CEMRACS 2023 mini-project proposal

Title: Conformal prediction for data-driven and physics-based prognostics algorithms

Main contacts: Vincent CHABRIDON(*) (vincent.chabridon@edf.fr - 01.30.87.71.19), Nicolas BRUNEL (nbrunel@quantmetry.com)

Autres contacts : Emmanuel REMY(*) (emmanuel.remy@edf.fr), Bertrand IOOSS(*) (bertrand.iooss@edf.fr), Michaël BAUDIN(*) (michael.baudin@edf.fr), Mathilde MOUGEOT(**) (mathilde.mougeot@ensiie.fr) and Didier LUCOR(***) (didier.lucor@lisn.upsaclay.fr)

Institution or company: (*) EDF R&D, (**) ENS-Paris Saclay and Centre Borelli, (***) CNRS and LIMSI

1. Context

In the context of risk assessment for an industrial critical system, various sources of information can be available to predict its behavior regarding a failure criterion or a remaining-useful life: data from laboratory experiments or from the system monitoring, which may be used to feed "machine learning" algorithms, and/or high-fidelity numerical simulations from physics-based models [1]. All of them embed their own uncertainties which are either intrinsic/irreducible (aleatory), or possibly reducible (epistemic) by adding more knowledge. In the context of numerical simulation, uncertainty quantification relies on a rather established and well-accepted mathematical framework with a large panel of tools (e.g., the use of surrogate models / metamodels such as Gaussian process regression or polynomial chaos expansions to emulate costly-to-evaluate simulators) [2]. Building data-driven predictive models using machine learning algorithms inevitably carries uncertainties from both data and model assumptions to the predictions. Therefore, one needs to set up a rational framework to handle these uncertainties in this statistical learning context. A possible solution is to use the "conformal prediction" paradigm [3,4].

2. Description and objectives

This work aims at investigating this paradigm in the context of data-driven prognostics since important decisions have to be made with respect to predictions. Thus, one needs to propose some metrics that enable to derive statistical guarantees for the decision-maker. Moreover, close links can be highlighted between conformal prediction and the validation of surrogate models in traditional uncertainty quantification for simulators (e.g., in Gaussian process regression, see [5,6,7]). This work will try to clearly exhibit and extend some of these links.

3. Proposed methodology

More precisely, a preliminary work program could be the following one:

- Propose a bibliography review about conformal prediction, its theoretical principles and a clear overview of the main challenges, in practice (starting by [3,4]).
- Perform a benchmark on a panel of datasets (both from computer experiments and real noisy data) in order to test several machine learning algorithms (linear and penalized regressions, random forests, neural networks) together with conformal prediction [8]. As a first assumption, one will focus on small to medium input dimensions problems (e.g., from 5 to 30 inputs), and to scalar and/or one dimensional field output variables (e.g., time series output).
- Provide an analysis with respect to the results one can obtain from Gaussian process regression (using validation criteria and robust prediction intervals as discussed in [5,6,7]).
- Apply the proposed methodology to both fictitious and industrial data.

4. Software requirements

As much as possible, the numerical developments will be implemented using Python, and will benefit from the existing libraries such as "Awesome Conformal Prediction" (https://github.com/valeman/awesome-conformal-prediction), "MAPIE - Model Agnostic Prediction Interval Estimator" (https://mapie.readthedocs.io/en/latest/index.html) or "Conformal Prediction" (https://github.com/aangelopoulos/conformal-prediction).
Regarding the validation of metamodels in uncertainty quantification framework, the implementation will mostly rely on the "OpenTURNS" software (https://openturns.github.io/www/).

5. Students funded for the project

- EDF R&D will sponsor Edgar Jaber (edgar.jaber@edf.fr), PhD student at EDF R&D and Université Paris-Saclay;
- Quantmetry will sponsor Vincent Blot (vblot@quantmetry.com), PhD student at Quantmetry and Université Paris-Saclay.

6. References

[1] Eker, O. F., Camci, F. and Jennions, I. K. (2012). "Major Challenges in Prognostics: Study on Benchmarking Prognostics Datasets". In: Proceedings of the First European Conference of the PHM Society 2012. Dresden, Germany.
[2] Le Gratiet, L., Marelli, S. and Sudret, B. (2017). "Metamodel-Based Sensitivity Analysis: Polynomial Chaos Expansions and Gaussian Processes". In: R. Ghanem, D. Higdon, H. Owhadi (Eds.), Handbook of Uncertainty Quantification, Springer International Publishing, pp. 1289-1325.
[3] Angelopoulos, A. N. and Bates, S. (2021). "A gentle introduction to conformal prediction and distribution-free uncertainty quantification". In: arXiv preprint, arXiv:2107.07511.
[4] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. and Wasserman, L. (2018). "Distribution-Free Predictive Inference for Regression". In: Journal of the American Statistical Association, 113:523, pp. 1094-1111.
[5] Acharki, N., Bertoncello, A. and Garnier, J. (2023). "Robust prediction interval estimation for Gaussian processes by cross-validation method". In: Computational Statistics and Data Analysis, 178, 107597.
[6] Demay, C., Iooss, B., Le Gratiet, L. and Marrel, A. (2021), "Model selection based on validation criteria for Gaussian process regression: An application with highlights on the predictive variance". In: *Quality and Reliability Engineering International,* 38, pp. 1482-1500.
[7] Wieskotten, M., Crozet, M., Iooss, B., Lacaux,C. and Marrel, A. (2022). "A comparison between Bayesian and ordinary kriging based on validation criteria: application to radiological characterisation". In: HAL preprint, hal-03806713.
[8] Mentch, L. and Hooker, G. (2016). "Quantifying uncertainty in random forests via confidence intervals and hypothesis tests," In: Journal of Machine Learning Research, 17, pp. 1-41.