Reliable Predictions? Counterfactual Predictions?
Equitable Treatment?
Some Recent Progress in Predictive Inference

Emmanuel Candès

Distinguished Lecture Series, Harvard’s Institute for Applied Computational Sciences (IACS)
October 2021
Collaborators

Lihua Lei

Romano, Patterson & Sabatti

Barber, Ramdas & Tibshirani
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?
Miscalibration

For validation purposes, cannot assume ML models are well calibrated

Guo et al. (’17)
Machine Bias

There's software used across the country to predict future criminals.
And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016
This lecture

1. How much have we really learned from past data?
2. How do we communicate what we have learned without hurting people?
3. How much do we know about counterfactuals?
Data ethics 101: convey uncertainty and reliable outcomes

Why don’t we see prediction intervals more often?

\[ P\{Y \in C(X)\} \approx 90\% \]

What have we really learned from past data/experience of others?
Today’s predictive algorithms

random forests, gradient boosting

[Diagram of a decision tree]

neural networks

[Diagram of a neural network]

Breiman and Friedman

LeCun, Hinton, Bengio, and Rumelhart
Conformalized Quantile Regression

*Wrapper* around any predictive algorithm with perfect predictive coverage guarantees

Yaniv Romano  
Evan Patterson
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**

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

- Any dist. \(P_{XY}\) (assumed unknown)
- 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”
Setting with perfect knowledge

\( P_{Y|X} \text{ known} \quad \leadsto \quad \text{can fit upper and lower quantile functions} \)
Setting with perfect knowledge

\[ P_{Y|X} \text{ known} \iff \text{can fit upper and lower quantile functions} \]

Length of interval can greatly vary
No perfect knowledge, only a few samples from $P_{Y|X}$!

REGRESSION QUANTILES

BY ROGER KOENKER AND GILBERT BASSETT, JR.
Formulate quantile estimation as a learning task

\[ f(\cdot) = \arg\min_{f \in \mathcal{F}} \sum_{i} \rho_\alpha(Y_i - f(X_i)) + \mathcal{R}(f) \]

- \( \mathcal{R}(f) \) is a possible regularizer
- \( \rho_\alpha \) is pinball loss
Validity for unseen data?

Valid? No (imagine training a neural net)

Target coverage level: 90%; Actual coverage: 66.77%
Split

Proper training set

Calibration set
Fit

Apply quantile regression

Calibration set
Apply quantile regression

Calibrate
Calibrate: how?

conformity scores are signed distances: $V_i \triangleq \max\{\text{lower}(X_i) - Y_i, Y_i - \text{upper}(X_i)\}$
Calibrate: how?

\[ C(x) = [\text{lower}(x) - Q, \text{upper}(x) + Q] \]
Validity on **new** data

Target coverage level: 90%; Actual coverage: 90.09%
Correctness of predicted range

Theorem (Romano, Patterson and C. 2019)

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

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

- \(m\) is size of calibration set
- Upper bound holds if scores \(E_i\) are distinct

✓ Any distribution \(P_{XY}\)
✓ Any sample size
✓ Regardless of choice or accuracy of quantile regression estimate
Learning by Transduction

A. Gammerman, V. Vovk, V. Vapnik
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX, UK
{alex,vovk,vladimir}@dcs.rhbnc.ac.uk

–UAI ’98
Some pioneers

Vladimir Vovk

Jing Lei

Larry Wasserman
Classic conformal prediction  Vovk et al. '99, Papadopoulos et al. '12, Lei et al. '18

- Fit **classic regression** function on proper training set

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

- Calibration using

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

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

**Valid prediction interval:**

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

- **Major limitation:** width is fixed, equal to $2Q$ (indep. of query point)
Comparison to split conformal: random forests regression

Split conformal

- Avg. Coverage: 91.4%
- Avg. Length: 2.91

CQR

- Avg. Coverage: 91.0%
- Avg. Length: 2.18

CQR is adaptive while split conformal is not
Approx. conditional coverage and adaptive length

CQR is largely the right thing to do  
Sesia and C. (’19)
Adaptive predictive intervals

- Kivaranovic, Johnson, Leeb ('19)
- Chernozhukov, Wüthrich, Zhu ('19)
- Gupta, Kuchibhotla, Ramdas
- Romano, Sesia, C. ('20)
- ...

