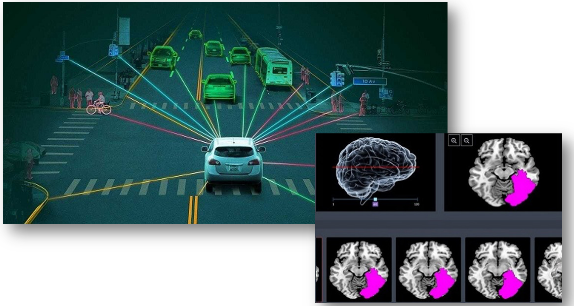
Conformal Prediction in 2022

Emmanuel Candès

Conference on Statistics for complex data, Padova, September 2022

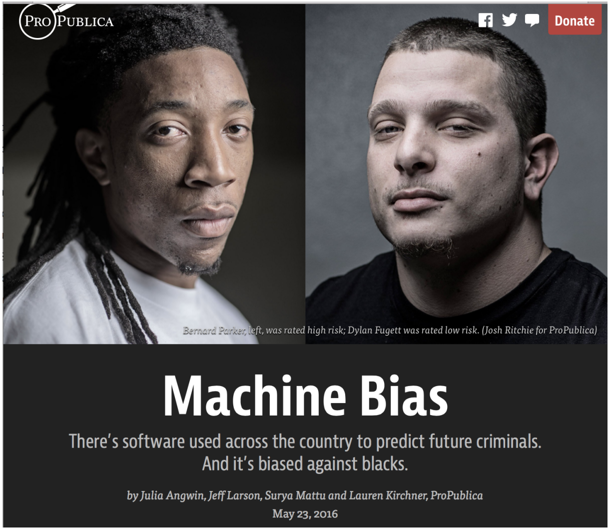
Machine learning in critical applications

- ML tools make potentially critical decisions: self-driving cars, disease diagnosis, ...

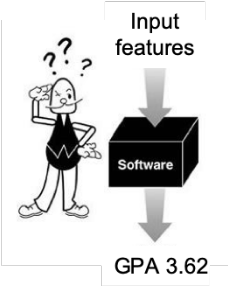


- Involves simultaneous predictions from observations (features), which triggers multiple decisions
- **Can we have confidence in these predictions?**

Growing pains



Data ethics 101: convey uncertainty and reliable outcomes



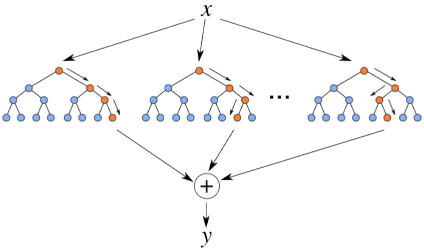
Why don't we see prediction intervals more often?

$$\mathbb{P}\{Y \in C(X)\} \approx 90\%$$

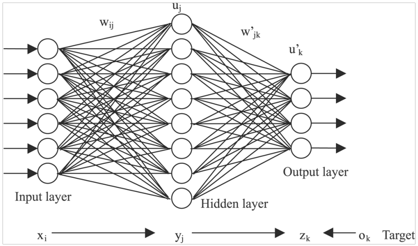
What have we really learned from past data/experience of others?

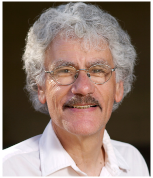
Today's predictive algorithms

random forests, gradient boosting



neural networks



Breiman and Friedman


LeCun, Hinton, Bengio, and Rumelhart

Prediction intervals

Training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and test point $(X_{n+1}, ?)$
(assumed exchangeable, e.g. i.i.d. from P_{XY})

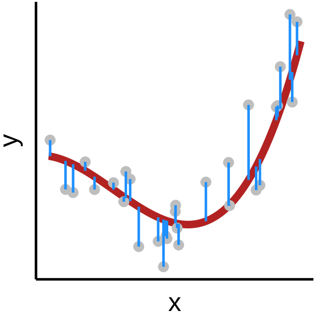
Goal: construct **marginal distribution free prediction interval**

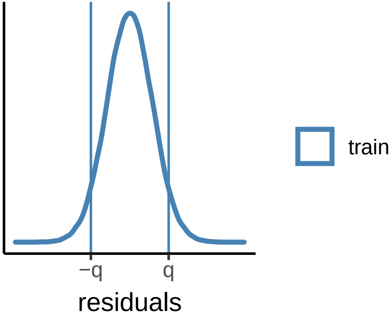
$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha$$

- Any dist. P_{XY} (assumed unkown)
- Any sample size n

"Based on the candidate's high school identifier and GPA, SAT scores, and other attributes, the college GPA is predicted to fall in the [3.4,3.8] range"

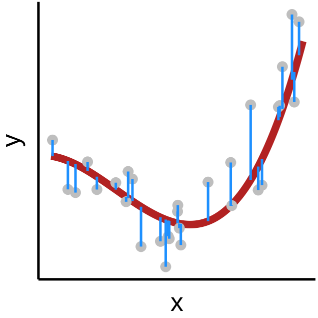
Predicting with confidence?

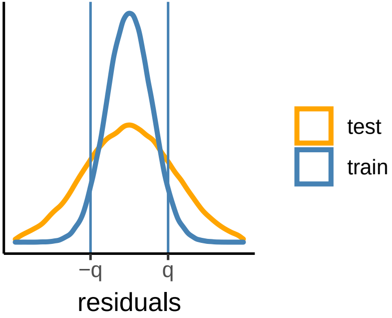




Naive approach: look at residuals and build predictive set $[\hat{\mu}(x) - q, \hat{\mu}(x) + q]$

Predicting with confidence?





Naive approach: look at residuals and build predictive set $[\hat{\mu}(x) - q, \hat{\mu}(x) + q]$

Doesn't work! residuals much smaller than on test points (extreme for neural nets)

(Jackknife is better, but still fails)

Enter conformal prediction: some pioneers

Predictive inference is possible under no assumptions!

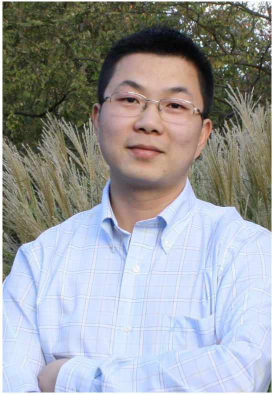
Vladimir Vovk

Glenn Shafer

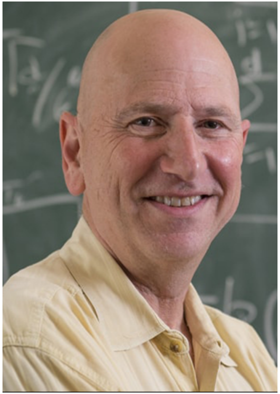
Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Papadopoulos, Proedrou, Vovk, Gammerman 2002, *Inductive Confidence Machines for Regression*

Some evangelists



Jing Lei



Larry Wasserman

Lei, Wasserman 2014, *Distribution-free prediction bands for non-parametric regression*

Lei, G'Sell, Rinaldo, Tibshirani, Wasserman 2018, *Distribution-free predictive inference for regression*

