Adaptive Conformal Prediction using Kernel Methods

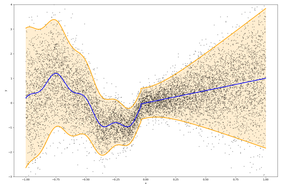
Louis Allain [1,2] Sébastien Da Veiga [1] Brian Staber [2]

[1]ENSAI-CREST [2]Safran Tech

Context

This work is part of an effort from Safran to enhance the conception of aero-spatial pieces, using machine learning and surrogate modeling. Our goal is to develop better uncertainty quantification for the prediction of those surrogate models. This uncertainty quantification must have the following characteristics:

- Marginal coverage
- Finite-sample coverage
- Distribution free
- Adaptive

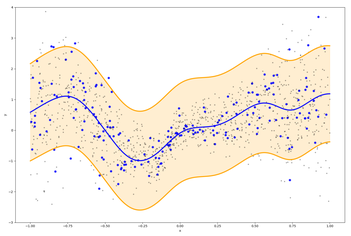


Example: ground truth

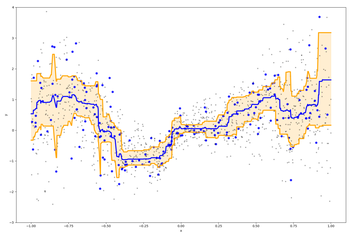
Usual models

	Marginal coverage	Finite-sample coverage	Distribution free	Adaptive
GP	✓	✗	✗	✓
RF	✓	✗	✗	✓
BNN	✗	✗	✗	✓

All those models construct very different prediction bands.



Gaussian Process



Random Forest

How can we construct prediction bands that are reliable, provide marginal coverage guarantee in finite sample and is distribution free?

Split conformal prediction

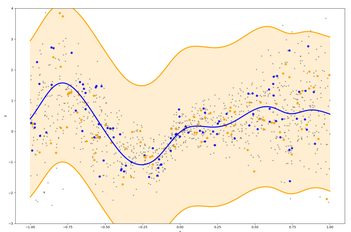
Conformal Prediction (CP) is an easy-to-use framework to learn prediction bands for machine learning (ML) models. The split CP [8] variant works as follows:

1. Train the surrogate model $\widehat{m}$ on the training dataset $\mathcal{D}_n$
2. Compute the model's errors on a holdout dataset: $|y_i - \widehat{m}(x_i)|, i \in \mathcal{D}_m$
3. Compute the α-quantile of those errors $\widehat{q}_\alpha$ (α: specified error-rate)
4. Prediction intervals are then given by $[\widehat{m}(X_{\text{test}}) \pm \widehat{q}_\alpha]$

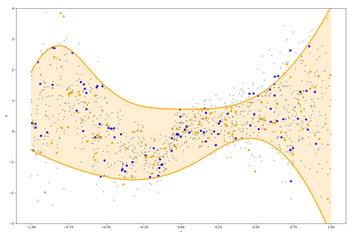
	Marginal coverage	Finite-sample coverage	Distribution free	Adaptive
CP	✓	✓	✓	✗

Conformal prediction imposes itself as the framework for easy and reliable prediction bands as long as one can make those bands adaptive.

More adaptive variants of split CP have been proposed:



Rescaled Score SCP Gaussian Process



CQR using splines

The quest for adaptivity

Improving adaptivity is crucial, and many works have addressed this challenge.

	CP	Adaptive	Small data	Big data	HP tuning	Goal oriented
CQR/OQR [9, 3]	✓	✓	✗	✓	∼	∼
RLCP [4]	✓	✓	✗	✓	✗	✓
Rescale [5]	✓	✓	✓	✓	✓	∼
SDP [6, 2]	✗	✓	✓	✗	✗	✓
Our	✓	✓	✓	✓	✓	✓

Problem setup

We propose a fundamentally different approach, inspired by [6, 2]:

