Selection by Prediction with (Weighted) Conformal p-values

Ying Jin

Joint work with Emmanuel Candès

Department of Statistics, Stanford University

International Seminar on Selective Inference, May 17, 2023

ML prediction assists decision

[VerVoe.com]

[smartdreamers.com]

[HeroHunt.ai]

- Job hiring: who to reach out to? who to select for interview?

ML prediction assists discovery

Deep Learning

Shortcuts to Simulation: How Deep Learning Accelerates Virtual Screening for Drug Discovery

May 11, 2020 14 min read

[DZone.com]

Automating Drug Discovery With Machine Learning

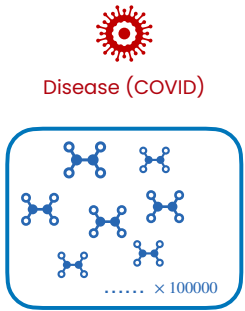
Article Published: April 16, 2021 | Neeta Ratanghayra, MPharm

[technologynetworks.com]

- Drug discovery: which molecules/compounds to proceed to screening and clinical trials?

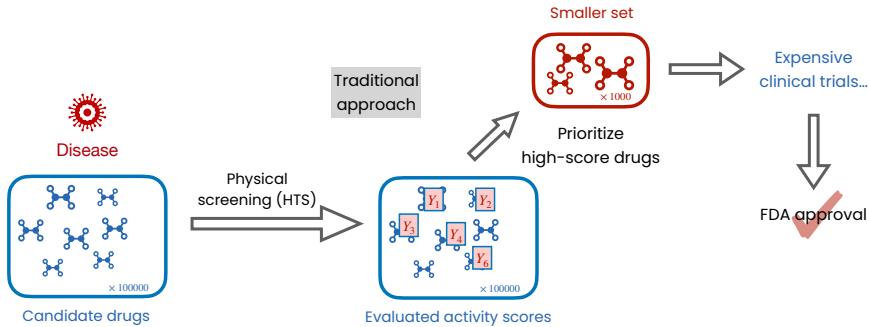
Decision and discovery processes

- Finding a few interesting cases from a huge pool



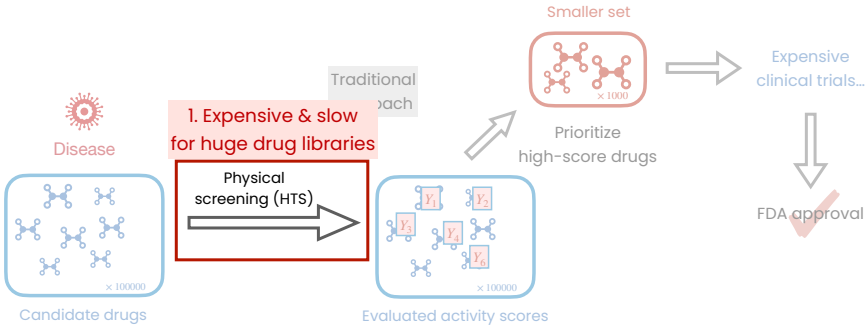
Decision and discovery processes

- Finding a few interesting cases from a huge pool



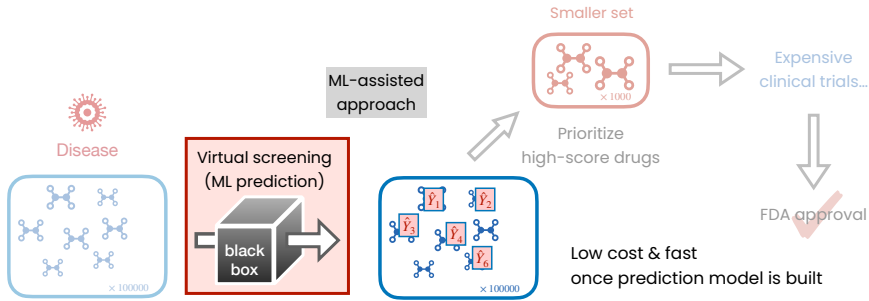
Decision and discovery processes

- Finding a few interesting cases from a huge pool



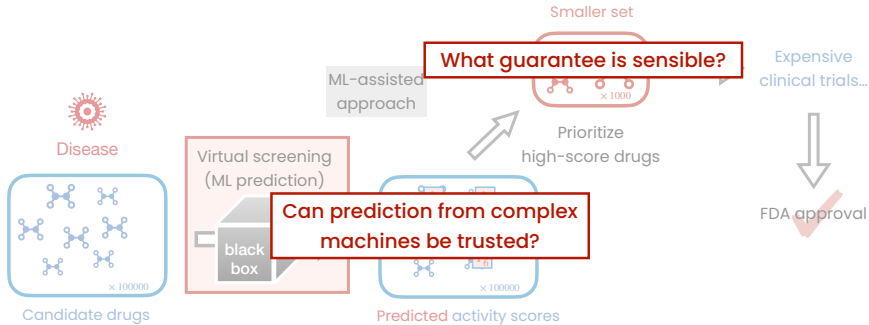
ML in decision and discovery processes

