Advances and Challenges in Conformal Prediction

Ryan Tibshirani
Department of Statistics
Berkeley
UNIVERSITY OF CALIFORNIA

My great collaborators: Rina Foygel Barber, Emmanuel Candes, Max G'Sell, Jing Lei, Aaditya Ramdas, Alessandro Rinaldo, Larry Wasserman

`http://www.stat.cmu.edu/~ryantibs/talks/conformal-2022.pdf`

A lofty goal?

Given i.i.d. pairs $(X_i, Y_i) \sim P$, $i = 1, \ldots, n$, from a distribution P on $\mathcal{X} \times \mathbb{R}$ (e.g., $\mathcal{X} = \mathbb{R}^d$)

Goal. Build prediction band $\hat{C}_n : \mathcal{X} \to \mathcal{P}(\mathbb{R})$, such that for a new i.i.d. pair (X_{n+1}, Y_{n+1}):

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

(where the probability is over all $n + 1$ pairs)

Can we do so without any assumptions on the distribution P, and hope for something nontrivial?

Starting simple

Suppose we have only i.i.d. $Y_i \sim P$, $i = 1, \ldots, n$ (no features). By taking the sample $(1-\alpha)$-quantile

$$\hat{q}_n = \text{Quantile}(1-\alpha; \{Y_i\}_{i=1}^n)$$

we have the approximate result

$$\mathbb{P}(Y_{n+1} \leq \hat{q}_n) \approx 1-\alpha$$

and this becomes exact as $n \to \infty$, under standard conditions

For some modified estimate $\hat{q}_n$, can we get finite-sample coverage

$$\mathbb{P}(Y_{n+1} \leq \hat{q}_n) \geq 1-\alpha?$$

Small tweak

With just a small tweak we can achieve this! Defining

$$\hat{q}_n = \text{Quantile}\left(\frac{\lceil (1-\alpha)(n+1) \rceil}{n}; \{Y_i\}_{i=1}^n\right)$$

we indeed get $\mathbb{P}(Y_{n+1} \leq \hat{q}_n) \geq 1 - \alpha$

Why? Note that $Y_{n+1} \leq \hat{q}_n$ is equivalent to

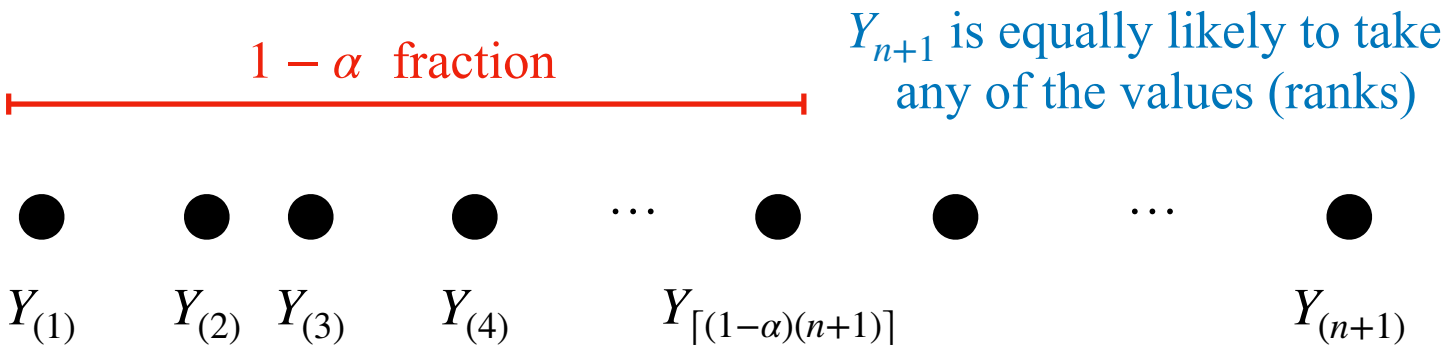
$$Y_{n+1} \leq \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } Y_1, \ldots, Y_n$$

which is in turn equivalent to

$$Y_{n+1} \text{ is among the } \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } Y_1, \ldots, Y_{n+1}$$

and by exchangeability this occurs with probability at least $1 - \alpha$

Simple illustration



Upper bound

As we learned from the picture, if there are almost surely no ties, then

$$Y_{n+1} \text{ is among the } \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } Y_1, \ldots, Y_{n+1}$$

occurs with probability exactly

$$\frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \leq 1 - \alpha + \frac{1}{n+1}$$

So under exchangeability and a continuous distribution for each Y_i (or suitable randomization), we get both lower and upper bounds:

$$\mathbb{P}(Y_{n+1} \leq \hat{q}_n) \in \left[1-\alpha,\, 1-\alpha+\tfrac{1}{n+1}\right]$$

Naive attempt

Back to our original problem: given (X_i, Y_i), $i = 1, \ldots, n$, suppose we compute $\hat{f}_n$, where $\hat{f}_n(x)$ estimates $\mathbb{E}[Y|X = x]$. Then let

$$R_i = |Y_i - \hat{f}_n(X_i)|, \; i = 1, \ldots, n$$

let $\hat{q}_n = \lceil (1-\alpha)(n+1) \rceil$ smallest of $R_1, \ldots, R_n$, and define

$$\hat{C}_n(x) = \left[\hat{f}_n(x) - \hat{q}_n, \; \hat{f}_n(x) + \hat{q}_n \right]$$

Note this will not work in general ...

Given new (X_{n+1}, Y_{n+1}), the residual $R_{n+1} = |Y_{n+1} - \hat{f}_n(X_{n+1})|$ is not exchangeable with the first n residuals—so we have broken symmetry

Conformal prediction: split

One way to preserve symmetry (achieve exchangeability): split the training data into proper training set D_1 and calibration set D_2

Leads to split conformal prediction:

- Compute $\hat{f}_{n_1}$ on proper training set (X_i, Y_i), $i \in D_1$
- Form calibration set residuals $R_i = |Y_i - \hat{f}_{n_1}(X_i)|$, $i \in D_2$
- Let $\hat{q}_{n_2} = \lceil (1-\alpha)(n_2+1) \rceil$ smallest of R_i, $i \in D_2$, and define

$$\hat{C}_n(x) = \left[\hat{f}_{n_1}(x) - \hat{q}_{n_2},\ \hat{f}_{n_1}(x) + \hat{q}_{n_2} \right]$$

By exchangeability, we have finite-sample coverage:

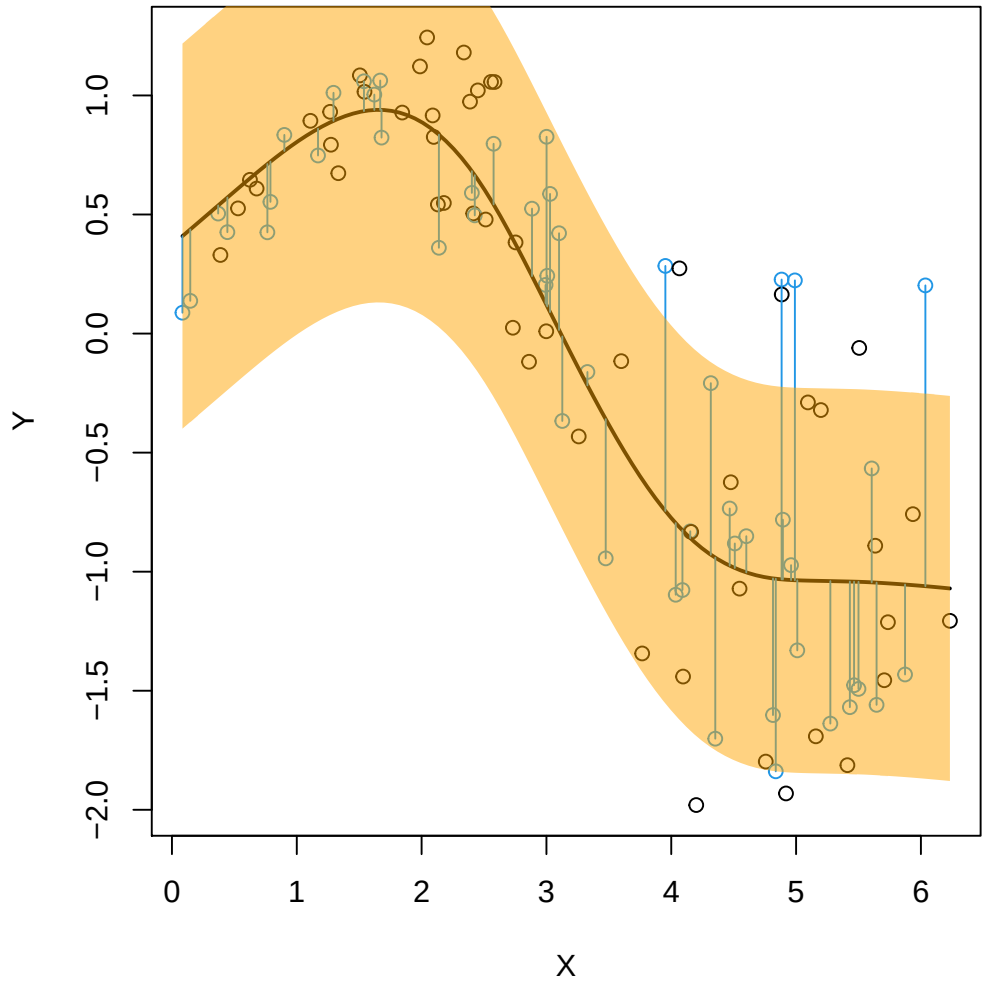
$$\mathbb{P}\Big(Y_{n+1} \in \hat{C}_n(X_{n+1}) \Big) \in \left[1-\alpha,\ 1-\alpha+\tfrac{1}{n_2+1} \right]$$

Some remarks

- No assumptions on P, no asymptotics
- Naive band is going to generally undercover. Close to correct when estimate $\hat{f}_n$ is accurate enough (requires assumptions)
- Split conformal band is protected against overfitting as test residual is is compared to calibration set residuals
- Gives prediction bands with exactly constant length in x
- Generally, the better the estimate $\hat{f}_{n_1}$, the tighter the band
- Multiple splits, say k splits, can be used, each at the nominal level $1 - \alpha/k$. Then $\hat{C}_n(x) = \bigcap_{j=1}^{k} \hat{C}_n^j(x)$ will have coverage at least $1 - \alpha$ (this is just Bonferroni)

Worked example: split

Example: split conformal, using smoothing spline with 5 df



Fit + quantiles to get split conformal band

Conditional coverage?

Calibration set residuals and test residuals are actually exchangeable conditional on the proper training set ... leads to

$$\mathbb{P}\Big(Y_{n+1} \in \hat{C}_n(X_{n+1}) \,\Big|\, \{(X_i, Y_i)\}_{i \in D_1}\Big) \in \Big[1-\alpha,\, 1-\alpha+\tfrac{1}{n_2+1}\Big]$$

What about coverage conditional on X_{n+1}? This **does not hold!** The split conformal coverage guarantee is marginal over X_{n+1}:

$$\int \mathbb{P}\Big(Y_{n+1} \in \hat{C}_n(x) \,\Big|\, \{(X_i, Y_i)\}_{i \in D_1},\, X_{n+1} = x\Big)\, dP_X(x) \in \Big[1-\alpha,\, 1-\alpha+\tfrac{1}{n_2+1}\Big]$$

Distribution-free, finite-sample coverage conditional on $X_{n+1} = x$, at each x, would be nice. Alas, this is asking for too much ...

Conformal prediction: full

Can we do this without splitting? Enter conformal prediction. Fix $x \in \mathcal{X}$. For each trial value $y \in \mathbb{R}$:

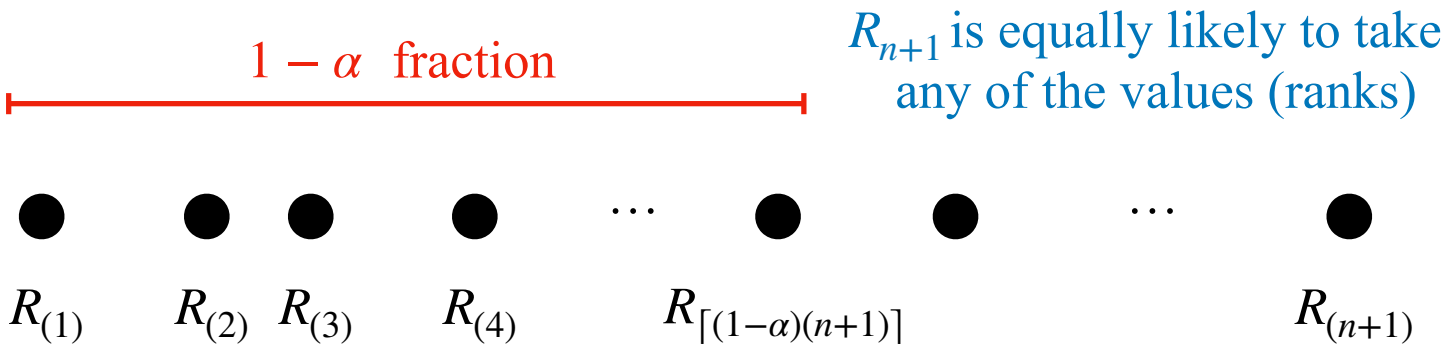
- Compute $\hat{f}_{n,(x,y)}$ on $(X_1, Y_1), \ldots, (X_n, Y_n), (x, y)$
- Form residuals $R_i^{(x,y)} = |Y_i - \hat{f}_{n,(x,y)}(X_i)|$, $i = 1, \ldots, n$, and $R_{n+1}^{(x,y)} = |y - \hat{f}_{n,(x,y)}(x)|$
- Define

$$\hat{C}_n(x) = \Big\{ y \in \mathbb{R} : R_{n+1}^{(x,y)} \leq \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } R_1^{(x,y)}, \ldots, R_n^{(x,y)} \Big\}$$

By exchangeability, we have finite-sample coverage:

$$\mathbb{P}\Big(Y_{n+1} \in \hat{C}_n(X_{n+1})\Big) \in \Big[1-\alpha,\, 1-\alpha+\tfrac{1}{n+1}\Big]$$

