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

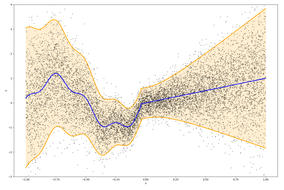
Louis Allain [1,2] Sébastien Da Veiga [1] Brian Staber [2]

[1]Univ Rennes, Ensai, CNRS, CREST [2]Safran Tech, Digital Sciences & Technologies

Context

In many applications, machine learning regression models require a trustworthy uncertainty quantification in their predictions.
This is especially true for high-stakes applications, in which we aim for prediction bands with the following properties:

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



Example: ground truth

Split conformal prediction

Conformal Prediction (CP) is an easy-to-use framework to learn prediction bands for machine learning (ML) models. The split CP [4] 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 $\widehat{C}(X_{\text{test}}) = [\widehat{m}(X_{\text{test}}) \pm \widehat{q}_\alpha]$

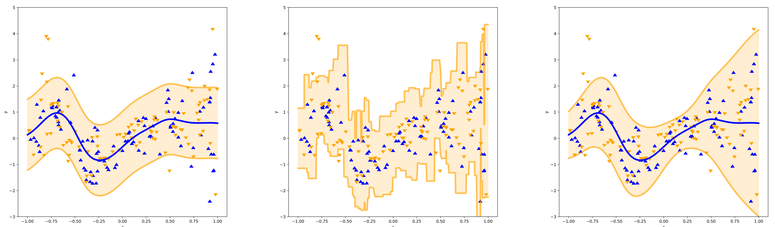
	Marginal coverage	Finite-sample coverage	Distribution free	Adaptive
CP	✓	✓	✓	✗

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

Adaptivity can be induced by changing the score function.

Score functions in CP

Absolute errors	Quantile regression	Normalization
$s(X_i, Y_i) \quad \lvert Y_i - \widehat{m}(X_i)\rvert$	$\max(\widehat{q}_l(X_i) - Y_i, Y_i - \widehat{q}_u(X_i))$	$\frac{(Y_i - \widehat{m}(X_i))^2}{\widehat{f}(X_i)}$



Absolute errors
Gaussian Process

Quantile regression
Random Forests

Normalized scores
Two Gaussian Processes

We aim to learn a normalized score function $\frac{(Y-m(X))^2}{f(X)}$, with $f \geq 0$ targeting both adaptivity and coverage.

Kernel Sum-of-Squares functions

Building on [3], we use the sum-of-squares framework to estimate the positive function f. Consider a RKHS $\mathcal{H}^f$ with associated feature map ϕ. We parameterize our positive function f by a Hilbert space operator $\mathcal{A} \in \mathcal{S}_+(\mathcal{H}^f)$:

$$f(x) = \langle \phi(x), \mathcal{A}\phi(x) \rangle_{\mathcal{H}^f}, \quad \mathcal{A} \succeq 0$$

The positivity of f is imposed by the positive semi definite property of operator $\mathcal{A}$.

Problem setup

We propose a fundamentally different approach, inspired by [2, 1]:

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

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

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

Important features for this problem are:

1. Accurate mean prediction
2. Prediction intervals with minimum mean width and an additional regularity penalty
3. 100% coverage on pre-training data, later calibrated on the calibration dataset

Representer theorem

This problem is infinite dimensional (m is a function and $\mathcal{A}$ an operator).
We prove a **representer theorem** that makes this problem tractable in practice, which writes as a **SDP** problem.

$$\inf_{\boldsymbol{\gamma} \in \mathbb{R}^n,\ \mathbf{A} \in \mathbb{S}_+^n} \ \frac{a}{n}\sum_{i=1}^n \left(Y_i - \boldsymbol{\gamma}^\top \mathbf{k}_{X_i}^m\right)^2 + \frac{b}{n}\sum_{i=1}^n \tilde{f}_{\mathbf{A}}(X_i) + \lambda_1 \|\mathbf{A}\|_\star + \lambda_2 \|\mathbf{A}\|_F^2$$
$$\text{s.t.} \quad \tilde{f}_{\mathbf{A}}(X_i) \geq \left(Y_i - \boldsymbol{\gamma}^\top \mathbf{k}_{X_i}^m\right)^2,\ i \in [n]\,,$$
$$\boldsymbol{\gamma}^\top \mathbf{K}^m \boldsymbol{\gamma} \leq s.$$

Dual formulation

Off-the-shelves SDP solvers have scale limitations: to overcome this, we exhibit the following dual formulation. The Lagrangian writes:

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

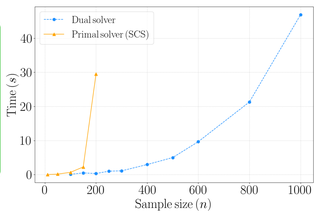
and the dual formulation is:

$$\sup_{\boldsymbol{\Gamma}, \theta \geq 0} D(\boldsymbol{\Gamma}, \theta) = \mathbf{r}(\boldsymbol{\Gamma}, \theta)^T \text{Diag}(\boldsymbol{\Gamma}_a)\mathbf{r}(\boldsymbol{\Gamma}, \theta) + \theta(\boldsymbol{\gamma}(\boldsymbol{\Gamma}, \theta)^T \mathbf{K}^m \boldsymbol{\gamma}(\boldsymbol{\Gamma}, \theta) - s) - \Omega^\star(\mathbf{V}\text{Diag}(\boldsymbol{\Gamma}_{-b})\mathbf{V}^T)$$

