Calibration and validation of a computer code

Pierre BARBILLON

AgroParisTech / INRA MIA UMR 518

13/11/13

Journée Validation
GDR MASCOT NUM

Calibration of a computer code

Computer experiments:

Computer model (simulator) $(\mathbf{x}^*, \boldsymbol{\theta}) \mapsto f(\mathbf{x}^*, \boldsymbol{\theta}) \in \mathbb{R}^s$ where

- **physical parameters**: $\mathbf{x}^* \in \mathbb{X} \subset \mathbb{R}^m$ observable and often controllable inputs
- **simulator parameters**: $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ non-observable parameters, required to run the simulator.

  2 types:
  - “calibration parameters”: physical meaning but unknown, necessary to make the code mimic the reality,
  - “tuning parameters”: no physical interpretation.

Goal:
Calibrate the code: finding “best” or “true” $\boldsymbol{\theta}$ from real observations / field data.

Validation

- Validation (rather than verification) is considered,
- Does the computer simulator correspond to field data ?
- The validation of the computer simulator depends on the known or unknown precision of the field data
- Biased computer model, no setting of calibrated parameters leads to outputs close to field data. What is the meaning of validation in that context?
- prediction after the calibration step ?

Outline

1. Context
   - Two kinds of data
   - Meta-modeling / emulator of the computer code
2. Bayesian calibration without discrepancy
   - Known σ^2, unlimited simulator runs
   - Unknown σ^2, unlimited simulator runs
   - Unknown σ^2, limited number of runs
3. Bayesian calibration with discrepancy
   - Calibration with discrepancy
4. Other topics and conclusion

Outline

Plan

Field data

- Field data provided by physical experiments:

$$\mathbf{y}^F = y^F(\mathbf{x}_1), \ldots, y^F(\mathbf{x}_n)\,,$$

- n is small, $\mathbf{x}_1, \ldots \mathbf{x}_n \in \mathbb{X}$ hard to set, sometimes uncontrollable, included in a small domain...

- Model:

$$y^F(\mathbf{x}_i) = \zeta(\mathbf{x}_i) + \epsilon(\mathbf{x}_i)\,,$$

  where
  - $\zeta(\cdot)$ real physical process (unknown),
  - $\epsilon(\mathbf{x}_i)$ often assumed i.i.d. $\mathcal{N}(0, \sigma^2)$,
  - σ^2 sometimes treated as known...

Computer model / simulator

$$(\mathbf{x}^*, \boldsymbol{\theta}) \mapsto f(\mathbf{x}^*, \boldsymbol{\theta}) \in \mathbb{R}^s$$

- **physical parameters**: $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^m$,
  - $\mathbf{x}^*$ same meaning as in field data,
  - extrapolation if $\mathbf{x}^* > \max(\mathbf{x}_i)$ or $\mathbf{x}^* < \min(\mathbf{x}_i)$.
- **simulator parameters** $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ non-observable parameters, required to run the simulator. No difference here between calibration and tuning.

The simulator is often an **expensive black-box function**.
$\Rightarrow$ limited number N_{run} of runs of the simulator.

Relationship between the simulator and the data

for $i = 1, \ldots, n$,

- if the simulator sufficiently represents the physical system:

$$y_i^F = f(\mathbf{x}_i, \boldsymbol{\theta}^*) + \epsilon(\mathbf{x}_i)\,,$$

  i.e. for the unknown value $\boldsymbol{\theta} = \boldsymbol{\theta}^*$: $f(\mathbf{x}, \theta^*) = \zeta(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{X}$,

- if the field observations are inconsistent with the simulations (irreducible model discrepancy):

$$y_i^F = f(\mathbf{x}_i, \boldsymbol{\theta}^*) + \delta(\mathbf{x}_i) + \epsilon(\mathbf{x}_i)\,.$$

  $\delta(\cdot)$ models the difference between the simulator and the physical system:

$$\delta(\mathbf{x}) = \zeta(\mathbf{x}) - f(\mathbf{x}, \theta^*)\,,$$

  but

  - What does θ^* mean ?
  - A best fitting ?
  - identifiability issues ?
  - usually assumed to be smoother than the real physical process $\zeta(\cdot)$

Ref.: Kennedy and O'Hagan (2001), Hidgon et al. (2005)...