- Cast the CP problem as a statistical learning problem. Goal: learning the best rescaled score
- Prediction bands and mean function are the solution of an optimization problem in RKHSs of functions $\mathcal{H}^f$ and $\mathcal{H}^m$:

$$\inf_{m\in\mathcal{H}^m,\ \mathcal{A}\in\mathcal{S}_+(\mathcal{H}^f)} \ \frac{a}{n}\sum_{i=1}^n (y_i - m(x_i))^2 + \frac{b}{n}\sum_{i=1}^n f_{\mathcal{A}}(x_i) + \lambda_1\|\mathcal{A}\|_\star + \lambda_2\|\mathcal{A}\|_F^2$$
$$\text{s.t.} \quad f_{\mathcal{A}}(x_i) \geq (y_i - m(x_i))^2,\ i \in [n],$$
$$\|m\|_{\mathcal{H}^m}^2 \leq s.$$

$$\left[\widehat{m}(X_{\text{test}}) \pm \sqrt{\widehat{q}_\alpha \widehat{f}_{\mathcal{A}}(X_{\text{test}})}\right] \quad \text{where } \widehat{q}_\alpha \text{ is computed on the errors } \frac{\left(y_i - \widehat{m}(x_i)\right)^2}{\widehat{f}_{\mathcal{A}}(x_i)},\ i \in \mathcal{D}_m$$

Function $f_{\mathcal{A}}$ can be looked at as similar to the conditional variance, thus must be *positive*.

Building on [7] we use the sum-of-squares framework to estimate this positive function. Similarly to the finite dimensional case, a sum-of-square function is positive if:

$$\forall u,\ u^T M u \geq 0 \iff M \succeq 0 \text{ (finite)} \quad | \quad \forall x,\ \phi(x)^T \mathcal{A} \phi(x) \geq 0 \iff \mathcal{A} \succeq 0 \text{ (infinite)}$$

Our positive function f is thus parametrized by a Hilbert space operator $\mathcal{A} \in \mathcal{S}_+\left(\mathcal{H}^f\right)$ and a feature map ϕ:

$$f(x) = \phi(x)^T \mathcal{A} \phi(x),\ \mathcal{A} \succeq 0$$

(ϕ: feature map associated to $\mathcal{H}^f$; $\mathcal{A}$: operator to learn)

Important features for this function are:

1. Minimize mean width
2. Smoothness, controlled through its norm
3. Learn well the variability, controlled through the constraints

From infinite problem to a tractable solution

This problem is infinite dimensional, indeed m is a function and $\mathcal{A}$ an operator.

Infinite problem (∞) $\Longrightarrow$ Representer theorem — Finite Semi-Definite problem ($n+n\times n$) $\Longrightarrow$ Lagrangian — Dual formulation ($n+1$)

The Lagrangian writes:

$$\mathcal{L} = \sum_{i=1}^n \left(\Gamma_i + \frac{a}{n}\right)(y_i - m(x_i))^2 + \theta\left(\|m\|_{\mathcal{H}^m}^2 - s\right) + \lambda_1\|\mathbf{A}\|_\star + \lambda_2\|\mathbf{A}\|_F^2 - \sum_{i=1}^n \left(\Gamma_i - \frac{b}{n}\right)\tilde{f}_{\mathbf{A}}(x_i)$$

The dual formulation writes:

$$D(\mathbf{\Gamma}, \theta) = \mathbf{r}(\mathbf{\Gamma}, \theta)^T \text{Diag}(\mathbf{\Gamma}_a)\mathbf{r}(\mathbf{\Gamma}, \theta) + \theta(\boldsymbol{\gamma}(\mathbf{\Gamma}, \theta)^T \mathbf{K}^m \boldsymbol{\gamma}(\mathbf{\Gamma}, \theta) - s) - \Omega^\star(\mathbf{V}\text{Diag}(\mathbf{\Gamma}_{-b})\mathbf{V}^T)$$

with

$$\mathbf{r}(\mathbf{\Gamma}, \theta) := \mathbf{y} - \mathbf{K}^m \boldsymbol{\gamma}(\mathbf{\Gamma}, \theta) \qquad \boldsymbol{\gamma}(\mathbf{\Gamma}, \theta) := \boldsymbol{C}(\mathbf{\Gamma}, \theta)^{-1}\text{Diag}(\mathbf{\Gamma}_a)\mathbf{y}$$
$$\boldsymbol{C}(\mathbf{\Gamma}, \theta) := \text{Diag}(\mathbf{\Gamma}_a)\mathbf{K}^m + \theta\mathbf{I}_n \qquad \Omega^\star(\mathbf{B}) := \frac{1}{4\lambda_2}\|[\mathbf{B} - \lambda_1\mathbf{I}_n]_+\|_F^2$$

We propose a new numerically efficient gradient ascent algorithm with Nesterov Acceleration which can scale up to several thousands samples, unlike previous SDP approaches that are limited to around 200 samples

