On Training Locally Adaptive Conformal Predictors

nicolo colombo *

September 14, 2023 - COPA2023

*nicolo.colombo@rhul.ac.uk

outline

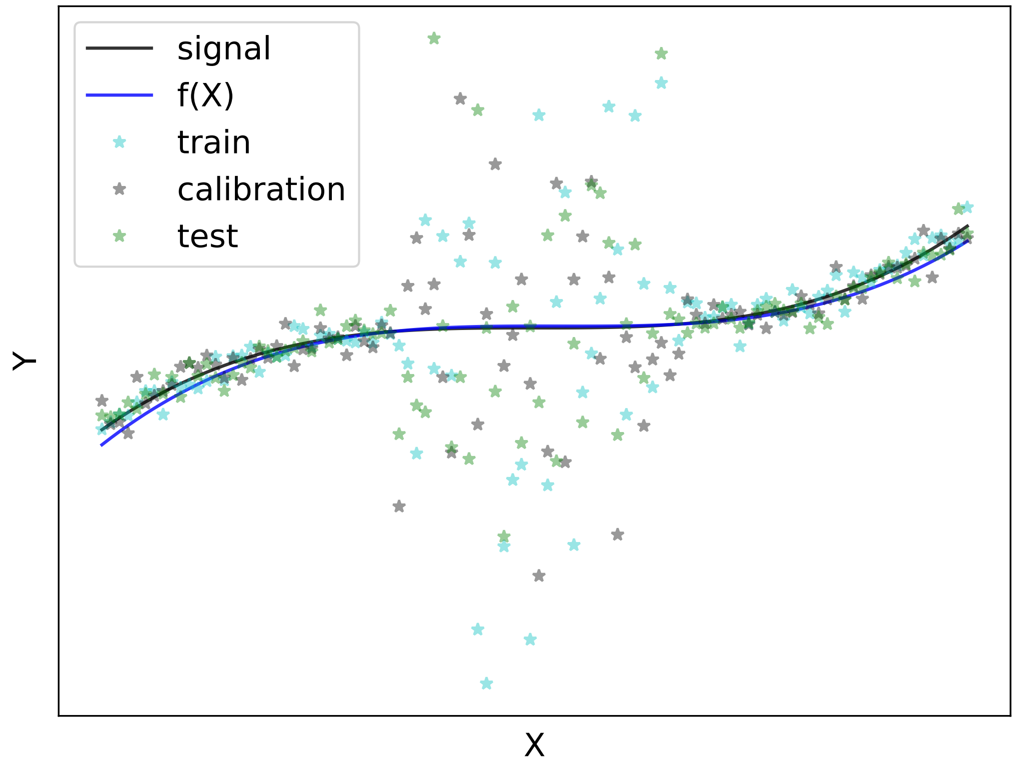
localizing CP

