

---

# Identifying Homogeneous and Interpretable Groups for Conformal Prediction (Supplementary Material)

---

## A APPENDIX

### A.1 PROOFS

Restatement of Proposition 5.1

**Proposition A.1.** *Given the objective in Eq. 8 if  $\mathcal{D}^1 = P(X, S)$  (infinite sample regime) and  $\theta_0$  in Algorithm 1 is the weakest admissible regularization, then  $\tau^* = \tau_{\theta_0}$ , which also minimizes pinball loss over all admissible regularizations  $\mathbb{E}_{\mathcal{D}}[\ell_{1-\alpha}(q_{\tau^*}(X), S)] \leq \mathbb{E}_{\mathcal{D}}[\ell_{1-\alpha}(q_{\tau_\theta}(X), S)]$ ,  $\forall \theta \in \Theta$  such that  $\theta \geq \theta_0$ .*

*Proof.* We first show that in the infinite sample regime the MCR is zero  $\forall \theta \in \Theta$ , making all  $\theta$  equivalent according to the MCR criteria. Then we show that Algorithm 1 would choose  $\theta^* = \theta_0$  and since  $\theta_0$  is the lowest regularization it achieves the smallest expected pinball loss.

Given access to the real distribution  $\mathcal{D}_1 = P(X, S)$  for any  $\theta \in \Theta$  we get a finite set partition  $\mathcal{G}_{\tau_\theta}$  such that the  $1 - \alpha$  quantile estimate  $q_{\tau_\theta}(X)$  is the exact group conditional quantile of the non-conformity score distribution for the group that contains the instance  $X$ .

$$q_{\tau_\theta}(X) = F_{S|G=g_{\tau_\theta}(X)}^{-1}(1 - \alpha) \quad (14)$$

where  $g_{\tau_\theta}(X) \in \mathcal{G}_{\tau_\theta}, \forall X \in \mathcal{X}$ . Then, in this asymptotic regime the group conditional miscoverage (Definition 4.1)  $MC_\alpha(q_{\tau_\theta}, g_{\tau_\theta}; g_j) = 0 \forall g \in \mathcal{G}_{\tau_\theta}, \forall g \in \mathcal{G}_{\tau_\theta}$  and  $\forall \theta \in \Theta$ . Then  $MCR_\alpha(\tau_\theta)$  as defined in Eq. 7 is 0  $\forall \theta \in \Theta$ .

Since Algorithm 1 terminates on the first  $\theta$  that achieves the minimum MCR then  $\theta^* = \theta_0$ . Since  $\theta_0$  is the weakest regularization, and we assume infinite sample regime to learn  $\tau_\theta \forall \theta \in \Theta$  then  $\mathbb{E}_{\mathcal{D}}[\ell_{1-\alpha}(q_{\tau^*}(X), S)] \leq \mathbb{E}_{\mathcal{D}}[\ell_{1-\alpha}(q_{\tau_\theta}(X), S)]$ ,  $\forall \theta \in \Theta$  such that  $\theta \geq \theta_0$ . □

### A.2 ADDITIONAL EXPERIMENTS

Figure 3b shows the decision trees that were obtained for the different datasets. We observe that the discovered regions have different prediction interval widths indicating that the model's prediction uncertainty is significantly different. Figure 4 shows the scatter and joint distribution between the prediction interval widths and coverage of the discovered groups. It extends Figure 2 in the main manuscript including all datasets and the groups discovered by the RF-G approach proposed by Amoukou and Brunel [2023]. Table 2 shows the same comparison presented in Table 1 but for a LASSO base model

