Multivariate Conformal Selection

Tian Bai Yue Zhao Xiang Yu Archer Y. Yang

ICML 2025

July 4, 2025

Problem Setup

- p-dimensional features $\boldsymbol{x}$
- d-dimensional response $\boldsymbol{y}$
- i.i.d. Data
  - **Training** $\mathcal{D}_{\text{train}} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{n}$
  - **Test** $\mathcal{D}_{\text{test}} = \{\boldsymbol{x}_{n+j}\}_{j=1}^{m}$, unobserved $\{\boldsymbol{y}_{n+j}\}_{j=1}^{m}$
- **Goal**: to identify a subset $\mathcal{S} \subseteq \{1, \ldots, m\}$ from $\mathcal{D}_{\text{test}}$, s.t. as many test obs. $j \in \mathcal{S}$ as possible satisfy

$$\boldsymbol{y}_{n+j} \in R$$

  where $R \in \mathbb{R}^d$ is a predefined region, with *FDR control*.
- Generalizes Jin and Candès, 2023, which works for univariate response ($d = 1$).

Problem Setup

- False discovery rate (FDR):

$$\text{FDR} = \mathbb{E}\left[\frac{|\mathcal{S} \cap \mathcal{H}_0|}{|\mathcal{S}|}\right] \leq q$$

  should be controlled, where $\mathcal{H}_0 = \{j : y_{n+j} \notin R\}$.

- A good selection procedure $\mathcal{S}$ gives high power:

$$\text{Power} = \mathbb{E}\left[\frac{|\mathcal{S} \cap \mathcal{H}_1|}{|\mathcal{H}_1|}\right]$$

  where $\mathcal{H}_1 = \{j : y_{n+j} \in R\}$.

Multivariate Conformal Selection

- For $j \in \{1, \ldots, m\}$, mCS performs

$$H_{0j} : \boldsymbol{y}_{n+j} \in R^c \quad \text{vs.} \quad H_{1j} : \boldsymbol{y}_{n+j} \in R$$

- mCS consists of three main steps:

1. **Training:** Construct a predictive model $\hat{\mu}$ for $\boldsymbol{y}$.
2. **Calibration:**
    1. Build a regionally monotone nonconformity score based on $\hat{\mu}$.
    2. Compute the conformal p-value for the tests
3. **Thresholding:** Apply the BH procedure

Oracle Conformal p-values

- Assuming a nonconformity score $V : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, a measure of atypicality of the pair (x, y),
- *Oracle* conformal p-values: if the true $\{\boldsymbol{y}_{n+j}\}_{j=1}^{m}$ were observed,

$$p_j^* = \frac{\sum_{i=1}^{n} \mathbb{1}\{V_i < V_{n+j}\} + 1}{n+1}$$

where $V_i = V(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for $i = 1, \ldots, n+m$.

Practical Conformal p-values

- *Oracle* conformal p-values requires knowing unobserved $\boldsymbol{y}_{n+j}$.
- In practice, replace V_{n+j} with

$$\widehat{V}_{n+j} = V(\boldsymbol{x}_{n+j}, \boldsymbol{r}_{n+j}),$$

  where $\boldsymbol{r}_{n+j}$ is an arbitrarily chosen in R.

- (Practical) conformal p-values

$$p_j = \frac{\sum_{i=1}^{n} \mathbb{1}\{V_i < \widehat{V}_{n+j}\} + 1}{n+1}.$$

Regional Monotonicity

- By i.i.d. data assumption, Oracle conformal p-values is super-uniform (Vovk et al., 2005)

$$\mathbb{P}(p_j^* \leq \alpha) \leq \alpha$$

- To ensure

$$\mathbb{P}(p_j \leq \alpha) \leq \alpha$$

  V must satisfy *regional monotonicity*.

