C^{[r+1]}, X_i^{[r]}) \;=\; J(\cdot \mid \theta^{[r+1]}, X_i^{[r]}) \tag{III.3}$$

where J is the instrumental distribution.

2. Take

$$X_i^{[r+1]} \;=\; \begin{cases} \widetilde{X}_i, & \text{with probability } \alpha\left(X_i^{[r]}, \widetilde{X}_i\right); \\ X_i^{[r]}, & \text{with probability } 1-\alpha\left(X_i^{[r]}, \widetilde{X}_i\right), \end{cases}$$

where

$$\alpha\left(X_i^{[r]}, \widetilde{X}_i\right) \;=\; \min\left\{ \frac{\pi_H(\widetilde{X}_i \mid Y_i, d_i, \theta^{[r+1]}, \rho)}{\pi_H(X_i^{[r]} \mid Y_i, d_i, \theta^{[r+1]}, \rho)} \, \frac{J(X_i^{[r]} | \widetilde{X}_i, \theta^{[r+1]})}{J(\widetilde{X}_i | X_i^{[r]}, \theta^{[r+1]})}, 1 \right\}. \tag{III.4}$$

As indicated in (90) (pp. 393-396), the arguments for this more rapid MH version are twofold :

1. its stationary distribution remains $\pi_H(m, C, \mathbf{X} \mid \mathbf{Y}, \mathbf{d}, \rho)$;
2. even without convergence at the MH step, Gibbs sampling also leads to an approximation of its target distribution.

In fact, the quality of the simulation at each iteration has no great effect on the validation of the iterative algorithm. Providing a more "precise" approximation of $\mathbf{X}^{[r+1]}$ in MH steps does not necessarily lead to a better approximation of the joint distribution $\pi_H(m, C, \mathbf{X} \mid \mathbf{Y}, \mathbf{d}, \rho)$.

Moreover, the replacement of the target full conditional posterior distribution $\pi_H(X_i^{[r+1]}|\dots)$ by the instrumental distribution $J(X_i^{[r+1]}|\dots)$ may even be beneficial for the speed of excursion of the chain. (See also Chen and Schmeiser, 1998, (16))

The choice of the instrumental distribution J in a MH algorithm is a critical issue. Although the convergence of the algorithm is ensured under some generic assumptions on J (see Section III.2), the chain can still suffer from a very slow rate of convergence, depending strongly on J. An efficient instrumental distribution J will sample candidates in regions where the target distribution $\pi_H(X_i^{[r+1]}|\dots)$ is high.

Based on the modified MH algorithm, several possible instrumental distributions J_1, J_2 and J_3 are considered and compared (see also Kuhn and Lavielle, 2004, (56)).

1. J could be the normal distribution $\mathcal{N}\left(m^{[r+1]}, C^{[r+1]}\right)$, which is *independent* from the previous value $X_i^{[r]}$:

$$J\left(\widetilde{X}_i \,|\, \theta^{[r+1]}, X_i^{[r]}\right) \;=\; J\left(\widetilde{X}_i \,|\, \theta^{[r+1]}\right),\ 1 \le i \le n. \tag{III.5}$$

The ratio term in $\alpha\left(X_i^{[r]}, \widetilde{X}_i\right)$ described by (III.4) can be simplified as follows:

$$\begin{aligned} & \frac{\pi_H\left(\widetilde{X}_i \mid Y_i, d_i, \theta^{[r+1]}, \rho\right)}{\pi_H\left(X_i^{[r]} \mid Y_i, d_i, \theta^{[r+1]}, \rho\right)} \frac{J\left(X_i^{[r]}|\widetilde{X}_i, \theta^{[r+1]}\right)}{J\left(\widetilde{X}_i|X_i^{[r]}, \theta^{[r+1]}\right)} \\ = \; & \frac{J\left(\widetilde{X}_i \,|\, \theta^{[r+1]}\right)\, \pi_H\left(Y_i \mid \widetilde{X}_i, d_i, \rho\right)}{J\left(X_i^{[r]} \,|\, \theta^{[r+1]}\right)\, \pi_H\left(Y_i \mid X_i^{[r]}, d_i, \rho\right)} \frac{J\left(X_i^{[r]}|\theta^{[r+1]}\right)}{J\left(\widetilde{X}_i|\theta^{[r+1]}\right)} \\ = \; & \frac{\pi_H\left(Y_i \mid \widetilde{X}_i, d_i, \rho\right)}{\pi_H\left(Y_i \mid X_i^{[r]}, d_i, \rho\right)} \end{aligned}$$

where

$$\pi_H\left(y \mid X_i, d_i, \rho\right) \;\propto\; \exp\left\{-\frac{1}{2}\left[\left(y - H(X_i, d_i)\right)^T R^{-1} (y - H(X_i, d_i))\right]\right\}, \tag{III.6}$$

as according to the model defined in (I.4) and the normal assumption on U_i, we have

$$Y_i \mid X_i, d_i, \rho \;\sim\; \mathcal{N}\left(H(X_i, d_i), R\right),\ 1 \le i \le n. \tag{III.7}$$

2. J could be the normal distribution $\mathcal{N}_q\left(X_i^{[r]}, \kappa C^{[r+1]}\right)$, with parameter κ to be fixed, which leads to the following balance relationship:

$$J\left(\widetilde{X}_i \mid X_i^{[r]}, \theta^{[r+1]}\right) = J\left(X_i^{[r]} \mid \widetilde{X}_i, \theta^{[r+1]}\right), \tag{III.8}$$

thanks to the *symmetry* property of the normal distribution. The ratio term in the

expression of $\alpha\left(X_i^{[r]}, \widetilde{X}_i\right)$ with respect to J can be simplified as follows:

$$\begin{aligned}
& \frac{\pi_H\left(\widetilde{X}_i \mid Y_i, d_i, \theta^{[r+1]}, \rho\right)}{\pi_H\left(X_i^{[r]} \mid Y_i, d_i, \theta^{[r+1]}, \rho\right)} \frac{J\left(X_i^{[r]} | \widetilde{X}_i, \theta^{[r+1]}\right)}{J\left(\widetilde{X}_i | X_i^{[r]}, \theta^{[r+1]}\right)} \\
= & \frac{\pi_H\left(\widetilde{X}_i \mid Y_i, d_i, \theta^{[r+1]}, \rho\right)}{\pi_H\left(X_i^{[r]} \mid Y_i, d_i, \theta^{[r+1]}, \rho\right)} \cdot 1 \\
= & \frac{\pi_H\left(\widetilde{X}_i, Y_i \mid d_i, \theta^{[r+1]}, \rho\right)}{\pi_H\left(X_i^{[r]}, Y_i \mid d_i, \theta^{[r+1]}, \rho\right)},
\end{aligned}$$

where the joint distribution of (X_i, Y_i) knowing $(d_i, \theta^{[r+1]}, \rho)$ can be calculated as follows, by applying Bayes' formula:

$$\begin{aligned}
& \pi_H\left(x, y \mid d_i, \theta^{[r+1]}, \rho\right) = \pi\left(x \mid \theta^{[r+1]}\right) \pi_H\left(y \mid x, d_i, \rho\right) \\
\propto & \exp\left\{-\frac{1}{2}\left[(x - m^{[r+1]})'(C^{[r+1]})^{-1}(x - m^{[r+1]}) + (y - H(x, d_i))'R^{-1}(y - H(x, d_i))\right]\right\}.
\end{aligned} \tag{III.9}$$

3. J could be a succession of q uni dimensional Gaussian random walks $\mathcal{N}\left(X_i^{[r]}(l), \kappa C^{[r+1]}(l, l)\right)$ with $l = 1, \ldots, q$, which means that each component of $\widetilde{X}_i$ is to be updated respectively. In more details, after the construction of the l-th component of $\widetilde{X}_i$, the current candidate is:

$$\widetilde{X}_i = \left(X_i(1), \ldots, X_i(l-1), \widetilde{X}_i(l), X_i^{[r]}(l+1), \ldots, X_i^{[r]}(q)\right)^T, \tag{III.10}$$

where the $(q - l)$ positions $(X_i^{[r]}(l+1), \ldots, X_i^{[r]}(q))$ have not yet been updated.

Assuming $\kappa = 2$, Figure III.1 provides us an illustration of J_1, J_2 and J_3 in an uni dimensional case. Remark that when X_i is uni dimensional, J_2 and J_3 are the same. J_1 is centered on $m^{[r+1]}$, which is relatively stable as it is simulated according to a normal distribution (see (II.12)) out of the MH step. J_2 is centered on the previous simulated value $X^{[r]}$, which is more variable and thus a larger variance $\kappa C^{[r+1]}$ is assumed to attenuate this variation.

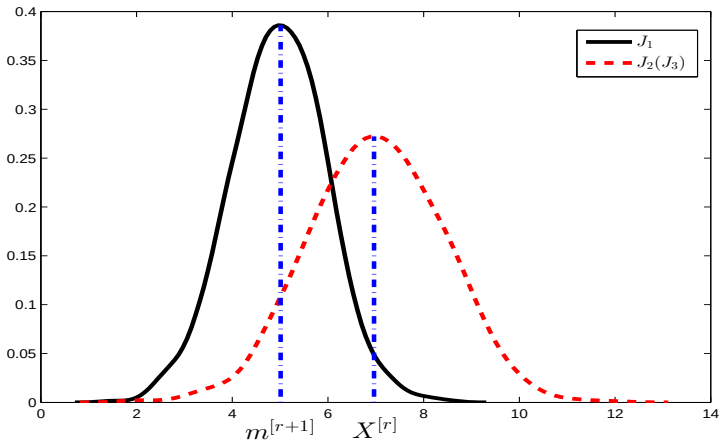


Figure III.1: Examples of instrumental distributions

Remark 17. *In Barbillon (2010, (3)), l_1 iterations with proposal 1, l_2 iterations with proposal 2 and l_3 iterations with proposal 3, such that $l_1+l_2+l_3=l$ of MH steps have been carried out for the simulation of the missing data $\mathbf{X}$. However, choosing l_1, l_2 and l_3 is a difficult issue, which is highly related to the numerical model. In Section III.3, we will discuss the choice of proposal distributions with two examples.*

III.1.3 Inner Metropolis-Hastings algorithm (the *poor man* version)

At $(r+1)$-th iteration, the missing data $\mathbf{X}^{[r+1]}$ is to be updated sequentially, as follows.

Modified Metropolis-Hastings algorithm (the *poor man* version, at the $(r+1)$-th iteration)

`Given` $\left(m^{[r+1]}, C^{[r+1]}, \mathbf{X}^{[r]}\right)$, `for each sample` $X_i^{[r]}$, $1 \leq i \leq n$:

1. Generate
$$\widetilde{X}_i \sim J(\cdot \mid m^{[r+1]}, C^{[r+1]}, X_i^{[r]}) = J(\cdot \mid \theta^{[r+1]}, X_i^{[r]}) \tag{III.11}$$
`where` J `is the instrumental distribution.`

2. `Let`
$$\alpha\left(X_i^{[r]}, \widetilde{X}_i\right) = \min\left\{\frac{\pi_{\widehat{H}}(\widetilde{\mathbf{X}}_i \mid \mathcal{Y}, \mathbf{d}, \theta^{[r+1]}, \rho, \mathbf{H}_D)}{\pi_{\widehat{H}}(\mathbf{X}_i^{[r]} \mid \mathcal{Y}, \mathbf{d}, \theta^{[r+1]}, \rho, \mathbf{H}_D)} \frac{J(X_i^{[r]}|\widetilde{X}_i, \theta^{[r+1]})}{J(\widetilde{X}_i|X_i^{[r]}, \theta^{[r+1]})}, 1\right\}, \tag{III.12}$$
`where`
$$\widetilde{\mathbf{X}}_i = \left(X_1^{[r+1]}, \ldots, X_{i-1}^{[r+1]}, \widetilde{X}_i, X_{i+1}^{[r]}, \ldots, X_n^{[r]}\right)^T, \tag{III.13}$$
$$\mathbf{X}_i^{[r]} = \left(X_1^{[r+1]}, \ldots, X_{i-1}^{[r+1]}, X_i^{[r]}, X_{i+1}^{[r]}, \ldots, X_n^{[r]}\right)^T. \tag{III.14}$$

3. `Take`
$$X_i^{[r+1]} = \begin{cases} \widetilde{X}_i, & \texttt{with probability } \alpha\left(X_i^{[r]}, \widetilde{X}_i\right); \\ X_i^{[r]}, & \texttt{with probability } 1-\alpha\left(X_i^{[r]}, \widetilde{X}_i\right). \end{cases}$$

4. `Update`

$$\mathbf{X}_i^{[r+1]} = \left(X_1^{[r+1]}, \ldots, X_i^{[r+1]}, X_{i+1}^{[r]}, \ldots, X_n^{[r]}\right)^T.$$

Remark 18. *In practice, the components of* $\mathbf{X}^{[r+1]}$ *can be simulated in a random order, as mentioned in Liu et al. (1995, (62)), in order to accelerate the convergence of the chain towards its stationary distribution* $\pi_{\widehat{H}}\left(m, C, \mathbf{X} \,|\, \ldots\right)$*, by increasing the "mixing" of the simulated Markov chain.*

In what follows, similar instrumental distributions J are to proposed as in the *rich man* version. However, the computations are somewhat different, as described now:

1. J could be the normal distribution $\mathcal{N}\left(m^{[r+1]}, C^{[r+1]}\right)$. which is independent from the previous value $X_i^{[r]}$:

$$J\left(\widetilde{X}_i \,|\, \theta^{[r+1]}, X_i^{[r]}\right) \;=\; J\left(\widetilde{X}_i \,|\, \theta^{[r+1]}\right),\ 1 \leq i \leq n. \tag{III.15}$$

Thus, the ratio term in $\alpha\left(X_i^{[r]}, \widetilde{X}_i\right)$ can be rewritten as:

$$\begin{aligned}
&\frac{\pi_{\widehat{H}}\left(\widetilde{\mathbf{X}}_i \mid \boldsymbol{\mathcal{Y}}, \mathbf{d}, \theta^{[r+1]}, \rho, \mathbf{H}_D\right)}{\pi_{\widehat{H}}\left(\mathbf{X}_i^{[r]} \mid \boldsymbol{\mathcal{Y}}, \mathbf{d}, \theta^{[r+1]}, \rho, \mathbf{H}_D\right)} \frac{J\left(X_i^{[r]} | \widetilde{X}_i, \theta^{[r+1]}\right)}{J\left(\widetilde{X}_i | X_i^{[r]}, \theta^{[r+1]}\right)} \\
= \;& \frac{\pi\left(\widetilde{\mathbf{X}}_i \mid \theta^{[r+1]}\right) \pi_{\widehat{H}}\left(\boldsymbol{\mathcal{Y}} \mid \widetilde{\mathbf{X}}_i, \mathbf{d}, \rho, \mathbf{H}_D\right)}{\pi\left(\mathbf{X}_i^{[r]} \mid \theta^{[r+1]}\right) \pi_{\widehat{H}}\left(\boldsymbol{\mathcal{Y}} \mid \mathbf{X}_i^{[r]}, \mathbf{d}, \rho, \mathbf{H}_D\right)} \frac{J\left(X_i^{[r]} \mid \theta^{[r+1]}\right)}{J\left(\widetilde{X}_i \mid \theta^{[r+1]}\right)} \\
= \;& \frac{\pi_{\widehat{H}}\left(\boldsymbol{\mathcal{Y}} \mid \widetilde{\mathbf{X}}_i, \mathbf{d}, \rho, \mathbf{H}_D\right)}{\pi_{\widehat{H}}\left(\boldsymbol{\mathcal{Y}} \mid \mathbf{X}_i^{[r]}, \mathbf{d}, \rho, \mathbf{H}_D\right)}
\end{aligned}$$

Remark 19.

$$\frac{\pi\left(\widetilde{\mathbf{X}}_i \mid \theta^{[r+1]}\right)}{\pi\left(\mathbf{X}_i^{[r]} \mid \theta^{[r+1]}\right)} \frac{J\left(X_i^{[r]} \mid \theta^{[r+1]}\right)}{J\left(\widetilde{X}_i \mid \theta^{[r+1]}\right)} \;=\; 1, \tag{III.16}$$

since using the notation $\widetilde{\mathbf{X}}_i$ *and* $\mathbf{X}_i^{[r]}$ *defined in* (III.13) *and* (III.14),

$$\begin{aligned}
\frac{\pi\left(\widetilde{\mathbf{X}}_i \mid \theta^{[r+1]}\right)}{\pi\left(\mathbf{X}_i^{[r]} \mid \theta^{[r+1]}\right)} &= \frac{\exp\left\{-\frac{1}{2}\left(\widetilde{X}_i - m^{[r+1]}\right)^T \left(C^{[r+1]}\right)^{-1}\left(\widetilde{X}_i - m^{[r+1]}\right)\right\}}{\exp\left\{-\frac{1}{2}\left(X_i^{[r]} - m^{[r+1]}\right)^T \left(C^{[r+1]}\right)^{-1}\left(X_i^{[r]} - m^{[r+1]}\right)\right\}} \\
&\cdot \frac{\exp\left\{-\frac{1}{2}\sum_{j=1}^{i-1}\left(X_j^{[r+1]} - m^{[r+1]}\right)^T \left(C^{[r+1]}\right)^{-1}\left(X_j^{[r+1]} - m^{[r+1]}\right)\right\}}{\exp\left\{-\frac{1}{2}\sum_{j=1}^{i-1}\left(X_j^{[r+1]} - m^{[r+1]}\right)^T \left(C^{[r+1]}\right)^{-1}\left(X_j^{[r+1]} - m^{[r+1]}\right)\right\}} \\
&\cdot \frac{\exp\left\{-\frac{1}{2}\sum_{j=i+1}^{n}\left(X_j^{[r]} - m^{[r+1]}\right)^T \left(C^{[r+1]}\right)^{-1}\left(X_j^{[r]} - m^{[r+1]}\right)\right\}}{\exp\left\{-\frac{1}{2}\sum_{j=i+1}^{n}\left(X_j^{[r]} - m^{[r+1]}\right)^T \left(C^{[r+1]}\right)^{-1}\left(X_j^{[r]} - m^{[r+1]}\right)\right\}} \\
&= \frac{J\left(\widetilde{X}_i \mid \theta^{[r+1]}\right)}{J\left(X_i^{[r]} \mid \theta^{[r+1]}\right)}.
\end{aligned}$$

Remark 20. *The large vector* $\boldsymbol{\mathcal{Y}}$ *(referring to Section II.1.2) is normally distributed knowing the variables* $\mathbf{X}, \mathbf{d}$ *and the observations* $\mathbf{H}_{D_N}$*:*

$$\boldsymbol{\mathcal{Y}} \,|\, \mathbf{X}, \mathbf{d}, \rho, \mathcal{H}_{D_N} = \mathbf{H}_{D_N} \;\sim\; \mathcal{N}\left[\widehat{H}(\mathbf{Z}), \mathbf{R} + MSE(\mathbf{Z})\right] \tag{III.17}$$

with $\mathbf{Z} = (\mathbf{X}, \mathbf{d})$*. Its conditional density is given by*

$$\begin{aligned}
&\pi_{\widehat{H}}(\boldsymbol{y} \,|\, \mathbf{X}, \mathbf{d}, \rho, H_D) \propto |\mathbf{R} + MSE(\mathbf{Z})|^{-\frac{1}{2}} \\
&\cdot \exp\left\{-\frac{1}{2}\left(\left(y_1 - \widehat{H}(Z_1)\right)', \ldots, \left(y_n - \widehat{H}(Z_n)\right)^T\right)\left(\mathbf{R} + MSE(\mathbf{Z})\right)^{-1}\begin{pmatrix} \left(y_1 - \widehat{H}(Z_1)\right) \\ \vdots \\ \left(y_n - \widehat{H}(Z_n)\right)\end{pmatrix}\right\}.
\end{aligned}$$

2. J could be the normal distribution $\mathcal{N}\left(X_i^{[r]}, \kappa C^{[r+1]}\right)$ with κ to be chosen, the so-called "symmetric" instrumental distribution since

$$J\left(\widetilde{X}_i \mid X_i^{[r]}, \theta^{[r+1]}\right) = J\left(X_i^{[r]} \mid \widetilde{X}_i, \theta^{[r+1]}\right). \tag{III.18}$$

The ratio term in $\alpha\left(X_i^{[r]}, \widetilde{X}_i\right)$ can then be simplified in a similar manner:

$$\begin{aligned}
&\frac{\pi_{\widehat{H}}\left(\widetilde{\mathbf{X}}_i \mid \boldsymbol{\mathcal{Y}}, \mathbf{d}, \theta^{[r+1]}, \rho, \mathbf{H}_D\right)}{\pi_{\widehat{H}}\left(\mathbf{X}_i^{[r]} \mid \boldsymbol{\mathcal{Y}}, \mathbf{d}, \theta^{[r+1]}, \rho, \mathbf{H}_D\right)} \frac{J\left(X_i^{[r]} | \widetilde{X}_i, \theta^{[r+1]}\right)}{J\left(\widetilde{X}_i | X_i^{[r]}, \theta^{[r+1]}\right)} \\
&= \frac{\pi_{\widehat{H}}\left(\widetilde{\mathbf{X}}_i \mid \boldsymbol{\mathcal{Y}}, \mathbf{d}, \theta^{[r+1]}, \rho, \mathbf{H}_D\right)}{\pi_{\widehat{H}}\left(\mathbf{X}_i^{[r]} \mid \boldsymbol{\mathcal{Y}}, \mathbf{d}, \theta^{[r+1]}, \rho, \mathbf{H}_D\right)} \cdot 1 \\
&= \frac{\pi_{\widehat{H}}\left(\widetilde{\mathbf{X}}_i, \boldsymbol{\mathcal{Y}} \mid \mathbf{d}, \theta^{[r+1]}, \rho, \mathbf{H}_D\right)}{\pi_{\widehat{H}}\left(\mathbf{X}_i^{[r]}, \boldsymbol{\mathcal{Y}} \mid \mathbf{d}, \theta^{[r+1]}, \rho, \mathbf{H}_D\right)},
\end{aligned}$$

where the joint distribution of $(\widetilde{\mathbf{X}}_i, \boldsymbol{\mathcal{Y}})$ given $(\mathbf{d}, \theta^{[r+1]}, \rho, \mathbf{H}_D)$ can be computed as:

$$\begin{aligned}
&\pi_{\widehat{H}}\left(\mathbf{x}, \boldsymbol{y} \mid \mathbf{d}, \theta^{[r+1]}, \rho, \mathbf{H}_D\right) = \pi\left(\mathbf{x} \mid \theta^{[r+1]}, \rho, \mathbf{d}, H_D\right) \pi_{\widehat{H}}\left(\boldsymbol{y} \mid \mathbf{x}, \mathbf{d}, \rho, \mathbf{H}_D\right) \\
\propto\ & |\mathbf{R} + \mathrm{MSE}(\mathbf{z})|^{-\frac{1}{2}} \cdot \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left[(x_i - m)' C^{-1} (x_i - m)\right] \right. \\
& \left. -\frac{1}{2}\left(\left(y_1 - \widehat{H}(z_1)\right)^T, \ldots, \left(y_n - \widehat{H}(z_n)\right)'\right)\left(\mathbf{R} + \mathrm{MSE}(\mathbf{z})\right)^{-1} \begin{pmatrix} \left(y_1 - \widehat{H}(z_1)\right) \\ \vdots \\ \left(y_n - \widehat{H}(z_n)\right) \end{pmatrix} \right\},
\end{aligned}$$

according to (II.11).

3. J could be a succession of q uni dimensional Gaussian random walks $\mathcal{N}\left(X_i^{[r]}(l), \kappa C^{[r+1]}(l,l)\right)$ with $l = 1, \ldots, q$, which updates $X_i^{[r+1]}$ dimension by dimension. Similarly to the *rich man* version, the current candidate $\widetilde{X}_i$ is, after the update of the l-th dimension :

$$\widetilde{X}_i = \left(X_i(1), \ldots, X_i(l-1), \widetilde{X}_i(l), X_i^{[r]}(l+1), \ldots, X_i^{[r]}(q)\right)^T, \quad \text{(III.19)}$$

where the $(q - l)$ positions $(X_i^{[r]}(l+1), \ldots, X_i^{[r]}(q))$ have not yet been modified.

III.2 Convergence issues of the MCMC algorithms

III.2.1 Two important theorems

The following theorems ensure the convergence of a Markov chain (see Gilks et al., 1996, (36)).

Theorem 1. *Suppose that $\left(\psi^{[t]}\right)$ is positive recurrent, with stationary distribution $\pi(\cdot)$, then for any real function $f \in \mathcal{L}^1(\pi)$,*

$$\bar{f}_N(\psi) \xrightarrow{\mathbb{P}} \mathbb{E}_\pi\left[f(\psi)\right], \quad \text{(III.20)}$$

where $\bar{f}_N(\psi)$ is the empirical mean of $f\left(\psi^{[t]}\right)_{t=1,\ldots,N}$ and $\mathbb{E}_\pi\Big[f(\psi)\Big]$ is the expectation of $f(\psi)$ with respect to $\pi(\cdot)$.

Theorem 2. *If, in addition, $\left(\psi^{[t]}\right)$ is aperiodic with transition kernel $P(\cdot,\cdot)$, then*

$$\left\|P^n(\psi^{(0)}, \cdot) - \pi(\cdot)\right\| \xrightarrow[n\to\infty]{} 0 \quad \text{(III.21)}$$

*for π-almost all starting point $\psi^{(0)}$, which is equivalent to the convergence on **total variation (TV) norm**:*

$$\lim_{n\to\infty} \left\| \int P^n(\psi^{(0)}, \cdot)\, \mu(d\psi) - \pi(\cdot)\right\|_{TV} = 0, \quad \text{(III.22)}$$

for any initial distribution μ, where $P^n(\psi^{(0)}, \cdot)$ denotes the kernel for n transitions with starting point $\psi^{(0)}$ defined as follows

$$P^n(\psi^{(0)}, A) \quad = \quad \int_{\Psi} P^{n-1}(y, A) P(\psi^{(0)}, dy). \tag{III.23}$$

Recall that the **total variation norm** is defined by

$$\left\|\mu_1 - \mu_2\right\|_{\mathrm{TV}} \quad = \quad \sup_{A} |\mu_1(A) - \mu_2(A)|. \tag{III.24}$$

III.2.2 Convergence of MH Markov chain

To verify the convergence of the Metropolis-Hastings chain is equivalent to verify the property of irreducibility and aperiodicity of the chain. With respect to its definition, the *irreducibility* follows from sufficient conditions such as the positivity of the instrumental distribution J:

$$J\left(x_1 | x_2\right) \quad > \quad 0, \ \forall (x_1, x_2) \in \mathcal{X} \times \mathcal{X}, \tag{III.25}$$

where $\mathcal{X}$ denotes the q first dimensions of Ω which corresponds to the domain for X_i. Moreover, a sufficient condition for the Metropolis-Hastings chain to be *aperiodic* is that the probability of events such that $\{X_i^{[r+1]} = X_i^{[r]}\}$ is not zero, and thus

$$\mathbb{P}\left[\frac{\pi_H\left(\widetilde{X}_i \mid Y_i, d_i, \theta^{[r+1]}, \rho\right)}{\pi_H\left(X_i^{[r]} \mid Y_i, d_i, \theta^{[r+1]}, \rho\right)} \frac{J\left(X_i^{[r]} | \widetilde{X}_i, \theta^{[r+1]}\right)}{J\left(\widetilde{X}_i | X_i^{[r]}, \theta^{[r+1]}\right)} \geq 1\right] \quad < \quad 1. \tag{III.26}$$

Thus, a MH Markov chain satisfying (III.25) and (III.26) converges to its target distribution since Theorems 1 and 2 hold.

III.2.3 Convergence of Metropolis-Hastings-within-Gibbs samplers

For the Markov chain $\left(\psi^{[t]}\right)_t = \left(m^{[t]}, C^{[t]}, \mathbf{X}^{[t]}\right)_t$ constructed in this chapter, the following theorem (Roberts and Casella, 2004, (90)) holds.

Theorem 3. *If one of the following conditions*

i) $\pi\left(\psi \,|\, \mathbf{y}, \mathbf{d}, \rho\right)$ *satisfies the positivity condition, i.e.*

$$\pi_i\left(\psi_i \,|\, \mathbf{y}, \mathbf{d}, \rho\right) \quad > \quad 0 \text{ for every } i = 1, 2, 3 \text{ implies that } \pi\left(\psi \,|\, \mathbf{y}, \mathbf{d}, \rho\right) > 0,$$

where π_i denotes the marginal distribution of ψ_i;

ii) *the transition kernel is absolutely continuous with respect to $\pi\left(\psi \,|\, \mathbf{y}, \mathbf{d}, \rho\right)$ (Tierney, 1994),*

is satisfied, the chain is irreducible.

Without the MH step, the condition of absolute continuity on the Gibbs kernel is satisfied by most decompositions. However, in the Metropolis-Hastings-within-Gibbs algorithm, the absolute continuity is lost (referring to Roberts and Casella, 2004, (90), pp. 380). The irreducibility property has to be established for the construction considered. It is then necessary to either study the positivity condition or the recursion properties of the chain. In our case, the positivity condition is well verified as the marginal posterior density is always positive. Thus the hybrid MCMC algorithm works well.

Moreover, for our irreducible Metropolis-Hastings-within-Gibbs Markov chain, the Harris recurrence property car be guaranteed thanks to the following proposition (Roberts and Rosenthal, 2006, (91)). This property allows us to replace the convergence from "almost all" starting points by "all" starting points.

Proposition 3. *Consider an irreducible Metropolis-Hastings-within-Gibbs Markov chain. Suppose that from any initial state x, with probability 1, the chain will eventually move at least once in each coordinate direction. Then the chain is Harris recurrent.*

III.2.4 Diagnosis of the convergence: the Brooks-Gelman statistic

Background: Statistic of Gelman and Rubin

In 1992, Gelman and Rubin ((33)) proposed a statistic to diagnose the convergence of the simulated Markov chain $\left(\psi^{[t]}\right) = \left(m^{[t]}, C^{[t]}, \mathbf{X}^{[t]}\right)$. The method is based on m parallel chains, generated from different initial values and only the M final simulations after the "*burn-in*" period are considered.

Gelman and Rubin's approach relies on the assumption of *normality*, which means that the behavior of potential inferences, i.e. the posterior distributions of the variables of interest, can be summarized by the mean and variance of the simulated draws. In the present work, the statistic is calculated at each dimension for each quantity of interest m and C ($\mathbf{X}$ itself is not a parameter to be explored), and a mean statistic is then computed. For each dimension of each component denoted by ξ, l parallel chains are simulated and M iterations are collected after the burn-in period, as follows:

$$\begin{aligned} \xi_1 &= \left\{\xi_1^1, \ldots, \xi_1^M\right\} \\ &\vdots \\ \xi_l &= \left\{\xi_l^1, \ldots, \xi_l^M\right\}. \end{aligned}$$

The between-chain variance B/M and the within-chain variance W can be calculated as follows:

$$\frac{B}{M} = \frac{1}{l-1}\sum_{j=1}^{l}\left(\overline{\xi}_j - \overline{\xi}\right)^2; \tag{III.27}$$

$$W = \frac{1}{l(M-1)}\sum_{j=1}^{l}\sum_{k=1}^{M}\left(\xi_j^k - \overline{\xi}_j\right)^2, \tag{III.28}$$

where $\overline{\xi}_j$ denotes mean of the M draws from the j-th chain and $\overline{\xi}$ denotes mean of the lM

draws from all the chains:

$$\overline{\xi}_j = \frac{1}{M}\sum_{k=1}^{M}\xi_j^k,\ j=1,\ldots,l; \tag{III.29}$$

$$\overline{\xi} = \frac{1}{lM}\sum_{j=1}^{l}\sum_{k=1}^{M}\xi_j^k. \tag{III.30}$$

Then, an unbiased estimator of the variance σ^2 can be given by a weighted average of B and W:

$$\widehat{\sigma^2} = \frac{M-1}{M}W + \frac{1}{M}B. \tag{III.31}$$

By ignoring the minor contribution to variability brought by the degrees of freedom, the Gelman and Rubin statistic $\widehat{R}_{GR}$ is the (*over*-)estimate of the ratio of between and within-sequence inferences:

$$\widehat{R}_{GR} = \frac{\widehat{\sigma^2} + \frac{B}{lM}}{W}, \tag{III.32}$$

where the denominator indicates the between posterior variance estimate accounting for the variability of the estimator $\overline{\xi}$. After simplification,

$$\widehat{R}_{GR} = \frac{M-1}{M} + \frac{l+1}{lM}\frac{B}{W} \geq 1. \tag{III.33}$$

The fact that $\widehat{R}_{GR}$ approaches to 1 is to say that the posterior distribution of the parallel chains is close to the target distribution, which indicates the convergence has been reached. Otherwise, a large $\widehat{R}_{GR}$ suggests that a longer time should be waited to increase W or to decrease $\widehat{\sigma^2}$.

Remark 21. *A great limit of this approach is the assumption of normality of the posterior distributions, which are thus summarized only by means and variances.*

Diagnosis: Brooks and Gelman statistic

In 1998, Brooks and Gelman ((11)) proposed a method derived from the Gelman and Rubin's approach (1992, (33)), which avoid the assumption of normality, for monitoring the convergence of iterative simulations. Once again, this statistic denoted by $\widehat{R}_{BG}$ is constructed on the final M iterations after the "burn-in" period from l parallel simulated chains,

$$\begin{aligned} \xi_1 &= \left\{\xi_1^1,\ldots,\xi_1^M\right\} \\ &\vdots \\ \xi_l &= \left\{\xi_l^1,\ldots,\xi_l^M\right\} \end{aligned}$$

described as follows:

1. For each j-th individual chain ξ_j, calculate the empirical $100(1-\alpha)\%$ interval which is the difference between the $100(1-\frac{\alpha}{2})\%$ and $100\frac{\alpha}{2}\%$ percentile of the M simulated points $\left\{\xi_j^1, \ldots, \xi_j^M\right\}$. Thus, form the l within-sequence interval length estimates.

2. For the entire set of lM simulated draws from all chains, calculate the empirical $100(1-\alpha)\%$ interval in the same way, to construct a total-sequence interval length estimate.

3. Evaluate the statistic $\widehat{R}_{BG}$ defined as

$$\widehat{R}_{BG} = \frac{\Delta}{\bar{\delta}},$$

where

- Δ is the total-sequence interval length computed in step 2:

$$\Delta = Q_{100(1-\frac{\alpha}{2})\%} - Q_{100\frac{\alpha}{2}\%}, \tag{III.34}$$

where Q_p denotes the p-quantile for the set of the lM simulations;

- $\bar{\delta} = \frac{1}{l}\sum_{j=1}^{l}\theta_j$, with θ_j the length of the within-sequence interval for the j-th chain calculated in step 1:

$$\theta_j = Q^j_{100(1-\frac{\alpha}{2})\%} - Q^j_{100\frac{\alpha}{2}\%}, \tag{III.35}$$

where Q_p^j denotes the p-quantile for the j-th chain.

Similarly, this statistic is calculated at each dimension for each of the parameters (m, C) and the mean of statistics is then computed, denoted by $\widehat{R}_{BG}$. The threshold value 1.2 is advocated by the authors to declare the convergence of the simulated Markov chains.

Remark 22. *In our experiments, we make use of a more conservative threshold to ensure that the MCMC algorithms have converged to their stationary distribution. A MCMC chain has been declared to have converged if the $\widehat{R}_{BG}$ statistics is smaller than 1.05 for 3,000 successive iterations.*

A practical trick : eliminating a troublesome chain To accelerate the convergence rate, in practice, a trick is as follows. Among l parallel simulated chains, a bad chain which is stuck and not converging may appear, while other chains perform well. In this case, to detect and eliminate this tricky chain, the following procedure is suggested.

1. `Simulate` l `Markov chains and continue` k `iterations after the supposed-to-be burn-in period.`

   Remark 23. *The true burn-in period should not be reached, otherwise we do not need this trick ...*

2. `For each quantity of interest` m `and` C`, calculate at each dimension the` $\widehat{R}_{BG}$ `statistics based on the` k `last simulations from the` l `parallel chains.`

– If $\widehat{R}_{BG} < 1.05$, do nothing as the convergence has been reached;

– else, calculate the $\widehat{R}_{BG}$ statistic by removing the i-th chain (*leave-one-out* procedure) with $1 \leq i \leq l$, denoted by $\widehat{R}_{BG,-i}$. If there exists i such that $\widehat{R}_{BG,-i} < 1.05$, the number of candidate $i^* = i$.

3. Continue the MCMC algorithm and repeat frequently the calculations of $\widehat{R}_{BG,-i}$ (e.g. every 50 iterations) for this extended period. If each time it indicates the same candidate the i-th chain, the i-th chain is thus to be eliminated.

Remark 24. *In this thesis, this strategy works well under the assumption that there is only a single bad chain.*

Some simulation results are given in Figures III.2-III.5. Comparing these figures we can see that once a wrong chain has been removed, the convergence is immediately reached, indicated by the statistic $\widehat{R}_{BG}$ which immediately decreases below 1.05.

