Conformal Prediction with Conditional Guarantees

Emmanuel Candès, *Stanford University*

Elisabeth Gassiat - a path in modern statistics, Orsay, May 31, 2023

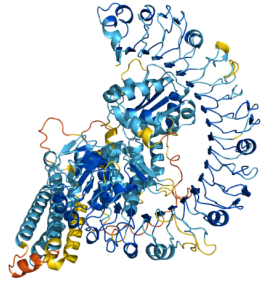
Collaborators

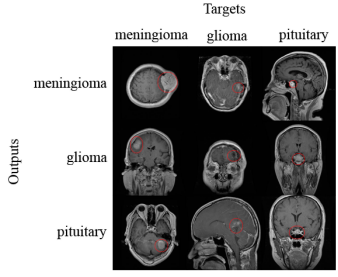
John Cherian

Isaac Gibbs

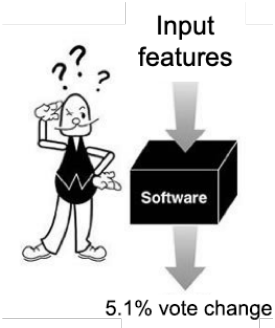
High-stakes ML

- Black boxes (random forests, NNs) can deliver excellent point predictions
- But are they trustworthy?





Data ethics 101: convey uncertainty



How certain am I of my prediction?

How does my level of uncertainty affect my decision?

Can my model be safely deployed?

Need to be explicit and clear about this to help users

One solution: prediction sets

Data $\{(X_i, Y_i)\}_{i=1}^{n+1} \overset{i.i.d.}{\sim} P \rightsquigarrow$ **want prediction interval $\hat{C}$ that covers unobserved Y_{n+1}**

Should be

- distribution-free
- capable of utilizing a complex black-box model
- conditionally valid

$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}) \mid X_{n+1}) = 1 - \alpha$$

One solution: prediction sets

Data $\{(X_i, Y_i)\}_{i=1}^{n+1} \overset{i.i.d.}{\sim} P \rightsquigarrow$ **want prediction interval $\hat{C}$ that covers unobserved Y_{n+1}**

Should be

- distribution-free
- capable of utilizing a complex black-box model
- conditionally valid

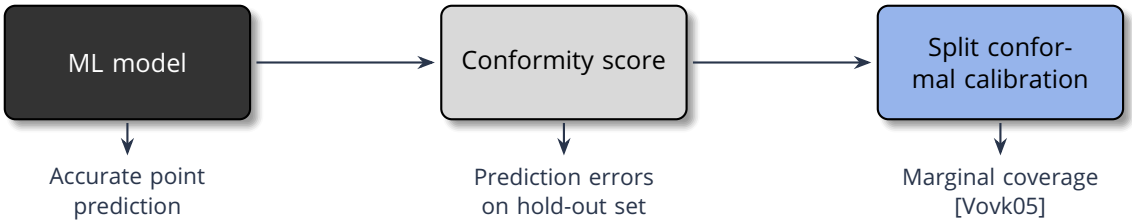
$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}) \mid X_{n+1}) = 1 - \alpha$$

Theorem (Vovk 2012)

If $\hat{C}$ is distribution-free and conditionally valid

$$\mathbb{E}[|\hat{C}(X_{n+1})|] = \infty$$

Split conformal prediction



Marginal coverage - Papadopoulos et al. (2002)

Assume that the conformity scores are almost surely distinct. Then

$$1 - \alpha \leq \mathbb{P}(Y_{n+1} \in \hat{C}_{\text{split}}(X_{n+1})) \leq 1 - \alpha + \frac{1}{n+1}$$

Conditional coverage: impossible?

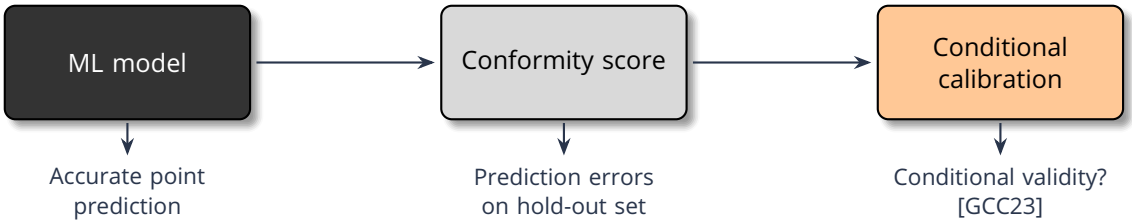
- True conditional coverage known to be impossible
- Conditional coverage over all “large sets” requires inflation

Coverage inflation - Barber–C–Ramdas–Tibshirani (2021)