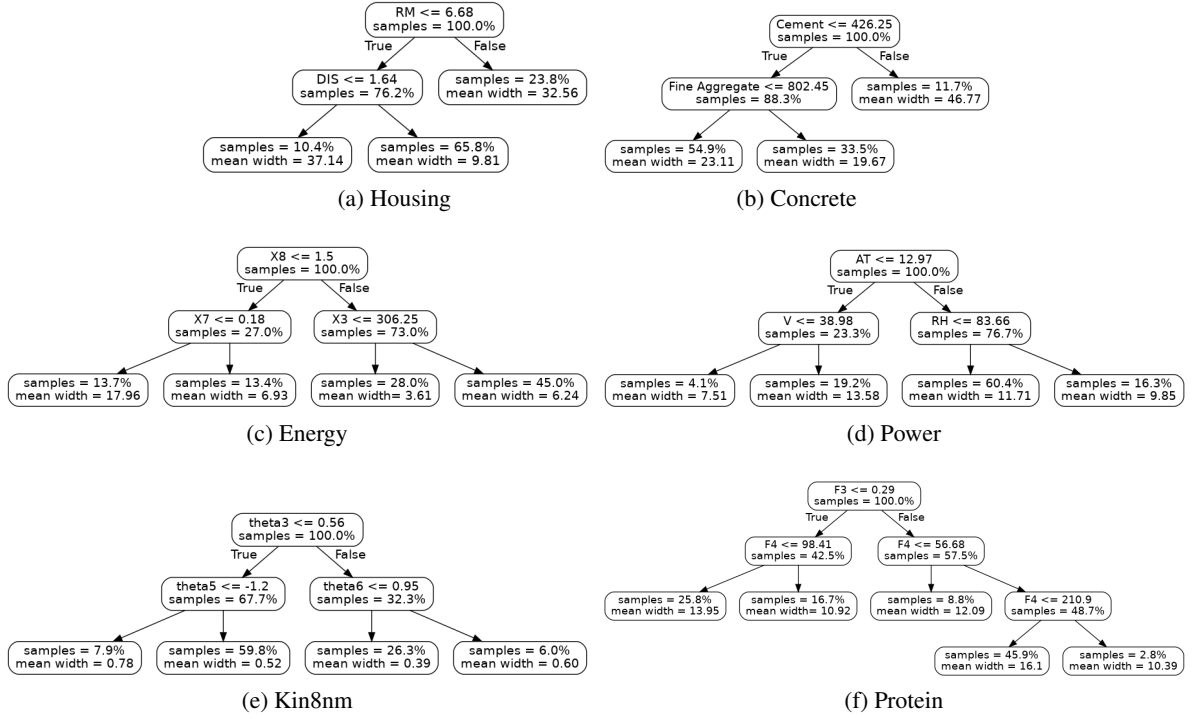


Figure 3: Example of decision trees identified for each regression dataset. (3a) In the Housing dataset groups are defined based on the features corresponding to average number of rooms per dwelling (RM) and weighted distances to five Boston employment centers (DIS). (3b) In the Concrete dataset the groups are defined based on the Cement and Fine Aggregate components ( $kg$  in a  $m^3$  mixture). (3c) the groups in the Energy dataset are defined based on Glazing Area Distribution (X8), Glazing Area (X7) and Wall Area (X3). (3d) In the Power dataset groups are defined based on Ambient Temperature (AT), Exhaust Vacuum (V) and Relative Humidity (RH). (3e) In the kin8nm dataset the groups are defined by the measurements on sensors from links 3, 5 and 6 from the robot arm. (3f) In the protein dataset the groups are defined by the features corresponding to fractional area of exposed non polar residue (F3) and fractional area of exposed non polar part of residue (F4).

