Uncertainty quantification and Bayesian calibration in computational geosciences

Olivier Le Maître[1]

[1]Centre de Mathématiques Appliquées, CNRS
Ecole Polytechnique, Palaiseau, France
https://http://olemaitre.perso.math.cnrs.fr/
olivier.le-maitre@polytechnique.edu
https://team.inria.fr/platon/

Waves & geosciences: Infrasound and beyond – LETMA, March 29th, 2022

Table of contents

Model Predictions

Predictions based on numerical model involve:

- Errors due to simplified **physical models**
- Errors due to **numerical approximations**
- Errors due to **unknown model parameters**
- Errors due to **misspecification of object of the prediction**

Although all these sources of prediction errors are separated, **their global treatment** is needed to

- Fairly assess the prediction quality
- Enable risk informed decision-making
- Ensure that no resources are inefficiently allocated
- Engage in prediction error reduction

Experimental data & observations are crucial to produce accurate model predictions

but are also compromised and subjected to multiple sources of error

Model Predictions & Experimental Data

Observations and experimental measurements are needed to

- Determine the model parameters (**e.g. calibration**)
- Validate the model and its predictions (**e.g. comparison**)

Since **the knowledge of the model parameters is never perfect** we adopt a probabilistic view:

- The predictions are random (uncertain) with a probability distribution induced by the probability distribution of the model
- One can assess the impact of different model parameters on the uncertain prediction (Forward UQ problem) and perform sensitivity analysis (ANOVA, HSIC,. . .)
- From an a priori parameter distribution, experimental observations can be incorporated to update the knowledge of the parameters

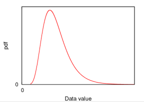
So-called Bayesian model calibration

Note: here, **we entirely focus on parametric uncertainty**. Other Bayesian techniques can compare the predictive capabilities of **different model structures**

Forward Uncertainty Propagation

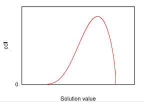
Propagation of data uncertainty

Data density



$$\mathcal{M}(S, D) = 0$$

Solution density



Many alternative methods

- Sampling based methods: Monte-Carlo, QMC, LHS,. . .
- Surrogate based methods: Regression and projection methods, Polynomial Chaos, Stochastic Multi-wavelets, Sparse-grid interpolation, Gaussian process regression,. . .
- Stochastic Galerkin methods

In geosciences, we have to deal with:

- Models with high to extremely high solution cost
- Possibly high dimensional parameters and predictions (fields, time series,. . .)
- Generally sparse and costly observations (although we are promised cheap data)

Calls for adapted techniques, mitigating numerical complexity with accuracy

Example UQ in large-scale seismic simulation

[Sochala, de Martin & OLM. Int. J. Unc. Quant., 2020]

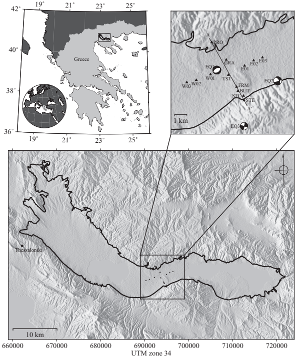


FIG. 1: Location of the domain of study in Greece close to the city of Thessaloniki. The free surface of the domain is depicted together with the internal boundary between the sedimentary basin and the bedrock (solid line).

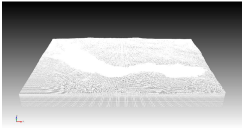


FIG. 4: Mesh $\mathcal{M}$ of the computational domain ($64 \times 42 \times 5$ km) composed of 7,053,889 hexahedron elements. The two levels of mesh refinement are noticeable at the top part of the domain (tripling technique) and inside the sedimentary basin.

Example UQ in large-scale seismic simulation

[Sochala, de Martin & OLM. Int. J. Unc. Quant., 2020]

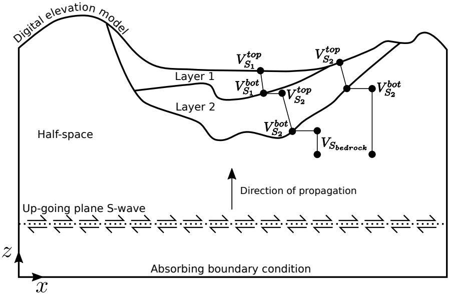


FIG. 6: Schematic representation of the shear wave velocity structure: in the sedimentary basis the profile is piecewise linear and defined by values at the top and bottom of the layers 1 and 2; in the bedrock the value is constant.

TABLE 1: The nominal values and ranges of the seven uncertain inputs retained in the uncertainty analysis

	$v_{s,1}^{top}$	$v_{s,1}^{bot}$	$v_{s,2}^{top}$	$v_{s,2}^{bot}$	$v_{s,bed}$	q_s	δd
Nominal	130 (m/s)	475 (m/s)	581 (m/s)	800 (m/s)	2,100 (m/s)	12.5 (–)	0 (m)
Range	[104, 156]	[428, 523]	[523, 639]	[640, 960]	[1920, 2280]	[5, 20]	[−10, 10]

Example UQ in large-scale seismic simulation

[Sochala, de Martin & OLM. Int. J. Unc. Quant., 2020]

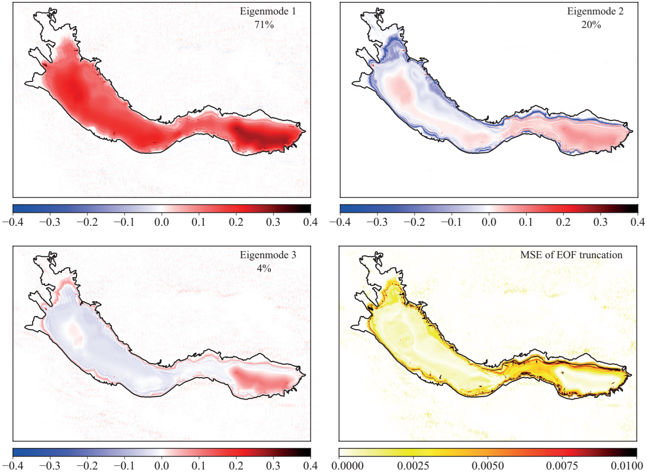


FIG. 10: Dominant rescaled eigenmodes $\mathrm{u}_k\sqrt{\lambda_k}$ of the EOF decomposition, and (componentwise) empirical $\mathrm{MSE}(\mathrm{u}, \mathrm{u}^r)$ for $r = 3$

Example UQ in large-scale seismic simulation

[Sochala, de Martin & OLM. Int. J. Unc. Quant., 2020]

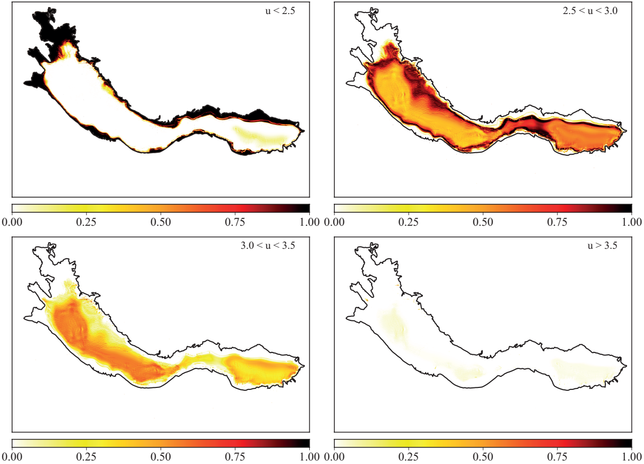


FIG. 16: PGM interval probability maps for different intervals obtained with a LHS of 10^4 realizations.

Table of contents

Bayesian inference

Parametric uncertainty

- incomplete knowledge of some model parameters: $\boldsymbol{q} \sim p(\boldsymbol{q})$
- uncertain model prediction $M(\boldsymbol{q})$
- uncertainty reduction strategies

Bayes formula

We want to update / infer a finite set of parameters $\boldsymbol{q} \in \mathbb{R}^q$, using

- a set $\mathcal{O} \doteq \{y_i \in \mathbb{R}, i = 1, \ldots, M\}$ of observations,
- the model prediction of the observations: $\mathbf{U}(\boldsymbol{q}) \in \mathbb{R}^M$

Bayesian rule to update our knowledge on $\boldsymbol{q}$:

$$p_{\mathrm{post}}(\boldsymbol{q}|\mathcal{O}) \propto L(\mathcal{O}|\boldsymbol{q})p(\boldsymbol{q}),$$

with

- $L(\mathcal{O}|\boldsymbol{q})$ is the likelihood of the measurements,
- $p(\boldsymbol{q})$ is the parameters' prior,
- $p_{\mathrm{post}}(\boldsymbol{q}|\mathcal{O})$ is the posterior.

Bayesian inference

Likelihood function (Gaussian example)

Model for the measurements error (noise):

$$Y_i = U_i(\boldsymbol{q}) + \epsilon_i, \quad \epsilon_i = N(0, \sigma_i^2),$$

The likelihood becomes:

$$L(\mathcal{O}|\boldsymbol{q}) \doteq \prod_{i=1}^{M} \exp\left[-\frac{|y_i - U_i(\boldsymbol{q})|^2}{2\sigma_i^2}\right].$$

Posterior sampled, for instance using Markov Chain Monte Carlo (MCMC)
Note: in reality needs hyper-parameters (*i.e.* noise variance)

Issues:

- **Rely heavily on multiple evaluations** of the model $\boldsymbol{q} \mapsto \boldsymbol{U}(\boldsymbol{q}) \doteq (U_1 \cdots U_M)(\boldsymbol{q})$: use of surrogate models
- Assumes the **measurements to be informative**: more is not always better, in particular in the absence of complete information regarding protocols
- Calls for the **selection of robust and informative observations**
- Model error?

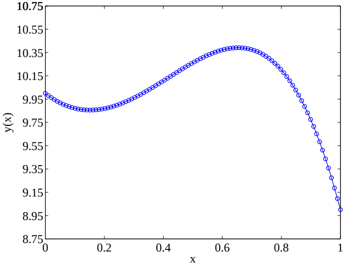
Example

- Suppose that we have the following polynomial model:

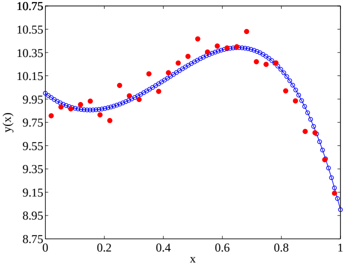
"True" polynomial

$$u(x) = 10 - 2x + 7.5x^2 - 3.3x^3 - 3.2x^4$$

observed at at N coordinates $\{x_i\}_{i=1}^{N} \in (0,1)$



- We perturb the observations with a Gaussian noise with mean zero and variance 0.01, i.e. $\mathcal{N}(0, 0.01)$.
- This yields a set of noisy observations, $(\{x_i, y_i\}_{i=1}^{N})$.
- For this example we have $N = 30$. (We will discuss the effect of the number of observations)



Example

- Objective: given the data $\mathcal{O} = \{y_i\}_{i=1}^N$, can we recover the original polynomial?
- We need to define a model (i.e. the hypothesis) to describe the data.
- Our model is a polynomial of certain order p:

$$M(x|\boldsymbol{q}) = \sum_{k=0}^{p} q_k x^k \tag{1}$$

- It follows that our set of parameters is:

$$\boldsymbol{q} = \{q_0, q_1, q_2, \ldots, q_p\} \tag{2}$$

Bayes' theorem

$$p_{\text{post}}(\{q_k\}_{k=0}^p | \{y_i\}_{i=1}^N) \propto L(\{y_i\}_{i=1}^N | \{q_k\}_{k=0}^p)\ p(\{q_k\}_{k=0}^p)$$

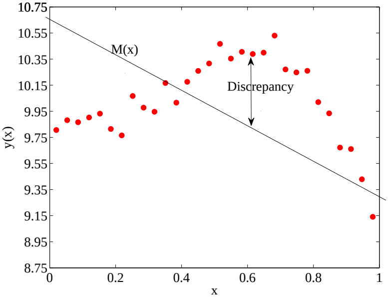
- We now need to define the likelihood and priors.

