Simultaneous calibration of a computer model and screening of its discrepancy function

Pierre Barbillon[1], Anabel Forte[2] and Rui Paulo[3]

[1] AgroParisTech/INRA, [2] Universitat de Valencia, [3] CEMAPRE/REM and ISEG Universidade de Lisboa

Workshop on calibration of numerical code
May 31st 2023 — Institut Henri Poincaré

Introduction

Computer model

Let $f(\boldsymbol{x}, \boldsymbol{\theta})$ denote the output of a real-valued, deterministic function, which implements a mathematical model aimed at reproducing a real phenomenon

- $\boldsymbol{x} = (x_1 \ldots, x_p)^\top$ are input variables describing controllable or observable aspects of the system (environmental variables)
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$ are model parameters which are unknown in the context of physical experiments

Example: Photovoltaic plant

12 photovoltaic panels connected together

$f(\boldsymbol{x}, \boldsymbol{\theta})$ is the instantaneous power delivered by the plant, where

- $\boldsymbol{x} = (t, I_g, Id, T_e)^\top$: t is the time since the beginning of the year, I_g is the global irradiation of the sun, Id is the diffuse irradiation of the sun, and T_e is the ambient temperature.
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6)^\top$ but only one is treated as unknown, the module photo-conversion efficiency. A sensitivity analysis has proven the other parameters to be of negligible importance.
- Carmassi et al. (2019)

Introduction

Field experiments

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ the configurations at which the field experiments are conducted; that is,

$$\boldsymbol{x}_i = (x_{1,i}, \ldots, x_{p,i})^\top$$

denotes the values of the input variables that have been set for the *i*th experiment (or that will be observed as part of that experiment, if corresponding to environmental variables)

Following Craig et al. (1996, 1997), Craig et al. (2001), but most notably Kennedy and O'Hagan (2001), model the field data as

$$y(\boldsymbol{x}_i) = f(\boldsymbol{x}_i, \boldsymbol{\theta}) + \delta(\boldsymbol{x}_i) + \varepsilon_i$$

Introduction

Accounting for various sources of uncertainty

$$y(\boldsymbol{x}_i) = f(\boldsymbol{x}_i, \boldsymbol{\theta}) + \delta(\boldsymbol{x}_i) + \varepsilon_i$$

- ε_i are independent $N(0, \sigma_0^2)$ random variables which represent measurement error
- $\boldsymbol{\theta}$ denotes the "true" but unknown value of the vector of model parameters
- $\delta(\boldsymbol{x}_i)$ denotes the **discrepancy function** and is meant to account for model inadequacy

Introduction

Uncertainty quantification

Bayarri et al. (2007), Higdon et al. (2004)

- Emulation: construction of a fast approximation for $f(\cdot,\cdot)$
- Calibration: estimation of $\boldsymbol{\theta}$
- Validation: how does $f(\cdot,\cdot)$ fare as a representation of the real phenomenon?
- Prediction, both inter- and extrapolation

Introduction

Our Problem:

- In most applications, the interest is not in $\delta(\cdot)$ itself
- By incorporating $\delta(\cdot)$ in the statistical model, one hopes for more meaningful calibration and to improve prediction — bias-corrected prediction
- **Our goal** is to ascertain which are the input variables that can be labeled as **active** in $\delta(\cdot)$
    - Those inputs are being mishandled in the computer model — need further attention
    - It's not recommend to extrapolate along those inputs
    - **Screening the discrepancy function**

Statistical framework

Gaussian process prior

We place a Gaussian process prior on $\delta(\cdot)$:

$$\delta(\cdot) \mid \sigma^2, \boldsymbol{\psi} \sim GP(0, \sigma^2 c(\cdot, \cdot \mid \boldsymbol{\psi}))$$

where

$$c(\boldsymbol{x}_i, \boldsymbol{x}_j) = \prod_{\ell=1}^{p} c(x_{\ell i}, x_{\ell j} \mid \psi_\ell)$$

with $\psi_\ell > 0$ being a range parameter.

