Asymmetric conformal prediction with penalized 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

Method

Context

- Operational ML requires a trustworthy uncertainty quantification in their predictions
- This necessitates strong properties such as distribution-free, finite-sample marginal coverage and adaptivity to the data
- **Split Conformal Prediction** achieves this by splitting the data into a training set $\mathcal{D}_n$ and a calibration set $\mathcal{D}_m$, with coverage property: $\mathbb{P}(Y \in \widehat{C}(X)) \geq 1-\alpha$

Symmetric vs Asymmetric intervals

- Symmetric bands with asym. noise become overly conservative, while asymmetric bands with sym. noise discard useful information and produce wider intervals
- Empirical residuals can diverge from the true noise: small samples induce artificial asymmetry, while predictive model bias skews observed residuals

Instead of assuming any symmetric structure:

We propose a continuum between symmetric and asymmetric bands	We select appropriate amount of symmetry in a data-driven way

- New asymmetric score function inspired by Conformalized Quantile Regression [4]

$$s(X,Y) = \max(\widehat{m}_{\mathrm{n}}(X) - f_{\mathrm{low}}(X) - Y, Y - \widehat{m}_{\mathrm{n}}(X) - f_{\mathrm{up}}(X))$$

- With $\widehat{q}_\alpha$ computed on $s(X_j,Y_j), (X_j,Y_j) \in \mathcal{D}_m$, the final prediction bands are

$$\widehat{C}_N(X) = [\widehat{m}_{\mathrm{n}}(X) - f_{\mathrm{low}}(X) - \widehat{q}_\alpha,\ \widehat{m}_{\mathrm{n}}(X) + f_{\mathrm{up}}(X) + \widehat{q}_\alpha]$$

- Following [1], we propose to learn **positive** f_{low} and f_{up} in a goal-oriented way

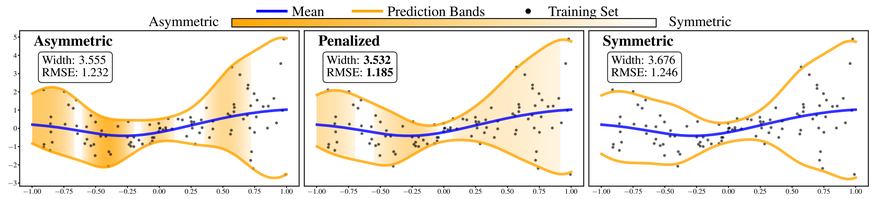
Kernel Sum-of-Squares functions

Building on [3], we use the sum-of-squares framework to estimate positive functions f_{low} and f_{up}. For each function, consider a RKHS $\mathcal{H}_{(\cdot)}$ with associated feature map $\phi_{(\cdot)}$. We parameterize our positive function $f_{(\cdot)}$ by a Hilbert space operator $\mathcal{A}_{(\cdot)} \in \mathcal{S}_+(\mathcal{H}_{(\cdot)})$:

$$f_{(\cdot)}(x) = \langle \phi_{(\cdot)}(x), \mathcal{A}_{(\cdot)}\phi_{(\cdot)}(x)\rangle_{\mathcal{H}_{(\cdot)}}, \quad \mathcal{A}_{(\cdot)} \succeq 0, \quad (\cdot) = \mathrm{low}, \mathrm{up}$$

The positivity of $f_{(\cdot)}$ is imposed by the positive semi definite property of operator $\mathcal{A}_{(\cdot)}$.

