Conformal Prediction Under Covariate Shift

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candes, Aaditya Ramdas

Carnegie Mellon University, University of Chicago, Stanford University, Carnegie Mellon University

Carnegie Mellon University

THE UNIVERSITY OF CHICAGO

Stanford University

Conformal prediction

Setup. Given i.i.d. samples $(X_i, Y_i) \sim P$, $i = 1, \dots, n$, where P is a distribution on $\mathbb{R}^d \times \mathbb{R}$. Goal: compute band $\widehat{C}_n : \mathbb{R}^d \to$ such that for a new i.i.d. point (X_{n+1}, Y_{n+1}),

$$\mathbb{P}\Big\{ Y_{n+1} \in \widehat{C}_n(X_{n+1}) \Big\} \geq 1 - \alpha \qquad (1)$$

for given miscoverage rate $\alpha \in (0, 1)$. No assumptions on P!

Quantile lemma. If $V_1, \dots, V_{n+1}$ are exchangeable, then for any $\beta \in (0, 1)$,

$$\mathbb{P}\Big\{ V_{n+1} \leq \mathrm{Quant}\big(\beta; V_{1:n} \cup \{\infty\}\big) \Big\} \geq \beta$$

Proof. Let $q = \mathrm{Quant}(\beta; F)$, where F has support points a_i, $i = 1, 2, \dots$. If we reassign points $a_i > q$ to any values strictly larger than q, then the level β quantile is unchanged. Thus

$$V_{n+1} > \mathrm{Quant}\big(\beta; V_{1:n} \cup \{\infty\}\big) \iff V_{n+1} > \mathrm{Quant}\big(\beta; V_{1:(n+1)}\big)$$

Equivalently with $\leq$. But $V_{n+1} \leq \mathrm{Quant}(\beta; V_{1:(n+1)}) \iff$ V_{n+1} is among $\lceil \beta(n+1) \rceil$ smallest of $V_1, \dots, V_{n+1}$. □

Conformal prediction. Due to Vovk et al. (2005). Denote $Z_i = (X_i, Y_i)$, $i = 1, \dots, n$. Choose a score function $\mathcal{S}$, e.g.,

$$\mathcal{S}\big((x, y), Z\big) = |y - \widehat{\mu}(x)|,$$

where $\widehat{\mu} : \mathbb{R}^d \to \mathbb{R}$ is fitted by running algorithm $\mathcal{A}$ on Z. For $x \in \mathbb{R}^d$, define $\widehat{C}_n(x)$ by computing for each $y \in \mathbb{R}$:

$$V_i^{(x,y)} = \mathcal{S}\big(Z_i, Z_{1:n} \cup \{(x, y)\}\big), \; i = 1, \dots, n$$
$$V_{n+1}^{(x,y)} = \mathcal{S}\big((x, y), Z_{1:n} \cup \{(x, y)\}\big)$$

We include y in $\widehat{C}_n(x)$ provided

$$V_{n+1}^{(x,y)} \leq \mathrm{Quant}\big(1 - \alpha; V_{1:n}^{(x,y)} \cup \{\infty\}\big) \iff$$
$V_{n+1}^{(x,y)}$ is among $\lceil (1-\alpha)(n+1) \rceil$ smallest of $V_1^{(x,y)}, \dots, V_{n+1}^{(x,y)}$

This construction gives distribution-free coverage as in (1)

Covariate shift

Setup. Suppose training and test data are not i.i.d., instead

$$(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P = P_X \times P_{Y|X}, \; i = 1, \dots, n,$$
$$(X_{n+1}, Y_{n+1}) \sim \widetilde{P} = \widetilde{P}_X \times P_{Y|X}, \text{ independently} \qquad (2)$$

Called covariate shift model. Conformal prediction no longer works (lack of exchangeability)

Suppose we knew likelihood ratio $w = \mathrm{d}\widetilde{P}_X / \mathrm{d}P_X$. By sampling from training set with probabilities $\propto w$, this would "look like" draws from test distribution, so we could use conformal

Key idea. can do this without sampling, with exact coverage. Compute scores as before, now let y in $\widehat{C}_n(x)$ provided

$$V_{n+1}^{(x,y)} \leq \mathrm{Quant}\bigg(1 - \alpha; \sum_{i=1}^n p_i^w(x) \delta_{V_i^{(x,y)}} + p_{n+1}^w(x) \delta_\infty\bigg)$$

Here $p_i^w(x) \propto w(X_i)$, $i = 1, \dots, n$, and $p_{n+1}^w(x) \propto w(x)$. This construction recovers (1) for the covariate shift model (2)