Likelihood

- To formulate the likelihood we assume the following relationship:

$$y_i = U_i(\boldsymbol{q}) + \epsilon_i, \quad U_i(\boldsymbol{q}) = M(x_i|\boldsymbol{q})$$

where ϵ_i **is a random variable** which represents the discrepancy between the i-th observation, y_i, and the model evaluated at the i-th coordinate, $M(x_i|\boldsymbol{q})$.



- Assuming N **independent** realizations and $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, N$, the likelihood can be written as

$$L \equiv p(\{y_i\}_{i=1}^N | \{q_k\}_{k=0}^p) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y_i - U_i(\boldsymbol{q}))^2}{2\sigma^2} \right)$$

- Objective: **jointly infer σ^2 and $\{q_k\}_{k=0}^p$**.

Prior selection

- The choice of a prior should be based, when possible, on some a priori knowledge about the parameters.
- We have $p+2$ unknowns, i.e. the $(p+1)$ coefficients $\{p_k\}_{k=0}^{p}$ and the variance σ^2.
- For each p_k, since we have limited information and for the purpose of this exercise, we choose a **uniform distribution**

$$p(q_k) = \begin{cases} \frac{1}{400} & \text{for } -200 < q_k \leq 200, \\ 0 & \text{otherwise} \ , \end{cases}$$

- In theory, these bounds can be made arbitrarily large.
- We know that σ^2 cannot be negative: this information is what we defined as *a priori* knowledge about a parameter. We assume a Jeffreys prior:

$$\mathcal{P}(\sigma^2) = \begin{cases} \frac{1}{\sigma^2} & \text{for } \sigma^2 > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Posterior

Final form of the joint posterior

$$p_{\text{post}}(\{q_k\}_{k=0}^{p}, \sigma^2 | \{y_i\}_{i=1}^{N}) \propto \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y_i - U_i(\boldsymbol{q}))^2}{2\sigma^2} \right) \right] \mathcal{P}(\sigma^2) \prod_{j=0}^{p} p(q_j)$$

- The problem now reduces to **simulate (sample) this posterior.**
- We are dealing with a $(p+2)$-dimensional probability distribution.
- For high-dimensional cases, which are also the only interesting ones, use Markov chain Monte Carlo (MCMC) methods.
- **MCMC**: class of algorithms suitable to sample high-dimensional probability distributions.
- Must pay attention to mixing ability, convergence...
- Important feature: the quality of the sample improves as a function of the number of steps.

Markov Chain Monte Carlo

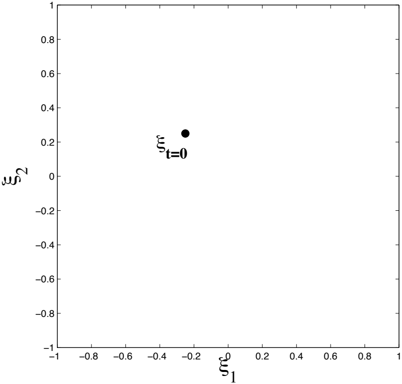
MCMC: features

- Markov chain Monte Carlo (MCMC) methods: class of algorithms aimed at simulating direct draws from some complex distribution $\pi(\xi)$.
- After a large number of steps the random state of the chain follows the desired distribution.
- The quality of the sample improves as a function of the number of steps.
- It is difficult to determine when the chain has converged to the stationary distribution: usually at least $\sim$ 10000 samples.
- The chain should be rapidly mixing, with the stationary distribution is reached quickly and the target probability is explored well and efficiently.
- **Focus on Metropolis-Hastings algorithm**: a random walk with proposal density and a method for accepting/rejecting proposed moves.

Metropolis (MH)

- MH algorithm can draw samples from a target probability distribution, π, requiring only the knowledge of a function proportional to the target PDF.
- It uses a proposal distribution, P, to generate (Markov chain) candidates that are accepted or rejected according to a certain rule. Let P be a Gaussian for simplicity.

1 Let $\boldsymbol{\xi}_{t=0}$ be an initial guess for a 2D problem.

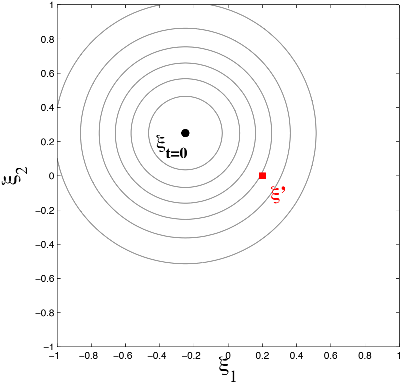


Metropolis (MH)

- MH algorithm can draw samples from a target probability distribution, π, requiring only the knowledge of a function proportional to the target PDF.
- It uses a proposal distribution, P, to generate (Markov chain) candidates that are accepted or rejected according to a certain rule. Let P be a Gaussian for simplicity.

1 Let $\boldsymbol{\xi}_{t=0}$ be an initial guess for a 2D problem.

2 Draw a candidate $\boldsymbol{\xi}'$ from a Gaussian centered on the current state: $\boldsymbol{\xi}' \sim \mathcal{N}(\boldsymbol{\xi}_0, Cov)$ where Cov is chosen a priori.



Metropolis (MH)

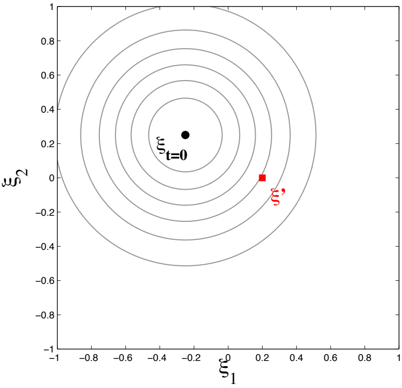
- MH algorithm can draw samples from a target probability distribution, π, requiring only the knowledge of a function proportional to the target PDF.
- It uses a proposal distribution, P, to generate (Markov chain) candidates that are accepted or rejected according to a certain rule. Let P be a Gaussian for simplicity.

1 Let $\boldsymbol{\xi}_{t=0}$ be an initial guess for a 2D problem.

2 Draw a candidate $\boldsymbol{\xi}'$ from a Gaussian centered on the current state: $\boldsymbol{\xi}' \sim \mathcal{N}(\boldsymbol{\xi}_0, Cov)$ where Cov is chosen a priori.

3 Calculate the ratio:

$$r = \frac{\pi(\boldsymbol{\xi}')}{\pi(\boldsymbol{\xi}_0)},$$



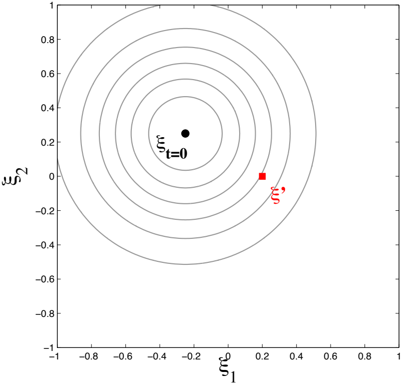
Metropolis (MH)

- MH algorithm can draw samples from a target probability distribution, π, requiring only the knowledge of a function proportional to the target PDF.
- It uses a proposal distribution, P, to generate (Markov chain) candidates that are accepted or rejected according to a certain rule. Let P be a Gaussian for simplicity.

1. Let $\boldsymbol{\xi}_{t=0}$ be an initial guess for a 2D problem.
2. Draw a candidate $\boldsymbol{\xi}'$ from a Gaussian centered on the current state: $\boldsymbol{\xi}' \sim \mathcal{N}(\boldsymbol{\xi}_0, Cov)$ where Cov is chosen a priori.
3. Calculate the ratio:

$$r = \frac{\pi(\boldsymbol{\xi}')}{\pi(\boldsymbol{\xi}_0)},$$

4. Draw a random number $\alpha \sim U(0, 1)$.



Metropolis (MH)

- MH algorithm can draw samples from a target probability distribution, π, requiring only the knowledge of a function proportional to the target PDF.
- It uses a proposal distribution, P, to generate (Markov chain) candidates that are accepted or rejected according to a certain rule. Let P be a Gaussian for simplicity.

1 Let $\boldsymbol{\xi}_{t=0}$ be an initial guess for a 2D problem.

2 Draw a candidate $\boldsymbol{\xi}'$ from a Gaussian centered on current state: $\boldsymbol{\xi}' \sim \mathcal{N}(\boldsymbol{\xi}_0, Cov)$ where Cov is chos a priori.

3 Calculate the ratio:

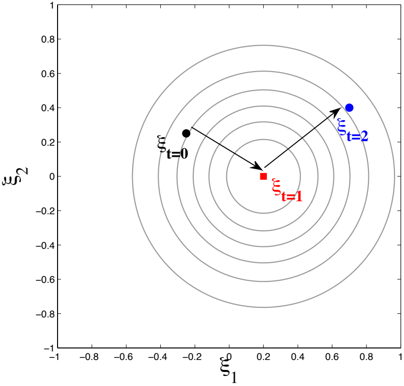
$$r = \frac{\pi(\boldsymbol{\xi}')}{\pi(\boldsymbol{\xi}_0)},$$

4 Draw a random number $\alpha \sim U(0, 1)$.

5 Chain moves (i.e. candidate is accepted/rejected) according to:

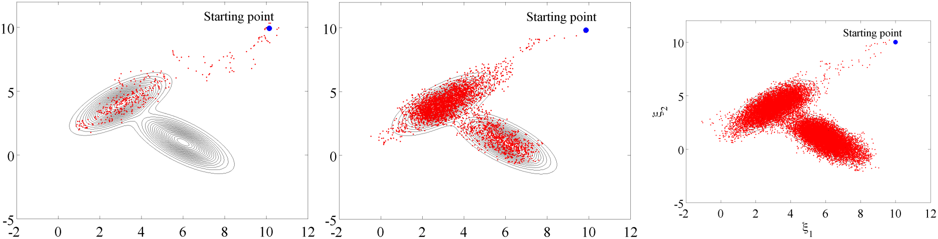
$$\boldsymbol{\xi}_1 = \begin{cases} \boldsymbol{\xi}' & \text{if} \quad \alpha < r, \\ \boldsymbol{\xi}_0 & \textit{otherwise}. \end{cases}$$

6 Repeat the loop.



Sensitivity to number of steps

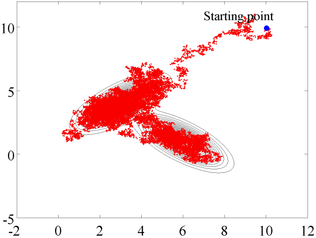
- The proposal distribution has covariance: $\Sigma_{prop} = 0.1 * \mathbf{I}_2$.
- Results for 3 different values of total steps $n = 500$, 5000 and 25000.
- The larger n, the better the approximation.



