Conformal Prediction for Surrogate Modeling with Gaussian Processes

Edgar Jaber

CP5 Session - 02/27/24 - SIAM-UQ 2024 - Trieste, Italy

Joint work with: V. Blot, N. Brunel (Quantmetry),
V. Chabridon, E. Remy, B. Iooss (EDF R&D),
D. Lucor (LISN), M. Mougeot (Centre Borelli)

Table of contents

Problem formulation

- Computer code $g : \mathcal{X} \to \mathcal{Y}$, with $\mathcal{X}, \mathcal{Y}$ measurable spaces, used in industry applications
- Uncertainty Quantification (UQ) methodology: how uncertainty on the inputs X affects our knowledge of the output $g(X)$?

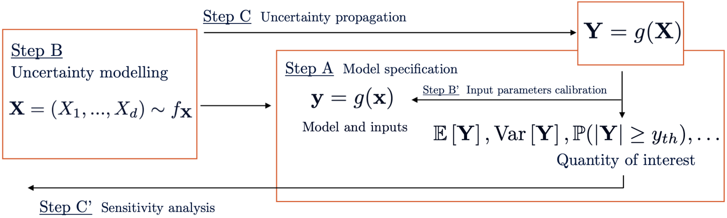


Figure: General UQ methodology [De Rocquigny et al., 2008].

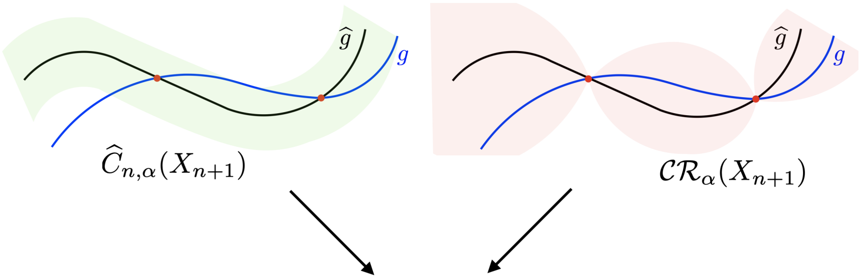
Problem formulation

- Some codes are time-costly $\rightarrow$ use of *surrogates* $\widehat{g}$
- Surrogates (or metamodels) facilitate heavy Monte Carlo batch runs and/or sensitivity analysis $\rightarrow$ in *Step C, C'*
- Assess the *quality* of these surrogates (decision making, trust in simulation outputs, ...) $\rightarrow$ metamodel *validation*
- Metamodel validation is still a debated topic and has open questions $\rightarrow$ no consensus on a best methodology
- Gaussian Processes (GPs) are a type of Bayesian metamodels $\rightarrow$ with a notion of *uncertainty* $\rightarrow$ but needs hypotheses to be interpreted
- Idea: use **Conformal Prediction** (CP) paradigm [Vovk et al., 2005] which is a generic, model-agnostic theory allowing to build *prediction sets* with frequentist coverage guarantees.
- Possibly useful for *qualifying* GP surrogate models?

Problem formulation

Conformal prediction

Gaussian processes



Adaptive conformal prediction for robust uncertainty quantification of Gaussian process surrogate models

Table of contents

CP in regression setting

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Definition

[Vovk et al., 2005] Let $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Let $n \in \mathbb{N}$ and $\mathcal{D}_n = \{Z_1, \ldots, Z_n\} \in \mathcal{Z}^n$ a training sample. For $\alpha \in (0,1)$, a conformal predictor of coverage $1-\alpha$ is any measurable function of the form:

$$\begin{aligned} C_\alpha \colon \mathcal{Z}^n \times \mathcal{X} &\to 2^{\mathcal{Y}} \\ (\mathcal{D}_n, X) &\mapsto C_{n,\alpha}(X), \end{aligned} \tag{1}$$

s.t. for any new couple of points $Z_{n+1} = (X_{n+1}, Y_{n+1}) \in \mathcal{Z}$ (marginal coverage property):

$$\mathbb{P}\left(Y_{n+1} \in C_{n,\alpha}(X_{n+1})\right) \geq 1-\alpha. \tag{2}$$