- Accelerating discovery via machine learning prediction



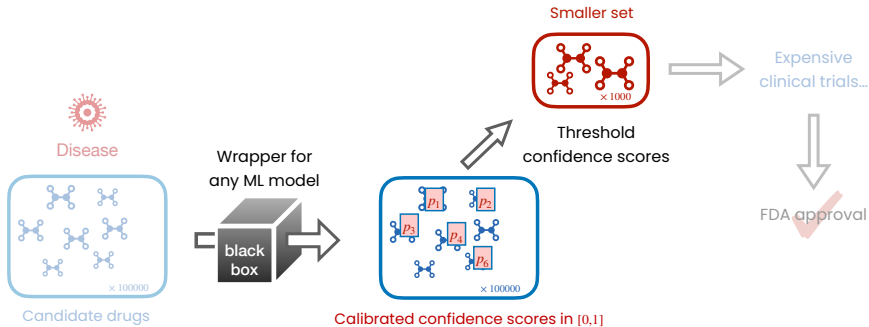
ML in decision and discovery processes

- Accelerating discovery via machine learning prediction



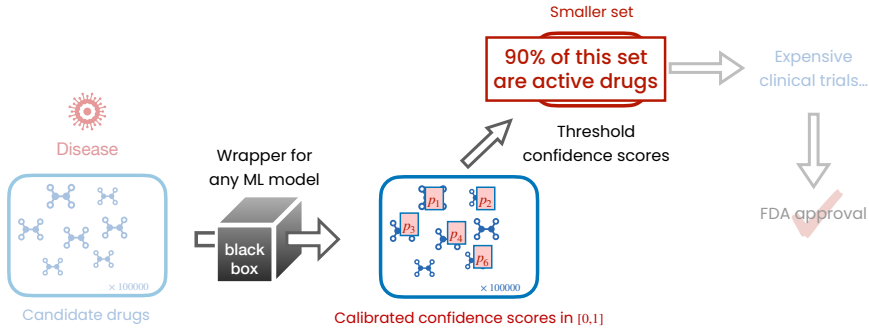
Our proposal

- Drug discovery with error control on the selected



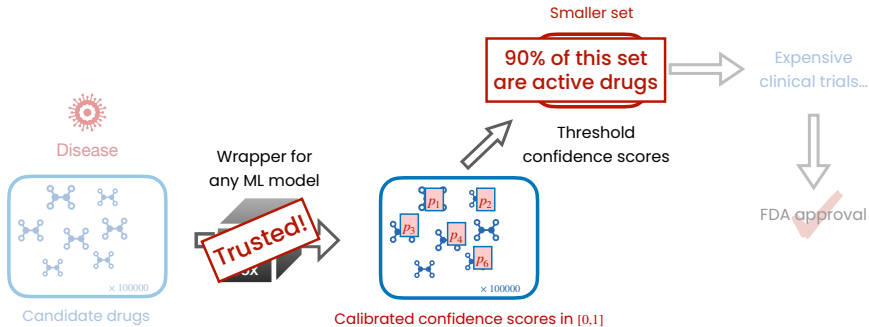
Our proposal

- Drug discovery with error control on the selected



Our proposal

- Drug discovery with error control on the selected



Identify a few interesting cases from a huge pool

- Problem setting
  - Any pre-trained model $\widehat{\mu}\colon \mathcal{X} \to \mathcal{Y}$
    - X: physical/chemical features of the drug
    - Y: activity score should we physically screen the drug
    - $Y \in \{0, 1\}$: whether the drug is active for the disease
    - $Y \in \mathbb{R}$: how active the drug is for the disease
  - Training data $(X_i, Y_i) \sim \mathbb{P}$, $i = 1, \ldots, n$. (already-screened drugs)
  - Test samples $(X_{n+j}, Y_{n+j}) \sim \mathbb{P}$, $j = 1, \ldots, m$. (new/other drugs in the library)
- Interesting $\Leftrightarrow$ the unseen outcome is large $Y_{n+j} > c_{n+j}$
  - highly competent candidates, highly effective drugs
  - c_{n+j}: how active should the drug Y_{n+j} be to be considered 'interesting' (pre-specified)