Plan

1. Context
   - Two kinds of data
   - Meta-modeling / emulator of the computer code
2. Bayesian calibration without discrepancy
   - Known σ^2, unlimited simulator runs
   - Unknown σ^2, unlimited simulator runs
   - Unknown σ^2, limited number of runs
3. Bayesian calibration with discrepancy
   - Calibration with discrepancy
4. Other topics and conclusion

Expensive black-box computer code

- Run the simulator for a given $(\mathbf{x}^*, \boldsymbol{\theta})$ is time-consuming / expensive.
- The simulator is a black-box, no intrusive methods are possible.

$\Rightarrow$ Only few runs of the simulator are possible then we cannot apply algorithms (as in Bayesian calibration) which make a massive use of simulator runs.

Using an emulator / metamodel / coarse model / approximation of the simulator which is fast to compute, but:

- loss on precision of prediction,
- new uncertainty source: accuracy of the model approximation,
- taken into account.

Choosing a design of experiments

Choose N_{run} couples

$$(\mathbf{x}_j^*, \theta_j)$$

- space filling for x,
- with respect to the prior distribution on θ,
- $\mathbf{x}_j^* = \mathbf{x}_i$?

where the simulator is called.

Emulator using Gaussian Process:

- Very popular in computer experiments.
- integrated in a Bayesian framework: appears in the likelihood function and a prior on the parameters of the Gaussian process are chosen.
- model uncertainty coming from approximation of f.
- After the calibration step, used in prediction for a new point $\mathbf{x}$.

Meta-modeling: prior distribution on f

Sacks et al. (1989).
f realization of a Gaussian process F:
$\forall(\mathbf{x}^*, \boldsymbol{\theta}) \in E$,

$$F((\mathbf{x}^*, \boldsymbol{\theta})) = \sum_{k=1}^{Q} \beta_k h_k((\mathbf{x}^*, \boldsymbol{\theta})) + Z((\mathbf{x}^*, \boldsymbol{\theta})) = H((\mathbf{x}^*, \boldsymbol{\theta}))^T \boldsymbol{\beta} + Z((\mathbf{x}^*, \boldsymbol{\theta})),$$

où

- $h_1, \ldots, h_Q$ regression functions and $\boldsymbol{\beta}$ parameters vector,
- Z centered Gaussians process with covariance function:

$$\mathrm{Cov}(Z((\mathbf{x}_1^*, \boldsymbol{\theta}_1)), Z((\mathbf{x}_2^*, \boldsymbol{\theta}_2))) = \sigma^2 K((\mathbf{x}_1^*, \boldsymbol{\theta}_1), (\mathbf{x}_2^*, \boldsymbol{\theta}_2)),$$

where K is correlation kernel.

Hypotheses

- $K((\mathbf{x}_1^*, \boldsymbol{\theta}_1), (\mathbf{x}_2^*, \boldsymbol{\theta}_2)) = \sigma_K^2 \exp(-\xi_{\mathbf{x}^*} \sum |\mathbf{x}_1^* - \mathbf{x}_2^*|^\alpha - \xi_{\boldsymbol{\theta}} \sum |\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|^\alpha)$
- parameters $\phi = (\boldsymbol{\beta}, \sigma^2, K \text{ parameters})$ assumed fixed (in practice, maximum likelihood estimators);

Meta-modeling: posterior

- $v_1 = f((\mathbf{x}^*, \boldsymbol{\theta})_1), \ldots, v_{N_{run}} = f((\mathbf{x}^*, \boldsymbol{\theta})_{N_{run}})$ evaluations of f on a design $D_{N_{run}}$
- Process $F^{D_{N_{run}}}$: Conditioning F to $F((\mathbf{x}_1^*, \boldsymbol{\theta}_1)) = v_1, \ldots, F(\mathbf{x}_{N_{run}}^*, \boldsymbol{\theta}_{N_{run}})) = v_{N_{run}}$.
  Gaussian Process with mean $m((\mathbf{x}^*, \boldsymbol{\theta}))$ and covariance $C((\mathbf{x}^*, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta})')\ \forall (\mathbf{x}^*, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta})'$.