We will use the power exponential correlation function:

$$c(x_{\ell i}, x_{\ell j} \mid \psi_\ell) = \exp\left(-|x_{\ell i} - x_{\ell j}|^a / \psi_\ell\right)$$

with $0 < a \leq 2$ fixed.

Statistical framework

Remarks

- There are known confounding issues between $\delta(\cdot)$ and $\boldsymbol{\theta}$ (e.g. Tuo and Wu, 2015)
- Brynjarsdóttir and O'Hagan (2014) (and others before) show how incorporating meaningful prior information on δ may be important
- Plumlee (2017) and Gu and Wang (2018) place more sophisticated priors on δ to ensure the separation between δ and $\boldsymbol{\theta}$
- **Important:** in what follows, we assume that $f(\cdot, \cdot)$ is fast to compute, although the methodology applies also to the case where we need to construct a surrogate model

Statistical framework

- $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are the configurations at which the field experiments are conducted
- $\boldsymbol{y}^\top = (y_1, \ldots, y_n)$, $y_i = y(\boldsymbol{x}_i)$
- $\boldsymbol{f}(\boldsymbol{\theta}) = (f(\boldsymbol{x}_i, \boldsymbol{\theta}),\ i = 1, \ldots, n)^\top$
- $\boldsymbol{R} = [c(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{\psi})]_{i,j=1,\ldots,n}$

$$\boldsymbol{y} \mid \boldsymbol{\psi}, \sigma^2, \sigma_0^2, \boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{\theta}) \sim N_n(\boldsymbol{f}(\boldsymbol{\theta}), \sigma^2 \boldsymbol{R} + \sigma_0^2\ \boldsymbol{I}_n)$$

Statistical framework

- Given the separable structure of the correlation function, as $\psi_\ell \to +\infty$ the effect of x_ℓ on $\boldsymbol{R}$ vanishes
- A possible solution: place priors on all unknowns and deem as inactive all x_ℓ that have "large" ψ_ℓ
- But how large is large enough? Always the case with estimation-based approaches to variable selection.

Linkletter's reparametrization

Linkletter et al. (2006) introduced the following reparametrization of the power exponential to address variable selection of a computer model:

$$\rho_\ell = \exp(-(1/2)^a/\psi_\ell)$$

which produces

$$c(x_{\ell i}, x_{\ell j} \mid \rho_\ell) = \rho_\ell^{2^a |x_{\ell i} - x_{\ell j}|^a}$$

with a fixed at some value in the range of $(0, 2]$.

Advantages:

- $0 \le \rho_\ell \le 1$
- x_ℓ is inert if $\rho_\ell = 1$

Our approach — model selection

Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top$ index all the 2^p models for $\delta(\cdot)$ that result from all possible subsets of $\{x_1, \ldots, x_p\}$ being active:

$$\gamma_\ell = \begin{cases} 1, & \text{if } x_\ell \text{ is active} \\ 0, & \text{if } x_\ell \text{ is inert} \end{cases}$$

Under model $\mathcal{M}_\gamma$,

$$\boldsymbol{y} \mid \boldsymbol{\rho}, \sigma^2, \sigma_0^2, \boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{\theta}) \sim N_n(\boldsymbol{f}(\boldsymbol{\theta}), \sigma^2 \boldsymbol{R}_\gamma + \sigma_0^2 \ \boldsymbol{I}_n)$$

with

$$\boldsymbol{R}_\gamma = \left[\prod_{\ell:\gamma_\ell=1} c(x_{\ell i}, x_{\ell j} \mid \rho_\ell) \right]_{i,j=1,\ldots,n}$$

that is,

$$\rho_\ell = 1 \Leftrightarrow \gamma_\ell = 0$$

Posterior model probabilities

A natural way to quantify model uncertainty is through the posterior model probabilities

$$\pi(\boldsymbol{\gamma} \mid \boldsymbol{y}) \propto m(\boldsymbol{y} \mid \gamma)\ \pi(\gamma)$$

where $\pi(\boldsymbol{\gamma}) = \mathbb{P}(\mathcal{M}_\gamma)$ and $\pi(\boldsymbol{\gamma} \mid \boldsymbol{y}) = \mathbb{P}(\mathcal{M}_\gamma \mid \boldsymbol{y})$ and

