Screening the Discrepancy Function of a Computer Model
arXiv:2109.02726

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

OBayes 2022

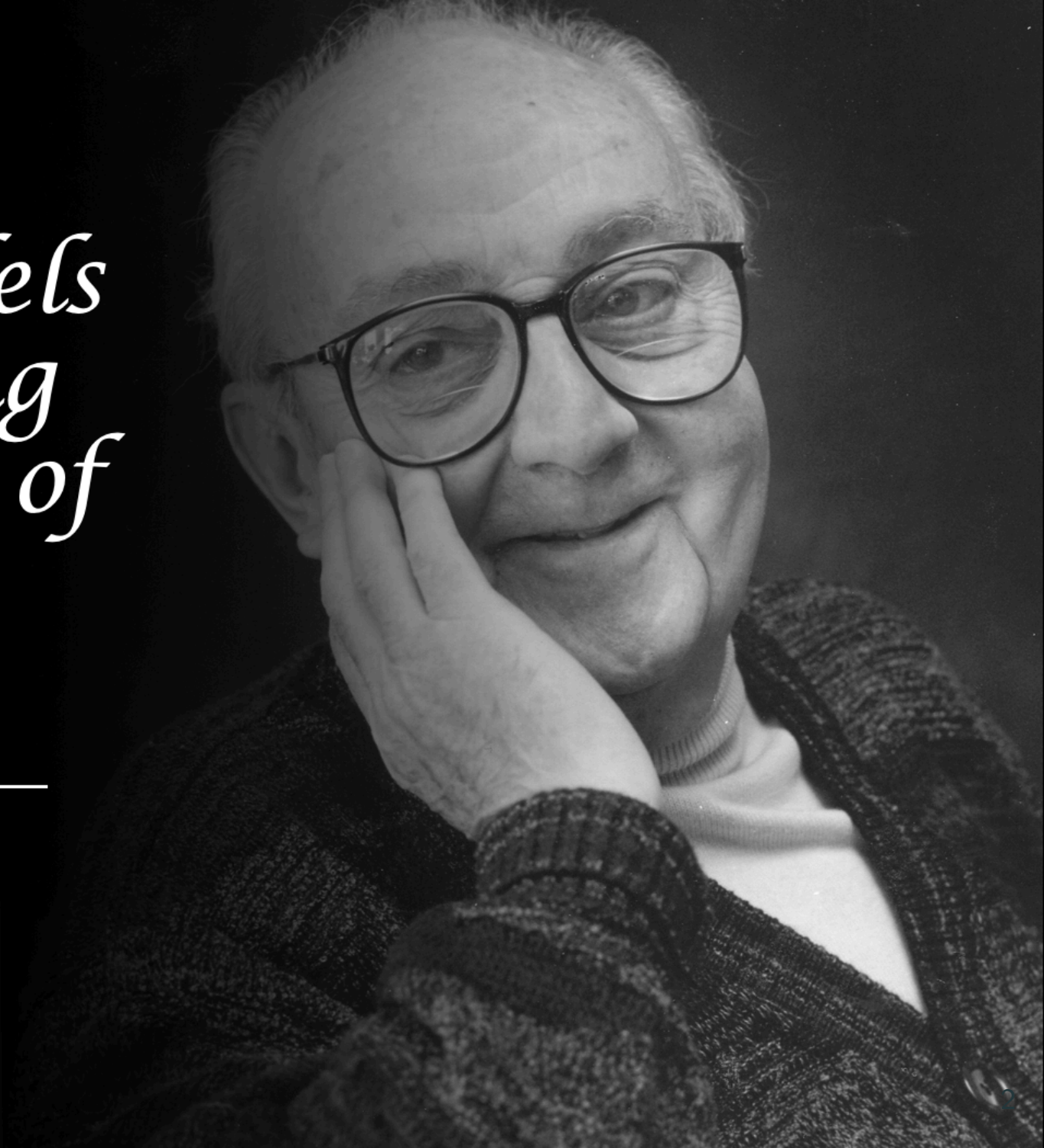
September 6 – 10, 2022 | Santa Cruz, CA

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

Motivation

"All models are wrong but some of them are useful"

George E.P. Box



Specially talking about:

Computer or mathematical models:

Let $y^M(\boldsymbol{x}, \boldsymbol{\theta})$ denote the output of a real-valued, <u>deterministic</u> 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

Motivating example

Example: A photovoltaic plant (PVP)

- Imagine a photovoltaic plant with 12 panels connected together.