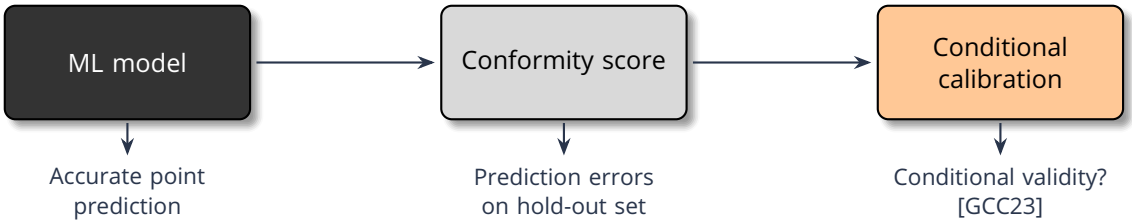
If $\hat{C}(\cdot)$ satisfies $\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}) \mid X \in A) \geq 1 - \alpha$ for all A such that $\mathbb{P}_X(A) \geq \delta$

$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha\delta$$

Conditional calibration



Conditional calibration



What is conditional validity?

Coverage spectrum

Conditional coverage

$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}) \mid X_{n+1}) = 1 - \alpha$$

$$\Longleftrightarrow$$

$$\mathbb{E}[(1\{Y_{n+1} \notin \hat{C}(X_{n+1})\} - \alpha) \cdot f(X)] = 0 \qquad \text{for all measurable } f$$

Marginal coverage

$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})) = 1 - \alpha$$

$$\Longleftrightarrow$$

$$\mathbb{E}[(1\{Y_{n+1} \notin \hat{C}(X_{n+1})\} - \alpha) \cdot \theta] = 0 \qquad \text{for all constant functions } \theta$$

Coverage spectrum

Conditional-*ish* coverage

$$\mathbb{E}[(1\{Y_{n+1} \notin \hat{C}(X_{n+1})\} - \alpha) \cdot f(X)] = 0 \qquad \text{for all } f \in \mathcal{F}$$

Coverage spectrum

Conditional-*ish* coverage

$$\mathbb{E}[(1\{Y_{n+1} \notin \hat{C}(X_{n+1})\} - \alpha) \cdot f(X)] = 0 \qquad \text{for all } f \in \mathcal{F}$$

Example

$$\mathcal{F} = \{\beta^\top \{1\{X \in G\}\}_{G \in \mathcal{G}} \mid \beta \in \mathbb{R}^{|\mathcal{G}|}\}$$

$$\Longrightarrow$$

$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}) \mid X_{n+1} \in G) = 1 - \alpha \quad \text{for all } G \in \mathcal{G}$$

Split conformal prediction

Inputs

- Prediction rule $f(\cdot)$ (think neural nets, random forests, XGBoost, etc.)
- Calibration set $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$
- Conformity score $S(\cdot, \cdot)$, e.g. $S(X, Y) := |Y - f(X)|$

Split conformal prediction

Inputs

- Prediction rule $f(\cdot)$ (think neural nets, random forests, XGBoost, etc.)
- Calibration set $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$
- Conformity score $S(\cdot, \cdot)$, e.g. $S(X, Y) := |Y - f(X)|$

Given scores

$$S_{(1)}, \ldots, S_{(n)}$$

where will S_{n+1} land?

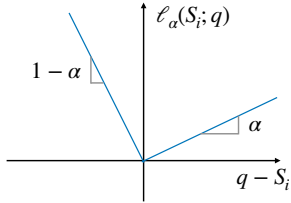
$$\hat{C}_{\text{split}}(X_{n+1}) := \left\{ y : S(X_{n+1}, y) \leq \text{Quantile}\left(\frac{\lceil (n+1)(1-\alpha) \rceil}{n}; \{S_i\}_{i=1}^n \right) \right\}$$

Split CP as quantile regression

Recall that

$$q^* = \text{argmin}_{q \in \mathbb{R}} \sum_{i=1}^{n} \underbrace{(1-\alpha) \cdot (S_i - q)_+ + \alpha \cdot (q - S_i)_+}_{\ell_\alpha(S_i; q)}$$

is the $(1-\alpha)$-quantile of $\{S_i\}_{i=1}^n$

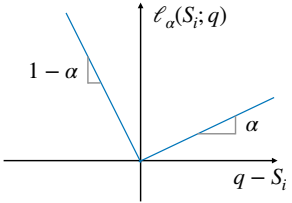


Split CP as quantile regression

Recall that

$$q^* = \text{argmin}_{q \in \mathbb{R}} \sum_{i=1}^{n} \underbrace{(1-\alpha) \cdot (S_i - q)_+ + \alpha \cdot (q - S_i)_+}_{\ell_\alpha(S_i; q)}$$

