University of Stuttgart

SimTech

BGE BUNDESGESELLSCHAFT FÜR ENDLAGERUNG

LS³

Input-Dimension Reduction for Surrogate Model Building: Application to Subsurface Transport Models

Maria Fernanda Morales Oreamuno, Sergey Oladyshkin, Wolfgang Nowak

Motivation

- Surrogate models ($\mathcal{S}$) are used to approximate a full-complexity model's (simulator's) outputs (y), at a fraction of the time.

$$y \equiv \hat{y} = \mathcal{S}(\omega, \theta)$$

- Subsurface systems are highly heterogeneous, and can include large number of processes: **high input dimension problem**
  - Geostatistical inputs: each grid cell corresponds to a model input (ω)
- High input dimension problems are a challenge for surrogate models
  - Needs more training points → Computational problems

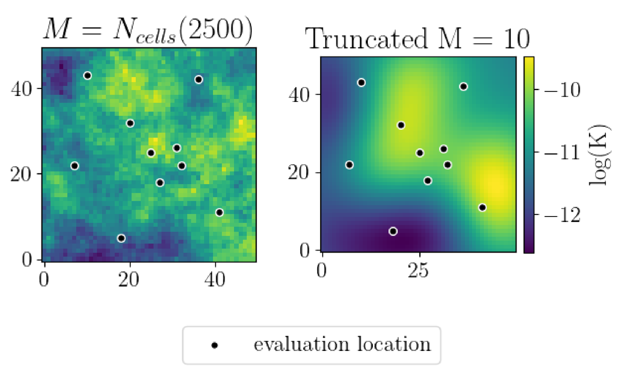
(Current) research questions

- What input dimension reduction (IDR) method should we use for geostatistically-dependent input parameters?
  - How do they behave with active learning methods, to reduce the number of training points needed?
- Can/should we consider an IDR error to account for the reduced amount of information being sent to the surrogate?
  - We want our surrogate prediction variance to account for the missing information and make sure the true (simulator) data is within the confidence intervals of the prediction.

Input dimension reduction method: Karhunen-Loéve decomposition (KLD)

Random field generation method + PCA approach

$$Z(x) = \mathrm{E}[Z(x)] + \sum_{i=1}^{M} \sqrt{\lambda_i} \cdot \varphi_i(x) \cdot \xi_i \qquad with\ \boldsymbol{M} \le N_{cells}$$

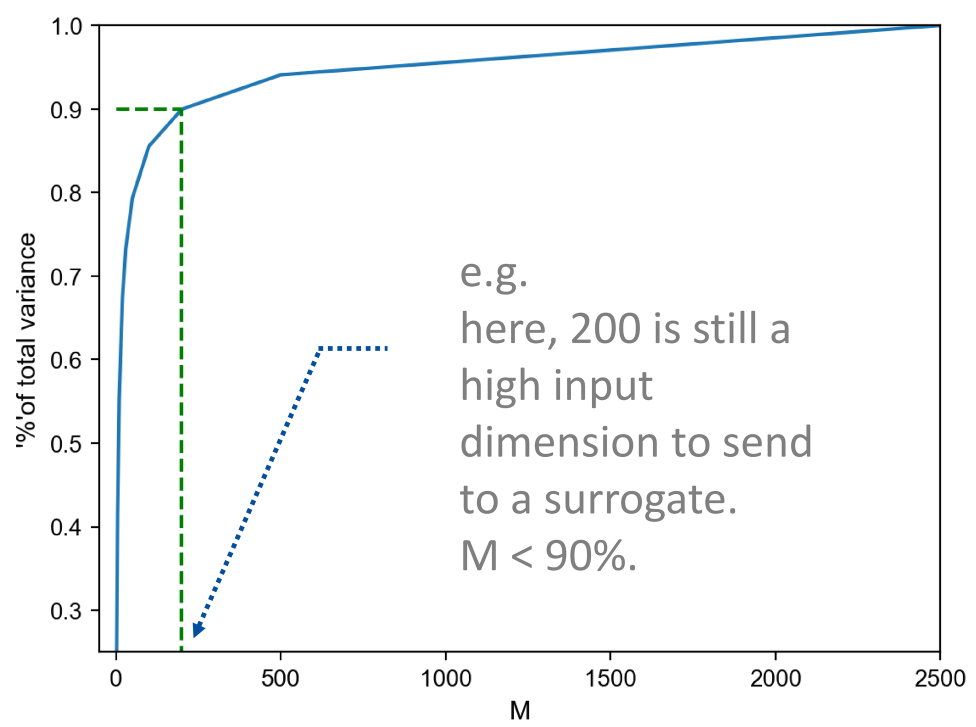


- $\boldsymbol{M} < N_{cells}$ = number of input parameters for surrogate
- Each "M" truncation value is associated to a percentage of the input variance

What happens when $M < N_{cells}$ represents a small percentage of input variance?

Is the surrogate, trained on the reduced input reproducing the behavior of the simulator as expected/desired?