Simple illustration



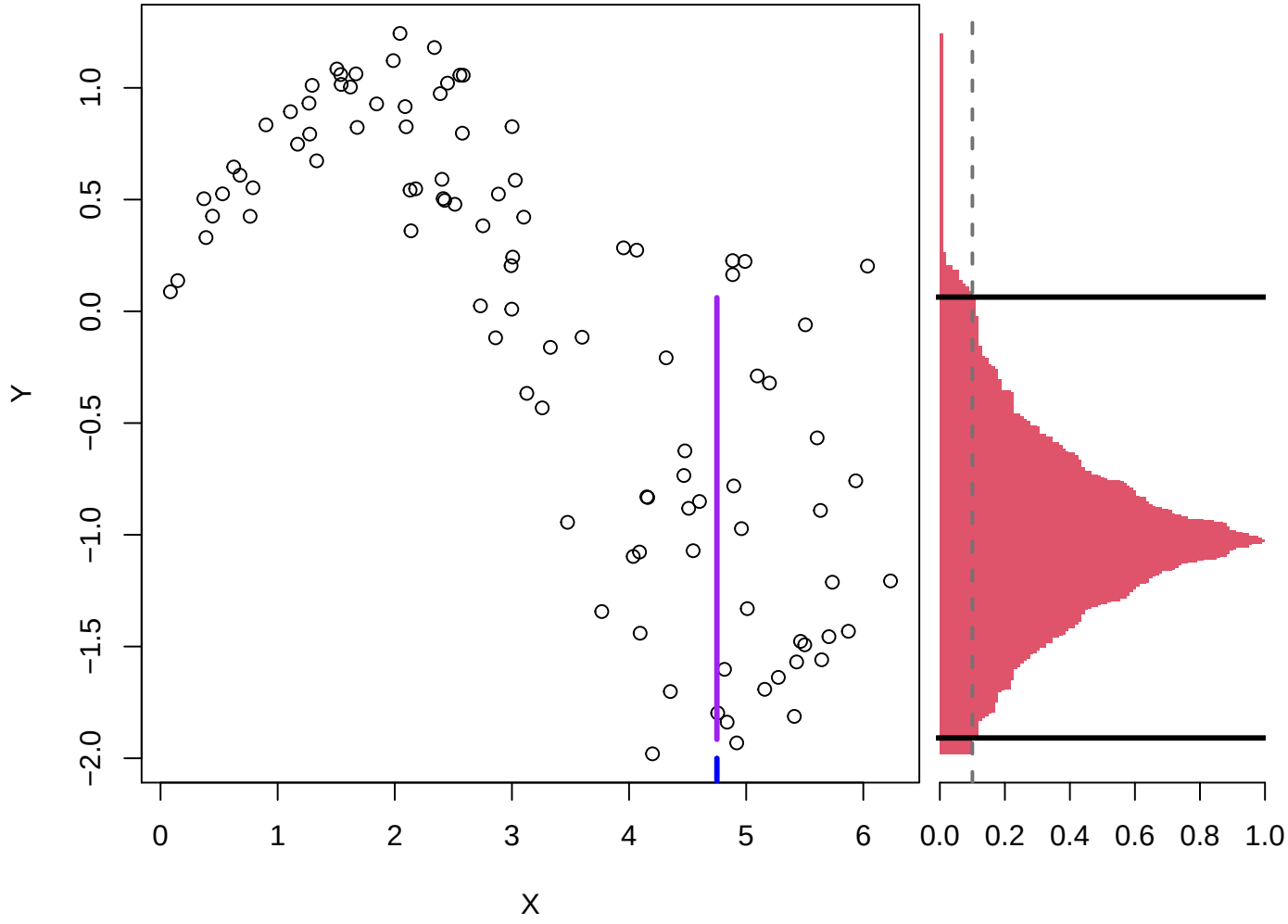
Here each $R_i = R_i^{(x,y)}$ and we are seeking y such that $R_{n+1}^{(x,y)}$ is among red points

Some remarks

- Again, no assumptions on P, no asymptotics
- Conformal band is protected against overfitting as computation of $\hat{f}_{n,(x,y)}$ involves new point (x, y)
- Again, the better the regression algorithm, the tighter the band
- Let $p(y)$ be fraction of residuals $R_1^{(x,y)}, \ldots, R_n^{(x,y)}$ larger than $R_{n+1}^{(x,y)}$. Essentially, $\hat{C}_n(x)$ contains all y such that $p(y) \geq \alpha$
- Informally, $p(y)$ is a p-value for testing $H_0 : Y_{n+1} = y$
- The residuals can be replaced by suitable nonconformity score, $R_i = \mathcal{S}((X_i, Y_i), \{(X_i, Y_i)\}_{i=1}^{n+1})$ (works for both full and split)

Worked example: full

Example: conformal prediction, using smoothing spline with 15 df



Threshold p-values to get full conformal interval

Conditional coverage?

As in split version, coverage guarantee in full conformal prediction is marginal over X_{n+1}, not conditional. Repeat: **not conditional!**

Theorem (Vovk, 2012; Lei and Wasserman, 2014). Let $\hat{C}_n$ be a prediction band satisfying:

$$\mathbb{P}\Big(Y_{n+1} \in \hat{C}_n(x) \,\Big|\, X_{n+1} = x\Big) \geq 1 - \alpha, \text{ for all } P, \text{ and a.e. } x.$$

Then for any P and any non-atom point x_0:

$$\mathbb{P}\left(\lim_{\delta \to 0} \sup_{x \in B_\delta(x_0)} \mu\big(\hat{C}_n(x)\big) = \infty\right) = 1$$

Note: we can get asymptotic conditional coverage, but this requires assumptions strong enough for consistency (Lei, G'Sell, Rinaldo, T., Wasserman 2018)

Approximate conditional coverage?

How about replacing conditional requirement by:

$$\mathbb{P}\Big(Y_{n+1} \in \hat{C}_n(x) \,\Big|\, X_{n+1} \in \mathcal{X}_0\Big) \geq 1-\alpha,$$

for all P, and $\mathcal{X}_0 \subseteq \mathcal{X}$ such that $P_X(\mathcal{X}_0) \geq \delta$

Barber, Candes, Ramdas, T. (2021) show that this is still too hard, in that the only solutions are "trivial"

- In order for this to be tractable, we must restrict our attention somehow, and cannot ask for every subset $\mathcal{X}_0 \subseteq \mathcal{X}$
- For example, a finite collection $\{\mathcal{X}_1, \ldots, \mathcal{X}_k\}$ (then we can just run local conformal prediction)
- Or, an infinite collection of "nice" subsets $\mathcal{X}_0$ (i.e., a collection with low VC dimension)

Conformal: locally-weighted

Using locally-weighted residuals in conformal methods can bring us closer to conditional coverage in practice, under heteroskedasticity:

$$\text{replace } \; R_i = |Y_i - \hat{f}_n(X_i)| \; \text{ by } \; V_i = \frac{|Y_i - \hat{f}_n(X_i)|}{\hat{\sigma}_n(X_i)}$$

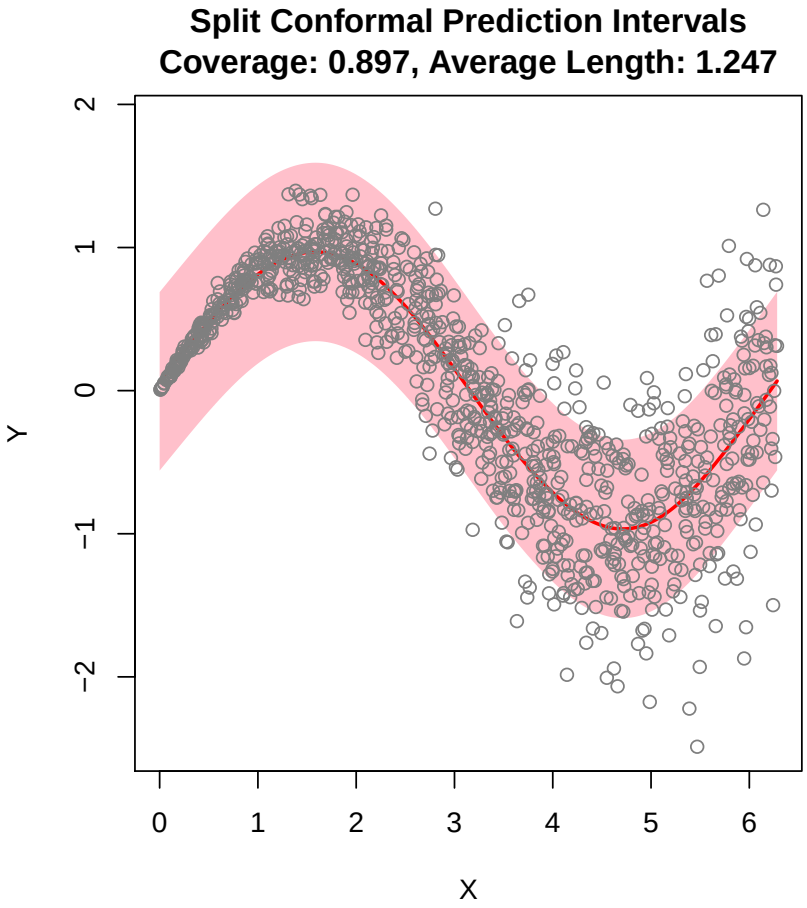
where $\hat{\sigma}_n^2(x)$ is an estimate of the variance function of the absolute residual $\mathrm{Var}(|Y - \hat{f}_n(X)| \,|\, X = x)$. (Note: $\hat{f}_n, \hat{\sigma}_n$ can be estimated jointly, or separately)

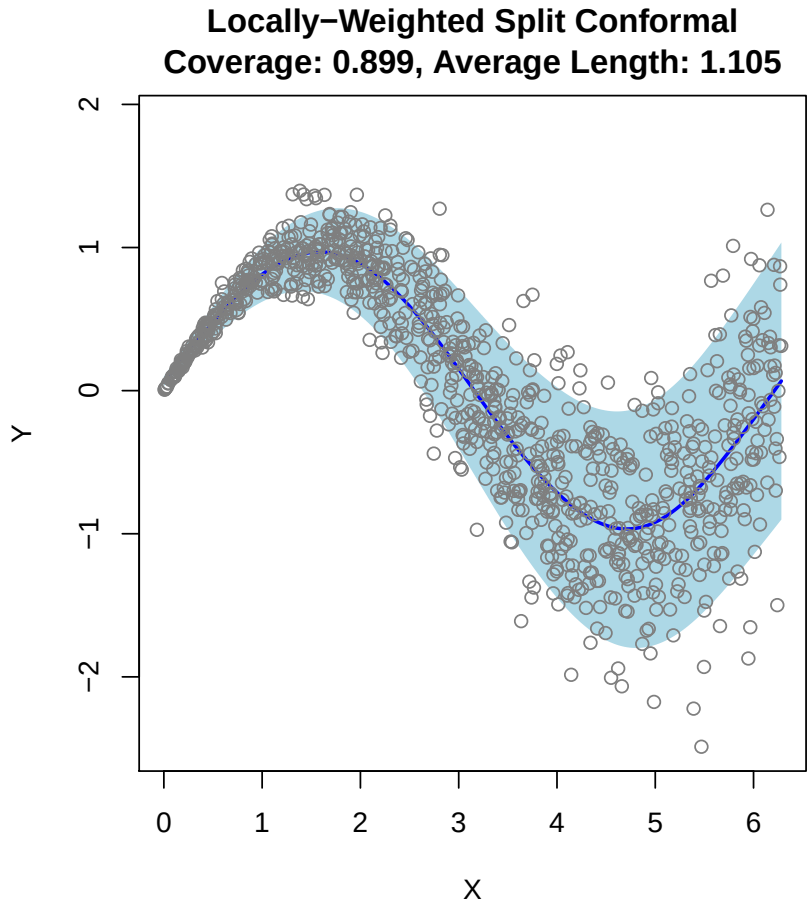
The effect of local-weighting on the prediction band: its local length can vary considerably, as needed. In the split version:

$$\hat{C}_n(x) = \Big[\hat{f}_{n_1}(x) - \hat{\sigma}_{n_1}(x)\hat{q}_{n_2},\; \hat{f}_{n_1}(x) + \hat{\sigma}_{n_1}(x)\hat{q}_{n_2}\Big],$$

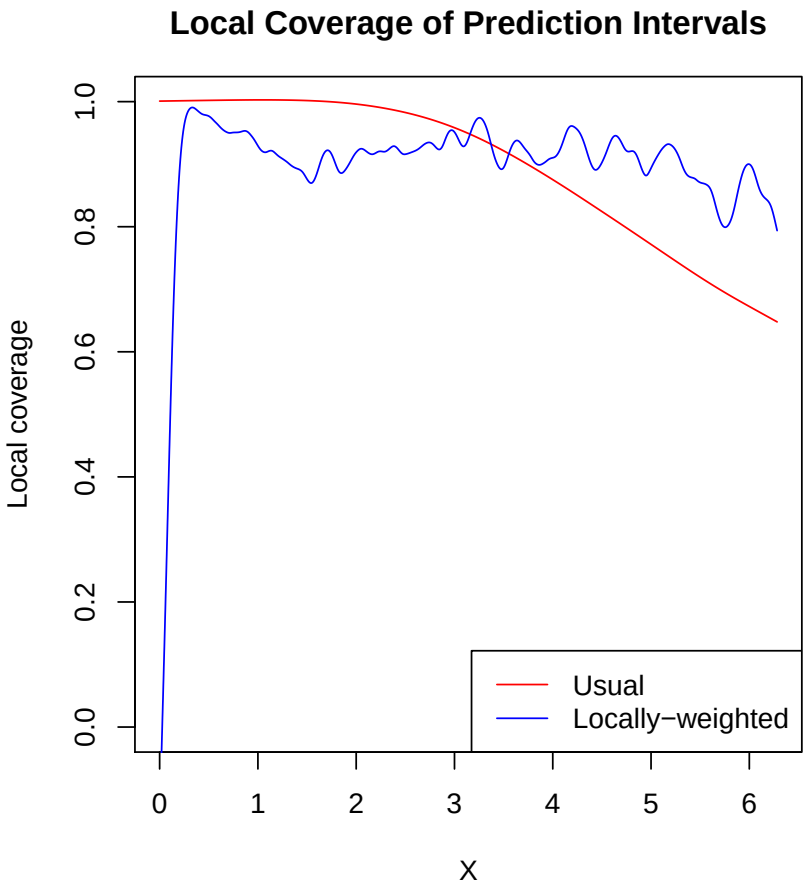
where $\hat{q}_{n_2} = \lceil (1-\alpha)(n_2+1) \rceil$ smallest of V_i, $i \in D_2$