Problem setup



Even with true symmetric noise, allowing some asymmetry can lead to better prediction bands

We learn the prediction bands with:

- small mean width (for adaptivity)
- 100% coverage on $\mathcal{D}_n$ to learn the bands shape (later calibrated on $\mathcal{D}_m$)
- regularization that allows to interpolate between asymmetry and symmetry with intensity λ_{pen}

Our penalized kSoS problem admits a **representer theorem** which guarantees the solution is characterized by two PSD matrices which solve the following **semi-definite problem** [2]:

$$\begin{aligned} \inf_{\substack{\mathbf{A}_{\mathrm{low}} \in \mathbb{S}_+^n \\ \mathbf{A}_{\mathrm{up}} \in \mathbb{S}_+^n}} \ & \frac{b}{n}\sum_{i=1}^n \left(f_{\mathbf{A}_{\mathrm{low}}}(X_i) + f_{\mathbf{A}_{\mathrm{up}}}(X_i)\right) + \Omega_{\mathrm{low}}(\mathbf{A}_{\mathrm{low}}) + \Omega_{\mathrm{up}}(\mathbf{A}_{\mathrm{up}}) + \lambda_{\mathrm{pen}}\sum_{i=1}^n \left(f_{\mathbf{A}_{\mathrm{low}}}(X_i) - f_{\mathbf{A}_{\mathrm{up}}}(X_i)\right)^2 \\ \text{s.t.} \ & \widehat{m}_n(X_i) - f_{\mathbf{A}_{\mathrm{low}}}(X_i) - Y_i \leq 0,\ i \in [n] \\ & Y_i - \widehat{m}_n(X_i) - f_{\mathbf{A}_{\mathrm{up}}}(X_i) \leq 0,\ i \in [n] \end{aligned}$$

where $\Omega(\mathbf{A}) = \lambda_1\|\mathbf{A}\|_\star + \lambda_2\|\mathbf{A}\|_{\mathrm{F}}^2$ is the elastic-net regularization for operators.

$\lambda_{\mathrm{pen}} \to 0$ allows for purely asymmetric model and $\lambda_{\mathrm{pen}} \to +\infty$ gives symmetric model close to [1].

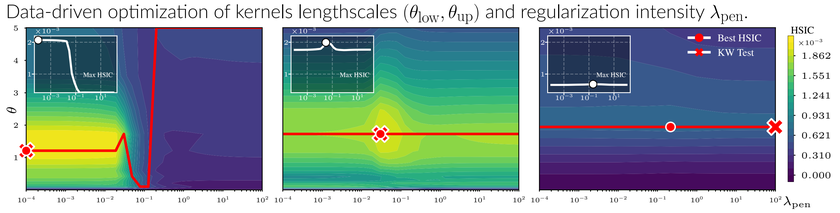
This kSoS problem admits a **dual formulation** with $\mathcal{O}(n)$ variables, which makes it scalable to larger datasets:

$$\sup_{\substack{\boldsymbol{\Gamma}_{\mathrm{low}} \in \mathbb{R}_+^n, \boldsymbol{\Gamma}_{\mathrm{up}} \in \mathbb{R}_+^n, \\ \boldsymbol{\alpha}_0 \in \mathbb{R}^n}} (\boldsymbol{\Gamma}_{\mathrm{up}} - \boldsymbol{\Gamma}_{\mathrm{low}})\mathbf{r}^\top - \frac{1}{4\lambda_{\mathrm{pen}}}\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^\top - \sum_{\mathrm{low,up}} \underbrace{\Omega_{(\cdot)}^\star(\mathbf{V}_{(\cdot)}\mathrm{Diag}(\boldsymbol{\Gamma}_{(\cdot)} \pm \boldsymbol{\alpha}_0 - b/n)\mathbf{V}_{(\cdot)}^\top)}_{\Omega^\star(\boldsymbol{B}) = \frac{1}{4\lambda_2}\|[\boldsymbol{B} - \lambda_1 \boldsymbol{I}_n]_+\|_{\mathrm{F}}^2,\ [\cdot]_+ \text{matrix positive part}}$$

To scale even further, we can leverage kernel **low-rank approximation** methods →

Hyperparameters selection

Data-driven optimization of kernels lengthscales $(\theta_{\mathrm{low}}, \theta_{\mathrm{up}})$ and regularization intensity λ_{pen}.