is the $(1-\alpha)$-quantile of $\{S_i\}_{i=1}^n$



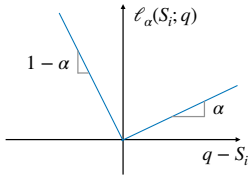
Alternative construction

$$\hat{C}_{\text{split}}(X_{n+1}) := \left\{ y : S(X_{n+1}, y) \leq \text{argmin}_q \left[\sum_{i=1}^{n} \ell_\alpha(S_i; q) + \ell_\alpha(S_{n+1}; q) \right] \right\}$$

Proof

KKT condition:

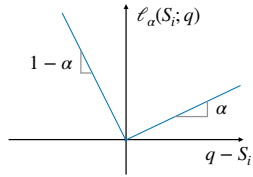
$$0 \in \left\{ \sum_{S_i \neq q^*} (\alpha - 1\{S_i > q^*\}) + \sum_{S_i = q^*} \lambda_i \mid \lambda_i \in [\alpha - 1, \alpha] \right\}$$



Proof

KKT condition:

$$0 \in \left\{ \sum_{S_i \neq q^*} (\alpha - 1\{S_i > q^*\}) + \sum_{S_i = q^*} \lambda_i \mid \lambda_i \in [\alpha - 1, \alpha] \right\}$$



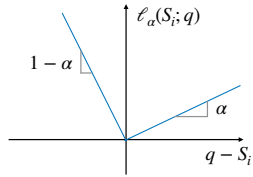
Rearranging

$$\frac{1}{n+1} \sum_{i=1}^{n+1} (\alpha - 1\{S_i > q^*\}) = \frac{1}{n+1} \sum_{S_i = q^*} (\alpha - \lambda_i^*) \geq 0$$

Proof

KKT condition:

$$0 \in \left\{ \sum_{S_i \neq q^*} (\alpha - 1\{S_i > q^*\}) + \sum_{S_i = q^*} \lambda_i \mid \lambda_i \in [\alpha - 1, \alpha] \right\}$$



Rearranging

$$\frac{1}{n+1} \sum_{i=1}^{n+1} (\alpha - 1\{S_i > q^*\}) = \frac{1}{n+1} \sum_{S_i = q^*} (\alpha - \lambda_i^*) \geq 0$$

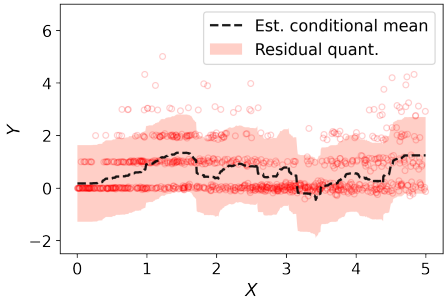
Taking expectations and using the exchangeability of the scores

$$\mathbb{E}\left[\frac{1}{n+1} \sum_{i=1}^{n+1} (\alpha - 1\{S_i > q^*\}) \right] = \alpha - \mathbb{P}(S_{n+1} > q^*) \geq 0$$

Marginal coverage is not conditional coverage

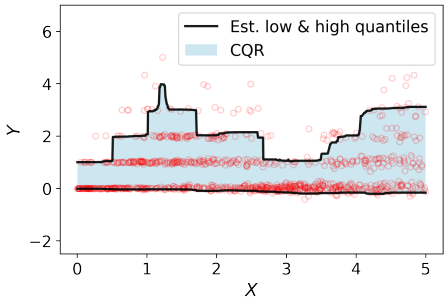
Romano, Sesia, C. 2019

Residual scores + random forests



Avg. Coverage 91.4%
Avg. Length 2.91
Fixed length

Conformalized quantile regression (CQR)



Avg. Coverage 91.0%
Avg. Length 2.18
Adaptive length

Finite dim. coverage

$$\mathcal{F} := \{\beta^\top \Phi(X) \mid \beta \in \mathbb{R}^d\}$$

Set

$$f_S := \operatorname{argmin}_{f \in \mathcal{F}} \left[\sum_{i=1}^{n} \ell_\alpha(S_i; f(X_i)) + \ell_\alpha(S; f(X_{n+1})) \right]$$

$$\implies \quad \hat{C}(X_{n+1}) := \{y : S(X_{n+1}, y) \leq \underbrace{f_{S(X_{n+1},y)}(X_{n+1})}_{\text{adaptive threshold}}\}$$