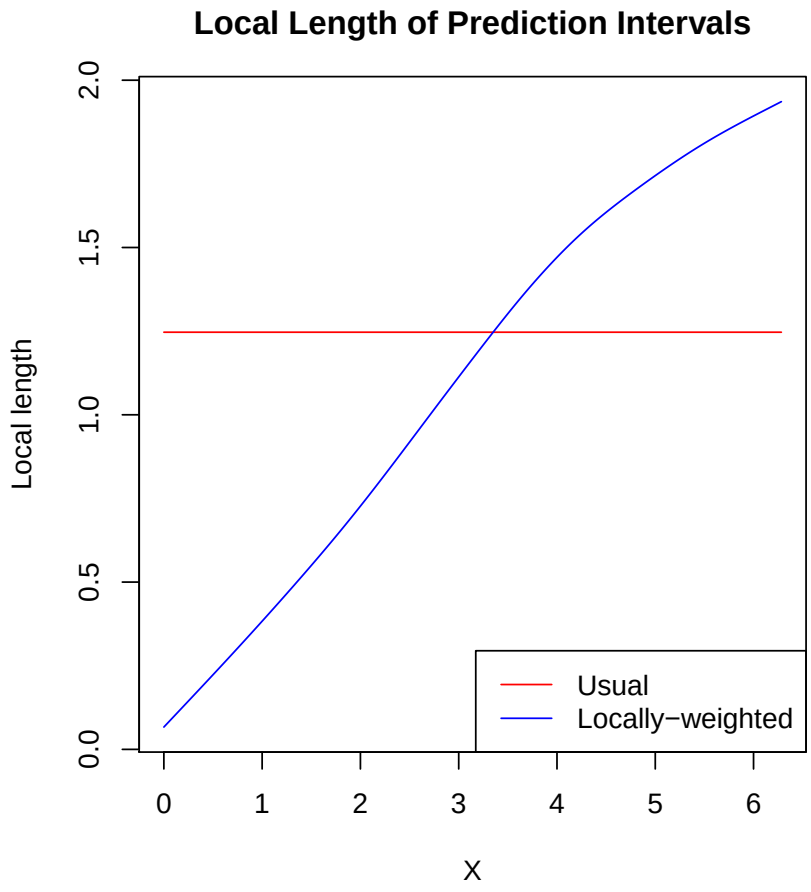
Example: local weighting (Lei et al. 2018)





Example: local weighting (Lei et al. 2018)





A little history

- The idea behind conformal prediction was apparently born out of conversations between Vovk, Gammerman, Vapnik in 1990s
- Definitive reference is Vovk, Gammerman, Shafer (2005)
- Vovk and collaborators still very active, 36 papers on this topic since 2009 (`http://alrw.net`)
- Lei and Wasserman sparked interest in statistics: conformal for nonparametric density estimation and regression (2013, 2014)
- Lei, G'Sell, Rinaldo, T., Wasserman (2018): conformal for high-dimensional regression, some new theory & methods
- Barber, Candes, Ramdas, T. (2019-22): conformal for covariate shift, jackknife+ and CV+, control beyond exchangeability
- Explosion of interest in ML (2020+): 1000s of new papers

Outline

Rest of talk:

- Experiments
- Jackknife+
- CQR
- Classification
- Covariate shift
- Distribution shift
- Conclusion

Experiments

Experimental setup

From Lei et al. (2018). Settings:

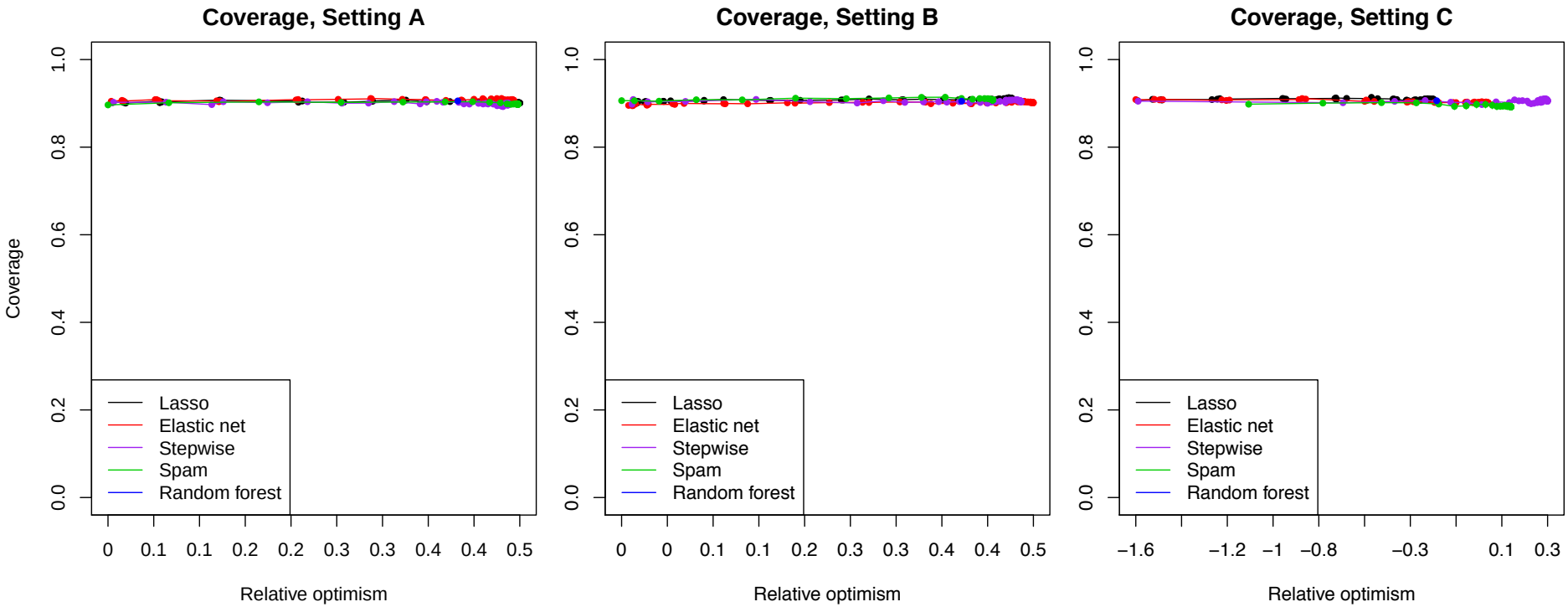
- A: linear mean, uncorrelated features, normal errors
- B: nonlinear mean, uncorrelated features, heavy-tailed errors
- C: linear mean, correlated features, heteroskedastic errors

For $n = 200$ and $d = 2000$, we'll compare coverages, lengths, and test errors across various base algorithms:

- lasso
- elastic net
- stepwise regression
- sparse additive models
- random forests

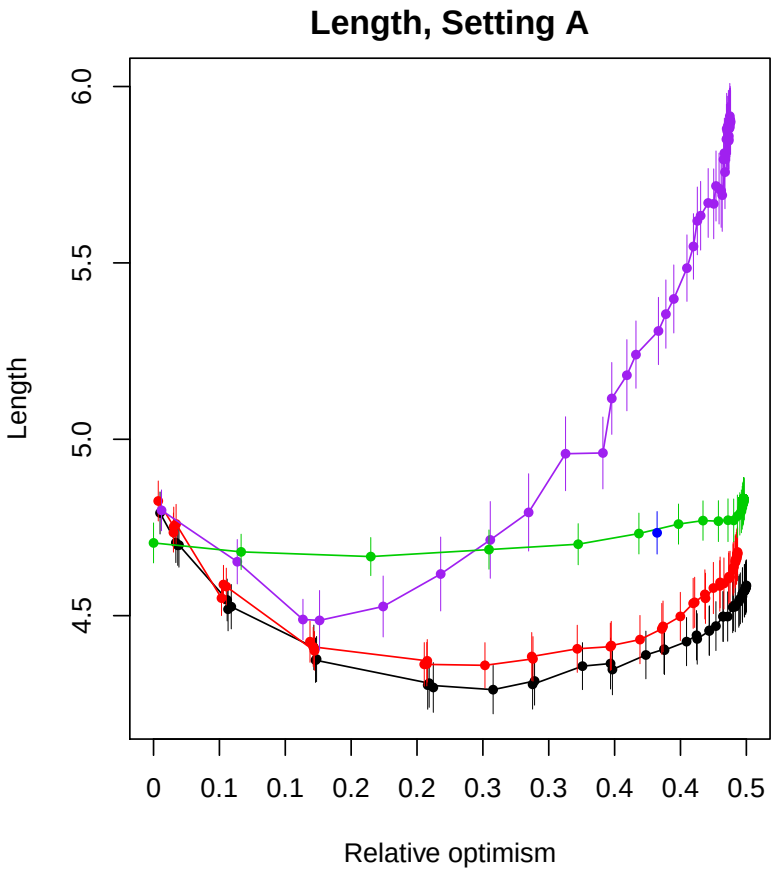
Coverage (Lei et al. 2018)

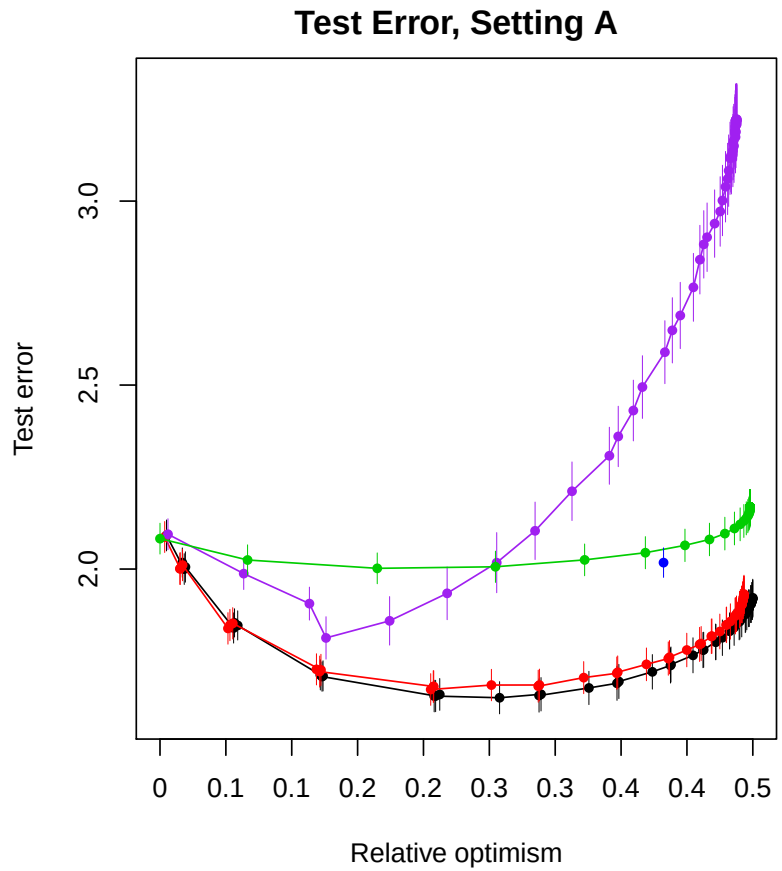
Coverages: pretty much exactly at the nominal level, $1 - \alpha = 0.9$



Length and test error (Lei et al. 2018)

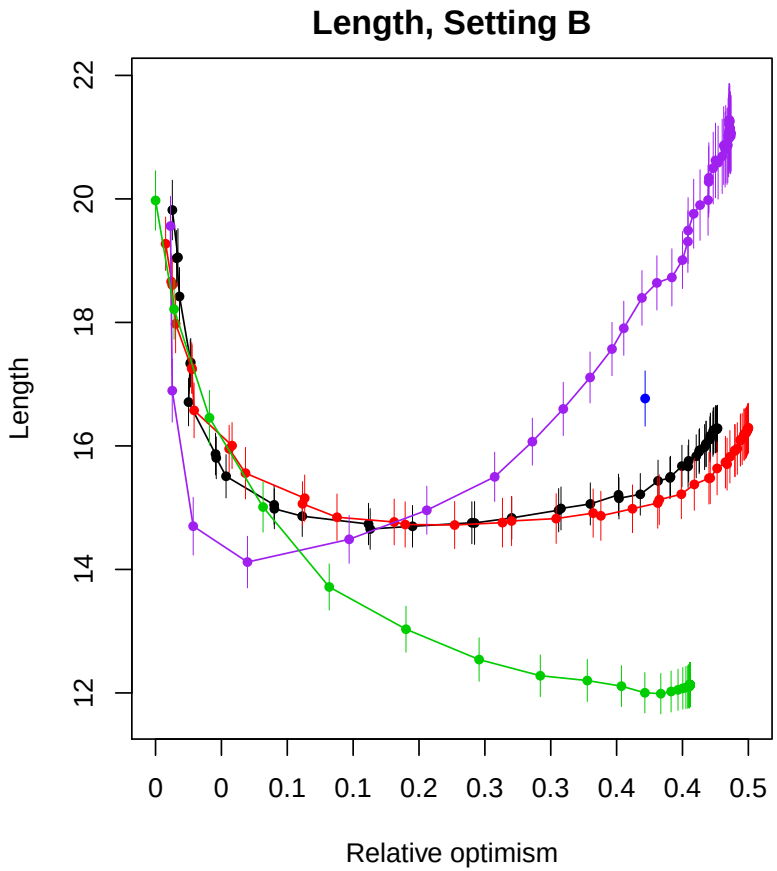
Setting A: linear mean, uncorrelated features, normal errors