  For all $(\mathbf{x}^*, \boldsymbol{\theta}) \in E$,
  - $m((\mathbf{x}^*, \boldsymbol{\theta}))$ approximates $f((\mathbf{x}^*, \boldsymbol{\theta}))$,
  - $C((\mathbf{x}^*, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}))$ uncertainty on this approximation.

For all $(\mathbf{x}_i^*, \boldsymbol{\theta}_i) \in D_{N_{run}}$,
- $m(\mathbf{x}_i^*, \boldsymbol{\theta}_i) = f(\mathbf{x}_i^*, \boldsymbol{\theta}_i)$,
- $C((\mathbf{x}_i^*, \boldsymbol{\theta}_i), (\mathbf{x}_i^*, \boldsymbol{\theta}_i)) = 0$.

Gaussian process emulator: illustration

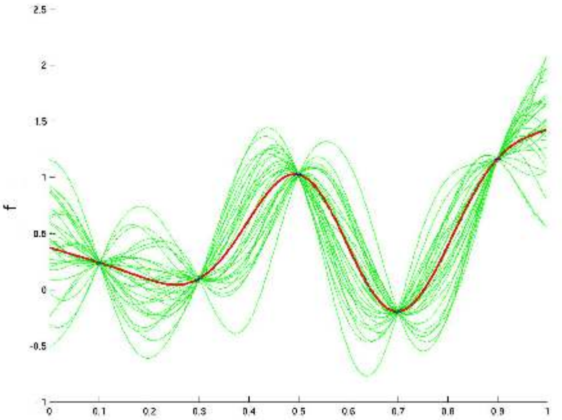


Figure: Posterior mean and realisations of the conditioned process

Outline

Plan

A calibration example

Hypotheses:

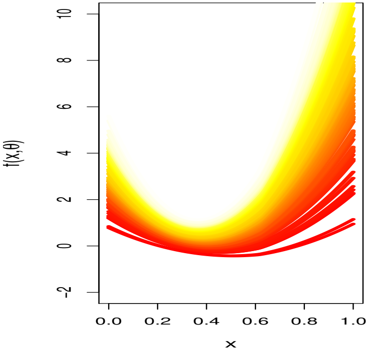
- The simulator represents sufficiently well the physical system:

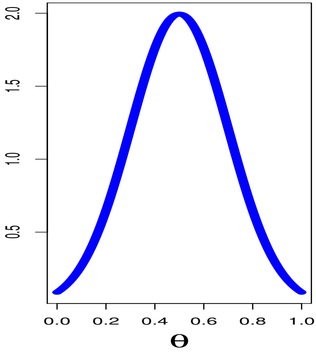
$$y(\mathbf{x}_i) = f(\mathbf{x}_i, \boldsymbol{\theta}^*) + \epsilon_i, \quad i = 1, \dots, n.$$

- But unknown $\boldsymbol{\theta}^*$.
- $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. with known σ^2.
- $\sigma^2 = 0.3$
- $n = 6$,
- $\theta^* = 0.6$

A calibration example

Prior:
prior distribution on unknown $\boldsymbol{\theta}$: $\pi(\cdot)$
from expert judgment, past experiments...
Possible choice $\pi(\boldsymbol{\theta}) = \mathcal{N}(\theta_0, \sigma_0^2) = \mathcal{N}(0.5, 0.04)$.





A calibration example

Data:
Couples $(\mathbf{x}_1, y_1^F), \ldots, (\mathbf{x}_n, y_n^F)$ from physical experiments.