Estimation of w. Given test covariates X_i, $i = n+1, \dots, m$, we can run any classifier (that outputs estimated probabilities) to (X_i, C_i), $i = 1, \dots, n+m$, where $C_i = 1\{i \leq n\}$. As

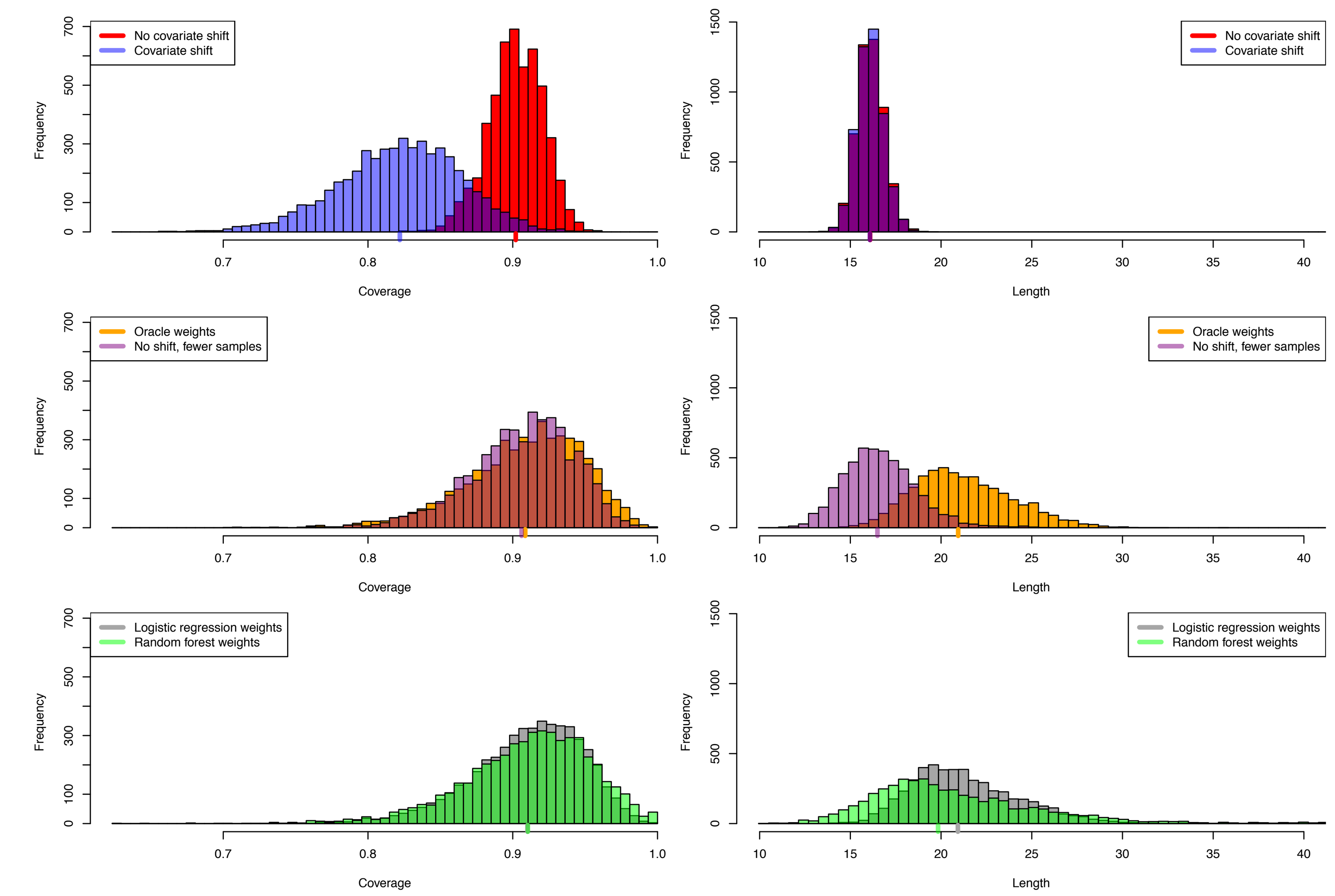
$$\frac{\mathbb{P}(C = 1 | X = x)}{\mathbb{P}(C = 0 | X = x)} = \frac{\mathbb{P}(C = 1)}{\mathbb{P}(C = 0)} \frac{\mathrm{d}\widetilde{P}_X}{\mathrm{d}P_X}(x),$$