Three main methods for estimating conformal-predictors: full-conformal, split-conformal (see [Angelopoulos and Bates, 2023]) and cross conformal estimators → focus on the latter

Jackknife+/minmax interval estimators

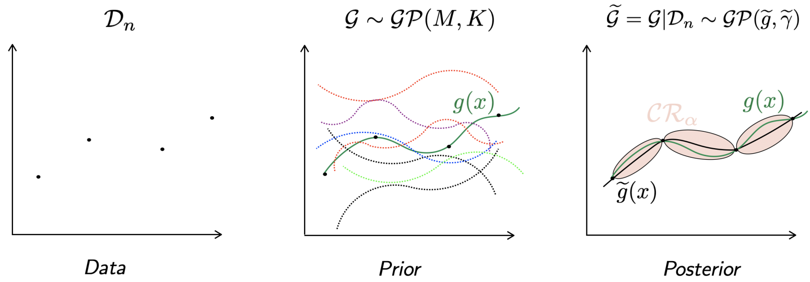
- $g : \mathcal{X} \subseteq \mathbb{R}^d \longrightarrow \mathcal{Y} \subseteq \mathbb{R}$
- $\mathcal{D}_n = \{(X_1, g(X_1)), \ldots, (X_n, g(X_n))\}$ an i.i.d. design of experiments
- $\widehat{g}$ a surrogate model trained on $\mathcal{D}_n$, $\widehat{g}_{-i}$ leave-one-out (LOO) surrogate model trained on $\mathcal{D}_n \backslash \{(X_i, g(X_i))\}$
- With empirical quantile of LOO residues $\rightarrow$ can build interval estimators: Jackknife+ $\widehat{C}_{n,\alpha}^{J+}$ & Jackknife-minmax $\widehat{C}_{n,\alpha}^{J-minmax}$

Estimators	$\widehat{C}_{n,\alpha}^{J+}$	$\widehat{C}_{n,\alpha}^{J-minmax}$
Marginal coverage	$\alpha \in (0, 1/2)$	$\alpha \in (0, 1)$
$+$	Cross-validation method Fast to compute	Coverage property
$-$	Constant width	Width is too large

Table: Main cross-conformal estimators pros and cons.

Table of contents

Gaussian process (GP) surrogates



- GP metamodel prediction is the *posterior mean* $\widehat{g} = \widetilde{g}$.
- Notion of uncertainty through the posterior covariance $\widetilde{\gamma}$ and the Gaussian structure of the metamodel.

Bayesian credibility intervals

- Credibility intervals of the posterior GP, for any new point $X_{n+1} \in \mathcal{X},\ \alpha \in (0,1)$:

$$\mathcal{CR}_\alpha(X_{n+1}) = \left[\, \widetilde{g}(X_{n+1}) \pm u_{1-\alpha/2}\widetilde{\gamma}(X_{n+1}) \,\right]. \quad (3)$$

- Has the conditional coverage property (stronger than marginal):

$$\mathbb{P}\left(g(X_{n+1}) \in \mathcal{CR}_\alpha(X_{n+1}) \mid \mathcal{D}_n\right) = 1 - \alpha. \quad (4)$$

- *However*, the above equality relies on **two** hypotheses:
  1. g is modeled by $\mathcal{G} \sim \mathcal{GP}(M, K)$
  2. priors mean and covariance functions M, K are well-specified.

- No generic way to test these two hypotheses when metamodeling black-box computer codes $\rightarrow$ *a real challenge for industrial application of UQ*.

Table of contents

The proposed J+GP estimator

- Optimize the hyperparameters of the GP kernel.
- Posterior mean $\widetilde{g}$ and standard-deviation $\widetilde{\gamma} = \widetilde{K}^{1/2}$.
- Define the Leave-One-Out-Gaussian (LOOγ) error:

$$R_i^{LOO\gamma} := \frac{|g(X_i) - \widetilde{g}_{-i}(X_i)|}{\widetilde{\gamma}_{-i}^{\beta}(X_i)}, \ \forall \beta \in \mathbb{N}. \tag{5}$$