- Its power production may depend on some metheorological conditions, but **which and how?**

Example: A simulator of the PVP

- A mathematical model (MM) has been developed by experts to mimic the electrical behavior of a PVP:

$$y^M : (\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^4 \times \mathbb{R}^6 \mapsto \mathbb{R}.$$

- Metheorological variables $\mathbf{x} = (t, I_g, I_d, T_e)^T$:
  - t the UTC time since the beginning of the year,
  - I_g the global irradiation of the sun,
  - I_d the diffuse irradiation of the sun and
  - T_e the ambient temperature.
- $y^M(\mathbf{x}, \boldsymbol{\theta})$: the instantaneous power following the MM.
- Carmassi et al. (2019)

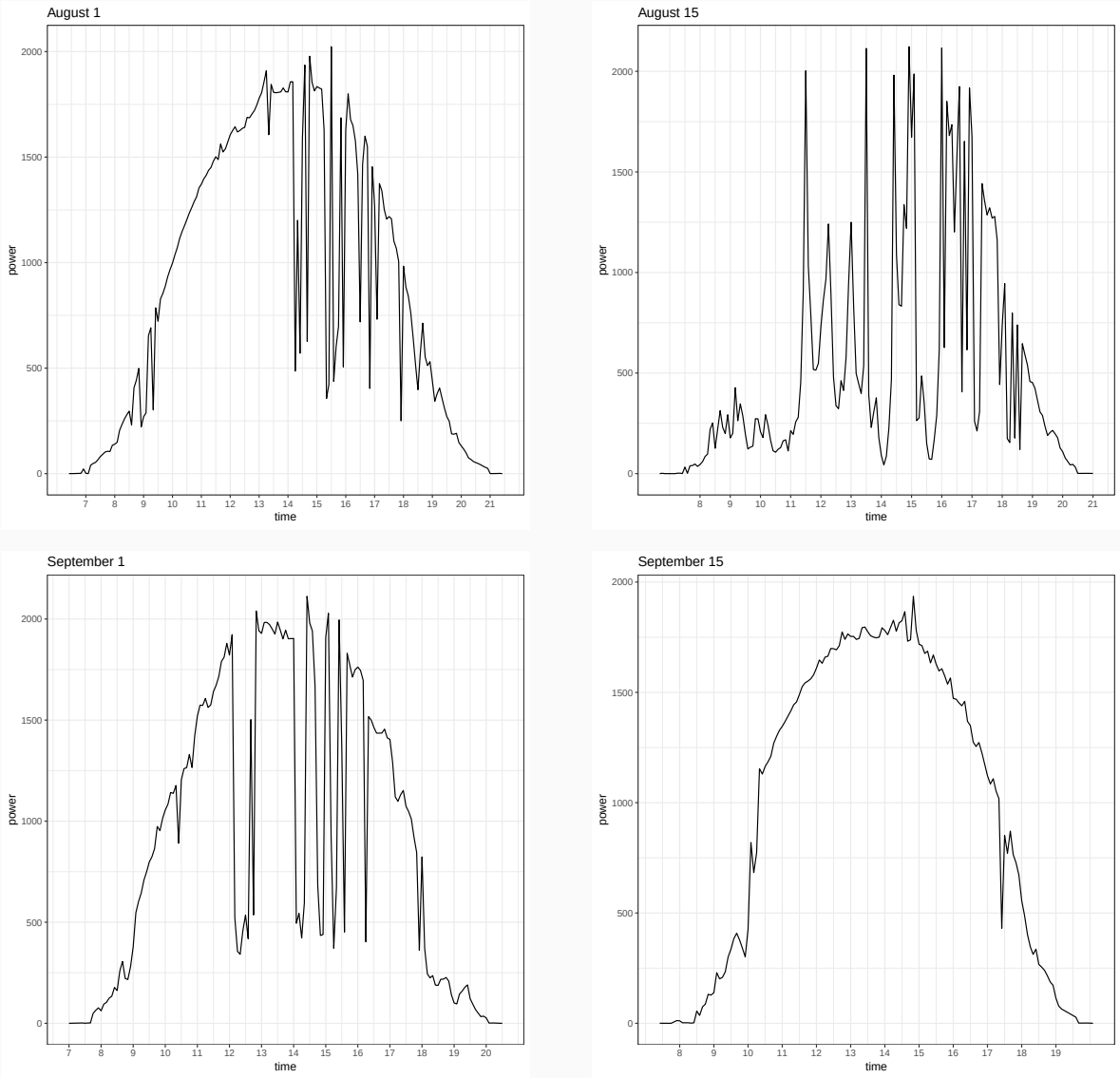
Example: Experimental Data

- Positive power production was recorded: y^F.
- Over two months (August and September).
- We just take an observation each 5 min (recorded every 10 sec).
- It is **reality** y^R **plus error**

$$y^F = y^R + \varepsilon \text{ where } \varepsilon \sim N(0, \frac{1}{\lambda^F}).$$

- Actual values for the covariates in $\mathbf{x}$ were also collected.
- Also, the temperature on the panel is recorded T_p.
- All the input and output data were normalized in $[0, 1]$

Example: Experimental Data



Example: Goal

- Understand if the MM $y^M(\mathbf{x}, \boldsymbol{\theta})$ is good enough to model reality y^R:
  - If the effect of metheorological covariates is well modelled through y^M.
  - If the temperature of the panel (not in the model) also affects the result.
- Notice that we just have field data y^F.
- In the world of MMs this process is usually known as screening.

Statistical Framework

Statistical framework

Field experiments

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

That is,

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

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

Following Kennedy and O'Hagan (2001), we model the field data as

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

Model discrepancy

