Conditional coverage diagnostics for conformal prediction

Sacha Braun[1,2], David Holzmüller[3], Michael I. Jordan[1,4], Francis Bach[1,2]

[1]Sierra team, Inria Paris [2]Ecole Normale Supérieure, PSL [3]Soda team, Inria Paris-Saclay [4]UC Berkeley

1. The goal: conditional coverage metrics

We want to estimate (among other metrics)

$$\mathbb{E}_X\Big[\,\big|\,\mathbb{P}(Y \in C_\alpha(X) \mid X) - (1-\alpha)\,\big|\,\Big].$$

2. Predicting conditional coverage through probabilistic classification

- Let $Z = \mathbf{1}\{Y \in C_\alpha(X)\}$.
- Train a classifier $h(X)$ to predict Z.
- Optimal predictor for this problem is $p(X) = \mathbb{P}(Y \in C_\alpha(X) \mid X)$.

3. Compare validation losses

For a proper loss ℓ, the **Excess Risk of the Target coverage (ℓ-ERT)** is:

$$\ell\text{-ERT}(p) := \mathbb{E}[\ell(1-\alpha, Z)] - \mathbb{E}[\ell(p(X), Z)].$$

Under conditional coverage, no predictor performs better than the constant $1-\alpha$.

The lower bound guarantee with cross-validation

For any learned classifier h, we get a **lower bound** on the true miscoverage:

$$\ell\text{-ERT}(h) \leq \ell\text{-ERT}(p).$$

We estimate this quantity using the empirical risk on held-out data, so the lower bound holds in expectation.

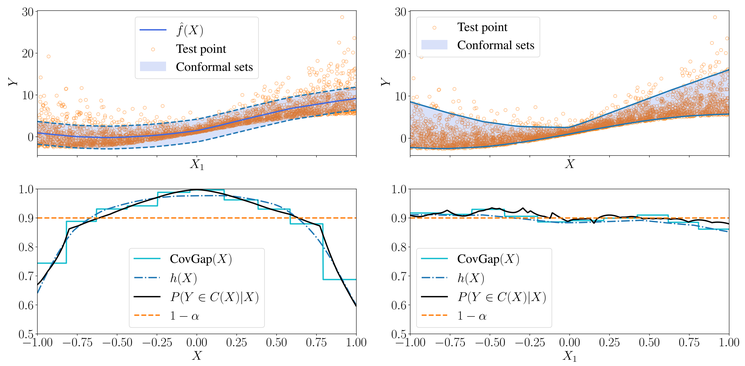
4. One loss for every distance function

Different proper scoring rules ℓ allow us to estimate specific, interpretable distances to perfect coverage:

Metric	Proper score $\ell(p, z)$	Estimated target quantity (ℓ-ERT)
L_1-ERT	$\operatorname{sgn}(p-(1-\alpha))(1-\alpha-z)$	$\mathbb{E}_X[\lvert 1-\alpha-p(X)\rvert]$
L_2-ERT	$(z-p)^2$	$\mathbb{E}_X[(1-\alpha-p(X))^2]$
KL-ERT	$-z\log p-(1-z)\log(1-p)$	$\mathbb{E}_X[D_{\mathrm{KL}}(p(X)\Vert 1-\alpha)]$
d-ERT	$-f_\alpha(p)-(y-p)f'_\alpha(p)$	$\mathbb{E}_X[d\big(1-\alpha, p(X)\big)]$

where $f_\alpha(q) = d(1-\alpha, q)$ is convex with $f_\alpha(1-\alpha) = 0$ and f'_α is a subderivative of f_α satisfying $f'_\alpha(1-\alpha) = 0$.