Main result and consequences [Jaber et al., 2024a]

$$\widehat{C}_{n,\alpha}^{J+GP}(X_{n+1}) = \left[\widehat{q}_{n,\alpha}^{\pm} \left\{ \widetilde{g}_{-i}(X_{n+1}) \pm R_i^{LOO\gamma} \times \widetilde{\gamma}_{-i}^{\beta}(X_{n+1}) \right\} \right] \tag{6}$$

- Coverage property still verified for $\alpha \in (0, 1/2)$
- Intervals have adaptive width $\rightarrow$ more informative
- No hypotheses for interpreting the interval!
- The J+GP-minmax variant has the same properties

Proposed methodology for GP qualification

On a test-set $\mathcal{D}_m \neq \mathcal{D}_n$: compare different GPs as well as classical cross-CP and Bayesian credibility sets by computing:

- the predictivity coefficient:

$$Q^2 = 1 - \sum_{i=1}^{m} \frac{|g(X_i) - \widetilde{g}(X_i)|^2}{\mathrm{Var}(g(X_i))}. \tag{7}$$

- the empirical coverage for usual α thresholds $(1\%, 5\%, 10\%)$:

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left\{g(X_i) \in \widehat{C}^{*}_{n,\alpha}(X_i)\right\} \gtrsim 1 - \alpha.$$

- the adaptivity, by use of the Spearman correlation coefficient r_s between the width of the interval and the metamodel error:

$$0 \ll r_s\left(\{(\ell(\widehat{C}^{*}_{n,\alpha}(X_i)), |g(X_i) - \widetilde{g}(X_i)|)\}_{i\in\{1,\ldots,m\}}\right)$$

Industrial use case

- The **steam generator** (SG) $\rightarrow$ heat exchanger between the primary and secondary circuits of a nuclear power plant (NPP).

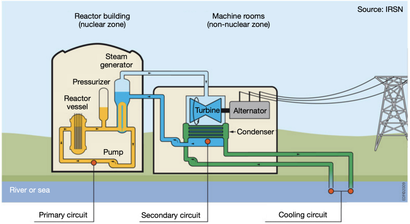


Figure: NPP Scheme

- Corrosion in the secondary circuit produces iron oxide impurities $\rightarrow$ **clogging** of the SG over time, requires **maintenance**.

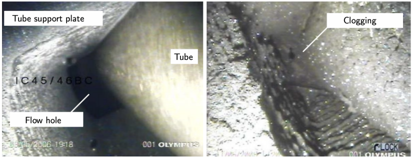


Figure: video examination during an PWR outage (© EDF)

Industrial use case

- EDF *time-costly* steam-generator clogging simulation code [Jaber et al., 2024b], input dimension $d = 7 \rightarrow$ GP metamodel with Matérn-ν covariance priors, with hyperparameters optimized (σ, θ) by MLE:

$$K(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|x - x'|}{\theta} \right)^{\nu} K_{\nu} \left(\sqrt{2\nu \frac{|x - x'|}{\theta}} \right).$$

- Parameter ν governs the *regularity* of the metamodel
- Crude Monte Carlo design of experiments of 10^3 points, 80% used for training and 20% for testing/qualification.

Component	Distribution	Component	Distribution
$X^{(1)}$	$\mathcal{N}(101.6, 4.0)$	$X^{(5)}$	$\mathcal{T}(0.5, 5.0, 10.0) \times 10^{-6}$
$X^{(2)}$	$\mathcal{N}(0.0233, 0.0005)$	$X^{(6)}$	$\mathcal{T}(1.0, 4.5, 8.0) \times 10^{-9}$
$X^{(3)}$	$\mathcal{T}(0.2, 0.3, 0.5)$	$X^{(7)}$	$\mathcal{T}(0.1, 7.8, 12) \times 10^{-4}$
$X^{(4)}$	$\mathcal{T}(0.01, 0.05, 0.3)$		

Table: Distributions of the input components of the clogging code.

How to quantify the quality of the GP-surrogate in prevision?

$\rightarrow$ UQ with adaptive conformal predictors!

Average widths / correlation, $\alpha = 0.05$, $\beta = 1/2$

✓/ ✗: empirical coverage is / is not achieved.

