Computing Full Conformal Prediction Set with Approximate Homotopy

Eugene Ndiaye and Ichiro Takeuchi

Riken AIP

If you are predicting the label y of a new object with $\hat{y}$,
how confident are you that $y = \hat{y}$?

Observations: $\mathcal{D}_n = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ iid $\sim \mathbb{P}$

New input data: x_{n+1}

Goal: build a set $\hat{\Gamma}(x_{n+1})$ that contains y_{n+1}

Observations: $\mathcal{D}_n = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ iid $\sim \mathbb{P}$

New input data: x_{n+1}

Goal: build a set $\hat{\Gamma}(x_{n+1})$ that contains y_{n+1}

Desirable property:

- $\mathbb{P}^{n+1}(y_{n+1} \in \hat{\Gamma}(x_{n+1})) \geq 1 - \alpha$ for $\alpha \in (0, 1)$
- size of $\hat{\Gamma}(x_{n+1})$ as small as possible

Main idea: Build a *conformity* function $\hat{\pi}$ such that

Given a confidence level $1 - \alpha$,

$$\hat{\pi}(y) > \text{ threshold}(\alpha) \text{ when } y \text{ is } \textit{"typical" w.r.t. } y_1, \cdots, y_n.$$

Main idea: Build a *conformity* function $\hat{\pi}$ such that

Given a confidence level $1 - \alpha$,

$$\hat{\pi}(y) > \text{ threshold}(\alpha) \text{ when } y \text{ is } \textit{"typical" w.r.t. } y_1, \cdots, y_n.$$

Somehow, $\hat{\pi}$ is a p-value function for testing $H_0 : y = y_{n+1}$

Framework

- **Learning algorithm e.g. ERM:**

$$\hat{\beta}(y_{n+1}) \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n+1} \ell(y_i, x_i^\top \beta) + \lambda \Omega(\beta)$$

(*e.g.* Lasso) $$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n+1} (y_i - x_i^\top \beta)^2 + \lambda \|\beta\|_1$$

Framework

- **Learning algorithm e.g. ERM:**

$$\hat{\beta}(y_{n+1}) \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n+1} \ell(y_i, x_i^\top \beta) + \lambda \Omega(\beta)$$

(*e.g.* Lasso) $$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n+1} (y_i - x_i^\top \beta)^2 + \lambda \|\beta\|_1$$

- **Measures the quality of a prediction (score function):**

$$\hat{R}_i(y_{n+1}) = \psi(y_i, x_i^\top \hat{\beta}(y_{n+1})) \quad \forall i \in [n+1]$$

(*e.g.* Lasso) $$\hat{R}_i(y_{n+1}) = |y_i - x_i^\top \hat{\beta}(y_{n+1})|$$

Main tools

Let $U_1, \cdots, U_n, U_{n+1}$ **iid**.

Order statistics: $U_{(1)} < \cdots < U_{(n)} < U_{(n+1)}$

$\mathrm{Rank}(U_{n+1}) = i$ when $U_{(i)} = U_{n+1}$.

$$\boxed{\mathrm{Rank}(U_{n+1}) \sim \mathcal{U}\{1, \cdots, n+1\}}$$

Main tools

Let $U_1, \cdots, U_n, U_{n+1}$ **iid**.

Order statistics: $U_{(1)} < \cdots < U_{(n)} < U_{(n+1)}$

$\mathrm{Rank}(U_{n+1}) = i$ when $U_{(i)} = U_{n+1}$.

$$\boxed{\mathrm{Rank}(U_{n+1}) \sim \mathcal{U}\{1, \cdots, n+1\}}$$

Assumption: ψ is any function that preserves **iid** structure:

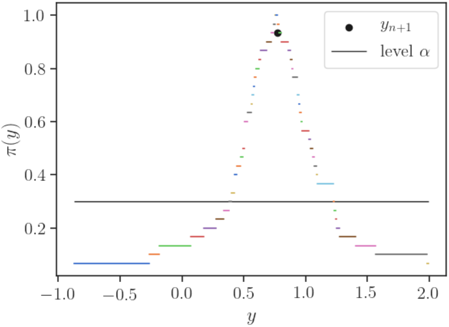
$(x_1, y_1), \cdots, (x_{n+1}, y_{n+1})$ iid $\implies \hat{R}_1(y_{n+1}), \cdots, \hat{R}_{n+1}(y_{n+1})$ iid

$$\boxed{\mathrm{Rank}(\hat{R}_{n+1}(y_{n+1})) \sim \mathcal{U}\{1, \cdots, n+1\} \perp\!\!\!\perp \mathbb{P} \quad !}$$