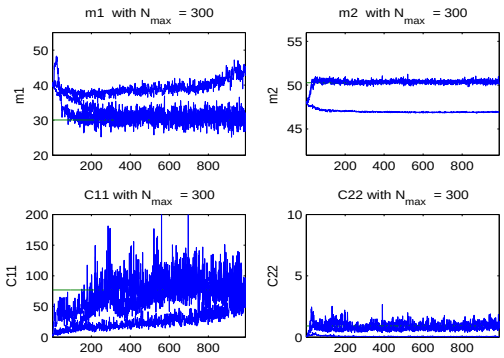


Figure III.2: Realization before removing the chain

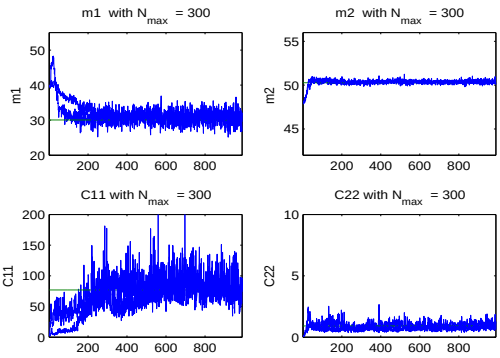


Figure III.3: Realization after removing the chain

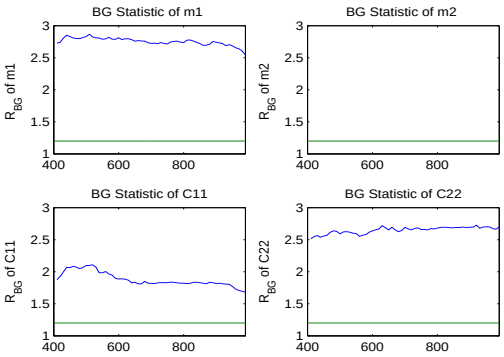


Figure III.4: $\widehat{R}_{BG}$ before elimination

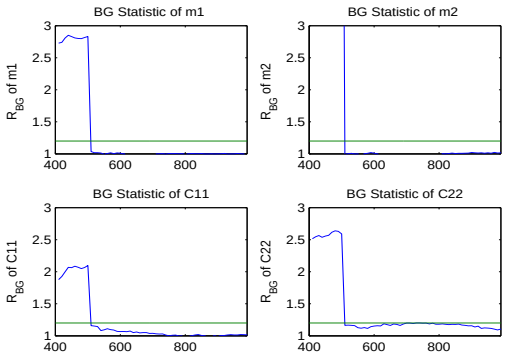


Figure III.5: $\widehat{R}_{BG}$ after elimination

III.3 First numerical results of the MCMC algorithm

This section deals with numerical experiments to illustrate the behavior of the hybrid MCMC algorithm. Let us recall the general form of inverse problems:

$$Y_i = H(X_i, d_i) + U_i,\ i = 1, \dots, n \tag{III.36}$$

where the non observed variable $X_i \sim \mathcal{N}_q(m, C)$, the error term $U_i \sim \mathcal{N}_p(\mathbf{0}, R)$ with a given R and the observed input d_i is assumed to be related to the experimental conditions. Moreover, as described in Chapter II, the prior distributions of the parameters $\theta = (m, C)$ are chosen as follows.

$$m|C \sim \mathcal{N}\left(\mu, \frac{C}{a}\right) \tag{III.37}$$

$$C \sim \mathcal{IW}\Big(\Lambda, \nu\Big) = \mathcal{IW}\left(t \cdot \widetilde{C}_{\text{Exp}}, \nu\right). \tag{III.38}$$

A Gibbs sampler is then construct to approximate the posterior distribution of θ. Two examples for the function H are provided and tested. The first example for a physical hydraulic engineering model mentioned in Section I.1.3. The second example is the classical Sobol function defined in a unit domain. For each example, different designs of experiments (DOEs) have been construct with different numbers of points within different experimental domains. Moreover, with a fixed DOE, different proposal distributions have been tried in the MH step. We compare the results of the Gibbs sampler following these different experiments.

III.3.1 Example 1: A hydraulic engineering model

The first example is the three-dimensional-input hydraulic model, where the function $H : \mathbb{R}^2 \otimes \mathbb{R}^1 \to \mathbb{R}^2$ is given by:

$$H(x, d) = \left(x_2 + \left(\frac{\sqrt{5000}}{300\sqrt{55 - x_2}} \times \frac{d}{x_1} \right)^{0.6}, \frac{d^{0.4} x_1^{0.6} (55 - x_2)^{0.3}}{300^{0.4} \times 5000^{0.3}} \right)^T, \tag{III.39}$$

with $x = (x_1, x_2)$. In our case study, the observations $\mathbf{y} = (y_1^T, \dots, y_n^T)^T$ are generated from the inverse problem model (III.36) where

$$X_i \sim \mathcal{N}\Big(\begin{pmatrix} 30 \\ 50 \end{pmatrix}, \begin{pmatrix} 5^2 & 0 \\ 0 & 1 \end{pmatrix} \Big), \tag{III.40}$$

$$d_i \sim \text{Gumbel}\Big(1013, -458\Big), \tag{III.41}$$

and the error term U_i is assumed to be normally distributed

$$U_i \sim \mathcal{N}\Big(\mathbf{0}, \begin{pmatrix} 10^{-5} & 0 \\ 0 & 10^{-5} \end{pmatrix} \Big). \tag{III.42}$$

Moreover, the mean and variance of d_i are

$$\mathbb{E}(d_i) = 1013 - \gamma \times 458 = 748.6$$

$$\text{Var}(d_i) = \frac{\pi^2}{6} \times 458^2 = 345050,$$

$\gamma = 0.5772$ being the Euler constant.

The advantage of this "*data generation*" is that the simulation results can be evaluated by comparing with the standard MLE results in the complete sequence problem. Following the prior choices (III.37) and (III.38), the hyperparameters are chosen as follows.

$$\begin{cases} a = 1, \\ t = 2, \\ \nu = 5, \\ \mu = \begin{pmatrix} 35 \\ 49 \end{pmatrix}, \\ \widetilde{C}_{\text{Exp}} = \begin{pmatrix} 7.5^2 & 0 \\ 0 & 1.5^2 \end{pmatrix}. \end{cases} \tag{III.43}$$

Moreover, the sample size n is fixed equal to 30 and the design domain Ω is given as

$$\Omega \quad = \quad [20, 40] \times [45, 55] \times [\min(\mathbf{d}), \max(\mathbf{d})], \tag{III.44}$$

with $\mathbf{d} = \{d_1, \ldots, d_{30}\}$.

The Brooks-Gelman statistic is applied to diagnose the convergence, where the criterion $\widehat{R}_{BG}$ is, in practice, calculated every 100 iterations for each coordinate m_1, m_2, C_{11} and C_{22} of θ and we control the maximum of these calculated statistics. The convergence is accepted if the maximal $\widehat{R}_{BG}$ remains smaller than 1.05 for 3,000 successive iterations.

Test $\mathcal{A}$. DOEs D_{20}, D_{100} vs. D_{500}

Three DOEs are generated as the standard *maximin*-Latin Hypercube Designs (LHDs) (see Chapter IV and V for more details), with 20, 100 and 500 points. The *rich man* version, which is not depending on the meta-modeling technique, is used as a benchmark.

The following numerical experiments aim at estimating the posterior distributions of θ with the help of the Gibbs sampler. We are interested in measuring the impact of the quality of the DOEs, more precisely, the number of points of the DOE, in the performance of Bayesian approach. As mentioned in the General Introduction, the meta-model brings an important emulator error in the posterior distributions of θ.

Figure III.6 shows the posterior distributions of different qualities with respect to different DOEs. For example, D_{100} and D_{500} brought a satisfying simulated distribution which is close to the benchmark (the non-kriging *rich man* version), while D_{20} appears not efficient enough. Thanks to Figure III.7, which displays the behavior of $\widehat{R}_{BG}$ for each component of m and C, we are ensured that the convergence of Markov chains has been reached for each case as all the $\widehat{R}_{BG}$s are below 1.05.

Test $\mathcal{B}$: proposal distributions J_1 vs. J_2

Two proposal distributions J_1 and J_2 in the Metropolis-Hastings algorithm (see Section III.1.3) have been tested, based on the same hydraulical model (III.39). The same *maximin*-LHD D_{100} has been built and the same hyperparameters have been chosen as described in

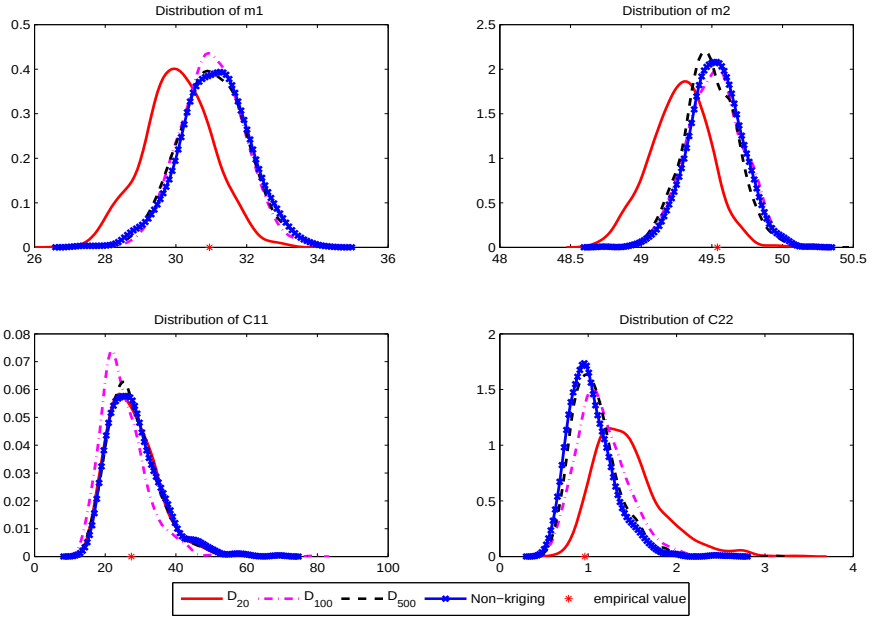


Figure III.6: Posterior distributions of θ with help of D_{20}, D_{100}, D_{500} and the non-kriging *rich man* version

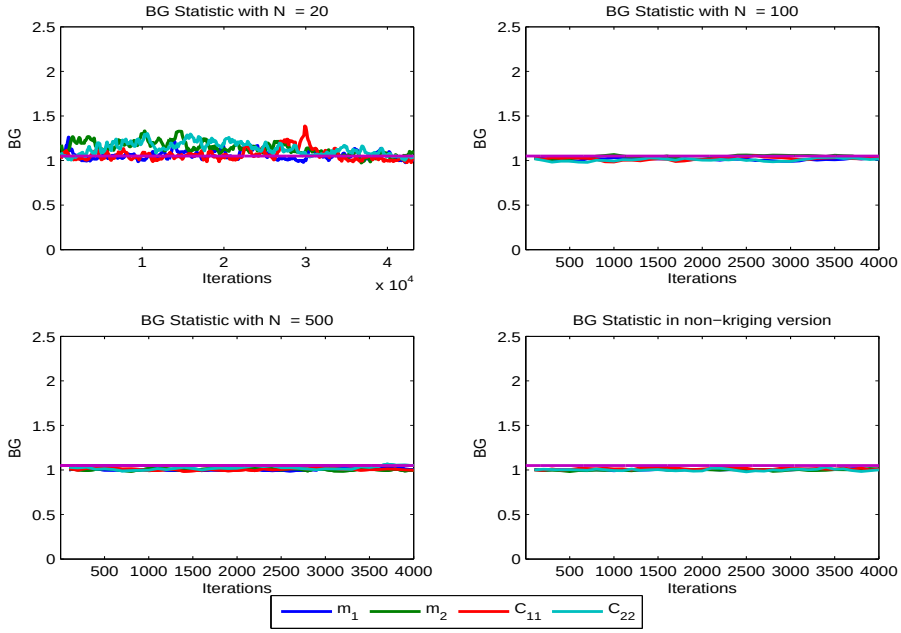


Figure III.7: $\widehat{R}_{BG}$ with help of D_{20}, D_{100}, D_{500} and the non-kriging *rich man* version

(III.43). At each $(r+1)$-th iteration, we choose

$$J_1 \quad : \quad X_i^{[r+1]} \sim \mathcal{N}\left(m^{[r+1]}, C^{[r+1]}\right); \tag{III.45}$$

$$J_2 \quad : \quad X_i^{[r+1]} \sim \mathcal{N}\left(X_i^{[r]}, \kappa C^{[r+1]}\right), \tag{III.46}$$

with $1 \leq i \leq n$ and $\kappa = 2$ for the burn-in period, $\kappa = 1$ for later iterations.

The posterior distributions of θ with respect to J_1 and J_2 are shown in Figure III.8 and the convergence in the two cases is ensured by Figure III.9 where all the $\widehat{R}_{BG}$s are below 1.05.

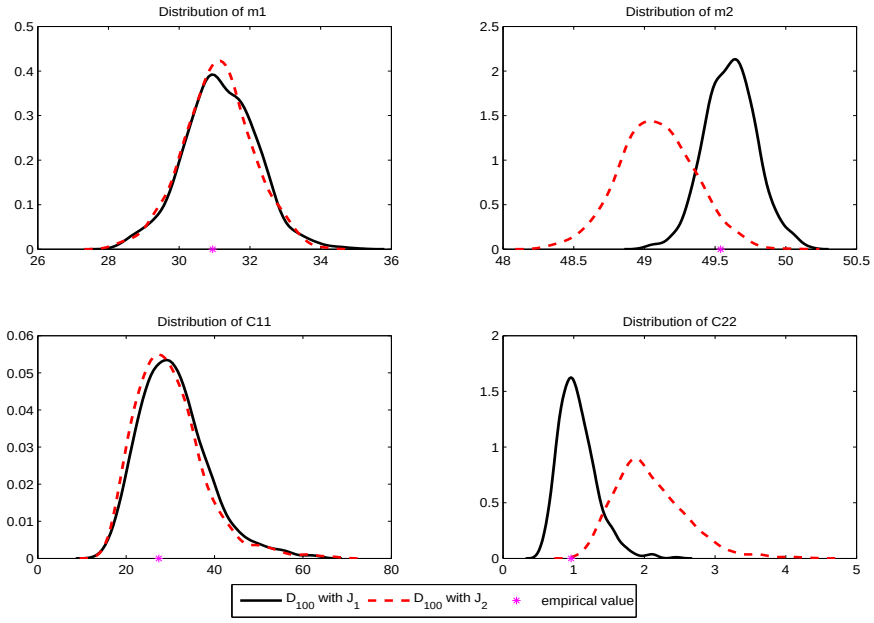


Figure III.8: Posterior distributions of θ with proposal distributions J_1 and J_2

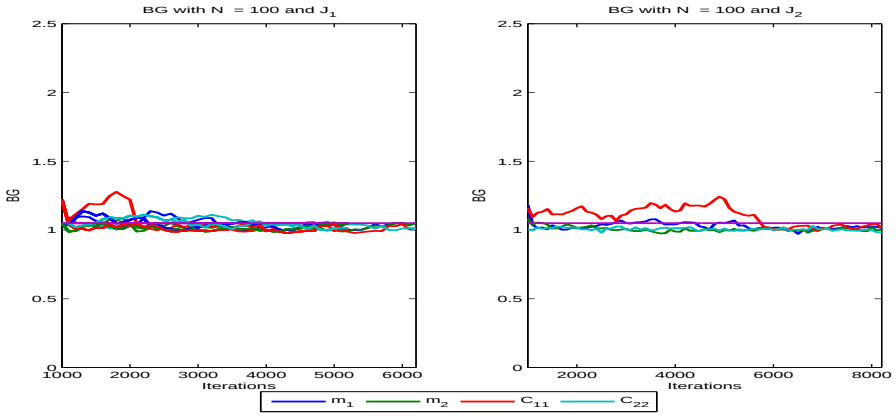


Figure III.9: $\widehat{R}_{BG}$ with help of D_{100} and proposal distributions J_1 and J_2

In this illustration, it is clear that the proposal distribution J_1 works better than J_2. Apart from their similar behaviors for m_1 and C_{11}, J_1 and J_2 give quite different results for m_2 and C_{22}. The distribution related to J_1 is well centered around the empirical value (the MLE estimator) and J_2 arises an algorithmic error, i.e. the Markov chains does converge but does not converge to the desired posterior distributions.

Test $\mathcal{C}$: domains Ω vs. Ω'

Apart from the kriging domain Ω described in (III.44), we introduce a much larger domain Ω' to illustrate the impact of kriging domains on the simulation results. The domain Ω' is chosen as follows:

$$\Omega' \quad = \quad [10, 100] \times [10, 55] \times [10, 2000]. \tag{III.47}$$

Figure III.10 illustrates the behaviors of the Gibbs sampler with respect to the two domains. The same DOE *maximin*-LHD with 100 points has been chosen. We see that the solid curve corresponding to the small domain Ω is well concentrated on the empirical value, which while the dashed curve related to the large domain Ω' is misleading.

That is why in the modeling procedure, we recommend to choose a kriging domain as small as possible while containing the eventual values of X_i with the greatest probability.

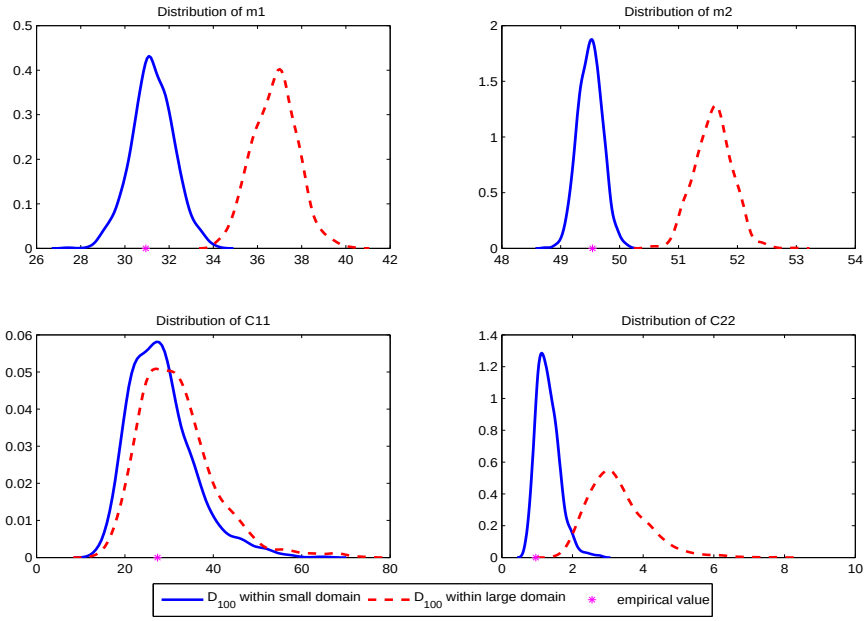


Figure III.10: Posterior distributions of θ in large and small kriging domains

Test $\mathcal{D}$. Good prior vs. bad prior vs. non-informative prior

To illustrate the impact of the prior distributions, we consider the following different prior distributions on the parameter θ. They are summarized in Table III.1.

Apart from the three mentioned informative priors, we also test the Jeffreys non informative prior, as presented in Section II.3. Figure III.11 displays the marginal posterior distributions based on 30 observations, with help of a *maximin*-LHD of 100 points. We obtain a huge different between the posterior distribution derived from the **BHV** prior and other priors. Especially, we notice that the Jeffreys non informative prior provides posterior results as good as the informative ones.

Prior	PLV	FHV	BHV
μ	$\{30, 50\}$	$\{35, 49\}$	$\{10, 54\}$
a	1	1	1
t	2	2	2
ν	5	5	5
$\widetilde{\mathbf{C}}_{\text{Exp}}$	$\begin{pmatrix} 1.5^2 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 7.5^2 & 0 \\ 0 & 1.5^2 \end{pmatrix}$	$\begin{pmatrix} 7.5^2 & 0 \\ 0 & 1.5^2 \end{pmatrix}$

Table III.1: Description of the three prior distributions: **PLV** = perfect mean and low variance, **FHV** = fair mean and high variance, **BHV** = bad mean and high variance.

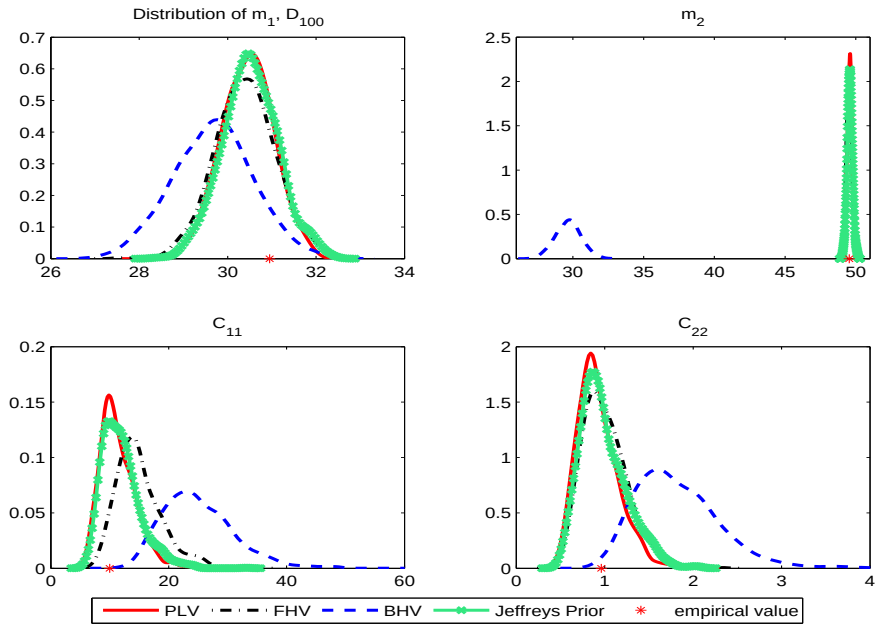


Figure III.11: Posterior distributions of θ with four types of prior distribution, based on 30 observations, with help of D_{100} in small domain

Test $\mathcal{E}$. Sample size $n = 5, 10$ vs. 30

Now we focus on the impact of the size of observations. We fix the prior distributions as described in (III.43) and a *maximin*-LHD of 100 points. By vary the sample size n from 5, 10 to 30, we obtain the following Figure III.12 from a Gibbs sampler.

As shown in it, each marginal posterior distribution of θ is quite sensible to the sample size. We notice that by enriching our observed data, our simulation can be largely improved. However, we should not forget that by adding observed data, the numerical computation becomes more and more expensive, especially resulting form the inversion of the matrix $\mathbf{R} + \text{MSE}(\mathbf{Z})$ of size $np \times np$ (cf. Chapter II).

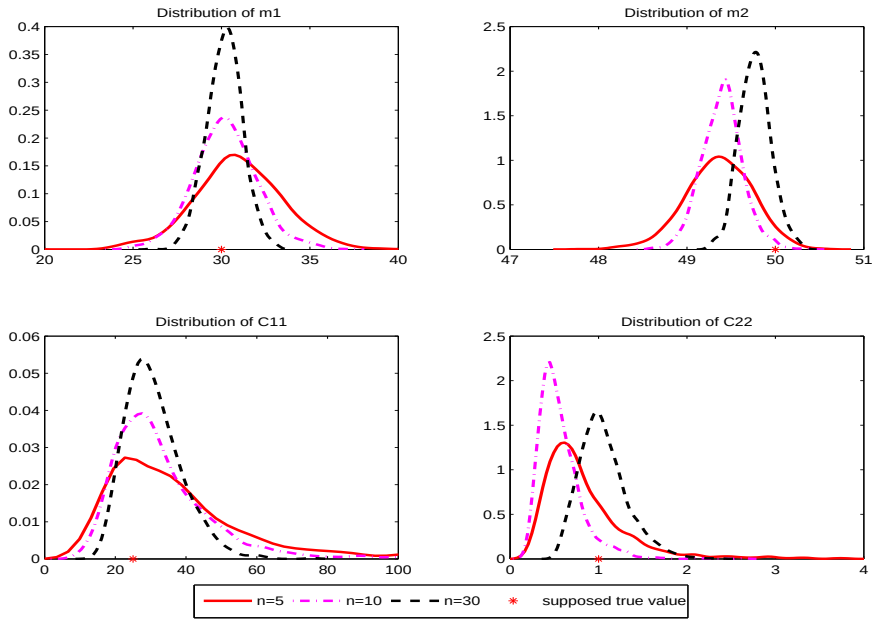


Figure III.12: Posterior distributions of θ based on 5, 10 and 30 observations, with help of D_{100} in small domain

III.3.2 Example 2: A classical Sobol function

The second example is the so-called g-function of Sobol, defined on $[0,1]^q$ as follows:

$$H(x_1, x_2) \quad = \quad g_1(x_1)\, g_2(x_2)\,, \text{ where } g_k(x) = \frac{|4x-2| + a_k}{1+a_k}, k = 1, 2. \tag{III.48}$$

In this function, $a_k \geq 0$ is called the weight coefficient, varying the contribution of each input x_k to the variability of the output. The small a_k is, the more significant the variable x_k is. In our experiments, for $\forall k$, a_k is fixed to 1. It is worth noting that this analytic function is highly nonlinear and non monotonic, as illustrated in Figure III.13.

The two-dimensional uncertainty model can be described as

$$Y_i \quad = \quad H(X_i) + U_i,\ i \in \{1, \ldots, n\}, \tag{III.49}$$

where the non observed variable $X_i = (X_i^1, X_i^2) \sim \mathcal{N}_2(m, C)$ and the error term $U_i \sim \mathcal{N}_1(0, R)$ with a given R. This time the design domain Ω is fixed to be $[0,1]^2$ and three *maximin*-LHDs with 20 points, 100 points and 500 points are generated.

By fixing the sample size n to 30, the dataset $\mathbf{y} = (y_i, i = 1, \ldots, 30)$ can be simulated from model (III.49) with the non observed input X_i generated from the following Gaussian distribution truncated in domain Ω

$$X_i \quad \sim \quad \mathbb{1}_\Omega \mathcal{N}_2\Big(\begin{pmatrix} 0.5 \\ 0.7 \end{pmatrix}, \begin{pmatrix} 0.15^2 & 0 \\ 0 & 0.4^2 \end{pmatrix} \Big), \tag{III.50}$$

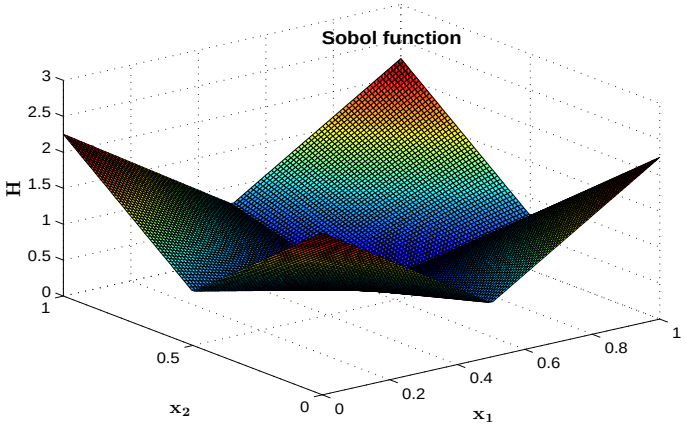


Figure III.13: Illustration of the Sobol function

and the error term U_i is generated from $\mathcal{N}_1(0, 10^{-5})$. Moreover, the hyperparameters are chosen as follows.

$$\begin{cases} a = 1, \\ \nu = 5, \\ \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \\ \Lambda = 2 \cdot \begin{pmatrix} 0.15^2 & 0 \\ 0 & 0.4^2 \end{pmatrix}. \end{cases}$$

In what follows, we aim at verifying the impacts of the quality of DOEs and the instrumental distribution required in the MH step on the posterior distributions of θ.

Test $\mathcal{A}$. DOEs D_{20}, D_{100} vs. D_{500}

Similarly to the previous hydraulic example, three standard *maximin*-LHDs D_{20}, D_{100} and D_{500} were generated in the same domain Ω. The Brooks-Gelman statistic $\widehat{R}_{BG}$ was calculated to verify the convergence of the Gibbs sampler not shown here.

Figure III.14 displays the posterior distributions of θ with different numerical DOEs. Once again, by increasing the number of points in the design, we improve the quality of the DOE and consequently, the distribution curve approaches to the benchmark. In this example, with the complex Sobol function, the difference between the posterior mean of m and the empirical mean (red point) illustrates the impact of the prior choice μ. We may increase the size of observations to reduce this impact to improve the simulation results.

Test $\mathcal{B}$: proposal distributions J_1 vs. J_2

Similarly to the first hydraulic case, the two proposal distributions J_1 and J_2, described in Section III.1.3, have been tested here. The same *maximin*-LHD D_{100} has been chosen for

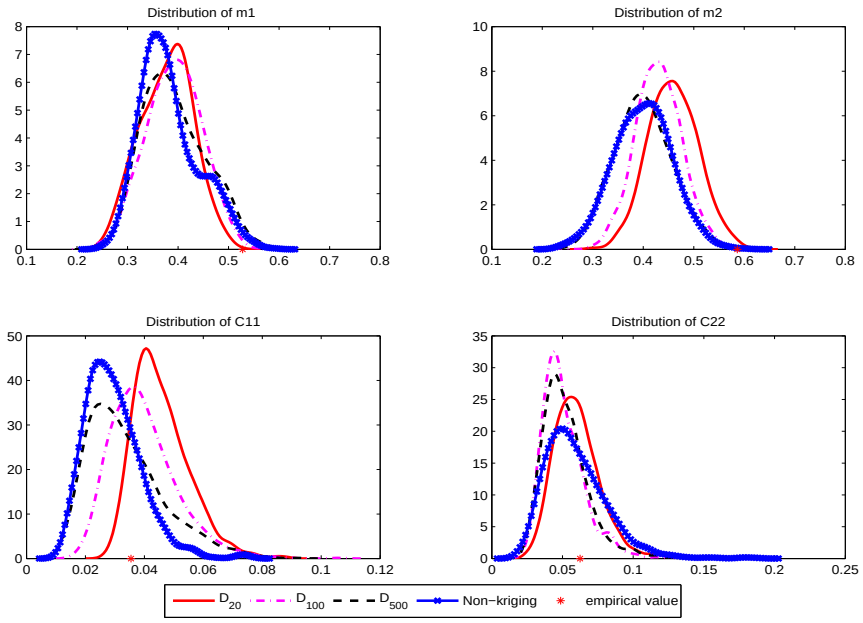


Figure III.14: Posterior distributions of θ with help of D_{20}, D_{100}, D_{500} and the non-kriging *rich man* version

both cases and the posterior distributions of θ after the convergence are displayed in Figure III.15. The BG statistic $\widehat{R}_{BG}$ ensures the convergence of the simulated Markov chains.

As shown in this figure, there are a significant difference between the two posterior distributions of θ by applying different proposal distributions J_1 and J_2. Contrary to the previous example, with this Sobol function, it is difficult to say which proposal distribution works better. It has been confirmed that the choice of the proposal distribution is highly related to numerical models.

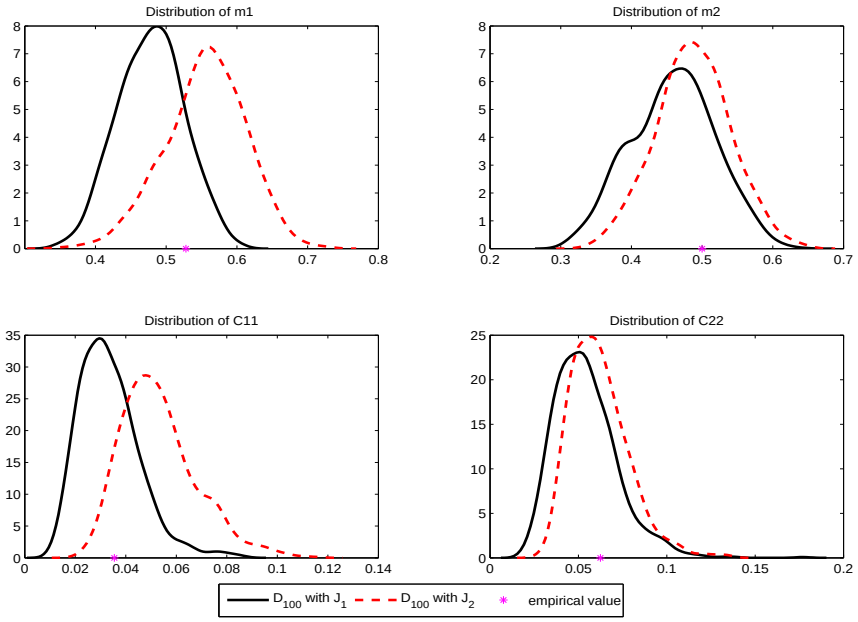


Figure III.15: Posterior distributions of parameters $\theta = (m, C)$ with proposal distributions J_1 and J_2

IV

Evaluation of the results and criteria of the quality of a design

Contents

This chapter is a collaboration with Gilles Celeux, Nicolas Bousquet and Mathieu Couplet. It has been published in INRIA, RR-7995.

The inverse problem considered here is to estimate the distribution of a non-observed random variable X from some noisy observed data Y linked to X through a time-consuming physical model H. Bayesian inference is considered to take into account prior expert knowledge on X in a small sample size setting. A Metropolis-Hastings-within-Gibbs algorithm is proposed to compute the posterior distribution of the parameters of X through a data augmentation process. Since calls to H are quite expensive, this inference is achieved by replacing H with a kriging emulator interpolating H from a numerical design of experiments. This approach involves several errors of different natures and, in this paper, we pay effort to measure and reduce the possible impact of those errors. In particular, we propose to use the so-called $\widetilde{DAC}$ criterion to assess in the same exercise the relevance of the numerical design and the prior distributions. After describing how computing this criterion for the emulator at hand, its behavior is illustrated on numerical experiments.

Keywords. Inverse problems, Bayesian analysis, Kriging, Design of Experiments, Assessment Error.

IV.1 Introduction

Probabilistic uncertainty treatment is gaining fast growing interest in the industrial field. Besides the uncertainty propagation challenges when dealing with complex and high CPU-time demanding physical models, one of the key issues regards the quantification of the sources of uncertainties. A key difficulty is linked to the highly-limited sampling information directly available on uncertain input variables. It can be highly beneficial (a) to integrate expert judgment, such as likely bounds on physical intervals or more elaborate probabilistic information, or (b) to integrate indirect information, such as data on other, more easily observable, parameters that are linkable to the uncertain variable of interest by a physical model. Methods for (b) are making use of probabilistic inverse methods since the recovering of indirect information involves generally the inversion of a physical model or a computer simulator H. It leads to the following uncertainty model

$$Y_i \quad = \quad H(X_i, d_i) + U_i, i \in \{1, \dots, n\}, \tag{IV.1}$$

where $X_i \in \mathbb{R}^q$ is a non-observed input, $d_i \in \mathbb{R}^{q_2}$ an observed input related to the experimental conditions and $U_i \in \mathbb{R}^p$ a measurement error. The error U_i and X_i are assumed to be independent for $i = 1, \dots, n$. Moreover the $(Y_i, i = 1, \dots, n)$ are independent. The purpose is to estimate the distribution of the random vectors X_is from the observations $(\mathbf{y}_i, i = 1, \dots, n)$, knowing that the function H (the physical model...) cannot be inverted. In what follows, the random vector X_i will be assumed to have a Gaussian distribution $\mathcal{N}_q(m, C)$, with mean m and variance matrix C to be estimated, and the error vector U_i will be assumed to have a Gaussian distribution $\mathcal{N}_p(\mathbf{0}, R)$, with known diagonal variance matrix R.

Many approaches are possible to approximate this inverse problem as linearizing the physical model H around a fixed point x_0 (see Celeux et al. 2010, (15)), or using a non linear approximation of the function H obtained through kriging and making use of a stochastic

procedure with this non linear approximation of H (see Barbillon et al. 2011, (4)). In this paper we opt for a Bayesian approach allowing to take into account prior knowledge that can be helpful, in particular, to avoid identifiability problems.

The estimation problem related to this inverse problem involves many possible errors:

- *Estimation error:* Usually the sample size n is small with respect to the dimension of the problem and the variance of the estimates could be expected to be large;
- *Emulator error:* Since H is too complex, there is the need to replace it with an emulator $\widehat{H}$ and the discrepancy between H and $\widehat{H}$ could induce an important error;
- *Algorithmic error:* To proceed to statistical inference, there is the need to use complex stochastic algorithms. In the Bayesian setting, those algorithms are Monte Carlo Markov Chains (MCMC) algorithms which produce Markov chains converging to the desired posterior distributions. But, controlling the convergence of the MCMC algorithms towards their limit distributions is important to get reliable estimates.
- *Prior error:* The prior knowledge on the parameters m and C is expected to produce regularized estimates of smaller variances than maximum likelihood estimates. But, if the prior distributions are irrelevant, it could jeopardize the statistical analysis.

Beyond the estimation problem, this paper is mainly concerned with the assessment of the quality of the proposed estimates. It implies to measure and control the above mentioned error sources. In this context, we focus on the *prior error* which received little attention and propose to measure it with a criterion (DAC) well-adapted for emulators defined on a compact set. Obviously those different error sources are linked and their relations for uncertainty analysis with small samples are discussed. The paper is organized as follows. In Section 2, the MCMC algorithm for a Bayesian estimation of an emulator of model (IV.1) is presented and the possible error sources are precisely described. Then, the DAC criterion to measure the prior error is presented in Section 3 as the resulting strategy for assessing both the emulator and the prior distribution. Numerical experiments, where different criteria assessing the different error sources are illustrated and compared, are presented in Section 4 and a Discussion section ends the paper.

IV.2 Bayesian inference with a Gaussian emulator

In the Bayesian framework, the first task is to choose a prior distribution $\pi(\theta)$ for the parameter $\theta = (m, C)$ to be estimated in the model (IV.1). A conjugate prior distribution has been selected

$$m \,|\, C \;\sim\; \mathcal{N}_q(\mu, C/a); \tag{IV.2}$$
$$C \;\sim\; \mathcal{IW}_q(\Lambda, \nu), \tag{IV.3}$$

the hyperparameters $\rho = (\mu, a, \Lambda, \nu)$ being specified by the user.

