Conformal prediction for surrogate modelling in the UQ framework

Edgar Jaber, Vincent Blot

ETICS 2023 - PhD Seminar

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

Sommaire

Problem formulation

- For a computer code $g : \mathcal{X} \to \mathcal{Y}$ used in industrial applications, the Uncertainty Quantification (UQ) methodology aims at evaluating how uncertainty on the inputs $X \in \mathcal{X}$ affects our knowledge of the output $g(X) \in \mathcal{Y}$ [De Rocquigny et al., 2008].

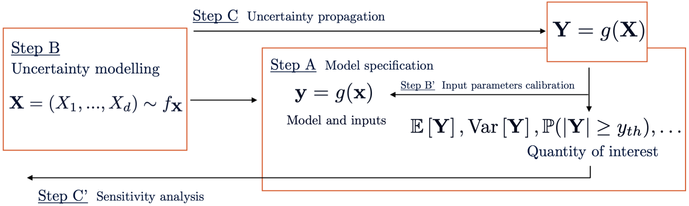


Figure – General UQ methodology.

Problem formulation

- Since these codes are time-costly, surrogates $\widetilde{g}$ are constructed for performing heavy simulations like Monte-Carlo batch runs for Step C, C'.
- For objective reasons, one needs to assess the *quality* of these surrogates.
- We propose to use the Conformal prediction paradigm [Vovk et al., 2005] which is a generic, model-agnostic theory allowing to build *prediction* sets to these surrogates with frequentist coverage guarantees.

Sommaire

CP in regression setting

[Vovk et al., 2005]. We fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition

Let $\mathcal{X}, \mathcal{Y}$ be metric spaces and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $n \in \mathbb{N}$ and $\mathcal{D} = \{Z_1, \ldots, Z_n\} \in 2^{\mathcal{Z}}$ a training sample. For $\alpha \in (0,1)$ a conformal predictor of coverage α is any measurable function of the form :

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

such that for a new point $Z_{n+1} = (X_{n+1}, Y_{n+1}) \in \mathcal{Z}$:

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

Three main methods exist to estimate these set-functions : full-conformal, split-conformal and cross-conformal estimators. We focus on the latter.

Full-Conformalized Ridge Regression (CRR)

[Vovk et al., 2005] A non-conformity score is any measurable function of the form :

$$\begin{aligned} A\colon 2^{\mathcal{Z}} \times \mathcal{Z} &\to \overline{\mathbb{R}} \\ (\mathcal{D}, Z) &\mapsto A(\mathcal{D}, Z). \end{aligned} \tag{3}$$

Assume $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and $R_{\mathcal{D}}$ a regression rule learned on $\mathcal{D} = \{Z_1, \ldots, Z_n\} = \{(X_1, Y_1), \ldots, (X_n, Y_n.)\}$. A straightforward non-conformity score is given by the difference between prediction and actual value, for all $i = 1, \ldots, n$:

$$A(\mathcal{D}, Z_i) = Y_i - R_{\mathcal{D}}(X_i) = Y_i - \widehat{Y}_i \tag{4}$$

Full-Conformalized Ridge Regression (CRR)

For a new input $X_{n+1} \in \mathcal{X}$ and output $Y \in \mathcal{Y}$, we denote by $\widetilde{\mathcal{D}} := \mathcal{D} \sqcup (X_{n+1}, Y)$. The full conformal predictor is given for all confidence level $0 < \alpha < 1$ by :

$$\widehat{C}_{n,\alpha}^{CRR}(X_{n+1}) = \left\{ Y \in \mathcal{Y} : \frac{\#\{i : A(\widetilde{\mathcal{D}}, Z_i) \geq A(\widetilde{\mathcal{D}}, (X_{n+1}, Y))\}}{n} > \alpha \right\}.$$

- This estimator is based on the "dissimilarity" of the new pair inside the updated dataset. It achieves the required coverage property.
- It is computationally intractable due to the full grid search on the label space $\mathcal{Y}$.
- To tackle this, more advanced estimators are proposed in the following.
- In the rest of this talk we suppose that $Y = g(X)$ where g is some deterministic function (e.g a numerical simulation code).

The Jackknife+ estimator

[Barber et al., 2021]. We train a metamodel $\widehat{g}$ on $\mathcal{D}$ and n leave-one-out metamodels $\widehat{g}_{-i}$ on $\mathcal{D}\backslash(X_i, g(X_i))$. The Jackknife+ estimator is given by :

$$\widehat{C}^{J+}_{n,\alpha}(X_{n+1}) = \left[q^{-}_{n,\alpha}\left\{\widehat{g}_{-i}(X_{n+1}) - R_i^{LOO}\right\}, q^{+}_{n,\alpha}\left\{\widehat{g}_{-i}(X_{n+1}) + R_i^{LOO}\right\}\right] \tag{5}$$

Where the leave-one-out error defined by :

$$R_i^{LOO} := |g(X_i) - \widehat{g}_{-i}(X_i)| \tag{6}$$

- With this estimator we have a coverage guarantee of $1 - 2\alpha$.
- However, these intervals have almost constant width for any new point.