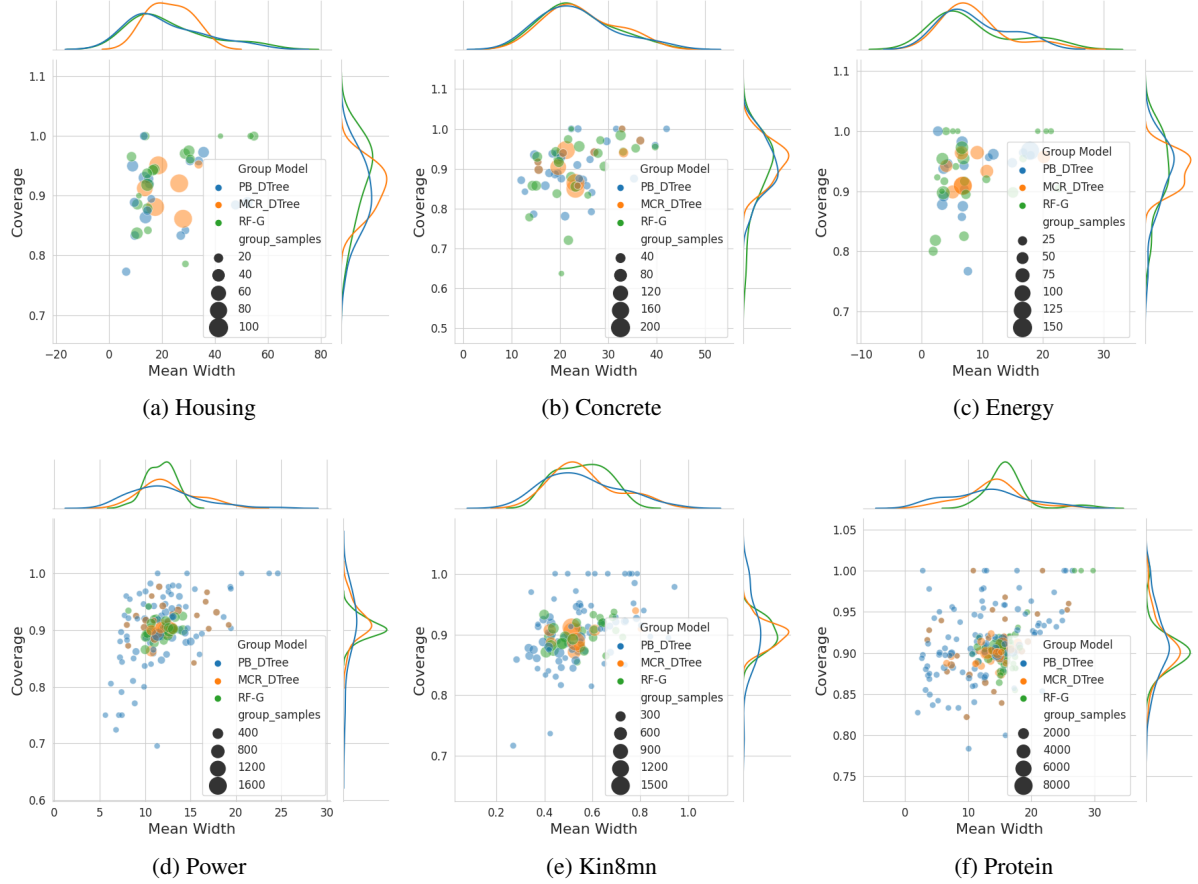


Figure 4: Scatter and distribution plot of the prediction interval widths (x-axis) versus coverage (y-axis) of the groups discovered by the proposed MCR\_DTREE, PB\_DTREE and RF-G methods across 6 datasets. Here we plot all the groups obtained across 5-Fold realizations. The size of the groups points represents the group size (number of samples). The target coverage is 0.9, we observe that MCR\_DTREE tends to identify a smaller number of groups of varying sizes, with group-conditional coverages concentrated around the 0.9 objective. Moreover, the identified groups show diversity in the range of interval widths. PB\_DTREE detects a significant larger number of (smaller) groups, with a larger variance in terms of group-conditional coverage.

