Confromalized Quantile Regression

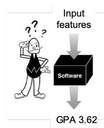
Yaniv Romano*, Evan Patterson*, Emmanuel J. Candés*†

* Department of Statistics, Stanford University
† Department of Mathematics, Stanford University

Reliability in high-stakes decision making

- Today's predictive algorithms (e.g. neural nets and random forests) show tantalizing performance in many high-stakes applications
- Predictive models are powerful but also complex and difficult to assess. Are the predictions really reliable?
- ML today: Who gets into college? Who gets paroled or makes bail?

Can we trust this?
$3.62 \pm$?



- Need reliable systems
- But still want powerful, complex models

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 joint distribution P_{XY})

- $X_i \in \mathbb{R}^p$: features (e.g. high-school GPA, SAT score...)
- $Y_i \in \mathbb{R}$: response (e.g. college GPA)

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

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

Must hold for any dist. P_{XY} (assumed unknown) and any sample size n.

"Based on the candidate's high-school GPA, SAT score, and other attributes, the college GPA is predicted to fall in the [3.4,3.8] range with 90% confidence"

Statistical efficiency

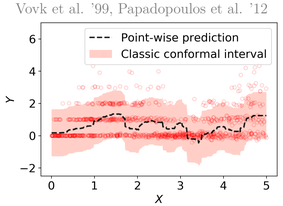
- Better predictions (via black box) $\implies$ shorter intervals!
  (So, want to use random forests, neural nets, and other complex models.)

"Candidate's expected GPA is in range $[3.4, 3.8]$*"*	**Vs.**	*"Candidate's expected GPA is in range* $[1.1, 4.3]$*"*
Informative		Uninformative

- For **heteroscedastic** data, to get *short* intervals, we need *adaptive* (variable-length) intervals.

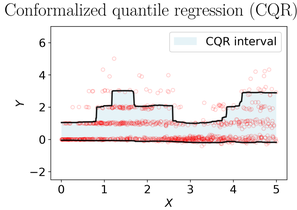
Fixed-length vs. adaptive prediction intervals

Conformal prediction
Vovk et al. '99, Papadopoulos et al. '12



Avg. Coverage 91.40%
Avg. Length 2.91

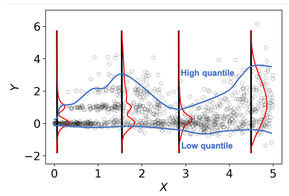
Proposed method
Conformalized quantile regression (CQR)



Avg. Coverage 91.06%
Avg. Length 1.99

Ideal setting: perfect knowledge

Conditional dist. $P_{Y|X}$ known $\implies$ perfect fit of upper and lower quantiles

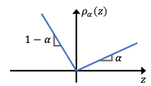


Practical setting: only finite samples from $P_{Y|X}$

Formulate quantile estimation as a learning problem:

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

- $\mathcal{R}(f)$ is regularization term
- ρ_α is pinball loss (Koenker & Bassett '78)



Validity on future data?

- No finite sample guarantees
- Asymptotic consistency ($n \to \infty$)?
  Only under regularity conditions and for specific models
  Zhou & Portnoy '96, Zhou & Portnoy '98, Takeuchi et al. '06, Meinshausen '06, Steinwart & Christmann '11

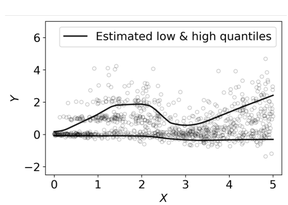
This work: quantile regression with finite-sample validity

Conformalized Quantile Regression (CQR)
Wraps any quantile regression algorithm (neural nets, random forests, etc) with valid coverage guarantee

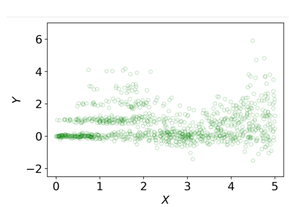
- In finite samples
- In any dimension
- Distribution (model) free
- Black box modeling

An implementation of CQR is available online at
`https://github.com/yromano/cqr`

Step 1: Split training data, then fit

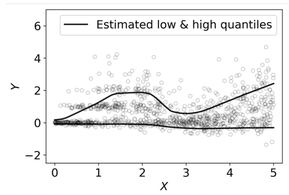


Quantile regression on proper training set

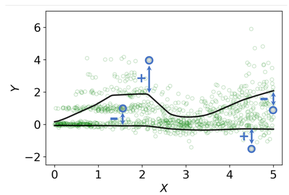


Calibration set

Step 2: Conformalization



Quantile regression on proper training set



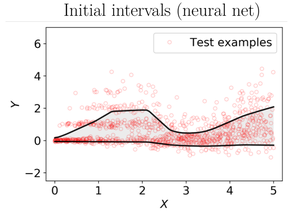
Conformalization on calibration set

$$E_i = \max\{\text{lower}(X_i) - Y_i, Y_i - \text{upper}(X_i)\} \quad i = 1, \ldots, m$$