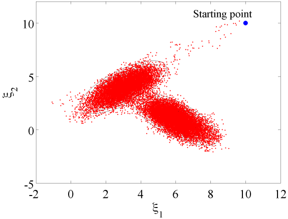
$n = 500$ $n = 5000$ $n = 25000$

Sensitivity to proposal amplitude

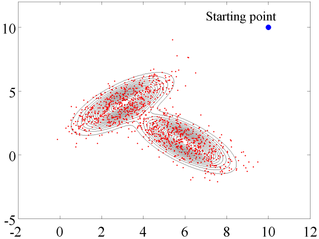
- The proposal amplitude must be tuned to obtain good exploration of the space and fast convergence of the chain toward the high-probability regions.
- Results shown for $0.005 * \mathbf{I}_2$, $0.1 * \mathbf{I}_2$ and $50 * \mathbf{I}_2$.
- The smaller the proposal amplitude, the larger the number of the accepted moves.
- Large proposals lead to small acceptance and slow exploration of the space.
- Ideally, the acceptance rate should be between 30 to 60%.



$amplitude = 0.005$
acceptance rate $\sim$ 95 %.



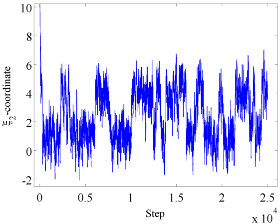
$amplitude = 0.1$
acceptance rate $\sim$ 50 %.



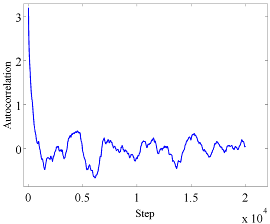
$amplitude = 50$
acceptance rate $\sim$ 4 %.

Sensitivity to proposal amplitude

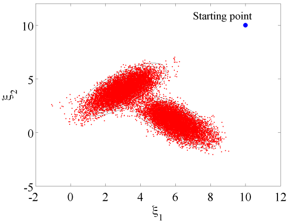
- To evaluate the mixing properties of a chain:
  - visually should look like a white noise.
  - the autocovariance should be rapidly decaying.
  - the acceptance rate should be 30 to 60%.
- Before computing statistics, the initial steps before convergence should be dropped: these steps are referred to as "burn-in" period.
- The burn-in period is estimated from the autocorrelation as the step at which it drops to and becomes oscillatory around zero: in this case it is about 3000 steps.



Chain for ξ_2 showing the bimodality.



Autocovariance for chain of ξ_2.



Chain samples after omitting 3000 steps.

Back to polynomial inference example

Zeroth-order model

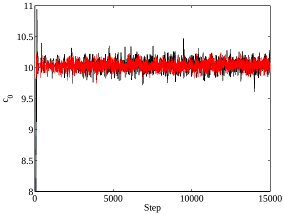
- Suppose that we infer a zeroth-order polynomial:

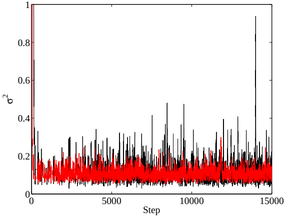
$$M(x|\boldsymbol{q}) = q_0$$

- We know that this is far from the true model defined before, which was a fourth-order polynomial.

Two-dimensional joint posterior

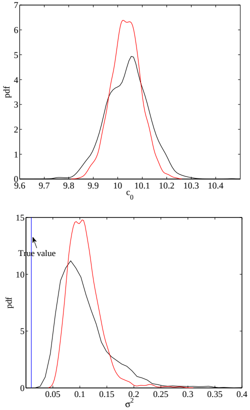
$$p_{\text{post}}(q_0, \sigma^2 | \{y_i\}_{i=1}^N) \propto \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y_i - q_0)^2}{2\sigma^2} \right) \right] \mathcal{P}(\sigma^2)\ p(q_0)$$

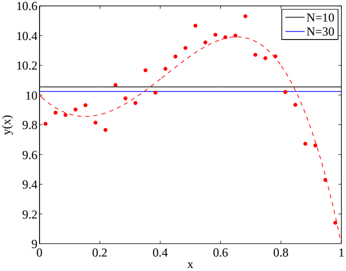




Posterior distributions

- Chain samples can be used to estimate the marginalized posteriors of the parameters via KDE.





This approach only allows us to infer the mean value.

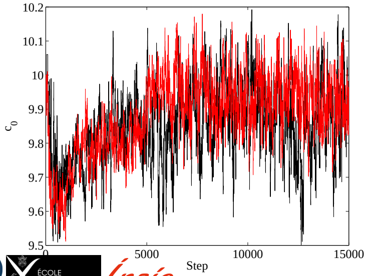
fourth-order model

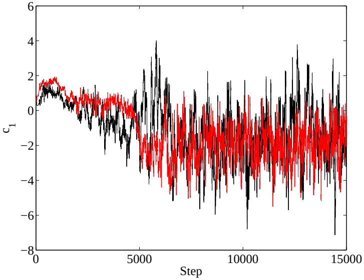
- Suppose that we infer a fourth-order polynomial:

$$M(x|\boldsymbol{q}) = q_0 + q_1 x + q_2 x^2 + q_3 x^3 + q_4 x^4$$

Six-dimensional joint posterior

$$p_{\text{post}}(\{q_k\}_{k=0}^4, \sigma^2 | \{y_i\}_{i=1}^N) \propto \left[\prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y_i - U_i(\boldsymbol{q}))^2}{2\sigma^2} \right) \right] \mathcal{P}(\sigma^2) \prod_{j=0}^p p(q_j)$$

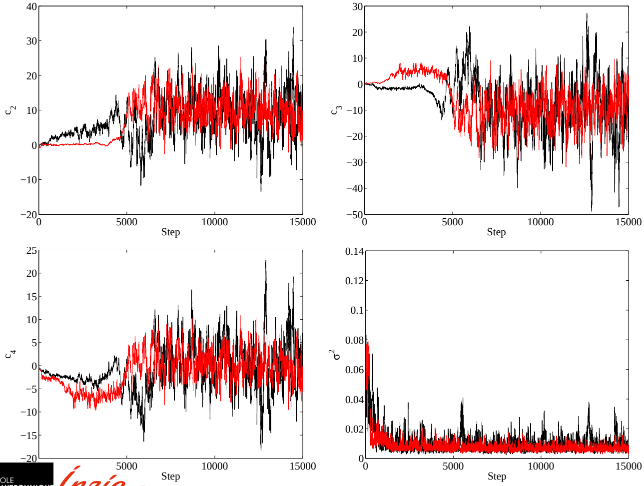




fourth-order model

- Suppose that we infer a fourth-order polynomial:

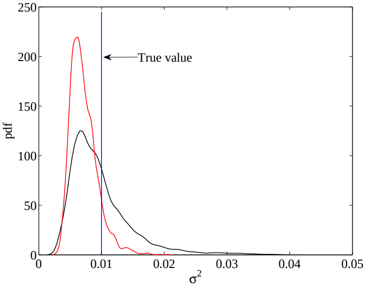
$$M(x|\boldsymbol{q}) = q_0 + q_1 x + q_2 x^2 + q_3 x^3 + q_4 x^4$$



fourth-order model

- Suppose that we infer a fourth-order polynomial:

$$M(x|\boldsymbol{q}) = q_0 + q_1 x + q_2 x^2 + q_3 x^3 + q_4 x^4$$



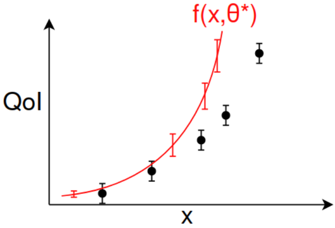
Inference with model error

Calibration with model error

Calibration equation:

$$\overbrace{y_{\text{obs}}(x)}^{\text{observations}} = \overbrace{f(x,\boldsymbol{\theta}^*)}^{\text{computer model}} + \overbrace{z(x)}^{\text{model discrepancy}} + \overbrace{\epsilon(x)}^{\text{measurement error}},$$

where $\boldsymbol{\theta}^*$ is the "best value" of the model parameters.



- $\epsilon(x)$ is Gaussian noise, $z(x)$ is a Gaussian Process.

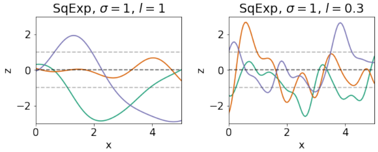
Gaussian Process model

- A Gaussian Process is a probability distribution over functions.
- It requires a mean function μ and a kernel c_ψ.
- Its hyperparameters are noted ψ.

- Squared Exponential kernel:

$$c_\psi(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right). \quad (3)$$

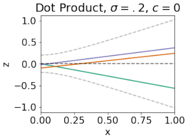
$\psi = (\sigma, l)$.



- Dot Product kernel:

$$c_\psi(x, x') = \sigma^2 + (x - c) * (x' - c). \quad (4)$$

$\psi = (\sigma, c)$.



- **One value of ψ ⇔ one "kind" of model error.**

Kennedy-O'Hagan calibaration

- Classical assumptions: $z|\psi \sim \mathrm{GP}(0, c_\psi)$; $\epsilon|\sigma^2_{\mathrm{mes}} \sim \mathrm{N}(0, \sigma^2_{\mathrm{mes}})$; $z_{\boldsymbol{\theta}} \perp\!\!\!\perp \epsilon$.
- The likelihood function writes:

$$\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \psi \sim \mathrm{N}(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{C}_\psi + \sigma^2_{\mathrm{mes}}\mathbf{I}_n),$$

where $\mathbf{f}_{\boldsymbol{\theta}}$ is the predictions of the computer model, $\mathbf{C}_\psi$ the covariance matrix of model error, $\mathbf{I}_n$ the identity matrix.