- **Regional Monotonicity (RM):**

$$V(\boldsymbol{x}, \boldsymbol{y}') \leq V(\boldsymbol{x}, \boldsymbol{y}) \quad \textbf{for any} \quad \boldsymbol{y}' \in R^c \textbf{ and } \boldsymbol{y} \in R$$

Choices of Nonconformity Score

- The selection power heavily depends on the quality of the chosen score.
- In the context of CP (Romano et al., 2019; Kivaranovic et al., 2020; Sesia & Candes, 2020).
- Limited focus for CS.

Two Types of RM Scores

- Distance-based scores (clipped score, Jin and Candes, 2023):

$$V(\boldsymbol{x}, \boldsymbol{y}) = M \cdot \mathbb{1}\{\boldsymbol{y} \notin R^c \cup \partial R\} - \inf_{\boldsymbol{s} \in R^c} \|\boldsymbol{y} - \boldsymbol{s}\|_p,$$

- Learning-based scores (Stutz et al., 2021, Xie et al., 2024):

$$V^\theta(\boldsymbol{x}, \boldsymbol{y}) = M \cdot \mathbb{1}\{\boldsymbol{y} \notin R^c \cup \partial R\} - f_\theta(\boldsymbol{x}, \boldsymbol{y}; R)$$

Distance-based Scores

- The second term $\inf_{\boldsymbol{s} \in R^c} \|\boldsymbol{y} - \boldsymbol{s}\|_p$ measures the distance between $\hat{\mu}(\boldsymbol{x})$ and R^c:
  - If $\hat{\mu}(\boldsymbol{x})$ moves away from R^c
  - Then the distance increases, leading to smaller test scores $\widehat{V}_{n+j}$ and smaller p-values
  - Thus, data with y in the interior of R are more likely to be selected by the BH.
- Selecting $\boldsymbol{r}_{n+j}$ on ∂R is optimal for power.

Learning-based Nonconformity Scores

For distance-based scores:

- Low power when R is a nonconvex;
- Constructing a closed-form distance function can be challenging when R is irregular.

Learning-based Nonconformity Scores

- `mCS-learn` learn an optimal nonconformity score within the family:

$$V^{\theta}(\boldsymbol{x}, \boldsymbol{y}) = M \cdot \mathbb{1}\{\boldsymbol{y} \notin R^{c} \cup \partial R\} - f_{\theta}(\boldsymbol{x}, \boldsymbol{y}; R)$$

- $f_{\theta} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a function from a specific ML class, e.g.
  - Kernel machines
  - Gradient boosting
  - Neural networks, etc.

Learning score function f_θ

- Introduce a differentiable loss function that mimics the non-differentiable mCS procedure.
- "hard" ranking is replaced with ***soft-ranking*** (Blondel et al., 2020; Cuturi et al., 2019).
- Use two hold-out datasets $\mathcal{D}_{f\text{-train}}$ and $\mathcal{D}_{f\text{-val}}$ (can be obtained by data splitting) for training f_θ.

Smooth conformal p-values

- Sample two disjoint subsets $\mathcal{D}_{f\text{-train1}}$ and $\mathcal{D}_{f\text{-train2}}$ from $\mathcal{D}_{f\text{-train}}$
- Let $n' = |\mathcal{D}_{f\text{-train1}}|$ and $m' = |\mathcal{D}_{f\text{-train2}}|$
- soft-rank$(a; A)$ ≡ the softened rank of element a within the set A.
- The smooth conformal p-values for $j = 1, \ldots, m'$

$$\bar{p}_j^{\theta} = \frac{\text{soft-rank}\big(\widehat{V}_{n'+j}^{\theta}; \{V_i^{\theta}\}_{i=1}^{n'} \cup \{\widehat{V}_{n'+j}^{\theta}\}\big)}{n'+1}.$$

- Loss function
  - $L(\theta) = \sum_{j=1}^{m'} \bar{p}_j^{\theta}\big[\mathbb{1}(\boldsymbol{y}_{n+j} \in R) - \gamma \cdot \mathbb{1}(\boldsymbol{y}_{n+j} \in R^c)\big]$.
  - $L(\theta) = -\bar{S}(\theta)$, the BH outcome with the smooth p-values.