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

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

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.

The most common choice for $c(\cdot, \cdot \mid \psi_\ell)$ is 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.

Remarks

- There are known confounding issues between $\delta(\cdot)$ and θ (e.g. Tuo and Wu, 2015)
- Brynjarsdóttir and O'Hagan (2014) shows 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 θ
- The goal is usually not δ itself but rather calibration, i.e., estimating $\boldsymbol{\theta}$, and improving prediction

Our approach

- We focus on this same scenario but with a different goal: help the modeler identify aspects of the computer model which might need improvement
- Variable selection procedure applied to $\delta(\boldsymbol{x})$: each model input will either be deemed active or inert — this is called **screening** in computer model jargon
- The inert inputs are the ones properly taken into account in $y^M(\boldsymbol{x}, \boldsymbol{\theta})$
- The active inputs are the one that need to examined
- In what follows, $y^M(\boldsymbol{x}, \boldsymbol{\theta})$ is fast to compute; the methodology can be extended to accommodate the situation where an emulator is needed

Screening the discrepancy

Variable selection for δ

- The sampling distribution of the data $\boldsymbol{y}^F = (y_1, \ldots, y_n)^\top$, $y_i = y^F(\boldsymbol{x}_i)$ is such that, with $\boldsymbol{f}(\boldsymbol{\theta}) = (y^M(\boldsymbol{x}_i, \boldsymbol{\theta}),\ i = 1, \ldots, n)$

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

  where $\boldsymbol{R}$ is a $n \times n$ matrix with entries $\boldsymbol{R} = [c(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{\psi})]_{i,j=1,\ldots,n}$ and $\boldsymbol{I}_n$ denotes the order-n identity matrix

- As $\psi_\ell \to +\infty$, $c(x_{\ell i}, x_{\ell j} \mid \psi_\ell) \to 1\ \forall i, j = 1, \ldots, n$ and $i \neq j$, so x_ℓ does not contribute to $\boldsymbol{R}$.