Finite dim. coverage

$$\mathcal{F} := \{\beta^\top \Phi(X) \mid \beta \in \mathbb{R}^d\}$$

Set

$$f_S := \text{argmin}_{f \in \mathcal{F}} \left[\sum_{i=1}^{n} \ell_\alpha(S_i; f(X_i)) + \ell_\alpha(S; f(X_{n+1})) \right]$$

$$\implies \quad \hat{C}(X_{n+1}) := \{y : S(X_{n+1}, y) \leq \underbrace{f_{S(X_{n+1},y)}(X_{n+1})}_{\text{adaptive threshold}}\}$$

Finite-dimensional coverage - Gibbs–Cherian–C (2023)

Assume that $S \mid X$ is continuous. Then for all $f \in \mathcal{F}$

$$\left| \mathbb{E}\left[(1\{Y_{n+1} \notin \hat{C}(X_{n+1})\} - \alpha) \cdot f(X_{n+1}) \right] \right| \leq \frac{d}{n+1} \mathbb{E}\left[\max_{1 \leq i \leq n+1} |f(X_i)| \right]$$

Finite dim. coverage

$$\mathcal{F} := \{\beta^\top \Phi(X) \mid \beta \in \mathbb{R}^d\}$$

Set

$$f_S := \operatorname{argmin}_{f \in \mathcal{F}} \left[\sum_{i=1}^{n} \ell_\alpha(S_i; f(X_i)) + \ell_\alpha(S; f(X_{n+1})) \right]$$

$$\implies \quad \hat{C}(X_{n+1}) := \{y : S(X_{n+1}, y) \leq \underbrace{f_{S(X_{n+1},y)}(X_{n+1})}_{\text{adaptive threshold}}\}$$

Group-conditional coverage - Gibbs–Cherian–C (2023)

Let $\mathcal{F} = \{\beta^\top \{1\{X \in G\}\}_{G \in \mathcal{G}} \mid \beta \in \mathbb{R}^{|\mathcal{G}|}\}$. Assume $S \mid X$ is continuous. Then

$$1 - \alpha \leq \mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}) \mid X_{n+1} \in G) \leq 1 - \alpha + \frac{|\mathcal{G}|}{n+1} \qquad \text{for all } G \in \mathcal{G}$$

Proof

By the KKT condition 0 must lie in the set

$$\left\{ \sum_{S_i \neq f_{S_{n+1}}(X_i)} 1\{X_i \in G\} \cdot (\alpha - 1\{S_i > f_{S_{n+1}}(X_i)\}) + \sum_{S_i = f_{S_{n+1}}(X_i)} 1\{X_i \in G\}\lambda_i \mid \lambda_i \in [\alpha - 1, \alpha] \right\}$$

Proof

By the KKT condition 0 must lie in the set

$$\left\{ \sum_{S_i \neq f_{S_{n+1}}(X_i)} 1\{X_i \in G\} \cdot (\alpha - 1\{S_i > f_{S_{n+1}}(X_i)\}) + \sum_{S_i = f_{S_{n+1}}(X_i)} 1\{X_i \in G\}\lambda_i \mid \lambda_i \in [\alpha - 1, \alpha] \right\}$$

Rearranging

$$\frac{1}{n+1} \sum_{i=1}^{n+1} 1\{X_i \in G\} \cdot (\alpha - 1\{S_i > f_{S_{n+1}}(X_i)\}) = \frac{1}{n+1} \sum_{S_i = f_{S_{n+1}}(X_i)} (\alpha - \lambda_i^*) 1\{X_i \in G\} \geq 0$$

Proof

By the KKT condition 0 must lie in the set

$$\left\{ \sum_{S_i \neq f_{S_{n+1}}(X_i)} 1\{X_i \in G\} \cdot (\alpha - 1\{S_i > f_{S_{n+1}}(X_i)\}) + \sum_{S_i = f_{S_{n+1}}(X_i)} 1\{X_i \in G\}\lambda_i \mid \lambda_i \in [\alpha - 1, \alpha] \right\}$$

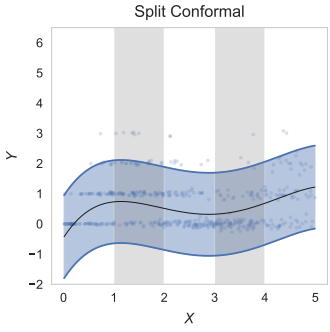
Rearranging

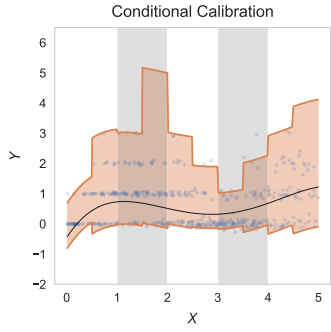
$$\frac{1}{n+1} \sum_{i=1}^{n+1} 1\{X_i \in G\} \cdot (\alpha - 1\{S_i > f_{S_{n+1}}(X_i)\}) = \frac{1}{n+1} \sum_{S_i = f_{S_{n+1}}(X_i)} (\alpha - \lambda_i^*) 1\{X_i \in G\} \geq 0$$