ν	Q^2	$\mathcal{CR}_\alpha$	$\widehat{C}_{n,\alpha}^{J-mm}$	$\widehat{C}_{n,\alpha}^{J-mm-GP}$
$1/2$	0.990	5.6 ✓	3.9 ✗	3.7 ✗
$3/2$	0.996	2.4 ✓	2.3 ✗	2.3 ✓
$5/2$	0.997	1.9 ✗	2.2 ✓	2.1 ✓

Table: Average widths of prediction intervals and Q^2.

ν	Q^2	$\mathcal{CR}_\alpha$	$\widehat{C}_{n,\alpha}^{J-mm}$	$\widehat{C}_{n,\alpha}^{J-mm-GP}$
$1/2$	0.990	0.46 ✓	0.66 ✗	0.63 ✗
$3/2$	0.996	0.35 ✓	0.65 ✗	0.55 ✓
$5/2$	0.997	0.21 ✗	0.60 ✓	0.45 ✓

Table: Correlation between widths and GP approximation error and Q^2.

$\rightarrow$ *A more robust validation of the Matérn-$5/2$ GP prior!*

Table of contents

Conclusion

- A robust uncertainty quantification methodology of GP surrogates with the help of CP can be deployed → better assessment of the metamodel quality for industrial studies
- GitHub Python module, implemented with MAPIE and OpenTURNS libraries
- Further work to include other kernel hyperparameters optimizations, including "nugget-effect", and extensions to other types of metamodels

Figure: ArXiV [2401.07733].

Figure: GitHub repository.

Thank you! Any question?

Table of contents

The Jackknife+ estimator

- $g : \mathcal{X} \subseteq \mathbb{R}^d \longrightarrow \mathcal{Y} \subseteq \mathbb{R}$
- $\mathcal{D}_n = \{(X_1, g(X_1)), \dots, (X_n, g(X_n))\}$ an i.i.d. design of experiments, X_{n+1} a new point
- $\widehat{g}$ a surrogate model trained on $\mathcal{D}_n$, $\widehat{g}_{-i}$ trained on $\mathcal{D}_n \backslash \{(X_i, g(X_i))\}$, and $\widehat{q}^{\pm}_{n,\alpha}(.)$ empirical α-quantile
- Leave-One-Out (LOO) error: $R_i^{LOO} := |g(X_i) - \widehat{g}_{-i}(X_i)|$

Definition

[Barber et al., 2021] The Jackknife+ estimator is given by:

$$\widehat{C}^{J+}_{n,\alpha}(X_{n+1}) = \left[\widehat{q}^{\pm}_{n,\alpha} \left\{ \widehat{g}_{-i}(X_{n+1}) \pm R_i^{LOO} \right\} \right]$$

- Coverage property verified only for $\alpha \in (0, 1/2)$
- Intervals have almost constant width for all points (including training points) $\rightarrow$ not that informative

Jacknife-minmax

- [Barber et al., 2021] Replace metamodel prediction with minimum (resp. maximum) of LOO error:

$$\widehat{C}_{n,\alpha}^{J-mm}(X_{n+1}) = \Bigg[\min_{i=1,\dots,n} \{\widehat{g}_{-i}(X_{n+1})\} - \widehat{q}_{n,\alpha}^{-}\left\{R_i^{LOO}\right\}, \max_{i=1,\dots,n} \{\widehat{g}_{-i}(X_{n+1})\} + \widehat{q}_{n,\alpha}^{+}\left\{R_i^{LOO}\right\}\Bigg]. \quad (8)$$

- Prediction intervals → not centered anymore but have *marginal coverage guarantee*:

$$\forall \alpha \in (0,1), \ \ \mathbb{P}\left(g(X_{n+1}) \in \widehat{C}_{n,\alpha}^{J-\mathrm{mm}}(X_{n+1})\right) \geq 1-\alpha. \quad (9)$$

- The resulting intervals will be more *conservative* i.e. with larger width.