Some collaborators

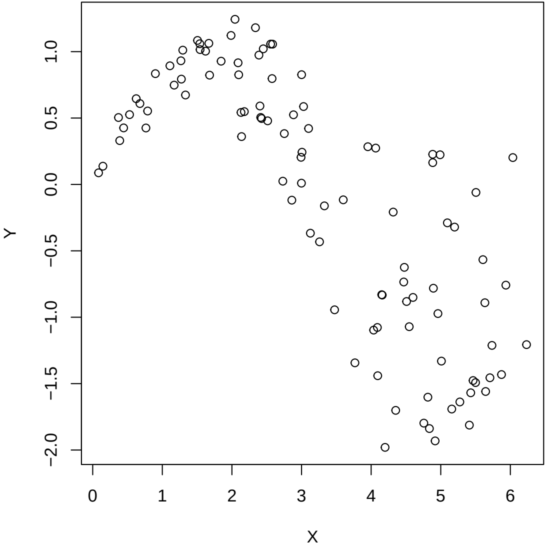
Rina Barber

Aaditya Ramdas

Ryan Tibshirani

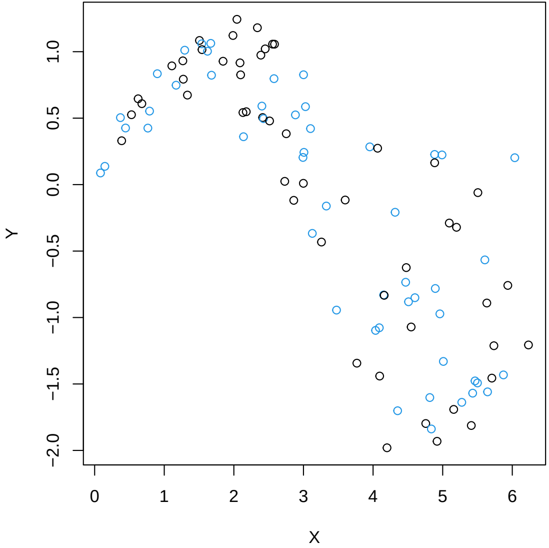
Split conformal prediction

Main idea: look at holdout residuals



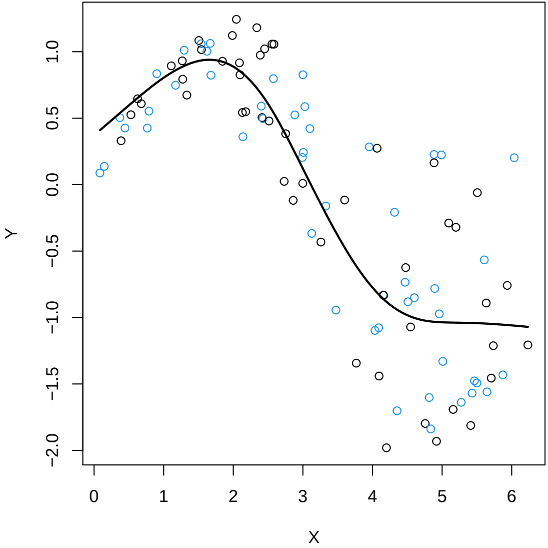
Split conformal prediction

Main idea: look at holdout residuals



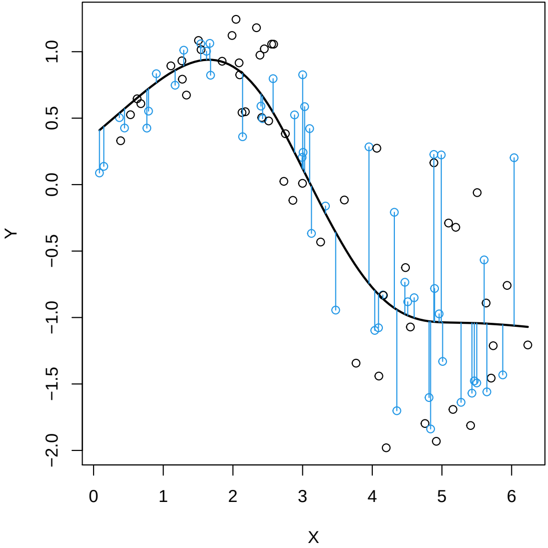
Split conformal prediction

Main idea: look at holdout residuals



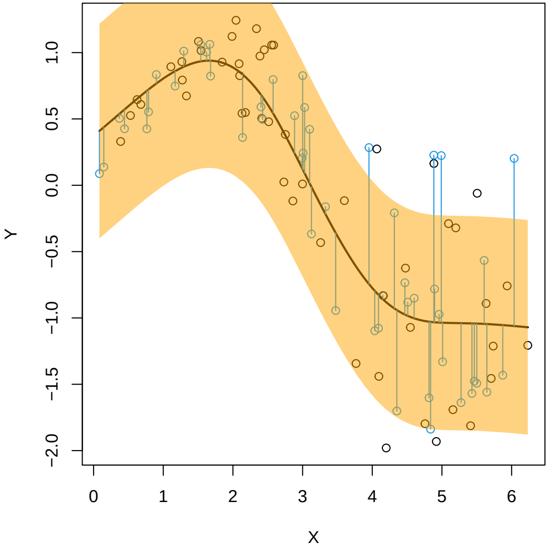
Split conformal prediction

Main idea: look at holdout residuals

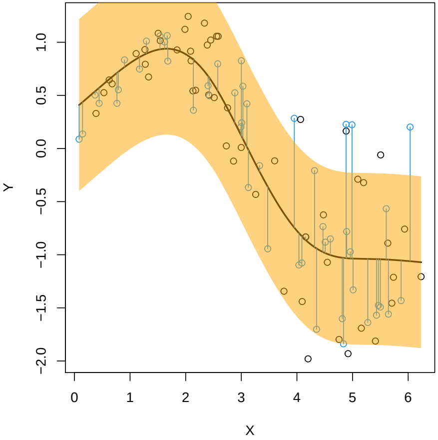


Split conformal prediction

Main idea: look at holdout residuals



Split conformal prediction



About 90% of future test points will fall within this band

q is 90th percentile of absolute residuals on calibration set (not used for model fitting)

$$\mathbb{P}\left\{Y_{n+1} \in [\hat{\mu}(X_{n+1}) - q, \hat{\mu}(X_{n+1}) + q]\right\} \geq 90\%$$

Papadopoulos, Proedrou, Vovk, Gammerman '02

Beyond residuals

- Just used $s(x, y) = |y - \hat{\mu}(x)|$
- Predictive set: $C(x) = \{y : s(x, y) \leq q\}$
- Why stop here? Can use *any conformity score* $s(x, y)$

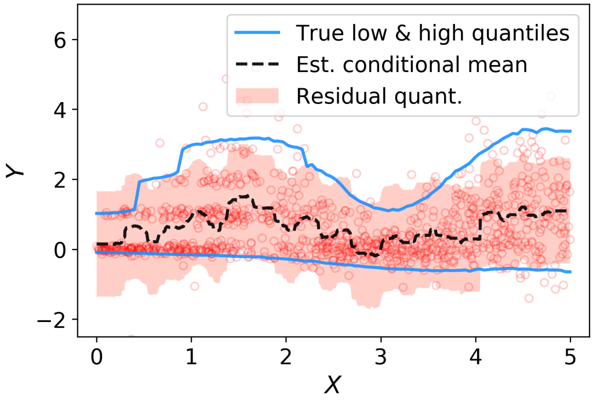
Beyond residuals

- Just used $s(x, y) = |y - \hat{\mu}(x)|$
- Predictive set: $C(x) = \{y : s(x, y) \leq q\}$
- Why stop here? Can use *any conformity score* $s(x, y)$

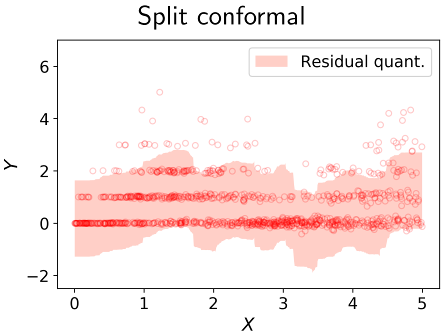
q is quantile of $s(X_i, Y_i)$ on calibration set. Then

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

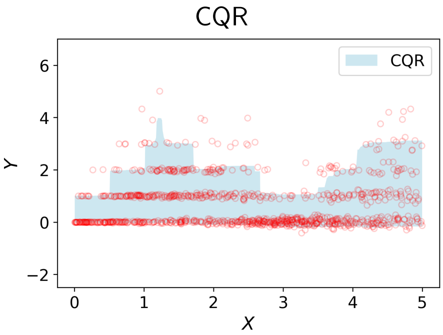
Fixed vs. adaptive intervals



Conformalized quantile regression (CQR) with random forests regression



Avg. Coverage 91.4%
Avg. Length 2.91



Avg. Coverage 91.0%
Avg. Length 2.18

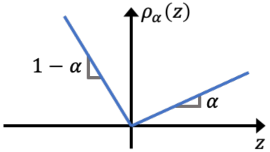
CQR is adaptive while split conformal is not

Conformalized quantile regression[1]

- Quantile regression

$$f(\cdot) = \underset{f \in \mathcal{F}}{\text{argmin}} \sum_i \rho_\alpha(Y_i - f(X_i)) + \mathcal{R}(f)$$

  - $\mathcal{R}(f)$ is a possible regularizer
  - ρ_α is pinball loss Koenker & Bassett 1978

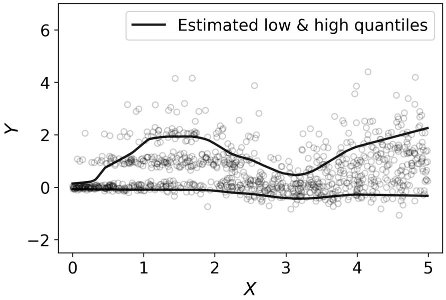


- Define conformity scores

$$S(x, y) = \max\{\hat{q}_{\alpha/2}(x) - y, y - \hat{q}_{1-\alpha/2}(x)\}$$

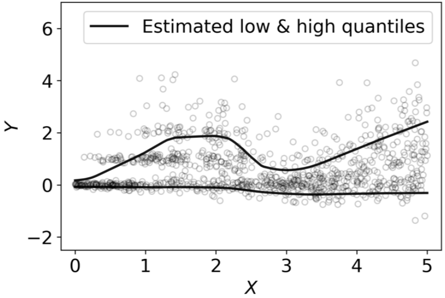
(many variations)

- Include y in predictive interval iff $S(X_{n+1}, y) \leq \text{quantile}\{S(X_i, Y_i)\}$

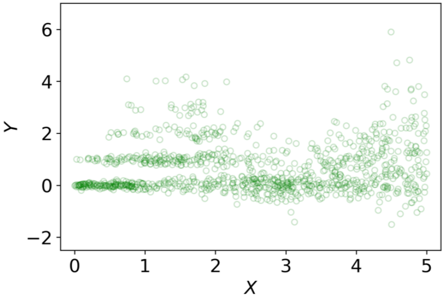


[1] Romano, Sesia & C. 2019, *Conformalized quantile regression*

Fit

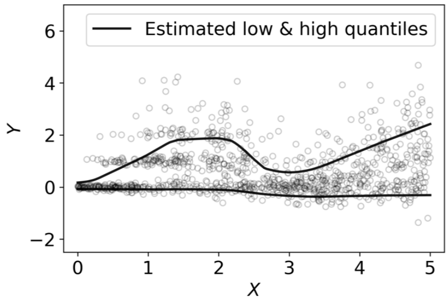


Apply quantile regression

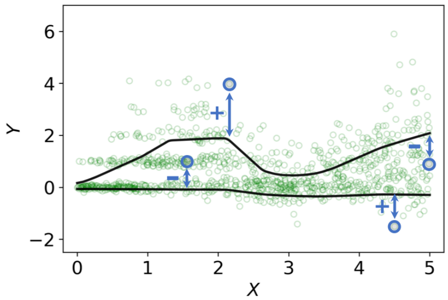


Calibration set

Calibration



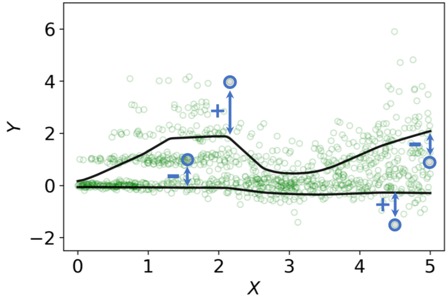
Apply quantile regression



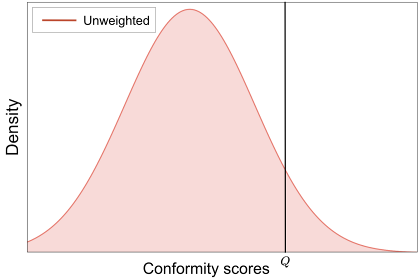
Calibrate

Conformity scores

conformity scores are signed distances: $S_i \triangleq \max\{\mathrm{lo}(X_i) - Y_i, Y_i - \mathrm{hi}(X_i)\}$