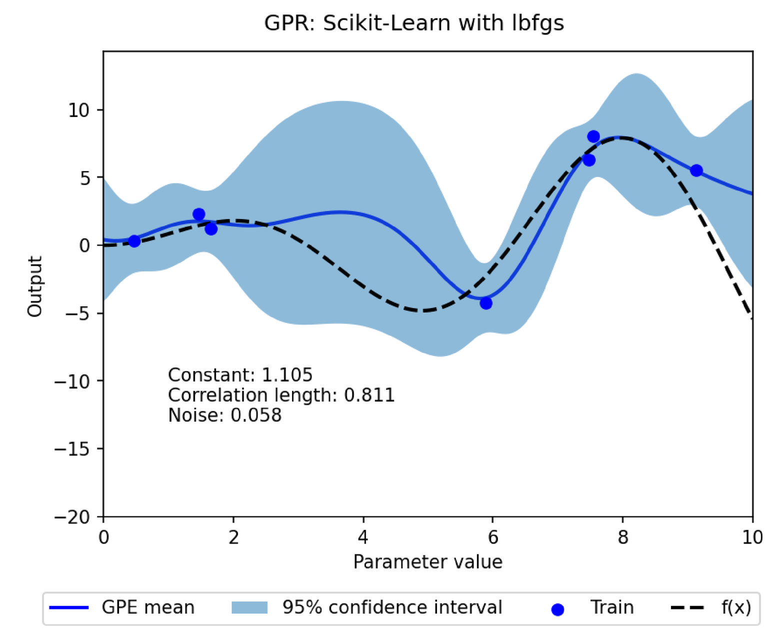
- Forward uncertainty quantification
- Posterior distributions



Input dimension reduction error

Using *Gaussian process regression (GPR)*, can we include an IDR error so the prediction variance compensates for a lack of information?

We want the prediction distribution to account for the IDR in the variance.



Approaches:

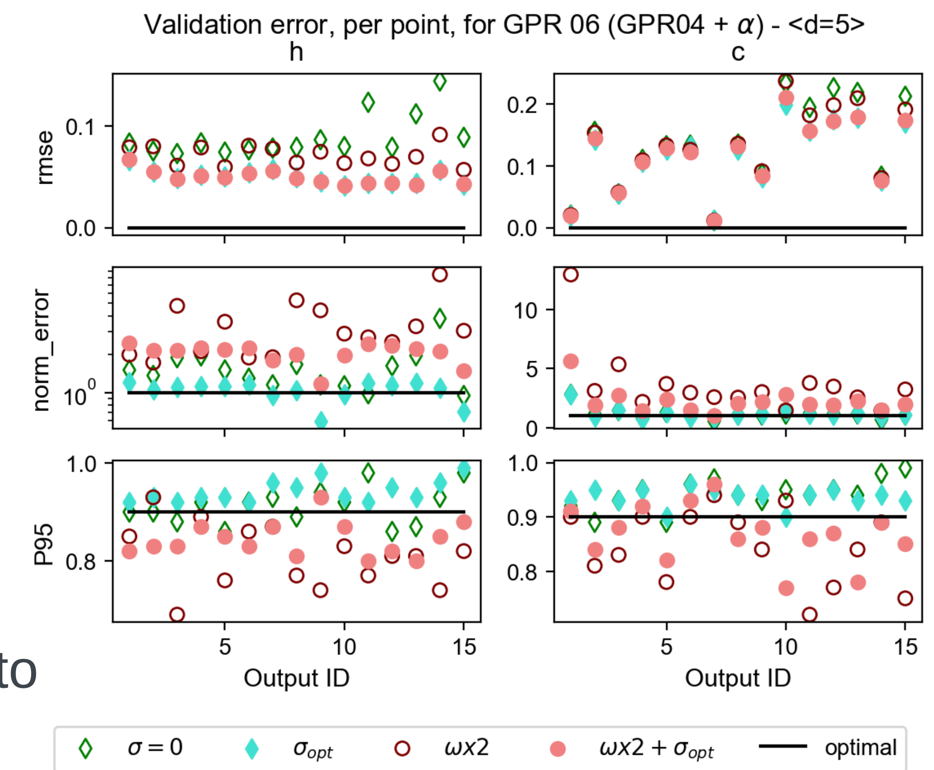
$\boldsymbol{\mu}^* = K(X, x^*)^T \cdot [K(X, X) + \boldsymbol{\sigma}^2 I]^{-1} y$

σ: optimizable parameter

σ: constant parameter, from simulator space

X: use input pairs, with IDR error from the simulator space

Preliminary results show how it is enough, but necessary, to consider an optimizable error, and the ML-approach compensates for the error.



Remaining questions:

- Is this the optimal way to consider IDR error?

What are the best method to validate surrogate, considering the distribution (variance) of our prediction?

Outlook

- With KLD: test for different truncation values (description lengths): How small is too small to train a surrogate?
- Test different IDR methods along with active learning methods
  - Variational auto-encoders
  - Pilot points
- Surrogate evaluation criteria
  - How to fairly implement Bayesian criteria to compare models
  - Include output+variance (output distribution) in evaluation criteria
- Application to independent input parameter sets: radioactive nuclide transport problem

References

- Oladyshkin, S., Mohammadi, F., Kroeker, I., & Nowak, W. (2020). Bayesian3 active learning for the Gaussian process emulator using information theory. Entropy, 22 (8), 890. doi: https://doi.org/10.3390/e22080890
- Sinsbeck, M., & Nowak, W. (2017). Sequential design of computer experiments for the solution of Bayesian inverse problems. SIAM/ASA Journal on Uncertainty Quantification, 5(1), 640-664. https://doi.org/10.1137/15M1047659
- Zhang, J., Zheng, Q., Chen, D., Wu, L., & Zeng, L. (2020). Surrogate-Based Bayesian Inverse Modeling of the Hydrological System: An Adaptive Approach Considering Surrogate Approximation Error. Water Resources Research, 56, e2019WR025721 https://doi.org/10.1029/2019WR025721