Scalable and adaptive prediction bands with kernel sum-of-squares

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

[1]Safran Tech, Digital Sciences & Technologies, 78114 Magny-Les-Hameaux, France
[2]Univ Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France
{louis.allain,brian.staber}@safrangroup.com
sebastien.da-veiga@ensai.fr

Prediction intervals

- In critical applications, we need confidence intervals around machine learning predictions with coverage guarantees:
  - The guarantee of **marginal coverage** at level α writes

$$\mathbb{P}(Y_{N+1} \in \widehat{C}(X_{N+1})) \geq 1 - \alpha$$

  for the true unknown value of the output Y_{N+1} at an unobserved point X_{N+1}

- Conformal prediction provides a great framework to target this problem

Conformal prediction

- Conformal prediction (CP): a rigorous method to construct prediction intervals with the following properties:
  - ✓ Marginal coverage
  - ✓ Finite sample
  - ✓ Distribution free
  - ✓ Model agnostic

- We focus on split CP, based on two independent datasets, a pre-training $\mathcal{D}_n$ and a calibration $\mathcal{D}_m$

- CP relies on a score function to evaluate the predictive quality of the model and adjusts the prediction bands accordingly

Score functions

	Absolute errors	Quantile regression[1]	Normalization[2]
$s(X_i, Y_i)$	$\lvert Y_i - \widehat{m}(X_i) \rvert$	$\max(\widehat{q}_{\mathrm{l}}(X_i) - Y_i, Y_i - \widehat{q}_{\mathrm{u}}(X_i))$	$\frac{(Y_i - \widehat{m}(X_i))^2}{\widehat{f}(X_i)}$

We propose here to *learn* a normalized score function in a way that targets both adaptivity and coverage

[1] [Romano et al. 2019]

[2] [Lei et al. 2014; Johansson et al. 2014; Papadopoulos 2024; Jaber et al. 2024]

Learning problem for a score function

- We consider a normalized score: $\frac{(Y-m(X))^2}{f(X)}$, with $\boldsymbol{f \geq 0}$
- As for all learning problems, we must first choose a search space for our functions, here we rely on **kernel methods**
  - m lives in the Reproducible Kernel Hilbert Space (RKHS) $\mathcal{H}^m$ with kernel k^m and lengthscales θ^m
  - f is a ***kernel sum-of-squares*** function parameterized by a semi-definite operator $\mathcal{A}$. This will impose its **positivity**

Learning the score function amounts to simultaneously learning

$$m \in \mathcal{H}^m,\ f \in \mathcal{S}\text{o}\mathcal{S}(\mathcal{H}^f) \quad \Leftrightarrow \quad m \in \mathcal{H}^m,\ \mathcal{A} \in \mathcal{S}_+(\mathcal{H}^f)$$

Learning problem for a score function

$$\inf_{m \in \mathcal{H}^m,\, \mathcal{A} \in \mathcal{S}_+(\mathcal{H}^f)} \quad \frac{a}{n}\sum_{i=1}^{n}(Y_i - m(X_i))^2 + \frac{b}{n}\sum_{i=1}^{n} f_{\mathcal{A}}(X_i) + \lambda_1 \|\mathcal{A}\|_\star + \lambda_2 \|\mathcal{A}\|_F^2$$

$$\text{s.t.} \quad f_{\mathcal{A}}(X_i) \geq (Y_i - m(X_i))^2,\ i \in [n],$$

$$\|m\|_{\mathcal{H}^m}^2 \leq s$$

Learning problem for a score function

$$\inf_{m\in\mathcal{H}^m,\,\mathcal{A}\in\mathcal{S}_+(\mathcal{H}^f)} \quad \frac{a}{n}\sum_{i=1}^{n}(Y_i - m(X_i))^2 + \frac{b}{n}\sum_{i=1}^{n} f_{\mathcal{A}}(X_i) + \lambda_1\|\mathcal{A}\|_\star + \lambda_2\|\mathcal{A}\|_F^2$$

$$\text{s.t.} \quad f_{\mathcal{A}}(X_i) \geq (Y_i - m(X_i))^2,\ i\in[n],$$

$$\|m\|_{\mathcal{H}^m}^2 \leq s$$

i) Faithful estimation of the mean function

Learning problem for a score function

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

i) Faithful estimation of the mean function
ii) 100% coverage on the training sample - **convex** constraint (later adjusted with split CP)

Learning problem for a score function

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

i) Faithful estimation of the mean function
ii) 100% coverage on the training sample - **convex** constraint (later adjusted with split CP)
iii) Minimization of the interval mean width