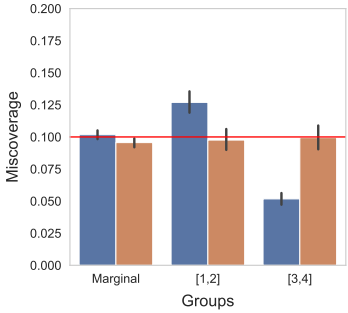
Taking expectations and using the exchangeability of the scores gives

$$\mathbb{P}\left(S_{n+1} > f_{S_{n+1}}(X_{n+1}) \mid X_{n+1} \in G\right) \leq \alpha$$

Group coverage: results







$$\mathcal{G} = \{[a, b] \mid a, b \in \{0, 0.5, \ldots, 5\}\}$$

Coverage under covariate shift

Consider $(X_{n+1}, Y_{n+1}) \sim P_f := Q_X \times P_{Y|X}$

$$dQ_X \propto f(X) \cdot dP_X$$

for some non-negative "tilt" f

Definition: Coverage under covariate shift

For $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$ and (X_{n+1}, Y_{n+1}) as above

$$\mathbb{P}_f(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha$$

Coverage under covariate shift

Consider $(X_{n+1}, Y_{n+1}) \sim P_f := Q_X \times P_{Y|X}$

$$dQ_X \propto f(X) \cdot dP_X$$

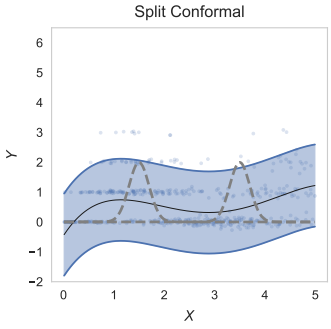
for some non-negative "tilt" f

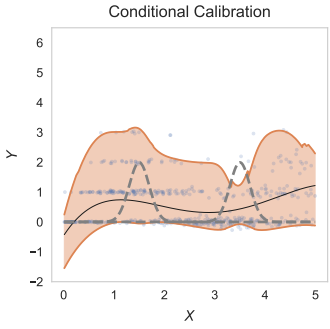
Coverage under covariate shifts - Gibbs–Cherian–C (2023)

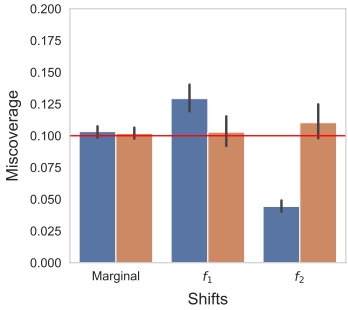
Let $\mathcal{F} = \{\beta^\top \Phi(X) : \beta \in \mathbb{R}^d\}$. Assume $S \mid X$ is continuous. Then for all non-negative functions $f \in \mathcal{F}$

$$1 - \alpha \leq \mathbb{P}_f(Y_{n+1} \in \hat{C}(X_{n+1})) \leq 1 - \alpha + \frac{d}{n+1}$$

Covariate shift: results







Infinite dim. coverage

$$f_S := \text{argmin}_{f \in \mathcal{F}} \left[\sum_{i=1}^{n} \ell_\alpha(S_i; f(X_i)) + \ell_\alpha(S; f(X_{n+1})) \right] + \lambda \cdot \|f\|_{\mathcal{F}}^2$$

$$\implies \hat{C}(X_{n+1}) := \{y : S(X_{n+1}, y) \leq f_{S(X_{n+1},y)}(X_{n+1})\}$$

Shift-agnostic coverage - Gibbs–Cherian–C (2023)

Let $\mathcal{F}$ be an RKHS. Assume $S \mid X$ is continuous. Then for all non-negative functions $f \in \mathcal{F}$

$$\mathbb{P}_f(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha - \frac{2 \cdot \lambda \cdot \mathbb{E}[\langle f_{S_{n+1}}, f \rangle]}{n+1}$$