(a) asymmetric noise: we automatically select a small λ_{pen}

(b) symmetric noise with sampling artefacts: we select an intermediate λ_{pen}

(c) symmetric noise: we select a large λ_{pen}

Criterion: perfectly adaptive bands guarantee **local coverage** (impossible to achieve in a distribution-free setting)

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

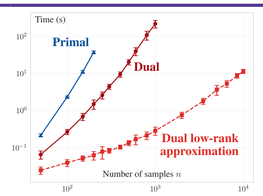
Relaxation: We replace conditional coverage with neighborhood coverage and show:

$$\mathbb{P}(Y \in \widehat{C}(X) \mid X \in \omega_X) \geq 1 - \alpha - \frac{1}{\delta}\sqrt{1 - \frac{\alpha_1}{1 - \alpha_2 \mathrm{HSIC}(\mathrm{res}, \mathrm{width})}}, \quad \mathbb{P}(X \in \omega_X) \geq \delta$$

- $\mathrm{res} = |Y - \widehat{m}_n(X) - (f_{\mathrm{up}}(X) - f_{\mathrm{low}}(X))/2|$ are the centered residuals
- $\mathrm{width} = (f_{\mathrm{up}}(X) + f_{\mathrm{low}}(X))/2$ is the width of the prediction bands

Goal: We expect better local coverage by maximizing HSIC

1. for each λ_{pen} we identify the best $(\theta_{\mathrm{low}}, \theta_{\mathrm{up}})$ with HSIC (—)
2. We select λ_{pen} with best opt. HSIC (●)
3. Except if we detect no significant difference in opt. HSIC along λ_{pen} (Kruskal-Wallis test), in which case we default to the simpler symmetric model (large λ_{pen}) (✕)



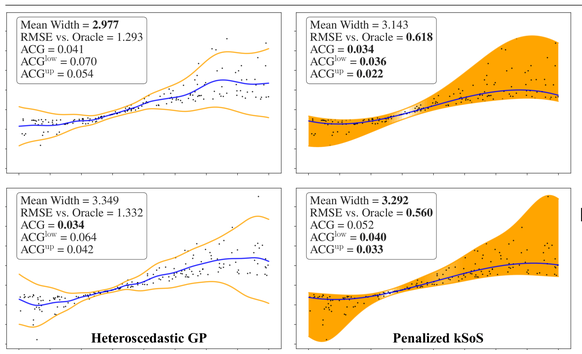
Low-rank approximation

Goal: Reduce $\Omega^\star$ complexity - positive part requires eigendecomposition, cost $\mathcal{O}(n^3)$

Solution: Replace the Cholesky factor $\mathbf{V}$ by a rank-r approximation (Nyström, Random Fourier Features or randomized SVD) → $\mathcal{O}(r^3)$

Results

Adaptivity Metrics



Mean width alone is misleading

Local coverage is not meaningful with asym. noise

Absolute Coverage Gap for analytical datasets:

$$\mathrm{ACG} = \frac{1}{n_X}\sum_{i=1}^{n_X}\left|\hat{p}(X_i) - (1-\alpha)\right|$$

with $\hat{p}$ an estimate of the local coverage

Worst Set Coverage for real world datasets:

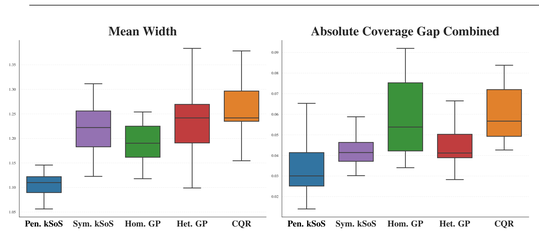
$$\mathrm{WSC} = \min_{l=1,\dots,L} \mathbb{P}\left(Y \in \widehat{C}_{\mathcal{D}_N}(X) \mid X \in \mathcal{R}_l\right)$$