Linkletter's reparametrization

Linkletter et al. (2006) introduced the following reparametrization 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:

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

Competing models

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

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

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

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

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

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

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

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

with

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

PIPS

Once $\pi(\gamma \mid \boldsymbol{y})$ 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**.

But...

We are still missing $\pi(\rho \mid \gamma)$

Existing methodology

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

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

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, its prior 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 γ from $\pi(\boldsymbol{\rho}, \gamma) = \pi(\boldsymbol{\rho} \mid \gamma)\ \pi(\gamma)$, 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 γ is no longer available so how to declare a variable inert?

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}

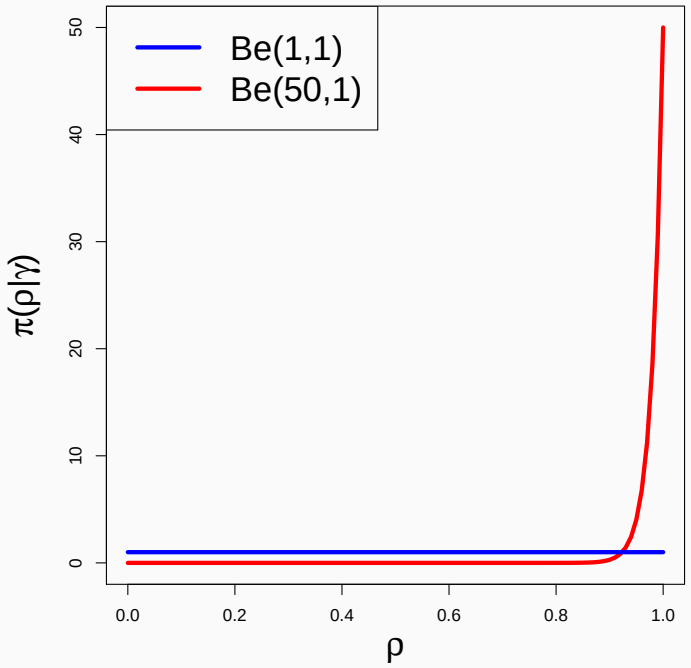
input x_ℓ if inert if the posterior median of ρ_ℓ exceeds a fixed lower percentile (say, the 10%) of the distribution of the posterior median of ρ_{new}.

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:



Computational advantages

Computation

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

$$\begin{aligned} B_\gamma &= E_{\mathbf{1}} \left[\frac{f(\boldsymbol{y} \mid \boldsymbol{\rho}, \boldsymbol{\eta}, \gamma)\ \pi(\boldsymbol{\rho}, \boldsymbol{\eta} \mid \gamma)}{f(\boldsymbol{y} \mid \boldsymbol{\rho}, \boldsymbol{\eta}, \gamma = \mathbf{1})\ \pi(\boldsymbol{\rho}, \boldsymbol{\eta} \mid \gamma = \mathbf{1})} \right] \\ &\approx \frac{1}{M} \sum_{r=1}^{M} \pi(\boldsymbol{\rho}^{(r)} \mid \boldsymbol{\gamma}) \end{aligned} \tag{1}$$

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

Simulated examples and comparisons

Simulation studies

- We compare RDVS and PIPS in the ability to detect active variables, both when $\boldsymbol{\theta}$ is fixed and when θ is calibrated
- Our method exhibits comparable performance but requires only one MCMC sample

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

Simulation studies

We also set idealized scenarios of computer model validation where 100 observations on 5 input variables $\boldsymbol{x}^\top = (x_1, x_2, x_3, x_4, x_5) \in [0,1]^5$ all simulated from independent uniform distributions except for x_3 and x_5, which are correlated.

x_1 is incorrectly modeled by the computer model and x_2 has no impact in reality