Extension to discrete labels: Romano, Sesia and C. ('20)
Predicting utilization of medical services

Medical Expenditure Panel Survey 2016

- $X_i$ – 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
- $\approx 140$ features
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, and Duchi ('20)
A more comprehensive study

Prediction intervals using quantile regression outperform existing conformal methods in 10/11 regression datasets
Counterfactual inference

Lihua Lei
Counterfactual inference

Assign treatment by a coin toss for each subject based on the propensity score $e(x)$

\[ \Pr(\text{treated} \mid X = x) = e(x) \]
\[ \Pr(\text{control} \mid X = x) = 1 - e(x) \]
Counterfactual inference

Each subject has potential outcomes \((Y(1), Y(0))\) and the observed outcome \(Y_{\text{obs}}\).
Counterfactual inference

How to infer $Y(1)$ of the circled person?
Counterfactual inference

Use observed treated units
The counterfactual inference problem and covariate shift
Adapting conformal inference to covariate shift

Goal: Use i.i.d. samples \((X_i, Y_i) \sim P_X \times P_{Y|X}\) to construct \(\hat{C}(x)\) with
\[
P(Y \in \hat{C}(X)) \geq 1 - \alpha \quad \text{with} \quad (X, Y) \sim Q_X \times P_{Y|X}
\]

**Covariate shift**
\[
w(x) \triangleq \frac{dQ_X}{dP_X}(x)
\]

**Counterfactual inference**
\[
w(x) \triangleq \frac{dP_{X|T=0}}{dP_{X|T=1}}(x) \propto \frac{1 - e(x)}{e(x)}
\]
Adapting conformal inference to covariate shift

Goal: Use i.i.d. samples \((X_i, Y_i) \sim P_X \times P_{Y|X}\) to construct \(\hat{C}(x)\) with
\[
P(Y \in \hat{C}(X)) \geq 1 - \alpha \quad \text{with } (X, Y) \sim Q_X \times P_{Y|X}
\]

**Covariate shift**
\[
w(x) \triangleq \frac{dQ_X}{dP_X}(x)
\]

**Counterfactual inference**
\[
w(x) \triangleq \frac{dP_{X|\tau=0}}{dP_{X|\tau=1}}(x) \propto \frac{1 - e(x)}{e(x)}
\]

Weighted Conformal Inference (Tibshirani, Barber, C., Ramdas '19)
Conformity scores are signed distances: 
\[ V_i \triangleq \max\{\hat{q}_{0.05}(X_i) - Y_i(1), Y_i(1) - \hat{q}_{0.95}(X_i)\} \]
Weighted conformal inference

Reweight distribution of conformity scores

\[ \sum_{i=1}^{n} p_i(x) \delta_{V_i} + p_{\infty}(x) \delta_{\infty} \]

\[ p_i(x) = \frac{w(X_i)}{\sum_{i=1}^{n} w(X_i) + w(x)} \]
Weighted CQR

Interval: $\hat{C}(x) = [\hat{q}_{0.05}(x) - Q(x), \hat{q}_{0.95}(x) + Q(x)]$
Theorem (Lei and C., 2020)

Set \( w(x) = \frac{1-e(x)}{e(x)} \) (\( e(x) \) known) in weighted conformal inference. Then

\[
1 - \alpha \leq \mathbb{P}(Y_{n+1}(1) \in \hat{C}(X_{n+1})) \leq 1 - \alpha + \frac{C}{n}
\]

- **Lower bound holds without extra assumption**
- **Upper bound holds if scores are a.s. distinct & an overlap condition holds**

- Applicable to randomized experiments with perfect compliance
- Holds approximately if either \( e(x) \) or \( q(Y(1) \mid X) \) are estimated well (double robustness)
Simulation: marginal coverage

- 100 covariates
- Smooth mean
- Heteroscedastic errors
- Smooth propensity score

![Empirical Coverage of Y(1)](chart.png)
Simulation: average interval length

Empirical Coverage of Y(1)

Average Length of intervals
Simulation: conditional coverage

Percentile of conditional variance

Conditional Coverage of $Y(1)$ (alpha = 0.05)
Conformal inference of individual treatment effects

Prediction interval for individual treatment effect $Y(1) - Y(0)$ of unseen individual

$$\mathbb{P}_{X \sim Q_X} (Y(1) - Y(0) \in \hat{\text{ITE}}(X)) \geq 1 - \alpha$$
Election Night at The Washington Post

John Cherian

Lenny Broner
Predicting county level election results

- Use demographic characteristics to predict normalized vote change

\[
\frac{\text{VotesBiden2020} - \text{VotesHillary2016}}{\text{VotesHillary2016}}
\]

in each county in the US

- Distribution shift: east coast counties are observed before west coast counties

Data not exchangeable! Reweight observed data to “match” the distribution of covariates in unseen counties?
Backtesting... from 2012 to 2016

- County coverage (CQR)
- State coverage (CQR)
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 50.6%
  - 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.
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.