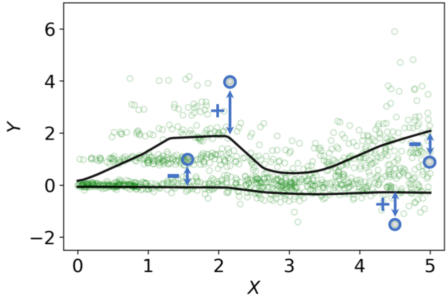
Calibrate



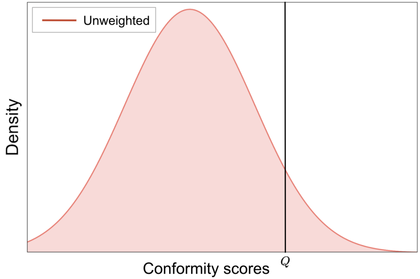
Histogram

Calibration

$$C(x) = [\mathrm{lo}(x) - Q,\ \mathrm{hi}(x) + Q]$$



Calibrate



Histogram

Predicting utilization of medical services

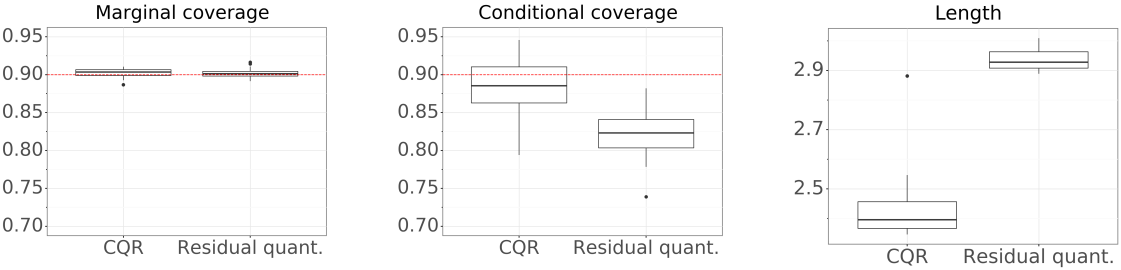
Medical Expenditure Panel Survey 2015

- X_i – age, marital status, race, poverty status, functional limitations, health status, health insurance type, ...
- Y_i – health care system utilization, reflecting # visits to doctor's office/hospital, ...
- $\approx 16,000$ subjects
- ≈ 140 features

AHRQ Agency for Healthcare Research and Quality
Advancing Excellence in Health Care

Results on MEPS data

- NNet regression (MSE or pinball loss)
- Average across 20 random train-test (80%/20%) splits



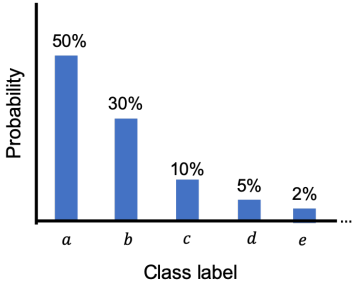
Better conditional coverage* and shorter intervals

*measured over the worst slab Cauchois, Gupta & Duchi 2020

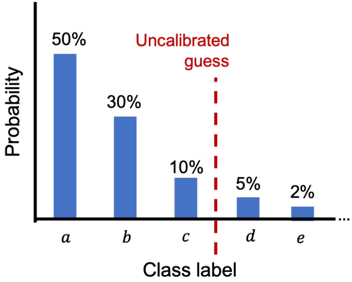
Discrete labels Romano, Sesia & C. 2020

- Estimate conditional probabilities $\hat{\pi}(y \mid x)$
  $\rightsquigarrow$ e.g., output of NNet's softmax layer

- Uncalibrated guess

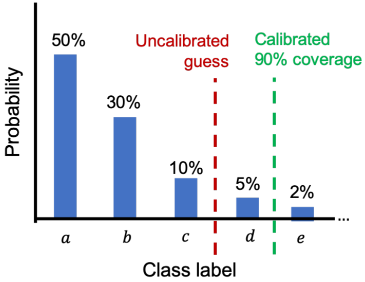


Sorted class probabilities

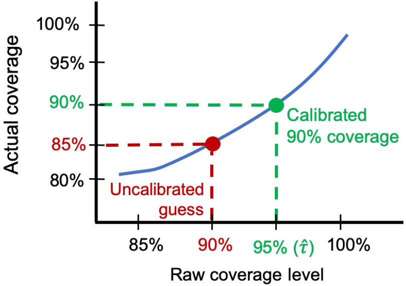


$C^{\text{naive}}(x, 90\%) = \{a, b, c\}$

Calibration via adaptive coverage



$$C^{\text{naive}}(x, 95\%) = \{a, b, c, d\}$$



Prediction set

$$C(x) = C^{\text{naive}}(x, \hat{\tau})$$

"Choose 95% nominal to get 90% coverage on test data"

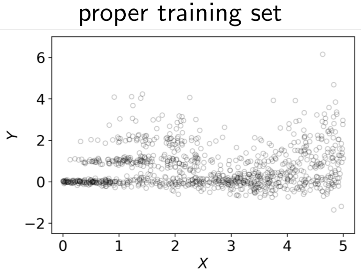
Examples

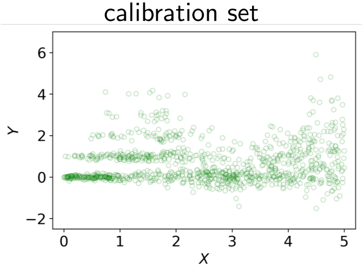
{ fox squirrel 0.99 }

{ 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 }

Partial summary

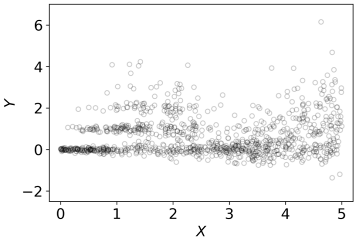




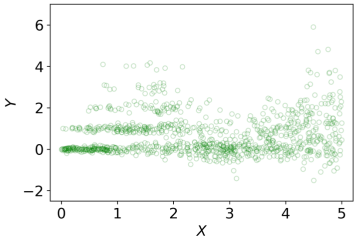
- *Training:* use $n/2$ data points to learn model $S(x, y)$
- *Validation:* use $n/2$ data points to learn distrib. of $S(X, Y)$
- *Calibrated prediction*: we can predict $S(X_{n+1}, Y_{n+1}) \rightsquigarrow$ can predict Y_{n+1}

Partial summary

proper training set



calibration set



Drawback: sample splitting $\rightsquigarrow$ only use half the data points to fit the model

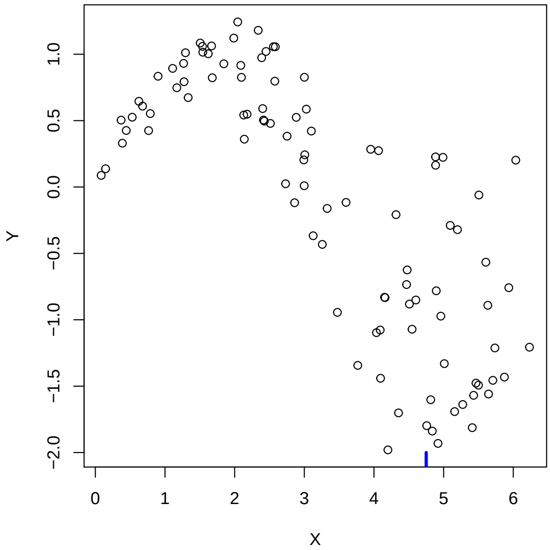
Full conformal prediction: use all data points for training & validation

Gammerman, Vovk, Vapnik, '98, Vovk, Gammerman, Shafer '05

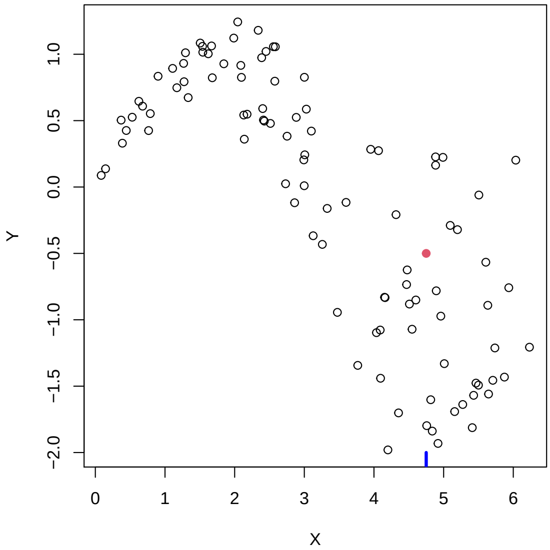
Jackknife+/CV+

Barber, C., Ramdas and Tibshirani '19

Full conformal: an example

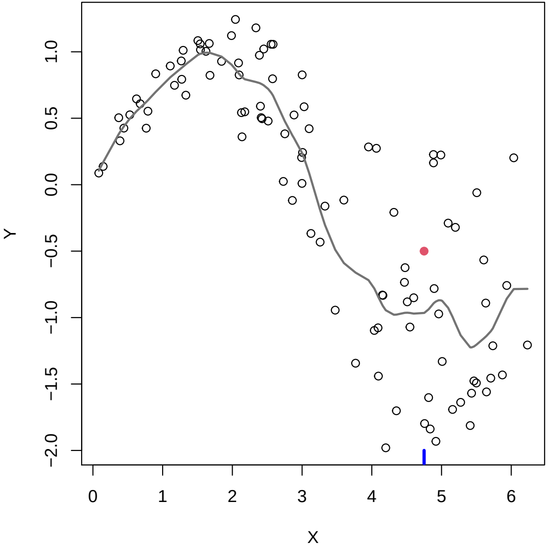


Full conformal: an example



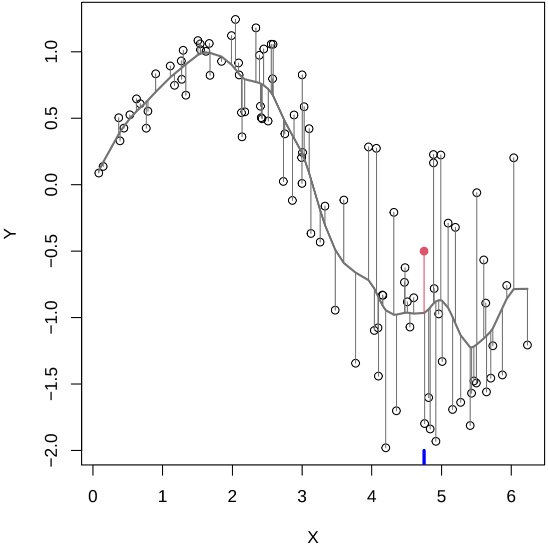
Iterate through trial y, compute p-value