$$m(\boldsymbol{y} \mid \gamma) = \int N(\boldsymbol{y} \mid \boldsymbol{f}(\boldsymbol{\theta}), \sigma^2\ \boldsymbol{R}_\gamma + \sigma_0^2\ \boldsymbol{I}_n) \\ \pi(\sigma^2, \sigma_0^2, \boldsymbol{\rho} \mid \boldsymbol{\gamma})\ \pi(\boldsymbol{\theta})\ d\sigma^2\ d\sigma_0^2\ d\boldsymbol{\rho}\ d\boldsymbol{\theta}\ .$$

with

- $\pi(\boldsymbol{\theta})$ specified using expert information
- $\pi(\sigma^2, \sigma_0^2, \boldsymbol{\rho} \mid \boldsymbol{\gamma}) = \pi(\sigma^2, \sigma_0^2)\ \pi(\boldsymbol{\rho} \mid \boldsymbol{\gamma})$

PIPS

Once $\pi(\boldsymbol{\gamma} \mid \boldsymbol{y})$ is computed for all γ, we can obtain the posterior inclusion probabilities of each input x_ℓ:

$$\pi(x_\ell \mid \boldsymbol{y}) = \sum_{\boldsymbol{\gamma}:\ \gamma_\ell = 1} \pi(\boldsymbol{\gamma} \mid \boldsymbol{y})$$

or even of pairs of inputs:

$$\pi(x_\ell \vee x_j \mid \boldsymbol{y}) = \pi(x_\ell \mid \boldsymbol{y}) + \pi(x_j \mid \boldsymbol{y}) - \sum_{\gamma:\, \gamma_\ell = 1, \gamma_j = 1} \pi(\gamma \mid \boldsymbol{y})$$

These quantities are central to our proposal: **posterior inclusion probability screening**.

Existing methodology

Savitsky et al. (2011) extends Linkletter et al. (2006) by proposes writing

$$\pi(\boldsymbol{\rho} \mid \boldsymbol{\gamma}) = \prod_{\ell=1}^{p} \left[\gamma_\ell \; I_{(0,1)}(\rho_\ell) + (1 - \gamma_\ell) \; \mathrm{Dir}_1(\rho_\ell)\right]$$

with Dir_1 representing the Dirac delta at 1.

(Discrete) spike and slab prior of Bayesian variable selection (Mitchell and Beauchamp, 1988):

> *if a variable is present in the model, the prior on ρ is the 'slab', a $U(0,1)$ here; otherwise it's a 'spike', a point mass at 1.*

Existing methodology

Additionally

$$\pi(\boldsymbol{\gamma}) = \prod_{\ell=1}^{p} {\tau_\ell}^{\gamma_\ell}(1-\tau_\ell)^{1-\gamma_\ell} \ ,$$

where τ_ℓ is a fixed number representing the prior probability that x_ℓ is active.

Fairly sophisticated MCMC schemes to sample from the posterior distribution of $(\boldsymbol{\rho}, \sigma^2, \sigma_0^2, \boldsymbol{\gamma})$. The selection of variables is made by inspecting the posterior on $(\boldsymbol{\rho}, \boldsymbol{\gamma})$.

Existing methodology

Linkletter et al. (2006): set $\tau_\ell = \tau$ and integrate out $\boldsymbol{\gamma}$ from $\pi(\boldsymbol{\rho}, \boldsymbol{\gamma}) = \pi(\boldsymbol{\rho} \mid \boldsymbol{\gamma})\ \pi(\boldsymbol{\gamma})$, resulting in

$$\pi(\boldsymbol{\rho}) = \prod_{\ell=1}^{p} \left[\tau I_{[0,1]}(\rho_\ell) + (1-\tau)\mathrm{Dir}_1(\rho_\ell)\right] \ .$$

Model indicator $\boldsymbol{\gamma}$ is no longer available so how to declare a variable inert? We revert back to the estimation-based approach!