conformity scores

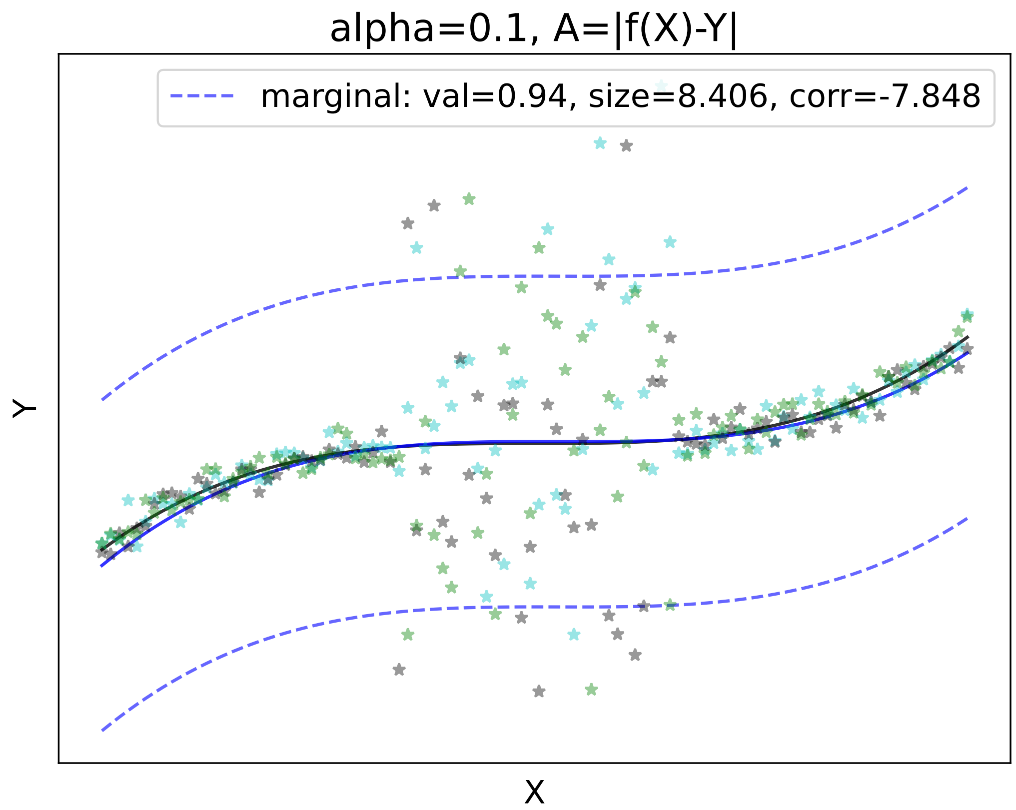
training

discussion

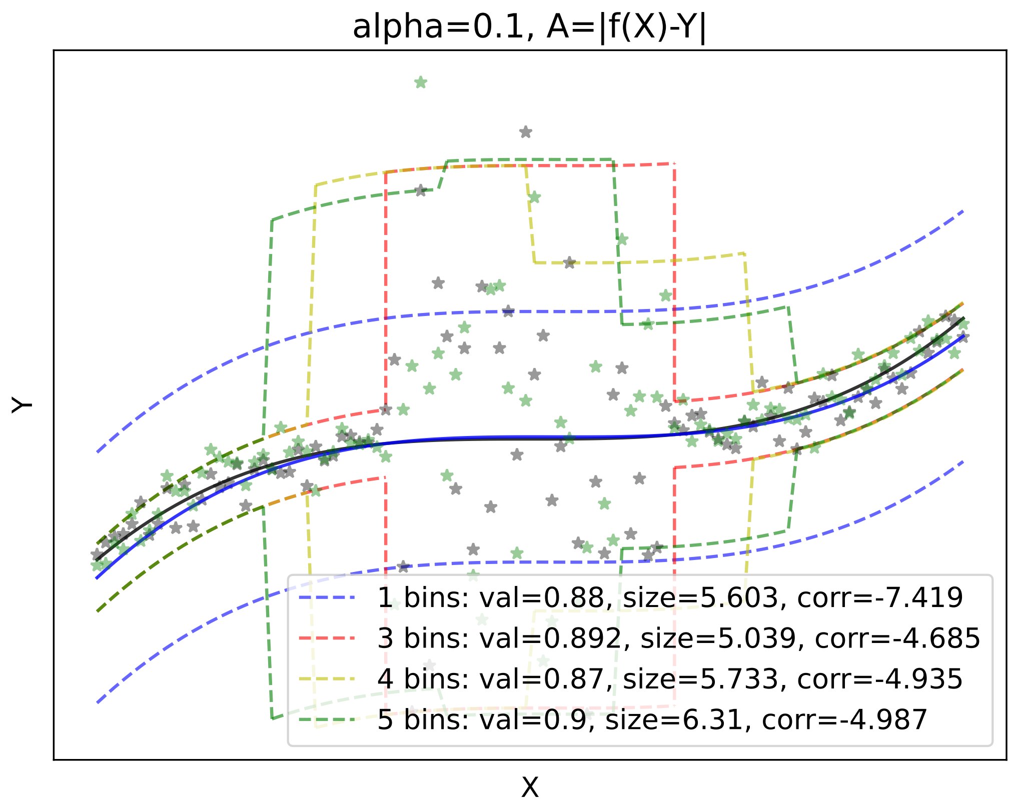
we have good point-prediction model $f(X) \approx \mathsf{E}_{Y|X}(Y)$