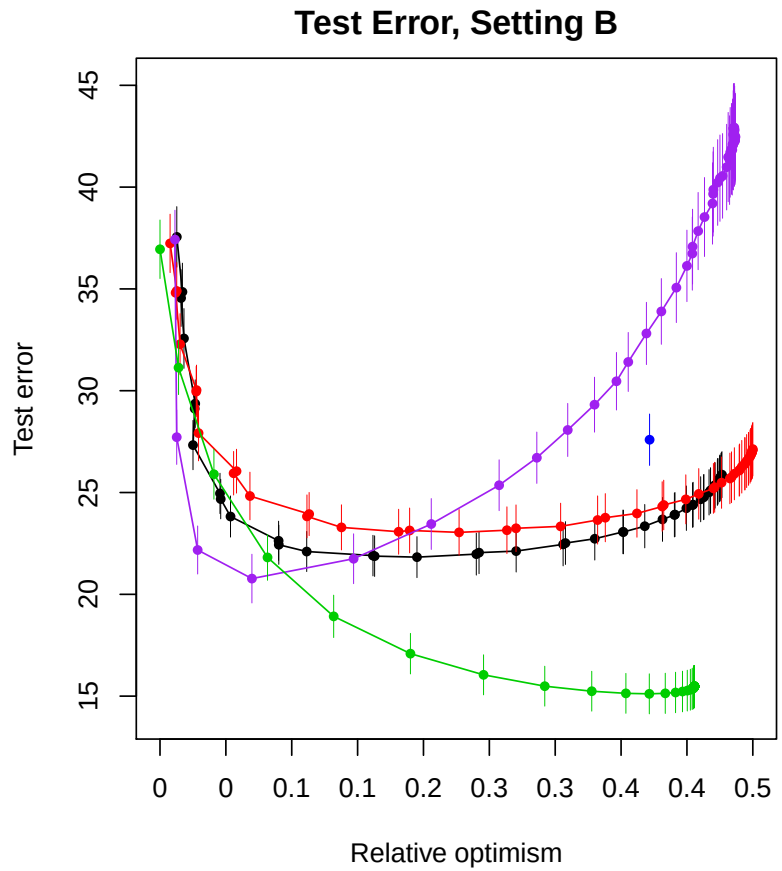




Length and test error (Lei et al. 2018)

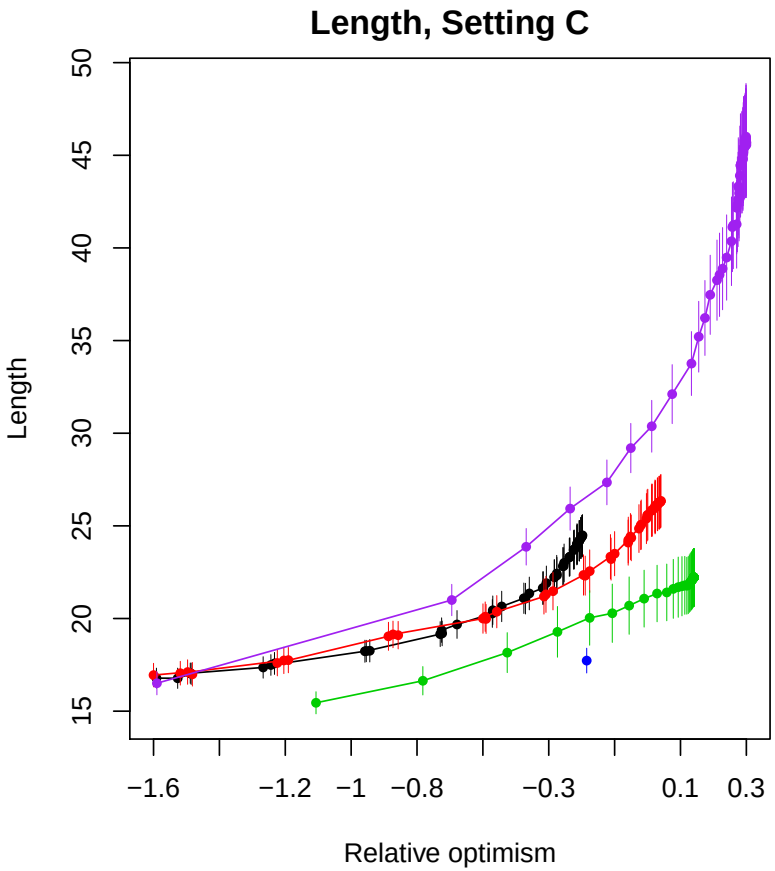
Setting B: nonlinear mean, uncorrelated features, heavy-tailed errors

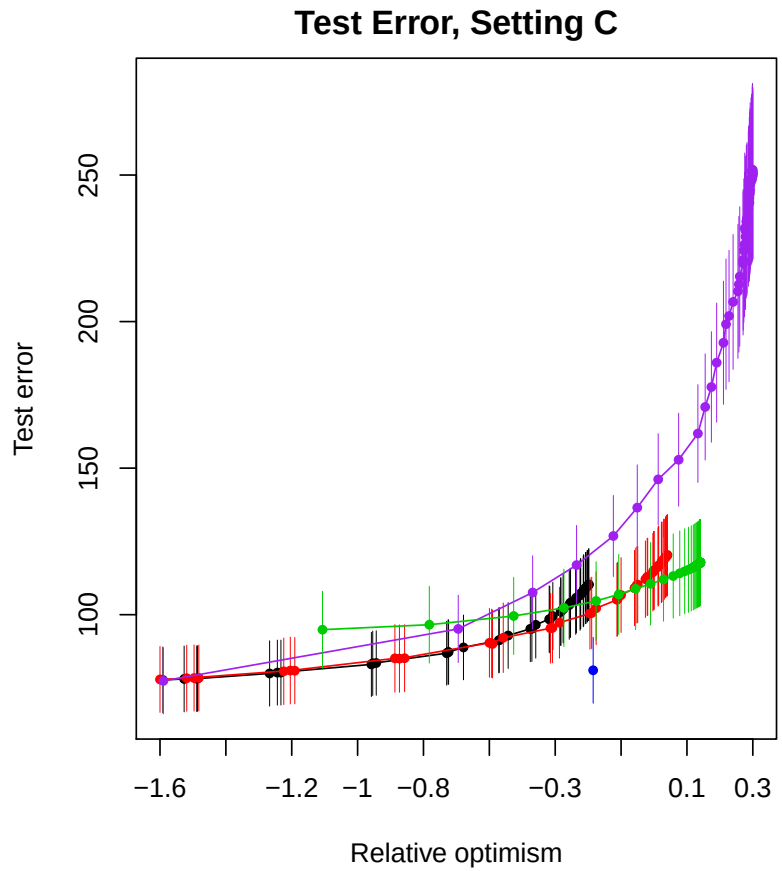




Length and test error (Lei et al. 2018)

Setting C: linear mean, correlated features, heteroskedastic errors





An observation

In all cases, average interval length correlates pretty strongly with test error (theory in Lei et al. 2018 supports this)

Both intuitive and surprising, considering that the coverage is nearly exact, regardless of the algorithm/choice of tuning parameter

How can this be? Average length is:

$$\mathbb{E}\Big(\mu\big(\hat{C}_n(X_{n+1})\big)\Big) = \mathbb{E}_{P^n}\bigg[\int\int_{\hat{C}_n(x)} d\mu(y)\, dP_X(x)\bigg]$$

Meanwhile, coverage is:

$$\mathbb{P}\Big(Y_{n+1} \in \hat{C}_n(X_{n+1})\Big) = \mathbb{E}_{P^n}\bigg[\int\int_{\hat{C}_n(x)} dP_{Y|X}(y)\, dP_X(x)\bigg]$$

Therefore an inefficient algorithm puts mass in low density regions of $P_{Y|X}$, which does not hurt its coverage, but inflates its length

Software tools

R package `conformalInference`, available at:

`https://github.com/ryantibs/conformal/`

- Conformal methods: full, split, jackknife, locally-weighted, ...
- Base algorithms: built-in or custom ones (functional approach)
- Reproduce all results from Lei et al. (2018) and T. et al. (2019)

Code example:

```
funs = rf.funs(ntree = 2000)
obj = conformal.pred(x, y, x0, alpha = 0.1,
                     train.fun = funs$train,
                     predict.fun = funs$predict)
```

See also Angelopoulos & Bates (2021) and references therein

Jackknife+

Classical jackknife

Jackknife has a rich history in statistics. Can we use it for predictive inference?

- For each $i = 1, \ldots, n$, fit estimate $\hat{f}_{n-1}^{(-i)}$ on (X_j, Y_j), $j \neq i$
- Form jackknife (leave-one-out) residuals $R_i = |Y_i - \hat{f}_{n-1}^{(-i)}(X_i)|$, $i = 1, \ldots, n$
- Let $\hat{q}_n = \lceil (1-\alpha)(n+1) \rceil$ smallest of $R_1, \ldots, R_n$, and define

$$\hat{C}_n(x) = \Big[\hat{f}_n(x) - \hat{q}_n,\ \hat{f}_n(x) + \hat{q}_n\Big]$$

Remarks:

- Computational cost is in between full and split conformal (can get big speed ups for some special linear smoothers)
- But unlike conformal methods, we do not have out-of-sample coverage guarantees ... jackknife is only in-sample symmetric

Just add a plus

Notation: $Q^+_{n,\alpha}\{v_i\} = \lceil (1-\alpha)(n+1) \rceil$ smallest of $v_1, \ldots, v_n$, and $Q^-_{n,\alpha}\{v_i\} = -q^-_{n,\alpha}\{-v_i\}$. Jackknife interval can be written as:

$$\hat{C}_n(x) = \Big[Q^-_{n,\alpha}\{\hat{f}_n(x) - R_i\},\, Q^+_{n,\alpha}\{\hat{f}_n(x) + R_i\} \Big]$$

Jackknife+ interval (Barber, Candes, Ramdas, T. 2021):

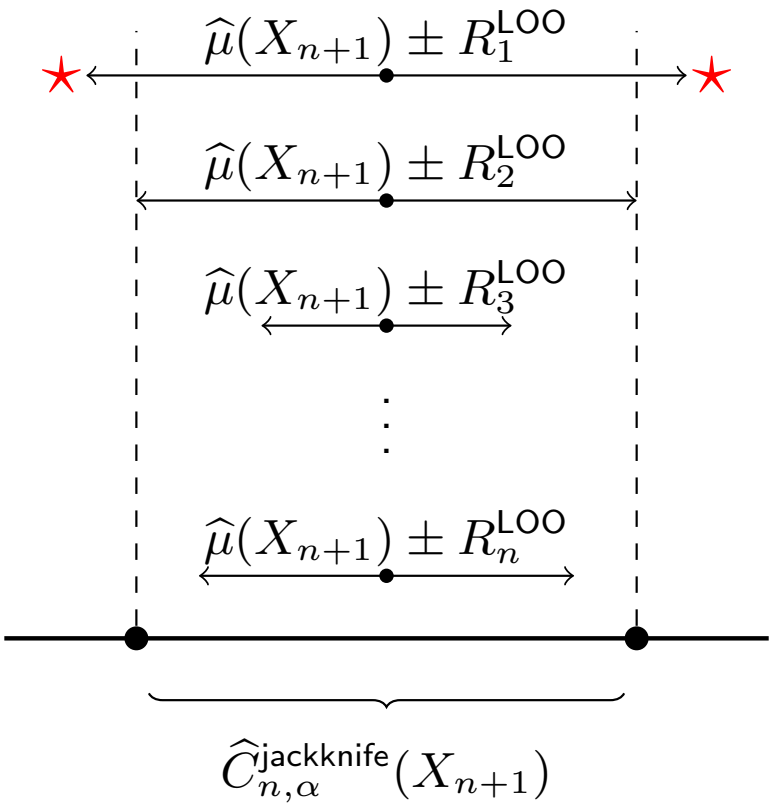
$$\hat{C}^+_n(x) = \Big[Q^-_{n,\alpha}\{\hat{f}^{(-i)}_n(x) - R_i\},\, Q^+_{n,\alpha}\{\hat{f}^{(-i)}_n(x) + R_i\} \Big]$$

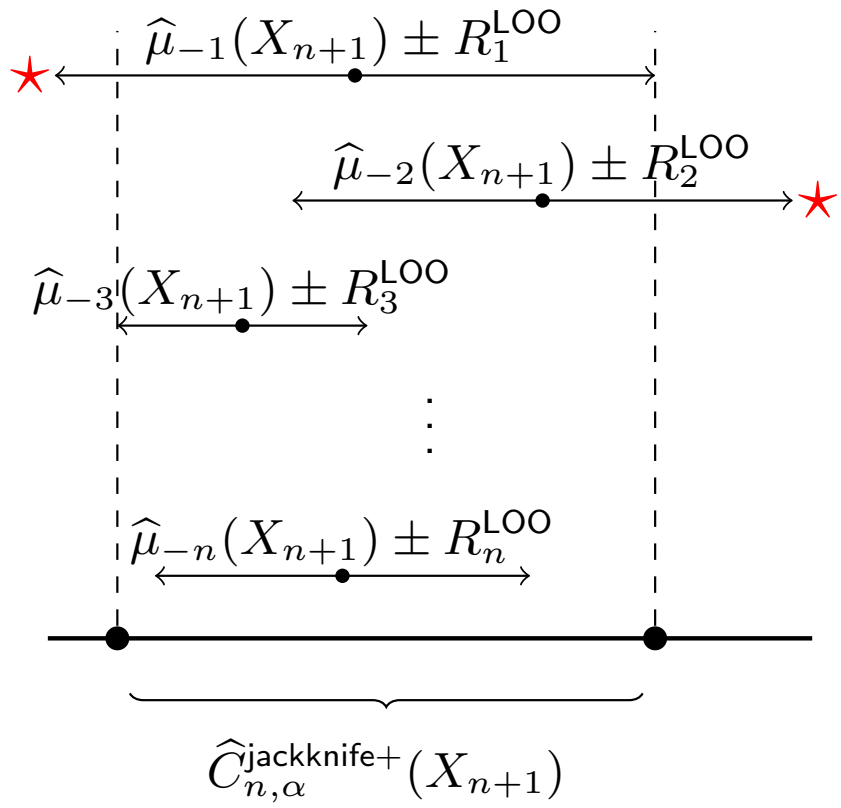
Under exchangeability, we have finite-sample coverage:

$$\mathbb{P}\Big(Y_{n+1} \in \hat{C}^+_n(X_{n+1}) \Big) \geq 1 - 2\alpha$$

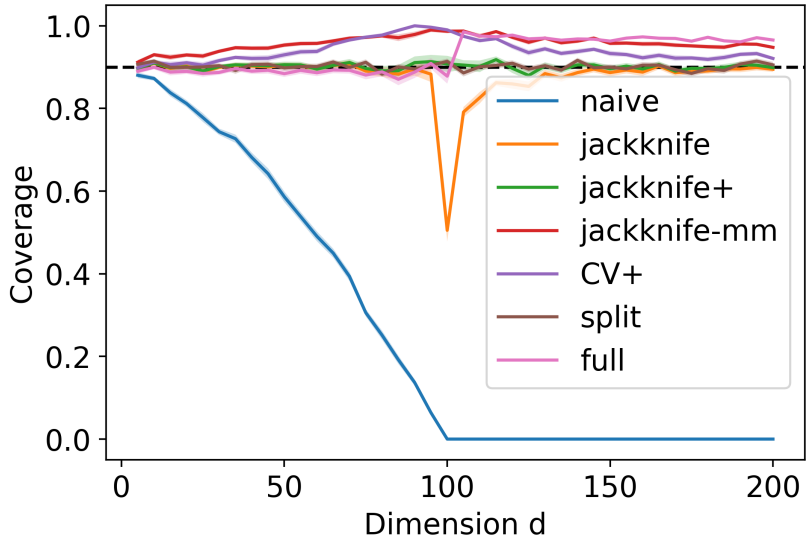
Barber et al. (2021) also derive cross-validation variant called CV+. Closely related to cross-conformal of Vovk (2015)