Predicting the unobserved responses: conformal prediction

- (Split) conformal inference [Vovk et al., 2005]
  - Find any nonconformity score $V\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ (such as $V(x, y) = -|y - \widehat{\mu}(x)|$)
  - Compute $V_i = V(X_i, Y_i)$ for $i = 1, \ldots, n$
  - Construct prediction intervals

$$\widehat{C}(X_{n+j}; \alpha) = \left\{ y\colon\ V(X_{n+j}, y) \geq \texttt{Quantile}\left(\alpha, \textstyle\sum_{i=1}^{n} \frac{1}{n+1} \delta_{V_i} + \frac{1}{n+1} \delta_{-\infty}\right) \right\}$$

  - Distribution-free guarantee

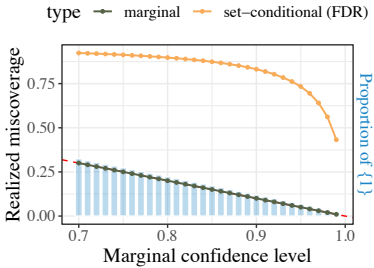
$$\mathbb{P}(Y_{n+j} \in \widehat{C}(X_{n+j}; \alpha)) \geq 1 - \alpha$$

    for each j (marginalized over $\{X_i, Y_i\}_{i=1}^{n}$ and (X_{n+j}, Y_{n+j}))

- A literature on conformal prediction for drug discovery [Norinder et al., 2014, Svensson et al., 2017, Ahlberg et al., 2017, Svensson et al., 2018, Cortes-Ciriano and Bender, 2019, Wang et al., 2022]
  - Build prediction sets and identify promising drugs

Is validity for one single point sufficient?

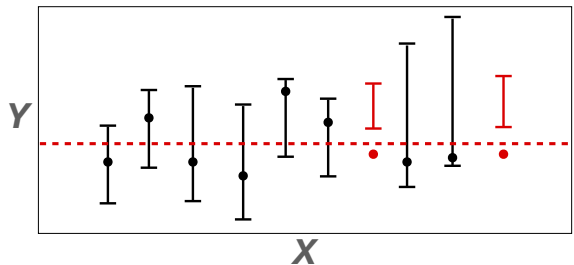
- Consider the binary case, where $Y = 1$ is of interest
  - Conformal prediction sets take the form $\{0\}, \{1\}, \{0, 1\}$
  - $\mathbb{P}(Y_{n+j} \in \widehat{C}(X_{n+j}; \alpha)) \geq 1 - \alpha$, over $\{X_i, Y_i\}_{i=1}^n$ and (X_{n+j}, Y_{n+j})
- What if we construct $\widehat{C}(X_{n+j}; \alpha)$ and choose those $\widehat{C}(X_{n+j}; \alpha) = \{1\}$?
  - Coverage on average does not imply coverage on selected ones



- x-axis is marginal coverage level $1 - \alpha$
- dark curve is miscoverage for all test data
- orange curve is miscoverage for those $\widehat{C}(X_{n+j}; \alpha) = \{1\}$

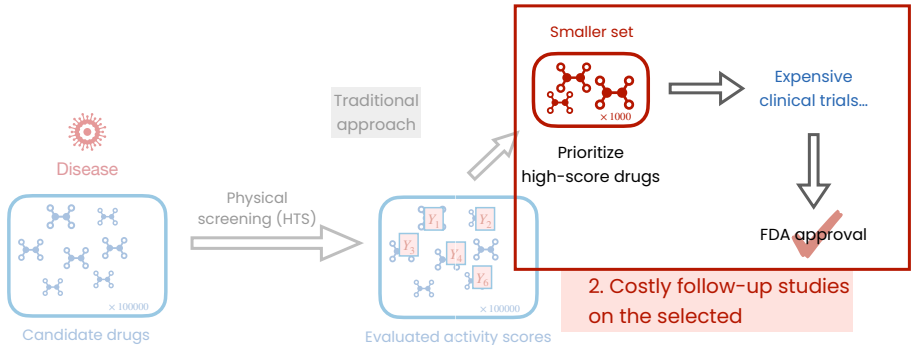
The selection issue with multiple decisions

- What if we construct $\widehat{C}(X_{n+j}; \alpha)$ and choose those seemingly promising ones?
  - Coverage on average does not imply coverage on selected ones