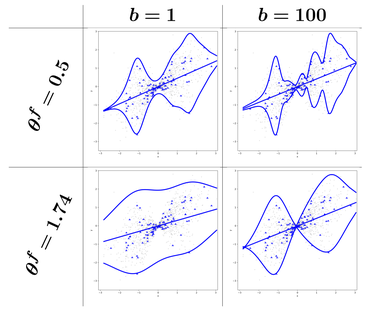
$$\mathbf{r}(\boldsymbol{\Gamma}, \theta) := \mathbf{y} - \mathbf{K}^m \boldsymbol{\gamma}(\boldsymbol{\Gamma}, \theta) \qquad \boldsymbol{\gamma}(\boldsymbol{\Gamma}, \theta) := \boldsymbol{C}(\boldsymbol{\Gamma}, \theta)^{-1} \text{Diag}(\boldsymbol{\Gamma}_a)\mathbf{y}$$
$$\boldsymbol{C}(\boldsymbol{\Gamma},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

We propose a new efficient AGD algorithm:

- Nesterov Acceleration with forward-backward step size search
- It scales up to several thousands samples, unlike previous SDP approaches that are limited to around 200 samples



Hyper-parameters selection



- When b increases, the mean width decreases
- When θ^f increases, the bands get smoother
- The diagonal shows a clear interaction between b and θ^f
- We propose to fix b and optimize θ^f only
- Criterion: perfectly adaptive bands guarantee **local coverage**

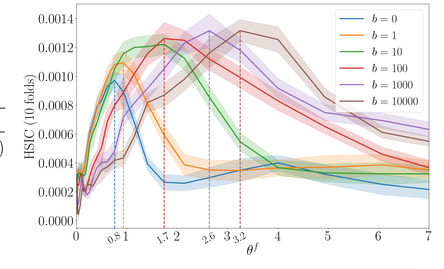
$$\mathbb{P}(Y \in \widehat{C}(X) \mid X = x) \geq 1 - \alpha$$

(impossible to reach if distribution-free)

- Relaxation: $\forall x \in \mathcal{X},\ \ \mathbb{P}(x \in \omega_X) \geq \delta$:

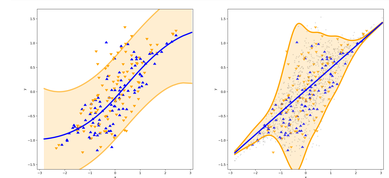
$$\mathbb{P}(Y \in \widehat{C}(X) \mid X \in \omega_X) \geq 1 - \alpha - \frac{1}{\delta}\sqrt{1 - \frac{\alpha_1}{1 - \alpha_2 \text{HSIC}(r_{\mathcal{D}_n}(X, Y), \widehat{f}_{\theta^f}(X))}}$$

- We estimate HSIC with cross-validation



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

Results



Homoscedastic GP
MW = 1.712

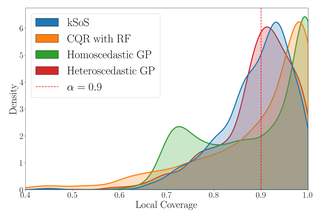
kSoS with Opt. HSIC
MW = 1.759

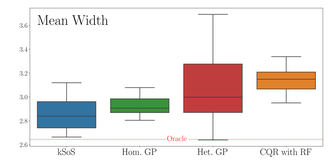
- Mean width is not a reliable adaptivity metric
- Instead, we look at local coverage for analytical datasets
- For real world datasets, we use worst-set coverage proposed by [5]

$$\min_{l=1,\dots,L} \mathbb{P}(Y_{N+1} \in \widehat{C}_{\mathcal{D}_N}(X_{N+1}) \mid X_{N+1} \in \mathcal{R}_l)$$

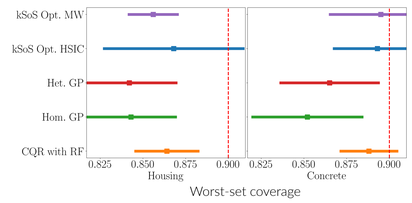
where $(\mathcal{R}_l)_{l=1}^L$ is a partition of $\mathcal{X}$

Analytical datasets





Real world datasets



Worst-set coverage

Dataset	CQR	Het GP	Hom GP	kSoS Opt. MW	kSoS Opt. HSIC
Concrete	0.586	**0.508**	0.543	0.556	0.568
Bike	1.114	1.000	**0.809**	**0.803**	**0.803**
Bio	**1.879**	2.21	2.194	2.03	–
Diabetes	188.62	191.24	190.58	**185.83**	187.6
MPG	9.89	9.70	9.71	**9.15**	9.36
Housing	1.816	1.585	**1.453**	**1.468**	1.586

Mean width on real-world datasets

- Kernel SoS produces intervals with mean widths comparable to existing methods
- More importantly, it delivers superior local and worst-set coverage than competitors
- These results demonstrate that kernel SoS is more adaptive, and with our dual formulation, now scalable to large-data settings

References

[1] Jianqing Fan, Jiawei Ge, and Debarghya Mukherjee. UTOPIA: Universally Trainable Optimal Prediction Intervals Aggregation, July 2024.
[2] Tengyuan Liang. Universal Prediction Band via Semi-Definite Programming. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, August 2022.
[3] Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Non-parametric Models for Non-negative Functions. *NeurIPS*, 2020.
[4] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive Confidence Machines for Regression. In *ECML*, 2002.
[5] Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. *PMLR*, July 2025.

NEURAL INFORMATION PROCESSING SYSTEMS
39th Conference on Neural Information Processing Systems (NeurIPS 2025)

louis.allain@safrangroup.com