CV+ estimator

- Let $K \in \{1, \ldots, n\}$. We divide $\mathcal{D}^{\text{train}}$ in K-disjoint subsets $S_1 \cup \ldots \cup S_K$.
- We fit K metamodels with the k-th fold removed : $\widehat{g}_{-S_k}$.
- We compute the conformity scores :

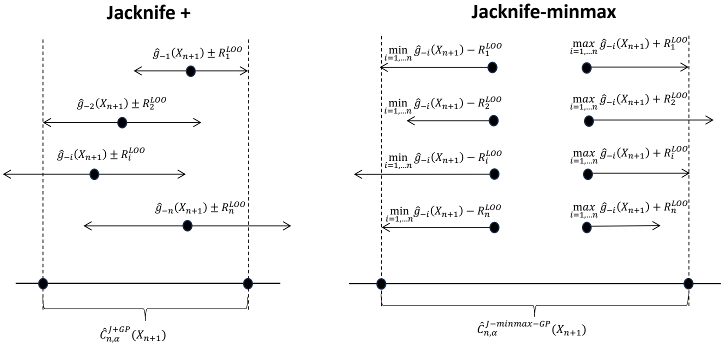
$$R_i^{CV} := |g(X_i) - \widehat{g}_{-S_{k(i)}}(X_i)|, \quad \forall i \in \{1, \ldots, n\}, \tag{7}$$

where $S_{k(i)}$ is the fold containing X_i.

- We estimate our prediction intervals :

$$\widehat{C}_{n,\alpha}^{\text{CV+}}(X_{n+1}) = \left[\widehat{q}_{n,\alpha}^{\pm} \left\{\widehat{g}_{-S_{k(i)}}(X_{n+1}) \pm R_i^{CV}\right\}\right] \tag{8}$$

J+/minmax



$$with\ R_i^{LOO} = |Y_i - \hat{g}_{-i}(X_i)|$$

Jacknife/CV minmax

Additionally, we can add minmax on both the J+ and the CV+ methods by replacing the metamodel prediction with :

$$\widehat{g}_{-i}(X_{n+1}) \longleftrightarrow \min_i \widehat{g}_{-i}(X_{n+1}), \max_i \widehat{g}_{-i}(X_{n+1}). \tag{9}$$

$$\widehat{g}_{-S_k(i)}(X_{n+1}) \longleftrightarrow \min_i \widehat{g}_{-S_k(i)}(X_{n+1}), \max_i \widehat{g}_{-S_k(i)}(X_{n+1}). \tag{10}$$

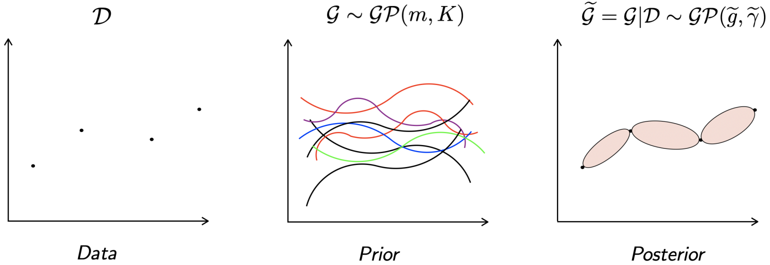
The intervals will not be centered anymore and we have the coverage guarantee [Barber et al., 2021] :

$$\mathbb{P}\left(g(X_{n+1}) \in \widehat{C}^{*+\text{-mm}}_{n,\alpha}(X_{n+1})\right) \geq 1 - \alpha. \tag{11}$$

Moreover, the resulting intervals will be more conservative.

Sommaire

Gaussian Process Regression (GPR)



Bayesian credibility intervals

Since we are in a Gaussian setting, we can have access to the confidence intervals for any new point $X_{n+1} \in \mathcal{X} \backslash \boldsymbol{X}$:

$$\mathcal{CR}_\alpha(X_{n+1}) = \left[\widetilde{g}(X_{n+1}) \pm F^{-1}\left(1 - \frac{\alpha}{2}\right)\widetilde{\gamma}(X_{n+1})\right] \tag{12}$$

which in our case are credibility intervals. Here F is the CDF of the standard normal distribution.

If g was *truly* drawn from our posterior $\mathcal{G}|(\boldsymbol{X}, g(\boldsymbol{X}))$, then we would have the exact coverage :

$$\mathbb{P}\left(g(X_{n+1}) \in \mathcal{CR}_\alpha(X_{n+1})\right) = 1 - \alpha, \tag{13}$$

and $\mathcal{CR}_\alpha(X_{n+1})$ would be the true prediction interval for the true g.