Decision and discovery processes

- Error on the selected is more of concern



Error control on the selected

- We want to select those $Y_{n+j} > c_{n+j}$ among test samples
  - Training data $(X_i, Y_i) \sim \mathbb{P}$, $i = 1, \ldots, n$. (already-screened drugs)
  - Test samples $(X_{n+j}, Y_{n+j}) \sim \mathbb{P}$, $j = 1, \ldots, m$. (new/other drugs in the library)
  - c_{n+j}: how active should the drug be to be considered 'interesting'
- Limiting the proportion of false selections: FDR control

$$\mathbb{E}\left[\frac{\sum_{j=1}^{m} \mathbb{1}\{Y_{n+j} \leq c_{n+j} \text{ but selected}\}}{1 \vee \sum_{j=1}^{m} \mathbb{1}\{Y_{n+j} \text{ selected}\}}\right] \leq q$$

- Why counting the error? Cost of follow-up studies, cost of interviews, cost of a missing patient...
- Why proportion? Tradeoff between costs and rewards

Selection by prediction with conformal p-values

- Testing random hypotheses

$$H_j\colon\ Y_{n+j} \le c_{n+j}, \quad j = 1, \dots, m.$$

- Rejecting H_j means claiming $Y_{n+j} > c_{n+j}$
- Our idea: construct p-values for these hypotheses and do classical

Selection by prediction with conformal p-values

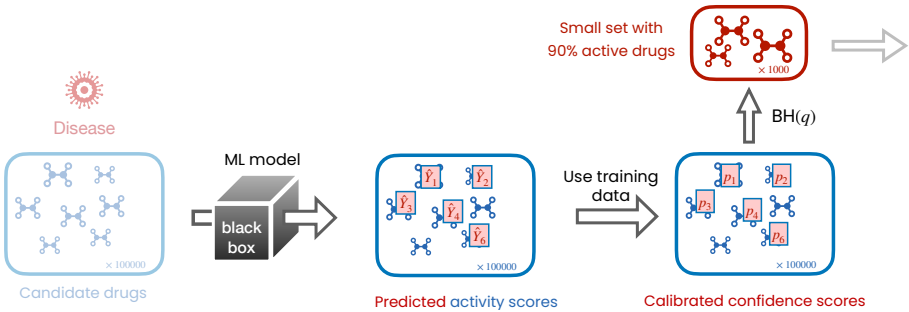
- A general strategy
  - Construct monotone nonconformity score $V(x, y)$, such that $y \leq y'$ implies $V(x, y) \leq V(x, y')$
    - One-sided residual $V(x, y) = y - \widehat{\mu}(x)$
    - Fitted cumulative distribution function $V(x, y) = \widehat{P}(Y \leq y \,|\, X = x)$
  - Construct training scores $V_i := V(X_i, Y_i)$, $i = 1, \ldots, n$
  - Construct test scores $\widehat{V}_{n+j} := V(X_{n+j}, c_{n+j})$, $j = 1, \ldots, m$
  - Obtain selection set by BH(q) procedure with conformal p-values (no ties)

$$p_j = \frac{\sum_{i=1}^{n} \mathbb{1}\{V_i < \widehat{V}_{n+j}\} + U_j}{n+1}, \quad U_j \sim \text{Unif}[0, 1]$$

  - That is, $\mathcal{R} = \{j\colon p_j \leq qk^*/m\}$, where $k^* = \max\{k\colon \sum_{j=1}^{m} \mathbb{1}\{p_j \leq qk/m\} \geq k\}$

The above procedure controls FDR below q for i.i.d. or exchangeable data

Back to the drug discovery pipeline



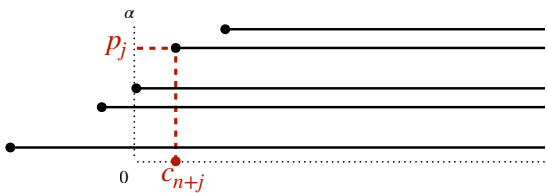
Conformal p-values via inverting conformal prediction intervals

- p_j is the smallest α such that one-sided $(1-\alpha)$ prediction interval excludes (all lies above) c_{n+j}

$$p_j = \inf\left\{\alpha\colon\, c_{n+j} \notin \widehat{C}(X_{n+j};\alpha)\right\}, \quad \text{where}$$

$$\widehat{C}(X_{n+1};\alpha) = \left\{y\colon\, V(X_{n+1},y) \geq \texttt{Quantile}\left(\alpha, \textstyle\sum_{i=1}^{n} \frac{1}{n+1}\delta_{V_i} + \frac{1}{n+1}\delta_{-\infty}\right)\right\}.$$

  - A small p-value indicates that c_{n+j} is smaller than the typical behavior of Y_{n+j}