Kennedy-O'Hagan calibaration

Recall the KOH calibration equation:

$$y_{\text{obs}}(x) = f(x, \boldsymbol{\theta}^*) + z(x) + \epsilon(x) \tag{5}$$

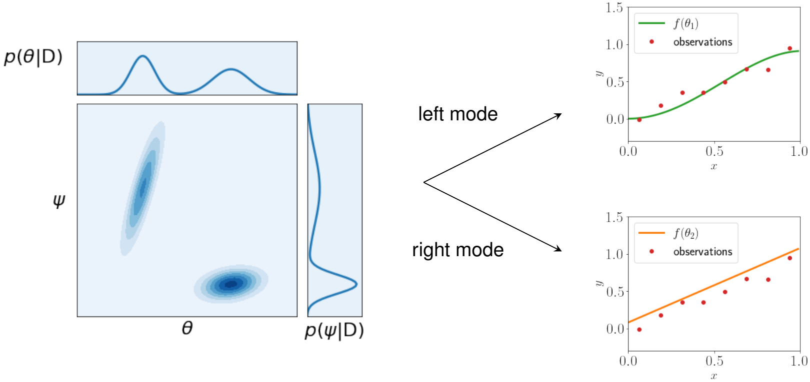
where $\boldsymbol{\theta}^*$ is the "best value" of the model parameters.
Hyperparameters of the model error are estimated with a single value:

$$\widehat{\psi}_{\text{KOH}} = \arg\max_{\psi} \; \text{p}(\psi | \mathbf{y}_{\text{obs}}). \tag{6}$$

- **Problem 1**: what is the meaning of a "best value" of model parameters ? $\rightarrow$ **Lack of identifiability** (F. Liu, Bayarri, and Berger 2009; Arendt et al. 2012).
- **Problem 2**: A single distribution is used for $z(x)$, inappropriate when different model parameters values can provide equally good representations of the data.

Comparison: Bayesian / Kennedy-O'Hagan / FMP

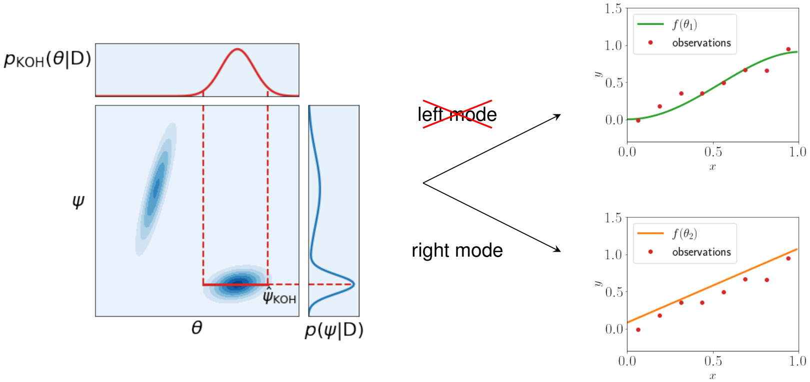
- Assumption: the posterior distribution p($\theta, \psi | D$) is a mixture of Gaussians with well-separated modes.



Full Bayesian solution

Comparison: Bayesian / Kennedy-O'Hagan / FMP

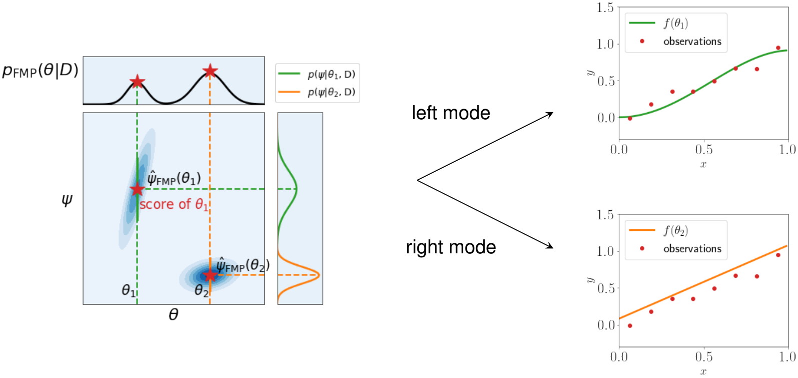
- Assumption: the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\psi}|D)$ is a mixture of Gaussians with well-separated modes.



Kennedy-O'Hagan solution

Comparison: Bayesian / Kennedy-O'Hagan / FMP

- Assumption: the posterior distribution $p(\theta, \psi | D)$ is a mixture of Gaussians with well-separated modes.



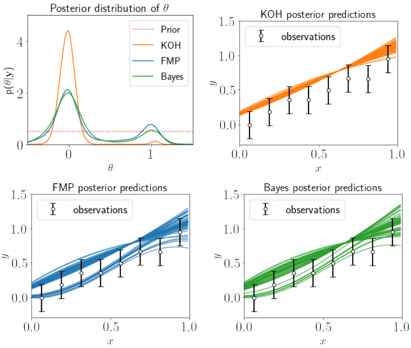
Full Maximum a Posteriori solution

Comparison: Bayesian / Kennedy-O'Hagan / FMP

10 noisy observations from the true process $y(x) = x$, with $x \in [0, 1]$, with computer model:

$$f(x, \theta) = x \sin(2\theta x) + (x + 0.15)(1 - \theta), \qquad (15)$$

with a single parameter $\theta \in [-0.5, 1.5]$.
Model error uses a squared exponential kernel with uniform priors.



- KOH estimation finds a single family of predictions, with $\theta \approx 0$.
- The FMP posterior finds the correct balance between the two interpretations of the data.
- The FMP method exhibits a more conservative behaviour.

comparison of posterior predictions

Table of contents

Standard approach

Inference of $\boldsymbol{q} \in \mathbb{R}^d$ from $\mathcal{O} \doteq \{y_i \in \mathbb{R}, i = 1, \dots, M\}$ (measurements)
Bayes' formula:

$$p_{\mathrm{post}}(\boldsymbol{q}|\mathcal{O}) \propto L(\mathcal{O}|\boldsymbol{q})p(\boldsymbol{q}),$$

with $p(\boldsymbol{q})$ (prior), $L(\mathcal{O}|\boldsymbol{q})$ (likelihood) and $p_{\mathrm{post}}(\boldsymbol{q}|\mathcal{O})$ (posterior)
Model for the measurement errors:

$$y_i = U_i(\boldsymbol{q}) + \epsilon_i, \quad \epsilon_i = N(0, \sigma_i^2),$$

$U_i(\boldsymbol{q})$ is the model prediction of the i-th measurement
Likelihood becomes:

$$L(\mathcal{O}|\boldsymbol{q}) \doteq \prod_{i=1}^{M} \exp\left[-\frac{|y_i - U_i(\boldsymbol{q})|^2}{2\sigma_i^2}\right].$$

Posterior sampled, for instance using **Markov Chain Monte Carlo (MCMC), rely heavily on multiple evaluations** of

$$\boldsymbol{q} \mapsto \boldsymbol{U}(\boldsymbol{q}) \doteq (U_1 \cdots U_M)(\boldsymbol{q})$$

Surrogate based posterior

Substitute costly model $\boldsymbol{U}$ with a surrogate $\hat{\boldsymbol{U}}$ with inexpensive evaluations.
The surrogate-based posterior becomes

$$\hat{p}_{\text{post}}(\boldsymbol{q}|\mathcal{O}) \propto \hat{L}(\mathcal{O}|\boldsymbol{q})p(\boldsymbol{q}), \quad \hat{L}(\mathcal{O}|\boldsymbol{q}) \doteq \prod_{i=1}^{M} \exp\left[-\frac{|y_i - \hat{U}_i(\boldsymbol{q})|^2}{2\sigma_i^2}\right].$$

Error estimate [Marzouk, Xiu, Najm, ...]

$$\text{KL}(p_{\text{post}}|\hat{p}_{\text{post}}) \doteq \int\cdots\int \log\frac{p_{\text{post}}(\boldsymbol{q}|\mathcal{O})}{\hat{p}_{\text{post}}(\boldsymbol{q}|\mathcal{O})} p_{\text{post}}(\boldsymbol{q}|\mathcal{O}) d\boldsymbol{q} \leq C(\mathcal{O}) \left(\sum_{i=1}^{M} \|U_i - \hat{U}_i\|^2_{L_2(p)}\right)^{1/2},$$

where

$$\|u\|^2_{L_2(p)} \doteq \int\cdots\int |u(\boldsymbol{q})|^2 p(\boldsymbol{q}) d\boldsymbol{q}$$

Motivate for surrogate minimizing $\|U_i - \hat{U}_i\|_{L_2(p)}$.
PC surrogates (off-line construction) [Marzouk, Najm]

$$U_i(\boldsymbol{q}) \approx \hat{U}_i(\boldsymbol{q}) \doteq \sum_{\alpha=1}^{P} [U_i]_\alpha \Psi_\alpha(\boldsymbol{q}),$$

Support the possibly high convergence rate of the approximation.

Surrogate based posterior

Substitute costly model $\boldsymbol{U}$ with a surrogate $\hat{\boldsymbol{U}}$ with inexpensive evaluations.
The surrogate-based posterior becomes

$$\hat{p}_{\mathrm{post}}(\boldsymbol{q}|\mathcal{O}) \propto \hat{L}(\mathcal{O}|\boldsymbol{q})p(\boldsymbol{q}), \quad \hat{L}(\mathcal{O}|\boldsymbol{q}) \doteq \prod_{i=1}^{M} \exp\left[-\frac{|y_i - \hat{U}_i(\boldsymbol{q})|^2}{2\sigma_i^2}\right].$$

Error estimate [Marzouk, Xiu, Najm, ...]

$$\mathrm{KL}(p_{\mathrm{post}}|\hat{p}_{\mathrm{post}}) \doteq \int\!\cdots\!\int \log \frac{p_{\mathrm{post}}(\boldsymbol{q}|\mathcal{O})}{\hat{p}_{\mathrm{post}}(\boldsymbol{q}|\mathcal{O})} p_{\mathrm{post}}(\boldsymbol{q}|\mathcal{O}) d\boldsymbol{q} \leq C(\mathcal{O}) \left(\sum_{i=1}^{M} \|U_i - \hat{U}_i\|_{L_2(p)}^2\right)^{1/2},$$

Constant $C(\mathcal{O})$ can be large if the observations are very informative:

$$\mathbb{E}_{p_{\mathrm{post}}}\left\{|U_i - \hat{U}_i|^2\right\} = \int\!\cdots\!\int |U_i(\boldsymbol{q}) - \hat{U}_i(\boldsymbol{q})|^2 p_{\mathrm{post}}(\boldsymbol{q}|\mathcal{O}) d\boldsymbol{q}.$$

But the posterior is unknown!