Conformity function:

$$\hat{\pi}(y_{n+1}) := 1 - \frac{1}{n+1}\mathrm{Rank}(\hat{R}_{y_{n+1},n+1})$$

Lemma: $\mathbb{P}^{n+1}(\hat{\pi}(y_{n+1}) \leq \alpha) \leq \alpha \quad \forall \alpha \in (0,1)$



Interpretation: $\hat{\pi}$ takes small value on non-conform/untypical data!

Conformal Prediction Set

$$\boxed{\textbf{Lemma:} \quad \mathbb{P}^{n+1}(\hat{\pi}(y_{n+1}) > \alpha) \geq 1 - \alpha \quad \forall \alpha \in (0, 1)}$$

[1](V. Vovk, A. Gammerman, and G. Shafer, 2005)
[2](G. Shafer and V. Vovk, 2008)
[3](J. Lei, M. G'Sell, A. Rinaldo, R.J. Tibshirani, and L. Wasserman, 2018)

Conformal Prediction Set

$$\boxed{\textbf{Lemma:} \quad \mathbb{P}^{n+1}(\hat{\pi}(y_{n+1}) > \alpha) \geq 1 - \alpha \quad \forall \alpha \in (0, 1)}$$

y_{n+1} is unknown !

[1](V. Vovk, A. Gammerman, and G. Shafer, 2005)
[2](G. Shafer and V. Vovk, 2008)
[3](J. Lei, M. G'Sell, A. Rinaldo, R.J. Tibshirani, and L. Wasserman, 2018)

Conformal Prediction Set

$$\boxed{\textbf{Lemma:} \quad \mathbb{P}^{n+1}(\hat{\pi}(y_{n+1}) > \alpha) \geq 1 - \alpha \quad \forall \alpha \in (0,1)}$$

y_{n+1} is unknown !

Idea: just test **all** the possibilities [1] [2] [3]

$$y_{n+1} \in \hat{\Gamma}(x_{n+1}) := \{y \in \mathbb{R} : \hat{\pi}(y) > \alpha\}$$

[1] (V. Vovk, A. Gammerman, and G. Shafer, 2005)
[2] (G. Shafer and V. Vovk, 2008)
[3] (J. Lei, M. G'Sell, A. Rinaldo, R.J. Tibshirani, and L. Wasserman, 2018)

Conformal Prediction Set

$$\boxed{\textbf{Lemma:} \quad \mathbb{P}^{n+1}(\hat{\pi}(y_{n+1}) > \alpha) \geq 1 - \alpha \quad \forall \alpha \in (0,1)}$$

y_{n+1} is unknown !

Idea: just test **all** the possibilities [1] [2] [3]

$$y_{n+1} \in \hat{\Gamma}(x_{n+1}) := \{y \in \mathbb{R} : \hat{\pi}(y) > \alpha\}$$

$$\boxed{\textbf{Proposition:} \quad \mathbb{P}^{n+1}(y_{n+1} \in \hat{\Gamma}(x_{n+1})) \geq 1 - \alpha \quad \forall \alpha \in (0,1)}$$

[1] (V. Vovk, A. Gammerman, and G. Shafer, 2005)
[2] (G. Shafer and V. Vovk, 2008)
[3] (J. Lei, M. G'Sell, A. Rinaldo, R.J. Tibshirani, and L. Wasserman, 2018)

Actual limitations

$$\hat{\Gamma}(x_{n+1}) := \{y \in \mathbb{R} : \hat{\pi}(y) > \alpha\}$$

Issue: compute $\hat{\pi}(y)$ *i.e.* refit the model $\hat{\beta}(y)$, $\forall y \in \mathbb{R}$.

Actual limitations

$$\hat{\Gamma}(x_{n+1}) := \{y \in \mathbb{R} : \hat{\pi}(y) > \alpha\}$$

Issue: compute $\hat{\pi}(y)$ *i.e.* refit the model $\hat{\beta}(y)$, $\forall y \in \mathbb{R}$.