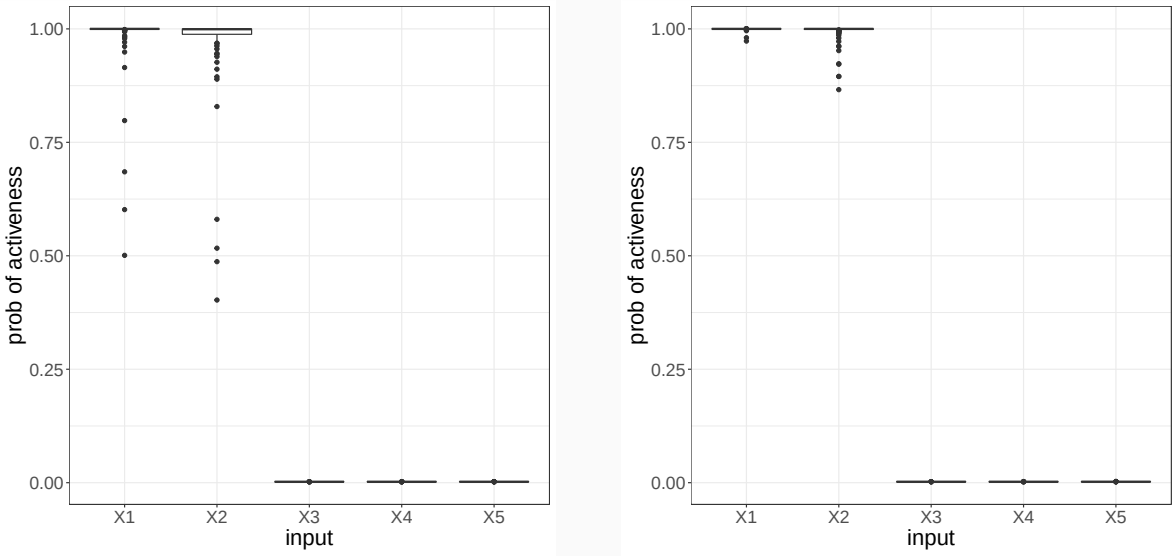


Figure 1: Boxplots of the probabilities of activeness over the 100 replications.

x_1 is incorrectly modeled by the computer model and x_3 is included instead of x_5

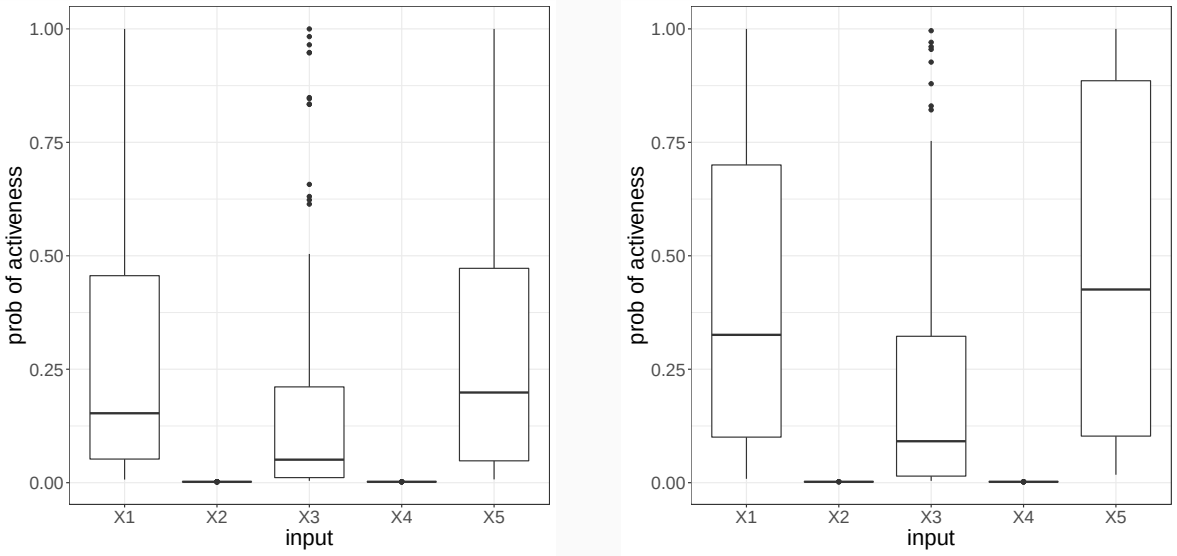


Figure 2: Boxplots of the probabilities of activeness over the 100 replications.