Infinite dim. coverage

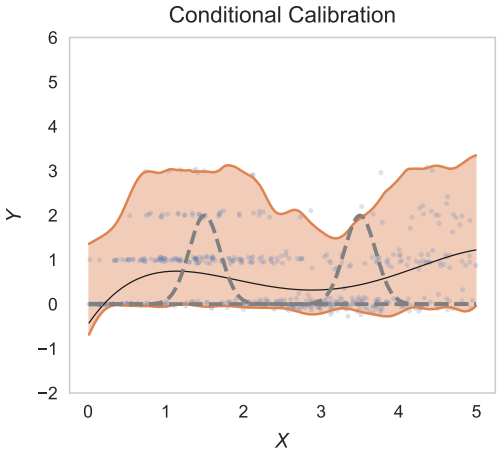
Shift-agnostic coverage (matching upper bound) - Gibbs–Cherian–C (2023)

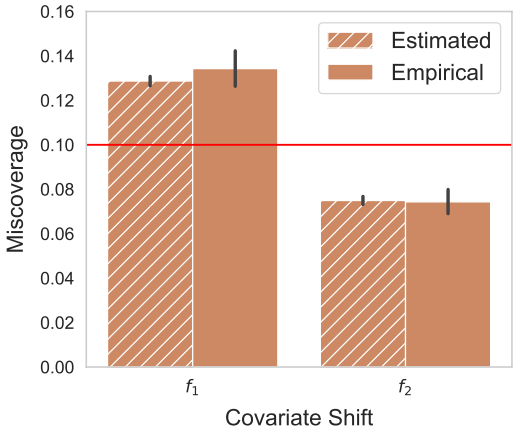
Let $\mathcal{F}$ be an RKHS. Assume $S \mid X$ is continuous. Then for all non-negative functions $f \in \mathcal{F}$

$$\mathbb{P}_f(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha - \frac{2 \cdot \lambda \cdot \mathbb{E}[\langle f_{S_{n+1}}, f \rangle]}{n+1}$$

$$\mathbb{P}_f(Y_{n+1} \in \hat{C}(X_{n+1})) \leq 1 - \alpha - \frac{2 \cdot \lambda \cdot \mathbb{E}[\langle f_{S_{n+1}}, f \rangle]}{n+1} + \underbrace{\mathbb{P}_f\left(S_{n+1} = f_{S_{n+1}}(X_{n+1})\right)}_{O\left(\frac{d \log(n)}{\lambda n}\right)}$$

Covariate shift: results





Computation

$$f_S := \text{argmin}_{f \in \mathcal{F}} \left[\sum_{i=1}^{n} \ell_\alpha(S_i; f(X_i)) + \ell_\alpha(S; f(X_{n+1})) \right] + \lambda \cdot \|f\|_{\mathcal{F}}^2$$

$$\hat{C}(X_{n+1}) := \{y : S(X_{n+1}, y) \leq f_{S(X_{n+1},y)}(X_{n+1})\}$$

How do we compute $\hat{C}(X_{n+1})$?

- Inefficiency of full conformal?
- Approximate methods?

Computation: convex duality

Primal

$$\begin{aligned} \min_{p,q,\beta} \quad & \sum_{i=1}^{n+1} (1-\alpha) p_i + \alpha q_i \\ \text{subject to} \quad & S_i - \Phi(X_i)^\top \beta - p_i + q_i = 0 \\ & S - \Phi(X_{n+1})^\top \beta - p_{n+1} + q_{n+1} = 0 \\ & p_i, q_i \geq 0 \end{aligned}$$

Dual

$$\begin{aligned} \max_{\eta} \quad & \sum_{i=1}^{n} \eta_i S_i + \eta_{n+1} S \\ \text{subject to} \quad & -\alpha \leq \eta_i \leq 1-\alpha \\ & \Phi_j(X)^\top \eta = 0 \end{aligned}$$

Computation: convex duality

Primal

$$\begin{aligned} \min_{p,q,\beta} \quad & \sum_{i=1}^{n+1} (1-\alpha) p_i + \alpha q_i \\ \text{subject to} \quad & S_i - \Phi(X_i)^\top \beta - p_i + q_i = 0 \\ & S - \Phi(X_{n+1})^\top \beta - p_{n+1} + q_{n+1} = 0 \\ & p_i, q_i \geq 0 \end{aligned}$$

Dual

$$\begin{aligned} \max_{\eta} \quad & \sum_{i=1}^{n} \eta_i S_i + \eta_{n+1} S \\ \text{subject to} \quad & -\alpha \leq \eta_i \leq 1-\alpha \\ & \Phi_j(X)^\top \eta = 0 \end{aligned}$$

Monotonicity of membership - Gibbs–Cherian–C (2023)