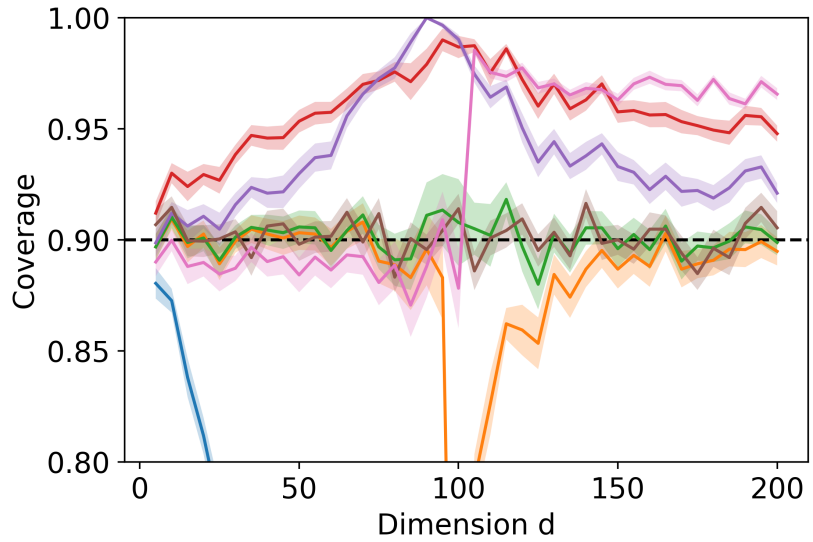
Jackknife+ illustration (Barber et al. 2021)

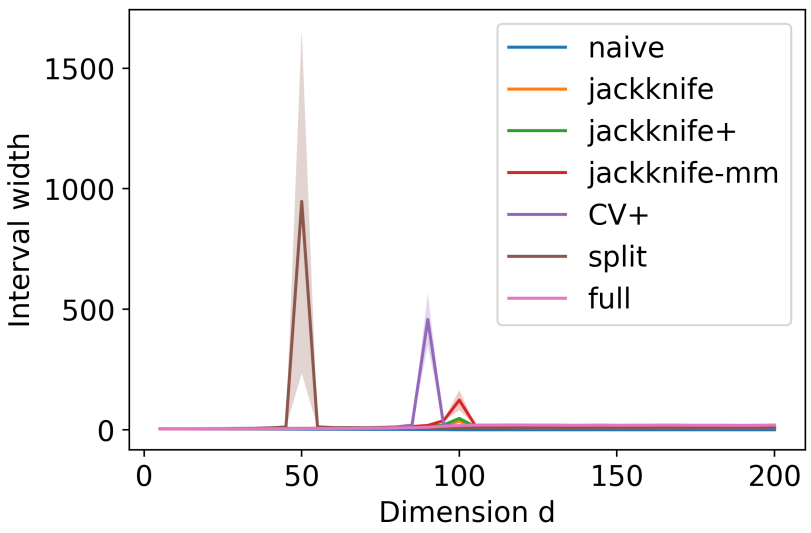


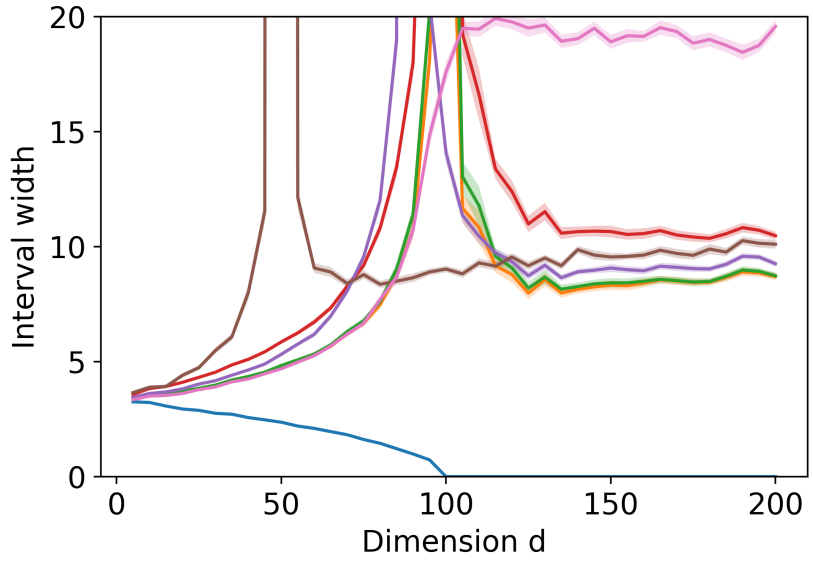


Jackknife+ examples (Barber et al. 2021)









CQR

Challenges of local weighting

When implementing local-weighted conformal, easiest approach is to fit $\hat{f}_n, \hat{\sigma}_n$ separately: first fit $\hat{f}_n$, then fit $\hat{\sigma}_n$ based on residuals to $\hat{f}$

Challenge: if using training residuals, and $\hat{f}_n$ is complex, then very little information is left in residuals to estimate $\hat{\sigma}_n$

To circumvent this, we must either:

- split further (undesirable); or
- train jointly, which is typically not easy to do out-of-the-box

Once jointly estimating $\hat{f}_n, \hat{\sigma}_n$, may as well estimate and calibrate a conditional quantile of $Y = X|x$, which leads us to ...

Conformal quantile regression

Conformal quantile regression (CQR, Romano et al. 2019): employs a particular nonconformity score based on quantile regression. In the split version:

- Compute $\hat{f}_{n_1}^{\alpha/2}$ and $\hat{f}_{n_1}^{1-\alpha/2}$ on (X_i, Y_i), $i \in D_1$, where $\hat{f}_{n_1}^{\tau}(x)$ estimates $\mathrm{Quantile}(\tau; Y|X = x)$
- Form calibration set "residuals"

$$R_i = \max\left\{\hat{f}_{n_1}^{\alpha/2}(X_i) - Y_i,\, Y_i - \hat{f}_{n_1}^{1-\alpha/2}(X_i)\right\},\ i \in D_2$$

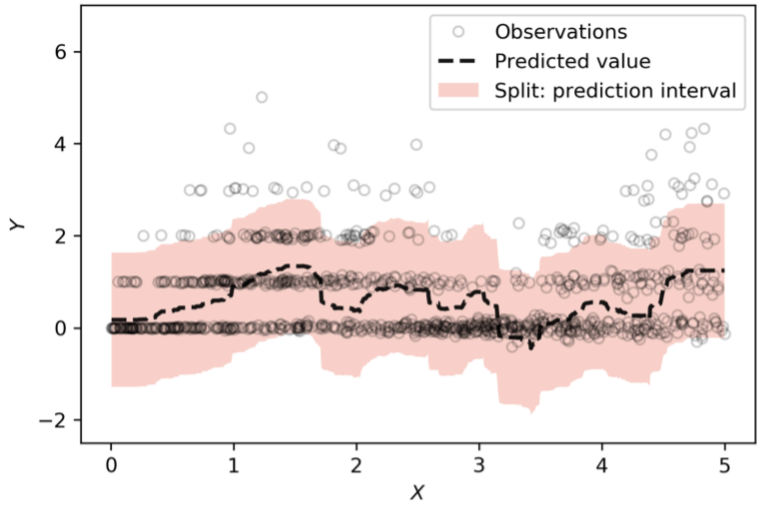
- Let $\hat{q}_{n_2} = \lceil (1-\alpha)(n_2+1) \rceil$ smallest of R_i, $i \in D_2$, and define

$$\hat{C}_n(x) = \left[\hat{f}_{n_1}^{\alpha/2}(x) - \hat{q}_{n_2},\, \hat{f}_{n_1}^{1-\alpha/2}(x) + \hat{q}_{n_2}\right]$$

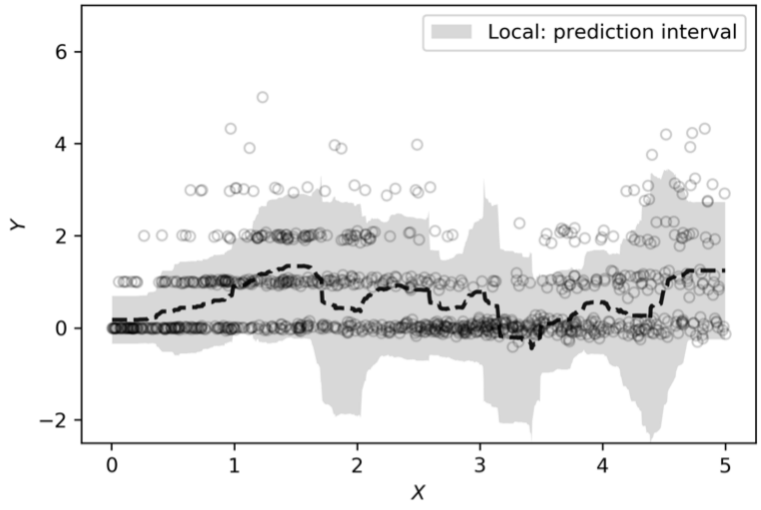
(This has the exact same guarantees as traditional split conformal)

CQR example (Romano et al. 2019)

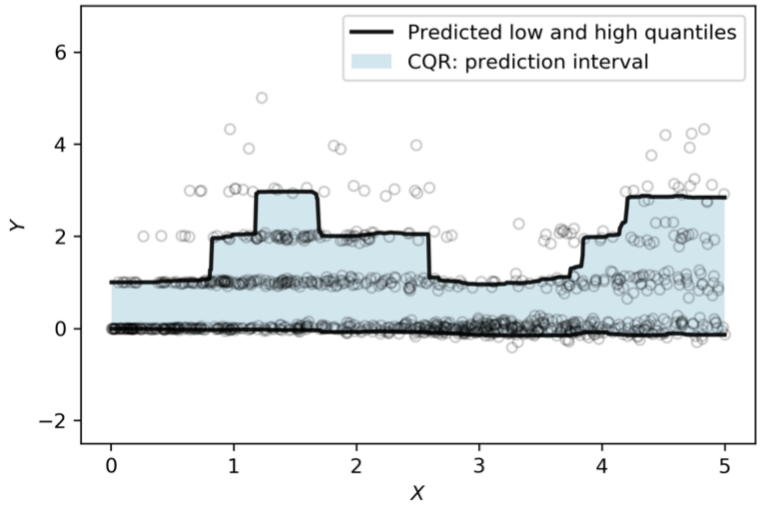
(a) Split: Avg. coverage 91.4%; Avg. length 2.91.



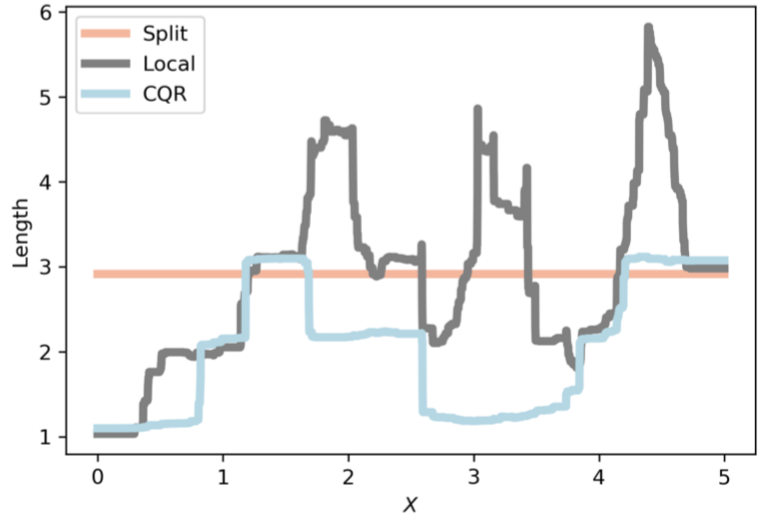
(b) Local: Avg. coverage 91.7%; Avg. length 2.86.



(c) CQR: Avg. coverage 91.06%; Avg. length 1.99.



(d) Length of prediction intervals.