Iterative surrogate construction

Iterative approach

Basic idea:

- a sequence of polynomial surrogates $\hat{\boldsymbol{U}}^{(k)}(\boldsymbol{q})$ incorporating progressively new observations of $\boldsymbol{U}$
- take new observations of the model to improve the surrogate error (in the posterior norm)

Denote $\mathcal{D} = \{(\boldsymbol{q}^j, \boldsymbol{U}^j, \rho^j), j = 1, \dots, n\}$ the set of collected model observations:

- $\boldsymbol{q}^j$ observation point
- $\boldsymbol{U}^j = \boldsymbol{U}(q^j)$ full model evaluation
- $\rho^j > 0$ trust index

Iterative approach

Basic idea:

- a sequence of polynomial surrogates $\hat{\boldsymbol{U}}^{(k)}(\boldsymbol{q})$ incorporating progressively new observations of $\boldsymbol{U}$
- take new observations of the model to improve the surrogate error (in the posterior norm)

Model construction:

- select a subset $\mathcal{I}^{(k)}$ of model observations indexes
- find the polynomial approximation

$$\boldsymbol{U}(\boldsymbol{q}) \approx \boldsymbol{U}^{(k)}(\boldsymbol{q}) = \sum_{\alpha=1}^{P} [\boldsymbol{U}]_{\alpha}^{(k)} \Psi_{\alpha}(\boldsymbol{\eta}^{(k)}(\boldsymbol{q})),$$

- solving a regularized regression problem of type

$$\boldsymbol{u} = \arg\min_{\boldsymbol{v} \in \mathbb{R}^P} \sum_{j \in \mathcal{I}} \rho^j \left| U^j - \sum_{\alpha=0}^{P} \Psi_{\alpha}(\boldsymbol{q}^j) v_{\alpha} \right|^2 + \lambda \sum_{\alpha=0}^{p} |v_{\alpha}|.$$

Iterative approach

Basic idea:

- a sequence of polynomial surrogates $\hat{\boldsymbol{U}}^{(k)}(\boldsymbol{q})$ incorporating progressively new observations of $\boldsymbol{U}$
- take new observations of the model to improve the surrogate error (in the posterior norm)

Resampling: (completing the model observations set)

$$\hat{p}^{(k)}_{\mathrm{post}}(\boldsymbol{q}|\mathcal{O}) \propto \exp\left[\sum_{i=1}^{M} -\frac{\left|y_i - \hat{U}^{(k)}_i(\boldsymbol{q})\right|^2}{2\sigma_i^2}\right] p(\boldsymbol{q}).$$

- Draw several independent samples $\boldsymbol{q}^j$ form $\hat{p}^{(k)}_{\mathrm{post}}$
- Compute model prediction $\boldsymbol{U}^j = \boldsymbol{U}(\boldsymbol{q}^j)$
- Define the trust index of the new observation as

$$(\Delta^j)^2 \doteq \sum_{i=1}^{M} \frac{|U_i^j - \hat{U}_i^{(k)}(\boldsymbol{q}^j)|^2}{2\sigma_i^2}, \rho^j \doteq \frac{1}{\max(\epsilon_t, \Delta^j)}.$$

General Iterative Algorithm

ALGORITHM 1: Iterative Procedure for the Construction of the Posterior Fitted Surrogate.

Require: Initial number of observations n_0, number of new observations at each step n_{add}, measurements set $\mathcal{O}$, maximal number of model evaluations $n_{\max}$

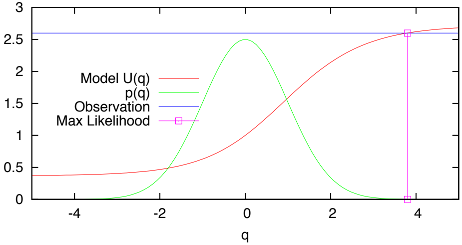
```
Initialization:
n = 1, D = ∅                                              ▷ Initialize the observations set
for j = 1, ... n_0 do                                     ▷ Generate the initial observations
    Draw q^n from p(q), D ← D ∪ {(q^n, U(q^n), ρ_0)}, n ← n + 1
end for
k = 0, construct Û^(0) with I^(0) = {1, ..., n}            ▷ Construct initial surrogate
while n < n_max do
    for j = 1, ... n_add do
        Draw q^n from p̂_post^(k)(q|O)                      ▷ Sample surrogate-based posterior
        Compute U(q^n) and observation weight ρ^n from (19) ▷ Set observation
        D ← D ∪ {(q^n, U(q^n), ρ_0)}, n ← n + 1            ▷ Update observation set
    end for
    k ← k + 1
    Define I^(k), construct Û^(k)                          ▷ Specify observations to use and compute surrogate
end while
Return Û^(k)                                              ▷ Return final surrogate
```

Elementary 1D problem

Simple one-dimensional test problem

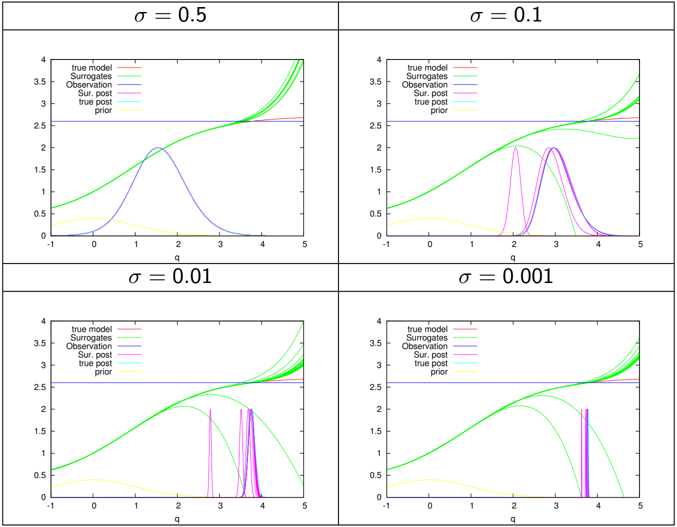
Problem settings

- ✓ $q \in \mathbb{R}^{d=1}$ and non-polynomial model: $U(q) = \exp[\tanh(q/2)]$
- ✓ standard Gaussian prior: $q \sim p(q) = \exp[-q^2/2]/\sqrt{2\pi}$
- ✓ single observation $O = 2.6$, likelihood maximized for $q = 3.8$



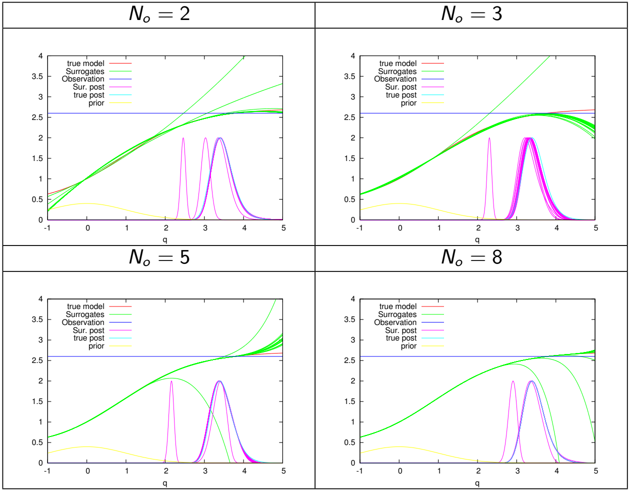
- ✓ for small noise level, $\sigma \ll 1$, prior and posterior are very distant
- ✓ high pol. order N_o required to globally approximate $U(q)$ over few std range

Elementary 1D problem



Elementary 1D problem

Effect of polynomial degree N_o (noise level $\sigma = 0.05$; sampling $|\mathcal{D}^{(k)}|_{k=1\ldots10} = 2N_o$)

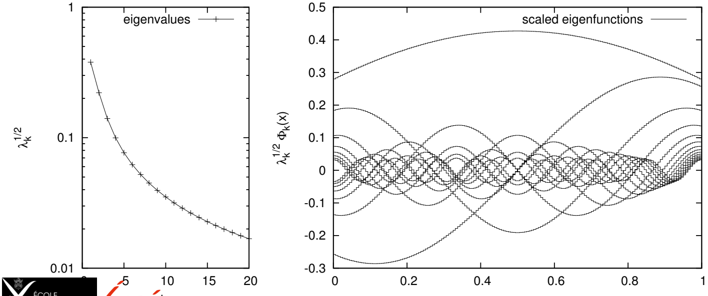


(1D) Elliptic problem

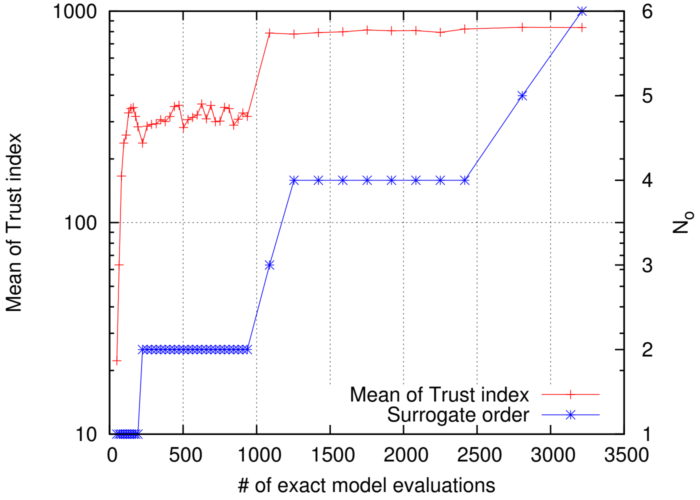
$$\partial \left(\kappa(x) \partial u(x) \right) = -g, \quad \forall x \in]0,1[$$

- Log-normal random field, exponential type covariance
- Retain the first 15 modes: $\boldsymbol{q} \in \mathbb{R}^{15}$

$$\log \kappa(x,\omega) = \sum_{l=1}^{l=15} \sqrt{\lambda_l} \phi_l(x) q_l(\omega), \quad \boldsymbol{q} \sim N(\mathbf{0}, \mathbf{I}).$$



Case of measurements from truth at $q = 0$ and $\sigma = 0.001$



Case of measurements from truth at $\boldsymbol{q} = 0$ and $\sigma = 0.001$

N_{max} ($\lvert\mathcal{D}\rvert$)	Iterative Surrogate			Global Surrogate			Error ratio
	$\epsilon^{(k)}$	${N_o}^{(k)}$	N_{PC}	ϵ^G	${N_o}^G$	N_{PC}	$\epsilon^{(k)}/\epsilon^G$
500 (503)	$3.1\ 10^{-3}$	2	16	$9.4\ 10^{-3}$	4	166	0.33
1000 (1088)	$3.8\ 10^{-4}$	4	166	$6.8\ 10^{-3}$	4	166	0.06
2000 (2084)	$3.7\ 10^{-4}$	4	166	$3.2\ 10^{-3}$	6	406	0.11
2500 (2807)	$2.9\ 10^{-4}$	6	406	$2.7\ 10^{-3}$	6	406	0.11
3000 (3213)	$4.1\ 10^{-4}$	6	406	$2.5\ 10^{-3}$	6	406	0.16

Table 1: Using ${N_o}^{(0)} = 1$, and different N_{max} as indicated. $\sigma = 0.01$.

Case of measurements from truth at $\boldsymbol{q} = 0$ and $\sigma = 0.001$

$N_{max} = 500$	$N_{max} = 2000$	$N_{max} = 3000$

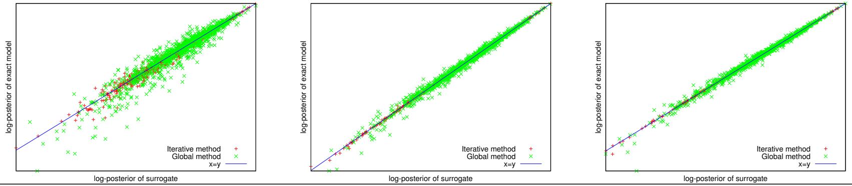


Figure 3: True log-posterior against surrogate log-posteriors values for 1000 sample points drawn from $\hat{p}^{(k)}_{post}$ (Iterative method) and $\hat{p}^{G}_{post}$ (Global method) respectively. Surrogates are constructed with different values of N_{max}, as indicated, and for $\sigma = 0, .01$, $\bar{\boldsymbol{q}} = 0$, ${N_o}^{(0)} = 1$.