x_1 is incorrectly modeled by the computer model and x_4 was forgoten in the computer model

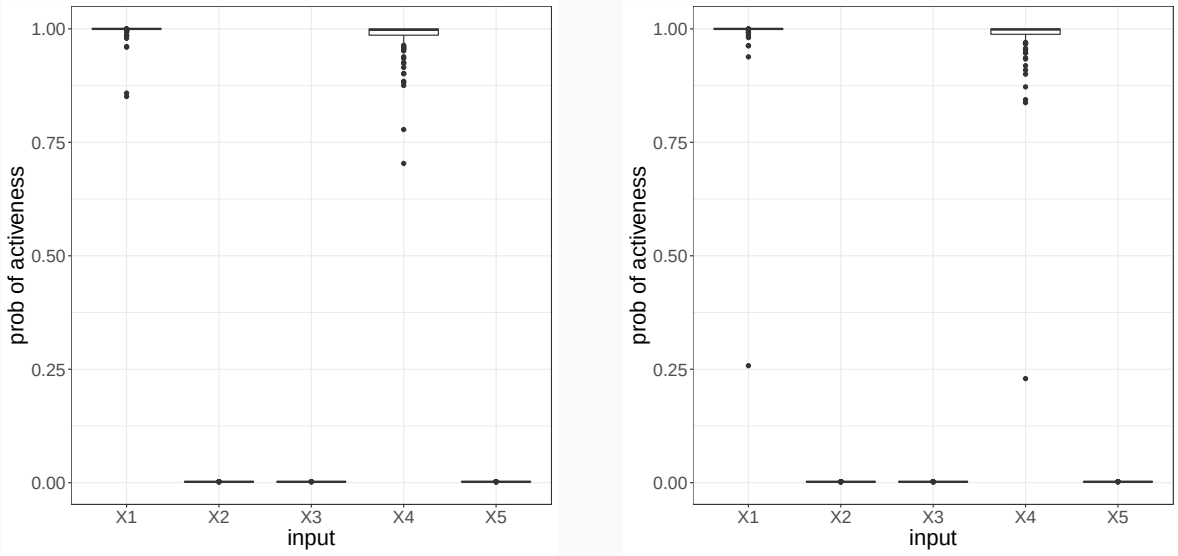


Figure 3: Boxplots of the probabilities of activeness over the 100 replications.