Existing methodology

Reference distribution variable selection: for a large number of times, say $T = 100$

- add a fictitious input x_{new} to the correlation kernel (along with ρ_{new}) and to the design set
- obtain the posterior distribution of $(\boldsymbol{\rho}, \rho_{\text{new}})$, record the posterior median of ρ_{new}

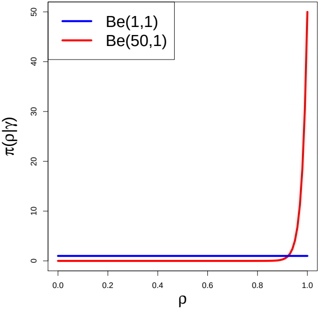
input x_ℓ if inert if the posterior median of ρ_ℓ exceeds a fixed lower percentile (say, the 10%) of the distribution of the posterior median of ρ_{new}.

Our approach

Continuous spike and slab (George and McCulloch, 1993)

$$\pi(\boldsymbol{\rho} \mid \boldsymbol{\gamma}) = \prod_{\ell=1}^{p} \left[\gamma_\ell \; I_{(0,1)}(\rho_\ell) + (1-\gamma_\ell) \; Be(\rho_\ell \mid \alpha_\ell, 1)\right]$$

where $Be\cdot \mid \alpha, \beta)$ represents the Beta density with positive shape parameters α and β. α_ℓ is a large value, typically larger than 50:



Computation

$\pi(\boldsymbol{\gamma} \mid \boldsymbol{y})$ can be written as a function of the Bayes factor

$$B_\gamma = \frac{m(\boldsymbol{y} \mid \gamma)}{m(\boldsymbol{y} \mid \gamma = \mathbf{1})}$$

which is a ratio of normalizing constants.
Ratio importance sampling of Chen and Shao, 1997

$$B_\gamma = E_{\mathbf{1}}\left[\frac{g(\boldsymbol{y} \mid \boldsymbol{\rho}, \boldsymbol{\eta}, \gamma)\ \pi(\boldsymbol{\rho}, \boldsymbol{\eta} \mid \gamma)}{g(\boldsymbol{y} \mid \boldsymbol{\rho}, \boldsymbol{\eta}, \gamma = \mathbf{1})\ \pi(\boldsymbol{\rho}, \boldsymbol{\eta} \mid \gamma = \mathbf{1})}\right] ,$$

which allows us to estimate all the Bayes factors using a sample from the posterior of the full model $\boldsymbol{\gamma} = \mathbf{1}$

$$\boldsymbol{\eta} = (\sigma^2, \sigma_0^2, \boldsymbol{\theta})^\top$$

Computation

If $\{\boldsymbol{\rho}^{(r)}, \boldsymbol{\eta}^{(r)},\ r = 1, \ldots, M\}$ is a sample from the posterior distribution of the unknowns for $\boldsymbol{\gamma} = \mathbf{1}$, then

$$B_\gamma \approx \frac{1}{M} \sum_{r=1}^{M} \frac{g(\boldsymbol{y} \mid \boldsymbol{\rho}^{(r)}, \boldsymbol{\eta}^{(r)}, \boldsymbol{\gamma})\ \pi(\boldsymbol{\rho}^{(r)}, \boldsymbol{\eta}^{(r)} \mid \boldsymbol{\gamma})}{g(\boldsymbol{y} \mid \boldsymbol{\rho}^{(r)}, \boldsymbol{\eta}^{(r)}, \boldsymbol{\gamma} = \mathbf{1})\ \pi(\boldsymbol{\rho}^{(r)}, \boldsymbol{\eta}^{(r)} \mid \boldsymbol{\gamma} = \mathbf{1})}$$

$$= \frac{1}{M} \sum_{r=1}^{M} \pi(\boldsymbol{\rho}^{(r)} \mid \boldsymbol{\gamma})$$

We show that this is a finite variance importance sampling estimator.