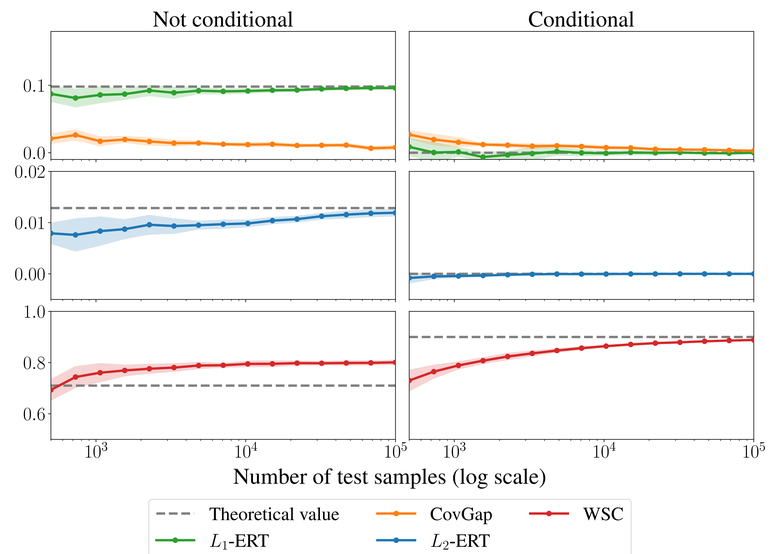
5. Visualizing ERT in 1D



$h(\cdot)$ is a neural network. Left: Non-conditional CP fails locally; $h(X)$ captures the variance (L_1-ERT $\approx 0.075 > 0$).
Right: More conditional sets; $h(X)$ stays flat at $1-\alpha$ (L_1-ERT ≈ 0).

6. Sample efficiency vs. baselines

ERT stabilizes with far fewer samples than WSC or CovGap, providing a reliable diagnostic rapidly without false positives.



$X \sim \mathcal{U}\left([-1,1]^8\right)$ and $Y \sim \mathcal{N}(0, \sigma(X_1))$.
Left: Marginal but not conditional setting. ERT detects failure instantly.
Right: Perfectly conditional setting (oracle). WSC hallucinated failures; ERT correctly stays near zero.

7. Estimating under- and over-coverage

By strategically clipping the predicted probability p within the proper loss function, we can decompose the miscoverage to isolate specific failures:

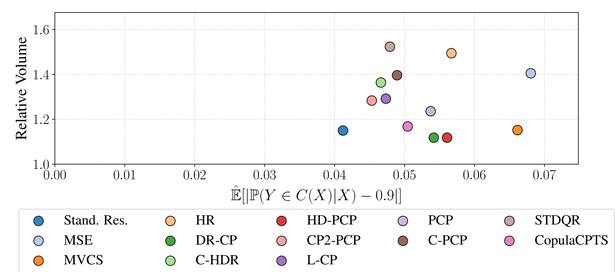
- **ℓ_+-ERT (over-coverage):** Computed using $\ell_+(p, y) = \ell(\max\{p, 1-\alpha\}, y)$.
- **ℓ_--ERT (under-coverage):** Computed using $\ell_-(p, y) = \ell(\min\{p, 1-\alpha\}, y)$.

8. Modern classifiers beat simple ones

Classifier	L_1-ERT	L_2-ERT	Time / 1K (s)
TabICLv1.1 (GPU)	$\mathbf{72.7}_{1.6}$	$\mathbf{54.0}_{1.4}$	$16.0_{0.0}$
RealTabPFN-2.5 (GPU)	$72.1_{1.4}$	$49.7_{1.1}$	$9.2_{0.0}$
LightGBM (CPU)	$68.9_{1.8}$	$46.4_{1.0}$	$2.4_{0.0}$
RandomForest (CPU)	$67.3_{2.0}$	$36.9_{2.4}$	$4.4_{0.0}$
PartitionWise (CPU)	$33.7_{1.9}$	$10.5_{0.7}$	$\mathbf{0.2}_{0.0}$

Average % of the maximum ERT recovered across datasets, for different number of test samples. Traditional partition-wise estimators (like CovGap) capture less than 40% of the actual miscoverage compared to gradient-boosted trees and foundation models.

9. Conformal prediction scores benchmark



Comparison of conditional coverage deviation and normalized prediction set volumes across various conformal prediction strategies ($\alpha = 0.1$). Volumes are scaled by the power $1/dim_{output}$ and normalized by the minimum observed volume. The x-axis displays the L_1-ERT metric, which estimates the conditional coverage deviation $\mathbb{E}[|\mathbb{P}(Y \in C(X) \mid X) - (1-\alpha)|]$.

10. Code & package :

```
pip install covmetrics

from covmetrics import ERT
ERT_value = ERT().evaluate(x, z, alpha)
```