New estimation algorithm for more reliable prediction in Gaussian process regression: application to an aquatic ecosystem model

Amandine Marrel, Bertrand Iooss

Speaker: Amandine Marrel

New estimation algorithm for more reliable prediction in Gaussian process regression: application to an aquatic ecosystem model

Amandine Marrel
CEA, DES, IRESNE, DER, Cadarache F-13108 Saint-Paul-Lez-Durance,
amandine.marrel@cea.fr

Bertrand Iooss
EDF R&D, 6 quai Watier, 78400, Chatou, France, bertrand.iooss@edf.fr

In the framework of emulation of numerical simulators with Gaussian process (GP) regression [2], we proposed in this work a new algorithm for the estimation of GP covariance parameters, referred to as GP hyperparameters. The objective is twofold: to ensure a GP as predictive as possible w.r.t. to the output of interest, but also with reliable prediction intervals, i.e. representative of its prediction error. To achieve this, we propose a new constrained multi-objective algorithm for the hyperparameter estimation. It jointly maximizes the likelihood of the observations as well as the empirical coverage function of GP prediction intervals, under the constraint of not degrading the GP predictivity [1]. Cross validation techniques and advantageous update GP formulas are notably used. The benefit brought by the algorithm compared to standard algorithms is illustrated on a large benchmark of analytical functions (up to twenty input variables). An application on a EDF R&D real data test case modeling an aquatic ecosystem is also proposed: a log-kriging approach embedding our algorithm is implemented to predict the biomass of the two species. In the framework of this particular modeling, this application shows the crucial interest of well-estimated and reliable prediction variances in GP regression.

[1] C. Demay, B. Iooss, L. L. Gratiet, and A. Marrel. Model selection for Gaussian process regression: an application with highlights on the model variance validation. *Quality and Reliability Engineering International Journal*, 38:1482–1500, 2022.

[2] A. Marrel, B. Iooss, and V. Chabridon. The ICSCREAM methodology: Identification of penalizing configurations in computer experiments using screening and metamodel – Applications in thermal-hydraulics. *Nuclear Science and Engineering*, 196:301–321, 2022.

New estimation algorithm for more reliable prediction in Gaussian process regression

Application to an aquatic ecosystem model

Amandine MARREL*, **Bertrand IOOSS‡**

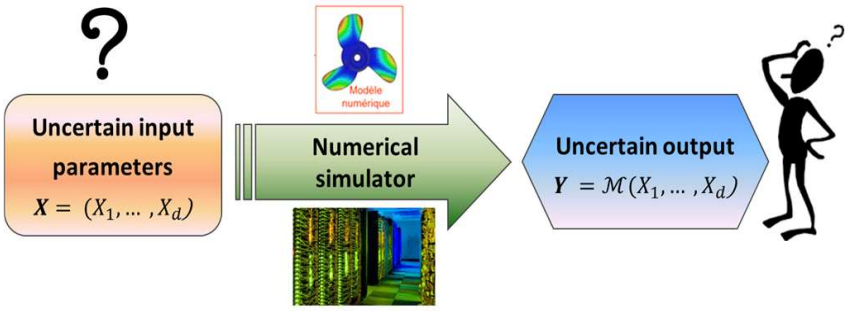
*CEA Energy Division, IRESNE, DER, Cadarache, France

‡EDF R&D, Chatou, France

ENBIS-23 Valencia Conference - September 2023

Risk assessment in nuclear accident analysis

- **Safety studies:** compute a failure risk (margins, rare events) and prioritize the risk indicators, with validated computer/numerical models
- **Numerical simulators:** fundamental tools to understand, model & predict physical phenomena
- **Large number of input parameters,** related to physical and numerical modelling
- **Uncertainty on some input parameters → impacts the uncertainty on the output, the evaluation of safety margins**
- **BEPU (Best Estimate Plus Uncertainties)**: realistic models + uncertain inputs → **Better assessment of the real margins**