Impact of measurement

$\sigma = 0.05$	$\sigma = 0.01$	$\sigma = 0.001$

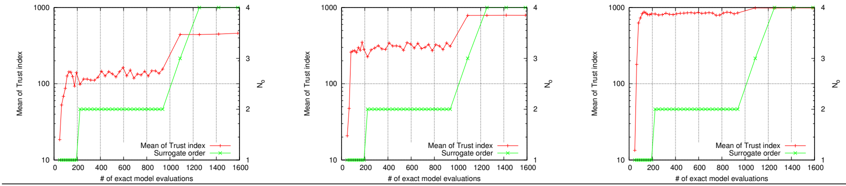


Figure 5: Evolutions of the averaged trust-index for $\bar{\boldsymbol{q}} = 0$, $\mathrm{N}_{\max} = 1500$, $\mathrm{N_o}^{(0)} = 1$ and different values for σ as indicated. Also shown are the evolutions of the polynomial order of the successive surrogates (left axis).

Impact of measurement

	$\overline{\Delta} = 0.5$	$\overline{\Delta} = 1.0$	$\overline{\Delta} = 2.0$	N_o	N_{PC}
$\epsilon^{(k)}$	$2.7\ 10^{-5}$	$7.5\ 10^{-6}$	$3.1\ 10^{-6}$	4	166
ϵ^{G}	$2.1\ 10^{-3}$	$7.6\ 10^{-3}$	$2.8\ 10^{-2}$	6	406
$\epsilon^{(k)}/\epsilon^{G}$	$1.3\ 10^{-2}$	$9.9\ 10^{-4}$	$1.1\ 10^{-4}$	-	-

Table 3: Using ${N_o}^{(0)} = 2$, $N_{max} = 1500$, $\sigma = 0.001$.

Impact of measurement

$\overline{\Delta} = 0.5$	$\overline{\Delta} = 1.0$	$\overline{\Delta} = 2.0$

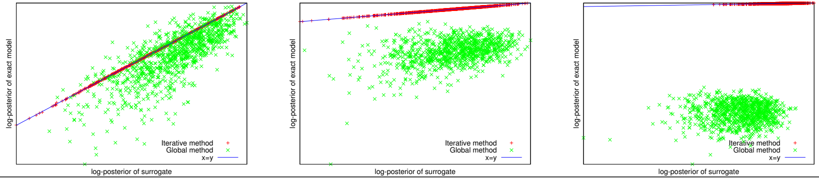
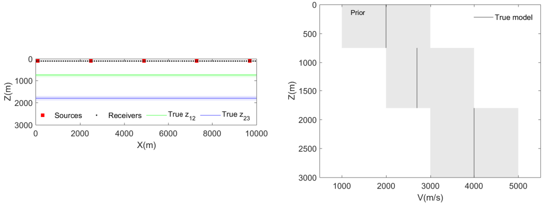


Figure 6: True log-posterior against surrogate log-posteriors values for 1000 sample points drawn from $\hat{p}^{(k)}_{\text{post}}$ (Iterative method) and $\hat{p}^{G}_{\text{post}}$ (Global method) respectively. Case of construction with $\mathrm{N}_{\text{max}} = 1500$, for $\bar{\boldsymbol{q}} = 0$, $\mathrm{N_o}^{(0)} = 1$ and different σ as indicated.

[OLM & D. Lucor. ESAIM Proc., 2018]

Application to tomography

[Sochala, Gesret & OLM. Int. J. Geomath., 2022]



(a) Acquisition geometry and interface locations

(b) Layers' velocities with the true values of z_{12} and z_{23}

Figure 12: Schematic description of the refraction test case

Parameter	v_1 (m/s)	v_2 (m/s)	v_3 (m/s)	z_{12} (m)	z_{23} (m)
Prior ranges	$[1,3] \times 10^3$	$[2,4] \times 10^3$	$[3,5] \times 10^3$	$[700, 850]$	$[1700, 1900]$
True values	2000	2700	4000	750	1800

Table 4: Refraction case: ranges of the uniform prior distributions of the velocity model parameters and true values used for the synthetic observations.

Application to tomography

[Sochala, Gesret & OLM. Int. J. Geomath., 2022]

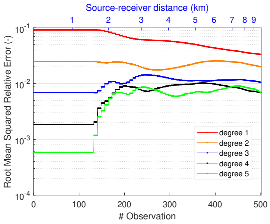


Figure 13: Refraction case: RMSRE of the PC surrogates of the traveltimes for maximum total degree d° from 1 to 5.

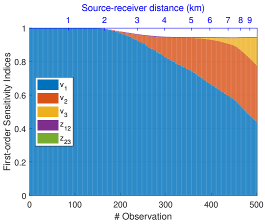
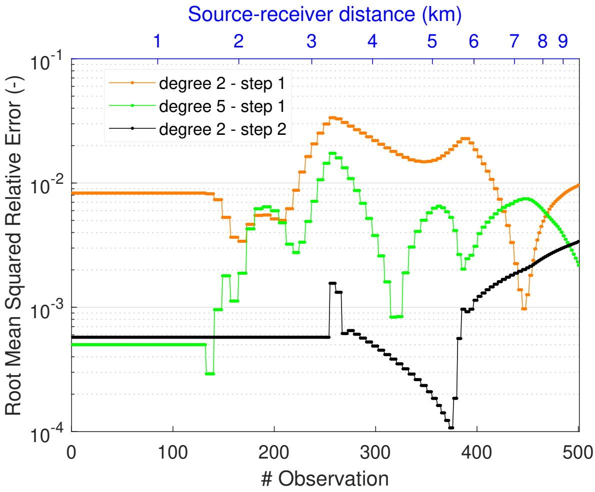


Figure 14: Refraction case: first-order sensitivity indices estimated with PC surrogates of maximum total degree $d^\circ = 5$.

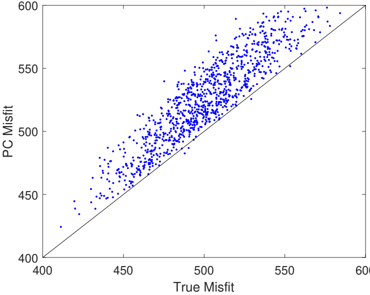
Application to tomography

[Sochala, Gesret & OLM. Int. J. Geomath., 2022]

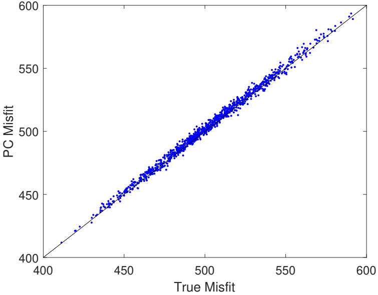


Application to tomography

[Sochala, Gesret & OLM. Int. J. Geomath., 2022]



(a) $d^\circ = 5$ (Step 1)

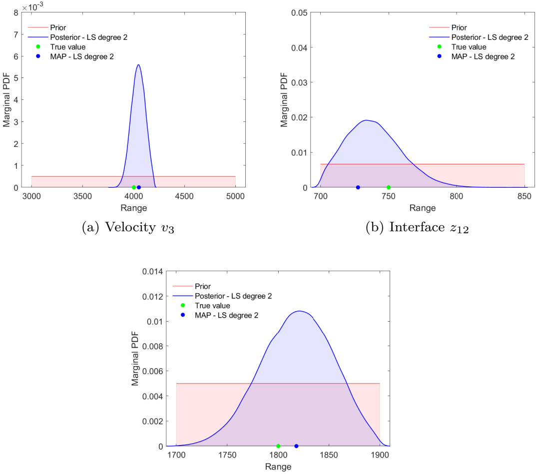


(b) $d^\circ = 5$ and $d_2^\circ = 2$ (Step 2)

Figure 17: Refraction case: true misfit function versus approximated PC misfit function plotted for 1, 000 independent values for $\boldsymbol{m}$ extracted from the MCMC sampling.

Application to tomography

[Sochala, Gesret & OLM. Int. J. Geomath., 2022]



(a) Velocity v_3

(b) Interface z_{12}

(c) Interface z_{23}

Figure 18: Refraction case: priors and posterior marginals of v_3, z_{12} and z_{23} based on the polynomial surrogate constructed with $d_2^\circ = 2$ on a sample set of 1,000 values drawn using the initial PC surrogates of the traveltimes with $d^\circ = 5$.

Table of contents

Selection of Observations: an example

Debris flow model

- Flow of debris (mud, gravels, small rocks, ...)
- Empirical / Phenomenological models
- Parameter calibration on experiments at USGS

Governing equations GeoClaw

$$
\begin{aligned}
&\frac{\partial h}{\partial t}+\frac{\partial (hu)}{\partial x}+\frac{\partial (hv)}{\partial y}=\varphi_1,\\
&\frac{\partial (hu)}{\partial t}+\frac{\partial}{\partial x}(hu^2)+\kappa\frac{\partial}{\partial y}(0.5g_z h^2)+\frac{\partial (huv)}{\partial y}+\frac{h(1-\kappa)}{\rho}\frac{\partial p_b}{\partial x}=\varphi_2,\\
&\frac{\partial (hv)}{\partial t}+\frac{\partial (huv)}{\partial x}+\frac{\partial}{\partial y}(hv^2)+\kappa\frac{\partial}{\partial y}(0.5g_z h^2)+\frac{h(1-\kappa)}{\rho}\frac{\partial p_b}{\partial y}=\varphi_3,\\
&\frac{\partial (hm)}{\partial t}+\frac{\partial (hum)}{\partial x}+\frac{\partial (hvm)}{\partial y}=\varphi_4,\\
&\frac{\partial p_b}{\partial t}-\chi u\frac{\partial h}{\partial x}+\chi\frac{\partial (hu)}{\partial x}+u\frac{\partial p_b}{\partial x}-\chi v\frac{\partial h}{\partial y}+\chi\frac{\partial hu}{\partial y}+v\frac{\partial p_b}{\partial y}=\varphi_5.
\end{aligned}
$$

Debris flow model

- Flow of debris (mud, gravels, small rocks, ...)
- Empirical / Phenomenological models
- Parameter calibration on experiments at USGS

Non-linear source terms [Iverson & George, 2014]

$$\varphi_1 = \frac{(\rho - \rho_f)}{\rho}\frac{-2k}{h\mu}(p_b - \rho_f g_z h),$$

$$\varphi_2 = h g_x + u\frac{(\rho - \rho_f)}{\rho}\frac{-2k}{h\mu}(p_b - \rho_f g_z h) - \frac{(\tau_{s,x} + \tau_{f,x})}{\rho},$$

$$\varphi_3 = h g_y + v\frac{(\rho - \rho_f)}{\rho}\frac{-2k}{h\mu}(p_b - \rho_f g_z h) - \frac{(\tau_{s,y} + \tau_{f,y})}{\rho},$$

$$\varphi_4 = \frac{2k}{hu}(p_b - \rho_f g_z h) m \frac{\rho_f}{\rho},$$

$$\varphi_5 = \zeta\frac{-2k}{h\mu}(p_b - \rho_f g_z h) - \frac{3}{\alpha h}\|\boldsymbol{u}\| \tan(\psi),$$

where

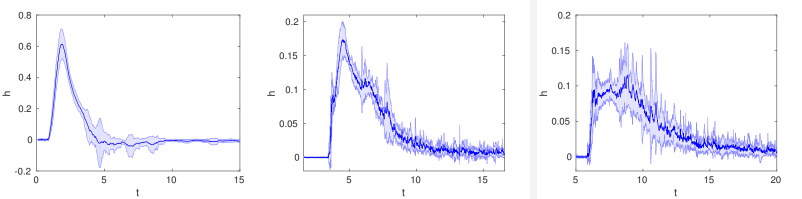
$$\zeta = \frac{3}{2\alpha h} + \frac{g_z \rho_f(\rho - \rho_f)}{4\rho}, \quad \alpha = \frac{a}{m(\rho g_z h - p_b + \sigma_0)}.$$

Debris flow experiment

Inference of model parameters [Iverson & George, 2014]

- static critical-state solid volume fraction (m_{crit})
- initial hydraulic permeability k_0
- pure-fluid viscosity μ
- steady friction contact angle ϕ
- compressibility constant a.