- Checking $S \leq f_S(X_{n+1})$ is (approx.) equivalent to checking $\eta_{n+1}^S < 1-\alpha$
- $S \mapsto \eta_{n+1}^S$ is non-decreasing

Computation: convex duality

Primal

$$\begin{aligned} \min_{p,q,\beta} \quad & \sum_{i=1}^{n+1} (1-\alpha) p_i + \alpha q_i \\ \text{subject to} \quad & S_i - \Phi(X_i)^\top \beta - p_i + q_i = 0 \\ & S - \Phi(X_{n+1})^\top \beta - p_{n+1} + q_{n+1} = 0 \\ & p_i, q_i \geq 0 \end{aligned}$$

Dual

$$\begin{aligned} \max_{\eta} \quad & \sum_{i=1}^{n} \eta_i S_i + \eta_{n+1} S \\ \text{subject to} \quad & -\alpha \leq \eta_i \leq 1-\alpha \\ & \Phi_j(X)^\top \eta = 0 \end{aligned}$$

Algorithm - Gibbs–Cherian–C (2023)

Binary search for largest S such that $\eta_{n+1}^S < 1 - \alpha$

Communities and Crime

- Predict crime rate from various demographics (age, income, race, unemployment, etc.)
- Four racial features: %Black, %White, %Asian, %Hispanic

Our approach

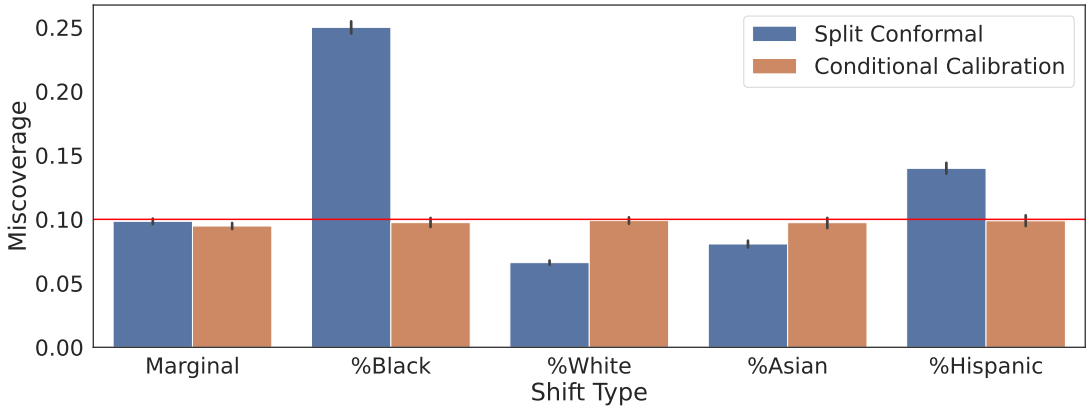
$$\mathcal{F} = \{f_k + \beta^\top \Phi(X) \mid f_k \in \mathcal{F}_K, \beta \in \mathbb{R}^5\}$$

$$K(x_i, x_j) = \exp\left(-4\|x_i - x_j\|_2^2\right), \ \Phi(X) = (\%\text{Black}, \%\text{White}, \%\text{Asian}, \%\text{Hispanic})$$

Communities and Crime: results

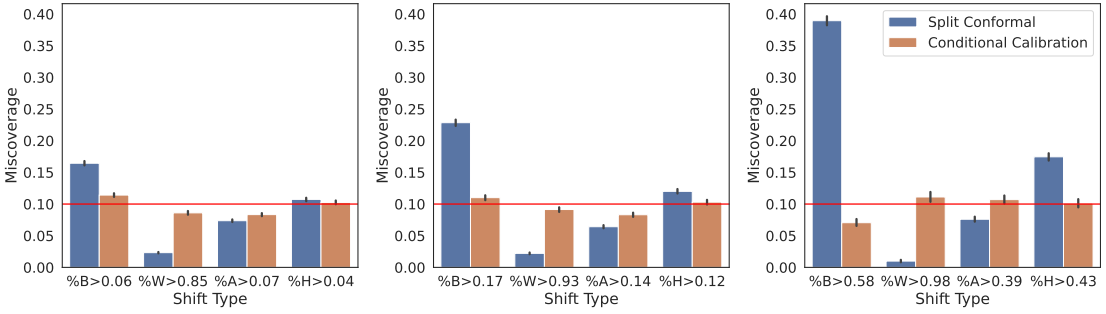
Consider linear shifts

$$\{x \mapsto 1,\ x \mapsto \%\text{Black},\ x \mapsto \%\text{White},\ x \mapsto \%\text{Asian},\ x \mapsto \%\text{Hispanic}\}$$