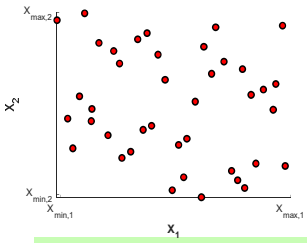


Risk assessment in nuclear accident analysis

- **How to deal with uncertainties in numerical simulation?**
  - → **Probabilistic framework and statistical methods**
  - → Monte Carlo-based approaches and data analysis ⇒ Data Sciences techniques
  - → CPU-expensive simulator ⇒ Essential use of **machine learning (metamodels)**

Crucial use of metamodel (*machine learning*)

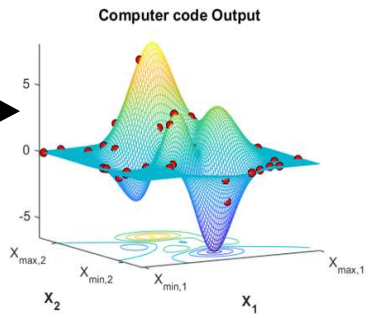
Design of numerical experiments → **Numerical simulations** → Analysis of simulator outputs



Simulator

$Y = \mathcal{M}(X_1, \ldots, X_d)$

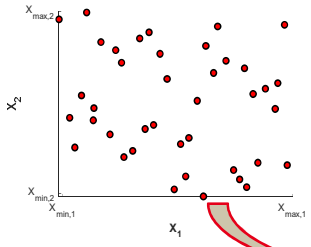
In case of costly $\mathcal{M}$:
Model reduction or Approximation with Machine Learning (ML)



Incertain inputs:
- Environmental variables
- Physical parameters
- Modeling and numerical parameters

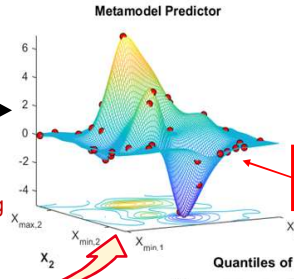
Crucial use of metamodel (*machine learning*)

Design of numerical experiments → **Numerical simulations** → Analysis of simulator outputs



Simulator

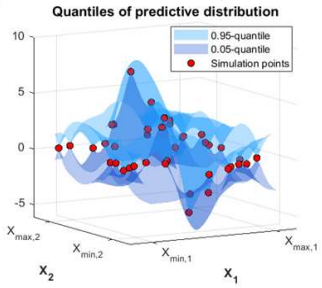
$Y = \mathcal{M}(X_1, \ldots, X_d)$



Probabilistic metamodel

In case of costly $\mathcal{M}$:
Approximation with Machine Learning

Metamodel: $Y_{app} \approx \mathcal{M}(X)$



- ✓ Build from the dataset a predictor that mimics the true model $\mathcal{M}$, with good prediction capabilities ⇒ to be controled
- ✓ **Bonus:** with reliable quantification of prediction error !

Ex : Polynomials, splines, neural networks, random forests… ***Gaussian process***

Reminders on Gaussian process metamodel

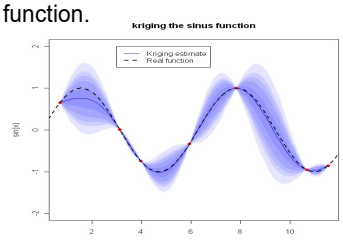
▶ **Only a *n*-sample of simulations is available** (Monte-Carlo sample, LHS, space-filling design, etc.)

$$\boldsymbol{D_S} = \left(\boldsymbol{X_S^{(j)}}, \boldsymbol{Y_S^{(j)}}\right)_{\mathbf{1 \le j \le n}} \text{ where } Y_S^{(j)} = \mathcal{M}\left(X_S^{(j)}\right)$$

▶ **Probabilistic surrogate model** : response is considered as a realization of a random GP field [RW05,Gra21]:

$$Y(\boldsymbol{x}) \sim GP(\mu(\boldsymbol{x}), k(\boldsymbol{x}', \boldsymbol{x}))$$

With $\mu(\boldsymbol{x})$ the mean and $k(\boldsymbol{x}', \boldsymbol{x})$ the covariance function.