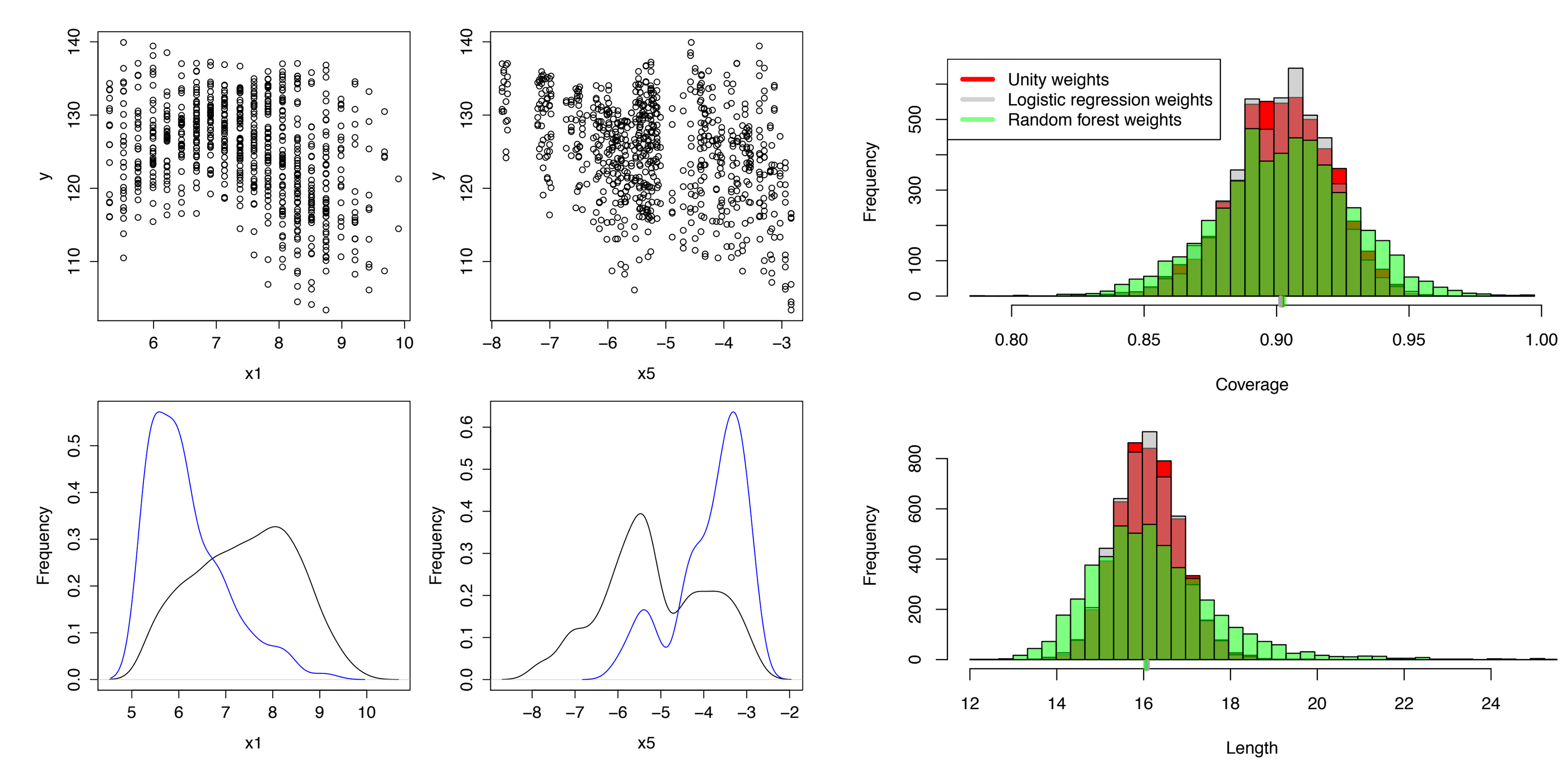
we can take $w(x) = \mathbb{P}(C = 1 | X = x) / \mathbb{P}(C = 0 | X = x)$

Simulated example

Airfoil data: $n = 1503$, $d = 5$. For $T = 5000$ trials, randomly partition the data into halves $D_{\text{train}}, D_{\text{test}}$, construct D_{shift}, by sampling from D_{test} proportionally to

$$w(x) = \exp(x^T \beta), \quad \text{where} \quad \beta = (-1, 0, 0, 0, 1)$$





Generalization

Weighted exchangeability. We say $V_1, \dots, V_n$ are weighted exchangeable if their joint density f factorizes as

$$f(v_1, \dots, v_n) = \prod_{i=1}^n w_i(v_i) \cdot g(v_1, \dots, v_n),$$

where g does not depend on the ordering of its inputs

Weighted conformal. Conformal prediction can be extended to (X_i, Y_i), $i = 1, \dots, n+1$ weighted exchangable. Special cases: i.i.d., exchangeable, covariate shift, covariate clusters