By monotonicity,
$\hat{C}(X_{n+j};\alpha) = [\eta(X_{n+j};\alpha), \infty)$

P-values for random hypotheses

- In conventional setting with deterministic hypotheses, we often rely on

$$\mathbb{P}(p_j \leq \alpha) \leq \alpha \quad \text{for } j \in \mathcal{H}_0$$

- Our p_j instead satisfies a generalized notion of "type-I error" control:

$$\mathbb{P}(p_j \leq \alpha, j \in \mathcal{H}_0) \leq \alpha,$$

  In particular, it obeys that for some "always null" $p_j^* \sim$ Unif[0,1],

$$p_j \geq p_j^* \quad \text{on the event } \{j \in \mathcal{H}_0\}.$$

- FDR control comes from this null property + PRDS among all p-values

Theory for FDR control

Write $Z_i = (X_i, Y_i)$ for $i = 1, \ldots, n+m$ and $\widetilde{Z}_{n+j} = (X_{n+j}, c_{n+j})$ for $j = 1, \ldots, m$.

Theorem (J. and Candès, 2022)

Suppose V is monotone, the training data $\{Z_i\}_{i=1}^n$ and test data $\{Z_{n+j}\}_{j=1}^m$ are i.i.d., and data in $\{Z_i\}_{i=1}^n \cup \{\widetilde{Z}_{n+\ell}\}_{\ell \neq j} \cup \{Z_{n+j}\}$ are mutually independent for any j. Then, for any $q \in (0,1)$, the output $\mathcal{R}$ of our procedure with input level q satisfies

$$FDR = \mathbb{E}\left[\frac{\sum_{j=1}^m I\{j \in \mathcal{R}, Y_{n+j} \leq c_{n+j}\}}{1 \vee |\mathcal{R}|}\right] \leq q.$$

Theory for FDR control

- Step 1: Leave-one-out
  - Define 'oracle' p-values $p_j^* = \frac{\sum_{i=1}^n \mathbb{1}\{V_i < V_{n+j}\} + U_j}{n+1}$, where $V_{n+j} = V(X_{n+j}, Y_{n+j})$ [Bates et al., 2021]
  - Let $\mathcal{R}_j^* = \mathrm{BH}(q; p_1, \ldots, p_{j-1}, p_j^*, p_{n+j}, \ldots, p_n)$
  - On the event $\{j \in \mathcal{R}, Y_{n+j} \le c_{n+j}\}$, one has $\mathcal{R} = \mathcal{R}_j^*$ and $p_j^* \le p_j$, hence

$$\mathbb{E}\left[\frac{\sum_{j=1}^m \mathbb{1}\{Y_{n+j} \le c_{n+j},\ j \in \mathcal{R}\}}{1 \vee |\mathcal{R}|}\right] \le \sum_{j=1}^m \mathbb{E}\left[\frac{\mathbb{1}\{Y_{n+j} \le c_{n+j},\ p_j \le q|\mathcal{R}_j^*|/m\}}{1 \vee |\mathcal{R}_j^*|}\right] \le \sum_{j=1}^m \mathbb{E}\left[\frac{\mathbb{1}\{p_j^* \le q|\mathcal{R}_j^*|/m\}}{1 \vee |\mathcal{R}_j^*|}\right]$$

- Step 2: PRDS for FDR control
  - For each j, $(p_1, \ldots, p_{j-1}, p_{j+1}, \ldots, p_m)$ is PRDS on p_j^*
  - Also, $p_j^* \sim \mathrm{Unif}[0,1]$. Thus

$$\sum_{j=1}^m \mathbb{E}\left[\frac{\mathbb{1}\{p_j^* \le q|\mathcal{R}_j^*|/m\}}{1 \vee |\mathcal{R}_j^*|}\right] \le \sum_{j=1}^m \frac{q}{m} = q.$$

Power considerations

- While FDR is controlled for any monotone score V, some is powerful
- If the thresholds are constant $c_{n+j} \equiv c$, a particularly powerful choice is the 'clipped' score

$$V(x, y) = +\infty \cdot \mathbb{1}\{y > c\} + c \cdot \mathbb{1}\{y \leq c\} - \widehat{\mu}(x)$$

- In the binary case with $c = 0$, an ideal score should be monotone in $\mathbb{P}(Y = 1 \,|\, X = x)$

Real data: Drug property prediction for HIV

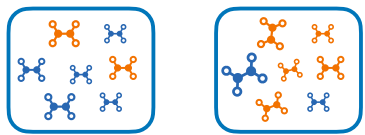
- $Y \in \{0, 1\}$: whether the drugs interact with the disease
- $n_{\text{tot}} = 41127$ in total, $6:2:2$ split, 3% in the training fold are active
- FDR level: $q \in \{0.1, 0.2, 0.5\}$
- Small neural network for illustration (can be more complicated)

	FDR			Power			$\lvert\mathcal{R}\rvert$		
Level q	0.1	0.2	0.5	0.1	0.2	0.5	0.1	0.2	0.5
`BH_clip`	0.0957	0.196	0.495	0.0788	0.174	0.410	26.5	64.2	240
`BH_res`	0.0989	0.196	0.494	0.0766	0.174	0.410	25.8	64.4	239

Table: FDR and power of the three methods averaged over $N = 100$ random splits.