Cross conformal predictors summary

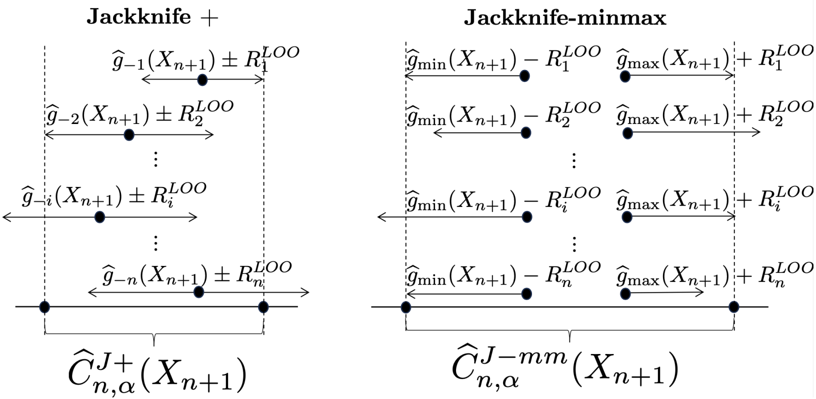


Figure: Jackknife+, and Jacknife-minmax schemes, adapted from [Barber et al., 2021].

The Burnaev-Wasserman program

[Burnaev and Vovk, 2014] Assume that $\mathcal{X} \subset \mathbb{R}^d$, for all i, $X_i \in L^2(\Omega)$ and the model g is truly Gaussian. The credibility sets have exact coverage and output an interval of the form :

$$\mathcal{CR}_\alpha(X_{n+1}) = [B_*, B^*]. \tag{10}$$

The CRR method with the GP rule outputs a prediction interval of the form:

$$\widehat{C}_{n,\alpha}^{CRR}(X_{n+1}) = [C_*, C^*]. \tag{11}$$

A natural question is to compare the differences of the bounds of these two intervals and their asymptotic behaviour

An asymptotic result

See [Burnaev and Vovk, 2014] for a proof of the following.

Theorem

Under the previous assumptions, we get:

$$\sqrt{n}\left(B^* - C^*\right) \xrightarrow[n\to\infty]{Law} \mathcal{N}\left(0, h(\alpha)\right), \qquad (12)$$

and similarly for the lower-bound.

Here h is a function of the $(1-\alpha/2)$-quantile of the standard normal distribution and of the mean and variance of the input distribution.

CPU dataset, $\alpha = 0.1$, $\beta = 1/2$

✓/ ✗: empirical coverage is / is not achieved.

ν	Q^2	$\mathcal{CR}_\alpha$	$\widehat{C}_{n,\alpha}^{J+GP}$	$\widehat{C}_{n,\alpha}^{J-mm-GP}$
1/2	0.845	91.6 ✓	28.14 ✓	47.3 ✓
3/2	0.856	72.9 ✓	29.9 ✓	47.1 ✓
5/2	0.854	70.9 ✓	32.1 ✓	49.0 ✓

Table: Average widths of prediction intervals and Q^2.

ν	Q^2	$\mathcal{CR}_\alpha$	$\widehat{C}_{n,\alpha}^{J+GP}$	$\widehat{C}_{n,\alpha}^{J-mm-GP}$
1/2	0.845	0.720 ✓	0.627 ✓	0.782 ✓
3/2	0.856	0.492 ✓	0.485 ✓	0.470 ✓
5/2	0.854	0.626 ✓	0.533 ✓	0.543 ✓

Table: Correlation between widths and GP approximation error and Q^2.

$\rightarrow$ *The metamodel with the lowest Q^2 is more robust to uncertainty!*

References I

Angelopoulos, A. N. and Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani., R. J. (2021). Predictive inference with the jackknife+. *Annals of Statistics, 486-507*, 49.

Burnaev, E. and Vovk, V. (2014). Efficiency of conformalized ridge regression. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 605–622.

De Rocquigny, E., Devictor, N., and Tarantola, S. (2008). *Uncertainty in industrial practice - A guide to quantitative uncertainty managment*. Wiley and Sons.

References II

Jaber, E., Blot, V., Brunel, N., Chabridon, V., Remy, E., Iooss, B., Lucor, D., Mougeot, M., and Leite, A. (2024a). Conformal approach to gaussian process surrogate evaluation with coverage guarantees.

Jaber, E., Chabridon, V., Remy, E., Baudin, M., Lucor, D., Mougeot, M., and Iooss, B. (2024b). Sensitivity analyses of a multi-physics long-term clogging model for steam generators.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer, springer edition.