The posterior distribution $\pi(\theta|\mathbf{y})$ is approximated with a Gibbs sampler including a Metropolis-Hastings step (see for instance Tierney, 1995, (109)). Actually, the calculation of the full con-

ditional posterior distributions of m, C and $\mathbf{X} = \{X_1, \ldots, X_n\}$ lead to the following Gibbs sampler (below the $(r+1)$-th iteration):

`Given` $(m^{[r]}, C^{[r]}, \mathbf{X}^{[r]})$ `for` $r = 0, 1, 2, \ldots,$ `generate`

1. $C^{[r+1]}|\cdots \sim \mathcal{IW}\Big(\Lambda + \sum_{i=1}^{n}(m^{[r]} - X_i^{[r]})(m^{[r]} - X_i^{[r]})' + a(m^{[r]} - \mu)(m^{[r]} - \mu)',\ \nu + n + 1\Big)$

2. $m^{[r+1]}|\cdots \sim \mathcal{N}\Big(\frac{a}{n+a}\mu + \frac{n}{n+a}\overline{\mathbf{X}_n^{[r]}},\ \frac{C^{[r+1]}}{n+a}\Big)$ where $\overline{\mathbf{X}_n^{[r]}}$ denotes the empirical mean of the n vectors $X_i^{[r]}, i = 1, \ldots, n$

3. $\mathbf{X}^{[r+1]}|\cdots \propto \exp\Big\{ -\frac{1}{2}\sum_{i=1}^{n}\Big[\Big(X_i^{[r+1]} - m^{[r+1]}\Big)' \big(C^{[r+1]}\big)^{-1} \Big(X_i^{[r+1]} - m^{[r+1]}\Big) + \Big(Y_i - H(X_i^{[r+1]}, d_i)\Big)' R^{-1} \Big(Y_i - H(X_i^{[r+1]}, d_i)\Big) \Big]\Big\}$
   which is not belonging to a closed form family of distributions. That is why a Metropolis-Hastings (MH) step is used to simulate $\mathbf{X}^{[r+1]}$ from its full conditional distribution.

Now, considering situations where extensive sampling of $H(X, d)$ is too time-consuming, we propose to replace H with a *maximin* LHD (Latin Hypercube Design) kriging emulator $\widehat{H}$, following Barbillon (2010, (3)). This emulator is briefly described below.

- *Kriging* is a geostatistical method (Matheron 1971, (65)) that has been adapted by Sacks and al. (1989b, (98)) to approximate a physical model H on a bounded hypercube Ω. This method has known a growing interest in meta-modeling since the works of Koehler and Owen (1996, (54)), Santner and al. (2003, (99)) and Fang and al. (2006, (28)), among others. According to this approach the function H is regarded as the realization of a Gaussian Process (GP) $\mathcal{H} \sim \mathrm{GP}(\mu, c)$, characterised by its mean and variance functions: $\mu(z) = \mathbb{E}[\mathcal{H}(z)]$ and $c(z, z') = \mathrm{Cov}[\mathcal{H}(z), \mathcal{H}(z')] = \sigma^2 K_\epsilon(\|z - z'\|)$

  for any $z = (x, d)$, K_ϵ being a symetric positive kernel such that $K_\epsilon(0) = 1$. In a Bayesian perspective, GP modelling can be interpreted as providing H with a prior (Rasmussen & Williams, 2006, (86)). The process $\mathcal{H}$ can be proved to be normally distributed knowing some evaluations $\mathbf{H}_{D_N} = \{H(z_{(1)}), \ldots, H(z_{(N)})\}$ on a *design of experiments* $D_N = \{z_{(1)}, \ldots, z_{(N)}\}$ of N points $z_{(j)} = (x_{(j)}, d_{(j)})$.

  The best MSPE (*Mean Squared Prediction Error*) predictor of H, denoted by $\widehat{H}$, is the conditional mean:

$$\widehat{H}(z) \quad = \quad \mathbb{E}\left(\mathcal{H}(z) \,|\, \mathbf{H}_{D_N}\right), \forall z \in \Omega.$$

  Then $\widehat{H}(z)$ is minimizing the conditional expectation of the loss function $(\mathcal{H}(z) - \widehat{H}(z))^2$, so-called MSE (Mean Squared Error) (see Johnson et al. 1990 for details, (46)),

$$\mathrm{MSE}(z) \quad = \quad \mathbb{E}\left((\mathcal{H}(z) - \widehat{H}(z))^2 \,|\, \mathbf{H}_{D_N}\right), \forall z \in \Omega.$$

- The set $D_N = \{z_{(1)}, \ldots, z_{(N)}\}$ is chosen on $\Omega \in \mathbb{R}^{q+q_2}$ according to a *maximin* LHD (see McKay, Beckman, and Conover 1979, (66)): each dimension of the multidimensional domain Ω is divided into N intervals of equal length and the set D_N of N points are selected such that when projected on any dimension, each interval contains one and only one of the N projected points. Moreover, D_N is chosen to be *maximin*, i.e. it maximizes

$$\delta_D \quad = \quad \min_{i \neq j} \|z_{(i)} - z_{(j)}\|$$

among the LHD of size N.

For the kriging version, considering the new *emulator error*, the conditional distribution of $\mathbf{X}$ is as follows

$$\begin{aligned} & \pi(\mathbf{X} \,|\, \mathbf{Y}, m, C, \rho, \mathbf{H}_{D_N}) \;\propto\; \pi(\mathbf{X} \,|\, m, C, \rho, \mathbf{H}_{D_N}) \cdot \pi(\mathbf{Y} \,|\, \mathbf{X}, m, C, \rho, \mathbf{H}_{D_N}) \\ = \quad & |\mathbf{R} + \mathrm{MSE}(\mathbf{Z})|^{-\frac{1}{2}} \cdot \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \Big[(X_i - m)' C^{-1} (X_i - m) \Big] \right. \\ & \left. -\frac{1}{2} \Big(\Big(Y_1 - \widehat{H}(Z_1) \Big)', \ldots, \Big(Y_n - \widehat{H}(Z_n) \Big)' \Big) \Big(\mathbf{R} + \mathrm{MSE}(\mathbf{Z}) \Big)^{-1} \begin{pmatrix} \Big(Y_1 - \widehat{H}(Z_1) \Big) \\ \vdots \\ \Big(Y_n - \widehat{H}(Z_n) \Big) \end{pmatrix} \right\}, \end{aligned} \tag{IV.4}$$

where

$$\mathbf{R} \quad = \quad \begin{pmatrix} R_{11} & & & & & \\ & \ddots & & & \mathbf{0} & \\ & & R_{11} & & & \\ & & & \ddots & & \\ & & & & R_{pp} & \\ & \mathbf{0} & & & & \ddots \\ & & & & & & R_{pp} \end{pmatrix} \begin{matrix} \\ \\ \Big\} \; n \text{ lines} \\ \\ \\ \Big\} \; n \text{ lines} \end{matrix},$$

with R_{ii} the $i-$th diagonal component of the diagonal variance matrix R, and $\mathrm{MSE}(\mathbf{Z})$ is the block diagonal matrix

$$\mathrm{MSE}(\mathbf{Z}) \quad = \quad \begin{pmatrix} \mathrm{MSE}_1(\mathbf{Z}) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathrm{MSE}_p(\mathbf{Z}) \end{pmatrix} \begin{matrix} \Big\} \; n \text{ lines} \\ \\ \Big\} \; n \text{ lines} \end{matrix}$$

composed with the variance matrices $\mathrm{MSE}_j(\mathbf{Z}) \in \mathcal{M}^{n\times n}$ described as

$$\mathrm{MSE}_j(\mathbf{Z}) \quad = \quad \mathbb{E}\left((\mathcal{H}_j(\mathbf{Z}) - \widehat{H}_j(\mathbf{Z}))^2 \,|\, \mathbf{H}_{D_N}\right),$$

for $j = 1, \ldots, p$, where $\mathbf{Z}$ denotes the n sample points $\{Z_1, \ldots, Z_n\}$ with $Z_i = (X_i, d_i)$, $\mathcal{H}_j$ denotes the jth dimension of the Gaussian process $\mathcal{H}$ and $\mathbf{H}_{D_N}$ is the evaluations of the function H on the design D_N. Simulating this conditional distribution of $\mathbf{X}$ requires the Metropolis-Hastings (MH) step described in Appendix A.

Controlling the algorithmic error An important problem when running MCMC algorithms is monitoring the convergence of the simulated Markov chain in order to minimize the above mentioned *algorithmic error*. Actually, MCMC algorithms can converge slowly and stopping a simulated chain too early could lead to a poor approximation of the target distribution. Monitoring the convergence of a MCMC algorithm is also a difficult problem. Despite many efforts have been paid on this question, there is not an absolute way to answer it. We chose to use the much employed Brooks-Gelman (BG) statistics (Brooks et Gelman, 1998, (11)) computed from five replications of the Monte Carlo Markov chain (see Appendix C). The MCMC algorithm is stopped if the BG statistics is smaller than 1.05. We select this severe threshold of 1.05, instead of the more standard 1.2 value suggested in (11), to make sure that a reasonable approximation of the target distribution has been reached.

Measuring the emulator error However, a good monitoring of the MCMC algorithm could be jeopardized if the emulator $\widehat{H}$ is too far from the model H (the *emulator error*). Typically, the emulator error can be large if the number of points N of the design D_N is too small. Two much employed criteria to measure the quality of a design are experimented here.

i) The coefficient of destructibility Q_2 (see Vanderpoorten and Palm, 2001, (111)) is

$$Q_2 \quad = \quad 1 - \frac{\mathrm{PRESS}(D^*)}{\left\|H(D^*) - \overline{H}(D^*)\right\|^2}, \tag{IV.5}$$

with

$$\mathrm{PRESS}(D^*) \quad = \quad \left\|H(D^*) - \widehat{H}(D^*)\right\|^2$$

the Euclidean distance between the true function value H and the approximated value $\widehat{H}$ on a validation sample $D^* = \{v_{(1)}, \ldots, v_{(N^*)}\}$, $\overline{H}(D^*)$ denoting the mean function value on D^*:

$$\overline{H}(D^*) \quad = \quad \frac{1}{N^*}\sum_{i=1}^{N^*} H(v_{(i)}).$$

A cheaper version of Q_2 can be obtained by cross-validation, as follows (*leave one out* procedure):

$$Q_{2\,\mathrm{CV}} \quad = \quad 1 - \frac{\mathrm{PRESS}_{\mathrm{CV}}}{\sum_{i=1}^{N}\left\|H(z_{(i)}) - \overline{H}_{D_N}\right\|^2}. \tag{IV.6}$$

where

$$\overline{H}_{D_N} = \frac{1}{N}\sum_{i=1}^{N} H(z_{(i)}),$$

and

$$\text{PRESS}_{\text{CV}} = \sum_{i=1}^{N} e_{(i)}^2 = \sum_{i=1}^{N} \left\| H(z_{(i)}) - \widehat{H}_{-i}(z_{(i)}) \right\|^2$$

with

- $e_{(i)}$ is the prediction error at $z_{(i)}$ of a fitted model without the point $z_{(i)}$;
- $\widehat{H}_{-i}(z_{(i)})$ is the approximation of H at $z_{(i)}$ derived from all the points of the design except $z_{(i)}$.

Both versions of Q_2 are related to the ratio of variance explained by an emulator. The closer Q_2 to 1, the smaller this ratio is and the better the quality of the design D_N is.

ii) An alternative criterion is the Mahalanobis distance (MD) (see Bastos and O'Hagan 2009, (5)), computed on a validation sample D^* with N^* points as follows:

$$\text{MD} = \Big(H(D^*) - \widehat{H}(D^*)\Big)' \Big(\text{MSE}(D^*)\Big)^{-1} \Big(H(D^*) - \widehat{H}(D^*)\Big), \qquad \text{(IV.7)}$$

where $\text{MSE}(D^*)$ (Mean Squared Error) is the conditional variance matrix of the design D^* knowing $H_{D^*} = \{H(v_{(1)}), \ldots, H(v_{(N^*)})\}$. An interest of this criterion is to take into account the correlations between the points through the $\text{MSE}(D^*)$ term. Obviously, the MD value is sensitive to the choice of D^*. D^* could be generated as a *maximin* LHD. A cheaper cross-validated version of MD is as follows:

$$\text{MD}_{\text{CV}} = \frac{1}{N}\sum_{i=1}^{N} \Big(H(z_{(i)}) - \widehat{H}_{-i}(z_{(i)})\Big)' \Big(\text{MSE}_{-i}(z_{(i)})\Big)^{-1} \Big(H(z_{(i)}) - \widehat{H}_{-i}(z_{(i)})\Big),$$

where $\widehat{H}_{-i}(z_{(i)})$ denotes the predictor of H at point $z_{(i)}$ by using the design $D_{-i} = \{z_{(1)}, \ldots, z_{(i-1)}, z_{(i+1)}, \ldots, z_{(N)}\}$ and $\text{MSE}_{-i}(z_{(i)})$ denotes the related squared error.

Now, the smaller the sample size n, the greater the *estimation error* is. The two above mentioned criteria are not aiming to measure the *estimation error*. But since H is complex, it is quite difficult to assess this error in an inverse modeling context. Bayesian inference could be expected to be helpful to reduce the estimation error when n is small and when reliable prior information is available. However, if the prior information is not relevant, the *prior error* will be large and Bayesian inference could be harmful. For this very reason, it is important to be able to measure the relevance of the prior information. In the present context, it is possible to use a promising criterion, the so-called DAC criterion (Bousquet 2008, (10)) for this task, as detailed in the next section.

IV.3 Assessing a prior distribution and a design

IV.3.1 The DAC criterion

The DAC criterion (Bousquet 2008, (10)) has been conceived as a measure of the discrepancy between a prior distribution of a model parameters and the data. Let $\mathbf{y}$ be a sample with pdf $f(\mathbf{y}|\theta)$. Let $\pi^J(\theta)$ be a benchmark non-informative prior (see for instance Yang and Berger 1998, (121)) and $\pi(\theta)$ the prior distribution derived from the prior information on θ. DAC is

$$\text{DAC}(\pi|\mathbf{y}) = \frac{\text{KL}(\pi^J(\theta|\mathbf{y})||\pi(\theta))}{\text{KL}(\pi^J(\theta|\mathbf{y})||\pi^J(\theta))}, \tag{IV.8}$$

where $\text{KL}(p||q)$ is denoting the Kullback-Leibler distance between the probability distributions p and q, which is defined as

$$\text{KL}(p||q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}\, dx, \tag{IV.9}$$

$\mathcal{X}$ being the set of all accessible values for x. The rationale underlying DAC criterion is as follows: the posterior distribution $\pi^J(\theta|\mathbf{y})$ derived from the non-informative prior can be regarded as an ideal prior distribution on θ in perfect agreement with the data $\mathbf{y}$. Thus, $\text{KL}(\pi^J(\theta|\mathbf{y})||\pi(\theta))$ is measuring the distance between the prior π to be assessed and the ideal prior $\pi^J(\cdot|\mathbf{y})$.

If $\text{DAC}(\pi|\mathbf{y}) \leq 1$, the informative prior π is closer to the ideal prior than the non-informative prior π^J, and the data $\mathbf{y}$ and the prior $\pi(\theta)$ are declared to be in agreement. Otherwise if $\text{DAC}(\pi|\mathbf{y}) > 1$, the data $\mathbf{y}$ and the prior $\pi(\theta)$ are declared to be discrepant. DAC has been proved to be efficient when the non-informative prior $\pi^J(\theta)$ is proper (see Bousquet 2008, (10)).

IV.3.2 The impact of the emulator

In the present context, a kriging emulator defined on a compact set Ω is used to compute an approximation of the posterior distribution of the parameter $\theta = (m, C)$. Since the emulator is defined on a compact set, the parameters m and C are also restricted to be in compact sets Ω_m and Ω_C. It allows us to define a proper non-informative prior $\pi^J(m, C)$ (chosen as the Jeffreys prior for the multivariate Gaussian model), then a tractable DAC. The technical precisions about Ω_m, Ω_C and the calculation of DAC are provided in Appendices D and E.

It is important to remark that the DAC criterion is depending on the design D_N. Denoting $\pi^J(\theta|\mathbf{y}, D_N)$ the posterior distribution of θ given the data $\mathbf{y}$ and the current design $D_N = \{z_{(1)}, \ldots, z_{(N)}\}$,

$$\text{DAC}(\pi|\mathbf{y}, \mathbf{H}_{D_N}) = \frac{\text{KL}\left(\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})||\pi(\theta)\right)}{\text{KL}\left(\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})||\pi^J(\theta)\right)}.$$

A DAC value greater than one is just indicating that there is something misleading between the data, the prior and the design. Thus, if the data and the prior are known (or assumed) to be relevant, DAC could regarded as a criterion to assess the design as Q_2 or MD.

IV.3.3 Computing DAC

Since

$$\frac{S}{T} \leq 1 \Longleftrightarrow S - T \leq 0, \text{ if } \geq 0, T > 0, \tag{IV.10}$$

a numerically more convenient version of DAC, denoted by $\widetilde{\text{DAC}}$ is

$$\widetilde{\text{DAC}}(\pi|\mathbf{y}, \mathbf{H}_{D_N}) = \text{KL}\left(\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})||\pi(\theta)\right) - \text{KL}\left(\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})||\pi^J(\theta)\right).$$

The critical value for $\widetilde{\text{DAC}}$ is 0. Since the support of $\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})$ is Ω, we have

$$\begin{aligned}\text{KL}\left(\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})||\pi(\theta)\right) &= \int_\Omega \pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N}) \log \frac{\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})}{\pi(\theta)}\, d\theta \\ &= \mathbb{E}_{\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})}\left[\log \pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})\right] - \mathbb{E}_{\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})}\left[\log \pi(\theta)\right],\end{aligned}$$

and

$$\begin{aligned}\text{KL}\left(\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})||\pi^J(\theta)\right) &= \int_\Omega \pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N}) \log \frac{\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})}{\pi^J(\theta)}\, d\theta \\ &= \mathbb{E}_{\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})}\left[\log \pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})\right] - \mathbb{E}_{\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})}\left[\log \pi^J(\theta)\right].\end{aligned}$$

Therefore, the transformed $\widetilde{\text{DAC}}$ can be written as:

$$\begin{aligned}\widetilde{\text{DAC}}(\pi|\mathbf{y}, \mathbf{H}_{D_N}) &= \text{KL}\left(\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})||\pi(\theta)\right) - \text{KL}\left(\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})||\pi^J(\theta)\right) \\ &= \mathbb{E}_{\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})}\left[\log \pi^J(\theta)\right] - \mathbb{E}_{\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})}\left[\log \pi(\theta)\right],\end{aligned}$$

and the $\widetilde{\text{DAC}}$ criterion can be computed using the outputs of a Gibbs sampler run with a non-informative prior $\pi^J(\cdot)$ (In practice, we chose a Jeffreys non-informative prior.)

$$\widetilde{\text{DAC}}(\pi|\mathbf{y}, \mathbf{H}_{D_N}) \backsimeq \frac{1}{R}\sum_{r=1}^{R} \log \pi^J(\theta^r) - \frac{1}{R}\sum_{r=1}^{R} \log \pi(\theta^r), \tag{IV.11}$$

where $\theta^r \sim \pi^J(\cdot|\mathbf{y}, \mathbf{H}_{D_N})$, $r \in \{1, ..., R\}$ is a simulated sequence obtained by Gibbs sampling. For the purpose of simplicity, in the following we use the notation

$$\widetilde{\text{DAC}}_N := \widetilde{\text{DAC}}(\pi|\mathbf{y}, \mathbf{H}_{D_N}).$$

$\widetilde{\text{DAC}}_N \leq 0$ means the prior distribution $\pi(\theta)$ and the couple $(\mathbf{y}, \mathbf{H}_{D_N})$ are declared compatible. Now, computing DAC criterion requires to run an additional Gibbs sampler with the non-informative prior distribution. Denoting $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, the full conditional distribution of m verifies

$$\pi^J(m \,|\, C, \mathbf{Y}, \mathbf{X}, \rho, H_D) \propto \mathbf{I}_{\Omega_m} \exp\Big[-\frac{1}{2}(m - \overline{X}_n)' \left(\frac{C}{n}\right)^{-1} (m - \overline{X}_n)\Big].$$

Thus, it is a normal distribution truncated on Ω_m: $\mathbf{I}_{\Omega_m} \cdot \mathcal{N}\left(\overline{X}_n, \frac{C}{n}\right)$. The full conditional distribution of the variance matrix C verifies

$$\pi^J(C \,|\, m, \mathbf{Y}, \mathbf{X}, \rho, H_D) \quad \propto \quad \mathbf{I}_{\Omega_C} \, |C|^{-\frac{n+q+2}{2}} \, \exp\Big[-\frac{1}{2}\mathrm{Tr}\left(n\,(m-\overline{X}_n)(m-\overline{X}_n)' \cdot C^{-1}\right)\Big]. \tag{IV.12}$$

Thus it is an Inverse-Wishart distribution truncated on Ω_C:

$$\mathbf{I}_{\Omega_C} \cdot \mathcal{IW}\left(n\,(m-\overline{X}_n)(m-\overline{X}_n)', n+1\right). \tag{IV.13}$$

Moreover, the full conditional distribution of the missing data $\mathbf{X}$ given the current parameters θ, the observation $\mathbf{y}, \mathbf{d}$ and the evaluations $\mathbf{H}_{D_N}$ is given by (IV.4).

Using those full conditional posterior distributions, the Gibbs sampler approximating the posterior distribution of (m, C) with a non-informative prior truncated to the domain $\Omega_m \times \Omega_C$ could be straightforwardly described. Obviously, it incorporates the MH step presented in Appendix A to simulate the missing data $\mathbf{X}$.

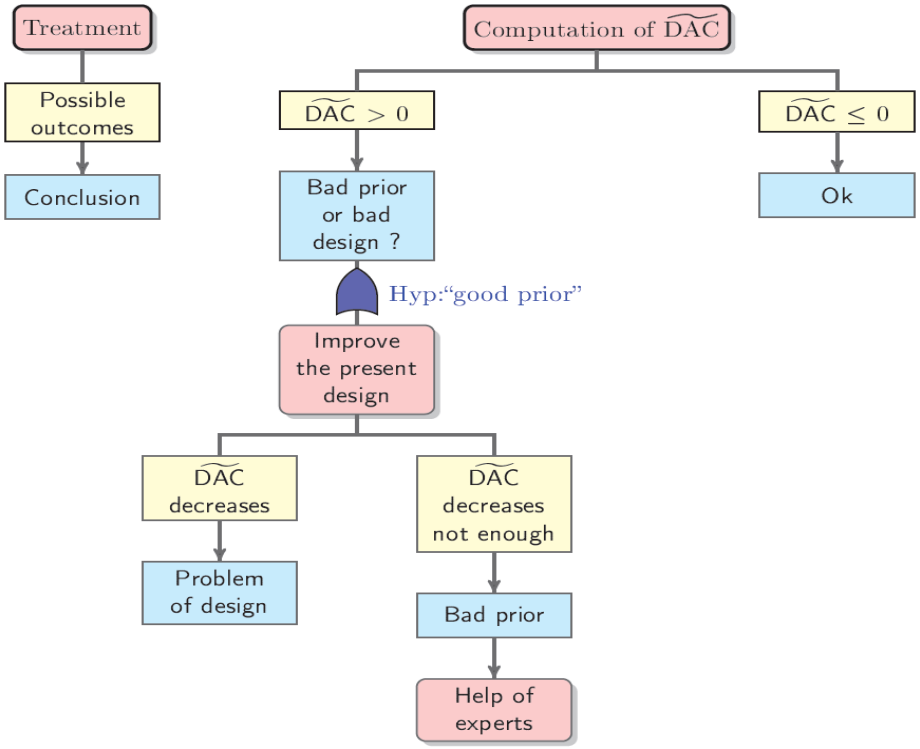
Remark: The simulation of C is difficult since $n\,(m-\overline{X}_n)(m-\overline{X}_n)'$ is not a definite but a semi-definite positive matrix and numerical problems can occur. However, up to an additive constant, the calculation (IV.12) is proper. For this reason, we recommend to use a Metropolis-Hastings algorithm for simulating C, as presented in Appendix B.

IV.3.4 Using the $\widetilde{\text{DAC}}$ criterion

By its very nature, the criterion $\widetilde{\text{DAC}}$ is measuring the agreement between the observed data and the prior distribution. As shown above, it could be computed without particular difficulties, despite it needs to run an additional Gibbs sampler, when the distribution H has been replaced by a kriging emulator $\widehat{H}$. Thus $\widetilde{\text{DAC}}$ is depending on the prior distribution and the design D_N. Hence $\widetilde{\text{DAC}}$ is a criterion allowing to assess both the prior and design relevances with respect to the observed data y. But this double assessment has to be done properly using the following procedure:

1. If $\widetilde{\text{DAC}} \leq 0$ then the prior and the design are declared to be acceptable.
2. If $\widetilde{\text{DAC}} > 0$, the following step is required:
   under a ''good prior'' assumption, efforts are paid to improve the design by increasing N or modifying Ω. If $\widetilde{\text{DAC}}$ is not decreasing under zero, it means that the prior information is questionable and there is the need to go back to the experts.

This procedure is depicted by the following diagram:



IV.4 Numerical experiments

In order to illustrate the behavior of the the above mentioned criteria, numerical experiments are performed from simulated data on a simplified version of a hydraulic model $Y = H(X, d) + U$ partly used in (74) where

$$H(X,d) = \left(X_2 + \left(\frac{\sqrt{5000}}{300\sqrt{55 - X_2}} \times \frac{d}{X_1} \right)^{0.6}, \frac{d^{0.4} X_1^{0.6} (55 - X_2)^{0.3}}{300^{0.4} \times 5000^{0.3}} \right),$$

with

$$\begin{aligned} X &\sim \mathcal{N}\left(\begin{pmatrix} 30 \\ 50 \end{pmatrix}, \begin{pmatrix} 5^2 & 0 \\ 0 & 1 \end{pmatrix} \right), \\ d &\sim \text{Gumbel}\Big(1013, -458\Big), \end{aligned}$$

and the error term $U \sim \mathcal{N}(\mathbf{0}, 10^{-5} \cdot I_2)$.

Since we are mainly concerned in analyzing the behavior of $\widetilde{\text{DAC}}$, six different prior distributions on the model parameters were considered. They are summarized in Table IV.1. Remind that the prior distributions or the parameters m and C are $m|C \sim \mathcal{N}(\mu, C/a)$ and $C \sim \mathcal{IW}(\Lambda, \nu)$ with $\Lambda = t \cdot \widetilde{C}_{\text{Exp}}$.

Prior	**PLV**			**PMV**	**PHV**	**FHV**			**BMV**	**BHV**
μ	$\{30, 50\}$			$\{30, 50\}$	$\{30, 50\}$	$\{35, 49\}$			$\{10, 54\}$	$\{10, 54\}$
a	1	10	10	1	1	1	5	10	1	1
t	2	2	30	2	2	2			2	2
ν	5	5	33	5	5	5			5	5
$\widetilde{\mathbf{C}}_{\text{Exp}}$	$\begin{pmatrix} 1.5^2 & 0 \\ 0 & 1 \end{pmatrix}$			$\begin{pmatrix} 5^2 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 7.5^2 & 0 \\ 0 & 1.5^2 \end{pmatrix}$	$\begin{pmatrix} 7.5^2 & 0 \\ 0 & 1.5^2 \end{pmatrix}$			$\begin{pmatrix} 5^2 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 7.5^2 & 0 \\ 0 & 1.5^2 \end{pmatrix}$

Table IV.1: Description of the six prior distributions: **PLV** = perfect mean and low variance, **PMV** = perfect mean and medium variance, **PHV** = perfect mean and high variance, **FHV** = fair mean and high variance, **BMV** = bad mean and medium variance, **BHV** = bad mean and high variance.

IV.4.1 Assessing the design

The first experiments are aiming to assess the ability of criteria Q_2 and MD to measure the quality of a design. In this purpose three different designs with 20 points, 100 points and 500 points have been considered on two different domains

$$\begin{aligned} \Omega_1 &= [25.1001, 34.8999] \times [48.0400, 51.9600] \times [40, 1800] \\ \Omega_2 &= [20, 40] \times [45, 55] \times [\min_i(d_i), \max_i(d_i)]. \end{aligned}$$

Ω_1 can be thought of as a realistic domain and Ω_2 is a larger domain. When using a validation sample D^*, we choose it as a *maximin* LHD with 100 points. Figures IV.1 and IV.2 give the box plots of $1 - Q_2$ based on 20 repetitions computed on a validation sample and by cross-validation respectively. The closer one and Q_2 are, the better the design is supposed to be. The observed differences on $1-Q_2$ according to the designs are relevant but hardly perceptible as even a small design of 20 points on the large domain Ω_2 produces small $1 - Q_2$ values. The difficulty with criterion Q_2 is to choose a sensible threshold to declare that a design is acceptable.

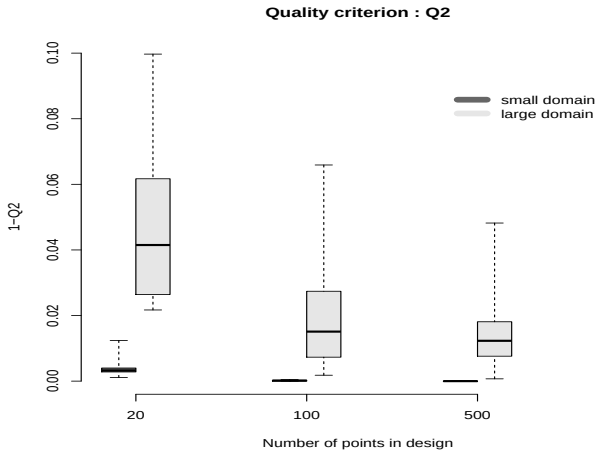


Figure IV.1: $1 - Q_2$ box plots calculated on a validation sample for six designs

Figures IV.3 and IV.4 display the box plots of log(MD) in the same conditions. As it could

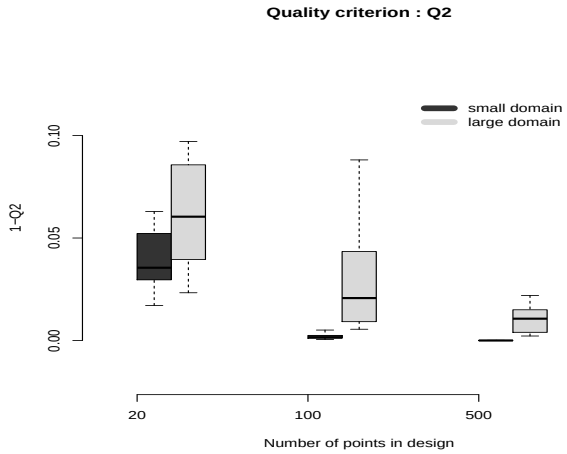


Figure IV.2: $1 - Q_2$ box plots calculated by cross-validation for six designs

be expected, this criterion is decreasing when the number of design points increases. The cross-validated MD does not seem very sensitive for the domain Ω_1 and the cross-validated MD values for the larger domain with a design of 500 points are amazing (see Figure IV.4). Moreover, contrary to Q_2 criterion, no reference value is available with MD and it seems difficult to use this more expensive criterion to assess a design (see Figure IV.3).

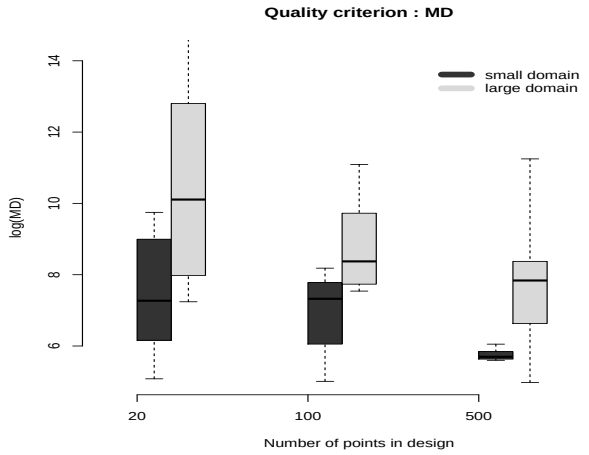


Figure IV.3: MD box plots calculated on a validation sample for six designs

IV.4.2 Assessing the prior and the design

The following numerical experiments aim to analyze the ability of $\widetilde{\text{DAC}}$ to assess either a design or a prior distribution.

Figures IV.5 and IV.6 depict the behavior of $\widetilde{\text{DAC}}$ for the domains Ω_1 and Ω_2, for 20 repetitions of the model with the six prior distributions and *maximin* LHDs with 20, 100 and 500 points. From those figures, it appears that the "bad" priors are discarded by $\widetilde{\text{DAC}}$ in all cases, but for the other priors even the design with 20 points seems acceptable. Obviously, for

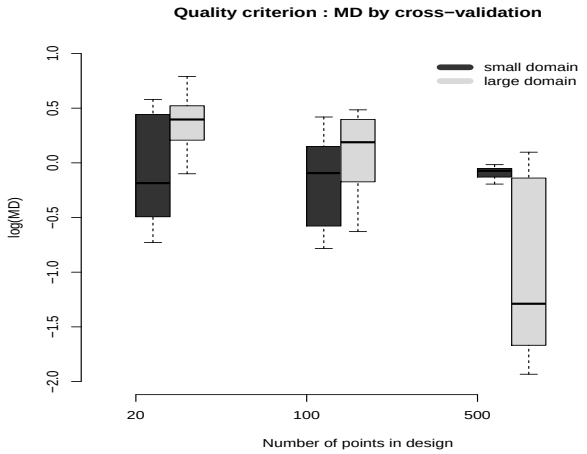


Figure IV.4: MD box plots calculated by cross-validation for six designs

this poor design the Gibbs sampler converges dramatically slower (2 000 iterations for D_{500} and 100 000 iterations for D_{20}), but in many situations it is not an issue. Actually, the main computational burden is computing the highly CPU-time demanding physical model H. In the present context, running a Gibbs sampler with a design of N points require N calls to the function H and it could be faster to run a Gibbs sampler on a D_{20} for 100 000 iterations than a Gibbs sampler with a D_{500} for 2 000 iterations... Moreover, the similar behavior of $\widetilde{\text{DAC}}$ in Figures IV.5 and IV.6 for both domains shows that the domain choice does not affect the agreement between the prior and the data. For this reason, we only reported the next experiments for the small domain Ω_1.

Figure IV.7 which displays the behavior of $\widetilde{\text{DAC}}$ for the PLV and FHV prior with different hyperparameters a and t values shows that those hyperparameters are sensitive and that too concentrated priors (related to large values of a and t) could lead to a doubtful Bayesian inference. For example, for the PLV prior increasing the value of a, which weights the prior mean μ, does not much change the value of $\widetilde{\text{DAC}}$ as μ is equal to the actual mean m; while for the FHV prior, a larger a results in a larger $\widetilde{\text{DAC}}$ value as in this "fair" case, μ and the actual mean m are different.

Figures IV.8 displays the marginal posterior distributions with a LHD-*maximum* design of 100 points and Figure IV.9 with a LHD-*maximum* design of 20 points. Those figures confirm the $\widetilde{\text{DAC}}$ diagnosis. There are great differences between the posteriors derived from "bad priors" and the other ones, including the posterior derived from the Jeffreys prior, which are quite similar. It is also important to notice than there is no sensitive differences between the posteriors derived from the 100 points and 20 points designs as indicated by the $\widetilde{\text{DAC}}$ criterion.

It seems that $\widetilde{\text{DAC}}$ is indicating that a reasonable prior can resist to a poor design. It is not always true. For instance a poor design of 18 randomly generated points on the faces of a cube (Three points were generated on each face.) has been considered with the same model

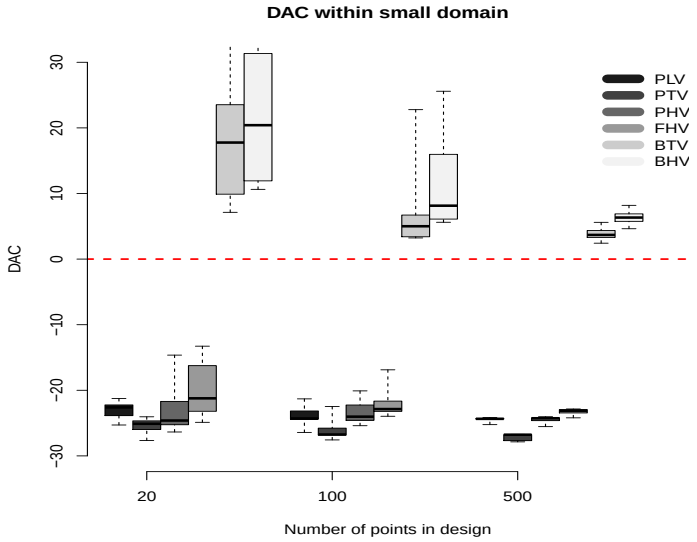


Figure IV.5: $\widetilde{\text{DAC}}$ in small domain with six priors and three designs

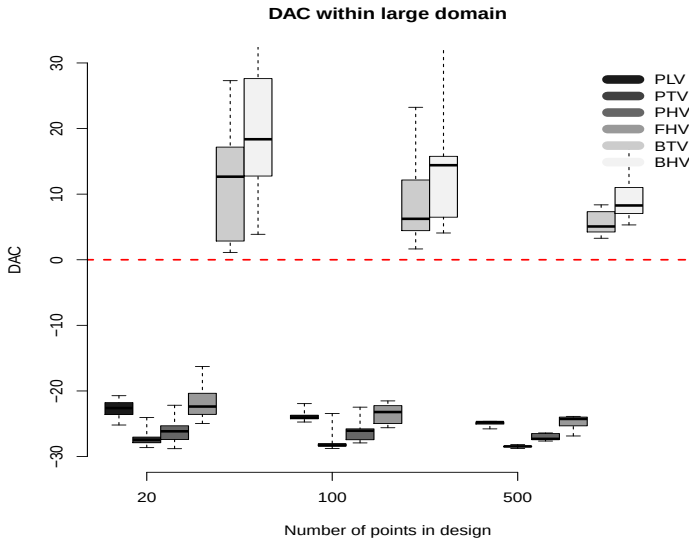


Figure IV.6: $\widetilde{\text{DAC}}$ in large domain with six priors and three designs

$Y = H(X, d) + U$ but with the Sobol function H

$$H(X,d) \quad = \quad \prod_{k=1}^{2} g_k\left(|\sin(X_k)|\right) g_3\left(|\sin(d)|\right), \text{ where } g_k(x) = \frac{|4x-2| + a_k}{1 + a_k},$$

with $a_k = 1$. A Gibbs sampler of 800 000 runs has been run to estimate the posterior distribution $\pi^J(\theta|\mathbf{y}, \mathbf{H}_{D_N})$. As shown in the left graph of Figure IV.10, $\widetilde{\text{DAC}}_{18}$ remains positive for the four prior choices indicating the need to improve the design.