So far, and next

- Reliable screening + selection procedure from any prediction model
- Works for i.i.d. or exchangeable (i.e., finite population) training and test samples
- Next: distribution shifts
  - Are my evaluated drugs comparable to the unknown drugs?



  - Similar concerns apply to job recruiting, health risk monitoring, etc

Selection by prediction under covariate shifts

- We assume that the test data $\{(X_{n+j}, Y_{n+j})\} \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$ for some unknown $\mathbb{Q}$
- The training (calibration) data are $\{(X_i, Y_i)\} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ that obeys

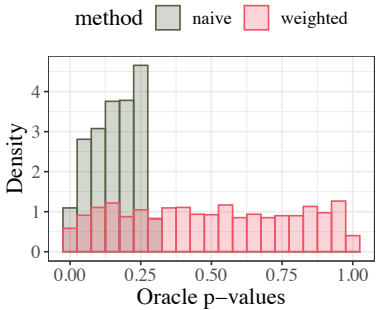
$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(x, y) = w(x)$$

  for some known weight function w: $\mathcal{X} \to \mathbb{R}^+$ [Tibshirani et al., 2019]

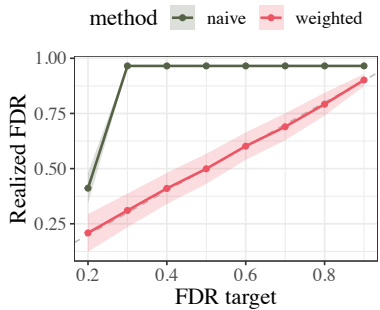
- Still want to find $Y_{n+j} > c_{n+j}$ with FDR control

Selection by prediction under covariate shifts

- If we apply the previous methods when there is actually covariate shift



P-values are no longer valid



FDR can be violated

Selection by prediction under covariate shifts

- Replace conformal p-values by weighted conformal p-values
  - Construct monotone nonconformity score $V(x, y)$, such that $y \leq y'$ implies $V(x, y) \leq V(x, y')$
  - Construct $\widehat{V}_{n+j} = V(X_{n+j}, c_{n+j})$, $j = 1, \ldots, m$ and $V_i = V(X_i, Y_i)$, $i = 1, \ldots, n$
  - Compute weighted conformal p-values (no ties)

$$p_j = \frac{\sum_{i=1}^{n} w(X_i)\, \mathbb{1}\{V_i < \widehat{V}_{n+j}\} + w(X_{n+j})}{\sum_{i=1}^{n} w(X_i) + w(X_{n+j})}$$

- We again have generalized type-I error control under covariate shift:

$$\mathbb{P}(p_j \leq \alpha, j \in \mathcal{H}_0) \leq \alpha, \quad \forall \alpha \in (0, 1)$$

- **Question: Does the previous recipe for FDR control apply?**

Weighted conformal p-values are not PRDS

Theorem (J. and Candès, in preparation, 2023+)

Suppose we construct p_j with $c_{n+j} = Y_{n+j}$. Then there exists a weight function $w(\cdot)$ and a monotone score function $V(\cdot, \cdot)$, such that the weighted conformal p-values are ***not*** *PRDS.*

- The PRDS property may fail when $V(X_i, Y_i)$ are negatively associated with $w(X_i)$
- Why?
  - Without weights, small $p_j \Leftrightarrow$ large training scores $\{V_i\} \Leftrightarrow$ small other p-values
  - With data-dependent weights, one cannot tell whether a small p_j is due to large training scores (hence other p-values are small) or small training weights (hence other p-values can be large)

Does BH + weighted p-values still work?