Full conformal: an example



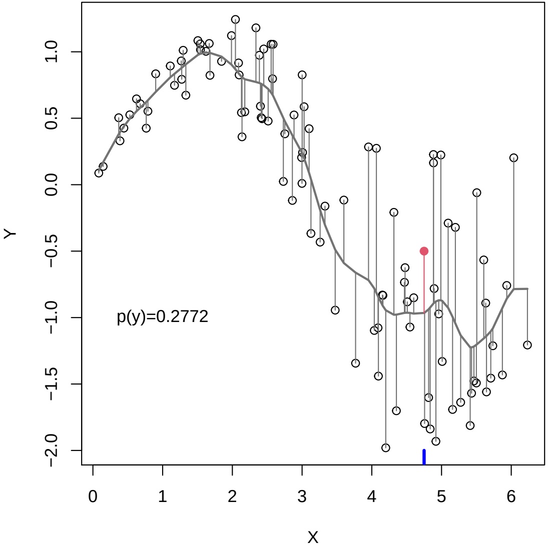
Iterate through trial y, compute p-value

Full conformal: an example



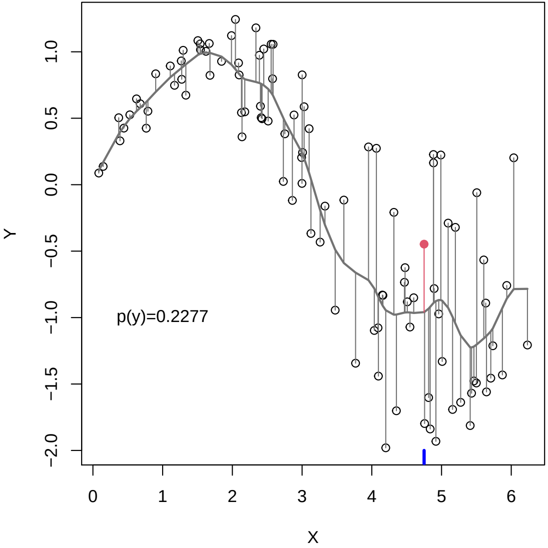
Iterate through trial y, compute p-value

Full conformal: an example



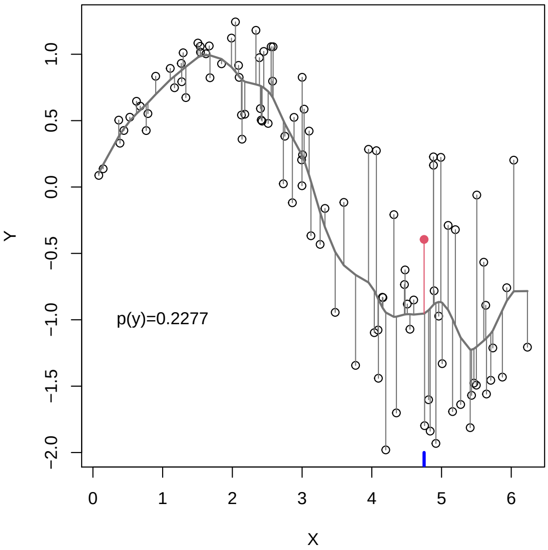
Iterate through trial y, compute p-value

Full conformal: an example



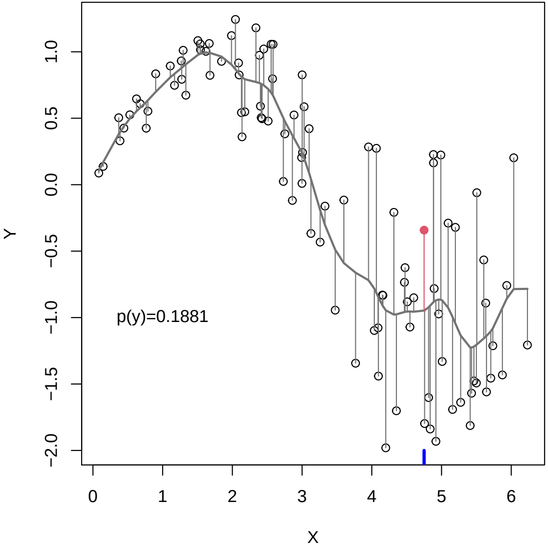
Iterate through trial y, compute p-value

Full conformal: an example



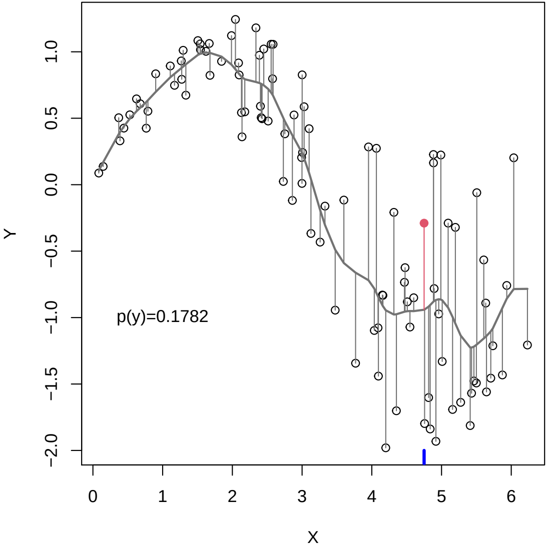
Iterate through trial y, compute p-value

Full conformal: an example



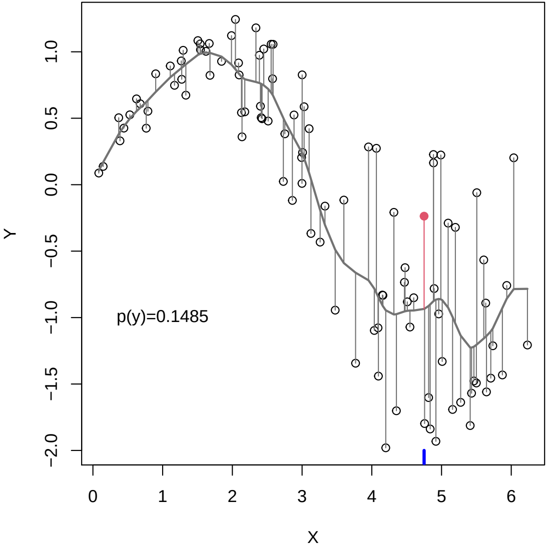
Iterate through trial y, compute p-value

Full conformal: an example



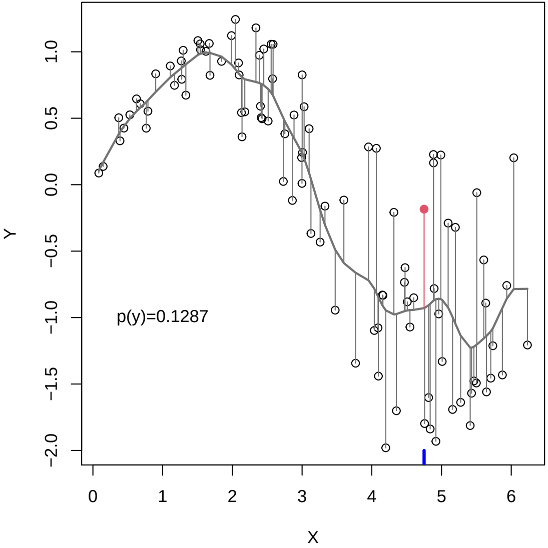
Iterate through trial y, compute p-value

Full conformal: an example



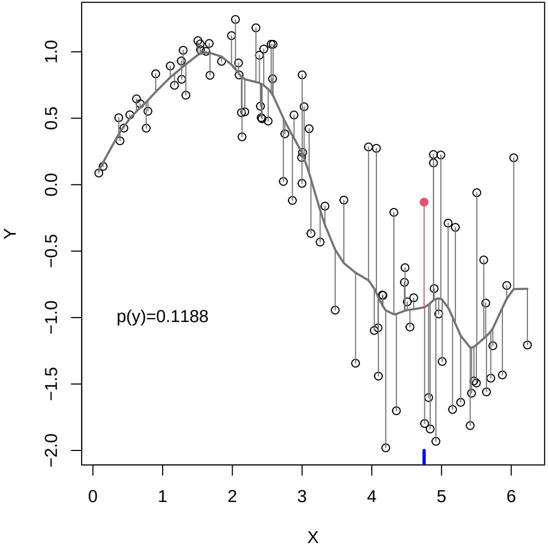
Iterate through trial y, compute p-value

Full conformal: an example



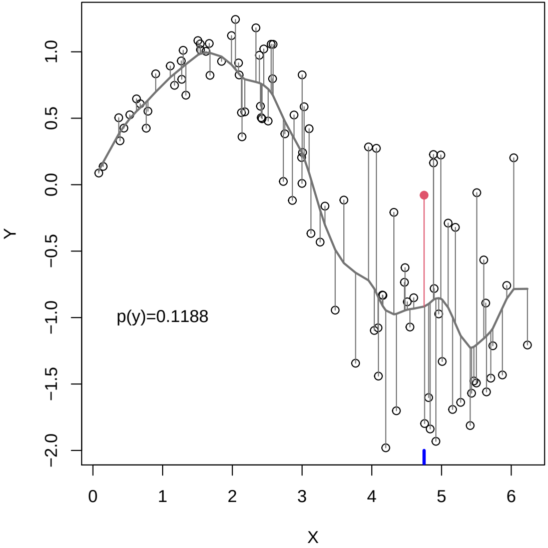
Iterate through trial y, compute p-value