Gate release experiments: available measurements

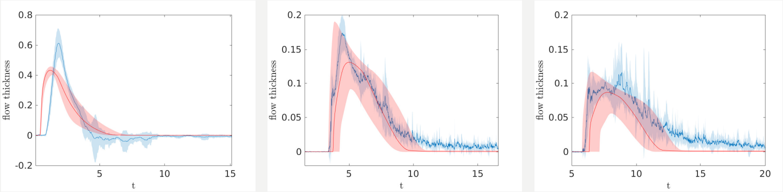


Debris flow model

A priori range of model parameters

$$m_{\text{crit}} \sim \mathscr{U}[0.62, 0.66], \quad k_0 \sim \mathscr{U}_{\log}[10^{-9}, 10^{-8}],$$
$$\mu \sim \mathscr{U}_{\log}[0.005, 0.05], \quad \phi \sim \mathscr{U}[0.62, 0.66], \quad a \sim \mathscr{U}[0.01, 0.05].$$

A priori analyis



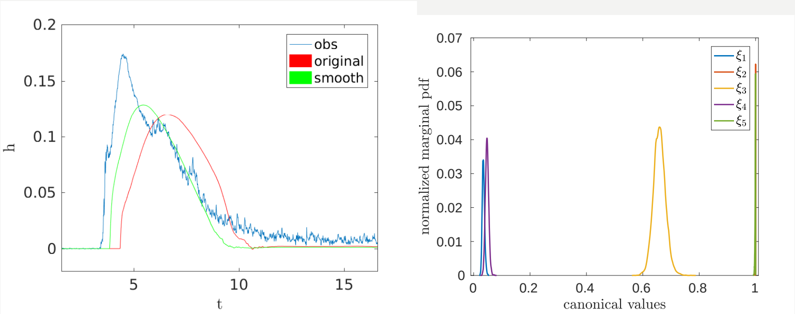
(a) $x = 2$ m (b) $x = 32$ m (c) $x = 66$ m

Independent measurement errors

Naive model: Gaussian likelihood

$$\mathcal{L}(\boldsymbol{d}|\boldsymbol{\xi}) = \prod_{i=1}^{m_d} \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left[-\frac{(\bar{h}_i - \widehat{h}_i(\boldsymbol{\xi}))^2}{2\sigma_i^2}\right]$$

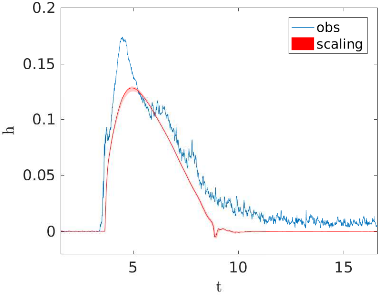
Independent / uncorrelated **"measurement noise"**

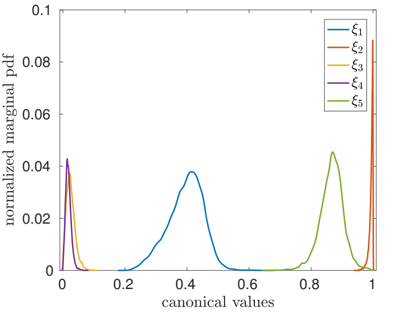


Appreciating inference quality

Trying to fit "important characteristics"

$$\ln(\mathcal{L}(\boldsymbol{d}|\boldsymbol{\xi})) \propto -\left(\frac{t_{\mathrm{arr}} - \widehat{t}_{\mathrm{arr}}(\boldsymbol{\xi})}{2\sigma_{arr}}\right)^2 - \left(\frac{t_{\max} - \widehat{t}_{\max}(\boldsymbol{\xi})}{2\sigma_{t_{\max}}}\right)^2 - \left(\frac{t_{\mathrm{dec}} - \widehat{t}_{\mathrm{dec}}(\boldsymbol{\xi})}{2\sigma_{\mathrm{dec}}}\right)^2 - \left(\frac{h_{\max} - \widehat{h}_{\max}(\boldsymbol{\xi})}{2\sigma_{h_{\max}}}\right)^2.$$



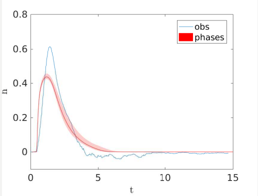


Limits of the model - experimental issues

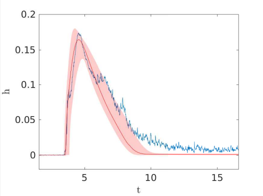
With feedback from experimentalist

Measurements were synchronized:

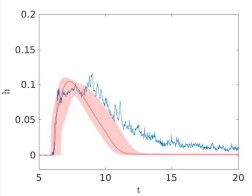
$$\ln(\mathcal{L}(\boldsymbol{d}|\boldsymbol{\xi})) \propto -\left(\frac{T_{\text{grw}} - \widehat{T_{\text{grw}}}(\boldsymbol{\xi})}{2\sigma_{T_{\text{grw}}}}\right)^2 - \left(\frac{T_{\text{dec}} - \widehat{T_{\text{dec}}}(\boldsymbol{\xi})}{2\sigma_{T_{\text{dec}}}}\right)^2 - \left(\frac{h_{\max} - \widehat{h}_{\max}(\boldsymbol{\xi})}{2\sigma_{h_{\max}}}\right)^2,$$



(a) $x = 2m$



(b) $x = 32m$



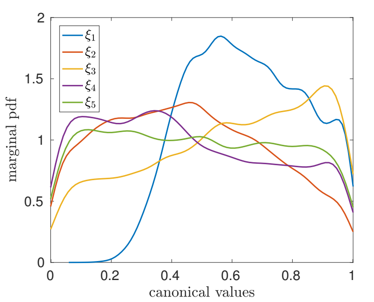
(c) $x = 66m$

Limits of the model - experimental issues

With feedback from experimentalist

Measurements were synchronized:

$$\ln(\mathcal{L}(\boldsymbol{d}|\boldsymbol{\xi})) \propto -\left(\frac{T_{\text{grw}} - \widehat{T_{\text{grw}}}(\boldsymbol{\xi})}{2\sigma_{T_{\text{grw}}}}\right)^2 - \left(\frac{T_{\text{dec}} - \widehat{T_{\text{dec}}}(\boldsymbol{\xi})}{2\sigma_{T_{\text{dec}}}}\right)^2 - \left(\frac{h_{\max} - \widehat{h}_{\max}(\boldsymbol{\xi})}{2\sigma_{h_{\max}}}\right)^2,$$



Take-away

What did we learn?

- **Experimental data may be biased**
- Raw measurements, or complete description of their treatments, are important
- Using all the available data may be counterproductive (yes!)
- If the model is poor, we should focus on basic features of interest, and not insist on obtaining global agreement
- **Models of model error are more robust and easier to propose & test for simple features**

How to select / reduce the experimental data to facilitate the inference problem?

[Navarro, OLM, Mandli, George, Hoteit and Knio. Comp. Geosciences, 2018.]

Optimal Reduction of Observations

Optimal Observations Reduction

Motivation

Bayesian inference in the case of overabundant data

- Weather forecasting
- Seismic wave inversion

Goal

Compute an optimal approximation

$$\min_V \mathscr{L}\left(P(Q \mid Y = y), P(Q \mid W = V^T y)\right)$$

- $\mathscr{L}$ a loss function
- n (random) observations $Y = (Y_i)_{i=1}^n$
- q parameters $Q = (Q_i)_{i=1}^{\mathrm{Nq}}$, $\mathrm{Nq} \ll n$
- r dimensional reduced space $V \in \mathbb{R}^{n \times r}$, $r \ll n$

[Giraldi, OLM, Hoteit and Knio. Comp. Stat. & Data An., 2018]

Linear Gaussian models

Gaussian model

$$Y = BQ + E,$$

- Observations: $Y \sim \mathcal{N}(m_Y, C_Y)$ with values in $\mathbb{R}^n$
- Parameter of interest: $Q \sim \mathcal{N}(m_Q, C_Q)$ with values in $\mathbb{R}^{\mathrm{Nq}}$
- Noise: $E \sim \mathcal{N}(m_E, C_E)$ with values in $\mathbb{R}^n$
- Design matrix: $B \in \mathbb{R}^{n \times \mathrm{Nq}}$
- Forward model: $A(Q) = BQ \sim \mathcal{N}(m_A, C_A)$, and $C_{AQ} = \mathrm{Cov}(A(Q), Q)$

Reduced model

$$W = V^T BQ + V^T E,$$

- Reduced observations: $W \sim \mathcal{N}(m_W, C_W)$ with values in $\mathbb{R}^r$
- Reduced space: $V \in \mathbb{R}^{n \times r}$

Posterior distributions

knowing the realization (a particular measurement) y of Y

Unreduced case

The posterior distribution is $P(Q \mid Y = y) \sim \mathcal{N}(m_\star, C_\star)$ where

$$C_\star = C_Q \left(C_Q + C_{AQ}^T C_E^{-1} C_{AQ} \right)^{-1} C_Q,$$
$$m_\star = C_{AQ}^T C_Y^{-1} (y - m_E) + C_\star C_Q^{-1} m_Q.$$

Reduced model

The posterior distribution is $P(Q \mid W = V^T y) \sim \mathcal{N}(m_V, C_V)$ where

$$C_V = C_Q \left(C_Q + C_{AQ}^T V \left(V^T C_E V \right)^{-1} V^T C_{AQ} \right)^{-1} C_Q,$$
$$m_V = C_{AQ}^T V (V^T C_Y V)^{-1} V^T (y - m_E) + C_V C_Q^{-1} m_Q.$$

CMAP cnrs ÉCOLE POLYTECHNIQUE Inria inventeurs du monde numérique

Invariance property

Proposition (Invariance property)

For all invertible matrices $M \in \mathbb{R}_*^{r \times r}$, we have

$$m_{VM} = m_V \quad \text{and} \quad C_{VM} = C_V.$$

- Posterior distribution invariant under rescaling, rotation or permutation of the observations
- Newton method can not be directly used
- range(V) is more important than V
- Use of a Riemannian trust region algorithm on the Grassmann manifolds Gr(r, n), the set of r-dimensional subspaces of $\mathbb{R}^n$ (see Absil et al. 2007, Manopt and Pymanopt libraries)