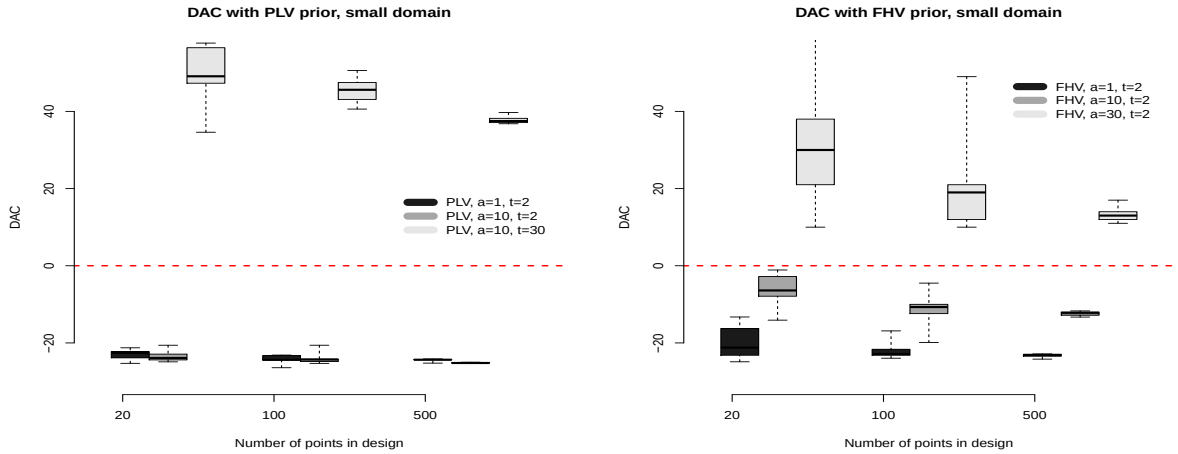


Figure IV.7: $\widetilde{\text{DAC}}$ with **PLV** and **FHV** priors for the small domain Ω_1 and different values of the hyperparameters a and t

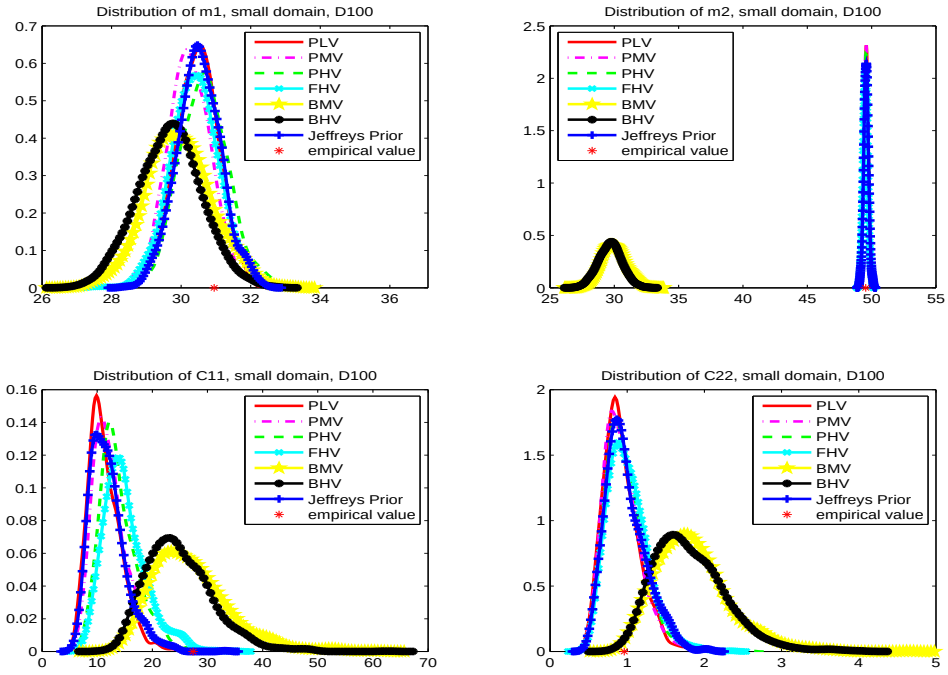


Figure IV.8: Posterior of θ with help of D_{100} in small domain Ω_1

IV.5 Discussion

We have shown that Bayesian analysis was possible and beneficial to solve inverse problems by estimating the parameters of highly complex uncertainty models. Bayesian analysis is possible thanks to MCMC algorithms such as Gibbs sampling and the approximation of the

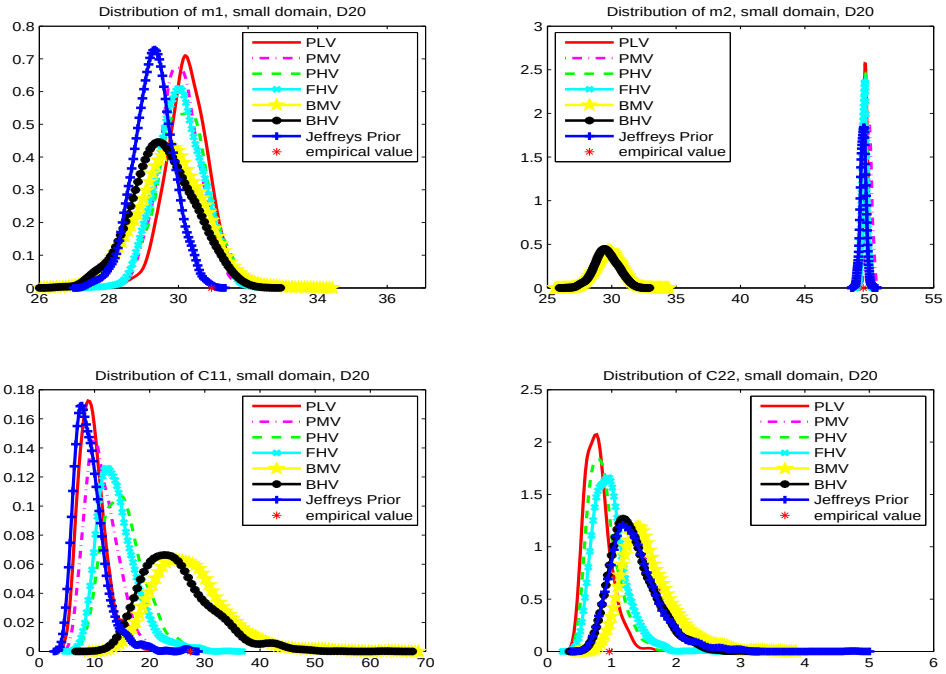


Figure IV.9: Posterior of θ with help of D_{20} in small domain Ω_1

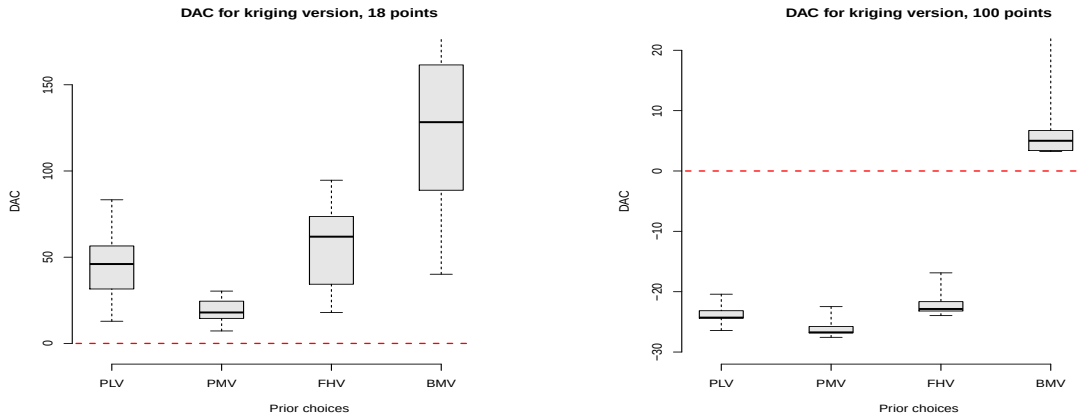


Figure IV.10: $\widetilde{\text{DAC}}$ for D_{18} and D_{100} with the Sobol function

physical model by a kriging emulator using a *maximin* LHD. Bayesian analysis is beneficial since it allows to take into account properly prior knowledge and avoids linearization of the physical model H. Our analysis has shown that Bayesian inference could be beneficial because MCMC algorithms could be hoped to be rapid even with a *maximin* LHD with few points in comparison to the huge time needed to compute H. From this point of view, it could be helpful to translate the time to get a realization of H as a number of iterations of the MCMC algorithm in order to choose the number of points of the emulator's design. Let us suppose that the computation time of one call to H equals the computation time of $L(N)$

iterations of the MCMC algorithm[1]. The integer $L(N)$ is expected to be quite large and is a decreasing function of the number N of points of the design D_N which is as well the number of "possible" calls to H. Our analysis proved that even when N is small, it is possible to increase the number of iterations of the MCMC algorithm to get a good approximation of the model parameter posterior distribution in an acceptable CPU time. For instance, with the hydraulic model, the CPU time (in seconds) has been 999 for $N = 500$, 1930 for $N = 100$ and 10 100 for $N = 20$ on a laptop PC, with two Intel P9700 cores of 2.80GHz

In this perspective, there is the need to control the four error sources listed in the introduction.

- By its very nature, Bayesian inference is helpful to control the *estimation error* when the number n of observations is small.
- The *algorithmic error* can be efficiently controlled with the BG statistics. To make sure that this error is not too big, we advocate a more severe threshold value 1.05 than the standard threshold 1.2.
- We propose to use the so-called $\widetilde{\text{DAC}}$ criterion which could be thought of as a relevant measure of the discrepancy between the observed sample and the prior distribution in order to control both the *emulator error* and the *prior error*. In our context, this criterion can be computed without major difficulties: the emulator is defined on a compact set and, consequently, proper non-informative priors are available. Our experiments show a promising behavior of this criterion. Obviously, computing $\widetilde{\text{DAC}}$ is not free since it involves to run an additional MCMC algorithm for non-informative priors. But we think that the result is worth the trouble. Moreover, as soon as the MCMC with a non-informative prior has been run, any informative prior can be assessed. On the other hand when $\widetilde{\text{DAC}}$ is greater than zero, it could be difficult to separate the *emulator* and the *prior* errors since both errors could be quite intricate. More experiments are needed to assess the relevance and sensibility of this criterion. But we think that it is a promising tool to drive Bayesian inference using an emulator for dealing with complex inverse problems in uncertainty analysis.

Finally, the conclusion of this study can be stated as follows. When the prior knowledge on the model parameters is relevant, Gibbs sampling or other MCMC algorithms on an appropriate emulator could be expected to lead to a sensible estimation of these parameters with well calibrated prior distributions while dramatically saving the number of calls to the expensive function H. And, the criterion $\widetilde{\text{DAC}}$ could be expected to be helpful to honestly calibrate the prior distributions and choose a good design for the emulator.

[1] Recall that N is the total allowed number of calls to H and it is also the number of points of the emulator design.

Appendix A. the Gibbs sampling with a Jeffreys non informative prior

Given $(m^{[r]}, C^{[r]}, \mathbf{X}^{[r]})$ for $r = 0, 1, 2, \ldots,$ generate

1. $C^{[r+1]} \,|\, m^{[r]}, \mathbf{X}^{[r]}, \boldsymbol{\mathcal{Y}}, \rho, \mathbf{H}_{D_N} \sim \mathcal{IW}\Big(\sum_{i=1}^{n} (m^{[r]} - X_i^{[r]})(m^{[r]} - X_i^{[r]})',\, n+1 \Big) \cdot \mathbb{1}_{\Omega_C}$;

2. $m^{[r+1]} \,|\, C^{[r+1]}, \mathbf{X}^{[r]}, \boldsymbol{\mathcal{Y}}, \rho, \mathbf{H}_{D_N} \sim \mathcal{N}\Big(\overline{\mathbf{X}^{[r]}},\, \frac{C^{[r+1]}}{n} \Big) \cdot \mathbb{1}_{\Omega_m}$,
   with $\overline{\mathbf{X}^{[r]}} = \sum_{i=1}^{n} X_i^{[r]}$;

3. $\mathbf{X}^{[r+1]} \,|\, m^{[r+1]}, C^{[r+1]}, \boldsymbol{\mathcal{Y}}, \rho, \mathbf{H}_{D_N} \Rightarrow$ Metropolis-Hastings algorithm
   More precisely, this full conditional posterior distribution is proportional to

$$|\mathbf{R} + \mathrm{MSE}^{[r+1]}|^{-\frac{1}{2}} \cdot \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (X_i^{[r+1]} - m^{[r+1]})' \Big[C^{[r+1]} \Big]^{-1} (X_i^{[r+1]} - m^{[r+1]}) \right.$$

$$\left. - \frac{1}{2} \Big(\Big(\mathcal{Y}_1 - \widehat{H}_{N,1}^{[r+1]} \Big)', \ldots, \Big(\mathcal{Y}_n - \widehat{H}_{N,n}^{[r+1]} \Big)' \Big) \Big(\mathbf{R} + \mathrm{MSE}^{[r+1]} \Big)^{-1} \begin{pmatrix} \mathcal{Y}_1 - \widehat{H}_{N,1}^{[r+1]} \\ \vdots \\ \mathcal{Y}_n - \widehat{H}_{N,n}^{[r+1]} \end{pmatrix} \right\},$$

   with $\widehat{H}_{N,i}^{[r+1]} = \widehat{H}_N(X_i^{[r+1]}, d)$ and $\mathrm{MSE}^{[r+1]} = \mathrm{MSE}(\mathbf{X}^{[r+1]}, d)$.

The Metropolis-Hastings step inside the Gibbs sampler At step $r+1$ of Gibbs sampling, after simulating $m^{[r+1]}$,$C^{[r+1]}$, the missing data $\mathbf{X}^{[r+1]}$ have to be updated with a Metropolis-Hasting (MH) algorithm. The MH step is updating $\mathbf{X}^{[r]} = (X_1^r, \ldots, X_n^r)'$ in the following way:

- For $i = 1, \ldots, n$
  1. Generate $\widetilde{X}_i \sim J(\cdot \mid X_i^r)$ where J is the proposal distribution.
  2. Let

$$\alpha(X_i^r, \widetilde{X}_i) \;=\; \min\Big(\frac{\pi_{\widehat{H}}(\widetilde{\mathbf{X}} \mid \boldsymbol{\mathcal{Y}}, \theta^{[r+1]}, \rho, \mathbf{d}, H_D)\, J(X_i^r | \widetilde{X}_i)}{\pi_{\widehat{H}}(\mathbf{X}^{[r]} \mid \boldsymbol{\mathcal{Y}}, \theta^{[r+1]}, \rho, \mathbf{d}, H_D)\, J(\widetilde{X}_i | X_i^r)}, 1 \Big),$$

     where

$$\begin{aligned} \widetilde{\mathbf{X}} &= \Big(X_1^{r+1}, \ldots, X_{i-1}^{r+1}, \widetilde{X}_i, X_{i+1}^r, \ldots, X_n^r \Big)' \\ \mathbf{X}^{[r]} &= \Big(X_1^r, \ldots, X_{i-1}^r, X_i^r, X_{i+1}^r, \ldots, X_n^r \Big)' \end{aligned}$$

  3. Take

$$X_i^{r+1} \;=\; \begin{cases} \widetilde{X}_i & \text{with probability } \alpha(X_i^r, \widetilde{X}_i), \\ X_i^{r+1} & \text{otherwise.} \end{cases}$$

Remarks:

- Many choices are possible for the proposal distribution J. It appears that choosing an independent MH sampler with J chosen to be the normal distribution $\mathcal{N}\left(m^{[r+1]}, C^{[r+1]}\right)$ give satisfying results for the model (IV.1).
- In practice, it can be beneficial to choose the order of the updates by a random permutation of $\{1, \ldots, n\}$ to accelerate the convergence of the Markov chain to its limit distribution.

Appendix B. Metropolis-Hasting (MH) algorithm for simulating C

1. `Iteration 0: Choose an arbitrary value` $C^{[0]} = C_0$
2. `Iteration` h: `Update` $C^{[h]}$ `as follows:`
   - `Generate` ξ `from the following proposal distribution` f^*, `which is adding a small correction` ϵI_q `to the semi-positive definite matrix` $(m-\overline{X}_n)(m-\overline{X}_n)'$

$$f^*(\xi) \quad = \quad \mathbf{I}_{\Omega_C}(\xi) \cdot \mathcal{IW}\left(n\,(m-\overline{X}_n)(m-\overline{X}_n)' + \epsilon I_q, n+1\right).$$

   - `Let`

$$\alpha(C^{[h-1]}, \xi) \quad = \quad \frac{g(\xi) f^*(C^{[h-1]})}{g(C^{[h-1]}) f^*(\xi)} \wedge 1, \tag{IV.14}$$

   `with` g `proportional to the target distribution which means the truncated Inverse-Wishart distribution` (IV.13)

$$g(C) \quad = \quad \mathbf{I}_{\Omega_C}(C) \cdot |C|^{-\frac{n+q+2}{2}} \exp\left[-\frac{1}{2}\mathrm{Tr}\left(n\,(m-\overline{X}_n)(m-\overline{X}_n)' \cdot C^{-1}\right)\right].$$

   - `Choose` $C^{[h]}$ `as follows`

$$C^{[h]} \quad = \quad \begin{cases} \xi & \text{with probability } \alpha(C^{[h-1]}, \xi), \\ C^{[h-1]} & \text{otherwise.} \end{cases} \tag{IV.15}$$

In this way, we obtain a Markov chain $(C^{[h]})$ which converges to the distribution (IV.13).

Appendix C. Brooks-Gelman Statistics

In 1998, Brooks and Gelman proposed a method derived from the method proposed by Gelman and Rubin (1992a), for monitoring the convergence of iterative simulations ((11)). Supposing m parallel chains have been simulated, the statistic $\widehat{R}_{BG}$ is constructed on the final M iterations after the "burn-in" period, as follows:

1. For each individual chain j, calculate the empirical $100(1-\alpha)\%$ interval δ_j, which is the difference between the $100(1-\frac{\alpha}{2})\%$ and $100\frac{\alpha}{2}\%$ percentile of the M simulated points. Thus, form the m within-sequence interval length estimates.

2. For the entire set of mM simulated draws from all chains, calculate the empirical $100(1-\alpha)\%$ interval to construct a total-sequence interval length estimate.

3. Evaluate the statistic $\widehat{R}_{BG}$ defined as

$$\widehat{R}_{BG} = \frac{\Delta}{\bar{\delta}}$$

- Δ the total-sequence interval length;
- $\bar{\delta} = \frac{1}{m}\sum_{j=1}^{m}\theta_j$, with θ_j the length of the within-sequence interval for the j-th chain.

The threshold value 1.2 is advocated by the authors ($\widehat{R}_{BG} < 1.2$) to declare that the simulation procedure has converged. In our experiments, we make use of a more conservative threshold and procedure to ensure that the MCMC algorithms have converged to their stationary distribution. A MCMC chain has been declared to have converged if the $\widehat{R}_{BG}$ statistics is smaller than 1.05 for 3,000 iterations.

Appendix D. Computing $\widetilde{DAC}$ for the kriging emulator

The compact set $\Omega_m = \Omega = \Omega_1 \times \ldots \times \Omega_q$ where Ω_i denotes the domain for the i-th coordinate of X. To determine the compact set Ω_C related to the variance matrix C, it is convenient to consider its eigenvalue decomposition $C = VDV^T$ where D is the diagonal matrix of eigenvalues of C with $|C| = |D|$ and V the orthogonal matrix of eigenvectors of C. For each dimension $i = 1, \ldots, q$, $X_i^2 \leq \beta_i = \max\left(\max \Omega_i)^2, \min \Omega_i)^2\right)$. On the other hand, recalling that R is the variance matrix of the measurement error in model (IV.1), it is reasonable to assume that the measurement error is smaller than the variance and thus $|R|^{1/p} \leq |C|^{1/q} = |D|^{1/q}$. Finally, the domain of variance Ω_C can be defined as follows:

$$\Omega_C = \left\{ C = VDV^T \in \mathcal{S}_q^+ \ st.\ |D| \geq |R|^{q/p},\ 0 \leq D_{ii} \leq \sqrt{\sum_{j=1}^{q} \beta_i^2},\ i = 1 \ldots, q \right\} \quad \text{(IV.16)}$$

where $\mathcal{S}_q^+$ is the set of symmetric positive definite matrices of rank q.

The benchmark prior $\pi^J(\theta)$ is chosen here as the Jeffreys prior for a multivariate Gaussian distribution restricted to Ω_m, ie.

$$\pi^J(\theta) = \frac{\mathbf{I}_{\Omega_m}(m)}{\mathrm{Vol}(\Omega_m)} \cdot \frac{\Delta_C}{|C|^{\frac{q+2}{2}}} \mathbf{I}_{\Omega_C}(C) \quad \text{(IV.17)}$$

with

$$\Delta_C = \left(\int_{\Omega_C} \frac{1}{|C|^{\frac{q+2}{2}}} dC \right)^{-1}.$$

Thus

$$\begin{aligned}\Delta_C^{-1} &= \int_{\Omega_C} \frac{1}{|C|^{\frac{q+2}{2}}} dC \\ &= \int_{\Omega_C} \frac{1}{|D|^{\frac{q+2}{2}}} d\left(VDV^T\right) \\ &= \int dV \left[\int_{\Omega_D} \frac{1}{|D|^{\frac{q+2}{2}}} dD\right],\end{aligned}$$

where

$$\Omega_D = \left\{D \in \mathcal{DS}_q^+ \ st. \ |D| \geq |R|^{q/p}, \ 0 \leq D_{ii} \leq \sqrt{\sum_{j=1}^{q} \beta_j^2}, \ i = 1 \ldots, q\right\}. \quad \text{(IV.18)}$$

Now, any orthogonal matrix V of dimension q is characterized by the composition of $q(q-1)/2$ rotations $(\psi_1, \ldots, \psi_{q(q-1)/2})$ (cf. Thiested 1988, (107)),

$$\int dV = \underbrace{\int_0^{\pi} \cdots \int_0^{\pi}}_{q(q-1)/2 \text{ times}} d\psi_1 \ldots d\psi_{q(q-1)/2} = \pi^{q(q-1)/2}.$$

Thus

$$\Delta_C^{-1} = \pi^{q(q-1)/2} \left[\int_{\Omega_D} \frac{1}{|D|^{\frac{q+2}{2}}} dD\right].$$

Finally, it remains to calculate the integral $\int_{\Omega_D} \frac{1}{|D|^{\frac{q+2}{2}}} dD$. Denoting it $I(q, a, \beta_1, \ldots, \beta_q)$, with $a = |R|^{q/p}$ it is derived by induction on q (the detailed calculation is given in Appendix IV.5).

$$I\left(q, a, \beta_1, \ldots, \beta_q\right) = \left(\frac{q-1}{q}\right)^{q-1} I\left(q-1, \left(\frac{a}{\beta_q}\right)^{\frac{q}{q-1}}, \beta_1^{\frac{q}{q-1}}, \ldots, \beta_{q-1}^{\frac{q}{q-1}}\right), \quad \text{(IV.19)}$$

and

$$I\left(2, a, \beta_1, \beta_2\right) = \frac{1}{a} \log \frac{\beta_1 \beta_2}{a} + \frac{1}{\beta_1 \beta_2} - \frac{1}{a}.$$

Appendix E. Computing the normalising constant of the diagonal variance matrix domain

We are aiming to calculate

$$I = \int_{\Omega_C} \frac{1}{|C|^{\frac{q+2}{2}}} dC, \quad \text{(IV.20)}$$

when the variance matrix C is diagonal and the domain Ω_C is defined as follows:

$$\Omega_C = \left\{C \in \mathcal{S}_q^+ \ st. \ |C| \geq |R|^{q/p}, \ |C_{ij}| \leq \sqrt{\beta_i \beta_j}, \ i, j = 1 \ldots, q\right\}. \quad \text{(IV.21)}$$

Since C is diagonal, the above definition is equivalent to

$$\left\{ \begin{array}{l} 0 \leq C_i \leq \beta_i \\ \prod_{i=1}^{q} C_i \geq a, \end{array} \right. \tag{IV.22}$$

where $\{C_i, 1 \leq i \leq q\}$ are the diagonal elements of C. Conditions (IV.22) are equivalent to the conditions

$$\left\{ \begin{array}{l} \frac{a}{\beta_2 \cdots \beta_q} \leq C_1 \leq \beta_1 \\ \frac{a}{C_1 \beta_3 \cdots \beta_q} \leq C_2 \leq \beta_2, \\ \vdots \\ \frac{a}{C_1 C_2 \cdots C_{q-1}} \leq C_q \leq \beta_q \end{array} \right. \tag{IV.23}$$

Considering I as a function of $(q, a, \beta_1, \ldots, \beta_q)$, the integral (IV.20) can be developed as follows

$$\begin{aligned} I(q, a, \beta_1, \ldots, \beta_q) &= \int_{\frac{a}{\beta_2 \cdots \beta_q}}^{\beta_1} \frac{1}{C_1^{\frac{q+2}{2}}} dC_1 \int_{\frac{a}{C_1 \beta_3 \cdots \beta_q}}^{\beta_2} \frac{1}{C_2^{\frac{q+2}{2}}} dC_2 \cdots \int_{\frac{a}{C_1 \cdots C_{q-1}}}^{\beta_q} \frac{1}{C_q^{\frac{q+2}{2}}} dC_q \\ &= \frac{2}{q a^{\frac{q}{2}}} \int_{\frac{a}{\beta_2 \cdots \beta_q}}^{\beta_1} \frac{1}{C_1} dC_1 \int_{\frac{a}{C_1 \beta_3 \cdots \beta_q}}^{\beta_2} \frac{1}{C_2} dC_2 \cdots \int_{\frac{a}{C_1 \cdots C_{q-2} \beta_q}}^{\beta_{q-1}} \frac{1}{C_{q-1}} dC_{q-1} \\ &- \frac{2}{q \beta_q^{\frac{q}{2}}} \int_{\frac{a}{\beta_2 \cdots \beta_q}}^{\beta_1} \frac{1}{C_1^{\frac{q+2}{2}}} dC_1 \int_{\frac{a}{C_1 \beta_3 \cdots \beta_q}}^{\beta_2} \frac{1}{C_2^{\frac{q+2}{2}}} dC_2 \cdots \int_{\frac{a}{C_1 \cdots C_{q-2} \beta_q}}^{\beta_{q-1}} \frac{1}{C_{q-1}^{\frac{q+2}{2}}} dC_{q-1} \\ &= \frac{2}{q a^{\frac{q}{2}}} I_{q-1} - \frac{2}{q \beta_q^{\frac{q}{2}}} \left(\frac{q-1}{q} \right)^{q-1} I\left(q-1, \left(\frac{a}{\beta_q} \right)^{\frac{q}{q-1}}, \beta_1^{\frac{q}{q-1}}, \ldots, \beta_{q-1}^{\frac{q}{q-1}} \right), \end{aligned} \tag{IV.24}$$

where

$$\begin{aligned} I_{q-1} &= \int_{\frac{a}{\beta_2 \cdots \beta_q}}^{\beta_1} \frac{1}{C_1} dC_1 \int_{\frac{a}{C_1 \beta_3 \cdots \beta_q}}^{\beta_2} \frac{1}{C_2} dC_2 \cdots \int_{\frac{a}{C_1 \cdots C_{q-2} \beta_q}}^{\beta_{q-1}} \frac{1}{C_{q-1}} dC_{q-1} \\ &= \frac{1}{(q-1)!} \left(\log \frac{\beta_1 \ldots \beta_q}{a} \right)^{q-1}, \end{aligned} \tag{IV.25}$$

is obtained by induction and

$$\begin{aligned} &\int_{\frac{a}{\beta_2 \cdots \beta_q}}^{\beta_1} \frac{1}{C_1^{\frac{q+2}{2}}} dC_1 \int_{\frac{a}{C_1 \beta_3 \cdots \beta_q}}^{\beta_2} \frac{1}{C_2^{\frac{q+2}{2}}} dC_2 \cdots \int_{\frac{a}{C_1 \cdots C_{q-2} \beta_q}}^{\beta_{q-1}} \frac{1}{C_{q-1}^{\frac{q+2}{2}}} dC_{q-1} \\ = &\left(\frac{q-1}{q} \right)^{q-1} I\left(q-1, \left(\frac{a}{\beta_q} \right)^{\frac{q}{q-1}}, \beta_1^{\frac{q}{q-1}}, \ldots, \beta_{q-1}^{\frac{q}{q-1}} \right), \end{aligned}$$

by the variable change

$$y_i = C_i^{\frac{q}{q-1}}.$$

Thus step by step thanks to equation (IV.24), the integral can be calculated when C is diagonal. For instance for $q = 2, 3, 4$ we get

$$\begin{aligned}
I(2, a, \beta_1, \beta_2) &= \frac{1}{a} \log \frac{\beta_1\beta_2}{a} + \frac{1}{\beta_1\beta_2} - \frac{1}{a}, \\
I(3, a, \beta_1, \beta_2, \beta_3) &= \frac{1}{3a^{\frac{3}{2}}} \left(\log \frac{\beta_1\beta_2\beta_3}{a} \right)^2 - \frac{4}{9a^{\frac{3}{2}}} \log \frac{\beta_1\beta_2\beta_3}{a} - \frac{8}{27\left(\beta_1\beta_2\beta_3\right)^{\frac{3}{2}}} + \frac{8}{27a^{\frac{3}{2}}}, \\
I(4, a, \beta_1, \beta_2, \beta_3, \beta_4) &= \frac{1}{12a^2} \left(\log \frac{\beta_1\beta_2\beta_3\beta_4}{a} \right)^3 - \frac{1}{8a^2} \left(\log \frac{\beta_1\beta_2\beta_3\beta_4}{a} \right)^2 + \frac{1}{8a^2} \left(\log \frac{\beta_1\beta_2\beta_3\beta_4}{a} \right) \\
&+ \frac{1}{16\left(\beta_1\beta_2\beta_3\beta_4\right)^2} - \frac{1}{16a^2}.
\end{aligned}$$

V

Adaptive design of experiments

Contents

This chapter is a collaboration with Mathieu Couplet and Nicolas Bousquet. It is in preparation for a submission.

This paper deals with the issue of building adaptive designs of experiments (DOEs) to solve inverse problems using meta-modeling (e.g. kriging). The inverse problem considered here is to estimate the distribution of a non-observed random variable X from some noisy observed data Y through a time-consuming physical model H. Bayesian inference is favored as it accounts for prior expert knowledge on X in a small sample size setting. Since it involves a high number of calls to H, the model is replaced by the kriging predictor $\widehat{H}$, along with the uncertainty of the meta-model. $\widehat{H}$ is calculated from a numerical DOE, following a methodology proposed in Fu et al. (2012, (29)). In this paper, we propose an adaptive method to build a DOE adapted to our particular purpose to estimate the posterior distribution of the parameters of X, by sequentially enriching the current DOE until a given computational budget is filled. Two Bayesian criteria, a weighted integrated mean square error (W-IMSE) and an expected conditional divergence (E-CD), are proposed. Several numerical experiments are conducted on examples. They show that such adaptive designs can significantly outperform the standard *maximin*-Latin Hypercube Designs (LHDs) and the E-CD criterion is recommended as it is successfully conducted and easier to adjust.

Keywords. Inverse problems; Bayesian inference; Kriging; Adaptive design of experiments; Markov models.

V.1 Introduction

In the past decades, science & engineering greatly benefited from the growing computational capability of computers, which allows making use of physical models with increasing levels of complexity. Complex numerical simulators are especially gaining fast development in the engineering field. However, good approximation precision is often paid for with high computational cost. Especially in fields such as statistic, optimization and probability (e.g. rare events), where many evaluations are required. Thus, so-called meta-modeling techniques have been developed to overcome the computational budget limitation (Sacks et al., 1989b, (98)).

In this paper, we focus on treating the inverse problems of the following kind:

$$Y_i \quad = \quad H(X_i, d_i) + U_i, i \in \{1, \dots, n\}, \tag{V.1}$$

where $X_i \in \mathbb{R}^q$ is a non-observed input, $d_i \in \mathbb{R}^{q_2}$ an observed input related to the experimental conditions and $U_i \in \mathbb{R}^p$ a measurement error. U_i and X_i are assumed to be independent for $i = 1, \dots, n$. Moreover, the observations Y_i are assumed to be independent for $i = 1, \dots, n$. The purpose is to estimate the distribution of the random vectors X_is from the observations $(y_i, i = 1, \dots, n)$, knowing that the physical model H cannot be inverted. In what follows, X_i will be assumed to follow a Gaussian distribution $\mathcal{N}_q(m, C)$, with mean m and variance matrix C to be estimated, and the error vector U_i will be assumed to follow a centered Gaussian distribution with known diagonal variance matrix R. The aim of the calibration is to estimate the parameters of interest $\theta = (m, C)$ from the observed data $(y_i, d_i, i = 1, \dots, n)$.

Many approaches are possible to approximate this inverse problem as linearizing the physical model H around a fixed point x_0 (see Celeux et al., 2010, (15)), or using a nonlinear approximation of the function H obtained through kriging and making use of a stochastic procedure with this approximation (see Barbillon et al., 2011, (4)). In this paper, Bayesian inference is privileged (see Fu et al., 2012, (29)). In such a Bayesian statistical framework, a Metropolis-Hastings-within-Gibbs algorithm is used to produce Markov chains converging towards the desired posterior distribution of θ.

However, as said previously, the numerical simulator H is often highly time-consuming. Therefore, using a meta-model is needed to reduce the cost. A meta-model (e.g. kriging) is an approximation of the original simulator built from its evaluations at a certain number of input values, a so-called design of experiments (DOE). The evaluation budget is usually limited, therefore it is crucial to develop methods to efficiently construct the DOEs, such that the obtained posterior distributions of θ is as close as possible to those with accessibility of the function H.

Inspired by the Efficient Global Optimization (EGO) algorithm (see Jones et al., 1998, (47)), this paper provides an adaptive method to sequentially build DOEs in such a way that new points are added to the current design according to some criterion. Two criteria are proposed here: Weighted-IMSE is minimizing the prediction error brought by the meta-model, especially in the regions of interest indicated by the current estimate of the posterior distribution of θ, while Expected-CD is maximizing the Kullback-Leibler divergence between two successive estimates of the posterior distributions of θ. A first global exploration is carried out by an initial space-filling *maximin*-Latin Hypercube Design (LHD).

The paper is organized as follows. Section V.2 gives details about kriging and the *maximin*-LHD. In Section V.3, the meta-modeling technique is combined in a Bayesian framework and the inversion is carried out using a Metropolis-Hastings-within-Gibbs algorithm. An important point is that the Gaussian Process (GP) obtained by kriging is fully embedded inside the model including the additional covariance. In the next two sections, the Expected-CD and Weighted-IMSE criteria to be optimized are described. Numerical studies are conducted on examples in Section V.6. It is important to underscore that even though the illustrating models are relatively cheap, the adaptive technique is addressed to really time-consuming functions. Finally, a discussion section sums up the main ideas and the methodology of the paper, and highlights several research aspects for further work.

Both methodologies are combined in a Bayesian inversion framework carried out using a Metropolis-Hastings-within-Gibbs algorithm in Section V.3. In the next two sections, the two criteria for deriving adaptive DOEs are described, involving an optimization problem to choose the best point to be added. Numerical studies are conducted on examples in Section V.6. It is wort Finally, a discussion section sums up the main ideas and the methodology of the paper, and highlights several research aspects for further work.

V.2 Kriging meta-model and design of experiments

This section provides a short review of meta-modeling techniques followed by a general presentation of the design of experiments, especially the construction of the *maximin*-Latin

Hypercube Design, which is chosen as the initial condition of the adaptive kriging methodology.

V.2.1 Kriging meta-model

Kriging is a geostatistical method (Matheron 1971, (65)) that has been adapted by Sacks and al. (1989b, (98)) to approximate a physical model H on a bounded hypercube Ω. This method has known a growing interest in meta-modeling since the works of Koehler and Owen (1996, (54)), Santner and al. (2003, (99)) and Fang and al. (2006, (28)), among others. More specifically, it consists of deriving a predictor $\widehat{H}(z)$ for any $z = (x, d) \in \Omega$, from the training set $\mathbf{H}_{D_N}$ evaluated from a DOE

$$D_N \;=\; \left(z_{(1)}^T, \ldots, z_{(N)}^T\right)^T, \tag{V.2}$$

with each $z_{(i)} = (x_{(i)}, d_{(i)})$. According to this approach, the function H is regarded as the realization of a Gaussian Process (GP) $\mathcal{H}$:

$$\forall z \in \Omega, \qquad \mathcal{H}(z) = \sum_{i=1}^{k} \beta_i f_i(z) + \mathbf{Z}(z), \tag{V.3}$$

where f_is are basis functions of linear regression corresponding to weight coefficients β_i, and $\mathbf{Z}$ is a Gaussian process with zero mean and stationary autocovariance

$$\mathrm{Cov}\left[\mathbf{Z}(z), \mathbf{Z}(z')\right] \;=\; \sigma^2 K_\psi(\|z - z'\|),\ \forall (z, z') \in \Omega^2, \tag{V.4}$$

with K_ψ a symmetric positive definite kernel such that $K_\psi(\mathbf{0}) = 1$. This is equivalent to assume a GP prior distribution for H in a Bayesian perspective (see Rasmussen & Williams, 2006, (86)). At any prediction point z, the process $\mathcal{H}$ can be proved to be normally distributed knowing the evaluations $\mathbf{H}_{D_N}$ (Santner et al. 2003, (99)):

$$\mathcal{H}(z) \,|\, \mathcal{H}_{D_N} = \mathbf{H}_{D_N} \;\sim\; \mathcal{N}\Big[\mu_H(z), \sigma_H^2(z)\Big]. \tag{V.5}$$

The best linear unbiased predictor (BLUP) for $H(z)$, denoted by $\widehat{H}(z)$, is the conditional kriging mean $\mu_H(z)$ and the conditional kriging variance $\sigma_H^2(z)$ is the so-called MSE (Mean Squared Error) (see Johnson et al. 1990 for details, (46)), which provides an estimate of the predicting accuracy. Assuming that the covariance parameters (σ^2, ψ) are known, MSE can be expressed by the following equation:

$$\mathrm{MSE}(z) \;= \sigma_H^2(z) \;=\; \sigma^2\Big(1 + \gamma(z)^T(\mathbf{F}_D^T \boldsymbol{\Sigma}_{DD}^{-1} \mathbf{F}_D)^{-1}\gamma(z) - \boldsymbol{\Sigma}_{zD}^T \boldsymbol{\Sigma}_{DD}^{-1} \boldsymbol{\Sigma}_{zD}\Big), \tag{V.6}$$

where $\gamma(z) = \mathbf{F}(z) - \mathbf{F}_D^T \boldsymbol{\Sigma}_{DD}^{-1} \boldsymbol{\Sigma}_{zD}$, and

- $\mathbf{F}(z) = [f_1(z), \ldots, f_k(z)]^T$ is a $k \times 1$ vector of basis functions,
- $\mathbf{F}_D = [\mathbf{F}(z_{(1)}), \ldots, \mathbf{F}(z_{(N)})]^T$ is a $N \times k$ regression matrix evaluated at design D_N,
- $\boldsymbol{\Sigma}_{zD} = [K_\psi(z, z_{(1)}), \ldots, K_\psi(z, z_{(N)})]^T$ is a $N \times 1$ vector of correlations between z and design points,

- $\mathbf{\Sigma}_{DD} = [K_\psi(z_{(i)}, z_{(j)})]_{1\leq i,j\leq N}$ is a $N \times N$ correlation matrix evaluated within the design of experiments.