- Ok if y_{n+1} has finite number of possibilities
- Ok for Ridge regression (and least square)
- Ok for Elastic net (and Lasso) very recently !
- Non linear regression and others: ???

Actual limitations

$$\hat{\Gamma}(x_{n+1}) := \{y \in \mathbb{R} : \hat{\pi}(y) > \alpha\}$$

Issue: compute $\hat{\pi}(y)$ *i.e.* refit the model $\hat{\beta}(y)$, $\forall y \in \mathbb{R}$.

- Ok if y_{n+1} has finite number of possibilities
- Ok for Ridge regression (and least square)
- Ok for Elastic net (and Lasso) very recently !
- Non linear regression and others: ???

Heuristic: *arbitrary* discretization of a large interval $[y_{\min}, y_{\max}]$.

Approximates the conformal set while keeping strong statistical and computational guarantee.

Approximated ERM

Given a candidate y

$$\hat{\beta}(y) \in \arg\min_{\beta \in \mathbb{R}^p} P_y(\beta) = \sum_{i=1}^{n} \ell(y_i, x_i^\top \beta) + \ell(y, x_{n+1}^\top \beta) + \lambda \Omega(\beta)$$

Approximated ERM

Given a candidate y

$$\hat{\beta}(y) \in \arg\min_{\beta \in \mathbb{R}^p} P_y(\beta) = \sum_{i=1}^{n} \ell(y_i, x_i^\top \beta) + \ell(y, x_{n+1}^\top \beta) + \lambda \Omega(\beta)$$

Approximate the conformal set $\hat{\Gamma}(x_{n+1})$ based on $\beta(y) \approx \hat{\beta}(y)$

$$P_y(\beta(y)) - P_y(\hat{\beta}(y)) \leq \epsilon \ .$$

Build a Solution Path: $\{y_{t_1}, \cdots, y_{T_\epsilon}\}$ such that

$$\forall y \in [y_{\min}, y_{\max}],\ \exists t_k \text{ s.t. } P_y(\beta(y_{t_k})) - P_y(\hat{\beta}(y)) \leq \epsilon$$

Build a Solution Path: $\{y_{t_1}, \cdots, y_{T_\epsilon}\}$ such that

$$\forall y \in [y_{\min}, y_{\max}],\ \exists t_k \text{ s.t. } P_y(\beta(y_{t_k})) - P_y(\hat{\beta}(y)) \leq \epsilon$$

Now we need to recompute the model only T_ϵ times (vs infinite times for the exact solution).

Duality gap bound

$$\hat{\beta}(y) \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \underbrace{\sum_{i=1}^{n} \ell(y_i, x_i^\top \beta) + \ell(y, x_{n+1}^\top \beta) + \lambda \Omega(\beta)}_{P_y(\beta)}$$

$$\hat{\theta}(y) \in \underset{\theta \in \mathbb{R}^{n+1}}{\arg\max} \underbrace{- \sum_{i=1}^{n} \ell^*(y_i, -\lambda\theta_i) - \ell^*(y, -\lambda\theta_{n+1}) - \lambda \Omega^*(X^\top \theta)}_{D_y(\theta)}$$

- **Bound on the approximation error:**

$$P_y(\beta(y)) - P_y(\hat{\beta}(y)) \leq G_y(\beta(y), \theta(y)) \ .$$

Duality gap bound

$$\hat{\beta}(y) \in \arg\min_{\beta \in \mathbb{R}^p} \underbrace{\sum_{i=1}^{n} \ell(y_i, x_i^\top \beta) + \ell(y, x_{n+1}^\top \beta) + \lambda \Omega(\beta)}_{P_y(\beta)}$$

$$\hat{\theta}(y) \in \arg\max_{\theta \in \mathbb{R}^{n+1}} \underbrace{-\sum_{i=1}^{n} \ell^*(y_i, -\lambda \theta_i) - \ell^*(y, -\lambda \theta_{n+1}) - \lambda \Omega^*(X^\top \theta)}_{D_y(\theta)}$$