⇒ Predictive GP is the GP conditioned by the observations (X_S, Y_S) :

$$Y(\boldsymbol{x}^*)_{|Y(X_S)=Y_S} \sim GP\left(\hat{\mu}(\boldsymbol{x}^*), \hat{s}(\boldsymbol{x}', \boldsymbol{x}^*)\right)$$

With analytical formulations for $\hat{\mu}(\boldsymbol{x}^*)$ *and* $\hat{s}(\boldsymbol{x}', \boldsymbol{x}^*)$

⇒ Conditional mean $\hat{\mu}(\boldsymbol{x}^*)$ serves as the predictor at location $\boldsymbol{x}^*$

⇒ Prediction variance (*i.e.* mean squared error) is given by conditional covariance $\hat{s}(\boldsymbol{x}^*, \boldsymbol{x}^*)$

⇒ Prediction interval of any level α can be built at any location $\boldsymbol{x}^*$

Reminders on Gaussian process metamodel

▶ **In practice:** parametric choices for trend function μ and covariance function k

$$Y(\boldsymbol{x}) \sim GP(\mu(\boldsymbol{x}), k(\boldsymbol{x}', \boldsymbol{x}))$$

⇒ For μ: either constant or linear basis

⇒ For k: tensorized 1-D covariance functions of ν-Matérn (with $h = |x - \tilde{x}|$)

1-Dim → $$k_{\sigma,\nu,\theta}(x, \tilde{x}) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} h}{\theta} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu} h}{\theta} \right)$$

d-Dim → $$k_{\sigma,\nu,\theta}(\mathbf{x}, \tilde{\mathbf{x}}) = \sigma^2 \prod_{i=1}^{d} k_{1,\nu,\theta_i}(x_i - \tilde{x}_i)$$

⇒ Need to estimate from the dataset the correlation hyperparameters $\boldsymbol{\theta} \in \mathbb{R}^{+,d}$

→ How to to ensure that the estimated hyperparameters $\boldsymbol{\theta}$ yield good predicitivity but also **reliable GP prediction intervals**? ⇒ Crucial for safety applications

→ Especially in large dimension (d>10) and small dataset (n~100)?

GP estimation and validation

▶ **Usual estimation methods** [KO22,Mur21,Pet22]

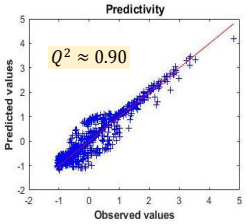
→ Maximum likelihood-based estimation (MLE) ⇔ minimization of the NLL (*Negative Log-Likelihood*)

→ Cross-validation-based estimation (Leave-One-Out or K-fold): minimization of $\mathrm{RMSE} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y(\mathbf{x}_i) - \hat{y}_{-i}(\mathbf{x}_i) \right)^2 \right\}^{1/2}$

→ Bayesian approaches (CPU ++)

▶ **Validation criteria computed by cross-validation (LOO)** [DIG+21]

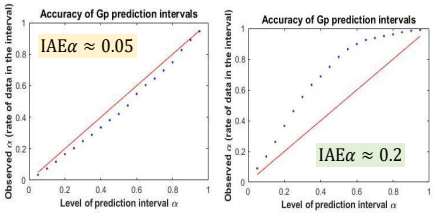
→ Accuracy of the GP predictor: $Q^2 = 1 - \frac{RMSE^2}{\frac{1}{n}\sum_{i=1}^{n} \left(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \right)^2}$



→ **Accuracy of the whole GP conditional distribution** ([ABG22])

Empirical coverage function for α ∈ [0,1]: $\hat{\Delta}(\alpha) = \frac{1}{n} \sum_{i=1}^{n} 1\left\{ y_i \in \mathcal{PI}_{\alpha,-i}(\mathbf{x}_i) \right\}$

⇒ **Integrated Absolute Error on $\hat{\Delta}(\alpha)$** $\mathrm{IAE}\alpha = \int_0^1 \left| \hat{\Delta}(\alpha) - \alpha \right| \mathrm{d}\alpha$



(RMSE : Root Mean squared Error)

New estimation algorithm for GP hyperparameters

▶ Study of criteria **NLL, Q^2 and $IAE\alpha$** on a large benchmark of analytical test functions

