On split conformal prediction

Theorem (marginal validity of SCP): Let $(X_i, Y_i)_{i=1}^{n+1}$ be exchangeable. SCP applied on $(X_i, Y_i)_{i=1}^{n}$ outputs $\hat{C}_\alpha(\cdot)$ such that:

$$\mathbb{P}(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})) \in [1-\alpha ; 1-\alpha + \frac{1}{n_{cal}+1}[,$$

where the upper bound holds if the scores $(S_i)_{i \in Cal} \cup \{S_{n+1}\}$ are a.s. distincts.

♡

Proof: A) To begin, let's write explicitly:

$$\{Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\} = \{\hat{\mu}(X_{n+1}) - q_{1-\alpha}(S) \leq Y_{n+1} \leq \hat{\mu}(X_{n+1}) + q_{1-\alpha}(S)\}$$

$$= \{|Y_{n+1} - \hat{\mu}(X_{n+1})| \leq q_{1-\alpha}(S)\}$$

$$\{Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\} = \{S_{n+1} \leq q_{1-\alpha}(S)\} \quad (♣)$$

B) (♣) makes us want to use the quantile lemma! But how do $q_{1-\alpha}(S)$ and $q_{1-\alpha}((S_i)_{i \in Cal}, S_{n+1})$ relate?

Recall that $q_{1-\alpha}(S) = q_{1-\alpha}((S_i)_{i \in Cal}, +\infty)$.

In fact, by a careful analysis of the order statistics, we have that $\{S_{n+1} \leq q_{1-\alpha}(S)\} = \{S_{n+1} \leq q_{1-\alpha}((S_i)_{i \in Cal}, S_{n+1})\}$.

Indeed: → noting that $q_{1-\alpha}((S_i)_{i \in Cal}, S_{n+1}) \leq q_{1-\alpha}(S) = q_{1-\alpha}((S_i)_{i \in Cal}, +\infty)$,

we have $\{S_{n+1} \leq q_{1-\alpha}((S_i)_{i \in Cal}, S_{n+1})\} \subseteq \{S_{n+1} \leq q_{1-\alpha}(S)\}$.

→ to prove the reverse inclusion, remark that if S_{n+1} is such that

$S_{n+1} > q_{1-\alpha}((S_i)_{i \in Cal}, S_{n+1})$, it must hold that

$$q_{1-\alpha}((S_i)_{i \in Cal}, S_{n+1}) = q_{1-\alpha}((S_i)_{i \in Cal}, +\infty) = q_{1-\alpha}(S).$$

$S_{n+1} > S_{(\lceil (1-\alpha)(n+1) \rceil)}$ so replacing it by $+\infty$ has no impact.

C) To conclude, note that by exchangeability of $(X_i, Y_i)_{i=1}^{n+1}$, we have that $(S_i)_{i \in \mathcal{C}al} \cup \{S_{n+1}\}$ are exchangeable (even conditionally on the training data).

for any $j \in [\![1, n+1]\!]$, we can write S_j as $g(X_j, Y_j)$, with g a deterministic function conditional on the training data (or with a remaining randomness independent of $(X_i, Y_i)_{i \in \mathcal{C}al} \cup (X_{n+1}, Y_{n+1})$).

Hence, we conclude using the quantile lemma on the scores:

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right) = \mathbb{P}\left(S_{n+1} \leq q_{1-\alpha}\left((S_i)_{i \in \mathcal{C}al}, S_{n+1}\right)\right)$$

$$\in \left[1-\alpha, 1-\alpha + \frac{1}{n_{\mathcal{C}al}+1}\right[$$

with the upper bound when $(S_i)_{i \in \mathcal{C}al} \cup \{+\infty\}$ are a.s. distincts.

$\square$

Lemma (CQR):

$$\{Y_{n+1} \notin \hat{C}_\alpha(X_{n+1})\} = \{Y_{n+1} < \widehat{QR}_{lower}(X_{n+1}) - q_{1-\alpha}(S)$$
$$\text{or } Y_{n+1} > \widehat{QR}_{upper}(X_{n+1}) + q_{1-\alpha}(S)\}$$
$$= \{\widehat{QR}_{lower}(X_{n+1}) - Y_{n+1} > q_{1-\alpha}(S)$$
$$\text{or } Y_{n+1} - \widehat{QR}_{upper}(X_{n+1}) > q_{1-\alpha}(S)\}$$
$$= \{\max\left(\widehat{QR}_{lower}(X_{n+1}) - Y_{n+1}, Y_{n+1} - \widehat{QR}_{upper}(X_{n+1})\right) > q_{1-\alpha}(S)\}$$

$$\{Y_{n+1} \notin \hat{C}_\alpha(X_{n+1})\} = \{S_{n+1} > q_{1-\alpha}(S)\}$$

$$\iff \{Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\} = \{S_{n+1} \leq q_{1-\alpha}(S)\}$$

☆