Frequentist prediction intervals

In practice we don't have access to the true distribution $\mathbb{P}_{(X,g(X))}$. If we dispose of a test dataset :

$$\mathcal{D}^{test} = \{(X_1, g(X_1)), \ldots, (X_m, g(X_m))\}, \tag{14}$$

different from $\mathcal{D}$, then in all generality we cannot expect to have :

$$\frac{1}{m}\sum_{i=1}^{m} 1\{g(X_i) \in \mathcal{CR}_\alpha(X_i)\} \geq 1-\alpha, \quad \text{a.s.} \tag{15}$$

Sommaire

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 :

$$\mathcal{CR}_\alpha(X_{n+1}) = [B_*, B^*] \tag{16}$$

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

$$\widehat{C}^{CRR}_{n,\alpha}(X_{n+1}) = [C_*, C^*]. \tag{17}$$

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

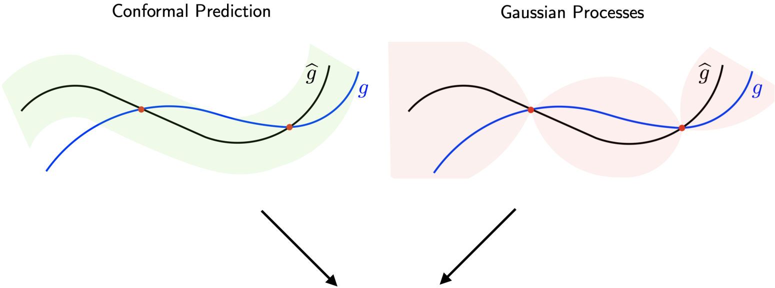
Under the above assumptions we get

$$\sqrt{n}\left(B^* - C^*\right) \xrightarrow[n\to\infty]{Law} \mathcal{N}\left(0, h(\alpha)\right), \tag{18}$$

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.

Problem formulation



Adaptive Conformal Prediction for better Uncertainty Quantification

$$\mathbb{P}\left(g(X) \in C_{\widehat{g},\alpha}(X)\right) \geq 1-\alpha. \tag{19}$$

J+GP

The idea is to adapt the Jackknife+ method presented earlier to GPR metamodels to have adaptive prediction intervals. We denote these predictors as $\widehat{C}_{n,\alpha}^{J+GP}$ and can prove the following theorem :

Theorem

Assume $\mathcal{D} = (\boldsymbol{X}, g(\boldsymbol{X}))$ is exchangeable. For a new point $X_{n+1} \in \mathcal{X} \backslash \boldsymbol{X}$ and a coverage level $\alpha \in (0,1)$:

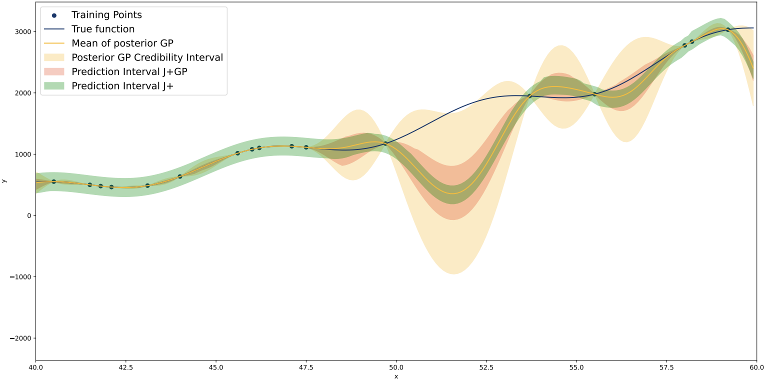
$$\mathbb{P}\left(g(X_{n+1}) \in \widehat{C}_{n,\alpha}^{J+GP}(X_{n+1})\right) \geq 1 - 2\alpha. \tag{20}$$

Example on a toy function

- The Vincent Blot© highly original g_{VB} :

$$g_{VB}(x) = 3x\sin(x) - 2x\cos(x) + \frac{x^3}{40} - \frac{1}{2}x^2 - 10x, \quad (21)$$

which we sample at ease.



Sommaire

Conclusion

- Conformal prediction is a method for performing distribution-free uncertainty quantification of machine learning algorithms.
- It can be used in a Bayesian regression setting for testing the soft-assumption on the original model.
- Prediction intervals for Gaussian Processes using an adaptation of the CP algorithms can be built (upcoming paper [Jaber and Blot, 2023] - in progress).
- A more robust uncertainty quantification of this type of surrogates can be achieved and can thus serve in assessing their quality for the purpose of performing industrial UQ.

References I

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.

Jaber, E. and Blot, V. (2023). Conformalizing Gaussian Processes For More Robust Uncertainty Quantification. *In progress*.

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