→ **Close behavior of NLL and Q^2 ⇒ keep NLL as the main estimation objective** to ensure the predictivity of the metamodel ⇒ Consistent with result of [PBF+22,Pet22]

→ **$IAE\alpha$ optimization is not guaranteed when optimizing NLL**

→ **But, in the neighborhood of the optimal NLL point, existence of better points θ (for $IAE\alpha$)**, however need to control the possible degradation of Q^2 value, which guarantees the predictivity

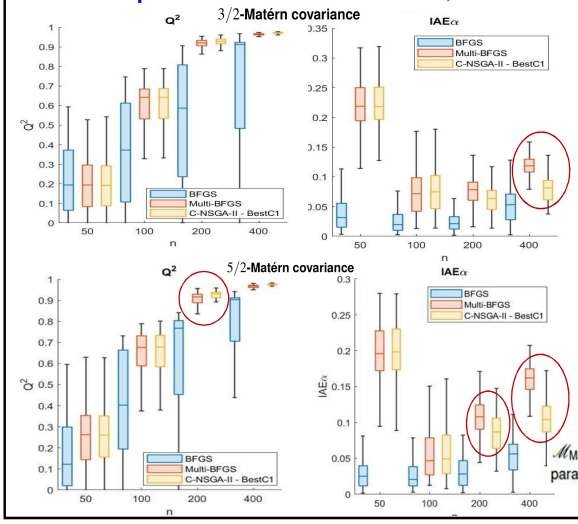
⇒ **Optimization based on NLL and $IAE\alpha$ + Control of Q^2**
($IAE\alpha$ and Q^2 estimated by cross validation + use of LOO Dubrule formulas)

⇒ **Proposition of a multi-objective NSGA-II algorithm with constraint on Q^2**

New estimation algorithm for GP hyperparameters

▶ Intensive benchmark on analytic functions

- Comparison with usual algorithms based on NLL optimization only (with standard algorithms)
- **Input dimension** from *d*=2 to 20, with ≠ covariance functions, ≠ sample sizes, ≠ design of experiments



⇒ **Predictivity with Constrained Multi-Objective algorithm (C-NSGA-II-BestC1) at least as good as simple NLL optimization**

⇒ Furthermore, **improvement of $IAE\alpha$ especially if** :

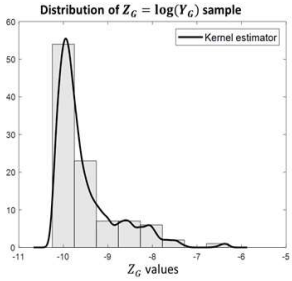
- **The model is misspecified: covariance does not match the regularity of the function.**
- **When the number of hyperparameters is large** (case of large dimension *d* + tensorized anisotropic stationary covariance)

Marrel-d20 Function – Evolution of validation criteria, according to sample size *n*, for different hyper-parameter estimation methods (GP with different covariances and without nugget effect).

New estimation algorithm for GP hyperparameters

▶ Application to an aquatic ecosystem data case (EDF test case)

- MELODY model: prey-predator chain in an aquatic ecosystem
- ***d = 20*** uncertain inputs. **Output of interest (Y_G)** : Biomass of grazers at day 60.
- Sample of ***n = 100*** simulations of the model MELODY (LHS design)
- Need of **preliminar logarithmic transformation**



⇒ **Lognormal-kriging modeling:**

- Emulation of $Z_G = \log(Y_G)$ with GP regression
- Lognormal-kriging back-transformations to obtain a metamodel for Y_G

$$\hat{y}_G(\mathbf{x}) = e^{\left(\hat{z}_G(\mathbf{x}) + 0.5\hat{s}^2_{z_G}(\mathbf{x})\right)}$$
$$\hat{s}^2_Y(\mathbf{x}) = \left(e^{\hat{s}^2_{z_G}(\mathbf{x})} - 1\right) e^{\left(2\hat{z}_G(\mathbf{x}) + \hat{s}^2_{z_G}(\mathbf{x})\right)}$$

- Additional comparison with **Bayesian RobustGaSP approach** (package of GU et al. [GWB18])