Photovoltaic plant

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

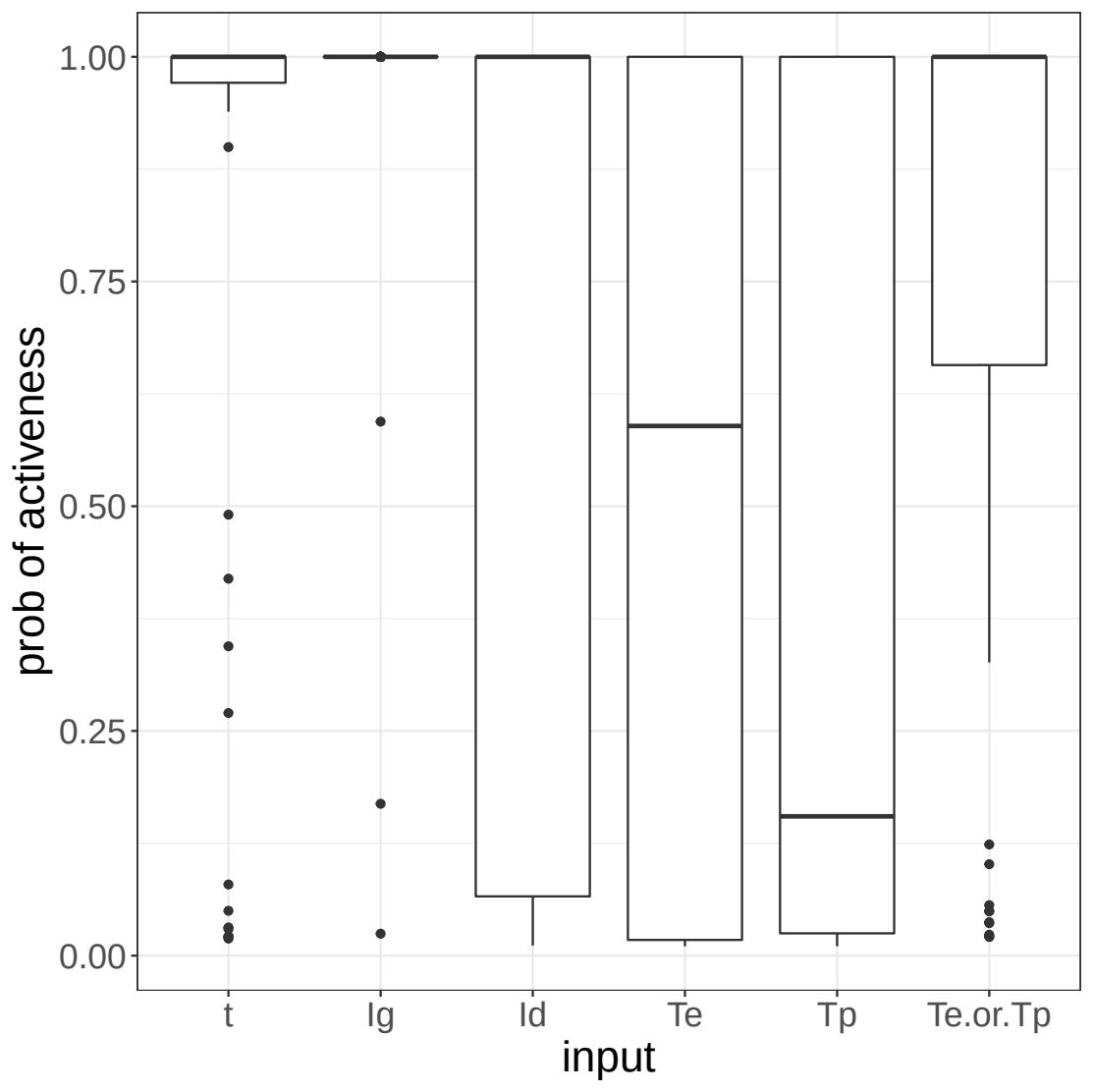


Figure 4: Boxplots of probabilities of activeness of the input variables in the discrepancy computed for the 60 days of data. The column (T_e or T_p) corresponds to the fact that at least one of two temperatures is active.

Discussion

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
- Moderate p requires exploring the model space as in Garcia-Donato and Martinez-Beneito (2013) — work in progress

References

- BARBILLON, P., FORTE, A. and PAULO, R. (2021). Screening the Discrepancy Function of a Computer Model. arXiv:2109.02726
- BRYNJARSDÓTIR, J. and O'HAGAN, A. (2014). Learning about physical parameters: the importance of model discrepancy. Inverse Problems 30, 114007.
- CHEN, M.-H. and SHAO, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics* 25, 1563–1594
- GEORGE, E. I. and R. E. McCULLOCH (1993). Variable selection via Gibbs sampling. Journal of the American Statistical Association 88, 881–-889.
- LINKLETTER, C., BINGHAM, D., HENGARTNER, N., HIGDON, D. and YE, K. Q. (2006). Variable selection for Gaussian process models in computer experiments. *Technometrics* 48 478–490.
- SAVITSKY, T., VANNUCCI, M., and SHA, N. (2011). Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statist. Sci.* 26 130–149

Thanks

This work has been partially funded by the Spanish government Grant PID2019-104790GB-I00