Classification

Conformal classification

Conformal prediction fluidly applies to classification problems, just by changing the nonconformity score

For example, in the split version, using a probabilistic classifier on K classes:

- Compute $\hat{f}_{n_1}$ on (X_i, Y_i), $i \in D_1$, where $\hat{f}_{n_1}(x; k)$ estimates $\mathbb{P}(Y = k|X = x)$, $k = 1, \ldots, K$
- Form calibration set scores $R_i = \hat{f}_{n_1}(X_i; Y_i)$, $i \in D_2$
- Let $\hat{q}_{n_2} = \lceil (1 - \alpha)(n_2 + 1) \rceil$ largest of R_i, $i \in D_2$, and define

$$\hat{C}_n(x) = \left\{ k : \hat{f}_{n_1}(x; k) \geq \hat{q}_{n_2} \right\}$$

(This has the exact same guarantees as traditional split conformal)

Adaptive prediction sets

Seeking to make this more adaptive: smaller sets for easy problems, larger for harder problems. Following Romano et al. (2020), define:

$$R_i = \sum_{j=1}^{k_i} \hat{f}_{n_1}(X_i; \pi_i(j)), \; i \in D_2$$

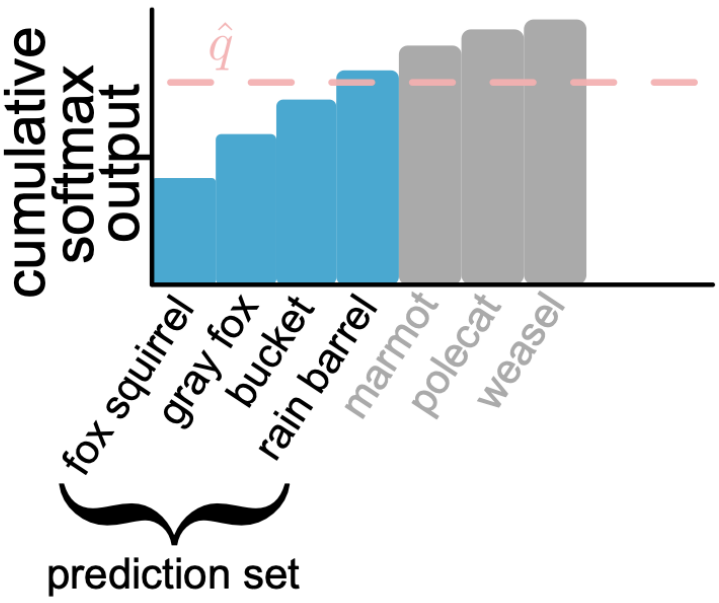
where π_i permutes $1, \ldots, K$ in decreasing order of $\hat{f}_{n_1}(X_i; k)$, and $\pi_i(k_i) = Y_i$

As before, we let $\hat{q}_{n_2} = \lceil (1-\alpha)(n_2+1) \rceil$ smallest of R_i, $i \in D_2$, and the conformal set becomes

$$\hat{C}_n(x) = \{\pi_x(1), \ldots, \pi_x(k_x)\},$$

$$\text{where } k_x = \min \left\{ k : \sum_{j=1}^{k} \hat{f}_{n_1}(x; \pi_x(j)) \leq \hat{q}_{n_2} \right\}$$

APS illustration (Angelopoulos and Bates 2021)



At test time, order class by decreasing predicted probability, include classes until cumulative probability exceeds $\hat{q}_{n_2}$

APS examples (Angelopoulos et al. 2021)

{ fox squirrel 0.99 }

{ fox squirrel 0.82, gray fox, 0.03, bucket, 0.02, rain barrel 0.02 }

{ marmot 0.30, fox squirrel 0.22, mink 0.18, weasel 0.16, beaver 0.03, polecat 0.01 }

Covariate shift

Challenges of covariate shift

Consider now i.i.d. training data $(X_i, Y_i) \sim P$, $i = 1, \ldots, n$, but $(X_{n+1}, Y_{n+1}) \sim \tilde{P}$. When $P = P_X \times P_{Y|X}$ and $\tilde{P} = \tilde{P}_X \times P_{Y|X}$, this is called covariate shift

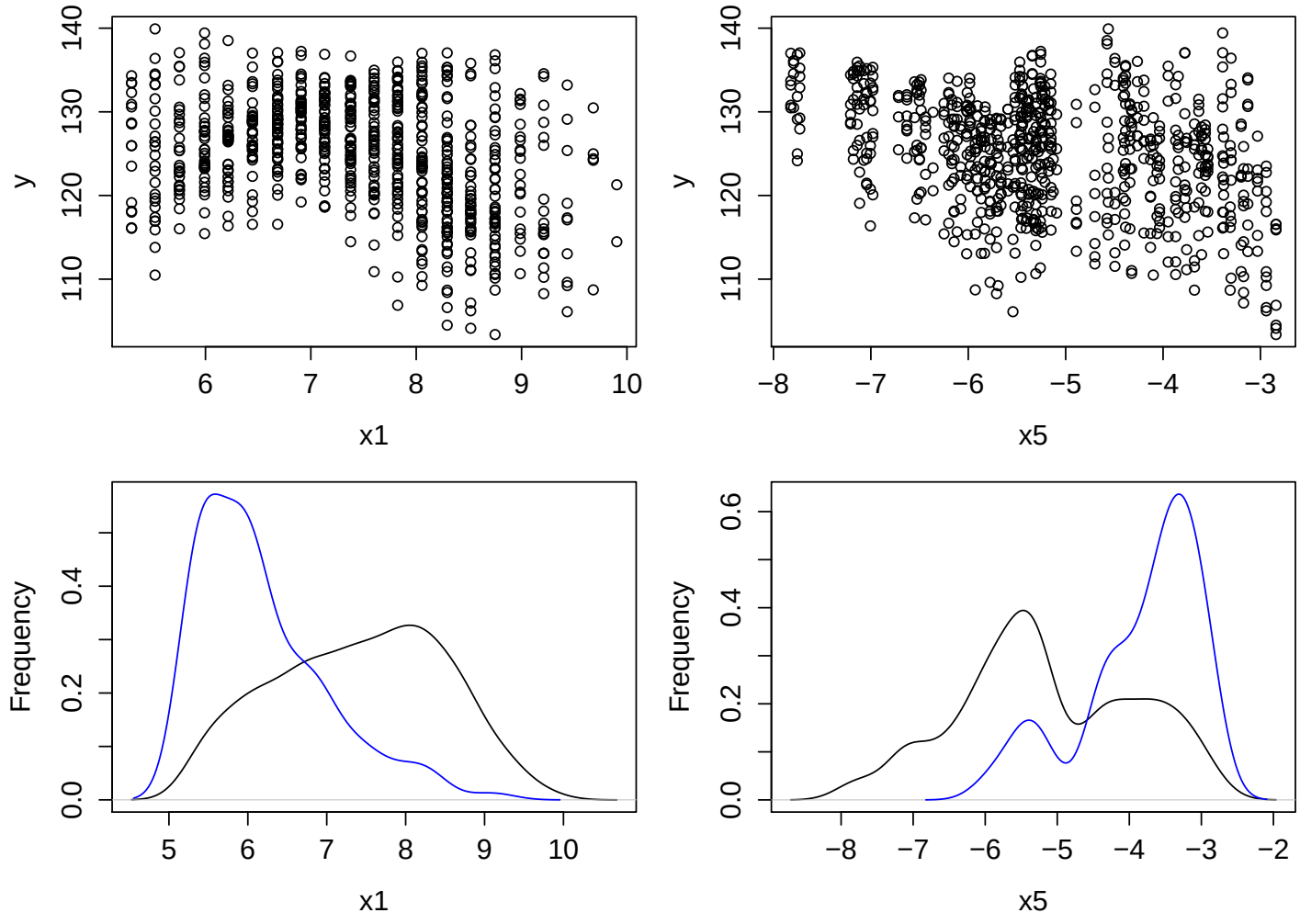
Conformal prediction:

$$\hat{C}_n(x) = \Bigg\{ y \in \mathbb{R} : R_{n+1}^{(x,y)} \leq \mathrm{Quantile}\bigg(1 - \alpha;\ \frac{1}{n+1} \sum_{i=1}^{n} \delta_{R_i^{(x,y)}} + \frac{1}{n+1} \delta_\infty\bigg) \Bigg\}$$

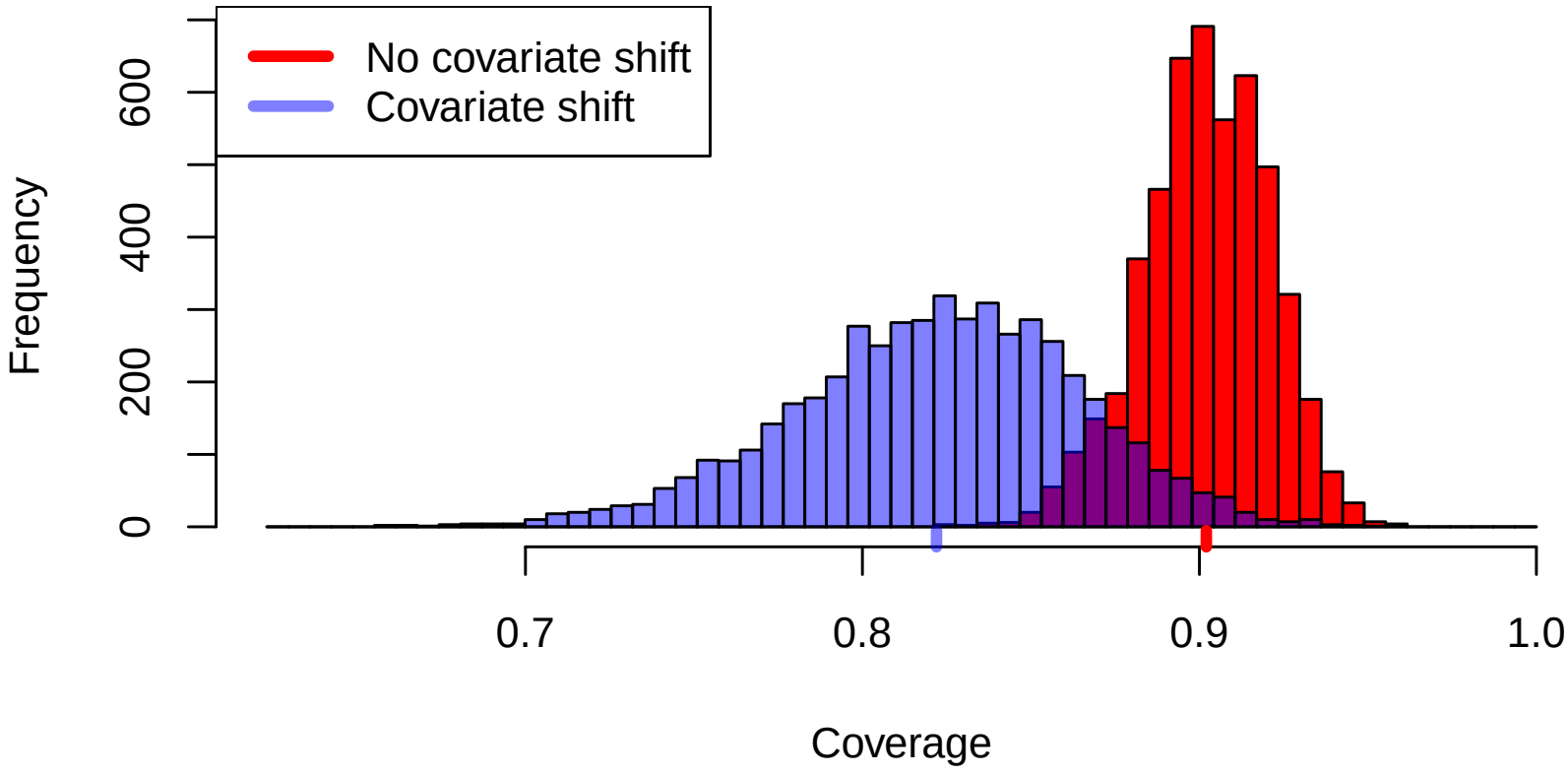
will in general fail, because the nonconformity scores will not be exchangeable. What to do?

Basic idea: from a set of test covariate points, could use importance sampling to get a subset that "looks like" it came from P_X

Covariate shift example (T. et al. 2019)



Covariate shift example (T. et al. 2019)



Conformal for covariate shift

Using importance weighting, this same idea can be applied without sampling (T., Barber, Candes, Ramdas 2019):