Full conformal: an example



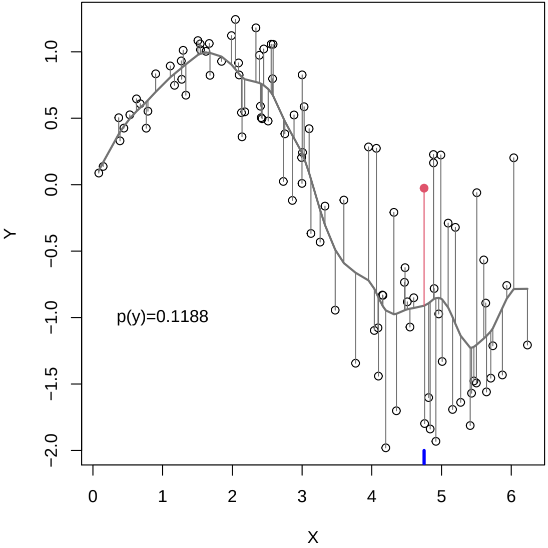
Iterate through trial y, compute p-value

Full conformal: an example



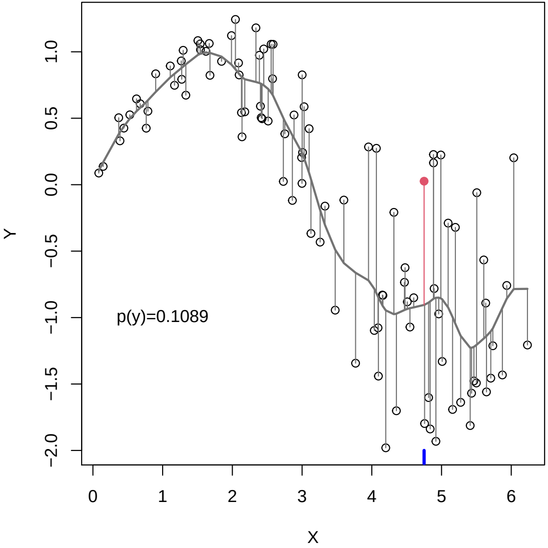
Iterate through trial y, compute p-value

Full conformal: an example



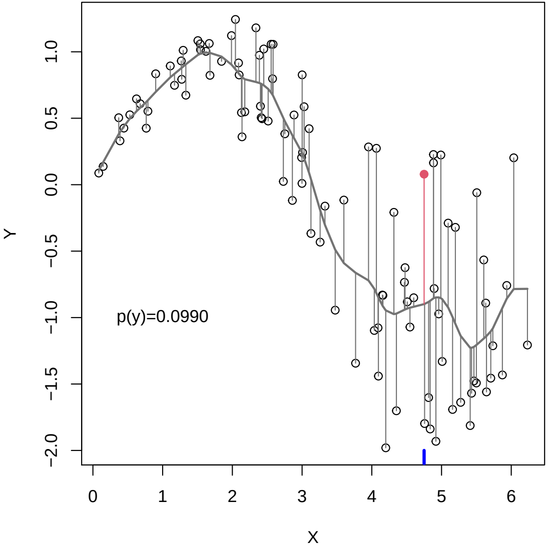
Iterate through trial y, compute p-value

Full conformal: an example



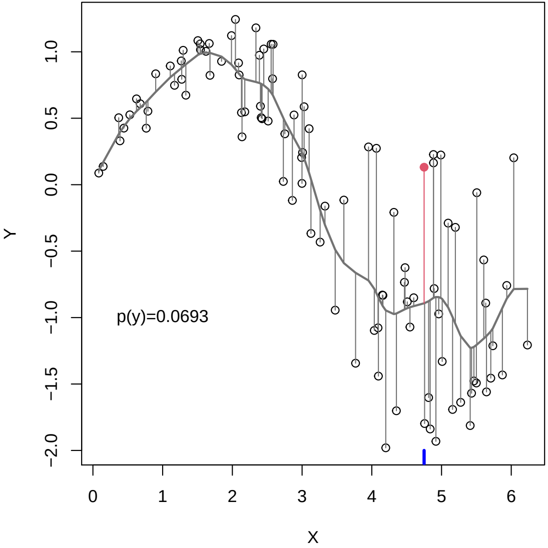
Iterate through trial y, compute p-value

Full conformal: an example



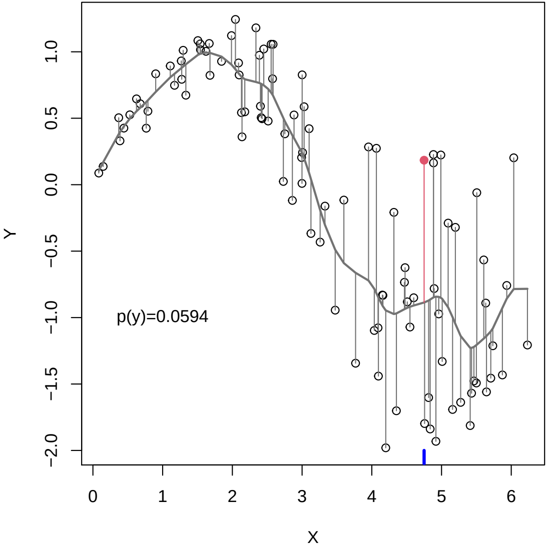
Iterate through trial y, compute p-value

Full conformal: an example



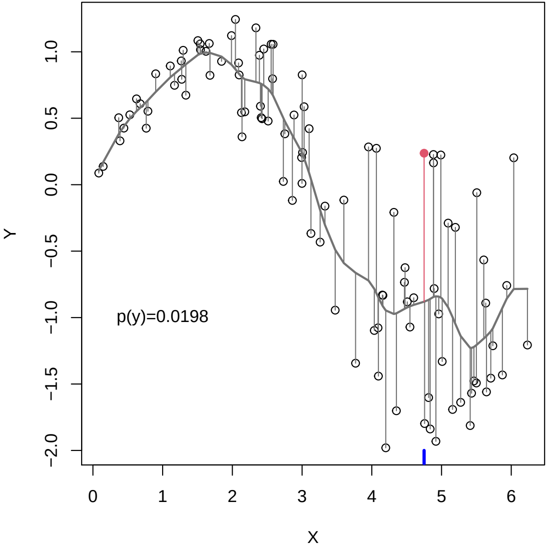
Iterate through trial y, compute p-value

Full conformal: an example



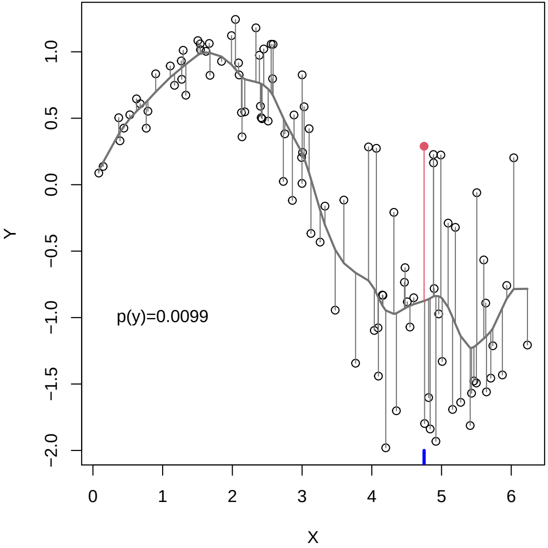
Iterate through trial y, compute p-value

Full conformal: an example



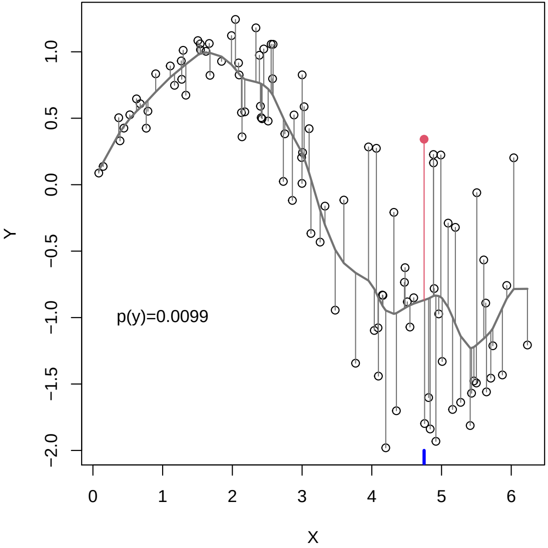
Iterate through trial y, compute p-value

Full conformal: an example



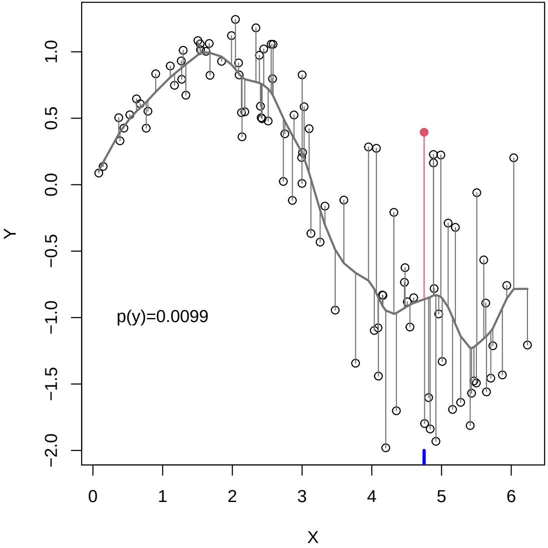
Iterate through trial y, compute p-value

Full conformal: an example



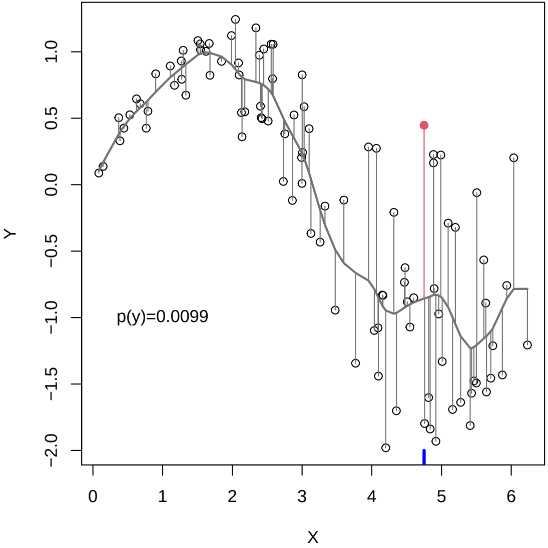
Iterate through trial y, compute p-value