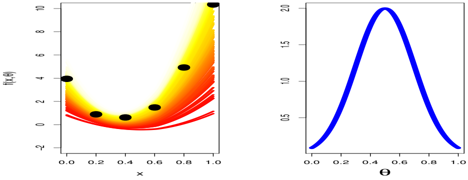
Posterior distribution:

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{y}^F) &\propto l(\boldsymbol{\theta}|\mathbf{y}^F)\cdot\pi(\boldsymbol{\theta})\\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y(\mathbf{x}_i)-f(\mathbf{x}_i,\boldsymbol{\theta}))^2-\frac{1}{2\sigma_0^2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^2\right)\end{aligned}$$

- Analytical posterior if $\boldsymbol{\theta}\mapsto f(\mathbf{x},\boldsymbol{\theta})$ is a linear map,
- Otherwise MH sampling to simulate according to the posterior distribution.

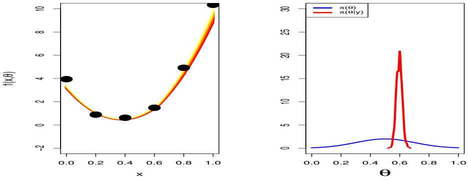
A calibration example

Prior with data:



$\Downarrow$ Metropolis-Hastings algorithm $\Downarrow$

Posterior on θ:



More details on the MH algorithm

Initialisation:
θ^0 chosen.

Update:
iterations $t = 1, \ldots,$

1. Proposal: $\tilde{\theta}^{t+1} = \theta^t + \mathcal{N}(0, \tau^2)$.
2. Compute
$$\alpha(\theta^t, \tilde{\theta}^{t+1}) = \frac{\pi(\tilde{\theta}^{t+1}|\mathbf{y}^F)}{\pi(\theta^t|\mathbf{y}^F)}$$
3. Acceptation:
$$\theta^{t+1} = \begin{cases} \tilde{\theta}^{t+1} & \text{with probability } \alpha(\theta^t, \tilde{\theta}^{t+1}) \\ \theta^t & \text{otherwise.} \end{cases}$$

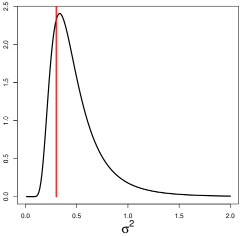
Note that the ratio $\alpha(\theta^t, \tilde{\theta}^{t+1})$ needs several computations of $f(\mathbf{x}, \boldsymbol{\theta})$ at each step since

$$\pi(\theta|\mathbf{y}^F) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y(\mathbf{x}_i) - f(\mathbf{x}_i, \boldsymbol{\theta}))^2 - \frac{1}{2\sigma_0^2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^2\right).$$

Plan

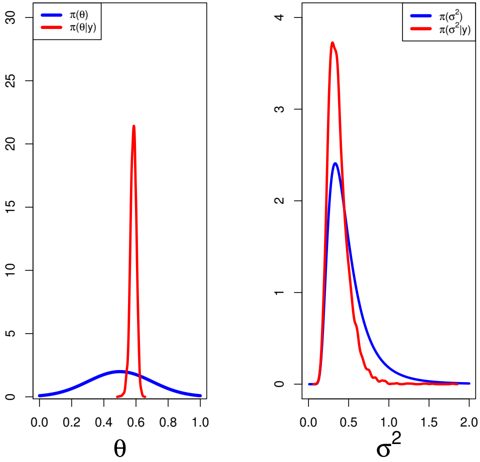
Unknown σ^2

- prior distribution on σ^2: $\pi(\sigma^2) = \mathcal{IG}(5, 2)$



- Gibbs algorithm to simulate couples $(\boldsymbol{\theta}, \sigma^2)$ from $\pi(\boldsymbol{\theta}, \sigma^2|\mathbf{y}^F)$. Iterate :
  1. MH algorithm to simulate $\boldsymbol{\theta}_t$ from $\pi(\cdot|\mathbf{y}^F, \sigma^2_{t-1})$,
  2. conditional simulation of σ^2_t from $\pi(\cdot|\mathbf{y}^F, \boldsymbol{\theta}_t)$.