marginal prediction intervals (PI) are not efficient



locally-defined PI may be do *better*



locally-defined PI approximate the *ideal* conditional validity

$$\text{Prob}(Y_{test} \in C | X_{test}) \geq 1 - \alpha$$

how to define the bins? what if somewhere the calibration set gets *too small*?

instead of partitioning the data, we *localize* the conformity function

$$A(Y, f(X)) \quad \rightarrow \quad B(A, X) = \phi_X(A)$$

to avoid overfitting, $\phi_X(A)$ has a globally defined functional form

predictions on different locations are evaluated *differently*, e.g.

$$A' < A \quad \not\Rightarrow \quad \phi_{X'}(A') < \phi_X(A)$$

the PI now depend on the transformed calibration set

$$\{B_n = \phi_{X_n}(|Y_n - f(X_n)|)\}_{n=1}^N$$

let Q_B be the $(1-\alpha)$th sample quantile of $\{B_n = \phi_{X_n}(|f(X_n) - Y_n|)\}_{n=1}^N$, i.e.

$$Q_B \quad \text{is such that} \quad |\{B_n \leq Q_B\}_{n=1}^N| = \lceil (1-\alpha)N \rceil$$

in the B-space, we have standard marginal PI

$$C_B = \{b \in \mathbb{R}, b \leq Q_B\}$$

in the label space, C_B *becomes* locally-adaptive

$$\begin{aligned} C_B \quad \sim \quad C &= \{y \in \mathbb{R}, |y - f(X_{test})| \leq \phi^{-1}_{X_{test}}(Q_B)\} \\ &= [f(X_{test}) - \phi^{-1}_{X_{test}}(Q_B), f(X_{test}) + \phi^{-1}_{X_{test}}(Q_B)] \end{aligned}$$

an example from the literature is the Locally Reweighted (LR) CP algorithm

$$B = \phi_X = \frac{A}{g^2(X) + \gamma}, \quad g(X) \approx \mathsf{E}_{A|X}(A), \quad \gamma > 0$$