Kullback-Leibler based loss functions

Kullback-Leibler divergence

Given two distributions $P(Z_0)$ and $P(Z_1)$ with densities f_{Z_0} and f_{Z_1},

$$\mathrm{D}_{\mathrm{KL}}\left(P(Z_0) \parallel P(Z_1)\right) = \mathbb{E}_{Z_0}\left(\log \frac{f_{Z_0}}{f_{Z_1}}\right).$$

- Quantify the “information lost when $[P(Z_1)]$ is used to approximate $[P(Z_0)]$” (Burnham and Anderson, 2003)
- Positive and null iff $P(Z_0) = P(Z_1)$
- Asymmetric quantity

Kullback-Leibler based loss functions

Kullback-Leibler divergence minimization

$$\min_{[V]\in \mathrm{Gr}(r,n)} \mathrm{D}_{\mathrm{KL}} \left(P(Q \mid Y = y) \parallel P(Q \mid W = V^T y) \right)$$

- Closed form of the functional available
- A solution to the optimization problem exists
- **A posteriori reduction** (measurement available)

Expected Kullback-Leibler divergence minimization

$$\min_{[V]\in \mathrm{Gr}(r,n)} \mathbb{E}_Y \left(\mathrm{D}_{\mathrm{KL}} \left(P(Q \mid Y) \parallel P(Q \mid W = V^T Y) \right) \right)$$

- Closed form of the functional available
- A solution to the optimization problem exists
- **A priori reduction**

Information-based loss function

Given random variables Z, Z_0, and Z_1,

Entropy

With $Z \sim P(Z)$,

$$\mathrm{H}(Z) = \mathbb{E}_Z(-\log(f_Z(Z))).$$

- Amount of information contained by $P(Z)$

Mutual information

With $Z_0 \sim P(Z_0)$ and $Z_1 \sim P(Z_1)$,

$$\mathcal{I}(Z_0, Z_1) = \mathrm{H}(Z_0) + \mathrm{H}(Z_1) - \mathrm{H}(Z_0, Z_1),$$

- Amount of information that $P(Z_0)$ contains about $P(Z_1)$
- Symmetric quantity

Mutual information maximization

Theorem (Mutual information maximization)

We have

$$\max_{V \in \mathbb{R}_*^{n\times r}} \mathcal{I}(W, Q) = \frac{1}{2} \sum_{i=1}^{r} \log \lambda_i,$$

where $(\lambda_i)_{i=1}^r$ are the r dominant eigenvalues of the problem

$$C_Y v = \lambda C_E v, \quad \lambda \in \mathbb{R}, \; v \in \mathbb{R}^n.$$

A solution to the optimization problem is given by the matrix V with columns being eigenvectors $(v_i)_{i=1}^r$ associated to the eigenvalues $(\lambda_i)_{i=1}^r$. (Error estimator)

Equivalences

The mutual information maximization is equivalent to:

- the maximization of the expected information gain
$$\max_{V \in \mathbb{R}_*^{n\times r}} \mathbb{E}_W \left(\mathrm{D}_{\mathrm{KL}} \left(P(Q|W) \parallel P(Q) \right) \right)$$
- the minimization of the entropy of the posterior distribution
$$\min_{V \in \mathbb{R}_*^{n\times r}} \mathrm{H} \left(P(Q|W = V^T y) \right)$$

Inference problem

Synthetic data

For $(t_i)_{i=1}^{n}$, $n = 500$, a uniformly drawn sample in $(-1, 1)$,

$$Y_{\text{ref}}(t_i) = A_{\text{ref}}(t_i) + E(t_i), \quad \forall i \in \{1, \dots, n\},$$

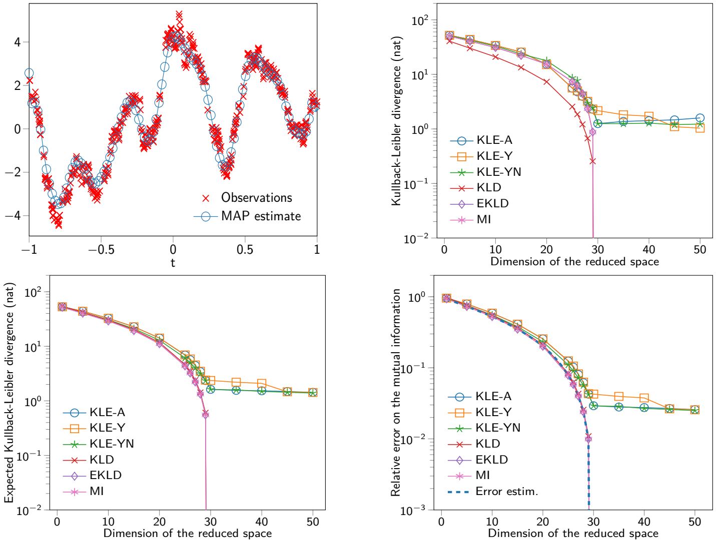
with $A_{\text{ref}} \sim \mathcal{N}(m_{\text{ref}}, C_{\text{ref}})$ and $E \sim \mathcal{N}(m_E, C_E)$.

Model

$$Y_i = \sum_{j=0}^{\mathrm{Nq}-1} T_j(t_i) Q_j + E(t_i), \quad \forall i \in \{1, \dots, n\},$$

with T_j the Chebyshev polynomial of order j and $\mathrm{Nq} = 30$.

Functionals versus the dimension of the reduced space



Inference problem: nonlinear models

Synthetic data

Given two random samples $(s_i)_{i=1}^n$ and $(t_i)_{i=1}^n$ being independent and uniformly distributed in $(-1, 1)$, with $n = 2000$,

$$Y_{\text{ref}}(s_i, t_i) = \exp(F_{\text{ref}}(s_i, t_i)) + E(s_i, t_i), \quad \forall i \in \{1, \dots, n\},$$

where $F_{\text{ref}} \sim \mathcal{N}(0, C_{\text{ref}})$, $E \sim \mathcal{N}(0, C_E)$.

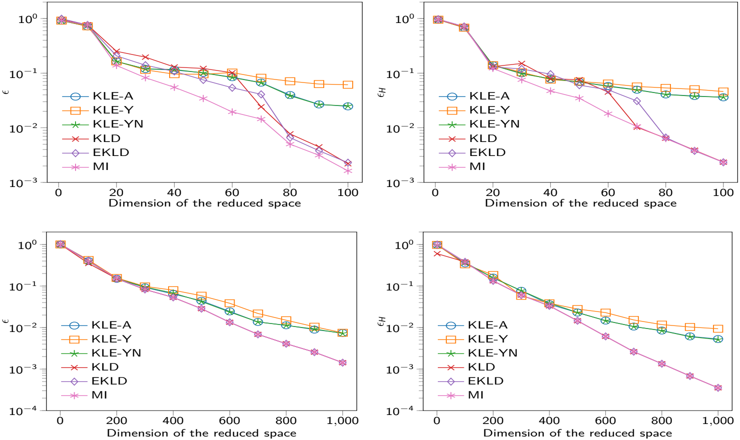
Model

$$Y_i = A_i(Q) + E(s_i, t_i), \quad \forall i \in \{1, \dots, n\},$$

where $A_i(Q) = \exp((BQ)_i)$, $Q \sim \mathcal{N}(0, C_Q)$, and $q = 30$.

- Columns of B: dominant eigenvectors of C_{ref}
- $C_Q = \text{diag}(\lambda_1, \dots, \lambda_q)$: dominant eigenvalues of C_{ref}

Errors versus the dimension of the reduced space $\sigma_{F_{\text{ref}}} = 0.301$ (top), $\sigma_{F_{\text{ref}}} = 1.501$ (bottom)



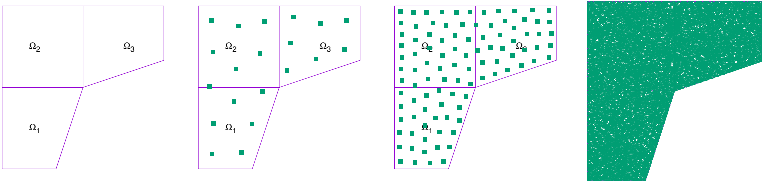
L_2 error on MAP point (left) and Frobenius error on Hessian at MAP.

Inference of conductivities

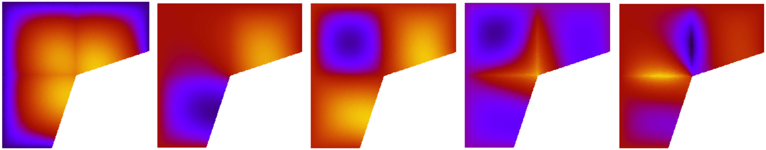
The model:

$$\nabla\left(\kappa(\boldsymbol{x})\nabla U(\boldsymbol{x})\right) = -1, \quad \kappa(\boldsymbol{x} \in \Omega_i) = \kappa_i,$$

where $\log \kappa_i \sim N(0,1)$. Observed at $n = 32,000$ points with Gaussian noise.



Dominant modes of the projection:

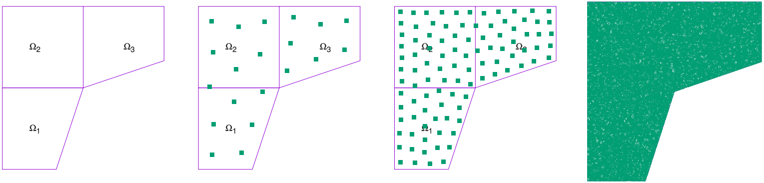


Inference of conductivities

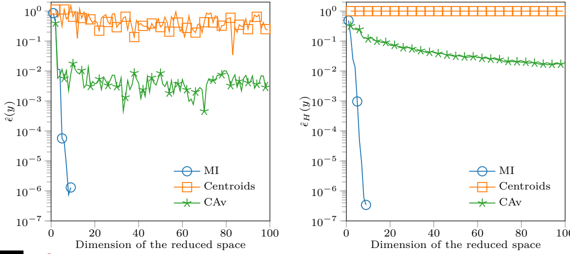
The model:

$$\nabla\left(\kappa(\boldsymbol{x})\nabla U(\boldsymbol{x})\right) = -1, \quad \kappa(\boldsymbol{x} \in \Omega_i) = \kappa_i,$$

where $\log \kappa_i \sim N(0,1)$. Observed at $n = 32,000$ points with Gaussian noise.



Convergence to unreduced MAP and Hessian:



[illegible]vergence with the reduction dimension of the MI, Centroids and Cluster Averages errors on MAP ($\hat{\epsilon}(y)$, left) and Hessian ($\hat{\epsilon}_H(y)$, right). Case of high noise level $\sigma_\epsilon = 0.5$.

Table of contents

Conclusions and Outlooks

Summary

- Reduction approaches are instrumental in UQ and inference
- May concern both the model and the observations
- Reduction strategies should be goal-oriented
- Information theoretic reduction approaches are promising

Outlooks

- Selection of **observation features** for Bayesian inference
- **Goal-oriented** design of model reduction and experiments
- Balancing all uncertainty and prediction error sources

Thank you