with $\{\mathcal{R}_l\}_{l=1,\dots,L}$ a partition of Ω

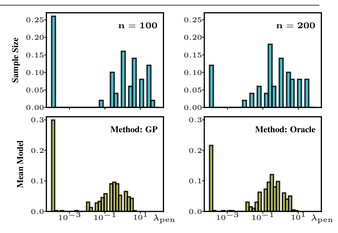
a. We extend both criteria in *low* and *up* versions with $1-\alpha/2$, which we combine into $\mathbf{ACG}^c$ and $\mathbf{WSC}^c$ to replace usual symmetric local coverage

b. Benchmarks: we look first for correct adaptivity, then break ties with mean width

Analytical Datasets



Mean Width and ACG^c for a dataset with true asymmetric noise



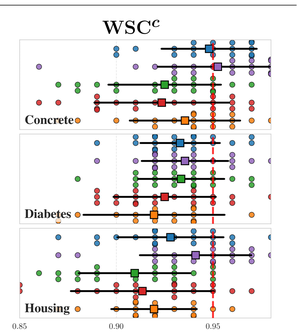
Sym. noise: histogram of selected λ_{pen} with two sample sizes and predictive models

- Penalized kSoS performs **better than or comparably to any competitor** on analytical datasets, regardless of noise symmetry
- When the sample size increases from $n = 100$ to $n = 200$, and when the predictive model changes from a Gaussian Process to the oracle, the purely asymmetric model is selected less often
- The asymmetric continuum thus allows Pen. kSoS to mitigate small samples artifacts and biased predictive models

Real World Datasets

Mean width

Dataset	CQR	Het. GP	Hom. GP	Sym. kSoS	Pen. kSoS
Concrete	22.66	21.53	**21.01**	28.00	21.74
Bike	217.98	199.23	169.57	177.53	**162.47**
Diabetes	**188.65**	196.25	193.21	195.80	**190.46**
Housing log	0.967	0.855	**0.768**	0.967	0.836
Housing	1.77	1.70	**1.60**	2.02	1.80
MPG	9.82	9.76	**9.33**	9.68	**9.32**
Boston	12.16	10.85	**9.76**	17.93	11.05
Energy	1.50	1.46	1.72	1.79	**1.39**
Miami	$31.4e^4$	$36.4e^4$	$29.5e^4$	$28.3e^4$	$\mathbf{26.9e^4}$
Sulfur	**0.051**	0.055	**0.050**	0.064	**0.050**
Power	13.29	**12.88**	13.05	13.62	**12.99**
Yacht	0.645	**0.595**	0.609	0.616	**0.599**
Kin8nm	0.607	0.425	**0.386**	0.453	0.393



- On real world datasets, Pen. kSoS is always **first or close second** when measured by WSC^c and mean width, a robustness never achieved by any other method

Takeaways

- Penalized kSoS **seamlessly interpolates between symmetric and asymmetric bands**, with representer theorems and scalable dual formulations
- **Fully data-driven** hyperparameter strategy that automatically detects whether asymmetry is beneficial
- Controlled asymmetry **compensates for small samples artifacts and model bias**, even when the true noise is symmetric

References

[1] L. Allain, S. Da Veiga, and B. Staber. Scalable and adaptive prediction bands with kernel sum-of-squares. *NeurIPS*, 2025.
[2] L. Allain, S. Da Veiga, and B. Staber. Asymmetric conformal prediction with penalized kernel sum-of-squares. *ICML*, 2026.
[3] U. Marteau-Ferey, F. Bach, and A. Rudi. Non-parametric Models for Non-negative Functions. *NeurIPS*, 2020.
[4] Y. Romano, E. Patterson, and E. Candes. Conformalized Quantile Regression. *NeurIPS*, 2019.

Link to Paper

louis.allain@safrangroup.com