In most general cases where the covariance parameters are unknown, several methods can be used to estimate them, e.g. maximum likelihood, restricted maximum likelihood, cross-validation or Bayesian kriging techniques. The predictor $\widehat{H}(z)$ and the variance MSE(z) which will not be detailed here can also be expressed explicitly by plugging in the estimated covariance terms.

Obviously, the predicting accuracy highly depends on the position of the prediction point $z \in \Omega$ with respect to the spatial structure of the DOE.

V.2.2 Design of experiments (*maximin*-Latin Hypercube Designs)

Following Picheny et al.(2010, (83)), it is possible to distinguish three kinds of DOEs:

- *Space-filling* designs, which aims to fill the input space with a finite number of points independently of models, the *maximin*-LHD for example;
- *model-oriented* designs, which attempts to construct a suited DOE accounting for the model H or the meta-model;
- *purpose-oriented* designs, which takes into account the final object to find the best adapted DOE, for example to find the best posterior distributions of θ in the current case study.

In this paper, we choose to build a *purpose-oriented* DOE in an adaptive way to solve the inverse problems. Namely, we get a first calibration of the covariance parameters from a space-filling design such as the *maximin*-Latin Hypercube Design (LHD), then the DOE is to be improved through some sequential strategies. Proposing such strategies is the central point of this paper. First, the standard *maximin*-LHD is reviewed. We choose a domain Ω to illustrate our explication.

The concept of LHD was introduced by McKay et al. (1979, (66)) and two distance criteria namely *maximin* and *minimax* were proposed by Johnson et al. (1990, (46)) to choose the optimal LHDs. *Maximin* means maximizing the minimum inter-site distance between the set of N points:

$$\delta_D \quad = \quad \min_{i\neq j} \|z_{(i)} - z_{(j)}\|_2,$$

while *minimax* means minimizing the maximum euclidean distance between any point $z \in \Omega$ and its nearest neighbor among the set of N design points

$$h_D \quad = \quad \sup_{z\in\Omega} \min_{1\leq i\leq N} \|z - z_{(i)}\|_2.$$

Ideally, the criterion *maximin* can avoid generating the points of design too close to each other and thus ensure good filling of the space, while *minimax* aims at making none of points in the domain Ω be far from the design. However, the optimization involved with the *minimax* is quite time-consuming, especially in large dimensions. For this reason, the *maximin*-LHD is chosen as our adaptive strategy.

V.3 Embedding the meta-model into Bayesian inference

In this section, we recall the Bayesian framework, introduced in Fu et al. (2012, (29)), within which the inverse problem of estimating the distribution of the missing-data $\mathbf{X}$ is considered. The following conditional conjugated prior distribution was chosen,

$$m \,|\, C \;\sim\; \mathcal{N}_q(\mu, C/a), \tag{V.7}$$
$$C \;\sim\; \mathcal{IW}_q(\Lambda, \nu), \tag{V.8}$$

$\rho = \{a, \mu, \Lambda, \nu\}$ being the hyperparameters . It can be noted that a can be regarded as the size of a *virtual* sample corresponding to the belief of the practitioner on the prior information (provided by experts). A Gibbs sampler involving data augmentation was proposed to compute its posterior distribution of the parameters. Actually, replacing the expensive-to-compute function H with a kriging emulator $\widehat{H}$ (see Barbillon, (2010), (3)) and introducing a new *emulator error* MSE, the Gibbs sampler can be adapted as follows:

Gibbs sampler (at the $(r+1)$-th iteration)

`Given` $(m^{[r]}, C^{[r]}, \mathbf{X}^{[r]})$ `for` $r = 0, 1, 2, \ldots,$ `generate:`

1. $C^{[r+1]}|\cdots \sim \mathcal{IW}\Big(\Lambda + \sum_{i=1}^n (m^{[r]} - X_i^{[r]})(m^{[r]} - X_i^{[r]})' + a(m^{[r]} - \mu)(m^{[r]} - \mu)',\ \nu + n + 1\Big)$,
2. $m^{[r+1]}|\cdots \sim \mathcal{N}\Big(\frac{a}{n+a}\mu + \frac{n}{n+a}\overline{\mathbf{X}_n^{[r]}},\ \frac{C^{[r+1]}}{n+a}\Big)$ where $\overline{\mathbf{X}_n^{[r]}} = n^{-1}\sum_{i=1}^n X_i^{[r]}$,
3. $\mathbf{X}^{[r+1]}|\cdots \propto |\mathbf{R} + \mathrm{MSE}^{[r+1]}|^{-\frac{1}{2}} \cdot \exp\Bigg\{ -\frac{1}{2}\sum_{i=1}^n (X_i^{[r+1]} - m^{[r+1]})' \Big[C^{[r+1]}\Big]^{-1} (X_i^{[r+1]} - m^{[r+1]}) -$

$$\frac{1}{2}\Big(\Big(\mathcal{Y}_1 - \widehat{H}_{N,1}^{[r+1]}\Big)', \ldots, \Big(\mathcal{Y}_n - \widehat{H}_{N,n}^{[r+1]}\Big)'\Big)\Big(\mathbf{R} + \mathrm{MSE}^{[r+1]}\Big)^{-1} \begin{pmatrix} \mathcal{Y}_1 - \widehat{H}_{N,1}^{[r+1]} \\ \vdots \\ \mathcal{Y}_n - \widehat{H}_{N,n}^{[r+1]} \end{pmatrix} \Bigg\}$$

`where` $\widehat{H}_{N,i}^{[r+1]} = \widehat{H}_N(X_i^{[r+1]}, d_i)$ `and` $\mathtt{MSE}^{[r+1]} = \mathtt{MSE}(\mathbf{X}^{[r+1]}, \mathbf{d})$ `is the block diagonal matrix`

$$\mathrm{MSE}(\mathbf{X}^{[r+1]}, \mathbf{d}) \;=\; \begin{pmatrix} \mathrm{MSE}_1(\mathbf{X}^{[r+1]}, \mathbf{d}) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathrm{MSE}_p(\mathbf{X}^{[r+1]}, \mathbf{d}) \end{pmatrix} \begin{matrix} \} \\ \\ \} \end{matrix} \begin{matrix} n \text{ lines} \\ \\ n \text{ lines} \end{matrix}$$

In the third step, the variance matrices $\mathrm{MSE}_j(\mathbf{X}^{[r+1]}, \mathbf{d}) \in \mathcal{M}^{n\times n}$ are defined by

$$\mathrm{MSE}_j(\mathbf{X}^{[r+1]}, \mathbf{d}) \;=\; \mathbb{E}\left(\Big(\mathcal{H}_j(\mathbf{X}^{[r+1]}, \mathbf{d}) - \widehat{H}_j(\mathbf{X}^{[r+1]}, \mathbf{d})\Big)^2 \,|\, \mathbf{H}_{D_N}\right),$$

for $j = 1, \ldots, p$, where $\mathcal{H}_j$ denotes the j-th dimension of the Gaussian process $\mathcal{H}$. Moreover,

$$\mathbf{R} \;=\; \begin{pmatrix} \mathbf{R}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{R}_p \end{pmatrix} \begin{matrix} \} \\ \\ \} \end{matrix} \begin{matrix} n \text{ lines} \\ \\ n \text{ lines} \end{matrix}, \text{ with } \mathbf{R}_i = \begin{pmatrix} R_{ii} & & 0 \\ & \ddots & \\ 0 & & R_{ii} \end{pmatrix},$$

where R_{ii} is the $i-$th diagonal component of the diagonal variance matrix R. It is worth noting that this third conditional distribution does not belong to any closed form family of distributions. Therefore a Metropolis-Hastings (MH) step is used to simulate $\mathbf{X}^{[r+1]}$ (see Appendix A).

As discussed in Fu et al. (2012, (29)), the use of the MCMC algorithms involves many possible errors. According to experimental trials, the accuracy of the meta-model plays a critical role in the the estimation problem. MCMC algorithms can produce Markov chains converging towards the desired posterior distribution. However, if the function H is really badly approximated, apart from the *algorithmic error* introduced by the MCMC algorithm, the result can also suffer from an *emulator error*.

V.4 The Expected-Conditional Divergence criterion for adaptive designs

The two following sections address the issue of adaptive designs of experiments. Two strategies are proposed. In this section, we propose a criterion called E-CD (Expected-Conditional Divergence) to define the adaptive procedure, which can be considered as a variation of the Expected-Improvement criterion proposed by Jones et al. (1998, (47)).

V.4.1 Principle

Ideally, the posterior distribution of the parameters $\theta = (m, C)$ after adding a new point $z_{(N+1)}$ to the current DOE D_N should be as close as possible to the posterior distribution knowing the original function H, i.e. a relevant discrepancy measure between the two relative distributions is to be minimized. Based on information-theoretical arguments given in Cover and Thomas (2006, (18)), the Kullback-Leibler divergence

$$\mathrm{KL}\Big(\pi(\theta|\mathbf{y},\mathbf{d},H)\,||\,\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N}\cup\{H(z)\})\Big), \tag{V.9}$$

is a good choice of discrepancy measure. Remind that given two densities $p(x)$ and $q(x)$ defined over the same space $\mathcal{X}$,

$$\mathrm{KL}(p||q) \;=\; \int_{\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}\,dx.$$

Assuming that this quantity can be calculated, we search for the next point $z_{(N+1)}$ within the feasible region Ω, as the global minimum of this divergence. Of course, the term $\pi(\theta|\mathbf{y},\mathbf{d},H)$ makes this formulation useless, but we can derive a tractable criterion from it.

$$\begin{aligned}
z_{(N+1)} &= \underset{z\in\Omega}{\mathrm{argmin}}\ \mathrm{KL}\Big(\pi(\theta|\mathbf{y},\mathbf{d},H)\,||\,\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N}\cup\{H(z)\})\Big)\\
&= \underset{z\in\Omega}{\mathrm{argmin}}\ \mathrm{KL}\Big(\pi(\theta|\mathbf{y},\mathbf{d},H)\,||\,\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N}\cup\{H(z)\})\Big) - \mathrm{KL}\Big(\pi(\theta|\mathbf{y},\mathbf{d},H)\,||\,\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N})\Big)\\
&= \underset{z\in\Omega}{\mathrm{argmax}}\ \int_{\theta\in\Omega}\pi(\theta|\mathbf{y},\mathbf{d},H)\,\log\frac{\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N}\cup\{H(z)\})}{\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N})}\,d\theta\\
&\simeq \underset{z\in\Omega}{\mathrm{argmax}}\ \int_{\theta\in\Omega}\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N}\cup\{H(z)\})\,\log\frac{\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N}\cup\{H(z)\})}{\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N})}\,d\theta\\
&= \underset{z\in\Omega}{\mathrm{argmax}}\ \mathrm{KL}\Big(\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N}\cup\{H(z)\})\,||\,\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N})\Big).
\end{aligned} \tag{V.10}$$

Under the assumption $\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N}\cup\{H(z)\}) \simeq \pi(\theta|\mathbf{y},\mathbf{d},H)$ to ensure the penultimate passage, minimizing the KL divergence (V.9) becomes equivalent to maximize in z the KL divergence given in the (V.10). In other words, this strategy aims at finding the optimal point $z_{(N+1)}$ which modify the actual distribution $\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N})$ as much as possible. Our choice appears asymptotically relevant.

The preceding formulation is not satisfactory yet, since one evaluation of the criterion requires one evaluation of H, which is time-consuming. However, in the spirit of EGO, it is possible to derive a new criterion from the former considering the following Gaussian process based on the available observations $\mathbf{H}_{D_N}$ instead of H:

$$h_N(z) \;:=\; \mathcal{H}(z)\,|\,\mathbf{H}_{D_N},$$

which follows the normal distribution given in (V.5). Thus, we define the expected distance criterion:

$$z_{(N+1)} \;=\; \underset{z\in\Omega}{\operatorname{argmax}}\,\mathbb{E}_{\pi(h_N)}\left[\mathrm{KL}\Big(\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N}\cup\{h_N(z)\})\,||\,\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N})\Big)\right]. \quad \text{(V.11)}$$

The idea of considering the Gaussian variable $h_N(z)$ rather than the predictor $\widehat{H}_N(z)$ allows us to take into account the uncertainty introduced by the kriging method, while it requires usual Monte Carlo methods to approximate the double integrals, i.e. the expectation and the Kullback-Leibler distance.

An practical problem with the so-defined criterion is that for one more search of the new point z, one additional Gibbs sampler has to be run to approximate the posterior distribution $\pi(\theta|\mathbf{y},\mathbf{d},\mathbf{H}_{D_N}\cup\{h_N(z)\})$, while Monte Carlo methods require thousands of searches. The CPU computational time is not negligible.

V.4.2 The Expected-Conditional Divergence criterion

Preliminary experiments showed that the criterion defined in (V.11) is generally too expensive to compute because of the Gibbs sampler. In this subsection, we provide the so-called Expected-Conditional Divergence (E-CD) criterion, which depends only on the intermediate full-conditional posterior distributions of θ. More precisely, at the $(r+1)$-th iteration of the Metropolis-Hastings-within-Gibbs algorithm, the criterion is described as:

$$z_{(N+1)} \;=\; \underset{z\in\Omega}{\operatorname{argmax}}\,\text{E-CD}(z)$$

with

$$\text{E-CD}(z) \;=\; \mathbb{E}_{\pi(h_N)}\left[\mathrm{KL}\Big(\pi(\theta|\tilde{\mathbf{X}}^{(r+1)}(z))\,||\,\pi(\theta|\mathbf{X}^{(r+1)})\Big)\right],$$

where $\mathbf{X}^{(r+1)}$ and $\tilde{\mathbf{X}}^{(r+1)}(z)$ denote the missing data samples simulated from

$$\begin{aligned}\mathbf{X}^{(r+1)} &\sim \pi\left(\cdot|\mathbf{y},\mathbf{d},\theta^{(r+1)},\mathbf{H}_{D_N}\right),\\ \tilde{\mathbf{X}}^{(r+1)}(z) &\sim \pi\left(\cdot|\mathbf{y},\mathbf{d},\theta^{(r+1)},\mathbf{H}_{D_N}\cup\{h_N(z)\}\right).\end{aligned}$$

It is worth noting that in the E-CD criterion, the final posterior distribution of θ is replaced by its sequential conditional posterior distribution at the $(r+1)$-th iteration. Thus, we do

not wait for the convergence of the simulated Markov chains to compute the criterion, with the purpose to speed up the research of a new relevant location $z_{(N+1)}$. Even if it may be less efficient, we do not need to call the expensive function H and we are free of any additional Gibbs sampling. This method can be expected to have good performances by exploring more candidates faster.

As presented before, Monte Carlo methods are recommended for the double integrals in (V.12). The empirical version of the CD strategy can then be summarized by the following five steps.

Expected-Conditional Divergence algorithm

At iteration $r+1$ (large enough), for each candidate z:

1. generate M samples $(h_N^1(z), \dots, h_N^M(z))$ according to (V.5) and build M corresponding emulators $(\widehat{H}_{N+1}^1(z), \dots, \widehat{H}_{N+1}^M(z))$ with $\widehat{H}_{N+1}^i(z)$ based on the dataset $\mathbf{H}_{D_N} \cup \{h_N^i(z)\}$.
2. for $i = 1, \dots, M$,

   (i) sample $\tilde{\mathbf{X}}^{(r+1),i}(z)$ from $\pi(\cdot|\mathbf{y}, \mathbf{d}, \theta^{(r+1)}, \widehat{H}_{N+1}^i(z))$,

   (ii) sample $\Theta^i = \{\theta_1^i, \dots, \theta_{L_1}^i\}$ with $\theta = (m_1, \dots, m_q, C_{11}, \dots, C_{qq})$ from $\pi(\cdot|\tilde{\mathbf{X}}^{(r+1),i}(z), \mathbf{y}, \mathbf{d})$,
3. sample $\Psi = \{\theta_1, \dots, \theta_{L_2}\}$ from $\pi(\cdot|\mathbf{X}^{(r+1)}, \mathbf{y}, \mathbf{d})$,
4. find the optimal point $z_{(N+1)}$ within the domain Ω

$$z_{(N+1)} = \underset{z\in\Omega}{\operatorname{argmax}} \frac{1}{M} \sum_{i=1}^{M} \widehat{\mathrm{KL}}\Big(\Theta^i \,||\, \Psi\Big) = \underset{z\in\Omega}{\operatorname{argmax}}\, \widehat{\text{E-CD}}_M(z), \tag{V.12}$$

   where $\widehat{\mathrm{KL}}(\Theta^i \,||\, \Psi)$ denotes the empirical KL divergence.
5. repeat 1)-4) until the maximum budget has been reached.

Remark 25. *Since we are in the context with an expensive code H, the stopping criterion based on the maximum budget appears a good practical choice.*

In the Expected-Conditional Divergence algorithm, (V.12) requires two computations : the empirical KL divergence and the inner optimization problem. Let us begin with the first calculation. As in Wang et al. (2006, (115)), a Nearest-Neighbor (NN) approach is applied to estimate the KL distance, as explained in Appendix B.

The optimization problem can be solved with help of the Simulated Annealing (SA) method, as presented in Appendix C. It is worth mentioning that to maximize the E-CD criterion is equivalent to minimize the $-$E-CD criterion.

V.5 The Weighted-IMSE criterion for adaptive designs

In this section, we provide another criterion as a modified version of the weighted-IMSE criterion (see Sacks et al., 1989a, (97), Picheny et al., 2010, (83)), adapted to inverse problems in Bayesian framework.

V.5.1 The Integrated MSE criterion

The Integrated Mean Square Error (IMSE) criterion (Sacks et al., 1989b, (98)) is a measure of the average accuracy of the kriging meta-model over the domain Ω:

$$\text{IMSE}(\Omega) \quad = \quad \int_{\Omega} \text{MSE}(z)\, dz,$$

where $\text{MSE}(z)$ follows (V.6). Aiming at ensuring the prediction accuracy in the regions of main interest, Picheny et al. (2010, (83)) proposed the W-IMSE criterion which consists in adding a weight function ω to indicate the critical regions. Based on the current design D_N of N points, W-IMSE is defined as follows:

$$\text{W-IMSE}(z^*) \quad = \quad \int_{\Omega} \text{MSE}\left(z|D_N \cup \{z^*\}\right) w\left(z|D_N, \mathbf{H}_{D_N}\right)\, dz, \tag{V.13}$$

where $\text{MSE}\left(z|D_N \cup \{z^*\}\right)$ denotes the prediction variance by adding the point $z^* = (x^*, d^*)$ into D_N and $w\left(z|D_N, \mathbf{H}_{D_N}\right)$ is a weight function emphasizing the MSE term over a region of interest. An adapted choice of ω to our Bayesian framework is to be given in the next section. It is worth noting that the calculation of MSE does not depend on the expensive evaluation $H(z^*)$ and the weight factor w only depends on the available observations $\mathbf{H}_{D_N}$. The next point to add to the DOE is thus defined by

$$z_{(N+1)} \quad = \quad \arg\min_{z\in\Omega} \text{W-IMSE}(z).$$

V.5.2 Adaptation to our purpose

Defining the regions of interest is the essential task in applying the W-IMSE criterion. As presented in previous sections, a probabilistic solution to inverse problems is to approximate the posterior distribution of the parameters $\theta = (m, C)$ using a Metropolis-Hastings-within-Gibbs algorithm (cf. Section V.3). Assuming that the $(N+1)-$th new point is added at the $(r+1)-$th iteration of the Gibbs sampling, the weight function is defined by the following formula:

$$\begin{aligned} w\left(z|D_N, \mathbf{H}_{D_N}\right) \quad &\propto \quad \prod_{i=1}^{n} \pi\left(x, d|y_i, \theta^{(r+1)}, D_N, \mathbf{H}_{D_N}\right) \\ &\propto \quad \prod_{i=1}^{n} |\mathbf{R} + \text{MSE}(x,d)|^{-\frac{1}{2}} \cdot \exp\left\{ -\frac{1}{2}(x - m^{(r+1)})' \left[C^{(r+1)}\right]^{-1} (x - m^{(r+1)}) \right. \\ &\quad \left. -\frac{1}{2}\left(y_i - \widehat{H}(x,d)\right)' \left(\mathbf{R} + \text{MSE}(x,d)\right)^{-1} \left(y_i - \widehat{H}(x,d)\right) \right\}, \end{aligned} \tag{V.14}$$

which is derived from the full conditional posterior distribution of $\mathbf{X}$ described in Section V.3 and W-IMSE defined in (V.13) can be considered as the posterior weighted prediction error up to a multiplicative constant. The advantage of this choice is twofold. First, such defined weight function ω indicates a potential position for the missing-data $\mathbf{X}$ where the accuracy of the meta-model should be improved. Second, this weight function depends on the observation sample $\mathbf{y} = \{y_1, \ldots, y_n\}$, which is coherent with the Bayesian inference and the advantage of building a *purpose-oriented* design can be emphasized.

Besides, as $w\left(z|D_N, \mathbf{H}_{D_N}\right)$ is of different nature from the prediction variance MSE, a compromising level α is introduced. We derive the following version of the W-IMSE criterion:

$$\text{W-IMSE}(z^*) \quad = \quad \int_\Omega \text{MSE}^\alpha\left(z|D_N \cup \{z^*\}\right) w^{1-\alpha}\left(z|D_N, \mathbf{H}_{D_N}\right) dz. \tag{V.15}$$

In this equation, α varying between 0 and 1 makes the criterion more flexible: if α is close to 1, the impact of the weight parameter ω disappears and the criterion becomes IMSE; if α is approached to 0, the prediction error MSE will not be taken into account. Experimental trails proved that the choice of α is critical.

However, such a chosen weight function w, defined as the product of n possible small densities, may cause numerical (underflow) problems. One solution to this issue is normalizing $w^{1-\alpha}$, which leads to the weight function $\frac{w^{1-\alpha}}{\int w^{1-\alpha}}$ a density of probability, as initially suggested by Picheny et al. (2004, (83)):

$$\text{W-IMSE}(z^*) \quad = \quad \int_\Omega \text{MSE}^\alpha\left(z|D_N \cup \{z^*\}\right) \frac{w^{1-\alpha}\left(z|D_N, \mathbf{H}_{D_N}\right)}{\int_\Omega w^{1-\alpha}\left(z'|D_N, \mathbf{H}_{D_N}\right) dz'} dz. \tag{V.16}$$

Another solution is dividing $w^{1-\alpha}$ by its maximum in domain Ω, which leads to

$$\text{W-IMSE}(z^*) \quad = \quad \int_\Omega \text{MSE}^\alpha\left(z|D_N \cup \{z^*\}\right) \frac{w^{1-\alpha}\left(z|D_N, \mathbf{H}_{D_N}\right)}{\max_{z' \in \Omega} w^{1-\alpha}\left(z'|D_N, \mathbf{H}_{D_N}\right)} dz. \tag{V.17}$$

Both solutions require a Monte Carlo approximation to calculate the normalizing constant.

For a DOE of dimension one or two, a Cartesian grid over the design space Ω can be used to solve the numerical integration and optimization problems (see Picheny, 2010, (83)). In more general cases of higher dimension, stochastic integration and global optimization techniques should be preferred, e.g. the Monte Carlo method and the Simulated Annealing (SA) algorithm (see Kirkpatrick et al., 1983, (53)). In this paper, the Monte Carlo method has been applied even for the two-dimensional example. The SA algorithm, as the main stochastic optimization method, is presented in Appendix C.

Next section is devoted to exploring on examples if the proposed methodology may be successfully conducted. Moreover, we compare them with a simple criterion which selects an optimal point $z_{(N+1)}$ to minimize the maximum MSE, the so-called MMSE, by adding this point to the current DOE:

$$\begin{aligned} z_{(N+1)} &= \operatorname*{argmin}_{z^* \in \Omega} \text{MMSE}(z^*) \\ &= \operatorname*{argmin}_{z^* \in \Omega} \max_{z \in \Omega} \text{MSE}\left(z|D_N \cup \{z^*\}\right). \end{aligned}$$

V.6 Numerical experiments

This section deals with numerical studies to check the performances of our adaptive kriging strategies. The first example is a simplified model with two inputs, and the second example is a more complex physical hydraulic engineering model. In both examples, the performance of the W-IMSE and E-CD criteria is compared with the standard *maximin*-LHD and the MMSE criterion, under the same evaluation budget which means the same number of calls to H. A good kriging meta-model with enough points has been built as a benchmark. Once again, we mention that despite of these simplified examples, our adaptive techniques aim at treating highly complex industrial codes.

V.6.1 Example: Two-input toy model

We begin with the following two-dimensional parametric function (see Bastos and O'Hagan, 2009, (5))

$$H(x_1, x_2) \quad = \quad \left(1 - \exp\left(-\frac{1}{2x_2}\right)\right)\left(\frac{2300x_1^3 + 1900x_1^2 + 2092x_1 + 60}{100x_1^3 + 500x_1^2 + 4x_1 + 20}\right),$$

with $x_i \in (0,1), i = 1, 2$. The two-dimensional uncertainty model can then be described as

$$Y_i \quad = \quad H(X_i) + U_i, i \in \{1, \ldots, n\}, \tag{V.18}$$

In our experimental trials, the design domain Ω is fixed to be $[0,1]^2$. Three types of DOE presented in Table V.6.1 are generated.

.

DOE 1	10-point-*maximin*-LHD	
DOE 2	5-point-*maximin*-LHD +	5-point-W-IMSE
		5-point-E-CD
		5-point-MMSE
DOE 3	100-point-*maximin*-LHD (*benchmark*)	

Table V.1: Description of the three types of DOE for the two-dimensional model

The dataset $\mathbf{Y} = (Y_i, i = 1, \ldots, 30)$ of size $n = 30$ is simulated from the uncertainty model (V.18) where the missing data X_i is generated with the following Gaussian distribution, truncated in domain Ω:

$$X_i \quad \sim \quad \mathbb{1}_\Omega \mathcal{N}_2\left(\begin{pmatrix} 0.5 \\ 0.7 \end{pmatrix}, \begin{pmatrix} 0.15^2 & 0 \\ 0 & 0.4^2 \end{pmatrix}\right), \tag{V.19}$$

and the error term U_i is the realization of a $\mathcal{N}_1(0, 10^{-5})$ random variable. Moreover, in (V.7) and (V.8), the hyperparameters are chosen as follows. Recall that a can be interpreted as the size of a *virtual* sample to be adjusted with respect to our belief on expert opinion.

$$\begin{cases} a = 1, \\ \nu = 5, \\ \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \\ \Lambda = 2 \cdot \begin{pmatrix} 0.15^2 & 0 \\ 0 & 0.4^2 \end{pmatrix}. \end{cases}$$

In practice, 3,000 iterations of the MCMC algorithm were used for the "pre" burn-in period, the relevance of which was controlled using the Brooks-Gelman diagnostic of convergence (see Brooks and Gelman, 1998, (11)). Here, the criterion $\widehat{R}_{BG}$ was calculated every 50 iterations and the convergence is accepted if $\widehat{R}_{BG}$ remains smaller than 1.05 for at least 3,000 successive iterations. The points $\{z_{(6)}, \ldots, z_{(10)}\}$ are added one by one into the current design every 100 iterations, after the pre burn-in period. For the SA algorithm, the initial point $x^{[0]}$ is fixed to be the current simulated missing data x, the initial temperature β_0 equals 100 and a large standard deviation $\sigma = 100$ is chosen for the instrumental distribution.

Weighted-IMSE criterion The first experiment concerns the performance of the W-IMSE criterion. Four values, $1, 0.8, 0.5$ and 0.2, have been chosen for the weight parameter α. The number of iterations of the SA algorithm is chosen equal to 1,000 and the number of iterations of the MC algorithm is limited to 1,000. Moreover, the initial DOE consists of a 5-point-*maximin*-LHD and 5 points are added iteratively to the DOE according to this criterion as presented in Table V.6.1.

Figure V.1 provides a comparison of the so-built W-IMSE designs with the standard 10-point-*maximin*-LHD, the E-CD design and the MMSE design. We can see that the added points are not far from the hypothesized mean $(0.5, 0.7)$ and the four W-IMSE-designs are similar.

However, the posterior distributions of the parameters θ are quite sensitive to the choice of α. Figure V.2 displays these posterior distributions with the different meta-models. We can see that the 10-point-*maximin*-LHD performs really poorly, which has been improved with help of the adaptive procedure. Moreover, the MMSE criterion performs correctly. The W-IMSE criterion improved the posterior distributions of m_2 and C_{22}, which are sensitive to the choice of α. Besides, for the posterior distribution of m_1 and C_{11}, $\alpha = 1, 0.5$ and 0.2 do not work well.

We can say that the interest of using the W-IMSE criterion is verified in this case study, but, the high dependence to α remains a great disadvantage. In what follows, α is fixed to be 0.8 in our case study.

Expected-CD criterion To compute an empirical version of this criterion, the number of generated GPs $h_N^i(z)$ is chosen equal to 100, the size of the sample Θ^i and Ψ is limited to 1,000 and the number of iterations of the SA algorithm is fixed to 1,000. As presented in Figure V.1, 5 points are added sequentially according to this criterion into a 5-point-*maximin*-LHD. Moreover, the same 10-point-*maximin*-LHD and the same benchmark have been used for this case study.

From Figure V.3, the E-CD design outperforms the 10-point-*maximin*-LHD under the same budget, i.e. the same number of calls to H and the E-CD criterion performs more efficiently that the MMSE criterion and similarly to the W-IMSE criterion with $\alpha = 0.8$. It is worth noting that it is free of the constraint of the choice of α, which can be considered as an important advantage.

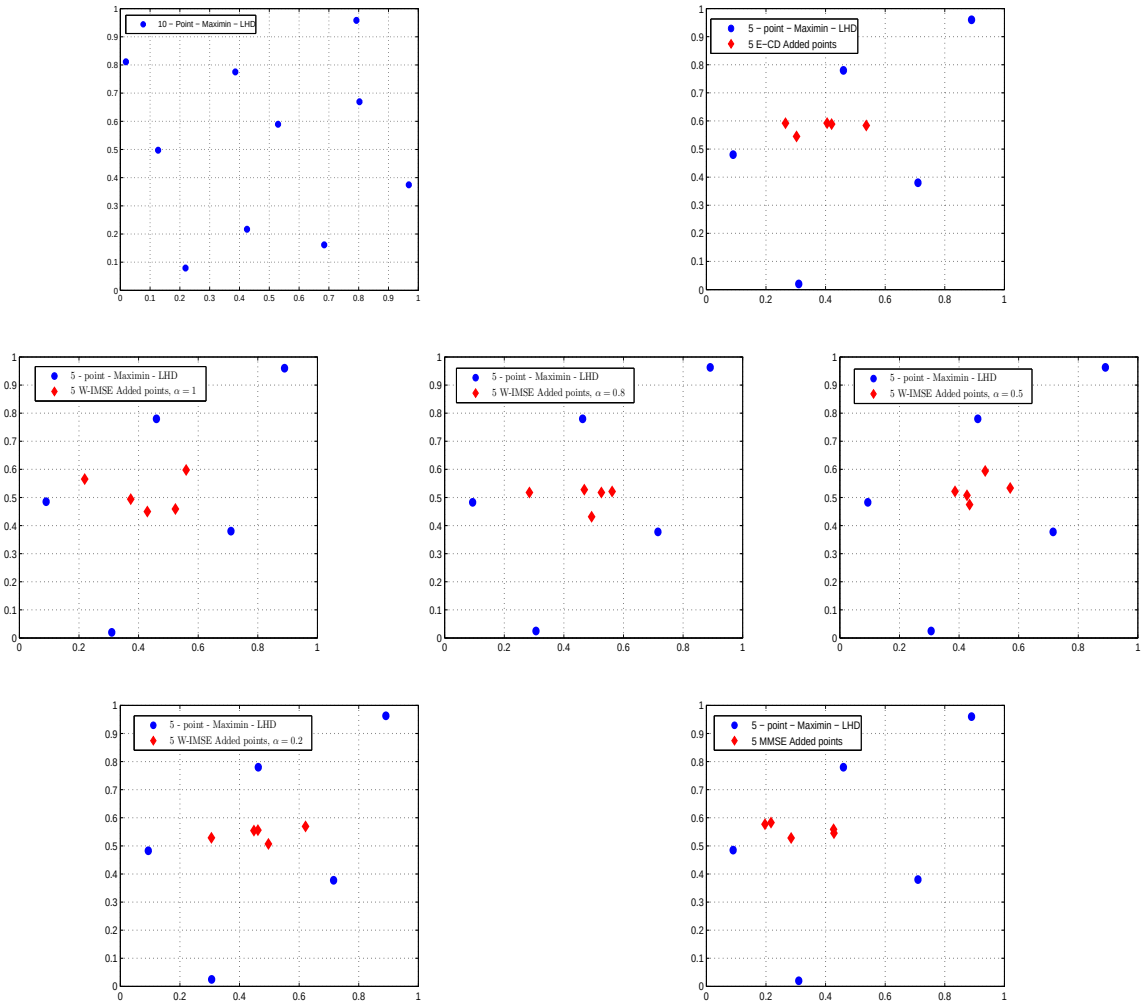


Figure V.1: Standard *maximin*-LHD, E-CD design, W-IMSE designs of experiments with $\alpha = 1, 0.8, 0.5, 0.2$ and MMSE design

V.6.2 Example: A hydraulic engineering model

The second example is a simplified three-dimensional-input model, which involves a hydraulic function described as:

$$H(x,d) \quad = \quad \left(x_2 + \left(\frac{\sqrt{5000}}{300\sqrt{55 - x_2}} \times \frac{d}{x_1} \right)^{0.6} , \frac{d^{0.4} x_1^{0.6} (55 - x_2)^{0.3}}{300^{0.4} \times 5000^{0.3}} \right)' , \qquad \text{(V.20)}$$

with $x = (x_1, x_2)$. This three-dimensional model takes the form defined in (V.1). Following similar Gaussian assumptions on X_i and U_i, the same prior distributions on $\theta = (m, C)$ have been chosen as in (V.7) and (V.8) and a Gibbs sampler was then run to approximate the posterior distributions of θ.

In this case study, the observations $\mathbf{y} = \{y_1, \ldots, y_{30}\}$ with the sample size $n = 30$ are

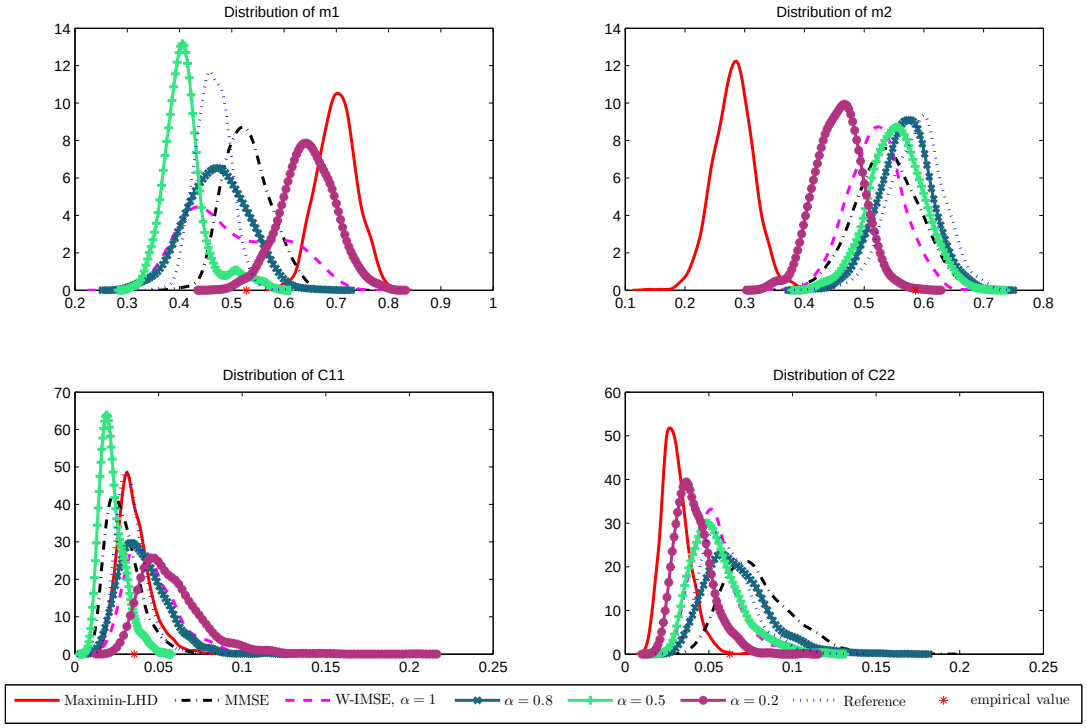


Figure V.2: Posterior distributions of θ with benchmark, standard *maximin*-LHD, MMSE design and W-IMSE designs with $\alpha = 1, 0.8, 0.5, 0.2$

generated from the hydraulic model (V.1), where

$$\begin{aligned} X_i &\sim \mathcal{N}\Big(\begin{pmatrix} 30 \\ 50 \end{pmatrix}, \begin{pmatrix} 5^2 & 0 \\ 0 & 1 \end{pmatrix} \Big), \\ d_i &\sim \text{Gumbel}\Big(1013, -458\Big), \end{aligned}$$

and the error term $U_i \sim \mathcal{N}(\mathbf{0}, 10^{-5} \cdot I_2)$. Several choices of the prior distribution of θ have been discussed and compared in Fu et al. (2012, (29)). As it is not the central point of this paper, here, we apply only the "FMHV" (Fair Mean High Variance) prior. The prior choice of hyperparameters has been inherited from the two-dimensional example except for the prior mean μ and the prior variance term Λ:

$$\begin{cases} \mu = \begin{pmatrix} 35 \\ 49 \end{pmatrix}, \\ \Lambda = 2 \cdot \begin{pmatrix} 7.5^2 & 0 \\ 0 & 1.5^2 \end{pmatrix}. \end{cases}$$

The design domain Ω is given as $[20, 40] \times [45, 55] \times [\min(\mathbf{d}), \max(\mathbf{d})]$, with $\mathbf{d} = \{d_1, \ldots, d_{30}\}$.