Posterior distributions



Comparison

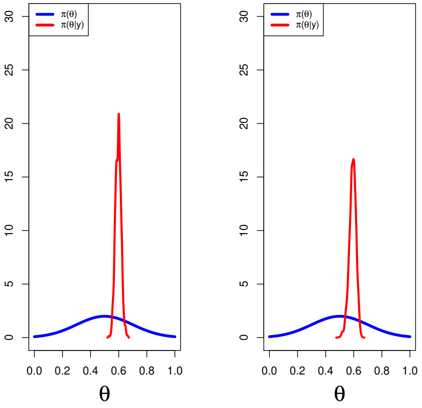


Figure: known σ^2 vs unknown σ^2

with a bad prior....

prior on θ: $\pi(\theta) = \mathcal{N}(0.2, 0.04)$ and $n = 12$ field data

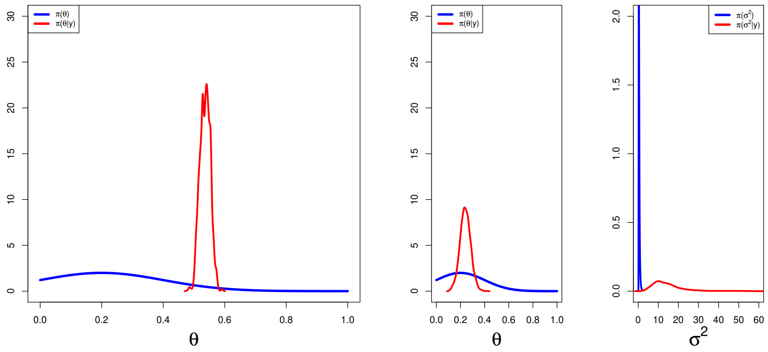


Figure: known σ^2 vs unknown σ^2

Plan

Likelihood with a Gaussian process hypothesis on f

- $\mathbf{z} = (\mathbf{y}_1^F, \ldots, \mathbf{y}_n^F, f(\mathbf{x}_1^*, \boldsymbol{\theta}_1), \ldots, f(\mathbf{x}_{N_{run}}^*, \boldsymbol{\theta}_{N_{run}}))$
- likelihood on $\mathbf{z}$

$$l(\boldsymbol{\theta}, \sigma^2 | \mathbf{z}) \propto |\Sigma_{\mathbf{z}}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{z} - \mu)^T \Sigma_{\mathbf{z}}^{-1} (\mathbf{z} - \mu)\right)$$

where

- μ is the mean of the Gaussian process,
- $$\Sigma_{\mathbf{z}} = \Sigma_f + \begin{pmatrix} \Sigma_y & 0 \\ 0 & 0 \end{pmatrix}$$

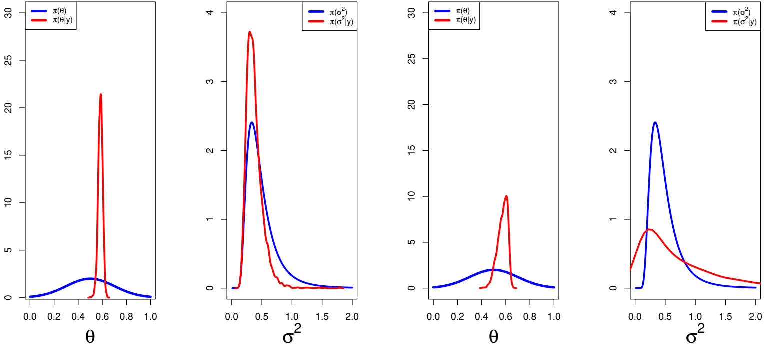
  with $\Sigma_y = \sigma^2 I_n$ and Σ_f is obtained as the covariance matrix corresponding to the points: $(\mathbf{x}_1, \boldsymbol{\theta}), \ldots, (\mathbf{x}_n, \boldsymbol{\theta}), (\mathbf{x}_1^*, \boldsymbol{\theta}_1), \ldots, (\mathbf{x}_{N_{run}}^*, \boldsymbol{\theta}_{N_{run}})$.