Learning problem for a score function

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

i) Faithful estimation of the mean function

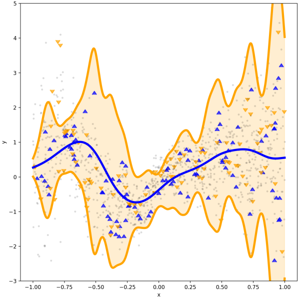
ii) 100% coverage on the training sample - **convex** constraint (later adjusted with split CP)

iii) Minimization of the interval mean width

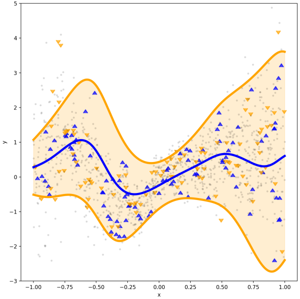
iv) Control of the regularity of the bands
- lasso-type norm $\|\mathcal{A}\|_\star$
- ridge-type norm $\|\mathcal{A}\|_F$

Representer theorem

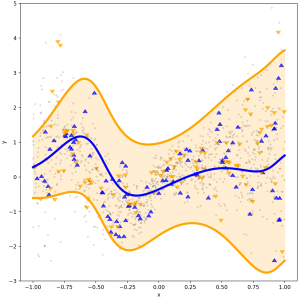
- We prove a representer theorem for this infinite dimensional problem
- It becomes a Semi-Definite Program (SDP) problem, solvable using off-the-shelves solvers



(a) $\theta^f = 0.1$



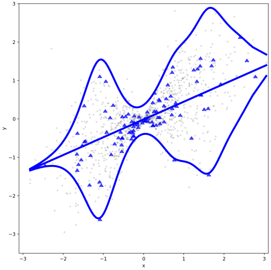
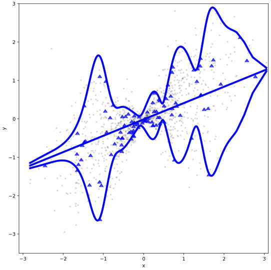
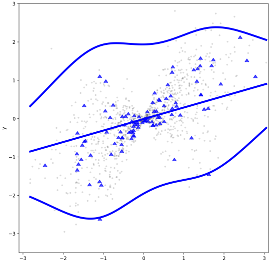
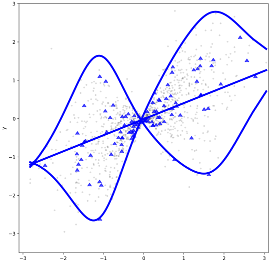
(b) $\theta^f = 0.5$



(c) $\theta^f = 0.9$

Note: θ^f is the vector of lengthscales for k^f, the kernel corresponding to $\mathcal{H}^f$

Hyperparameter tuning

	$b = 1$	$b = 100$
$\theta^f = 0.5$		
$\theta^f = 1.74$		

Before calibration

Hyperparameter tuning

- Perfectly adaptive bands guarantee local coverage

$$\mathbb{P}(Y_{N+1} \in \widehat{C}(X_{N+1}) \mid X_{N+1} = x) \geq 1 - \alpha$$

- Without hypothesis on the data, satisfying this local coverage leads to infinitely wide prediction bands [Vovk 2012; Barber et al. 2021]
- We can relax the local coverage by considering X in a small neighbourhood ω_X, such that $\forall x \in \mathcal{X},\ \ \mathbb{P}(x \in \omega_X) \geq \delta$:

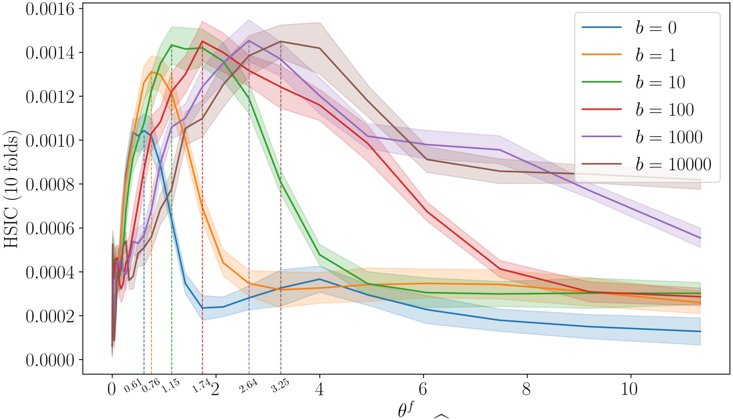
$$p_{\mathcal{D}_N} := \mathbb{P}(Y_{N+1} \in \widehat{C}(X_{N+1}) | X_{N+1} \in \omega_X) \geq 1 - \alpha$$