New estimation algorithm for GP hyperparameters

▶ Application to an aquatic ecosystem data case (EDF test case)

⇒ **With** nugget effect (included in the set of GP hyperparameters to be estimated)

Data	Covariance	Predictivity Coefficient Q^2			IAEα		
		Multi-BFGS	C-NSGA-II-BestC1	RobustGaSP	Multi BFGS	C-NSGA-II-BestC1	RobustGaSP
Y_2	Matern3/2	0,70	0,74	0,25	0,10	0,07	0,04
	Matern5/2	0,77	**0,82**	0,66	0,09	**0,02**	0,07
	Gaussian	0,75	**0,79**	0,66	0,08	**0,02**	0,06

⇒ Best results with **Constr-NSGA-II algorithm**: better Q^2 and IAEα

⇒ **Without** nugget effect*

Data	Covariance	Predictivity Coefficient Q^2			IAEα		
		Multi-BFGS	C-NSGA-II-BestC1	RobustGaSP	Multi BFGS	C-NSGA-II-BestC1	RobustGaSP
Y_2	Matern3/2	0,70	**0,75**	0,47	0,10	0,06	0,03
	Matern5/2	0,78	**0,84**	**0,83**	0,08	**0,02**	0,07
	Gaussian	0,70	**0,72**	**0,89**	0,06	**0,03**	0,06

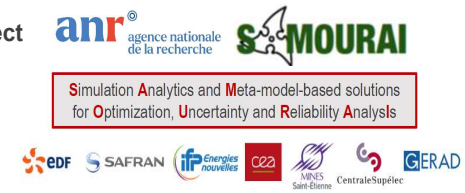
⇒ **Better behavior of RobustGasp without nugget** : best Q^2 but not IAEα

⇒ **Constr-NSGA-II algorithm more robust to modeling choices**

**additional white noise ⇒ one additional variance parameter (to be estimated)*

Conclusions

- Demonstrates the benefits of considering (in addition to NLL) some criteria assessing the accuracy of **whole GP distribution** when estimating hyperparameters ⇒ **More robust estimation !**
- Particular attention must be paid to **GP validation**
- Part of a more general effort to ensure confidence in machine learning
- Some improvement: **combining our approach with RobustGasp** with tractable approximation of robust prior proposed by [GWB18]
- **Work partly funded by ANR SAMOURAI research project**



References

[ABG23] Acharki, N., Bertoncello, A., and Garnier, J. (2023). Robust prediction interval estimation for GP by cross-validation method. Computational Statistics Data Analysis, 178:107597

[CCC12] Ciric, C., Ciffroy, P., and Charles, S. (2012). Use of sensitivity analysis to discriminate non-influential and influential parameters within an aquatic ecosystem model. Ecological Modelling, 246:119–130.

[DIG+21] Demay, C., Iooss, B., Gratiet, L., and Marrel, A. (2022). Model selection for GP regression: an application with highlights on the model variance validation. QREI Journal, 38:1482-1500.

[Gra21] B. Gramacy. Gaussian Process Modeling, Design, and Optimization for the Applied Sciences. Chapman and Hall/CRC, 2021.

[GWB18] Gu, M., Wang, X., and Berger, J. O. (2018). Robust gaussian stochastic process emulation. The Annals of Statistics, 46(6A):3038 – 3066.

[KO22] Karvonen & Oates (2022). Maximum Likelihood Estimation in GP is ill-posed. Preprint.

[Mur21] Muré (2021). Propriety of the reference posterior GP distribution. The Annals of Statistics. 49(4):2356-2377.

[Pet22] Petit S. (2022). Improved Gaussian process modeling. Application to Bayesian optimization. PhD University Paris-Saclay.

[PBF+22] Petit, S., Bect, J., Feliot, P., and Vazquez, E. (2022). Model parameters in GP interpolation: an empirical study of selection criteria. Preprint - https://hal-centralesupelec.archives-ouvertes.fr/hal-03285513.

[RW05] C.E. Rasmussen and C.K.I. Williams. Gaussian processes for machine learning. MIT Press, 2006.