Full conformal: an example



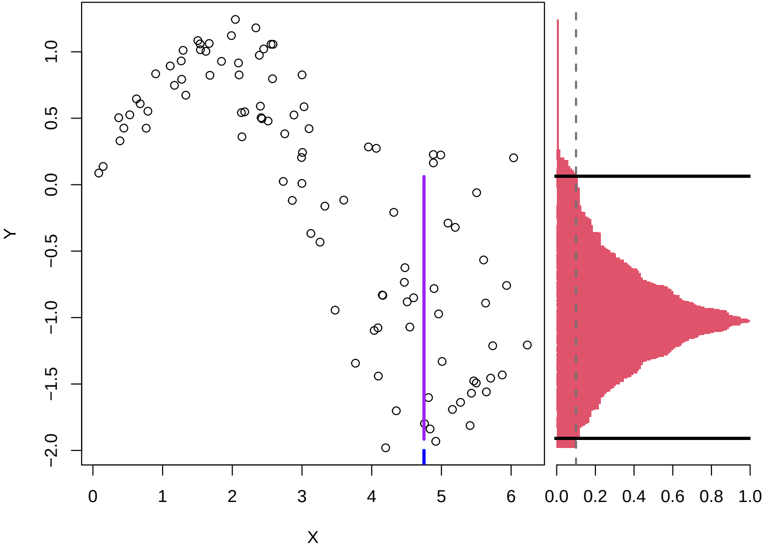
Iterate through trial y, compute p-value

Full conformal: an example



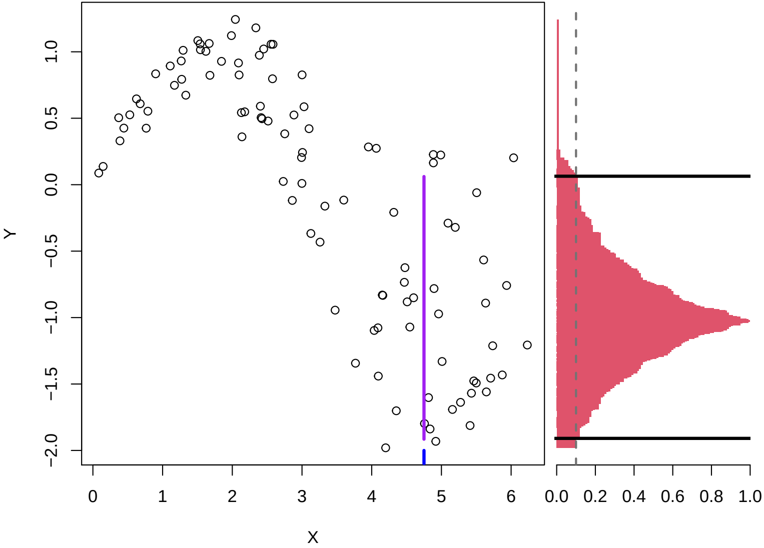
Iterate through trial y, compute p-value

Full conformal: an example



Threshold p-values to get full conformal interval

Full conformal: an example



Drawback: computationally expensive $\rightsquigarrow$ Jackknife+/CV+

Barber, C., Ramdas and Tibshirani '19

(Full) conformal

- Observe training + test data:

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$$

- Fit model $\widehat{\mu}$ to all $n+1$ data points via symmetric algorithm & get residuals

$$R_i = Y_i - \widehat{\mu}(X_i), i = 1, \ldots, n, \quad R_{n+1} = Y_{n+1} - \widehat{\mu}(X_{n+1})$$

- Check if $|R_{n+1}| \leq [(1-\alpha)$ quantile of $|R_1|, \ldots, |R_n|, |R_{n+1}|]$

By exchangeability of $R_1, \ldots, R_{n+1}$
this occurs with prob. $\geq 1 - \alpha$

(Full) conformal

- Assume we observe training + test data:

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$

- Fit model $\widehat{\mu}$ to all $n+1$ data points via symmetric algorithm & get residuals

$$R_i = Y_i - \widehat{\mu}(X_i), i = 1, \ldots, n, \quad R_{n+1} = y - \widehat{\mu}(X_{n+1})$$

- Check if $|R_{n+1}| \leq [(1-\alpha)$ quantile of $|R_1|, \ldots, |R_n|, |R_{n+1}|]$

↖
By exchangeability of $R_1, \ldots, R_{n+1}$
this occurs with prob. $\geq 1 - \alpha$ if we plug $y = Y_{n+1}$

(Full) conformal prediction

- Propose test value $y \in \mathbb{R}$

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$

- Fit model $\widehat{\mu}$ to all $n+1$ data points via symmetric algorithm & get residuals

$$R_i = Y_i - \widehat{\mu}(X_i), i = 1, \ldots, n, \quad R_{n+1} = y - \widehat{\mu}(X_{n+1})$$

- Check if $|R_{n+1}| \leq [(1-\alpha)$ quantile of $|R_1|, \ldots, |R_n|, |R_{n+1}|]$
- $y \overset{\widehat{\mu},\alpha}{\rightsquigarrow} \{$ Yes, No $\}$
- Include y in $\widehat{C}(X_{n+1})$ iff answer is Yes (iff it conforms)

Theorem

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} = \mathbb{P}\{\text{for test value } y = Y_{n+1}, \text{answer is Yes}\} \geq 1 - \alpha$$

(Full) conformal prediction

- Propose test value $y \in \mathbb{R}$

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$$

- Fit model $\widehat{\mu}$ to all $n+1$ data points via symmetric algorithm & get residuals

$$R_i = Y_i - \widehat{\mu}(X_i), i = 1, \ldots, n, \quad R_{n+1} = y - \widehat{\mu}(X_{n+1})$$

- Check if $|R_{n+1}| \leq [(1-\alpha)$ quantile of $|R_1|, \ldots, |R_n|, |R_{n+1}|]$
- $y \overset{\widehat{\mu},\alpha}{\rightsquigarrow} \{$ Yes, No $\}$
- Include y in $\widehat{C}(X_{n+1})$ iff answer is Yes (iff it conforms)

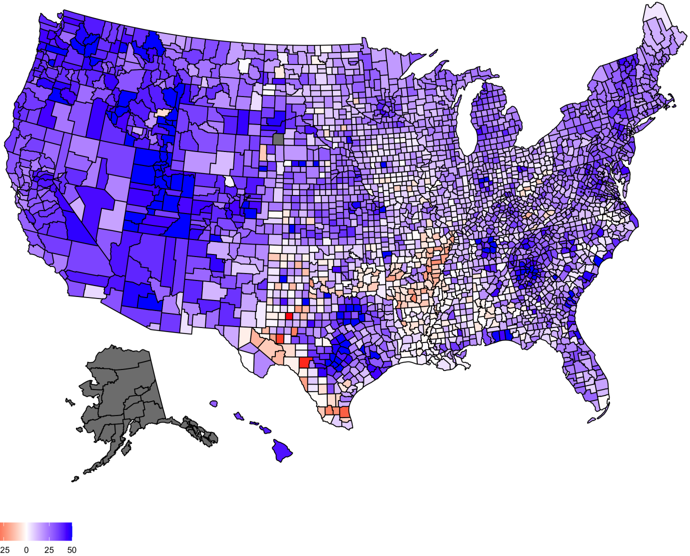
Theorem

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} = \mathbb{P}\{\text{for test value } y = Y_{n+1}, \text{answer is Yes}\} \geq 1 - \alpha$$

Extends to arbitrary conformity scores computed in a symmetric fashion

Forecasting 2020 US Presidential Election Results County by County

2020 US Presidential Election results county by county



Problem statement

Data (X_i, Y_i), for each reporting county i

- X_i county features (demographic, socio-economic, ... variables)
- Interested in normalized vote change Y_i:

  $\#$ Republican or Democratic votes $R_i^{(20)}$ or $D_i^{(20)}$ $\qquad Y_i = (R_i^{(20)} - R_i^{(16)})/R_i^{(16)}$

- Use reported counties to forecast unreported counties

Interlude: Election Night at *The Washington Post*

Variation on *weighted conformalized quantile regression* used by WP as forecast

John Cherian

Lenny Broner

Pennsylvania

20 ELECTORAL VOTES

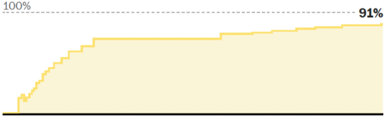
LIVE: Donald Trump (R) is leading. An estimated 91 percent of votes have been counted.

■ Biden
48.1%
3,051,555

■ Trump
50.7%
3,215,969

How much of the vote has been counted in Pennsylvania?

The Post estimates **91%** of votes cast have been counted here.



U.S. House District 10

Perry	54.9%	
DePasquale	45.1%	

An estimated 88% of votes have been counted

U.S. House District 17

Lamb	50.5%	
Parnell	49.5%	

An estimated 92% of votes have been counted

Pennsylvania has 18 U.S. House races. Jump to results

Note: Map colors on this page won't indicate a lead for a candidate until an estimated 35 percent of the vote has been reported there. Results updated at 3:30 a.m. ET

The Washington Post 5 November 2020, 12:50 AM

Pennsylvania

20 ELECTORAL VOTES

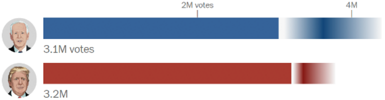
LIVE: Donald Trump (R) is leading. An estimated 91 percent of votes have been counted.

Where the vote could end up

These estimates are calculated based on past election returns as well as votes counted in the presidential race so far. View details

We estimate that 91 percent of the total votes cast have been counted. We're estimating ranges of possible outcomes, and these are the most likely ones.