intuitively, LR *works* because B is almost uniformely distributed for all X

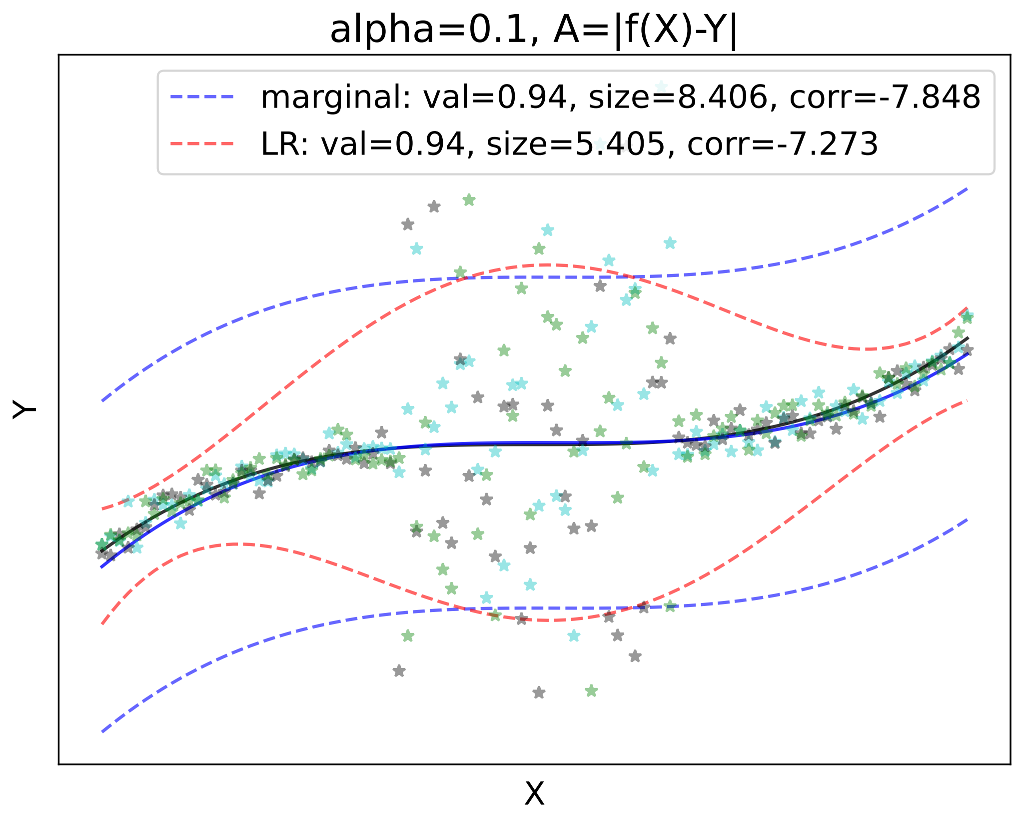
the label-space PI are

$$C = [f(X_{test}) - \phi^{-1}_{X_{test}}(Q_B), f(X_{test}) + \phi^{-1}_{X_{test}}(Q_B)]$$

which have locally-adaptive sizes

$$|C| = Q_B(g^2(X_{test}) + \gamma)$$

the obtained PI are marginally valid by construction



we *extend* the LR idea in two ways

1 - we let $\phi_X(A)$ be general monotonic functions of A

2 - we train ϕ_X to maximize the *efficiency of the PI*

i.e. we define a model class $\Phi = \{\phi_X(A,\theta), X \in \mathcal{X}, \theta \in \mathbb{R}^d\}$ and minimize the average size of the PI

$$\ell_{\texttt{size}}(\theta) = \mathsf{E}_{\alpha, X_{test}, D_{cal}} \left(\phi^{-1}_{X_{test}}(Q_B, \theta)\right)$$

for example, let

$$\phi_X = A\sigma(\theta_1(1-\theta_2 X^2)), \quad \sigma(t) = \frac{1}{1+e^{-t}}$$