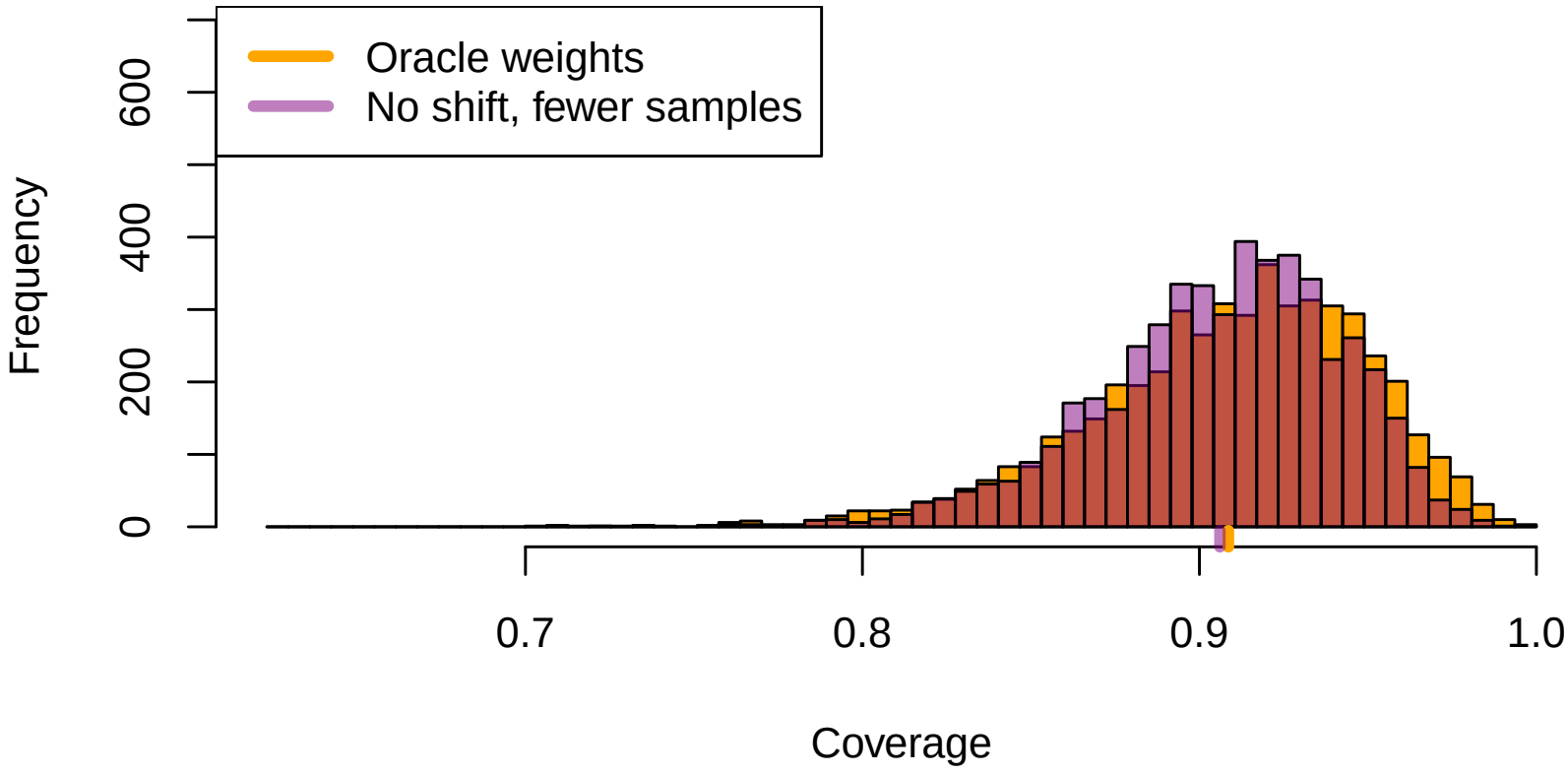
$$\hat{C}_n(x) = \Bigg\{ y \in \mathbb{R} : R_{n+1}^{(x,y)} \leq \text{Quantile}\Bigg(1-\alpha; \frac{\sum_{i=1}^n w(X_i)\delta_{R_i^{(x,y)}} + w(x)\delta_\infty}{\sum_{i=1}^n w(X_i) + w(x)}\Bigg)\Bigg\}$$

where $w = d\tilde{P}_X / dP_X$. Leads to familiar conclusion:

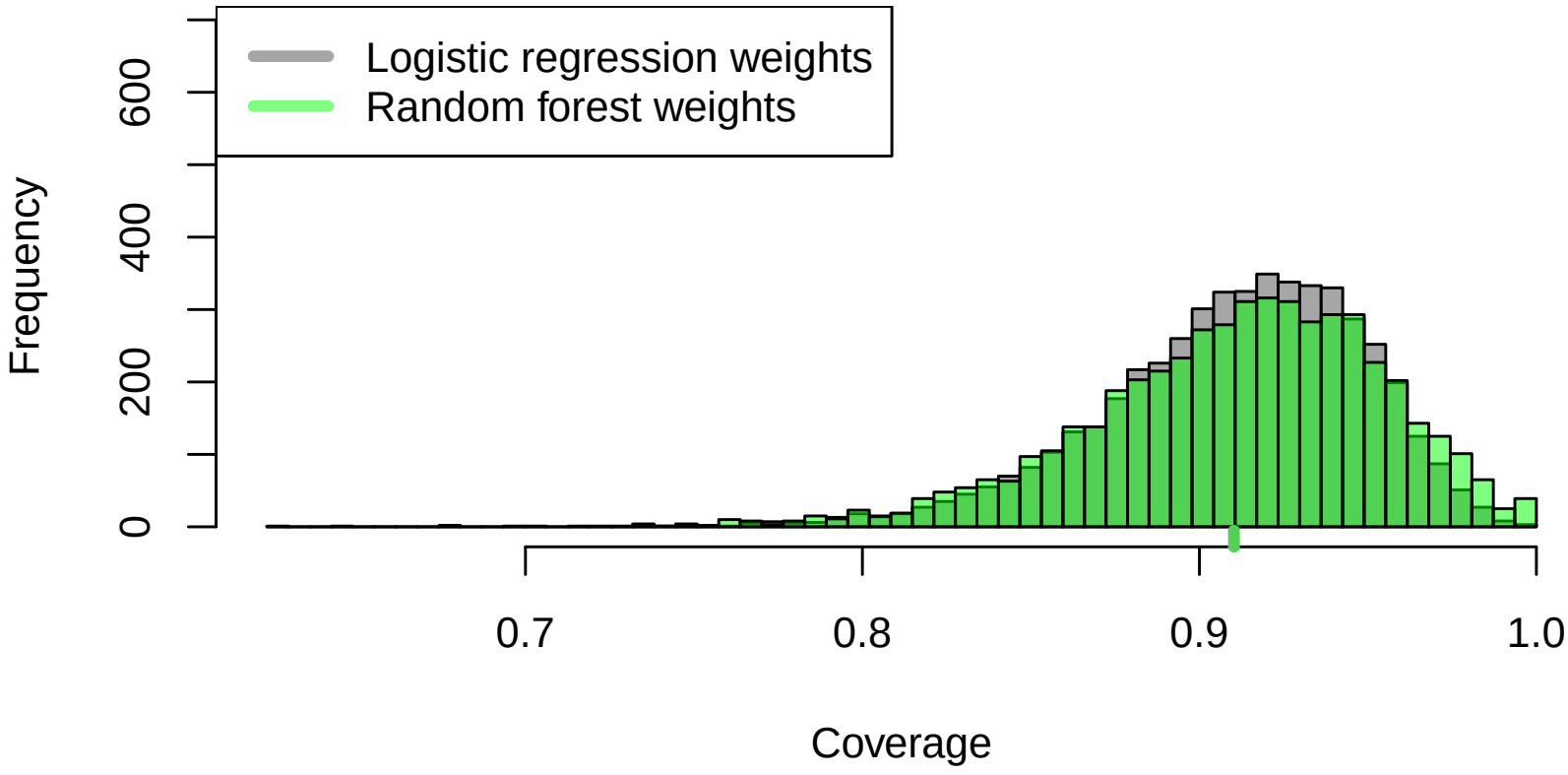
$$\mathbb{P}\Big(Y_{n+1} \in \hat{C}_n(X_{n+1})\Big) \geq 1-\alpha$$

(Note that probability is taken over $(X_i, Y_i) \sim P$, $i = 1, \ldots, n$, and $(X_{n+1}, Y_{n+1}) \sim \tilde{P}$ ... interestingly, upper bound does not translate)

Covariate shift example (T. et al. 2019)



Covariate shift example (T. et al. 2019)



Distribution shift

Conformal under distribution shift

What can we say beyond specific joint dependence conditions (exch, covariate shift, etc.)? Consider conformal prediction sets built from custom-weighted quantiles (arbitrary $0 \leq w_i \leq 1$, $i = 1, \ldots, n$):

$$\hat{C}_n^w(x) = \Bigg\{ y \in \mathbb{R} : R_{n+1}^{(x,y)} \leq \mathrm{Quantile}\bigg(1-\alpha; \frac{\sum_{i=1}^n w_i \delta_{R_i^{(x,y)}} + \delta_\infty}{\sum_{i=1}^n w_i + 1}\bigg) \Bigg\}$$

Barber, Candes, Ramdas, T. (2022) prove that for any joint law of $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$,

$$\mathbb{P}\Big(Y_{n+1} \in \hat{C}_n^w(X_{n+1})\Big) \geq 1 - \alpha - \sum_{i=1}^n w_i \cdot d_{\mathrm{TV}}(Z, Z^i)$$

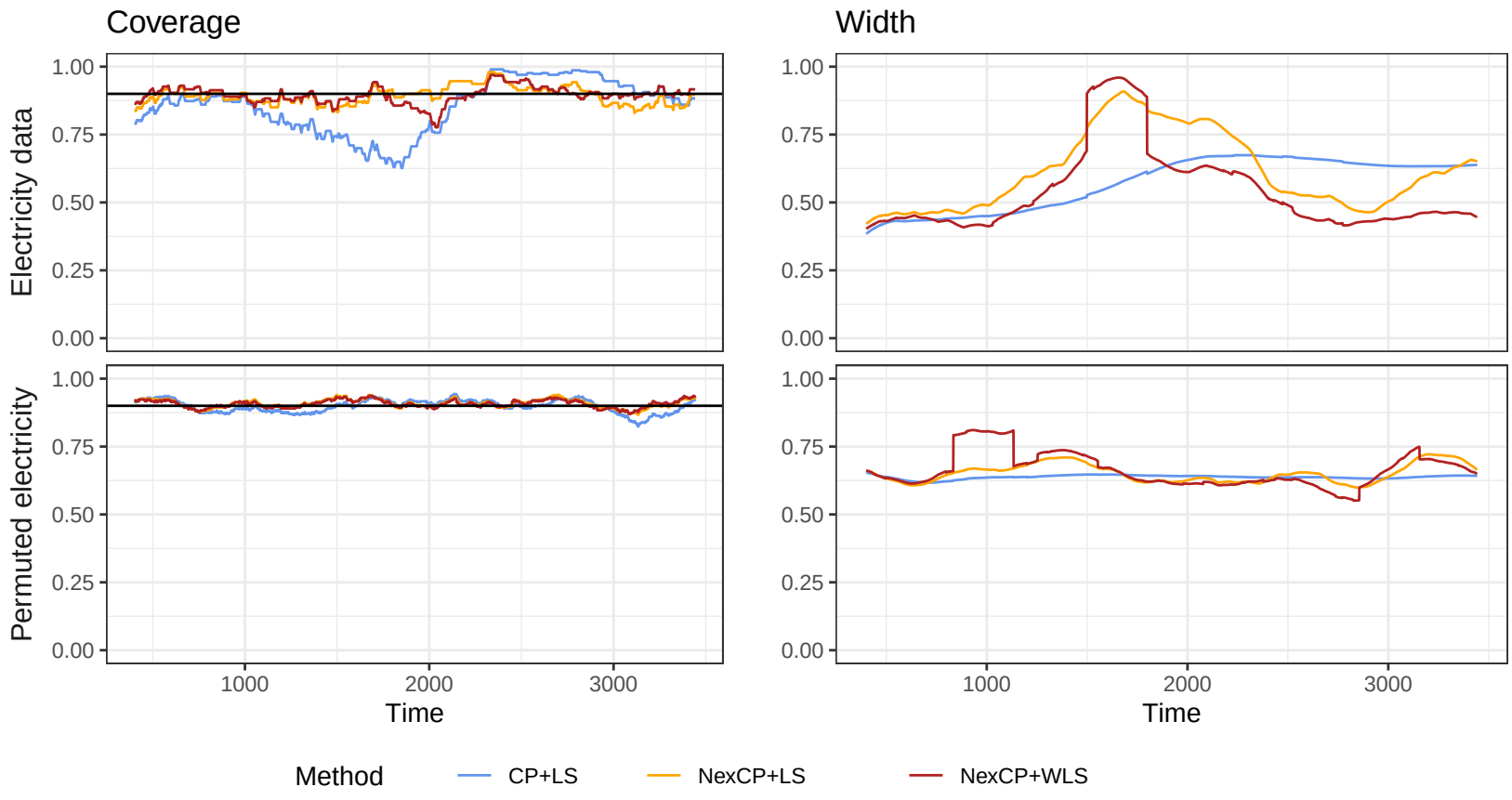
where $Z = (Z_1, \ldots, Z_{n+1})$, and Z^i is the result of swapping $i, n+1$

Some remarks

- In exchangeable case, each $d_{\mathrm{TV}}(Z, Z^i) = 0$, which reveals that arbitrarily-weighted conformal has $1 - \alpha$ coverage
- In independent case, each $d_{\mathrm{TV}}(Z, Z^i) \leq 2d_{\mathrm{TV}}(Z_i, Z_{n+1})$, so if we "guess right" (lower w_i on Z_i farther from Z_{n+1}), then we get good coverage
- In general, the coverage gap is $\leq \sum_{i=1}^n w_i \cdot d_{\mathrm{TV}}(R(Z), R(Z^i))$ where $R(Z)$ is the vector of scores on Z, and same for $R(Z^i)$
- In fact, Barber et al. (2022) even allow the nonconformity score to be nonsymmetric ... uses internal randomization scheme for conformal quantiles

Choosing weights in practice is highly nontrivial, more work should be done; but simple schemes seem to work pretty well in general

NexCP example (Barber. et al. 2022)



Conclusion

Summary

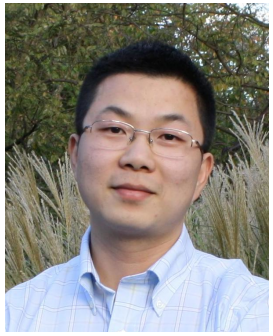
- Conformal inference, pioneered by Vovk (and others) acts as a wrapper on top of an arbitrary prediction algorithm: it maps

  $$\text{predictions} \mapsto \text{prediction sets}$$

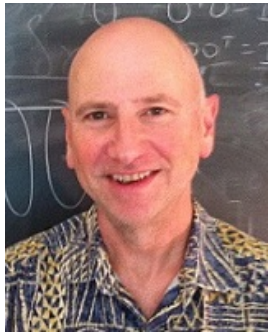
  with finite-sample validity for i.i.d. or exchangeable data
- Split conformal is much faster and simpler, has same guarantee
- The better the base algorithm, the smaller the prediction set
- The sets have average coverage over the feature space; getting conditional coverage is a much harder problem
- Local weighting, quantile scores, other methods can help here
- Many other extensions available, and yet still lots of work to be done, especially beyond exchangeability

Acknowledgments


M. G'Sell


J. Lei


A. Rinaldo


L. Wasserman


R. F. Barber


E. Candes


A. Ramdas

`http://www.stat.cmu.edu/~ryantibs/talks/conformal-2022.pdf`
`https://github.com/ryantibs/conformal/`

Thank you for listening!