Alternative approach

- Joseph and Yan (2015) also proposes screening the discrepancy function
- Fit the KOH model; plug the estimated values in the posterior mean of the discrepancy function, $\delta(\cdot)$; next, screen the estimated discrepancy function using sensitivity analysis
- This approach hinges on a single estimate of $\boldsymbol{\theta}$ whereas ours is fully Bayesian
- Given the confounding between $\delta(\cdot)$ and $\boldsymbol{\theta}$, relying on a estimate of $\boldsymbol{\theta}$ is dangerous
- We have an example where we empirically demonstrate potential pitfalls of the two-step approach that our methodology, by relying on the joint distribution of $\boldsymbol{\theta}$ and $\delta(\cdot)$ is able to circumvent.

Simulation studies

The paper includes

- A comparison between RDVS and PIPS in the ability to detect active variables, both when $\boldsymbol{\theta}$ is fixed and when $\boldsymbol{\theta}$ is calibrated
- Our method exhibits comparable performance but requires only one MCMC sample
- Savitsky et al. (2011) is hard to implement and tune
- $f(\boldsymbol{x}_i, \boldsymbol{\theta}) = \sum_{\ell=1}^{4} \frac{|4x_{\ell i} - 2| + \theta_\ell}{1 + \theta_\ell}$
- $\delta(\boldsymbol{x}_i) = \sin(2\pi x_{1i} \cdot x_{5i}) + x_{2i}^3 + (1 - x_{6i})^3$

With $\boldsymbol{\theta}$ calibrated:

		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
RDVS	q5%	1.00	1.00	0.03	0.03	1.00	1.00	0.03	0.00
	q10%	1.00	1.00	0.07	0.05	1.00	1.00	0.03	0.00
	q15%	1.00	1.00	0.12	0.05	1.00	1.00	0.03	0.00
PIPS	th0.1	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00
	th0.5	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00
	th0.9	1.00	0.98	0.00	0.00	1.00	1.00	0.00	0.00

Table: Proportion of detection for a variable to be active when using RDVS and PIPS methods when the parameters $\boldsymbol{\theta}$ are calibrated.

Simulation studies

The paper includes idealized scenarios of computer model validation where

- a variable is incorrectly modeled by the computer model
- a variable appears in the computer model but not in the real phenomenon
- a variable appears in the real phenomenon but not in the computer model
- the wrong input is modeled in the computer model

Simulation studies

Scenarios:

$$f(\boldsymbol{x}_i, \boldsymbol{\theta}) = \sum_{\ell=1}^{3} \frac{|4x_{\ell i} - 2| + \theta_\ell}{1 + \theta_\ell}$$

Real phenomenon is

1. $\zeta(\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{|4x_{1i}^2 - 2|}{1+\theta_1} + \frac{|4x_{3i}-2|+\theta_3}{1+\theta_3}$
2. $\zeta(\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{|4x_{1i}^2 - 2|}{1+\theta_1} + \sum_{\ell=2}^{4} \frac{|4x_{\ell i}-2|+\theta_\ell}{1+\theta_\ell}$
3. $\zeta(\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{|4x_{1i}^2 - 2|}{1+\theta_1} + \frac{|4x_{2i}-2|+\theta_2}{1+\theta_2} + \frac{|4x_{5i}-2|+\theta_5}{1+\theta_5}$

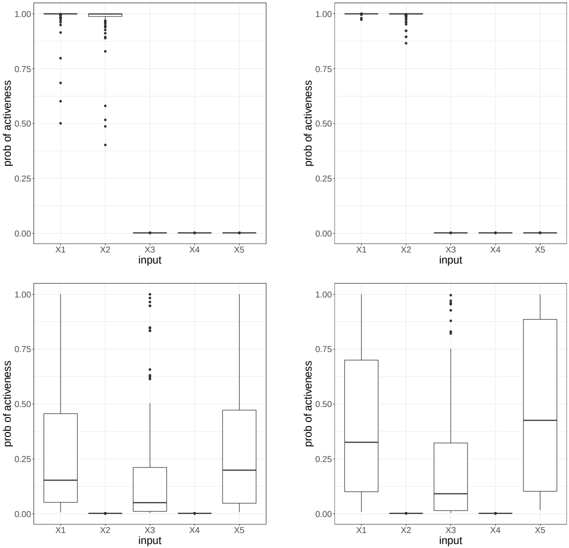


Figure: Boxplots of the probabilities of activeness over the 100 replications for cases 1 and 3. x_3 and x_5 are correlated in the simulations.