- We can show applying BH(q) to weighted conformal p-values controls FDR *asympototically*
  - For fixed m and $n \to \infty$, or $m, n \to \infty$ when data are i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$
  - It also empirically controls the FDR in most of our numerical experiments
- But we recently observe violated FDR in a large-scale drug discovery task (finite population)
- Theoretically, it is still an open problem

A new approach to exact FDR control

- Compute V_i, $\widehat{V}_{n+j}$, and p_j as before
- Calibrate the rejection threshold of p_j via 'auxiliary p-values'
  - For each j, for all $\ell \neq j$, define

$$p_\ell^{(j)} = \frac{\sum_{i=1}^n w(X_i)\,\mathbb{1}\{V_i < \widehat{V}_{n+\ell}\} + w(X_{n+j})\mathbb{1}\{\widehat{V}_{n+j} < \widehat{V}_{n+\ell}\}}{\sum_{i=1}^n w(X_i) + w(X_{n+j})}$$

$$\text{(as opposed to)} \quad p_\ell = \frac{\sum_{i=1}^n w(X_i)\,\mathbb{1}\{V_i < \widehat{V}_{n+\ell}\} + w(X_{n+\ell})}{\sum_{i=1}^n w(X_i) + w(X_{n+\ell})}$$

  - Let $\widehat{\mathcal{R}}_j$ be the rejection set of BH(q) applied to $\{0\} \cup \{p_\ell^{(j)}\}_{\ell \neq j}$
  - Set the rejection threshold $s_j = q|\widehat{\mathcal{R}}_j|/m$
- Obtain the final rejection set

$$\mathcal{R} := \left\{j\colon p_j \leq s_j,\ \xi_j|\widehat{\mathcal{R}}_j| \leq r^*\right\}, \quad r^* := \max\{r\colon \textstyle\sum_{j=1}^m \mathbb{1}\{p_j \leq s_j,\ \xi_j|\widehat{\mathcal{R}}_j| \leq r\} \geq r\}$$

where either $\xi_j \equiv 1$, $\xi_j \equiv \xi \sim \mathrm{Unif}[0,1]$, or $\xi_j \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0,1]$.

Exact FDR control

Theorem (J. and Candès, in preparation, 2023+)

Suppose $\{Z_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ *and* $\{Z_{n+j}\}_{j=1}^m \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$ *for* $Z_i = (X_i, Y_i)$*, and the covariate shift holds for* $w(\cdot)$*. Assume that for each* $j = 1, \ldots, m$*, data in* $\{Z_1, \ldots, Z_n, Z_{n+j}\} \cup \{\widetilde{Z}_{n+\ell}\}_{\ell \neq j}$ *are mutually independent for* $\widetilde{Z}_{n+\ell} = (X_{n+\ell}, c_{n+\ell})$*. Then all three choices of* $\{\xi_j\}$ *lead to*

$$\mathbb{E}\left[\frac{\sum_{j=1}^m \mathbb{1}\left\{j \in \mathcal{R}, j \in \mathcal{H}_0\right\}}{1 \vee |\mathcal{R}|}\right] \leq q,$$

where the expectation is taken over both calibration and test data.

Theory: step I

- Proof step 1: Extending the conditional calibration idea [Fithian and Lei, 2022], one can show that with all three choices of $\{\xi_j\}$,

$$\mathbb{E}\left[\frac{\sum_{j=1}^{m} \mathbb{1}\left\{j \in \mathcal{R}, j \in \mathcal{H}_0\right\}}{1 \vee |\mathcal{R}|}\right] \leq \sum_{j=1}^{m} \mathbb{E}\left[\frac{\mathbb{1}\left\{p_j \leq s_j, Y_{n+j} \leq c_{n+j}\right\}}{|\widehat{\mathcal{R}}_j|}\right].$$

Theory: step II, leave-one-out analysis

- Proof step 2: Leave-one-out analysis. We relate p_j and $p_\ell^{(j)}$ to

$$p_j^* = \frac{\sum_{i=1}^n w(X_i)\,\mathbb{1}\{V_i < V_{n+j}\} + w(X_{n+j})}{\sum_{i=1}^n w(X_i) + w(X_{n+j})},$$
$$p_\ell^{*,(j)} = \frac{\sum_{i=1}^n w(X_i)\,\mathbb{1}\{V_i < \widehat{V}_{n+\ell}\} + w(X_{n+j})\,\mathbb{1}\{V_{n+j} < \widehat{V}_{n+\ell}\}}{\sum_{i=1}^n w(X_i) + w(X_{n+j})}$$

  The only distinction between them is whether we used $\widehat{V}_{n+j}$ or V_{n+j}

- Define a 'proxy' rejection set $\mathcal{R}^*_{j\to 0} = \mathrm{BH}\big(q; p_1^{*,(j)}, \cdots, p_{j-1}^{*,(j)}, 0, p_{j+1}^{*,(j)}, \cdots, p_m^{*,(j)}\big)$