Bonus time

Split conformal: more coverage details

Recall in split conformal we get coverage conditional on the proper training set:

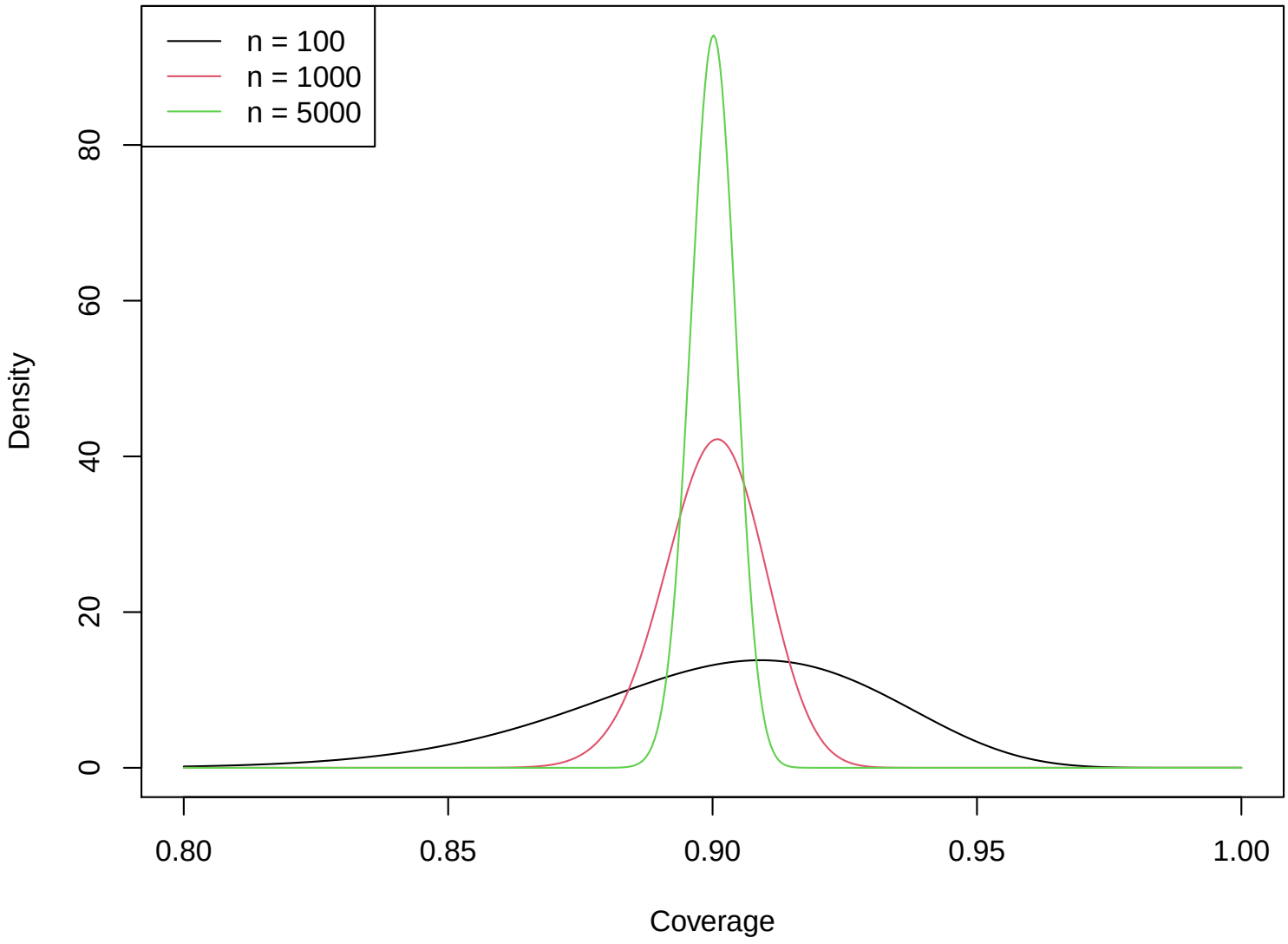
$$\mathbb{P}\Big(Y_{n+1} \in \hat{C}_n(X_{n+1}) \,\Big|\, \{(X_i, Y_i)\}_{i \in D_1}\Big) \in \Big[1-\alpha,\, 1-\alpha+\tfrac{1}{n_2+1}\Big]$$

What about coverage conditional on the entire training set? We get

$$\mathbb{P}\Big(Y_{n+1} \in \hat{C}_n(X_{n+1}) \,\Big|\, \{(X_i, Y_i)\}_{i \in D_1}, \{(X_i, Y_i)\}_{i \in D_2}, \Big) \sim \mathrm{Beta}(k_\alpha, n_2 + 1 - k_\alpha)$$

where $k_\alpha = \lceil (1-\alpha)(n_2+1) \rceil$ (and randomness is over X_{n+1}, Y_{n+1})

Beta coverage illustration



Adaptive conformal inference

In sequential setting, let $\hat{C}_t(\alpha)$ denote level $1-\alpha$ prediction set at time t (formed using data at times $s < t$)

Gibbs and Candes (2021) propose simple iterative scheme to achieve nominal $1-\alpha$ coverage:

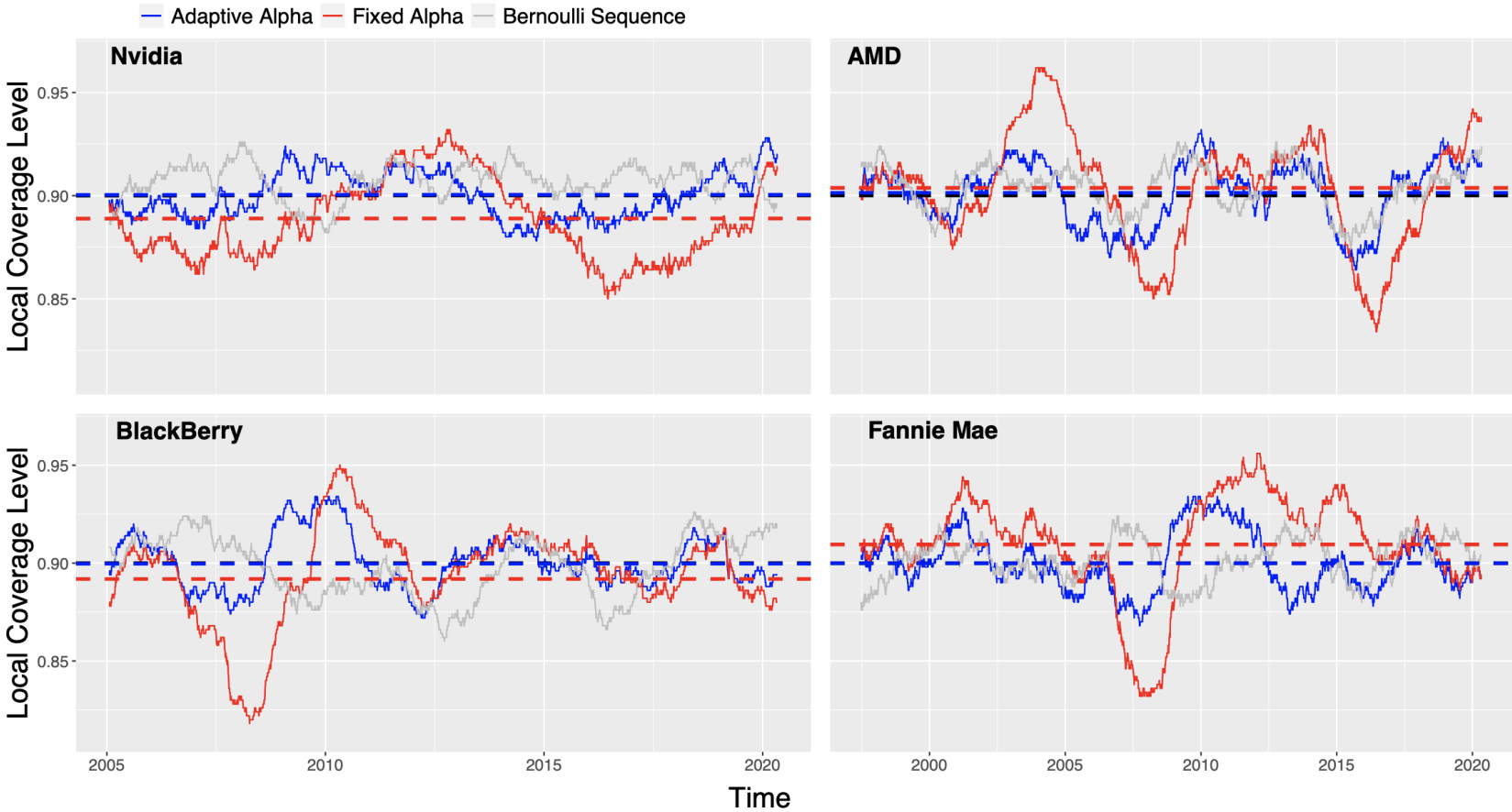
$$\mathrm{err}_t = 1\{Y_t \notin \hat{C}_t(\alpha_t)\}$$
$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t)$$

Called adaptive conformal inference (ACI), it can be seen as online gradient descent on a certain loss function. They show:

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathrm{err_t} = \alpha \quad \text{almost surely}$$

... under essentially no conditions on the family of prediction sets!

ACI example (Gibbs and Candes 2021)



Conformal feature importance

Let:

- $\hat{f}_n$ be an estimate of interest fit on n training points
- $\hat{f}_n^{(-j)}$ be the estimate fit without access to feature j
- $\hat{\Delta}_j(x, y) = |y - \hat{f}_n^{(-j)}(x)| - |y - \hat{f}_n(x)|$, error inflation

(As a concrete example, consider the lasso with λ selected by CV)

Let $\hat{C}_n(x)$ denote prediction band from (split) conformal inference. Define:

$$W_j(x) = \Big\{ |y - \hat{f}_n^{(-j)}(x)| - |y - \hat{f}_n(x)| : y \in \hat{C}_n(x) \Big\}$$

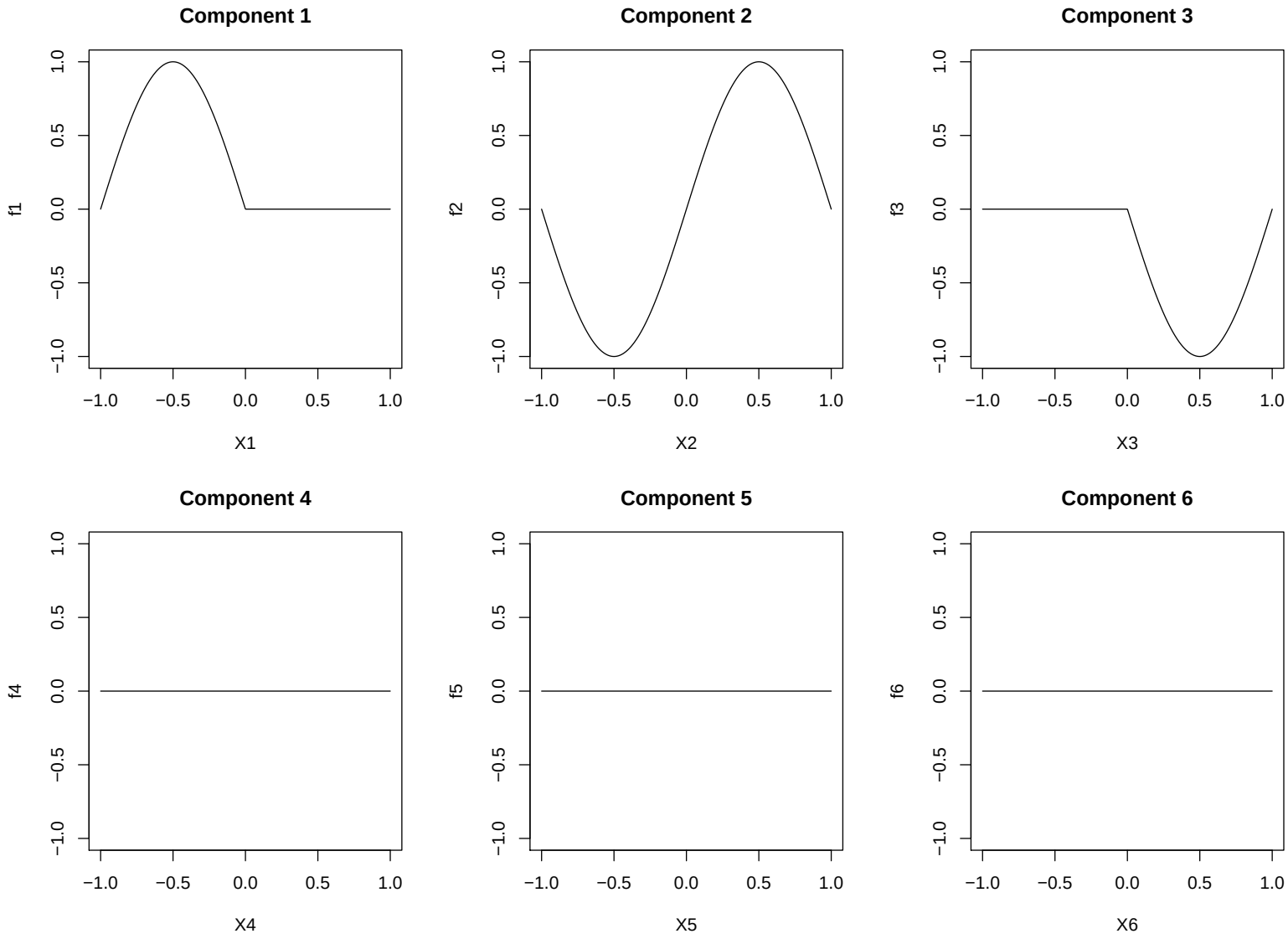
Then:

$$\mathbb{P}\Big(\hat{\Delta}_j(X_{n+1}, Y_{n+1}) \in W_j(X_{n+1}),\ j = 1, \ldots, d \Big) \geq 1 - \alpha$$

Importantly, note the simultaneity over j (hence holds for random j)

Feature importance example (Lei et al. 2018)

Additive model in $d = 6$ dimensions with $f_4 = f_5 = f_6 = 0$



Feature importance example (Lei et al. 2018)

Feature importance intervals $W_j(X_i)$, for $j = 1, \ldots, d$