Learning-based mCS Algorithm

1: Initialize parameters $\theta = \theta_0$.

2: **for** epoch $t = 1, \ldots, T$ **do**

3: Sample two disjoint subsets $\mathcal{D}^{(t)}_{f\text{-train1}}$ and $\mathcal{D}^{(t)}_{f\text{-train2}}$.

4: Use the current f_θ to obtain V_i^θ from $\mathcal{D}^{(t)}_{f\text{-train1}}$ and $\widehat{V}^\theta_{n+j}$ from $\mathcal{D}^{(t)}_{f\text{-train1}}$.

5: Compute the smooth conformal p-values $\bar{p}_j^\theta$ and the loss function.

6: Update model parameters $\theta = \theta_t$.

7: Applying mCS on $\mathcal{D}_{f\text{-val}}$ k times and record the average power.

8: **end for**

9: Use $\mathcal{D}_{f\text{-val}}$ for validation to obtain the optimal epoch t^*.

10: Return $f_{\theta_{t^*}}$.

ADMET Data

- ADMET dataset, compiled from various public sources (Wenzel et al., 2019; Iwata et al., 2022; Kim et al., 2023; Watanabe et al., 2018; Falcon-Cano et al., 2022; Esposito et al., 2020; Braga et al., 2015; Aliagas et al., 2022; Perryman et al., 2020; Meng et al., 2022; Vermeire et al., 2022).
- $n = 20K \sim 200K$
- Biological activities $\boldsymbol{y} \in \mathbb{R}^d$, $d = 15$
- Molecular structure-derived features $\boldsymbol{x} \in \mathbb{R}^{1024}$
- Two selection tasks
  1. The (shifted) first orthant, $R = \{\boldsymbol{y} : y_k \geq c_k \quad \forall k\}$
  2. A sphere centered at $\boldsymbol{c}$, $R = \{\boldsymbol{y} : ||\boldsymbol{y} - \boldsymbol{c}||_2 \leq r\}$

Baseline methods

- `CS_int` Rectangular target region $\mathcal{S} = \cap_{k=1}^{d} \mathcal{S}_k$, each dimension controlled by $q_k = q$
- `CS_ib` Like `CS_int`, but controlled by $q_k = q/d$ (too conservative)
- `CS_is` Like `CS_int`, but controlled by an adaptive q_k (Sheridan)
- `binary` Univariate CS with pseudo outcomes $\tilde{y}_i = \mathbb{1}(\boldsymbol{y}_i \in R)$

Performance Comparison

Table 19: Observed FDR of different methods for the first drug discovery task.

q	CS_int	CS_ib	CS_is	bi	mCS-d, score (7)	mCS-d, score (8)	mCS-l
0.3	0.760	0.000	0.303	0.038	0.290	0.304	0.275
0.5	0.782	0.393	0.496	0.040	0.417	0.499	0.488

Table 20: Observed power of different methods for the first drug discovery task.

q	CS_int	CS_ib	CS_is	bi	mCS-d, score (7)	mCS-d, score (8)	mCS-l
0.3	0.993	0.000	0.019	0.000	0.003	0.006	0.010
0.5	1.000	0.003	0.225	0.000	0.159	0.433	0.193

Table 21: Observed FDR of different methods for the second drug discovery task.

q	bi	mCS-d, score (7)	mCS-d, score (8)	mCS-l
0.3	0.000	0.207	0.300	0.293
0.5	0.000	0.338	0.499	0.498

Table 22: Observed power of different methods for the second drug discovery task.

q	bi	mCS-d, score (7)	mCS-d, score (8)	mCS-l
0.3	0.000	0.139	0.278	0.086
0.5	0.000	0.382	0.759	0.515

Thank you!