Similarly, the Brooks-Gelman statistic $\widehat{R}_{BG}$ was calculated to verify the convergence of the Gibbs sampler and the same parameterization was assumed for the SA algorithm. Once again, the performances of the W-IMSE and E-CD criteria are verified with help of a benchmark as

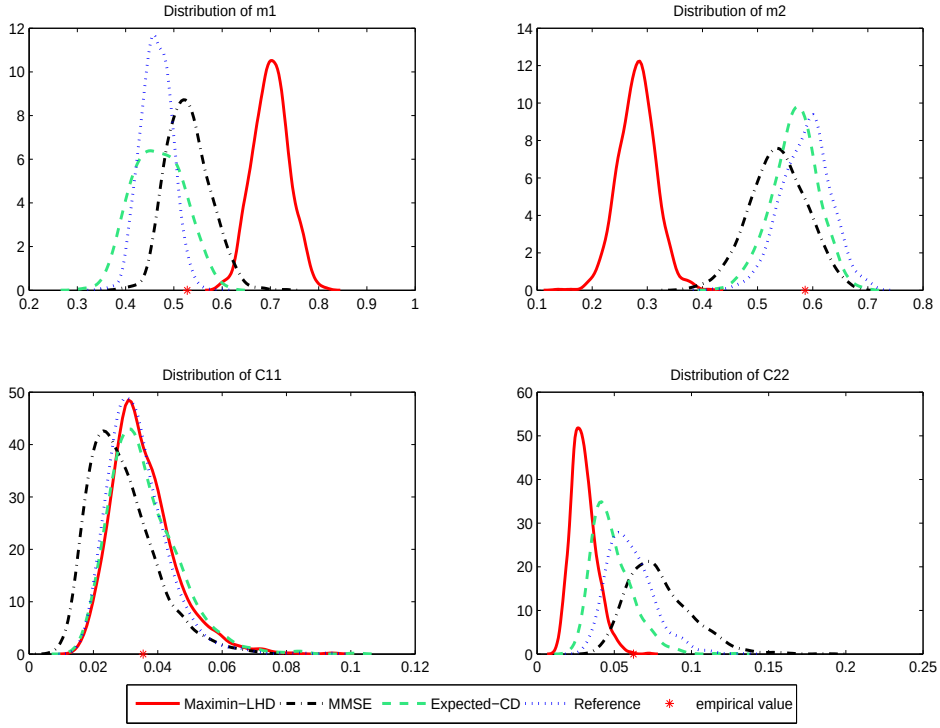


Figure V.3: Posterior distributions of θ with benchmark, standard *maximin*-LHD, MMSE design and E-CD design

well as a standard *maximin*-LHD and the MMSE design under the same computing budget. The so-built DOEs are summarized in Table V.6.2.

.

DOE 1	20-point-*maximin*-LHD	
DOE 2	10-point-*maximin*-LHD +	10-point-W-IMSE
		10-point-E-CD
		10-point-MMSE
DOE 3	500-point-*maximin*-LHD (*benchmark*)	

Table V.2: Description of the three types of DOE for the three-dimensional model

Moreover, the important parameters for the adaptive procedure were as follows:

1. Weighted-IMSE criterion:
   - the weight parameter $\alpha = 0.8$,
   - the number of iterations of the SA algorithm $L = 5,000$,
   - the size of the MC algorithm $M = 5,000$.
2. Expected-CD criterion:
   - the number of the GPs $M = 50$,

- the size of the sample Θ^i and Ψ $L_1 = L_2 = 5,000$,
- the number of iterations of the SA algorithm $L = 5,000$.

Figures V.4-V.6 illustrate the behavior of each criterion. Figure V.4 displays the posterior distribution of θ with respect to the three DOEs (Table V.6.2). We can see both the W-IMSE and E-CD criteria work well to improve the standard space-filling technique, better than the MMSE criterion. However, between the two criteria, it is difficult to say which criterion works better in this example.

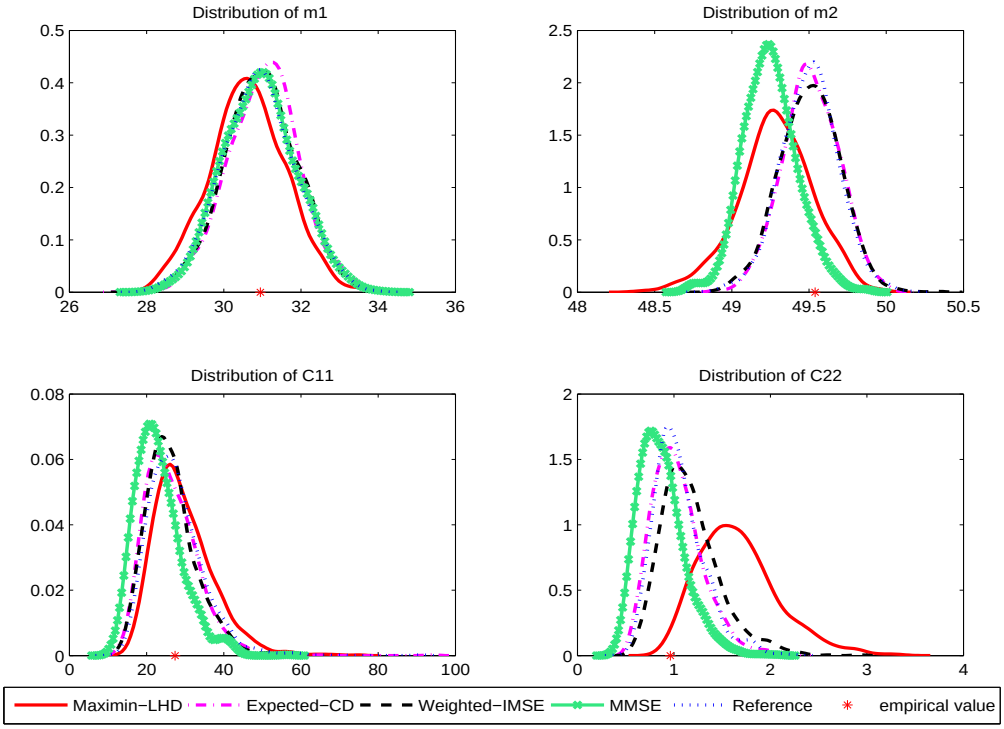


Figure V.4: Posterior distributions of θ with benchmark, standard *maximin*-LHD, W-IMSE ($\alpha = 0.8$)and E-CD designs

Now, we compare the *emulator errors* yielded by the so-built DOEs, using the coefficient of predictibility Q_2 (see Vanderpoorten and Palm, 2001, (111), and Appendix D). In this paper, we use a cross-validation *leave-one-out* version for computational simplicity. The closer Q_2 to 1, the smaller the variance explained by the emulator and the better the quality of the DOE.

Displayed on Figure V.5, the Q_2 coefficient related to the *maximin*-LHD D_{20} equals 0.9745 and the benchmark Q_2 corresponding to the D_{500} equals 0.9933. Moreover, by adding 10 points iteratively to the initial design D_{10} according to our two proposed criteria, we obtain an increasing coefficient Q_2 and the E-CD criterion provides a slightly better Q_2 value.

On Figure V.6, six iso-joint probability densities of $X_i = (X_i^1, X_i^2)$ are displayed, which correspond to their prior and posterior predictive distributions following the five DOEs described in Table V.6.2. The prior predictive distribution can be proved to be the following

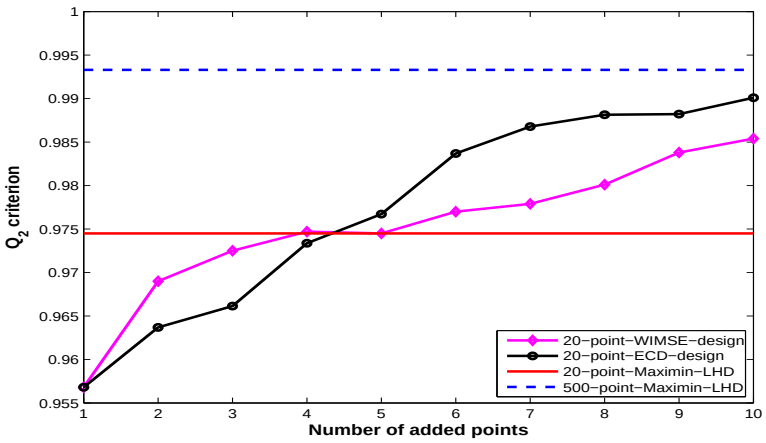


Figure V.5: Comparaison of the quality of different DOEs (Q_2 criterion)

multivariate Student (see Chapter II)

$$X_i \sim \mathrm{St}_q\Big(\mu, \frac{a+1}{a(\nu+1-q)}\Lambda, \nu+1-q\Big).$$

To generate M samples of X_i from their posterior predictive distribution, which means marginalizing the joint posterior distribution $\pi(m, C, \mathbf{X}|\mathbf{y}, \mathbf{d})$ by integrating the parameters $\theta = (m, C)$ out, we propose the following procedure:

1. `simulate` $X_i^{(r)} \sim \mathcal{N}\left(m^{[r]}, C^{[r]}\right)$ `with` $r = 1, \ldots, M$, `where` $m^{[r]}$ `and` $C^{[r]}$ `belong to the final` M `simulated samples from the hybrid MCMC algorithm,`
2. `estimate the probability density` f `of` X_i `from` $\{X_i^{(1)}, \ldots, X_i^{(M)}\}$,
3. `draw the iso-curves of` f.

It confirms that with the help of adaptive procedures, the posterior joint probability of X_i^1 and X_i^2 become more concentrated on the posterior values and the E-CD criterion works better than the IMSE and the MMSE criteria, as it brings iso-curves more similar to those with the benchmark. It is worth noting that on Figure A), we set a quite concentrated prior probability. To conclude, the adaptive designs outperform the standard space-filling DOE with the same number of calls to H and the E-CD criterion seems the most efficient in this example.

V.7 Discussion

This article aims to provide an adaptive methodology to improve the space-filling design of experiments, typically the *maximin*-Latin Hypercube Design, such that the meta-model yields a better trade-off between the reduction of the global uncertainty and the exploration of the

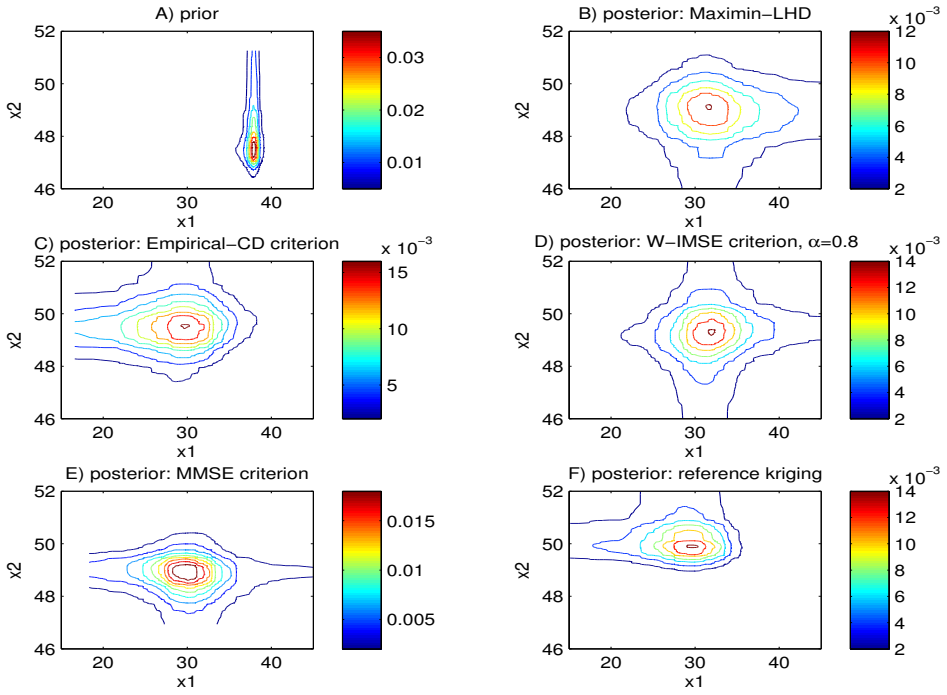


Figure V.6: Posterior cumulative distribution functions of X with different designs

regions of interest. The resulting posterior distributions can thus be improved to provide a more convenient solution to inverse problems.

In this methodology, two adaptive criteria have been proposed to sequentially complete the current design. The first one is a modified version of the standard Weighted-IMSE criterion in Bayesian framework. It is obtained by weighting the MSE term over a region of interest indicated by the current full conditional posterior distributions. The other criterion, called Expected-CD, focuses on minimizing the Kullback-Leibler divergence between the posterior distribution related to the DOE and the desired distribution.

In the second time, numerical experiments have highlighted, on two examples, that applying this adaptive procedure can reduce the prediction error and improve the accuracy of the meta-modeling approximation, compared with the standard space-filling DOE.

Although the two criteria work well in our case study, several limitations and importance points still appear, which also give us directions of further research:

- Both criteria involve expensive numerical integration. The E-CD criterion is a little more expensive than the W-IMSE criterion since it requires the calculation of the empirical KL divergence.
- Furthermore, it is worth noting that in the definition of IMSE, the choice of α is quite important. As the second weight function is globally much smaller that the first prediction error, this balance parameter permits us to find a good behavior of this

criterion. In this paper, we have not systematically studied this important parameter.

- Another point of interest would be the adaptation of the methodology to more complicated hydraulic models, for example, the important MASCARET and TELEMAC-2D codes, both of which are based the French Garonne river and play a critical role in flood risk assessment.

In conclusion, such adaptive procedures can be useful when the CPU time required to compute an occurrence of the simulator H of physical models is dramatically greater than the time required to run a Gibbs sampler, a Monte Carlo integration or to perform an optimization with a Simulated Annealing procedure.

Acknowledgments

This work was partially supported by the French Ministry of Economy in the context of the CSDL (*Complex Systems Design Lab*) project of the Business Cluster System@tic Paris-Région.

Appendix A. The Metropolis-Hastings step inside the Gibbs sampler

At step $r+1$ of Gibbs sampling, after simulating $m^{[r+1]}$,$C^{[r+1]}$, the missing data $\mathbf{X}^{[r+1]}$ have to be updated with a Metropolis-Hasting (MH) algorithm. The MH step is updating $\mathbf{X}^{[r]} = (X_1^r, \ldots, X_n^r)'$ in the following way:

- `For` $i = 1, \ldots, n$
  1. `Generate` $\widetilde{X}_i \sim J(\cdot \mid X_i^r)$ `where` J `is the proposal distribution.`
  2. `Let`

$$\alpha(X_i^r, \widetilde{X}_i) \;=\; \min\Big(\frac{\pi_{\widehat{H}}(\widetilde{\mathbf{X}} \mid \boldsymbol{\mathcal{Y}}, \theta^{[r+1]}, \rho, \mathbf{d}, H_D)\, J(X_i^r|\widetilde{X}_i)}{\pi_{\widehat{H}}(\mathbf{X}^{[r]} \mid \boldsymbol{\mathcal{Y}}, \theta^{[r+1]}, \rho, \mathbf{d}, H_D)\, J(\widetilde{X}_i|X_i^r)}, 1\Big),$$

  where

$$\begin{aligned}\widetilde{\mathbf{X}} &= \Big(X_1^{r+1}, \ldots, X_{i-1}^{r+1}, \widetilde{X}_i, X_{i+1}^r, \ldots, X_n^r\Big)'\\ \mathbf{X}^{[r]} &= \Big(X_1^r, \ldots, X_{i-1}^r, X_i^r, X_{i+1}^r, \ldots, X_n^r\Big)'\end{aligned}$$

  3. `Take`

$$X_i^{r+1} \;=\; \begin{cases} \widetilde{X}_i & \texttt{with probability } \alpha(X_i^r, \widetilde{X}_i),\\ X_i^{r+1} & \texttt{otherwise.}\end{cases}$$

Remarks:

- Many choices are possible for the proposal distribution J. It appears that choosing an independent MH sampler with J chosen to be the normal distribution $\mathcal{N}\Big(m^{[r+1]}, C^{[r+1]}\Big)$ give satisfying results for the model (V.1).
- In practice, it can be beneficial to choose the order of the updates by a random permutation of $\{1, \ldots, n\}$ to accelerate the convergence of the Markov chain to its limit distribution.

Appendix B. Nearest-Neighbor approach

$$\widehat{\mathrm{KL}}_{L_1,L_2}(\Theta^i \,||\, \Psi) \;=\; \frac{d}{L_1}\sum_{j=1}^{L_1} \log\frac{\nu_{L_2}(\theta_j^i)}{\rho_{L_1}^i(\theta_j^i)} + \log\frac{L_2}{L_1 - 1}, \tag{V.21}$$

where d denotes the dimension of the parameter θ ($2q$ in our case), $\nu_{L_2}(\theta_j^i)$ denotes the (Euclidean) distance between $\theta_j^i \in \Theta^i$ and its nearest neighbor in sample Ψ

$$\nu_{L_2}(\theta_j^i) \;=\; \min_{r=1,\ldots,L_2} ||\theta_r - \theta_j^i||_2,$$

and $\rho^i_{L_1}(\theta^i_j)$ denotes the (Euclidean) distance of θ^i_j to its nearest neighbor in sample Θ^i except itself (as it is also included in Θ^i)

$$\rho^i_{L_1}(\theta^i_j) \quad = \quad \min_{l=1,\ldots,L_1; l\neq j} ||\theta^i_l - \theta^i_j||_2.$$

It has been proved in Wang et al. (2006, (115)) that under some regularity conditions on the samples Θ^i and Ψ, the estimator $\widehat{\mathrm{KL}}_{L_1,L_2}(\Theta^i \,||\, \Psi)$ is consistent in the sense that

$$\lim_{L_1,L_2\to\infty} \mathbb{E}\left(\widehat{\mathrm{KL}}_{L_1,L_2}(\Theta^i \,||\, \Psi) - \mathrm{KL}(\Theta^i \,||\, \Psi)\right)^2 \quad = \quad 0, \tag{V.22}$$

and asymptotically unbiased, i.e.

$$\lim_{L,R\to\infty} \mathbb{E}\left[\widehat{\mathrm{KL}}_{L_1,L_2}(\Theta^i \,||\, \Psi)\right] \quad = \quad \mathrm{KL}(\Theta^i \,||\, \Psi). \tag{V.23}$$

Appendix C. Simulated Annealing algorithm (searching for the minimum of a function f)

Given the current point $z^{(k)}$, at iteration $k+1$:

1. Generate $\widetilde{z} \sim \mathcal{N}\left(z^{(k)}, \sigma^2\right)$, with a certain fixed variance σ^2.
2. Let

$$\lambda\left(z^{(k)}, \widetilde{z}\right) \quad = \quad \min\left(1, \exp\left(\frac{f(z^{(k)}) - f(\widetilde{z})}{\beta_{k+1}}\right)\right),$$

   where β_{k+1} is the current temperature at step $k+1$.
3. Accept

$$z^{[k+1]} \quad = \quad \begin{cases} \widetilde{z}, & \text{with probability } \lambda\left(z^{(k)}, \widetilde{z}\right), \\ z^{(k)}, & \text{otherwise.} \end{cases}$$

4. Update $\beta_{k+1} = 0.99 \times \beta_k$.

Appendix D. Coefficient of predictibility Q_2

The cross-validation *leave-one-out* version of the coefficient of predictibility Q_2 (see Vanderpoorten and Palm, 2001, (111)) is

$$Q_2 \quad = \quad 1 - \frac{\mathrm{PRESS}}{\sum_{i=1}^N \left\|H(z_{(i)}) - \overline{H}_{D_N}\right\|^2}.$$

where

$$\overline{H}_{D_N} \quad = \quad \frac{1}{N}\sum_{i=1}^N H(z_{(i)}),$$

and

$$\text{PRESS} \quad = \quad \sum_{i=1}^{N} e_{(i)}^2 \; = \; \sum_{i=1}^{N} \left\| H(z_{(i)}) - \widehat{H}_{-i}(z_{(i)}) \right\|^2$$

with

- $e_{(i)}$ is the prediction error at $z_{(i)}$ of a fitted model without the point $z_{(i)}$;
- $\widehat{H}_{-i}(z_{(i)})$ is the approximation of H at $z_{(i)}$ derived from all the points of the design except $z_{(i)}$.

Remark: Q_2 is related to the ratio of variance explained by an emulator. The closer Q_2 to 1 and the better the quality of the DOE is.

VI

Uncertainty analysis in flood risk assessment

Contents

This chapter is dedicated to the application of the methodologies previously presented to a real case-study of uncertainty management in the field of hydraulic engineering. More precisely, we aim at treating the uncertainty related to the flood risk assessment. Two hydraulic EDF codes are considered in this section: the uni-dimensional MASCARET code and the two-dimensional TELEMAC-2D code.

VI.1 Introduction

The hydraulic engineering models treated in this thesis come from Rocquigny et al. (2008, (23)). The mathematical problem is defined in terms of uncertainty analysis, and can be illustrated in the following diagram.

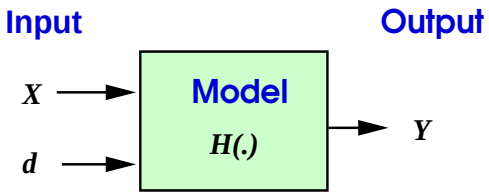


Figure VI.1: Diagram of the problem of uncertainty

In Figure VI.1, we observe the output Y which is related to two types of input with the help of the function H representing the hydraulic model: X denotes a random input with uncertainty and d denotes a fixed input. In the current flood risk assessment, the Garonne river spreading over about 50 km between Tonneins and La Réole is considered. The output of the function H is the water level, the observed input of the model is the flow Q and the most important uncertainty source comes from the Strickler coefficient K_s which measures the friction of the river bed. This coefficient has a physical sense since it is directly related to the water volume from the bottom of the river, which is usually determined by the calibration of the water level and flow. The related uncertainty is both epistemic due to the small sample size used for the calibration and random due to the change of the water level and flow during a flood.

The final aim of the treatment of uncertainty is to predict the risk of dyke overflow during a flood event and so, we are placed in the step B: "Quantification of the uncertainty sources" of the general scheme of treating uncertainties, presented in the Introduction (Figure 1). Taking into account the measurement error, which is considered as the second source of uncertainty, the hydraulic model is as follows:

$$Y = H(K_s, Q) + U. \tag{VI.1}$$

The uncertainty treatment consists of calibrating the Strickler coefficient K_s from the observed couple (Y, d). The hydraulic engineering function H results from the complex St-Venant equation, solved through the finite difference method (for the MASCARET code) and the finite element method (for the TELEMAC-2D code), where the limit conditions are provided by the flow of the river Q.

Many studies have already been addressed to calibrate the non observed Strickler coefficient K_s. In Horrit (2000, (43)), Goutal et al. (2005, (38)) and Bernardara et al. (2008, (8)) for example, K_s was assumed constant and only the uncertainty introduced by the error term U was taken into account. The main idea of the solution is as follows. Given a sample of observations (y_i, d_i) with $i = 1, \ldots, n$, the estimate of the Stickler coefficient K_s is minimizing the cost function $\mathcal{C}$ which is the sum of the squares of the differences between the observations and the results of the model, so-called the *Least squares*:

$$\widehat{K_s} \quad = \quad \underset{K_s}{\operatorname{argmin}} \sum_{i=1}^{n} \Big(y_i - H(K_s, d_i) \Big)^2. \tag{VI.2}$$

This deterministic solution not taking into account the variability of K_s is restrictive and harmful. In this thesis, the variability of K_s is well considered in the Bayesian statistical framework. It would be interesting to compare our calibration results with the existing solutions.

VI.1.1 Uncertainty source in the MASCARET code

The MASCARET code is developed at EDF-LNHE in collaboration with the Centre d'Etudes Techniques Maritimes and River (CETMF), which brings together the computer code of free surface in LNHE. In this uni dimensional model, the Garonne river is described by a hydraulic axis corresponding to the main direction of flow. We are interested in the MASCARET code combined with compound channel, where there are two main sources of uncertainty: the friction on the riverbed and the friction on the floodplain. Both frictions are assumed to be homogeneous with respect to the physical position. Figure VI.2 gives an illustration of the riverbed and the floodplain on a cross-section of a river channel.

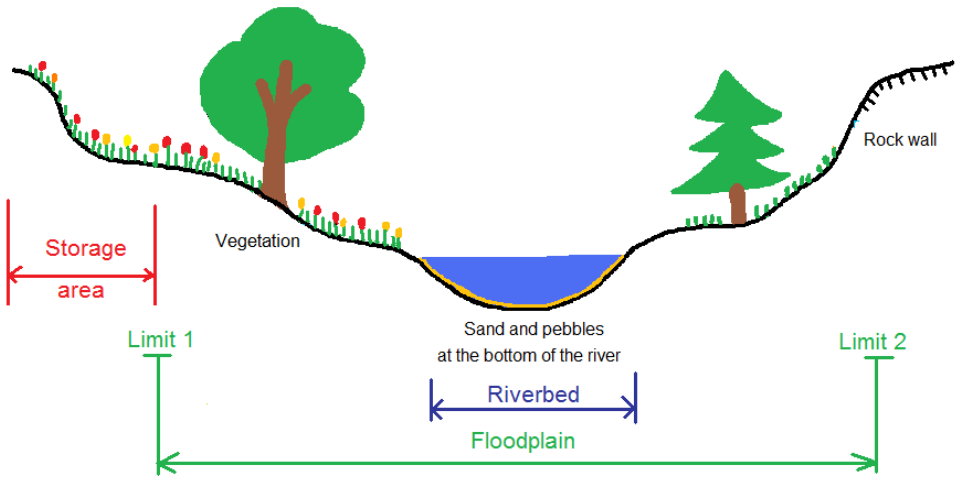


Figure VI.2: Cross-section of a river channel

The riverbed is the main area of flow out of flood periods. The floodplain is the secondary zone of flow in flood, when the dimension of the water passes over the crest of the bank. The third zone is the so-called storage area. It is considered as a reservoir filling up with a flood,

which does not participate in the actual flow since the speed in the direction of the flow axis is assumed zero. Nevertheless it interacts with the flood plain in relation to water extraction. The following notation has been chosen for the two types of friction:

$$\begin{aligned} K_{s,\text{min}} &= \text{Strickler coefficient on the riverbed} \\ K_{s,\text{maj}} &= \text{Strickler coefficient on the floodplain.} \end{aligned}$$

Besides, in MASCARET model, the corresponding kriging domain of inputs can be described as

$$\Omega = \Omega_Q \times \Omega_{\text{maj}} \times \Omega_{\text{min}}. \tag{VI.3}$$

VI.1.2 Uncertainty source in the TELEMAC-2D code

In this two-dimensional TELEMAC model, the river area is divided into a mesh of about 41,000 nodes, as shown in Figure VI.3.

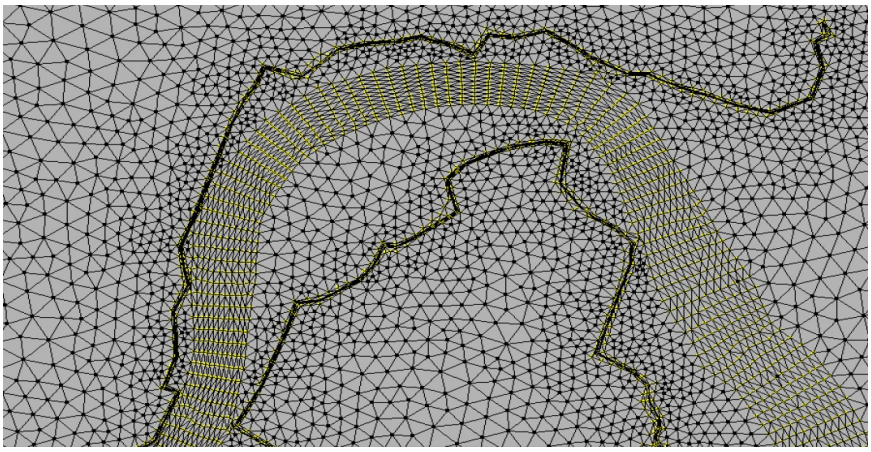


Figure VI.3: Mesh made in the Telemac-2D code

Ideally, the dimension q of the source of uncertainty K_s would be 41,000 to take into account the variability of K_s with each node. However, in practice, the dimension of K_s can be reduced down to four based on the facts that:

1. the topography is relatively homogeneous over large areas around the riverbed of the Garonne;

2. none of the water levels is available at the floodplain at the beginning of the upstream section, which makes impossible to calibrate the Strickler coefficient on the floodplain at this section.

In this case, the Garonne river can be divided into three sections, separated by Tonneins, Mas d'Augenais, Marmande and La Réole, as shown in Figure VI.4.

Figure VI.4: Profiles across the bed of the Garonne river

The variables of interest in this model would be three Strickler coefficients on the riverbed of each section and one global Strickler coefficient on the floodplain. We use the following notation:

$$
\begin{aligned}
K_{s,\min_{TA}} &= \text{Strickler coefficient on the riverbed of section 1 (Tonneins - Mas d'Augenais)} \\
K_{s,\min_{AA}} &= \text{Strickler coefficient on the riverbed of section 2 (Mas d'Augenais - Marmande)} \\
K_{s,\min_{AL}} &= \text{Strickler coefficient on the riverbed of section 3 (Marmande - La Réole)} \\
K_{s,\text{maj}} &= \text{Strickler coefficient on the floodplain of the three sections.}
\end{aligned}
$$

Thus, the kriging domain required by the meta-modeling technique can be noted as:

$$\Omega = \Omega_Q \times \Omega_{\text{maj}} \times \Omega_{\min_{TA}} \times \Omega_{\min_{AA}} \times \Omega_{\min_{AL}}. \tag{VI.4}$$

In the following section, we will discuss how to select a reasonable domain Ω from the available knowledge.

VI.2 Choosing the kriging domain and dyke positions

VI.2.1 Domain of the Strickler coefficients

The Strickler coefficient K_s is well-known to be variant with respect to different types of river channels. A complete international bibliography is available on the subject, and provides the following table.

Nature of the majority surface within the considered area	**Strickler coefficient (in $m^{1/3} \cdot s^{-1}$)**
smooth concrete	75-90
coated bottom (concrete)	70-80
channel in the ground, not grassed	60
channel in the ground, grassed	50
river plain, without shrub	35-40
river plain, sparse vegetation	30
river banks, very narrow vegetated	10-15
river natural bottom	30-50
floodplain grassland	20-30
river bottom cluttered with obstacles	10-30
flood plain	10-30
floodplain in vines or bushes	10-15
urbanized floodplain	10-15
floodplain forest	<10
algal blooms	3.3 - 12.5

Table VI.1: Several orders of magnitude of the Strickler coefficient, taken from (24; 114; 101; 113) and stated in (27).

Taking into account the variability of K_s indicated in Table VI.1, the kriging domain Ω of the Strickler coefficient on the riverbed can be chosen as $[20, 70]$ and for the Strickler coefficient on the floodplain, the domain is a priori selected as $[10, 30]$.

VI.2.2 Domain of the flow of the river

Moreover, frequent values of the flow Q can be found in the note EDF-LNHE (Laboratoire National d'Hydraulique et Environnement) (2008, (7)), where 19 measurements between 1914 and 1987 are available, summarized in Figure VI.5 with its fitting distribution. A large enough empirical domain $[700, 4800]$ for this fixed input can thus be reasonably chosen.

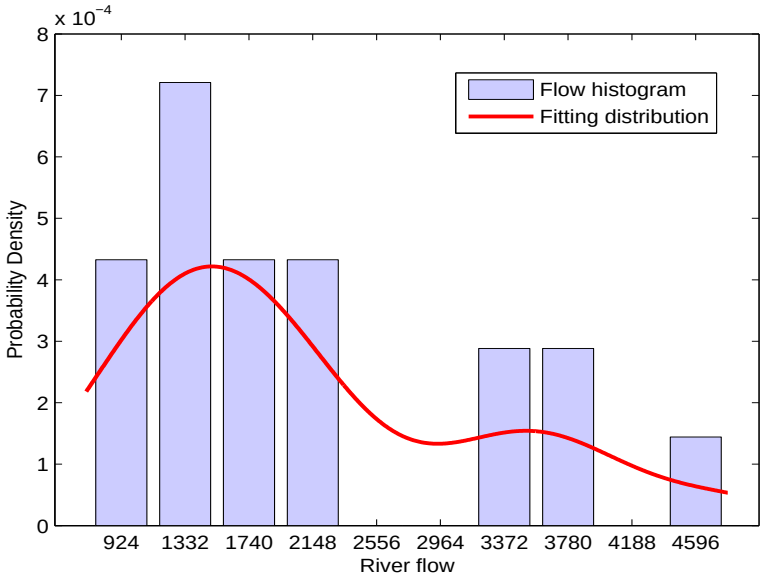


Figure VI.5: Flow histogram of the observations (1914-1987).

Finally, summarizing the domain choice and the diagram in Figure VI.5, the kriging domain Ω can be specified as :

$$\Omega \quad = \quad [700, 4800] \times [10, 30] \times [20, 70] \tag{VI.5}$$

for the MASCARET code and

$$\Omega \quad = \quad [700, 4800] \times [10, 30] \times [20, 70] \times [20, 70] \times [20, 70], \tag{VI.6}$$

for the TELEMAC-2D code.

VI.2.3 Dyke positions - Sensitivity analysis

In the hydraulic model, the water level Y is observed at several dyke positions and composed as a vector of outputs. In the note EDF-LNHE (2008, (7)), a sensitivity analysis has been provided, which shows that among the four stations of Figure VI.4, at Marmande and Tonneins the water level is more sensitive to the variability of the Strickler coefficient K_s. These two stations are thus chosen to ensure that the inverse problem is well-posed to be free of identifiability problem.

The x and y coordinates of the stations are:

$$\begin{aligned} \mathcal{O}(\text{Marmande}) &= (426627, 246567)\,, \\ \mathcal{O}(\text{Tonneins}) &= (438007, 234088)\,. \end{aligned} \tag{VI.7}$$

Here we provide a more general test of sensitivity. We generate M samples of the Strickler coefficient $\mathbf{K}_s = \{K_s^1, \ldots, K_s^M\}$ from the normal distribution

$$K_s^i \quad \sim \quad \mathcal{N}\left(\mu, C_{\text{Exp}}\right), \tag{VI.8}$$

with $i = 1, \ldots, M$ and M samples of the river flow $\mathbf{Q} = \{Q^1, \ldots, Q^M\}$ from the Gumbel distribution with mean 2000 and standard deviation 1000. It is worth noting that the prior choice of $\mathbf{Q}$ is coherent with the histogram shown in Figure VI.5 and its fitting distribution.

At each random position p, M observations of the water level $\mathbf{Y}_p = \{Y_p^1, \ldots, Y_p^M\}$ can be obtained with each component

$$Y_p^i = \delta_p \circ H(K_s^i, Q^i) + U^i, \tag{VI.9}$$

where

$$\delta_p = \begin{cases} 1, & \text{if it is observed at the position p;} \\ 0, & \text{else.} \end{cases} \tag{VI.10}$$

If the empirical variance is such that $\mathrm{Var}(\mathbf{Y}_p) > \mathrm{Var}(\mathbf{Y}_{p'})$, we can say that the position p is better than the position p' to measure the water level, as it is more sensitive to the Strickler coefficient K_s as well as the flow of river Q.

The numerical results of the sensibility analysis are summarized in Figure VI.6.

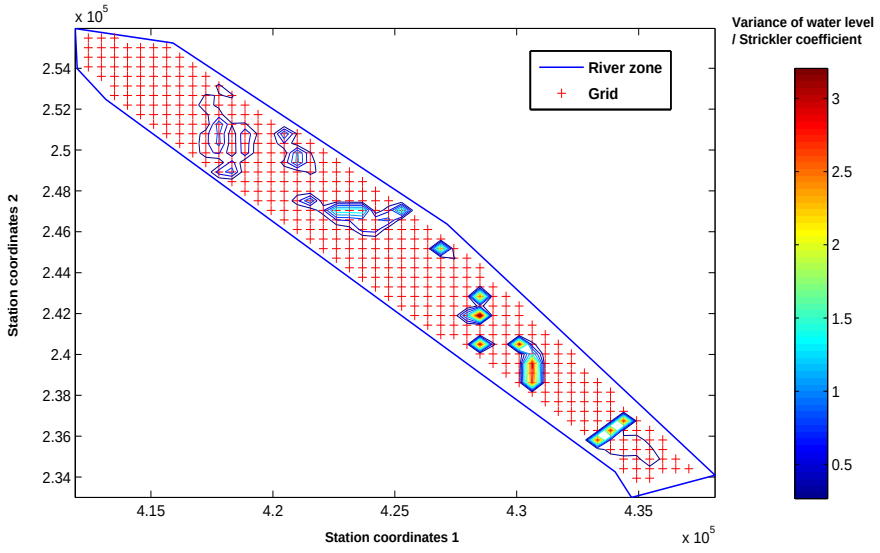


Figure VI.6: Variability of the water level / Strickler coefficient of the Garonne river.