Hyper-parameters selection

All model's hyper-parameters						
Name	MSE	MW	m-lengthscale	m-regularization	$\mathcal{A}$-lengthscale	$\mathcal{A}$-regularization
Symbol	a	b	θ^m	s	θ^f	λ_1, λ_2
Solution	-	-	GP	GP	HSIC	- -

For m hyper-parameters, we propose to fit a Gaussian Process on training data $\mathcal{D}_n$ and choose:

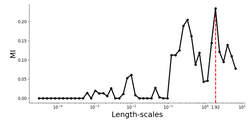
$$\theta^m = \theta^{\text{GP}} \qquad \text{and} \qquad s = \boldsymbol{\gamma}_{\text{GP}}^T \mathbf{K}^m \boldsymbol{\gamma}_{\text{GP}}$$

We now present our novel solution to tune the length-scale θ^f associated to the operator $\mathcal{A}$.

Choosing the length-scale to improve adaptivity

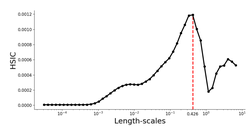
- Our goal: choosing θ^f to target adaptive prediction bands
- Better adaptivity may be attained by targeting conditional coverage
- [1] shows that minimizing $\text{MI}(X, s)$ leads closer to conditional coverage

$$\mathbb{P}\left(Y \in \widehat{C}_{\mathcal{D}}(X) \mid X \in \omega_X\right) \geq 1 - \alpha - \frac{\sqrt{1 - e^{-\text{MI}(X,s)}}}{\delta}$$



- We can show that minimizing the *multi-dimensional* $\text{MI}(X, s)$ can be achieved by maximizing the *one-dimensional* $\text{MI}(f, (y-m)^2)$
- Robust MI estimation is hard to achieve, we propose instead to rely on HSIC
- In practice the estimation is more robust, and we have also shown that a similar conditional guarantee holds:

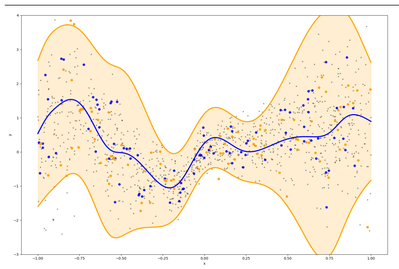
$$\mathbb{P}\left(Y \in \widehat{C}_{\mathcal{D}}(X) \mid X \in \omega_X\right) \geq 1 - \alpha - \frac{\sqrt{1 - \frac{\alpha_2}{1-\alpha_1\text{HSIC}}}}{\delta}$$



- Thus our proposition is the following:

$$\max_{\theta^f \in \mathbb{R}^d} \ \text{HSIC}\left(\widehat{f}(x_i), \left(y_i - \widehat{m}(x_i)\right)^2\right),\ i \in \mathcal{D}_n, \quad \text{with} \quad k_E(x, y) = \frac{1}{2}\left(\|x\| + \|y\| - \|x - y\|\right)$$

Results



Our method

Summary of contributions:

- Statistical learning framework for best rescaled score
- Efficient dual formulation, solvable using gradient ascent
- Improving adaptiveness by choosing the right length-scale through HSIC metric

References

[1] Nicolas Deutschmann, Mattia Rigotti, and Maria Rodriguez Martinez. Adaptive conformal regression with jackknife+ rescaled scores, 2023. URL https://arxiv.org/abs/2305.19901.

[2] Jianqing Fan, Jiawei Ge, and Debarghya Mukherjee. UTOPIA: Universally Trainable Optimal Prediction Intervals Aggregation, July 2024. URL http://arxiv.org/abs/2306.16549.

[3] Shai Feldman, Stephen Bates, and Yaniv Romano. Improving conditional coverage via orthogonal quantile regression, 2021. URL https://arxiv.org/abs/2106.00394.

[4] Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables local guarantees, October 2024. URL http://arxiv.org/abs/2310.07850.

[5] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression, 2017. URL https://arxiv.org/abs/1604.04173.

[6] Tengyuan Liang. Universal Prediction Band via Semi-Definite Programming, January 2023. URL http://arxiv.org/abs/2103.17203.

[7] Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Non-parametric Models for Non-negative Functions, July 2020. URL http://arxiv.org/abs/2007.03926.

[8] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Machine Learning: ECML 2002*, pages 345–356, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[9] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized Quantile Regression, May 2019. URL http://arxiv.org/abs/1905.03222.

louis.allain@safrangroup.com