Dealing with GP parameters

- prior distribution on μ and covariance parameters Hidgon et al. (2005) $\Rightarrow$ MCMC inference
- MLE estimators Kennedy and O'Hagan (2001)
  - treated as fixed,
  - only computer data $f(\mathbf{x}_1^*, \boldsymbol{\theta}_1), \ldots, f(\mathbf{x}_{N_{run}}^*, \boldsymbol{\theta}_{N_{run}})$ are used ($n < N_{run}$) for MLE
  - likelihood $l(\boldsymbol{\theta}, \sigma^2|\mathbf{z})$:

$$l(\boldsymbol{\theta}, \sigma^2|\mathbf{z}) \propto |\tilde{\Sigma}_{\mathbf{y}^F}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y}^F - m(\mathbf{x}, \boldsymbol{\theta}))^T \tilde{\Sigma}_{\mathbf{y}^F}^{-1}(\mathbf{y}^F - m(\mathbf{x}, \boldsymbol{\theta}))\right)$$

  where
    - $m(\cdot)$ is the mean of the GP conditioned to simulator data,
    - $\tilde{\Sigma}_{\mathbf{y}^F} = \Sigma_{\mathbf{y}^F} + \tilde{\Sigma}_f = \sigma^2 I_n + \tilde{\Sigma}_f$ where $\tilde{\Sigma}_f$ is constructed with the covariance function C of the conditioned GP.



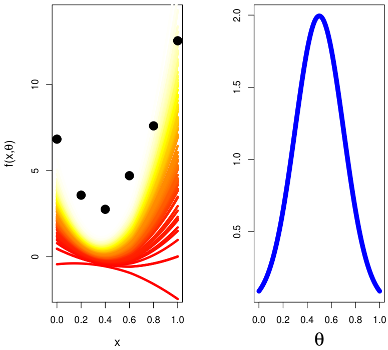
unlimited runs versus $N_{run} = 12$

Outline

Plan

Model discrepancy

$$y_i^F = f(\mathbf{x}_i, \boldsymbol{\theta}^*) + \delta(\mathbf{x}_i) + \epsilon(\mathbf{x}_i)\,.$$



No value of θ makes the simulator corresponding to the fied data

Modelisation of δ:

Sensible to assume: $\delta(\mathbf{x}) \approx \delta(\mathbf{x} + d\mathbf{x})$

Gaussian Process hypothesis on δ with possible:

- zero mean,
- smooth a priori on covariance function,
- combining with Gaussian process hypothesis on f.

Meaning of θ:

- few information on $\boldsymbol{\theta}$ if there is a systematic discrepancy ?
- the model $f(\mathbf{x}, \boldsymbol{\theta})$ is informative through $\boldsymbol{\theta}$ on the shape of the physical phenomenon $\zeta(\cdot)$?

Prior specification on δ

- $\mathbb{E}(\delta(\cdot)) = 0$,
- Covariance function:

$$K_\delta(\mathbf{x}, \mathbf{x}') = \sigma_\delta^2 \exp\left(-\xi_\delta \|\mathbf{x} - \mathbf{x}'\|^2\right)$$

- $\pi(\sigma^2) = \mathcal{IG}(3, 1)$
- $\pi(\xi_\delta) \propto (1 - \exp(-\xi_\delta))^{-0.6} \exp(-\xi_\delta)$

- Kennedy and O'Hagan (2001)proposed a Gaussian approximation of $\pi(\mathbf{y}^F|\xi_\delta, \sigma_\delta^2)$ to use ML estimators for $\xi_\delta, \sigma_\delta^2$.

Likelihood

$$l(\boldsymbol{\theta}, \sigma_\delta^2, \xi_\delta | \mathbf{y}^F) \propto |\tilde{\Sigma}_{\mathbf{y}^F}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y}^F - m(\mathbf{x}, \boldsymbol{\theta}))^T \tilde{\Sigma}_{\mathbf{y}^F}^{-1} (\mathbf{y}^F - m(\mathbf{x}, \boldsymbol{\theta}))\right),$$

where

- $m(\cdot)$ is the mean of the GP conditioned to simulator data,
- $\tilde{\Sigma}_{\mathbf{y}^F} = \sigma^2 I_n + \tilde{\Sigma}_f + \Sigma_\delta$ where $\tilde{\Sigma}_f$ is constructed with the covariance function C of the conditioned GP on f and Σ_δ is constructed with the covariance function of the GP on δ.