One can see that the most sensitive positions are near the upstream of the river. Under the assumption that the water level depends only on the Strickler coefficient at its downstream, it is understandable that at La Réole, the most downstream dyke position (see Figure VI.2), the water level is not at all sensitive to the Strickler coefficient.

According to the results of our sensitivity analysis, two dyke quite sensitive but not too close to each other positions have been chosen:

$$\begin{aligned} \mathcal{O}(p_1) &= (428493, 242837), \\ \mathcal{O}(p_2) &= (434384, 236751)\,. \end{aligned}$$

Remark 26. *The two chosen positions are not quite far from the Marmande and Mas d'Agenais described in* (VI.7)*, according to the note EDF-LNHE (2008,* (7)*)*

VI.3 Eliciting the prior distributions

In the Bayesian framework, apart from the observed data, prior information and expert knowledge can be taken into account to improve the estimation. In the hydraulic applications, several types of information sources are available. This section addresses the issue of the prior elicitation of hyperparameters $\rho = (\mu, a, \Lambda, \nu)$ from the available information.

VI.3.1 Statistical modeling

In Chapter II, it has been proven that the prior predictive distribution of K_s, which is the marginal distribution of K_s by integrating the mean and variance parameters out, is the following multivariate Student distribution:

$$K_s \sim \mathrm{St}_q\Big(\mu, \frac{a+1}{a(\nu+1-q)}\Lambda, \nu+1-q\Big). \tag{VI.11}$$

Under the assumptions that

$$\Lambda = t \cdot C_{\mathrm{Exp}} \tag{VI.12}$$
$$t = \nu - q - 1, \tag{VI.13}$$

(VI.11) leads to

$$\mathbb{E}[K_s] = \mu, \tag{VI.14}$$
$$\mathrm{Var}[K_s] = \frac{a+1}{a} C_{\mathrm{Exp}}, \tag{VI.15}$$

where C_{Exp} denotes the prior variance matrix of K_s which represents the expert opinion (see Section II.2.2). These hyperparameters can thus be elicited from the predictive distribution.

In hydraulic models, the Strickler coefficient K_s on the floodplain can be reasonably a priori assumed to be independent from the friction on the riverbed, as the topographic areas are of quite different nature (by the shrub-land, the non-vegetated land...) For the TELEMAC-2D code, two Strickler coefficients on the riverbed are most likely correlated. However, as no prior information is available on this correlation and only the "marginal" knowledge about the frictions is known, all the Strickler coefficients are a priori assumed to be independent[1]. In summary, the prior variance matrices for the two models are assumed to be diagonal under the following form:

$$C_{\mathrm{Exp}} = \begin{pmatrix} \sigma^2_{\mathrm{maj}} & 0 \\ 0 & \sigma^2_{\mathrm{min}} \end{pmatrix}, \tag{VI.16}$$

[1] See numerical experiments in Section VI.4.2 for the validation of this assumption.

for the MASCARET code and

$$C_{\text{Exp}} = \begin{pmatrix} \sigma^2_{\text{maj}} & 0 & 0 & 0 \\ 0 & \sigma^2_{\text{min}_{TA}} & 0 & 0 \\ 0 & 0 & \sigma^2_{\text{min}_{AA}} & 0 \\ 0 & 0 & 0 & \sigma^2_{\text{min}_{AL}} \end{pmatrix}, \tag{VI.17}$$

for the TELEMAC-2D code. The next two subsections are addressed to select the prior mean μ and the prior variance C_{Exp}.

VI.3.2 prior calibration of μ and a from expert knowledge

In this section, we focus on calibrating the prior mean μ of the variable of interest K_s. We begin with the Strickler coefficient on the floodplain.

As presented in Section VI.1, the knowledge on the friction on the floodplain is generally not available by lack of data. However, a methodology developed by EDF, carrying out the calculations of Cotes Majorées de Sécurité (CMS), has succeeded in determining this coefficient. In the case of the Garonne river, an empirical value of 17 $m^{1/3} \cdot s^{-1}$ has been proposed by Besnard et al. (2008, (6; 7)), which can be considered as the prior mean for the friction on the floodplain. This value is characteristic of areas with low vegetation.

On the other hand, several types of information allow us to assess the Strickler coefficients on the riverbed. Table VI.1 in the Section VI.2 summarizes some related documented values. Based on numerical tests and preliminary studies, Besnard et al. (2008 (6; 7)) proposed the following estimates of the frictions on the riverbed, which can also be considered as the prior values in a Bayesian approach:

- section 1 (Tonneins - Mas d'Augenais) : 45 $m^{1/3}.s^{-1}$
- section 2 (Mas d'Augenais - Marmande) : 38 $m^{1/3}.s^{-1}$
- section 3 (Marmande - La Réole) : 40 $m^{1/3}.s^{-1}$

The prior value for the mean μ can be found in Section VI.3.4. Moreover, the size a of the *virtual sample* (refer to Section II.2.2 for the interpretation) can be chosen equal to 1, as at least one sample is required to specify a median/mode/mean. It is worth noting that this choice is variable, which depends on our confidence on the expert judgement, the complexity of the variable of interest K_s and so on. The last hyperparameter to calibrate is t. As explained in Section II.2.2, it can be chosen as:

$$t = a + 1. \tag{VI.18}$$

VI.3.3 prior calibration of C_{Exp} through statistical analysis

In this section, we aim at calibrating the prior variance of K_s which measures the uncertainty brought by this Strickler coefficient. The prior knowledge is available under the form of the

calibration curve *average - standard deviation* of the Manning coefficient, proposed by U.S. Army Corps of Engineers (1996, (110)) and recalled in Liu (2009, (61)). It is shown in Figure VI.7.

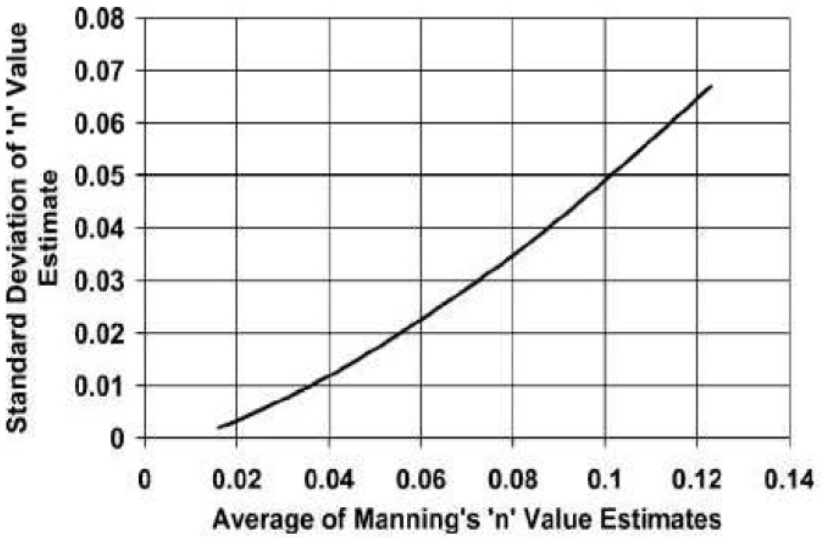


Figure VI.7: Uncertainty in the estimates of Manning coefficient $M = 1/K_s$, Figure 3.5 in (61), originally from (110).

The main idea is to calibrate the variance C_{Exp} by deriving the mean-variance relationship on K_s with the help of the mean and variance of M, proceeding as follows. Let us note that $\widehat{M}$ is the estimator of M based on observed data, σ_M is its proper prior standard deviation, $K_s = 1/M$ denotes the corresponding Strickler coefficient and $\widehat{K_s}$ denotes its prior estimator. Under the assumption that the sample size is large enough, the Central Limit Theorem (CLT) leads to the following convergence in distribution

$$\sigma_M^{-1}\left(\widehat{M} - M\right) \quad \xrightarrow{\mathcal{L}} \quad \mathcal{N}(0,1)\,, \tag{VI.19}$$

which can be transformed into the convergence of $\widehat{K_s}$ by applying the Delta method. In detail, with the help of the function $f(x) = 1/x$ and $f'(x) = -1/x^2$, the Delta method states

$$\sigma_M^{-1}\left(f(\widehat{M}) - f(M)\right) \quad \overset{\mathcal{L}}{\rightsquigarrow} \quad \mathcal{N}\left(0, \left[f'(M)\right]^2\right) \tag{VI.20}$$

$$\iff \quad \sigma_M^{-1}\left(\widehat{M}^{-1} - M^{-1}\right) \quad \overset{\mathcal{L}}{\rightsquigarrow} \quad \mathcal{N}\left(0, M^{-4}\right) \tag{VI.21}$$

$$\iff \quad M^2\sigma_M^{-1}\left(\widehat{M}^{-1} - M^{-1}\right) \quad \overset{\mathcal{L}}{\rightsquigarrow} \quad \mathcal{N}(0,1)\,. \tag{VI.22}$$

As $M = 1/K_s$, it is equivalent to

$$K_s{}^{-2}\sigma_M^{-1}\left(\widehat{K_s} - K_s\right) \quad \overset{\mathcal{L}}{\rightsquigarrow} \quad \mathcal{N}(0,1)\,, \tag{VI.23}$$

which leads to the following approximation:

$$\mathrm{Var}[\widehat{K_s}] \quad \simeq \quad \widehat{K_s}^4\sigma_M^2. \tag{VI.24}$$

By replacing $\widehat{K_s}$ by the prior mean μ and using Figure VI.7 to fix the value of σ_M^2, the predictive prior variance of K_s can be computed as follows.

$$\mathrm{Var}[K_s] = \begin{pmatrix} 34 & 0 \\ 0 & 100 \end{pmatrix}, \tag{VI.25}$$

for the MASCARET code and

$$\mathrm{Var}[K_s] = \begin{pmatrix} 34 & 0 & 0 & 0 \\ 0 & 100 & 0 & 0 \\ 0 & 0 & 100 & 0 \\ 0 & 0 & 0 & 100 \end{pmatrix}, \tag{VI.26}$$

for the TELEMAC-2D code. Moreover, using $a = 1$ and (VI.15) can lead to the prior value of the matrix C_{Exp}, which will be detailed in the next section.

VI.3.4 Summary of the prior elicitation

In summary, we propose in Table VI.2 the elicitation of the hyperparameters with respect to the two hydraulic models.

Hyperparameters	MASCARET	TELEMAC-2D
μ	$\begin{pmatrix} 17 \\ 40 \end{pmatrix}$	$\begin{pmatrix} 17 \\ 45 \\ 38 \\ 40 \end{pmatrix}$
a	1	1
t	2	2
ν	5	7
C_{Exp}	$\begin{pmatrix} 4.1^2 & 0 \\ 0 & 7.1^2 \end{pmatrix}$	$\begin{pmatrix} 4.1^2 & 0 & 0 & 0 \\ 0 & 7.1^2 & 0 & 0 \\ 0 & 0 & 7.1^2 & 0 \\ 0 & 0 & 0 & 7.1^2 \end{pmatrix}$

Table VI.2: Prior calibration of the hyperparameters for two models

VI.4 Numerical experiments

It is worth mentioning in this section that, the experimental trials validating our methodology have been carried out from data simulated from the true physical code.

The first example is the uni dimensional MASCARET model with two Strickler coefficients to estimate, and the second example is the more complex TELEMAC-2D model with four Strickler coefficients. For each one, two *maximin*-LHDs with 20 and 200 points are generated and there are two data samples with 10 and 50 observations. Moreover, in order to make the

prior assumptions on the model parameters described in Table VI.2 being one *perfect* and one *fair* priors, we consider two kinds of distributions to generate the data sample. The next two subsections provide their formulas.

VI.4.1 First model: the MASCARET code

The first experiments aim at assessing the ability of the Bayesian methodology to calibrate the uncertainty in the MASCARET code, accounting for the impacts of the prior assumptions, the sample sizes and the quality of the DOEs.

Under the assumption of the "perfect prior", the missing data K_s is simulated from

$$K_s \sim \mathcal{N}\left(\mu, C_{\text{Exp}}\right) = \mathcal{N}\left(\begin{pmatrix} 17 \\ 40 \end{pmatrix}, \begin{pmatrix} 4.1^2 & 0 \\ 0 & 7.1^2 \end{pmatrix}\right). \tag{VI.27}$$

By adding a small difference Δ to the mean the the variance term, we generate K_s for the "fair prior" that

$$K_s \sim \mathcal{N}\left(\mu + \Delta_\mu, C_{\text{Exp}} + \Delta_{C_{\text{Exp}}}\right) = \mathcal{N}\left(\begin{pmatrix} 18 \\ 35 \end{pmatrix}, \begin{pmatrix} 3^2 & 0 \\ 0 & 5^2 \end{pmatrix}\right). \tag{VI.28}$$

Remark 27. *Under the "fair prior" assumption, we prefer a smaller variance in the data generation than the prior variance C_{Exp} as it is natural to assume less uncertainty introduced by the observation than the prior elicitation.*

The observation samples $\mathbf{y}_n = \{y_1, \ldots, y_n\}$ with the sample size $n = 10$ and 50 can thus be generated from

$$Y = H(K_s, Q) + U, \tag{VI.29}$$

with $Q \sim \text{Gumbel}(1550, 780)$ and H the MASCARET code which is an approximating solution to the St-Venant equation through finite difference methods.

Remark 28. *In the Gumbel distribution, the parameters 1550 and 780 have been chosen according to the empirical mean* $2,000$ *and the empirical standard deviation* $1,000$ *of Q, as explained in Section VI.2.3 (cf. Figure VI.5).*

Figure VI.8 displays the marginal posterior distributions of θ with the "perfect prior" assumption. Four cases have been analyzed with respect to the sample size $n = 10, 50$ and the number of points of the *maximin*-LHD $N = 20,200$. The normal distribution $\mathcal{N}(m^{[r+1]}, C^{[r+1]})$ has been chosen as the instrumental distribution to simulate $X^{[r+1]}$ in the MH algorithm. The curve with respect to D_{200} and 50 observations seems to perform the best, which is understandable as it profits of the largest amount of information and is closed to the "true" value.

Figure VI.9 shows the posterior distributions of θ under the same experimental conditions except for the "fair prior" assumption. The curve with D_{200} and 50 observations works the best whereas the difference with other curves become less significant. Moreover, we remark that as expected, the posterior distributions with a "fair prior" are globally less concentrated than with a "perfect prior".

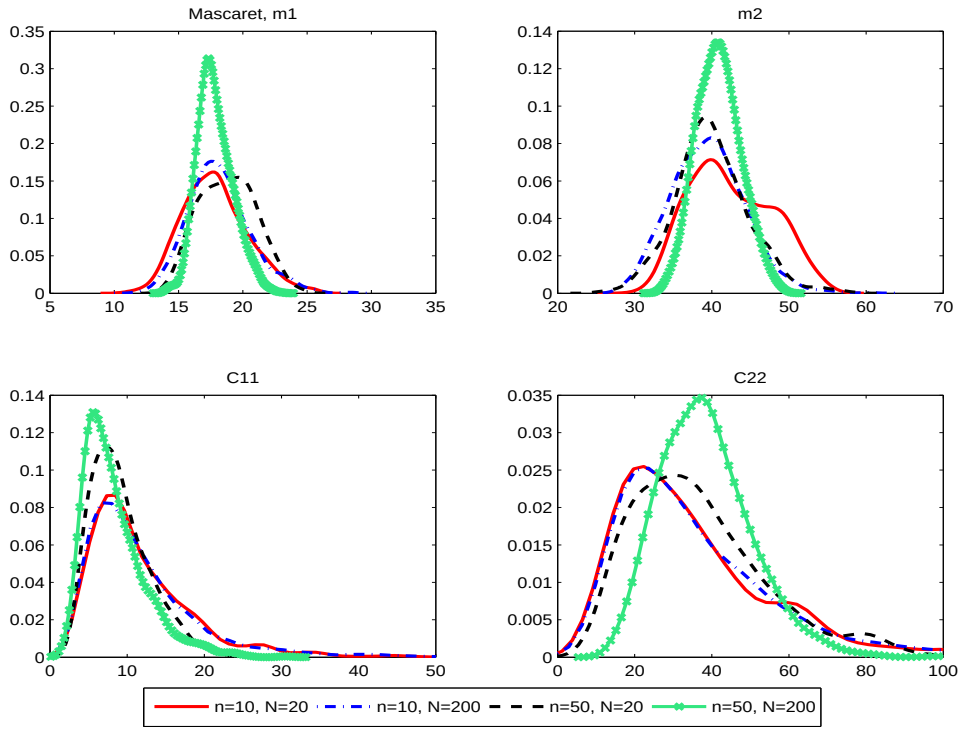


Figure VI.8: Posterior distributions of θ with perfect prior, the sample size $n = 10, 50$ and *maximin*-LHDs D_{20}, D_{200}

VI.4.2 Second model: the TELEMAC-2D code

In the second time, we focus on validating our Bayesian methodology in uncertainty treatment with the help of the TELEMAC-2D code. Alike in the first example, we simulate the observation samples in the following way. Under the "perfect prior" assumptions, the missing data K_s is simulated from

$$K_s \sim \mathcal{N}\left(\mu, \tilde{C}_{\text{Exp}}\right) = \mathcal{N}\left(\begin{pmatrix} 17 \\ 45 \\ 38 \\ 40 \end{pmatrix}, \begin{pmatrix} 4.1^2 & 0 & 0 & 0 \\ 0 & 7.1^2 & 0.2 \times 7.1^2 & 0.2 \times 7.1^2 \\ 0 & 0.2 \times 7.1^2 & 7.1^2 & 0.2 \times 7.1^2 \\ 0 & 0.2 \times 7.1^2 & 0.2 \times 7.1^2 & 7.1^2 \end{pmatrix}\right), \tag{VI.30}$$

and by introducing a difference Δ to the prior distribution (referring to Table VI.2), we generate K_s under the "fair prior" assumption:

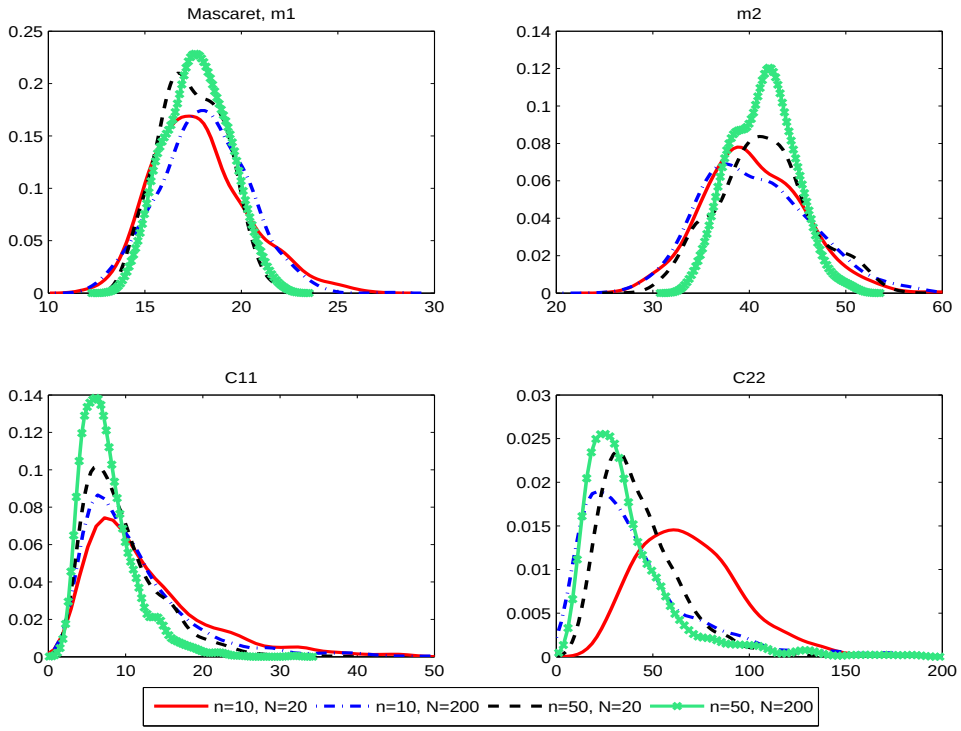


Figure VI.9: Posterior distributions of θ with fair prior, the sample size $n = 10, 50$ and *maximin*-LHDs D_{20}, D_{200}

$$K_s \sim \mathcal{N}\left(\mu + \Delta_\mu, \tilde{C}_{\text{Exp}} + \Delta_{\tilde{C}_{\text{Exp}}}\right) = \mathcal{N}\Big(\begin{pmatrix} 18 \\ 42 \\ 35 \\ 44 \end{pmatrix}, \begin{pmatrix} 4.1^2 & 0 & 0 & 0 \\ 0 & 7.1^2 & 0.2 \times 7.1^2 & 0.2 \times 7.1^2 \\ 0 & 0.2 \times 7.1^2 & 7.1^2 & 0.2 \times 7.1^2 \\ 0 & 0.2 \times 7.1^2 & 0.2 \times 7.1^2 & 7.1^2 \end{pmatrix}\Big). \tag{VI.31}$$

Remark 29. *A correlation equal to 0.2 between $K_{s,min}$s has been assumed under both assumptions as in the reality, the frictions on the riverbed is often correlated. This causes a variance different from C_{Exp} described in Table VI.2. That is why it is noted $\tilde{C}_{Exp}$ instead of C_{Exp}.*

Similarly, two observation samples $\mathbf{y}_n = \{y_1, \ldots, y_n\}$ of size $n = 10$ and 50 are generated from

$$Y_p = \delta_p \circ H(K_s, Q) + U, \tag{VI.32}$$

where $Q \sim \text{Gumbel}(1550, 780)$, H is the TELEMAC-2D code as an approximating solution to the St-Venant differential equation through finite element methods and p denotes the dyke position chosen to observe the river level Y. As presented in Section VI.2.3, two dyke positions are chosen according to our sensitivity analysis, which are not far from the Marmande and Mas d'Agenais. In the following experiments, the two observations positions are chosen to be Marmande and Mas d'Agenais for some reliability reasons.

In Figure VI.10, four marginal posterior distributions of θ under the assumption of "perfect prior" are displayed, with respect to the sample size $n = 10, 50$ and two *maximin*-LHDs D_{20} and D_{200}. In the MH algorithm, the instrumental distribution has been chosen as the normal distribution $\mathcal{N}(m^{[r+1]}, C^{[r+1]})$. We find that the quality of posterior distributions are influenced by both the observations and the choice of DOEs.

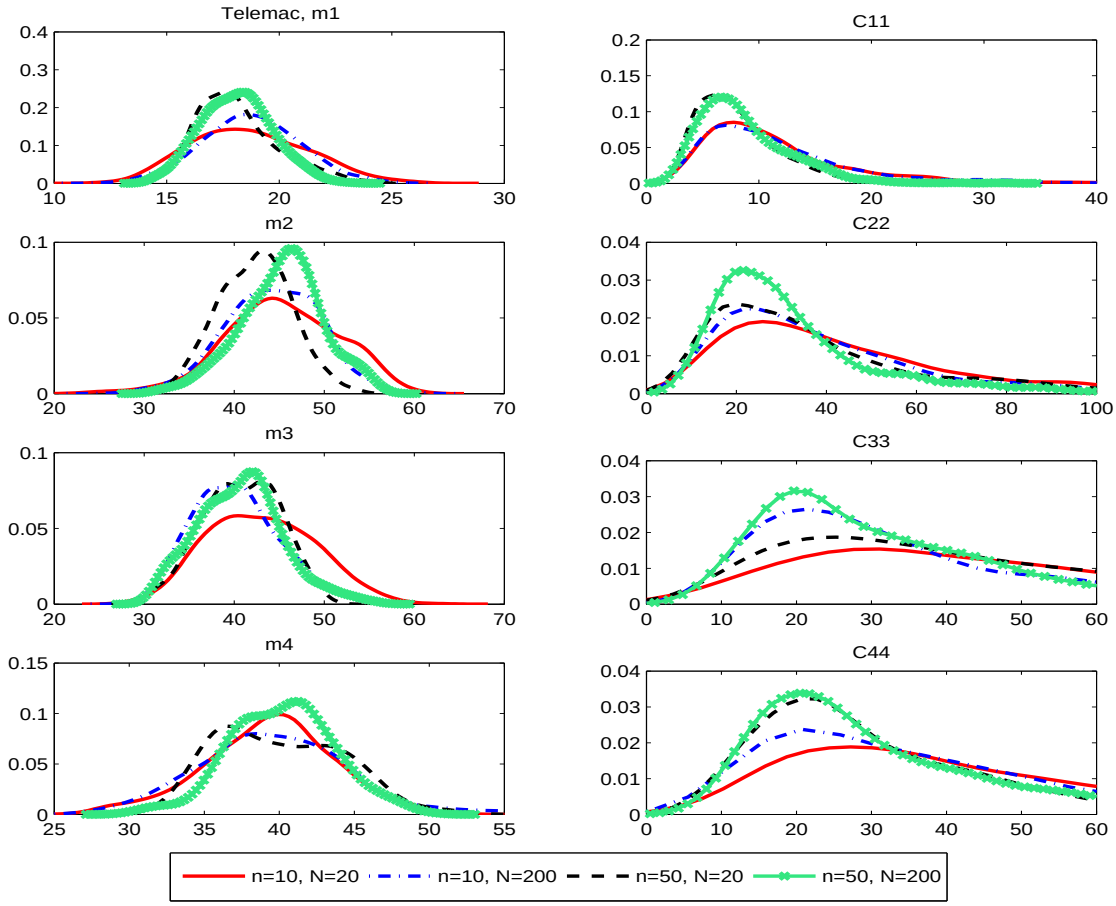


Figure VI.10: Posterior distributions of θ with perfect prior, the sample size $n = 10, 50$ and *maximin*-LHDs D_{20}, D_{200}

Figure VI.11 shows the posterior distributions of θ with the "fair prior" assumption. Different from the MASCARET code, in this example, the performance of the MCMC algorithm under the "fair prior" assumption is quite similar to that under the "perfect prior" assumption. This indifference to the prior distributions makes reasonable the application of the Bayesian inference in the TELEMAC-2D code.

VI.4.3 Test: Checking the $\widetilde{\text{DAC}}$ criterion

This section focuses on verifying the performance of the $\widetilde{\text{DAC}}$ criterion, which is applied to assess the prior and design relevance with respect to the observations (cf. Chapter IV). For the purpose of simplicity, the simpler computational code MASCARET has been chosen.

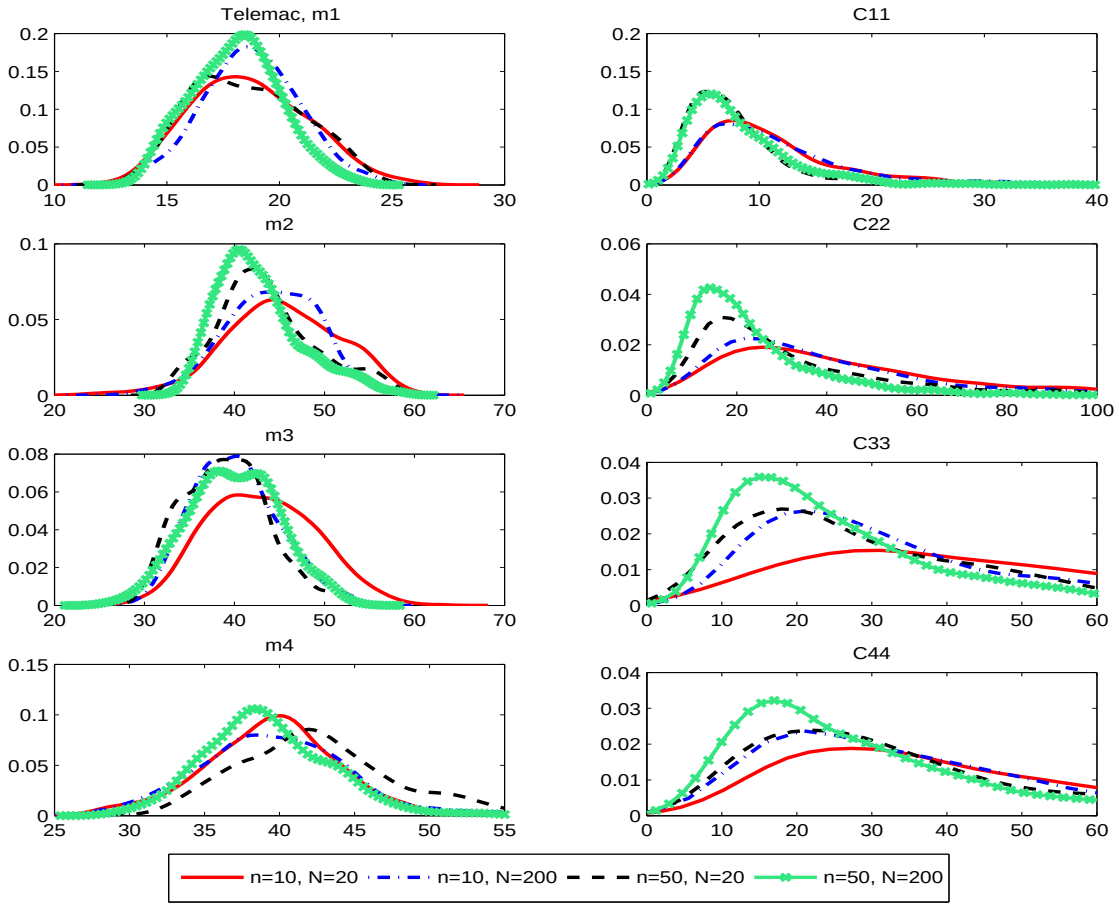


Figure VI.11: Posterior distributions of θ with fair prior, the sample size $n = 10, 50$ and *maximin*-LHDs D_{20}, D_{200}

To that end, the first step is to display a Gibbs sampling to obtain a sample $\theta^r \sim \pi^J(\cdot|\mathbf{y}_n, H_{D_N})$ $(r = 1, \ldots, R)$ with a non informative prior π^J, chosen here as the Jeffreys prior:

$$\pi^J(\theta) \quad = \quad \frac{\mathbf{I}_{\Omega_m}(m)}{\mathrm{Vol}(\Omega_m)} \cdot \frac{\Delta_C}{|C|^{\frac{q+2}{2}}} \mathbf{I}_{\Omega_C}(C). \tag{VI.33}$$

Choosing an instrumental distribution The Gibbs sampling with a Jeffreys non informative prior has been described in Appendix A in Chapter IV. In the MH step which updates the missing data from $\mathbf{X}^{[r]}$ to $\mathbf{X}^{[r+1]}$, the following two instrumental distributions have been considered (referring to Chapter III)

$$J_1 : \quad \mathcal{N}\left(m^{[r+1]}, C^{[r+1]}\right), \tag{VI.34}$$

$$J_2 : \quad \mathcal{N}\left(\mathbf{X}^{[r]}, 5C^{[r+1]}\right). \tag{VI.35}$$

Similarly to previous examples, the sample size n has been fixed to be 10 and 50, and two *maximin*-LHDs D_{20} and D_{200} have been chosen. A sample of the posterior distribution of θ

is obtained by the Gibbs sampler once the convergence has been reached. The corresponding "posterior" distributions are displayed in Figures VI.12 and VI.14. It is worth noting that for J_1, the quotation marks are necessary. In fact, with the first instrumental distribution which seems to work badly, it is quite difficult to reach the convergence. Figure VI.12 represents only the current distributions of θ after 30 000 iterations of Gibbs sampling.

Its bad behavior has been confirmed in Figure VI.13. We provide an example of the simulation of m_1. Based on 50 observations and the *maximin*-LHD D_{20}, our 10 parallel simulated chains behaved all differently and the BG statistics remained high.

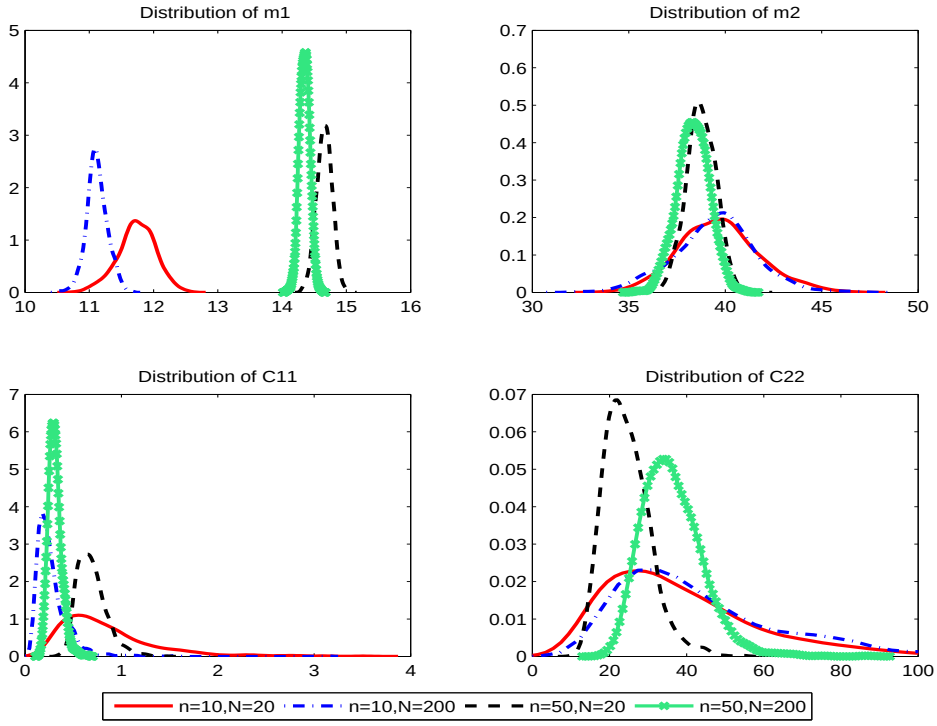


Figure VI.12: Posterior distributions of θ with Jeffreys non informative prior and the instrumental distribution J_1, the sample size $n = 10, 50$ and *maximin*-LHDs D_{20}, D_{200}

In the contrary, as shown in Figure VI.14, J_2 provides good results for all the parameters. The convergence of the simulated Markov chains has been checked in Figure VI.14. These figures illustrate the importance of choosing a suitable instrumental distribution. In the following, we choose the second instrumental distribution J_2 to calculate the $\widetilde{\text{DAC}}$ criterion.

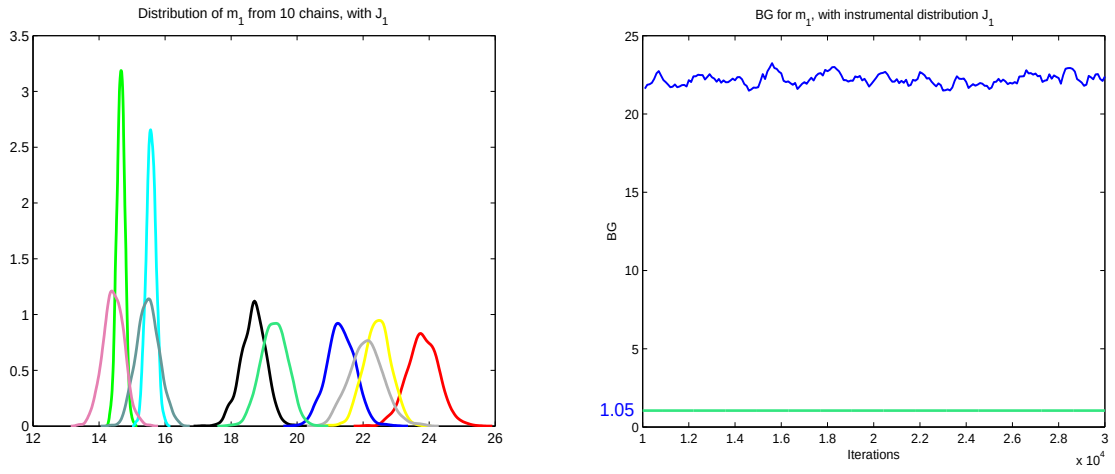


Figure VI.13: Distributions of m_1 and the corresponding BG statistics, resumed from 10 parallel chains, with 50 observed data, the *maximin*-LHDs D_{20} and the instrumental distribution J_1, for the MASCARET code

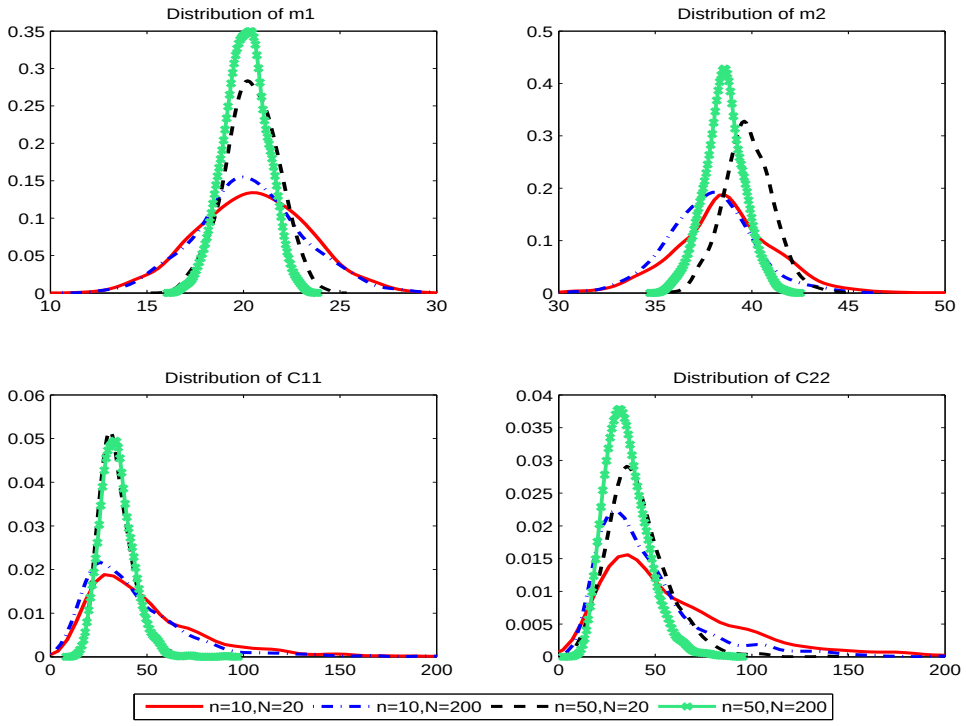


Figure VI.14: Posterior distributions of θ with Jeffreys non informative prior and the instrumental distribution J_2, the sample size $n = 10, 50$ and *maximin*-LHDs D_{20}, D_{200}

Using the $\widetilde{\text{DAC}}$ criterion To check the behavior of $\widetilde{\text{DAC}}$, two different prior distributions on the MASCARET model hyperparameters are considered and summarized in Table VI.3 (cf. Chapter II for the meaning of the hyperparameters).