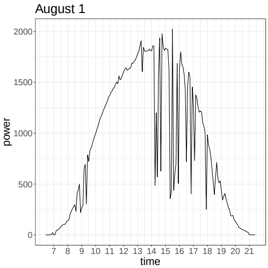
Example: Photovoltaic plant

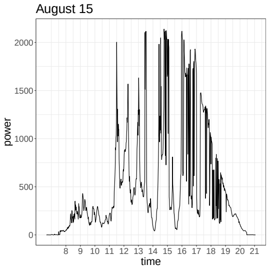
12 photovoltaic panels connected together. $f(\boldsymbol{x}, \boldsymbol{\theta})$ is the instantaneous power delivered by the plant, where

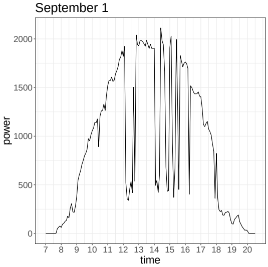
- $\boldsymbol{x} = (t, I_g, Id, T_e)^\top$: t is the time since the beginning of the year, I_g is the global irradiation of the sun, Id is the diffuse irradiation of the sun, and T_e is the ambient temperature.
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6)^\top$ but only one is treated as unknown, the module photo-conversion efficiency. A sensitivity analysis has proven the other parameters to be of negligible importance.

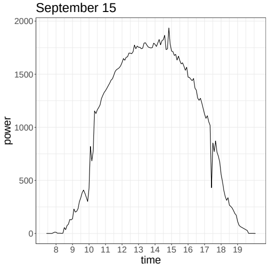
A photovoltaic plant computer model

- Instantaneous power delivered by the 12 panels was collected over a period of 2 months every 10 seconds
- $\boldsymbol{x} = (t, I_g, Id, T_e)^\top$
- The temperature on the panel T_p was measured and is tested as a potential active variable in $\delta(\cdot)$
- Considered measurements every 5 minutes
- Methodology is applied to each of the 60 days, between 99 and 178 data per day
- Boxplots of inclusion probabilities over the 60 days


August 1
power
2000
1500
1000
500
0
time
7 8 9 10 11 12 13 14 15 16 17 18 19 20 21


August 15
power
2000
1500
1000
500
0
time
8 9 10 11 12 13 14 15 16 17 18 19 20 21


September 1
power
2000
1500
1000
500
0
time
7 8 9 10 11 12 13 14 15 16 17 18 19 20


September 15
power
2000
1500
1000
500
0
time
8 9 10 11 12 13 14 15 16 17 18 19

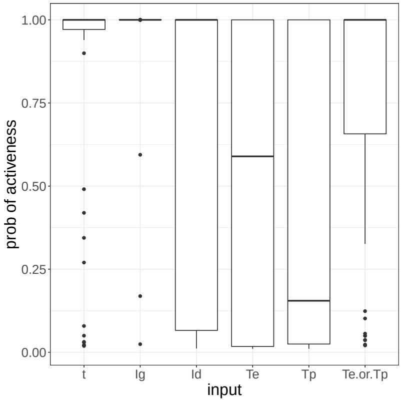


Figure: Boxplots of probabilities of activeness of the input variables in the discrepancy computed for the 60 days of data. The column (T_e or T_p) corresponds to the fact that at least one of two temperatures is active.

Discussion

- Screening the discrepancy function may provide the practitioner with a better understanding of the flaws of the computer model
- Cast this problem into the more general problem of variable selection for GaSP regression
- PIPS is computationally attractive as it relies on a single MCMC sample
- Posterior inclusion probabilities are easy to interpret
- By relying on the joint distribution of $\delta(\cdot)$ and $\boldsymbol{\theta}$ there is evidence that we alleviate the consequences of the confounding
- Moderate p requires exploring the model space as in Garcia-Donato and Martinez-Beneito (2013) — work in progress