Hyperparameter tuning

- Using information theory results and recent inequalities result between the TV distance and the MMD, we prove a new bound

$$p_{\mathcal{D}_N} \geq 1 - \alpha - \frac{1}{\delta}\sqrt{1 - \frac{\alpha_1}{1 - \alpha_2 \mathrm{HSIC}(r_{\mathcal{D}_n}(X_{N+1}, Y_{N+1}), \widehat{f}_{\theta^f}(X_{N+1}))}}$$

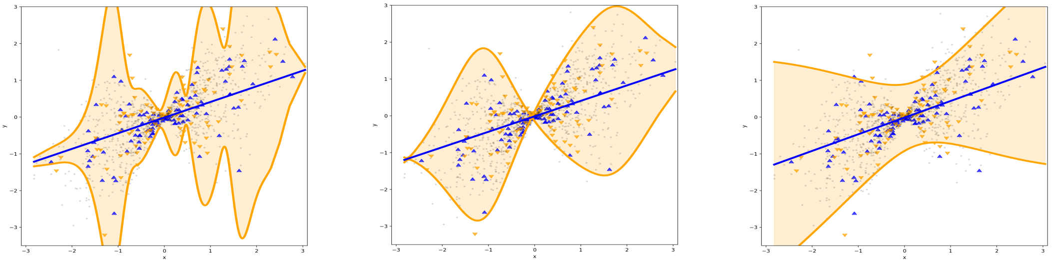
Hyperparameter tuning



$$\mathrm{HSIC}(r_{\mathcal{D}_n}(X, Y), \widehat{f}_{\theta^f}(X))$$

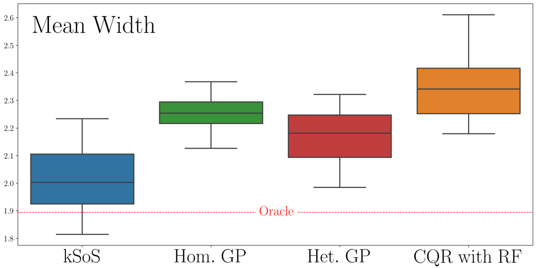
Maximizing this HSIC, i.e. the dependence between the residuals and the interval widths, allows to target better local coverage

Analytical case



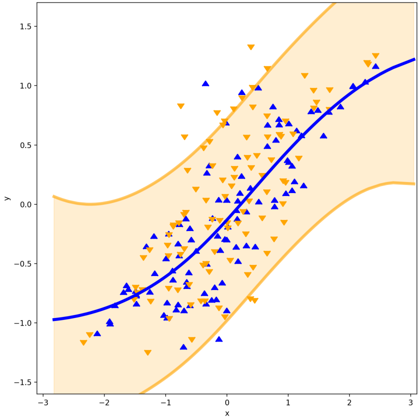
(a) $\theta^f = 0.5$ (b) $\boldsymbol{\theta^f_{\mathrm{HSIC}}} = 1.74$ (c) $\theta^f = 10$

Mean width metric

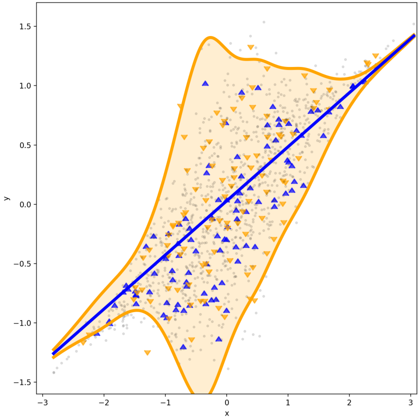


- A common measure for adaptive prediction bands in the literature is **mean width**, which should be minimized
- kSoS leads to better or as good mean width as competitors
- **However, mean width does not always tell the full story**

Mean width metric

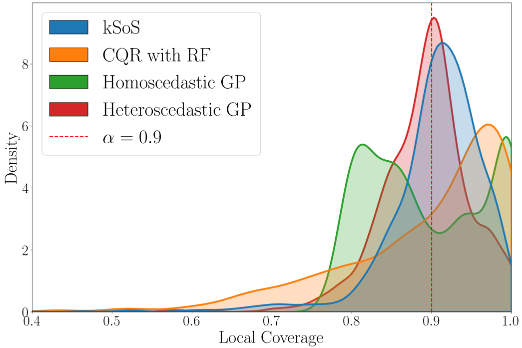


(a) Homoscedastic GP
MW = 1.712



(b) kSoS with Opt. HSIC
MW = 1.759

Local coverage metric



- The best measure of adaptivity is **local coverage**
- The target for local coverage is a Dirac at $1 - \alpha = 0.9$
- kSoS leads to better concentrated local coverage in general

Contributions

- Learning setting for a score function in the context of split CP
- Representer theorem to make the problem tractable
- Dual formulation with AGD to handle thousands of points
- Brand new adaptivity measure based on HSIC, that allows to automatically choose hyperparameters of the model