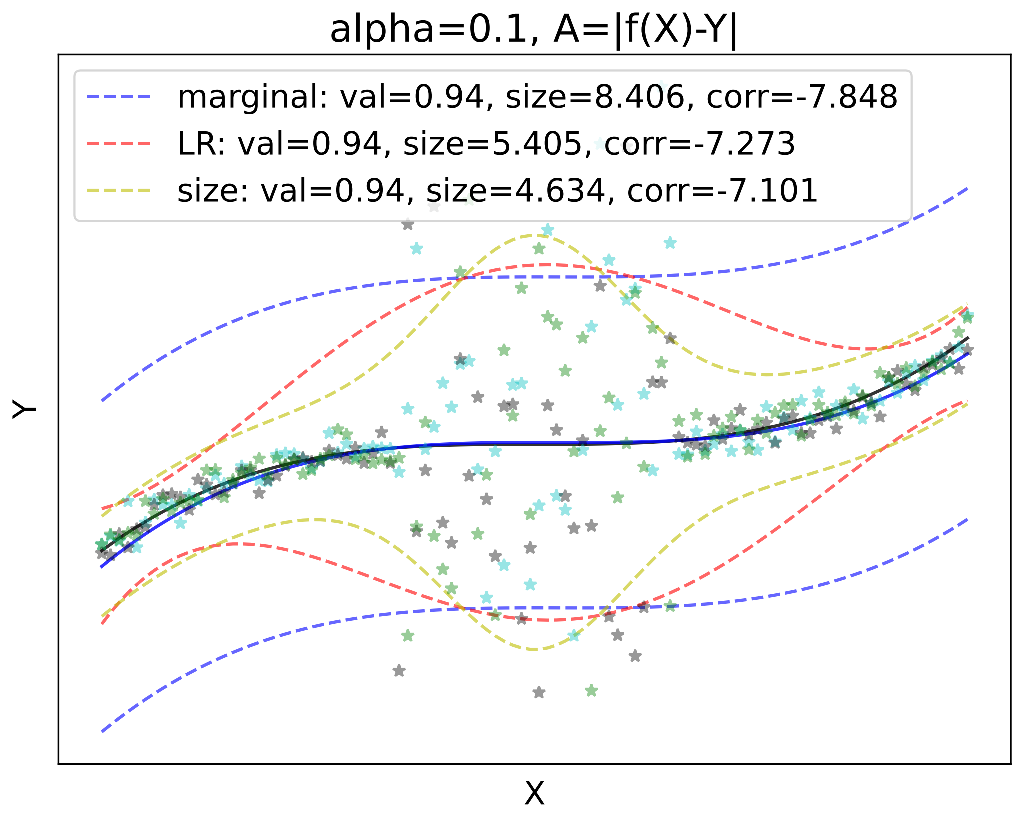
and search for the *optimal* $\theta = (\theta_1, \theta_2)$

$\sigma(\theta_1(1-\theta_2 X^2))$ does *not* need to be a model of the conditional residuals

the obtained locally-adaptive PI are

$$C = \left\{ y \in \mathbb{R}, |f(X) - y| \leq \frac{Q_B}{\sigma(\theta_1(1-\theta_2 X^2))} \right\}$$

again, the PI are marginally valid by construction



we obtain θ by minimizing

$$\ell_{\texttt{size}}(\theta) \approx \sum_{n \neq n'} \phi_{X_n}^{-1}\left(\phi_{X_{n'}}(A_{n'}, \theta), \theta)\right) = \sum_{n \neq n'} A_{n'} \frac{\sigma(\theta_1(1 - \theta_2 X_{n'}^2)}{\sigma(\theta_1(1 - \theta_2 X_n^2))}$$

with gradient descent updates *

$$\theta \leftarrow \theta - \eta \sum_{n \neq n'} d\phi_{X_n}^{-1}\left(\phi_{X_{n'}}(A_{n'}, \theta), \theta\right)$$

the derivatives of ϕ_X^{-1} are obtained implicitly from

$$d_\theta \phi_X^{-1}(\phi_X(A, \theta)) = 0$$

and

$$\partial_B \left(\phi_X\left(\phi_X^{-1}(B, \theta), \theta\right)\right) = 1$$

$^* d(\psi \circ \zeta) = \nabla \psi \circ \zeta + (\psi' \circ \zeta)\, \nabla \zeta$

how flexible is the scheme?

other possible model classes are

$$\phi_X = A \exp(g(X)), \quad \phi_X = \log A + g(X), \quad \phi_X = \sigma(\log A + g(X))$$

the models fulfil a domain-codomain assumption

$$\phi_X : \mathbb{R}_+ \to \mathcal{B}, \quad \phi_X^{-1} : \mathcal{B} \to \mathbb{R}_+ \text{ for all } X$$

for example, $\phi_X = \sigma(A + g^2(X))$ *is not allowed* because $\mathrm{logit}(\sigma(A + g^2(X)) - g^2(X')$ may be negative for some $X, X' \in \mathbb{R}$ and $A = |f(X) - Y|$

thank you!