Prediction interval: $C(X_{n+1}) = [\text{lower}(X_{n+1}) - Q,\ \text{upper}(X_{n+1}) + Q]$,
where Q is $(1-\alpha)m$-th largest value of E_i

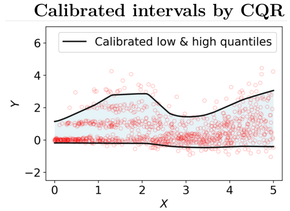
Validity on new data

Initial intervals (neural net)



Target coverage level: 90%
Actual coverage: 66.77%

Calibrated intervals by CQR



Target coverage level: 90%
Actual coverage: 90.09%

Validity of CQR prediction intervals

Theorem. If samples (X_i, Y_i), $i = 1, \ldots, n+1$ are exchangeable, then

$$1 - \alpha \leq \mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \leq 1 - \alpha + 1/(m+1),$$

where m is size of calibration set. Upper bound holds if scores E_i are distinct.

- Any joint distribution P_{XY}
- Any sample size
- Regardless of choice or accuracy of quantile regression estimate

Inspiration: Split conformal prediction

Vovk et al. '99, Papadopoulos et al. '12, Lei et al. '18

- **Classical (conditional mean) regression** on proper training set:

$$\mu(\cdot) = \underset{\mu \in \mathcal{F}}{\operatorname{argmin}} \sum_i (Y_i - \mu(X_i))^2 + \mathcal{R}(\mu)$$

- Conformalize using

$$R_i = |Y_i - \text{mean}(X_i)| \quad i = 1, \ldots, m$$

$$Q \text{ is } (1-\alpha)\text{-th quantile of } R_i$$

Prediction interval: $C(X_{n+1}) = \text{mean}(X_{n+1}) \pm Q$

- **Major limitation**: *interval width is fixed*, equal to $2Q$ (indep. of query point).

Related work: Locally weighted conformal

Papadopoulos et al. '08, Lei et al. '18
To handle heteroscedasticity, use scaled residuals

$$\tilde{R}_i = \frac{|Y_i - \text{mean}(X_i)|}{\hat{\sigma}(X_i)} = \frac{R_i}{\hat{\sigma}(X_i)},$$

where $\hat{\sigma}(X_i)$ is measure of residual dispersion; e.g., estimate the conditional MAD of $|Y_i - \text{mean}(x)|$ at $X_i = x$

Prediction interval:

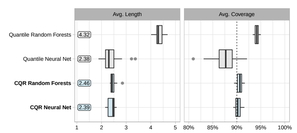
$$C(X_{n+1}) = \text{mean}(X_{n+1}) \pm Q\hat{\sigma}(X_{n+1}), \quad Q \text{ is } (1-\alpha)\text{-th quantile of } \tilde{R}_i$$

Major limitation: $\hat{\sigma}$ *underestimates prediction errors* since it uses residuals on training data.
In neural nets, overfitting is common, so that proper training residuals ≈ 0
$\implies$ forces Q to be large
$\implies$ intervals are less adaptive, hence longer

Application: Predicting usage 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 utilization, reflecting # visits to doctor's office/hospital
- $n \approx 16,000$ subjects, $p \approx 140$ features



- CQR achieves exact coverage while black-box does not!
- Competitive intervals (in fact shorter)

Further experiments

- 11 popular regression datasets
- Comparing classic split conformal, local conformal, CQR, and *non-conformalized quantile regression
- 20 random training-test splits (80%-20%)
- 10% target coverage rate
- Summarizing all 2,200 experiments:

Method	Avg. Length	Avg. Coverage
Ridge	3.07	90.08
Ridge Local	2.93	90.14
Random Forests	2.24	90.00
Random Forests Local	1.82	89.99
Neural Net	2.20	89.95
Neural Net Local	1.79	90.02
CQR Random Forests	**1.40**	**90.34**
CQR Neural Net	**1.40**	**90.02**
*Quantile Random Forests	*2.21	*92.62
*Quantile Neural Net	*1.50	*88.87

- Local conformal is better than classic conformal prediction
- CQR outperforms all methods, including the non-conformalized ones
- *Quantile Random Forest is too conservative
  *Quantile Neural Net fails to achieve valid coverage

Acknowledgements. E. C. was partially supported by the Office of Naval Research (ONR) under grant N00014-16-1-2712, by the Army Research Office (ARO) under grant W911NF-17-1-0304, by the Math + X award from the Simons Foundation and by a generous gift from TwoSigma. E. P. and Y. R. were partially supported by the ARO grant. Y. R. was also supported by the same Math + X award. Y. R. thanks the Zuckerman Institute, ISEF Foundation and the Viterbi Fellowship, Technion, for providing additional research support.