Gibbs algorithm

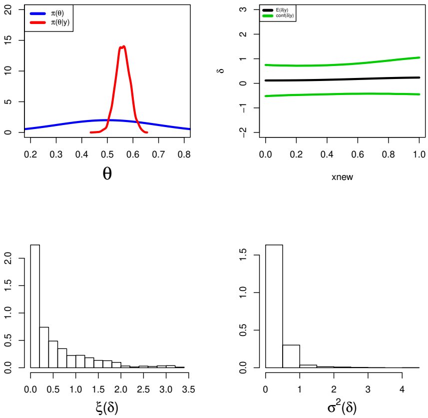
Iterate :

1. MH algorithm to simulate $\boldsymbol{\theta}_t$ from $\pi(\cdot|\mathbf{y}^F, \xi_{\delta,t-1}, \sigma^2_{\delta,t-1})$,
2. MH algorithm to simulate $\xi_{\delta,t-1}$ from $\pi(\cdot|\mathbf{y}^F, \boldsymbol{\theta}_t, \sigma^2_{\delta,t-1})$,
3. MH algorithm to simulate $\sigma^2_{\delta,t-1}$ from $\pi(\cdot|\mathbf{y}^F, \boldsymbol{\theta}_t, \xi_{\delta,t})$.

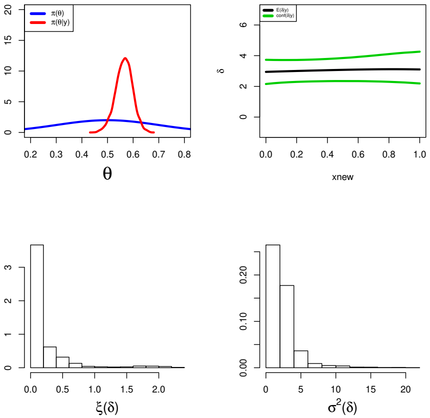
An example

- $n = 6$,
- $N_{run} = 12$,
- σ^2 assumed known,
- different bias :
  1. $\delta(\mathbf{x}) = 0$
  2. $\delta(\mathbf{x}) = 3$
  3. $\delta(\mathbf{x}) = 2 - \mathbf{x}$

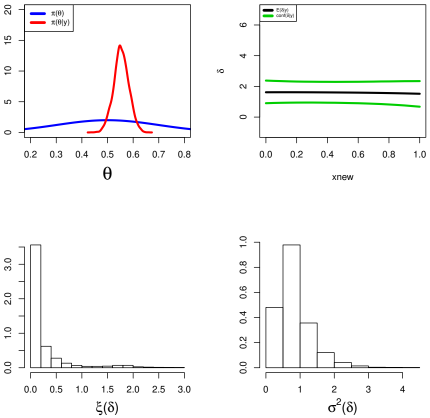
$\delta(\mathbf{x}) = 0$



$\delta(\mathbf{x}) = 3$



$\delta(\mathbf{x}) = 2 - \mathbf{x}$



Remarks

- same difficulties with bad prior,
- validation if the bias can be considered flat and equal to 0 ?
- difficulties to identify a non constant bias...
- not tested with unknown σ^2

Some considerations on bias

- Brynjarsdóttir and O'Hagan (2013)advocated for taken into account a constraint form for the bias.
- Bachoc et al. proposed a validation method where the calibration makes use of a linearisation of the simulator.

Outline

Prediction with a calibrated simulator:

Once the model is calibrated:

Posterior distribution on $\boldsymbol{\theta}$: $\pi(\cdot|\mathbf{y}^F, \ldots)$

Prediction of the physical phenomenon $\zeta(\cdot)$, for $\mathbf{x}^{new}$?