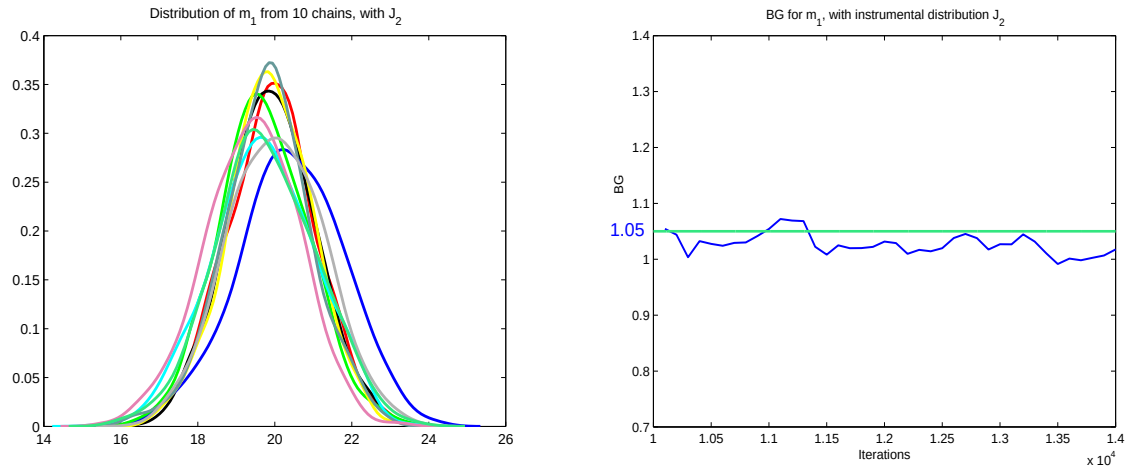


Figure VI.15: Distributions of m_1 and the corresponding BG statistics, resumed from 10 parallel chains, with 50 observed data, the *maximin*-LHDs D_{20} and the instrumental distribution J_2, for the MASCARET code

Prior	**PMV**	**BLV**
μ	$\{17, 40\}$	$\{5, 60\}$
a	1	1
t	2	2
ν	5	5
$\widetilde{\mathbf{C}}_{\text{Exp}}$	$\begin{pmatrix} 4.1^2 & 0 \\ 0 & 7.1^2 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

Table VI.3: Description of the two prior distributions: **PMV** denotes perfect mean and medium variance, **BLV** denotes bad mean and low variance.

Figure VI.16 displays the behaviors of $\widetilde{\text{DAC}}$ for 20 repetitions with the **PMV** and **BLV** priors, in four different cases with the sample size $n = 10, 50$ and the *maximin*-LHDs D_{20} and D_{200}. It appears that the "bad" value **BLV** prior is rejected by $\widetilde{\text{DAC}}$ in all the four cases as this criterion remains positive, and it seems almost acceptable for the last case with 50 observed data and 200 points in design, as $\widetilde{\text{DAC}}$ is quite near zero. Moreover, the **PMV** prior is obviously acceptable in each case study, thanks to the negative $\widetilde{\text{DAC}}$s, which shows a perfect agreement between the prior, the data and the design.

Figure VI.17 displays the marginal posterior distributions of θ with respect to the two priors, based on 10 and 50 observed data and two *maximin*-LHDs D_{20} and D_{200}. It confirms the performance of $\widetilde{\text{DAC}}$. Just as it indicated, the **BLV** prior remains far from the "supposed true" values, according to which our observations have been generated, especially for C_{11} and C_{22}. However, the **PMV** prior provides us reasonable posterior values in all of the four cases.

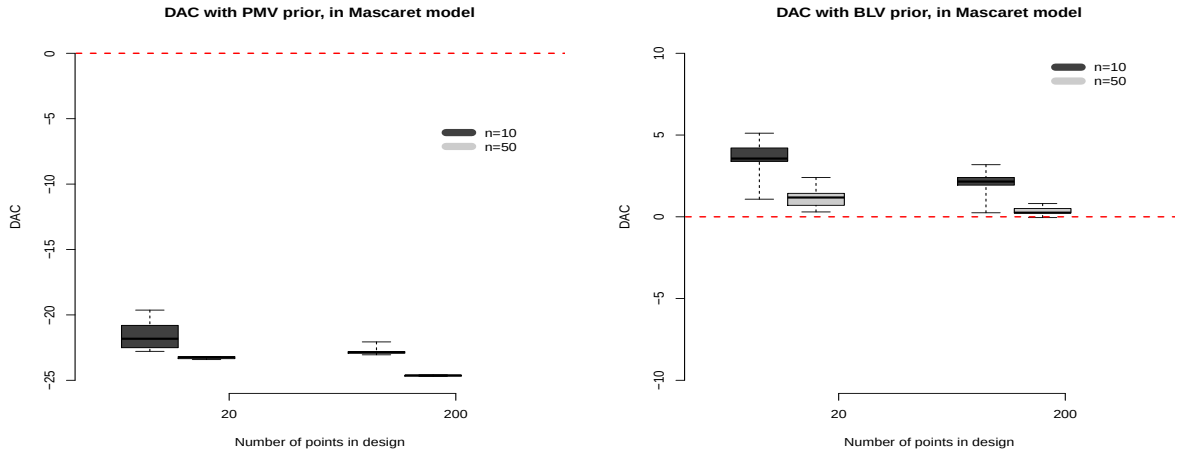


Figure VI.16: $\widetilde{\mathrm{DAC}}$ with **PMV** and **BLV** priors, based on $n = 10, 50$ observations and two *maximin*-LHDs D_{20} and D_{200}, in the MASCARET code

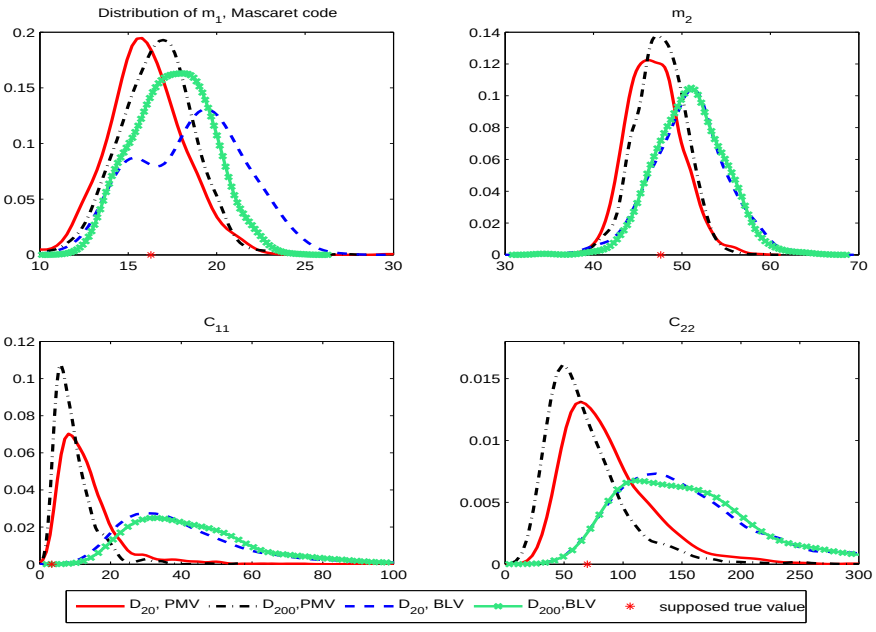


Figure VI.17: Posterior distributions of θ with two informative priors, based on $n = 10, 50$ observations, and two *maximin*-LHDs D_{20} and D_{200}

VII

Conclusion and perspective

$\mathcal{A}$. Summary and main contributions

The work presented in this thesis was aimed at providing a Bayesian solution to inverse problems under the Gaussian assumption on the variable of interest, with an numerical simulator expensive to compute. One call to the simulator may require several hours of computation. This goal has been achieved in the light of the industrial examples presented in Chapter VI.

To reach the goal with a small sample setting and available expert knowledge, a Bayesian framework has been chosen and it has been decided to apply the kriging technique to approximate the time-consuming function. Chapter I provides a review of the main tools required in this work. Besides, several probability estimation techniques for frequentist inference have been discussed along with their pros and cons.

The construction of the Bayesian model has been discussed in Chapter II. It focused on the elicitation of prior distributions, which can be considered as the basis of the Bayesian analysis. Moreover, by introducing the meta-modeling methodology, another source of uncertainty was added and the original uncertainty model was modified to adapt to the Bayesian framework combined with the kriging prediction.

Chapter III is devoted to the description of the main algorithm: the Metropolis-Hastings-within-Gibbs algorithm. It provides an exhaustive presentation by distinguishing two versions depending on the availability of the time-consuming function. In addition, the main properties of the so-built Markov chains and the convergence issues have been studied.

Chapter IV is a key chapter on the evaluation of the results following such a Gibbs sampling. It resumes the *estimation error*, *emulator error*, *algorithmic error* and *prior error*, which can affect the posterior results. In this chapter, a Bayesian criterion was proposed to control the impact of some of those errors by assessing the relevance between the numerical design and the prior distribution. The need to improve the numerical design in order to reduce the emulator error is thus highlighted.

For this reason, Chapter V is devoted to proposing an adaptive methodology to construct the numerical design in a sequential way. A purpose-oriented DOE has been defined which takes

into account the observations $\mathbf{y}$. Two Bayesian criteria were provided, whose performances were compared in view of two examples.

The final Chapter VI has described a real industrial application. Two codes from EDF used by engineers for hydraulic applications were exercised to yield the interest of applying the methods developed in this thesis work.

$\mathcal{B}$. Further investigations

We hope that the contributions of this thesis are helpful to solve inverse problems in uncertainty analysis. Nevertheless, there are still some points and some directions which need to be developed.

1. Bayesian model In the main model (I.4), it has been assumed in our case study that the input d_i related to the experimental conditions is observed and the measurement error U_i follows a centered normal distribution with a known variance. In order to propose a more general model, we introduce another variable input Z_i with known distributions. Namely the model (I.4) could become:

$$Y_i \quad = \quad H(X_i, Z_i, d_i) \; + \; U_i. \tag{VII.1}$$

Moreover, there is no reason except for the purpose of simplicity to assume that the variability of the error term U_i is known. In future researches, this assumption could be removed to make the model more reliable, keeping in mind that an unknown variance R might cause nonidentifiability problems.

2. MCMC algorithms Second, in the Metropolis-Hastings-within-Gibbs algorithm, it would be interesting to develop an Adaptive Metropolis-Hastings (AMH) algorithm (see e.g. Pasanisi et al., 2012, (76), Roberts and Rosenthal, 2007, 2009, (92), (93)) to accelerate the convergence of the simulated Markov chain. More precisely, as mentioned in Chapter III that the choice of the proposal distribution J plays a critical role in the performance of the MCMC algorithms, choosing a proper J accounting for the evaluation of the current full conditional posterior distribution to approach the stationary distribution is a promising research subject.

Besides, due to the observation cost or other limitations, the sample $\mathbf{y}$ may be partially observed, which indicates a missing data framework. The MCMC algorithms, especially the Gibbs sampler, are used to deal with such missing data schemes.

3. Adaptive kriging method A first study could be to test the robustness of the proposed adaptive methodology in Bayesian approach, by trying a form of the initial DOE different from the *maximin*-LHD to compare the posterior results.

Moreover, if we have several available calculators, it would be effective to make several evaluations of the time-consuming function H at one time. For example, we use the E-CD criterion to find add several optimal points by controlling the distance between them.

4. Multi-fidelity meta-modeling In the treatment of the hydraulic applications, we have considered two industrial codes: MASCARET and TELEMAC-2D, where the evaluation time of the first code is negligible compared to the second code. This highlights the interest of applying the multi-fidelity meta-modeling technique (see Kennedy and O'Hagan, 2000, (52)). We introduce the main idea here. Consider two levels of code $Z_1(\cdot)$ and $Z_2(\cdot)$, where $Z_2(\cdot)$ is the higher level code. We assume that for $\forall x$, given the point $Z_1(x)$, we can learn no more about $Z_2(x)$ from any other run $Z_1(x')$ for $x' \neq x$. Under a prior assumption that each output of the code can be considered as the realization of a Gaussian Process (GP), we describe the multi-level model as follows:

$$\begin{cases} Z_2(x) = \rho_1(x) Z_1(x) + \delta_2(x) \\ Z_1(x) \perp\!\!\!\perp \delta_2(x), \end{cases} \tag{VII.2}$$

where we define that

$$\delta_2(x) \sim \mathcal{N}\left(f_2^T(x)\beta_2, \sigma_2^2 K_{r_2}\right), \tag{VII.3}$$

and

$$Z_1(x) \sim \mathcal{N}\left(f_1^T(x)\beta_1, \sigma_1^2 K_{r_1}\right). \tag{VII.4}$$

We let $\beta = (\beta_1, \beta_2)$, $\phi = (\sigma^1, \sigma^2, r_1, r_2, \rho)$ and $Z = (Z_1, Z_2)$, the process $[Z_2(x) \mid Z, \beta, \phi]$ after integrating over β is shown to be a Gaussian process

$$[Z_2(x) \mid Z, \phi] \sim \mathcal{N}\left(m_{Z_2}(x), S^2_{Z_2}(x)\right), \tag{VII.5}$$

where the posterior mean function $m_{Z_2}(x)$ is a cheap approximation for the expensive code $Z_2(x)$.

5. Polynomial chaos Introduced by Ghanem et Spanos (1991, (35)), the polynomial chaos focuses on approximating the random variables by Gaussian polynomials or Wiener-Hermite expansion, to characterize the uncertainty in dynamical system. We use this idea to develop a spectral representation of X_i.

The Gaussian assumption on the non observed variable X_i can be replaced by a weaken one, say any second-order distributions, thanks to the Wiener-Hermite parameterization. In fact, apart from the Gaussian distribution, any second-order distribution can be approximated by Hermite polynomials defined in standard Gaussian probability space. The so-called Wiener-Hermite representation (see e.g. Wiener, 1938, (119)) is as follows:

$$X_i \simeq X_i^{P,M} = \sum_{j=0}^{P-1} z_{i,j} P_{i,j}(\xi_i), \tag{VII.6}$$

where $P_{i,j}$s denote the multivariate Hermite polynomials of degree lower than p (the degree of the expansion), at a sequence of independent standard normal random variables $\xi_i = (\xi_{i,1}, \ldots, \xi_{i,M})$ and $z_{i,j}$s $(j = 0, \ldots, P-1)$ denote P integer coefficients to be estimated for each X_i, with

$$P = \frac{(M+p)!}{M!\,p!}. \tag{VII.7}$$

The same representation can be done for the measurement error U_i. The advantage is that we can remove the Gaussian hypothesis on X_i and U_i to make a more general modeling while a Gaussian space can still be used.

Based on the Wiener-Hermite representation, an interesting parameterization (see Perrin, 2008, (80), Rachdi, 2011,(85)) transforms the model into the standard Gaussian probability space

$$Y_i \quad = \quad h(\xi_i; \theta_i), i \in \{1, \ldots, n\}, \tag{VII.8}$$

where ξ_i, as previously defined, denotes a $M-$dimensional independent standard normal random variable, θ_i denotes the related coefficients in the Wiener-Hermite expansion and h corresponds to a variation of the original function H adapted to this new parameterization. Calibrating the posterior distribution of the variable X_i becomes to estimate the coefficients θ_i. It should be noted that a larger P requires more prior eliciting work.

Bibliographie

[1] Ababou, R., Bagtzoglou, A.C. and Wood, E.F. (1994). On the condition number of covariance matrices in kriging, estimation, and simulation of random fields, *Mathematical Geology*, **26**, 99-133.

[2] Allen, D. (1971). The prediction sum of squares as a criterion for selecting prediction variables, *Technical Report*, **23**, Dept. of Statistics, University of Kentucky.

[3] Barbillon P. (2010). Méthodes d'interpolation noyaux pour l'approximation de fonctions type boîte noire coûteuses, *Ph.D thesis*, Université Paris-Sud 11, 106-121.

[4] Barbillon, P., Celeux, G., Grimaud, A., Lefebvre, Y. and De Rocquigny, E. (2011), Non linear methods for inverse statistical problems, *Computational Statistics & Data Analysis*, **55**, 132-142.

[5] Bastos, L.S. and O'Hagan, A. (2009), Diagnostics for Gaussian Process Emulators, *Technical report*, University of Sheffield.

[6] Besnard, A., Dranguet, M. (2008). Intercomparaison de modèles hydrauliques 1D et 2D sur la Garonne, *Note EDF LNHE*, H-P73-2007-02253-FR.

[7] Besnard A., Goutal N. (2008) Comparison between 1D and 2D models for hydraulic modeling on a flood plain : Case of Garonne river, *Proc. Int. Conf. River Flow 2008.*

[8] Bernardara, P., De Rocquigny, E., Goutal, N. and Passoni, G. (2008). Flood risk assessment: Model calibration under uncertainty, *Journal of Hydraulic Engineering.*

[9] Bousquet, N. (2006). Subjective Bayesian statistics: agreement between prior and data, *Technical report*, INRIA.

[10] Bousquet, N. (2008), Diagnostics of prior-data agreement in applied Bayesian analysis, *J. Appl. Statist.*, **35**, 1011-1029.

[11] Brooks, S.P. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations, *Journal of Computational and Graphical Statistics*, **7**, 434-455.

[12] Browne, W.J. and Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models, *Computational Statistics*, **15** (3), 391-420.

[13] Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, *Computational Statistics Quarterly*, **2**, 73-82.

[14] Celeux, G. and Diebolt, J. (1987). A probabilistic teacher algorithm for iterative maximum likelihood estimation, *Classification and related methods of Data Analysis*, 617-623.

[15] Celeux, G., Grimaud, A., Lefebvre, Y., and De Rocquigny, E. (2010). Identifying intrinsic variability in multivariate systems through linearised inverse methods, *Inverse Problems in Engineering*, **18**, 401-415.

[16] Chen, M. and Schmeiser, B. (1998). Towards black-box sampling, *J. Comput. Graph. Statist.*, **7**, 1-22.

[17] Clarke, B.S. (1996). Implications of reference priors for prior information and for sample size, *J. Amer. Statis. Assoc.*, **91**, 173-184.

[18] Cover, T.M. and Thomas, J.A. (2006). Elements of Information Theory (2nd Edition), WILEY.

[19] Cowles, M.K. and Carlin, B.P. (1996). Markov Chain Monte Carlo Convergence Diagnostic: A Comparative Review, *Journal of the American Statistical Association*, **91** (434), 883-904.

[20] De Crecy, A. (1996). Determination of the uncertainties of the constitutive relationships in the Cathare 2 Code, *Proceedings of the 1996 4th ASME/JSME International Conference on Nuclear Engineering.*

[21] De Crécy, A. (2001). Determination of the Uncertainties of the Constitutive Relationships of the CATHARE 2 Code, *Communication to M & C 2001*, Salt Lake City, Utah, USA.

[22] De Rocquigny, E. and Cambier, S. (2009). Inverse probabilistic modelling of the sources of uncertainty: a non-parametric simulated-likelihood method with application to an industrial turbine vibration assessment, *Inverse Problems in Science and Engineering*, **17**(7).

[23] De Rocquigny, E., Devictor, N. and Tarantola, S., editors (2008). Uncertainty in industrial practice - A guide to quantitative uncertainty management, *Wiley*, ISBN 978-0-470-99447-4.

[24] Degoutte, G. (2006). Aide-mémoire d'hydraulique à surface libre. In *Diagnostic, aménagement et gestion des rivières*, Tec & Doc.

[25] Dempster, E. J., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm, *Annals of the Royal Statistical Society*, Series B, **39**, 1-38.

[26] Dubourg, V. (2011). Méta-modèles adaptatifs pour l'analyse de fiabilité et l'optimisation sous contrainte fiabiliste, *Ph.D thesis*, Université Blaise Pascal - Clermont II, 19-33.

[27] Esclaffer, T. (2003). Étude théorique de la formation des débits de crues à l'échelle du versant, *DEA report in Sciences and Techniques of the Environment*, University Paris XII. 15-16.

[28] Fang, K.-T., Li, R. and Sudjianto, A. (2006). Design and Modeling for Computer Experiments, *Computer Science and Data Analysis*, Chapman & Hall/CRC.

[29] Fu, S., Celeux, G., Bousquet, N. and Couplet, M. (2012) Bayesian inference for inverse problems occuring in uncertainty analysis, *Technical report*, INRIA.

[30] Fuh, C.D. (1993). Statistical inquiry for Markov chains by bootstrap method, *Statistica Sinica*, **3**, 53-66.

[31] Garthwaite, P.H., Kadane, J.B. and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions, *Journal of the American Statistical Association*, **100**, 680-700.

[32] Gelfand, A.E. and Sahu, S.K. (1994). On Markov chain Monte Carlo acceleration, *Journal of the American Statistical Association*, **3** (3), 261-276.

[33] Gelman, A. and Rubin, D. (1992). Inference from Iterative Simulation using Multiple Sequences, *Statistical Science*, **7**, 457-511.

[34] Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**(6): 721-741. doi:10.1109/TPAMI.1984.4767596

[35] Ghanem, R. et Spanos, P. (1991). Stochastic finite elements - A spectral approach, *Springer Verlag.*

[36] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). Markov Chain Monte Carlo In Practice, 45-65. Chapman & Hall.

[37] Girard, P., Parent, E. (2004). The deductive phase of statistical analysis via predictive simulations: test, validation and control of a linear model with autocorrelated errors representing a food process, *Journal of Statistical Planning and Inference*, **124**, 99-120.

[38] Goutal, N. (2005). Calage automatique du coefficient de Strickler en régime permanent sur un seul bief, *Technical report EDF LNHE*, HP-75/05/021/A.

[39] Hastie, T., Tibshirani, R. and Friedman, J. (2001). The elements of statistical learning, *Springer Series in Statistics*, Springer, New York.

[40] Hermite, C.(1864). Sur un nouveau développment en série de fonctions, *C. R. Acad. Sci. Paris*, **58** (1864), 93-100; Oeuvres II, 293-303.

[41] Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55-67.

[42] Horn, R. A. and Johnson, C. R. (1991), *Topics in Matrix Analysis*, Cambridge University Press.

[43] Horrit, M.S. (2000). Development of physically based meshes for two-dimensional models of meandering channel flows, *International Journal of Numerical Methods in Engineering*, **41**, 2109-2137.

[44] Huang, D., Allen, T.T., Notz, W.I. and Zeng, N. (2006). Global optimization of stochastic black-box systems via sequential kriging meta-models, *Journal of Global Optimization*, **34** (3), 441-466.

[45] Idier, J. (2001). Approche bayésienne pour les problèmes inverses, *Hermès Science Publications*, Paris, 25-40.

[46] Johnson, M. E., Moore, L. M. et Ylvisaker, D. (1990), Minimax and maximin distance designs, *Journal of Statistical Planning and Inference*, **26**, 131-148.

[47] Jones, D. R., Schonlau, M. et Welch, W. J. (1998). Effcient global optimization of expensive black-box functions, *Journal of Global Optimization*, **13**(4), 455-492.

[48] Joseph V.R., and Hung Y. (2008), Orthogonal-Maximin Latin Hypercube Designs, *Statistica Sinica*, **18**, 171-186.

[49] Kadane, J.B. and Wolfson, J.A. (1998). Experiences in elicitation, *The Statistician*, **47**, 3-19.

[50] Kárný, M., Nedoma, P., Khailova, N., Pavelková, L. (2003). Prior information in structure estimation, *IEEE Proceedings in Control Theory and Applications*, **150**, 643-653.

[51] Kass, R.E. and Wasserman, L. (1996). The selection of prior distribution by formal rules, *J. Amer. Statist. Assoc.*, **91**, 1343-1370.

[52] Kennedy, M.C. and O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available, , *Biometrika*, **87**, 1-13.

[53] Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983). Optimization by Simulated Annealing, *Science*, New Series, **220** (4598), 671-680.

[54] Koehler, J.R. and Owen, A.B. (1996). Computer experiments. In: Ghosh, S., Rao, C.R. (Eds.), , *Handbook of Statistics*, Elsevier, Amsterdam, 261-308.

[55] Kuhn, E. (2003). Estimation par maximum de vraisemblance dans des problèmes inverses non linéaires, *Ph.D thesis*, Université Paris-Sud 11, 13-28.

[56] Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with a MCMC procedure, *ESAIM P & S*, **8**, 633-648.

[57] Le Gratiet, L. and Garnier, J. (2012). Bayesian analysis of hierarchical codes with different levels of accuracy. , *Mascot Meeting*.

[58] Leroy, O. (2010). Estimation d'incertitudes pour la propagation acoustique en milieu extérieur, *Ph.D thesis*, Université du Maine.

[59] Li, R. and Sudjianto, A. (2005). Analysis of computer experiments using penalized likelihood in gaussian kriging models, *Technometrics*, **47**, 111-120.

[60] Liu, C. and Rubin, D.B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence, *Biometrika*, **81**, 633-648.

[61] Liu, D. (2009). Uncertainty Quantification with Shallow Water Equations, *Ph.D thesis*, Carl-Friedrich-Gauss Faculty, University of Braunschweig.

[62] Liu, J., Wong, W. and Kong, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans, *J. Royal Statist. Soc.*, Series B, **57**, 157-169.

[63] Lophaven, S., Nielsen, H. and Sϕndergaard, J. (2002). DACE, *A Matlab Kriging Toolbox*, Techinical University of Denmark.

[64] Marrel, A. (2008). Mise en oeuvre et utilisation du métamodèle processus gaussien pour l'analyse de sensibilité de modèles numériques, *Ph.D thesis*, L'Institut national des sciences appliquées de Toulouse, 28-35.

[65] Matheron, G. (1971). The theory of regionalised variables and its applications, *Ph.D thesis*, École Nationale Supérieure des Mines de Paris.

[66] McKay, M.D., Beckman, R.J., and Conover, W.J. (1979), A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, *Technometrics*, **21**, 239-245.

[67] McLachlan, G.J. and Krishnan, T. (2008). The EM algorithm and extensions, *Wiley Series in Probability and Statistics*, 77-85.

[68] Memarsadeghi, N., Raykar, V.C., Duraiswami, R. and Mount, D.M. (2008). Efficient Kriging via Fast Matrix-Vector Products. University of Maryland, College Park.

[69] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics*, **21**(6), 1087-1092.

[70] Meyn, S.P. and Tweedie, R.L. (1993). Markov Chains and Stochastic Stability, *London: Springer-Verlag.*

[71] Mitchell, T., Morris, M. and Ylvisaker, D. (1990), Existence of smoothed stationary processes on an interval, *Stochastic Processes and Their Applications*, **35**, 109-119.

[72] Morita, S., Thall, P.F. and Mueller, P. (2007). Determining the effective sample size of a parametric prior, *UT MD Anderson Cancer Center Department of Biostatistics*, Working Paper Series. Working Paper 36.

[73] Muller, P. (1991). A generic approach to posterior integration and Gibbs sampling, *Technical report*, Purdue Univ., West Lafayette, Indiana.

[74] Munoz-Zuniga, M., Garnier, J., Remy, E. and de Rocquigny, E. (2011). Analysis of adaptive directional stratification for the controlled estimation of rare event probabilities, *Statistics & Computing*, **22**, 809-821.

[75] Parent, E., Lebdi, F., Hurand, P. (1991). Stochastic modeling of a water resource system: analytical techniques versus synthetic approaches. In: Ganoulis, J. (Ed.), Water Resources Engineering Risk Assessment, *Springer-Verlag*, Heidelberg, 415-434.

[76] Pasanisi, A., Fu, S. and Bousquet, N. (2012). Estimating discrete Markov models from various incomplete data schemes, *Computational Statistics and Data Analysis*, **56**(9), 2609-2625.

[77] Pasanisi, A., Keller, M., Parent, E. (2011). Réflexions sur l'analyse d'incertitudes dans un contexte industriel : information disponible et enjeux décisionnels, *Journal de la SFdS*, **152**, 60-77.

[78] Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika*, **58**, 545-554.

[79] Paulino, C.D.M. and Pereira, C.A.B. (1994). On identifiability of parametric statistical models, *Statistical Methods & Applications*, **3**(1), 125-151.

[80] Perrin, F. (2008). Prise en compte des données expérimentales dans les modèles probabilistes pour la prévision de la durée de vie des structures, *Ph.D thesis*, Université Blaise Pascal (Clermont II).

[81] Petelet M., Iooss. B., Asserin O., Marrel, A.(2010), Latin hypercube sampling with inequality constraints, *Advances in Statistical Analysis*, **94**, 325-339.

[82] Picheny, V. (2009). Improving accuracy and compensating for uncertainty in surrogate modeling, *Ph.D thesis*, École Nationale Supérieure des Mines de Saint-Étienne.

[83] Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R.T. and Kim, N-H. (2010). Adaptive designs of experiments for accurate approximation of a target region, *Journal of Mechanical Design*, **132**(7), 071008.

[84] Puolamaki, K., Kaski, S. (2009). Bayesian solutions to the label switching problem. In: Adams, N., Robardet, C., Siebes, A., Boulicaut, J.F. (Eds.), Advances in Intelligent Data Analysis VIII. Proceedings of the 8th International Symposium on Intelligent Data Analysis, , *IDA* , Springer, Berlin, 381-392.

[85] Rachdi, N. (2011). Apprentissage statistique et computer experiments, *Ph.D thesis*, Université Toulouse III, 126-148.

[86] Rasmussen, C.E. and Williams, C,K,I. (2006). Gaussian Processes for Machine Learning, *The MIT Press*, Cambridge, MA.

[87] Raykar, V.C. (2007). Scalable machine learning for massive datasets: Fast summation algorithms, University of Maryland, College Park, MD, 20742.

[88] Richard, F. J.P. and Samson, A. (2007). Metropolis-Hasting techniques for finite element-based registration, *CVPR Workshop*, Minneapolis, Minnesota, USA.

[89] Robert, C.P. (2001). The Bayesian Choice: From Decision-Theoretic Motivations to Computational Implementation, *Springer*, New York.

[90] Robert, C.P. and Casella, G. (2004). Monte Carlo Statistical Methods. Second edition, *Springer*, 267-286, 371-407.

[91] Roberts, G.O. and Rosenthal, J.S. (2006). Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains, *The Annals of Applied Probability*, **16**(4), 2123-2139.

[92] Roberts, G.O., Rosenthal, J.S. (2007). Coupling and ergodicity of adaptive MCMC, *Journal of Applied Probabilities*, 44, 458-475.

[93] Roberts, G.O., Rosenthal, J.S.(2009). Examples of adaptive MCMC, *Journal of Computational and Graphical Statistics*, 18, 349-367.

[94] Roberts, G.O. and Smith, A.F.M. (1994). Simple conditions for the convergence of the Gibbs sampler and Hastings-Metropolis algorithms, *Stoch. Proc. Appl.*, **49**, 207-216.

[95] Roustant, O., Ginsbourger, D. and Deville, Y. (2010). DiceKriging: Kriging methods for computer experiments, *R package version 1.1.*

[96] Rubino, G., Tuffin, B. (2009). Rare event simulation using Monte Carlo methods, *Wiley.*

[97] Sacks, J., Schiller, S.B., Mitchell, T.J. and Wynn, H.P. (1989a). Design and analysis of computer experiments (with discussion), *Statistica Sinica*, **4**, 409-435.

[98] Sacks, J., Schiller, S.B. and Welch, W.J. (1989b). Designs for computer experiments, *Technometrics*, **31**(1), 41-47.

[99] Santner, T.J., Williams, B. and Notz, W. (2003). The Design and Analysis of Computer Experiments, *Spring-Verlag.*

[100] Sekhon, J. and Mebane, W. (2011). Genetic optimization using derivatives: the rgenoud package, *R. J. Stat. Software*, **42**(11), 1-26.

[101] Sellin, R.H.J., Keast, J., Van Beeston, D. (1997). Seasonal Variation in River Channel Hydraulic Roughness, *27th IAHR Congress*, B2, 1390-1396.

[102] Stein, M.L. (1999). Interpolation of Spatial Data: Some Theory for Kriging, *Springer*, New York.

[103] Stein, M.L., Chi, Z. and Welty, L.J. (2004). Approximating likelihoods for large spatial data sets, *Journal of the Royal Statistical Society*, **66** (2), 275-296.

[104] Stone, C.J., Hansen, M.H., Kooperberg, C. and Truong, Y.K. (1997). Polynomial splines and their tensor products in extended linear modeling, *Annals of Statistics*, **25**, 1371-1470.

[105] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *J. Royal Stat. Soc.*, Series B, **36**, 111-147.

[106] Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation, *J. American Statist. Assoc.*, **82**, 528-550.

[107] Thisted, R.A. (1988), *Elements of Statistical Computing*, Chapman & Hall.

[108] Tibshirani, R. (1994). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, Series B, **58**, 267-288.

[109] Tierney, L. (1995), Introduction to general state-space Markov chain theory, *Markov Chain Monte Carlo in Practice*, 59-74, Chapman & Hall.

[110] U.S. Army Corps of Engineers (1996). Risk-analysis for flood damage reduction studies, *Technical report*, No. EM 1110-2-1619.

[111] Vanderpoorten, A. and Palm, R. (2001). Compared regression methods for inferring ammonium nitrogen concentrations in rivers from aquatic bryophyte assemblages, *Hydrobiologia*, **452**, 181-190.

[112] Vazquez, E. (2005) Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications, *Ph.D thesis*, Université Paris-Sud Orsay.

[113] Viollet, P.-L., Chabard, J.-P., Esposito, P. and Laurence, D. (1998). *Mécanique des Fluides Appliquée*, Presses of l'École Nationale des Ponts et Chaussées.

[114] Walesh, S.G. (1989). *Urban Water Surface Management*, John Wiley and Sons.

[115] Wang, Q., Kulkarni, S.R. and Verdú, S. (2006). A Nearest-Neighbor approach to estimating divergence between continuous random vectors, *IEEE International symposium on information theory*, 242-246.

[116] Watzenig, D. (2007). Bayesian inference for inverse problems-statistical inversion, Springer-Verlag.

[117] Wei, G.C.G. and Tanner, M.A. (1990a). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, *Journal of the American Statistical Association*, **85**, 699-704.

[118] Wei, G.C.G. and Tanner, M.A. (1990b). Posterior computations for censored regression data, *Journal of the American Statistical Association*, **85**, 829-839.

[119] Wiener, N. (1938). The homogeneous chaos, *American Journal of Mathematics*, **60** (4), 897-936.

[120] Wu, C.F.J. (1983). On the convergence properties of the EM algorithm, *Annals of Statistics*, **11**, 95-103.

[121] Yang, R. and Berger, J.O. (1998), A Catalog of Non-informative Priors, *ISDS Discussion Paper*, 97-42.

[122] Zellner, A. (1986). On assessing Prior Distributions and Bayesian Regression analysis with g-prior distribution regression using Bayesian variable selection, *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 233-243, North-Holland, Elsevier.

Abstract

This thesis provides a probabilistic solution to inverse problems through Bayesian techniques. The inverse problem considered here is to estimate the distribution of a non-observed random variable X from some noisy observed data Y following a time-consuming physical model H. In general, such inverse problems are encountered in treatment of uncertainties. Bayesian inference is favored as it accounts for prior expert knowledge on X in a small sample size setting. A Metropolis-Hastings-within-Gibbs algorithm is proposed to compute the posterior distribution of the parameters of X through a data augmentation process. Since it requires a high number of calls to the expensive function H, the model is replaced by a kriging meta-model $\widehat{H}$. This approach involves several errors of different natures and we focus on measuring and reducing the possible impact of those errors. A $\widetilde{\text{DAC}}$ criterion has been proposed to assess the relevance of the numerical design of experiments and the prior assumption, taking into account the observed data. Another contribution is the construction of adaptive designs of experiments adapted to our particular purpose in Bayesian framework. The main methodology presented in this thesis has been applied to a real hydraulic engineering case-study.

Keywords: inverse problem, Bayesian inference, expert opinion, Markov model, hybrid MCMC algorithm, Kriging, assessment error, prior-data conflict, adaptive design of experiments.

Résumé

Ce travail de recherche propose une solution aux problèmes inverses probabilistes avec des outils de la statistique bayésienne. Le problème inverse considéré est d'estimer la distribution d'une variable aléatoire non observée X à partir d'observations bruitées Y suivant un mod 'ele physique coûteux H. En général, de tels problèmes inverses sont rencontrés dans le traitement des incertitudes. Le cadre bayésien nous permet de prendre en compte les connaissances préalables d'experts surtout avec peu de données disponibles. Un algorithme de Metropolis-Hastings-within-Gibbs est proposé pour approcher la distribution a posteriori des paramètres de X avec un processus d'augmentation des données. A cause d'un nombre élevé d'appels, la fonction coûteuse H est remplacée par un émulateur de krigeage (méta-modèle) $\widehat{H}$. Cette approche implique plusieurs erreurs de nature différente et, dans ce travail, nous nous attachons à estimer et réduire l'impact de ces erreurs. Le critère $\widetilde{\text{DAC}}$ a été proposé pour évaluer la pertinence du plan d'expérience (design) et le choix de la loi a priori, en tenant compte des observations. Une autre contribution est la construction du design adaptatif adapté à notre objectif particulier dans le cadre bayésien. La principale méthodologie présentée dans ce travail a été appliquée à un cas d'étude d'ingénierie hydraulique.

Mots-clés: problème inverse, inférence bayésienne, expert industriel, modèle de Markov, algorithme MCMC hybride, krigeage, erreur d'évaluation, conflit entre données et *a priori*, plans d'expérience adaptatifs.