model	MCR		coverage		num groups
		average	max group	min group	
Housing: nsamples = 506, nfeatures = 13   LASSO-Regressor R2 = 0.69 ± 0.04					
LCP-RF-G	2.71±0.77	0.8±0.06	0.91±0.08	0.75±0.07	2.6±0.55
RF-G	0.42±0.38	<b>0.91±0.03</b>	0.96±0.03	0.81±0.15	3.2±0.45
PB-KMEANS	1.47±0.49	0.86±0.03	0.98±0.03	0.44±0.43	14.2±15.02
MCR-KMEANS	1.35±0.74	0.88±0.04	0.97±0.03	0.69±0.38	7.4±11.52
PB_DTREE	0.32±0.21	0.88±0.03	0.98±0.05	0.83±0.05	4.0±1.87
MCR_DTREE	<b>0.25±0.39</b>	0.89±0.04	<b>0.95±0.04</b>	<b>0.84±0.07</b>	3.6±2.07
Concrete: nsamples = 1030, nfeatures = 8   LASSO-Regressor R2 = 0.60 ± 0.05					
LCP-RF-G	1.37±1.12	0.83±0.02	<b>0.96±0.04</b>	0.7±0.05	5.4±0.55
RF-G	0.29 ± 0.15	0.91±0.02	0.98±0.03	0.8±0.08	5.0±0.71
PB-KMEANS	0.89±0.48	<b>0.9±0.05</b>	1.0±0.0	0.26±0.37	37.2±16.93
MCR-KMEANS	0.43±0.43	0.92±0.02	0.97±0.03	0.7±0.3	15.8±18.98
PB_DTREE	0.25±0.14	<b>0.9±0.03</b>	1.0±0.0	0.8±0.07	7.0±2.24
MCR_DTREE	<b>0.15±0.09</b>	<b>0.9±0.03</b>	1.0±0.0	<b>0.84±0.04</b>	6.8±2.39
Energy: nsamples = 768, nfeatures = 8   LASSO-Regressor R2 = 0.91 ± 0.005					
LCP-RF-G	0.38±0.19	0.88±0.05	<b>0.98±0.03</b>	0.8±0.08	4.8±0.45
RF-G	0.12±0.12	<b>0.94±0.02</b>	1.0±0.0	0.87±0.06	5.0±0.71
PB-KMEANS	1.07±0.77	0.87±0.04	0.99±0.02	0.18±0.4	38.2±19.15
MCR-KMEANS	0.32±0.41	<b>0.94±0.03</b>	<b>0.98±0.04</b>	0.83±0.13	13.0±11.92
PB_DTREE	0.12±0.16	<b>0.94±0.02</b>	0.99±0.03	0.84±0.11	9.0±3.46
MCR_DTREE	<b>0.05±0.09</b>	<b>0.94±0.02</b>	<b>0.98±0.02</b>	<b>0.89±0.03</b>	6.0±3.24
Power: nsamples = 9568, nfeatures = 4   LASSO-Regressor R2 = 0.93 ± 0.003					
LCP-RF-G	2.04±1.26	0.82±0.05	0.86±0.08	0.78±0.05	6.0±2.24
RF-G	0.83±0.57	<b>0.9±0.0</b>	<b>0.93±0.02</b>	0.87±0.01	5.2±0.84
PB-KMEANS	0.73±0.27	0.91±0.01	0.99±0.02	0.78±0.05	37.2±5.22
MCR-KMEANS	0.46±0.15	<b>0.9±0.0</b>	<b>0.93±0.03</b>	<b>0.88±0.03</b>	6.0±7.28
PB_DTREE	0.08±0.05	<b>0.9±0.01</b>	0.94±0.03	0.87±0.02	6.4±4.16
MCR_DTREE	<b>0.06±0.05</b>	<b>0.9±0.0</b>	0.94±0.01	<b>0.88±0.02</b>	7.4±3.71
Protein: : nsamples = 45730, nfeatures = 9   LASSO-Regressor R2 = 0.28 ± 0.01					
LCP-RF-G	0.89±0.56	0.87±0.03	<b>0.92±0.02</b>	0.75±0.04	5.8±1.6
RF-G	0.44±0.37	<b>0.9±0.0</b>	0.95±0.05	0.87±0.02	6.00±1.59
PB-KMEANS	0.71±0.75	<b>0.9±0.0</b>	1.0±0.0	0.65±0.21	42.6±7.86
MCR-KMEANS	0.52±0.21	<b>0.9±0.0</b>	0.96±0.05	0.76±0.24	16.2±12.91
PB_DTREE	0.44±0.37	<b>0.9±0.0</b>	1.0±0.0	0.83±0.02	15.6±0.89
MCR_DTREE	<b>0.2±0.08</b>	<b>0.9±0.0</b>	0.93±0.03	<b>0.89±0.01</b>	5.6±2.19
kin8mn: : nsamples = 8192, nfeatures = 8   LASSO-Regressor R2 = 0.40 ± 0.007					
LCP-RF-G	1.68±0.29	0.79±0.01	0.81±0.01	0.77±0.01	3.0±0.0
RF-G	<b>0.21±0.04</b>	<b>0.9±0.01</b>	<b>0.91±0.01</b>	<b>0.88±0.0</b>	3.2±0.45
PB-KMEANS	0.67±0.16	0.92±0.01	0.99±0.01	0.76±0.04	39.4±14.06
MCR-KMEANS	0.44±0.37	<b>0.9±0.01</b>	0.93±0.04	0.87±0.05	11.6±21.47
PB_DTREE	0.41±0.36	0.89±0.01	0.98±0.04	0.82±0.07	14.2±3.03
MCR_DTREE	0.24±0.18	<b>0.9±0.01</b>	0.94±0.04	<b>0.88±0.02</b>	6.4±5.37

Table 2: Comparison between the group discovery partition methods. We show MCR, marginal, minimum, and maximum coverage group coverage on the identified partition. We also report the number of groups per approach. Standard deviations are computed across 5 data splits. The proposed MCR\_DTREE is consistently better in terms of MCR, with values consistently below 1, indicating that the discovered groups improve worst-group under-coverage w.r.t. to single threshold SCP. Every dataset uses a LASSO regressor as the base model. We highlight the lowest MCR and the smallest average coverage above the objective (0.9). For methods that achieved the marginal coverage objective we highlight the max and min group coverage closest to the 0.9 objective.