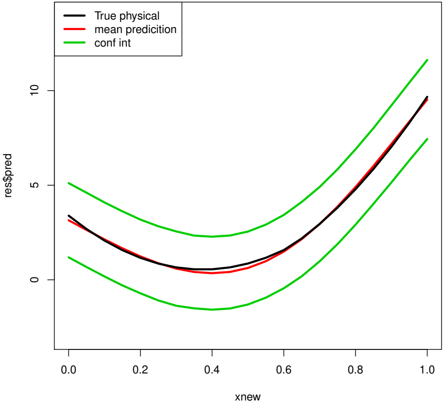
- If no discrepancy, no emulator, $\zeta(\mathbf{x}^{new})$ can be estimated through

$$\hat{\zeta}(\mathbf{x}^{new}) = \int_{\Theta} f(\mathbf{x}^{new}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}^F)d\boldsymbol{\theta}\,.$$

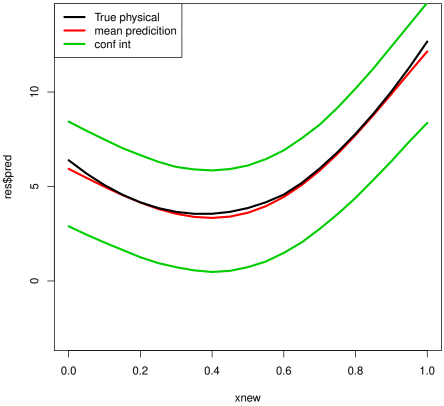
- otherwise $\zeta(\mathbf{x}^{new})$ has a Gaussian process as posterior distribution with mean and covariance depending on $\boldsymbol{\theta}$.
  $\Rightarrow$ combining this distribution with $\pi(\cdot|\mathbf{y}^F, (f(\mathbf{x}_j^*, \boldsymbol{\theta}_j))_j)$
  integration of the posterior mean of $\zeta(\mathbf{x}^{new})$:

$$\int_{\Theta} \mathbb{E}(\zeta(\mathbf{x}^{new})|\mathbf{y}^F, (f(\mathbf{x}_j^*, \boldsymbol{\theta}_j))_j, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}^F, (f(\mathbf{x}_j^*, \boldsymbol{\theta}_j))_j)\,.$$

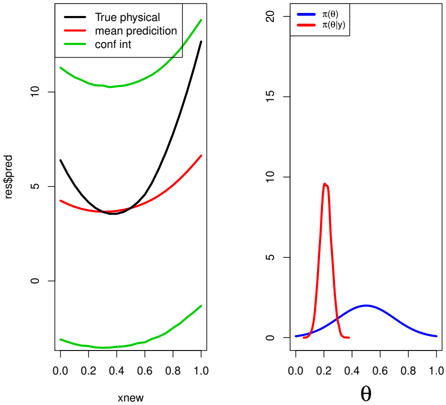
Prediction without discrepancy, with emulator



Prediction with discrepancy, with emulator



Prediction with discrepancy, with emulator, bad prior



Concerns and questions

Identifiability concerns

- If there is discrepancy, very little information on $\boldsymbol{\theta}$ and meaning of “best” or “true” $\boldsymbol{\theta}$?
- If measurement error distribution ($\epsilon_i \sim \mathcal{N}(0, \sigma^2)$) unknown $\Rightarrow$ lack of identifiability.
- Prediction can be accurate in a non-identifiable model...

Validation ?

- Validate with unknown σ^2 ?
- Validate with model discrepancy ?
- Incorporate a bias and validate if the bias can be assumed identically null.
- Discrepancy between prior on **calibration** parameters and posterior.

MCMC issues

- Gibbs on a potentially big number of parameters,
- each MH chain has to be tune.

References

Calibration of computer models

- Dave Hidgon et al., 2005. Combining Field Data and Computer Simulations for Calibration and Prediction. SIAM 26(2).
- Marc Kennedy and Anthony O'Hagan, 2001. Bayesian Calibration of Computer Models. Journal of the Royal Statistical Society B 68.
- Jenny Brynjarsdóttir and Anthony O'Hagan, 2013. J. of Uncertainty Quantification.

Validation of computer models

- Susie Bayarri et al., 2002. A Framework for Validation of Computer Models. Technometrics 49(2).
- François Bachoc et al., 2013.Gaussian process computer model validation method

Gaussian Process emulator:

- Thomas Santner et al., 2003. The Design and Analysis of Computer Experiments. Springer-Verlag.
- Kai-Tai Fang et al., 2006. Design and Modeling for Computer Experiments. Computer Science and Data Analysis. Chapman & Hall/CRC.