Breaking down the estimates

Urban counties

Suburban counties

Rural counties

The Washington Post 5 November 2020, 12:50 AM
Addressing Distribution Shifts

Gibbs and Candès (2021)

Isaac Gibbs
Observing data stream 
\{(X_t, Y_t)\}_{t=0,1,...}

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 \(Y_t\)
Adapting conformal to distribution shift

Under exchangeability the empirical and true distributions will approximately align.
Adapting conformal to distribution shift

Distribution shift can cause the true distribution to shift to the right or left.
Adapting conformal to distribution shift

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

Coverage

Quantile

$1 - \alpha$
Adapting conformal to distribution shift

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

- If oracle knowledge, would use $\alpha_t^*$
- Key idea: learn $\alpha_t^*$
Learning $\alpha^*_t$

Fit $\alpha_t$ using online update

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

$\text{err}_t$ acts as an unbiased estimate of the current miscoverage probability

$$\text{err}_t := \begin{cases} 
1 & Y_t \not\in \hat{C}_t \\
0 & Y_t \in \hat{C}_t 
\end{cases}$$
Learning $\alpha^*_t$

Fit $\alpha_t$ using online update

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

$\text{err}_t$ 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 (online gradient descent)
- Connection to control theory (P-only controller)
Predicting county level election results

- Adaptive Alpha
- Fixed Alpha

![Graph showing local coverage level over time for Adaptive Alpha and Fixed Alpha.](image-url)
Estimating volatility in the stock market

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

<table>
<thead>
<tr>
<th>Company</th>
<th>2005</th>
<th>2010</th>
<th>2015</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nvidia</td>
<td>0.85</td>
<td>0.90</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>AMD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fannie Mae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BlackBerry</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bernoulli</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Time

Graphs of local coverage levels for different companies and time periods.
Estimating volatility in the stock market
Equitable Treatment?

With Malice Towards None:
Assessing Uncertainty via Equalized Coverage

Yaniv Romano*  Rina Foygel Barber†  Chiara Sabatti* ‡ Emmanuel J. Candès* §
Growing pains

Machine Bias
There's software used across the country to predict future criminals.
And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016
Growing pains

Design AI so that it’s fair

Identify sources of inequity, de-bias training data and develop algorithms that are robust to skews in data, urge James Zou and Londa Schiebinger.
On the use of ML to support important decisions

- How do we communicate uncertainty to decision makers?
- How do we not overstate what can be inferred from the black box?
- How do we treat everyone equitably?

**Our take:**

Decouple the statistical problem from the policy problem

Corbett-Davis and Goel, '19

Somewhat against current thinking in “algorithmic fairness in ML”
Current thinking in “algorithmic fairness in ML”

Tendency to position algorithm as decision maker

- Statistical parity
- Calibration (used by COMPAS)
- Equalized odds Hardt, Price and Srebro ’16

Review
Dwork et al. ’12, Chouldechova ’16, Berk et al. ’17, Zafar et al. ’17

Incompatible
Chouldechova ’17, Kleinberg, Mullainathan and Raghavan ’17
RECONCILING LEGAL AND TECHNICAL APPROACHES TO ALGORITHMIC BIAS

By Alice Xiang*

This Article Has Been Accepted for Publication in the Tennessee Law Review, Volume 88, Number 3 (Spring 2021).
Equalized coverage

Goal: construct perfectly calibrated intervals across all groups

\[ \mathbb{P}\{ Y \in \hat{C}(X) \mid A = \text{♂} \} \geq 90\% \]
\[ \mathbb{P}\{ Y \in \hat{C}(X) \mid A = \text{♀} \} \geq 90\% \]

Summarizes what we have learned from ML s.t.

- Rigorously quantifies uncertainty
  - Honest reporting: interval is long? \( \leadsto \) model can say little
- Treats individuals equitably
Joint training + separate calibration
Predicting utilization of medical services

MEPS 2016 data set

- $X_i$ – 140 features including 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, ...

- $\approx 9,600$ non-white individuals

- $\approx 6,000$ white individuals
Some observations on 2016 MEPS data set + fix

Fit a neural network regression function:

- NNet overestimates the response of the non-white group
- NNet underestimates the response of the white group
Some observations on 2016 MEPS data set + fix

Fit a neural network regression function:
- NNet overestimates the response of the non-white group
- NNet underestimates the response of the white group

<table>
<thead>
<tr>
<th>Method</th>
<th>Group</th>
<th>Avg. Coverage</th>
<th>Avg. Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>CQR (joint train.)</td>
<td>Non-white</td>
<td>0.902</td>
<td>2.527</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>0.901</td>
<td>3.102</td>
</tr>
</tbody>
</table>

Average across 40 random train-test (80%/20%) splits
• Effective conformity scores: https://sites.google.com/view/cqr/
• Counterfactual and individual treatment effects: https://lihualei71.github.io/cfcausal/index.html
Summary

*How much have we really learned from the experience of others?*

- **Reliable and informative predictions?**
  conformalized quantile regression

- **What if?**
  conformalized counterfactual inference

- **Equitable treatment?**
  equalized coverage

- **Small datasets & data reuse?**
  cross conformal  Vovk et al. '18
  jackknife+/CV+ Barber, Candès, Ramdas and Tibshirani '19