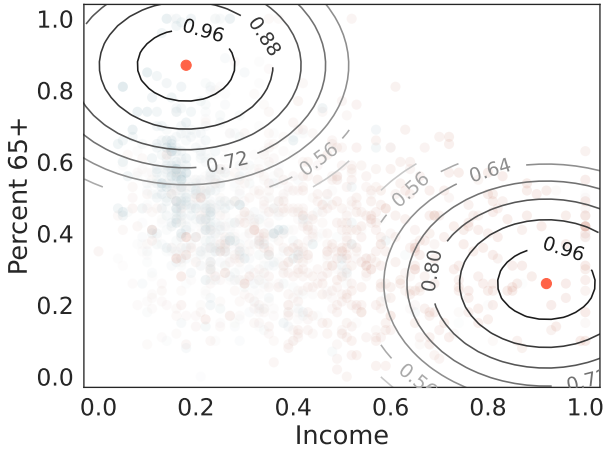
Communities and Crime: results

Consider groups defined by %-thresholds, e.g., all communities with %Black > 0.06

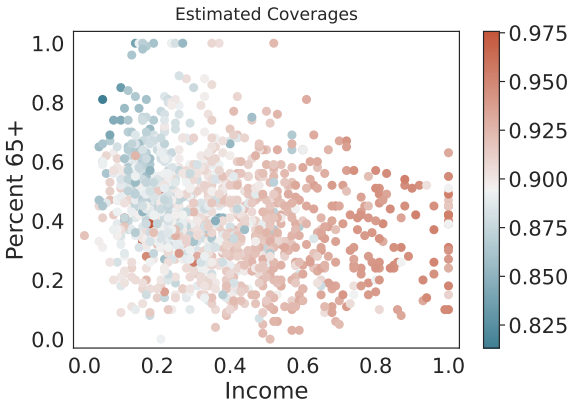


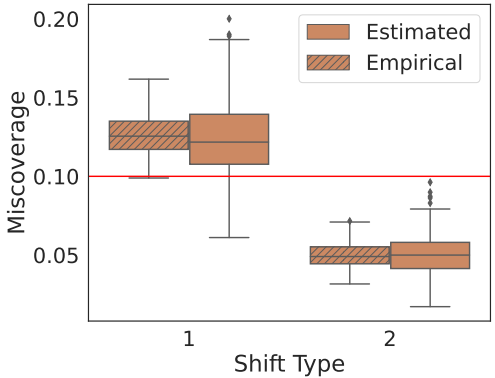
%-thresholds defined as the p-th percentile of group membership over all communities ($p \in \{50, 70, 90\}$)

Communities and Crime: results



Communities and Crime: results





Biomolecular data

We are given images of cells obtained using fluorescent microscopy and we must predict which one of the 1339 genetic treatments the cells received. 51 different experiments run across four different cell types

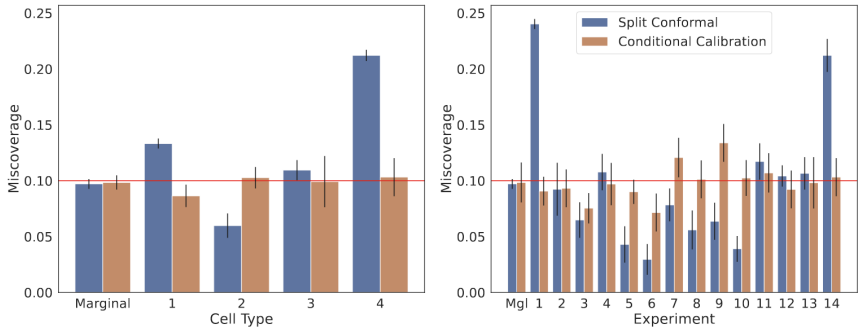


Figure 5.4: Empirical conditional miscoverage of our method (orange) and split conformal (blue) across cell types and experiments. Red lines indicate the target level of $\alpha = 0.1$ and black error bars show 95% binomial confidence intervals for the calibration-conditional miscoverage $\mathbb{P}(Y_{n+1} \notin \hat{C}(X_{n+1}) | D_{\text{train}}, D_{\text{cal}})$, where D_{train} and D_{cal} denote the training dataset used to learn the feature representation and the calibration dataset used to implement our method, respectively.

Summary

I. Gibbs, J. Cherian, E. Candès, Conformal Prediction With Conditional Guarantees (2023)

- Second layer of protection against lack of calibration
- Flexible: can be applied to any conformal predictor
- Provides tight conditional guarantees
- Miscoverage can be estimated accurately
- Code: `https://github.com/jjcherian/conditional-conformal`

Happy Birthday Elisabeth!