- **Bound on the approximation error:**

$$P_y(\beta(y)) - P_y(\hat{\beta}(y)) \leq G_y(\beta(y), \theta(y)) \enspace .$$

- **Variation of the gap = Variation of the loss:**

$$G_y(\beta, \theta) - G_{y_0}(\beta, \theta) = \ell(y, x_{n+1}^\top \beta) - \ell(y_0, x_{n+1}^\top \beta) \enspace .$$

Achievements

- If the loss ℓ is smooth, we can guarantee that

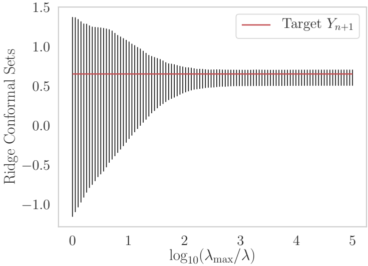
$$\hat{\Gamma}(x_{n+1}) \subset \Gamma^{(\epsilon)}(x_{n+1})$$

- Without smoothness, we can still provide a valid conformal set using ϵ-solution.

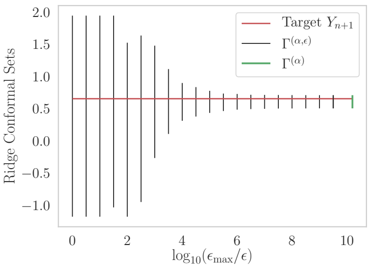
- Computational complexity: upper and lower bound on T_ϵ *w.r.t.* to the regularity of the loss:

  e.g. $T_\epsilon \in O(1/\sqrt{\epsilon})$ for smooth loss.

Experiment 1



(a) Exact solution



(b) Approximation

Figure: Illustration for Ridge regression.

Experiment 2

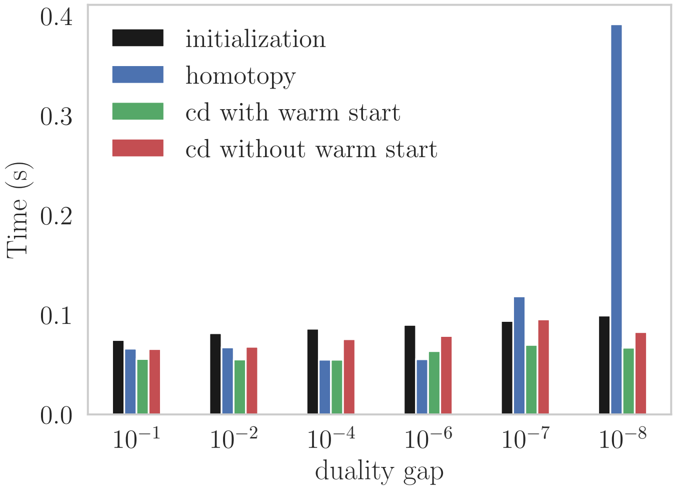


Figure: Evaluate the computational time for Lasso

Experiment 3

	Coverage	Length	Time
Oracle	0.9	1.685	0.59
Split	0.9	3.111	0.26
1e-2	0.9	1.767	2.17
1e-4	0.9	1.727	8.02
1e-6	0.9	1.724	45.94
1e-8	0.9	1.722	312.56

Table: Empirical coverage

Experiment 4

	Oracle	Split	1e-2	1e-4	1e-6	1e-8
`Smooth Chebychev`						
Coverage	0.92	0.95	0.92	0.92	0.92	0.92
Length	1.940	2.271	1.998	1.990	1.987	1.981
Time	0.019	0.016	0.073	0.409	3.742	36.977
`Linex regression`						
Coverage	0.91	0.93	0.91	0.91	0.91	0.91
Length	2.189	2.447	2.231	2.209	2.205	2.199
Time	0.013	0.012	0.050	0.234	2.054	20.712

Table: Regression problem with different loss function regularized with Ridge penalty on Boston and Diabetes dataset.

- $\ell(a, b) = \gamma \log\cosh((a - b)/\gamma)$ is a smooth approx. of $\|\cdot\|_\infty$.
- $\ell(a, b) = \exp(\gamma(a - b)) - \gamma(a - b) - 1$ is an "asymmetric version" of the quadratic loss.

Implementation available at

https://github.com/EugeneNdiaye/homotopy_conformal_prediction