- A more complicated leave-one-out analysis yields

$$\sum_{j=1}^m \mathbb{E}\left[\frac{\mathbb{1}\{p_j \le s_j, Y_{n+j} \le c_{n+j}\}}{|\widehat{\mathcal{R}}_j|}\right] \le \sum_{j=1}^m \mathbb{E}\left[\frac{\mathbb{1}\{p_j^* \le q|\mathcal{R}^*_{j\to 0}|/m\}}{|\mathcal{R}^*_{j\to 0}|}\right]$$

Theory: step III, conditional independence

- Proof step 3: Due to covariate shift,

$$p_j^* \perp\!\!\!\perp \left|\mathcal{R}^*_{j\to 0}\right| \;\Big|\; \mathbf{Z}_j, \quad \forall j$$

  for the unordered set $\mathbf{Z}_j = [Z_1, \ldots, Z_n, Z_{n+j}]$, where $Z_i = (X_i, Y_i)$

- A rough argument:
  - $|\mathcal{R}^*_{j\to 0}|$ only depends on the unordered set $\mathbf{Z}_j$ and $\{\widehat{V}_{n+\ell}\}_{\ell\neq j}$
  - p_j^* and $\mathbf{Z}_j$ are independent of $\{\widehat{V}_{n+\ell}\}_{\ell\neq j}$

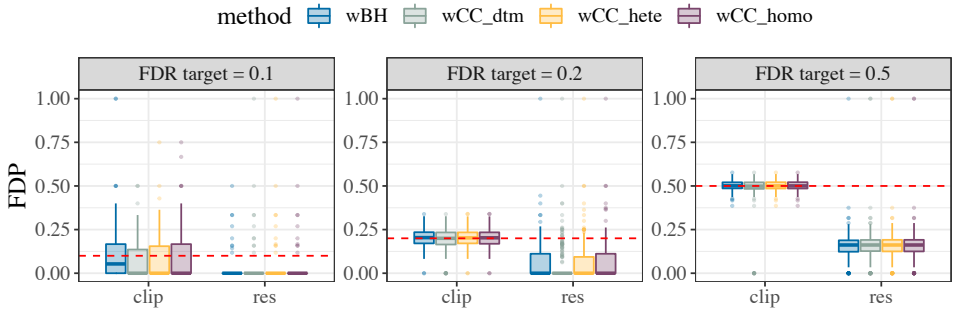
- Also, $p_j^* \mid \mathbf{Z}_j$ stochastically dominates Unif$[0,1]$. This gives

$$\sum_{j=1}^{m} \mathbb{E}\left[\frac{\mathbb{1}\left\{p_j^* \le q|\mathcal{R}^*_{j\to 0}|/m\right\}}{|\mathcal{R}^*_{j\to 0}|}\right] \le \sum_{j=1}^{m} \mathbb{E}\left[\frac{q|\mathcal{R}^*_{j\to 0}|/m}{|\mathcal{R}^*_{j\to 0}|}\right] = q.$$

- Connection to conditional calibration [Fithian and Lei, 2022]: $\mathbf{Z}_j$ serves as the 'sufficient statistic'

Real data: drug-target interaction prediction under biased sampling

- DAVIS dataset, $Y \in \mathbb{R}$ continuous binding affinities, X feature for drug-target pairs
- $n_{\text{tot}} = 30060$ drug-target pairs in total, $2:2:6$ split
- Covariate shift created by preferring high-prediction drugs in calibration data
- $c_{n+j}=$ the q_{pop}-th quantile of the outcomes of the training samples with the same binding target as sample j, where $q_{\text{pop}} \in \{0.7, 0.8, 0.9\}$. FDR level: $q \in \{0.1, 0.2, 0.5\}$



Other applications of this framework

- Detecting positive individual treatment effects
  - $\Delta = O(1) - O(0)$ is the difference between outcome under treatment $O(1)$ and under control $O(0)$
  - Our method allows for finding $O_{n+j}(1) > O_{n+j}(0)$ test units in the control group (so that $O_{n+j}(0)$ is observed, but $O_{n+j}(1)$ is not) with FDR control
  - It is equivalent to taking $Y_{n+j} = O_{n+j}(1)$ and $c_{n+j} = O_{n+j}(0)$
  - Works even though two quantities are never observed for calibration data
- Detecting outliers/concept drifts under covariate shift

Summary

- We argue FDR as a sensible error criterion in prediction-assisted screening and discovery
- Methods that turns *any* prediction model into a reliable selection procedure
  - P-value and multiple testing for random hypotheses
- Extend to settings with covariate shifts
  - Some more complicated methodology & theory

(first part) arXiv: 2210.01408