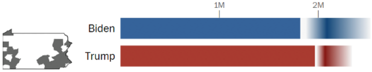
Counted votes — Estimates of final vote tally
Lighter colors are less likely outcomes



Breaking down the estimates

Urban counties

Suburban counties



Rural counties

The Washington Post 5 November 2020, 12:50 AM

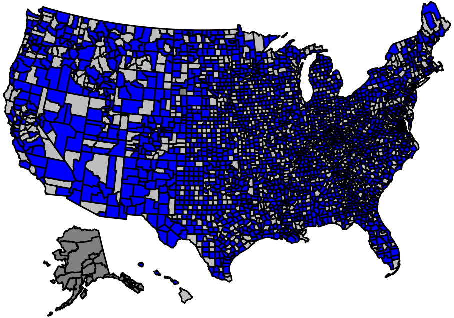
Problem setup

- Randomly split $n = 3076$ counties into training $|\mathcal{D}_{\text{train}}| = 1200$ and test $|\mathcal{D}_{\text{test}}| = 1876$ samples
  $\rightsquigarrow$ exchangeability and $\therefore$ theorem hold
- For each test sample $j \in \mathcal{D}_{\text{test}}$, run the *full conformal procedure* with $\mathcal{D}_{\text{train}} \cup \{j\}$ to predict Y_j
- Coverage target $\alpha = 0.1$. Nonconformity scores

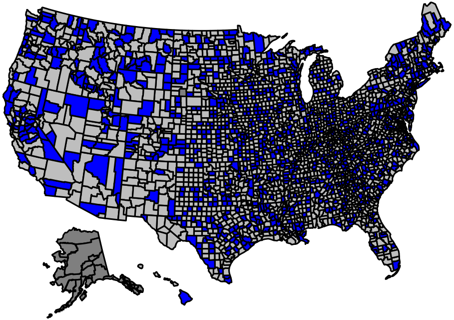
  QR: $S(x, y) = \max\{\hat{q}_{1-\alpha/2}(x) - y, y - \hat{q}_{\alpha/2}(x)\}$ for fitted β-conditional quantiles $\hat{q}_\beta(x)$

  LM: $S(x, y) = |y - \hat{\mu}(x)|$ for linear OLS prediction $\hat{\mu}(x) = \hat{\theta}^\top x$

Drawing counties



Training



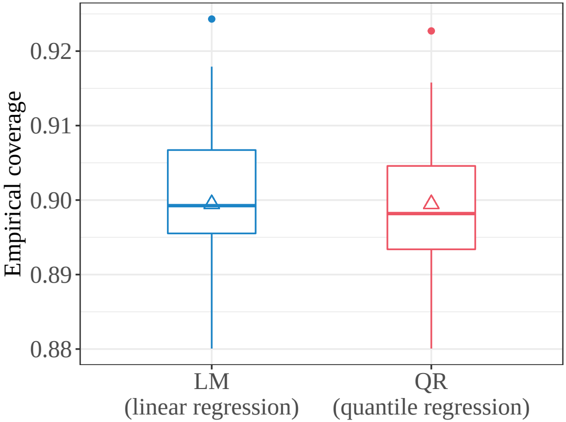
Testing

Coverage on test samples

- 1st run: QR 0.8982, LM 0.8955
- 2nd run: QR 0.8945, LM 0.9019
- 3rd run: QR 0.8827, LM 0.8992

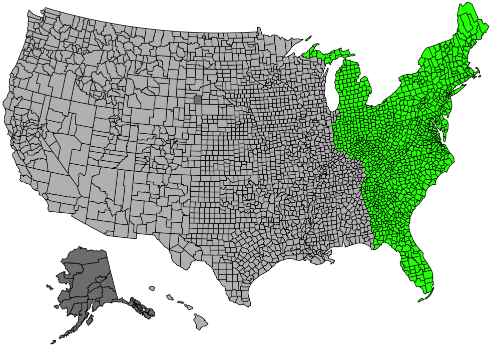
Coverage over $N = 25$ independent runs

- Empirical coverage on $\mathcal{D}_{\text{test}}$ over $N = 25$ independent runs (Δ represents average across runs)



Is my data exchangeable?

Are eastern counties representative of other counties?



Beyond exchangeability: what if ...?

- Want to deploy model in a new environment? e.g. a diagnostic model trained in America on French patients

  Cauchois et. al. '20, Tibshirani, Barber, C. and Ramdas '19

- Environment is dynamic? e.g. stock market behaviour may shift in response to world events

  Gibbs and C. '21 & '22

  Barber, C. Ramdas and Tibshirani '22

Adaptive conformal inference

Isaac Gibbs

Online methods?

- Observe data stream $\{(X_t, Y_t)\}_{t=0,1,\ldots}$
- Perhaps $(X_t, Y_t) \sim P_t$ with P_t varying across time
- At time t, want to use past data along with X_t to form a prediction set $\hat{C}_t$ for Y_t

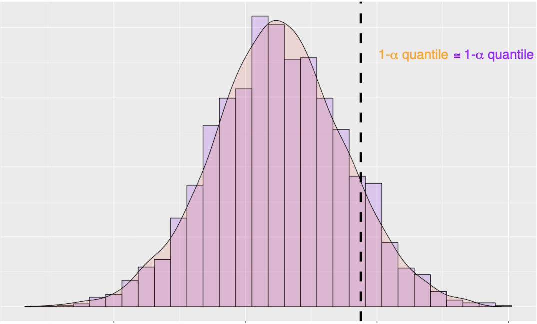
Goals

- **Minimum:** guarantee that $Y_t \in \hat{C}_t$ at least a $1-\alpha$ fraction of the time
- **Ambitious:** guarantee that $\mathbb{P}(Y_t \in \hat{C}_t) \cong 1-\alpha$ for all t

Adapting conformal to distribution shift

$$\hat{C}_t(\alpha) := \left\{y : S_t(X_t, y) \leq \text{Quantile}\left(1-\alpha, \{S_t(X_\ell, Y_\ell)\}_{(X_\ell, Y_\ell) \in \mathcal{D}_{\text{cal},t}}\right)\right\}$$

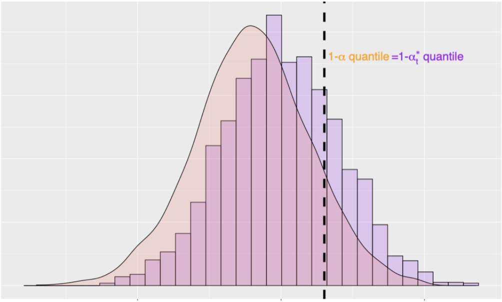
Under the i.i.d. assumption the empirical and true distributions will approximately align

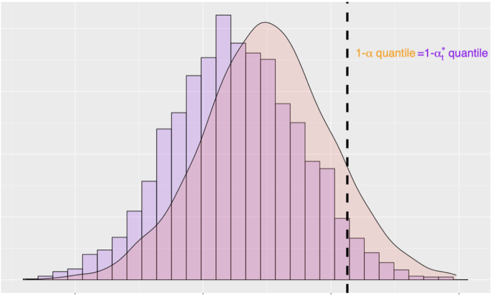


Adapting conformal to distribution shift

$$\hat{C}_t(\alpha) := \left\{ y : S_t(X_t, y) \leq \text{Quantile}\left(1-\alpha, \{S_t(X_\ell, Y_\ell)\}_{(X_\ell, Y_\ell)\in\mathcal{D}_{\text{cal},t}}\right)\right\}$$

Distribution shift can cause the true distribution to shift to the right or left

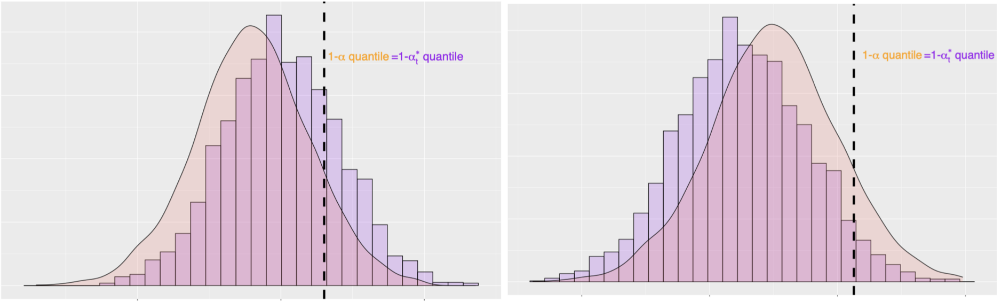




Adapting conformal to distribution shift

$$\hat{C}_t(\alpha) := \left\{y : S_t(X_t, y) \leq \text{Quantile}\left(1-\alpha, \{S_t(X_\ell, Y_\ell)\}_{(X_\ell, Y_\ell) \in \mathcal{D}_{\text{cal},t}}\right)\right\}$$

Distribution shift can cause the true distribution to shift to the right or left



Key Idea: Learn α_t^*

Learning α_t^*

Fit α_t using online update

$$\alpha_{t+1} := \alpha_t + \gamma(\alpha - \mathrm{err}_t)$$

err_t acts as an unbiased estimate of the current miscoverage probability

$$\mathrm{err}_t := \begin{cases} 1 & Y_t \notin \hat{C}_t \\ 0 & Y_t \in \hat{C}_t \end{cases}$$

Learning α_t^*

Fit α_t using online update

$$\alpha_{t+1} := \alpha_t + \gamma(\alpha - \text{err}_t)$$

err_t acts as an unbiased estimate of the current miscoverage probability

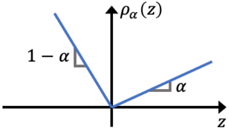
$$\text{err}_t := \begin{cases} 1 & Y_t \notin \hat{C}_t \\ 0 & Y_t \in \hat{C}_t \end{cases}$$

Connection to online learning

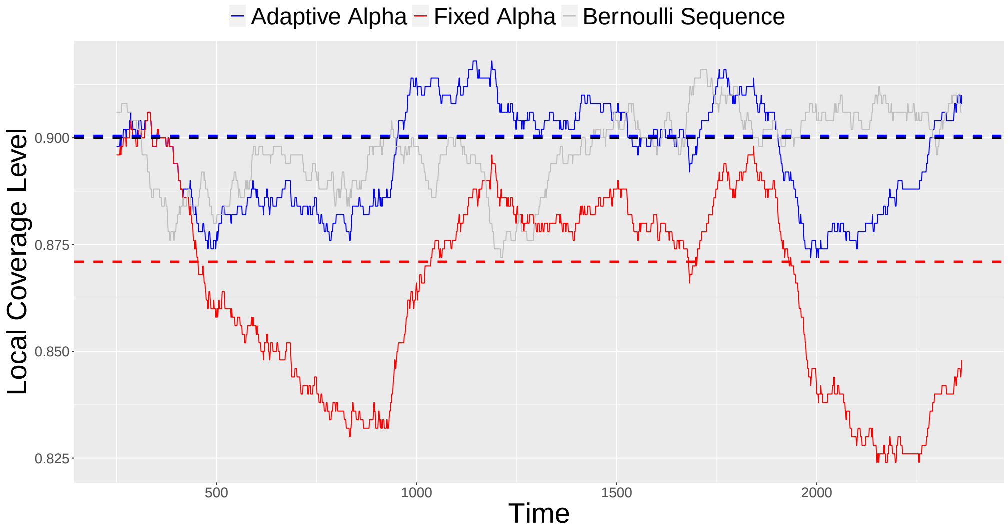
$$\beta_t := \max\{\beta : Y_t \in \hat{C}_t(\alpha_t := \beta)\}$$

Update can be reformulated as online gradient descent wrt. target β_t and pinball loss

$$\ell_\alpha(\alpha_t, \beta_t) = \rho_\alpha(\beta_t - \alpha_t)$$



Predicting county level election results: East to West



Distribution free theory

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \text{err}_t)$$

Under no assumptions on the data-generating process

$$\left| \frac{1}{T} \sum_{t=1}^{T} \text{err}_t - \alpha \right| \leq \frac{\max\{\alpha_1, 1 - \alpha_1\} + \gamma}{T\gamma}$$

and thus

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \text{err}_t \overset{a.s.}{=} \alpha$$

Additional theory re. $\mathbb{P}(Y_t \in \hat{C}_t) \cong 1 - \alpha, \ \forall t$ Gibbs and C. '21

Estimating volatility in the stock market

$$\text{Volatility} \qquad V_t = R_t^2 = \left(\frac{\text{Price}(t) - \text{Price}(t-1)}{\text{Price}(t-1)} \right)^2$$

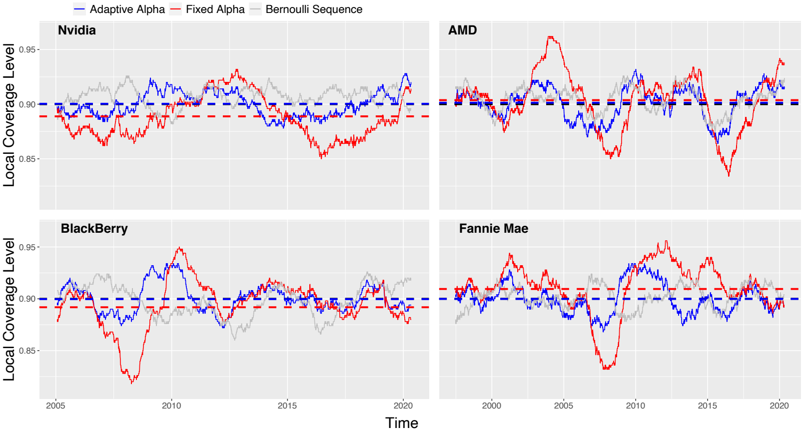
Use Garch(1,1) model to predict $\sigma_t^2 = \mathbb{E}[V_t | \ldots]$ and get prediction sets using conformity score

$$S_t := \frac{|V_t - \hat{\sigma}_t^2|}{\hat{\sigma}_t^2}$$

If the model was perfect S_t would be stationary...

Estimating volatility in the stock market

$$\text{LocalCov}_t := 1 - \frac{1}{500} \sum_{\ell=t-250+1}^{t+250} \text{err}_\ell$$



Accounting for unknown or changing shift size

Previous methodology/theory require γ *a priori*

Accounting for unknown or changing shift size

Previous methodology/theory require γ *a priori*

New algorithm:

1. Start with candidate gammas $\{\gamma^e\}_{e\in E} \quad \rightsquigarrow \quad \{\alpha_t^e\}_{e\in E}$
2. To judge α_t^e use past losses $\{\ell_\alpha(\beta_s, \alpha_s^e)\}_{s<t}$ to construct weights w_t^e
3. Output $\alpha_t := \sum_{e\in E} \frac{w_t^e}{\sum_{e'} w_t^{e'}} \alpha_t^e$

Accounting for unknown or changing shift size

Previous methodology/theory require γ *a priori*

New algorithm:

1. Start with candidate gammas $\{\gamma^e\}_{e \in E} \quad \rightsquigarrow \quad \{\alpha_t^e\}_{e \in E}$
2. To judge α_t^e use past losses $\{\ell_\alpha(\beta_s, \alpha_s^e)\}_{s<t}$ to construct weights w_t^e
3. Output $\alpha_t := \sum_{e \in E} \frac{w_t^e}{\sum_{e'} w_t^{e'}} \alpha_t^e$

Obtain w_t^e by

$$w_{t+1}^e := (1-\sigma)\exp(-\eta \ell_\alpha(\beta_t, \alpha_t))w_t^e + \frac{\sigma}{|E|} \sum_{e'} w_t^{e'}$$

Accounting for unknown or changing shift size

Previous methodology/theory require γ *a priori*

New algorithm:

1. Start with candidate gammas $\{\gamma^e\}_{e \in E} \quad \rightsquigarrow \quad \{\alpha_t^e\}_{e \in E}$
2. To judge α_t^e use past losses $\{\ell_\alpha(\beta_s, \alpha_s^e)\}_{s<t}$ to construct weights w_t^e
3. Output $\alpha_t := \sum_{e \in E} \frac{w_t^e}{\sum_{e'} w_t^{e'}} \alpha_t^e$

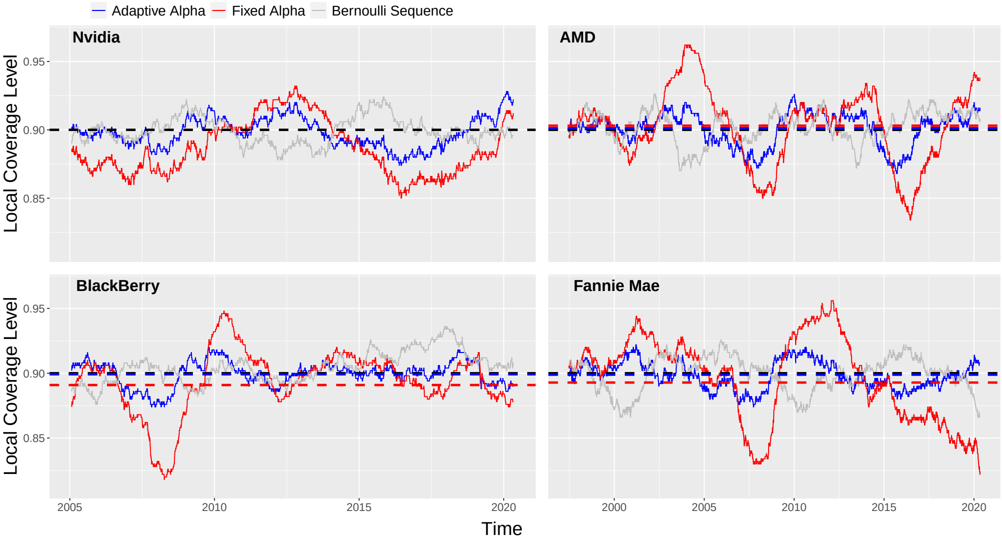
Obtain w_t^e by

$$w_{t+1}^e := (1-\sigma)\exp(-\eta \ell_\alpha(\beta_t, \alpha_t)) w_t^e + \frac{\sigma}{|E|} \sum_{e'} w_t^{e'}$$

Lots of theory Gibbs and C. '22

Returning to stock example

Results for new algorithm same as for gradient descent



Returning to stock example

In the stock market example used conformity score

$$S_t = \frac{|V_t - \hat{\sigma}_t^2|}{\hat{\sigma}_t^2}$$

and modelled $V_t \sim \sigma_t^2 \chi_1^2$ so hopefully $S_t \dot{\sim} |\chi_1^2 - 1|$ is stationary

Returning to stock example

In the stock market example used conformity score

$$S_t = \frac{|V_t - \hat{\sigma}_t^2|}{\hat{\sigma}_t^2}$$

and modelled $V_t \sim \sigma_t^2 \chi_1^2$ so hopefully $S_t \dot{\sim} |\chi_1^2 - 1|$ is stationary

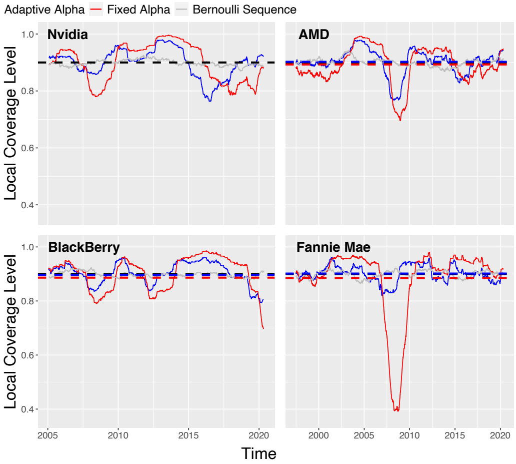
A bad idea would be to use

$$\tilde{S}_t = |V_t - \hat{\sigma}_t^2| \dot{\sim} \sigma_t^2 |\chi_1^2 - 1|$$

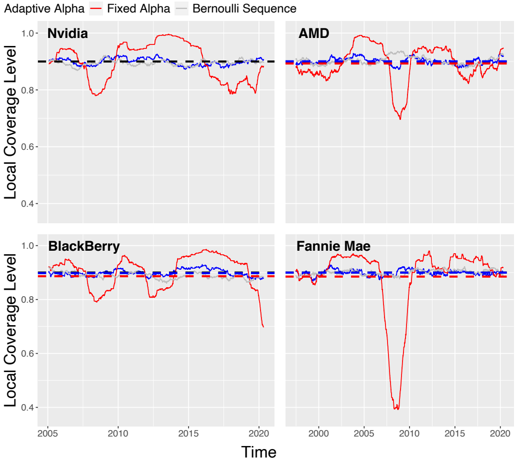
which is far from stationary

Results with "bad" conformity score

Gradient descent:



New Algorithm:



Summary

- New tools for uncertainty quantification
- No modeling assumptions whatsoever (except for exchangeability)
- Explosion of interest in academia & industry
  - Thousands of papers/year
  - Conformalized predictions in production at major tech companies
  - ...
- Resources available online

  Breiman Award Lecture at Neurips 2022 will exclusively feature conformal prediction